Direct Selling Business Lead Prediction by Social Media Data Mining

by

Ahmed Balfagih

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

Dalhousie University
Halifax, Nova Scotia
April 2016

Table of Contents

List of Tables

List of Figures

Abstract

Business leads are new potential customers and networkers from the direct selling business point of view, which are the marketing backbone of the direct selling industry. People who work in the direct selling business always want to enrich their contacts to promote their business. Today, with the huge increase of using internet technology by most people in the world, and with their activity information available on social media websites, it is possible to discover more suitable models for predicting potential people to contact. This thesis investigates some suitable data mining solutions for building a business lead prediction system framework over available social media data to suggest new potential customers and agents for supporting direct selling business. The information on Facebook friends' list provides the networkers with business leads to help them to promote direct selling marketing. This research uses Facebook transactions as a case study for social media based lead prediction data mining because of its wide global usage. A set of data mining methods and algorithms are investigated and compared in determining the most suitable option based on feature analysis and selection of the social media data. Extensive experiments demonstrate and justify the proposed lead prediction system framework for supporting direct selling marketing promotion.

List of Abbreviations and Symbols Used

B2B	Business 2 Business
B2C	Business 2 Customer
IR	Independent Representative
DSA	Direct Selling Association
MLM	Multi-Level Marketing
NB	Naïve Bayes algorithm
RF	Random Forest algorithm
KNN	K-Nearest Neighbor algorithm
SVM	Support Victor Machine algorithm
BF	Best First search method
GA	Genetic Algorithm search method
GS	Greedy Stepwise search method
3C	Three Classes dataset
2C	Two Classes dataset, with lead and non-lead classes
2CP	Two Classes dataset, with Partner and non-lead classes
2CC	Two Classes dataset, with Customer and non-lead classes
FS	Feature Selection
MFS	Manual Feature Selection
T3C	Transformed dataset with Three Classes
FT3C	Factor Table dataset with Three Classes

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Qigang Gao of the Computer Science Department at Dalhousie University. The door to Dr. Gao office was always open whenever I had a question about my research or writing, and for his understanding, elegant treating, useful comments, remarks, engagement through the learning process and support of this master thesis.

I would also like to acknowledge Dr. Hai Wang of the Business School at Saint Mary's University and Dr. Peter Bodorik of the Computer Science department at Dalhousie University for being the readers of this thesis, and I am gratefully indebted to them for their very valuable comments on this thesis.

Furthermore I would like to thank the participants in my survey. Without their passionate participation and input, the validation survey could not have been successfully conducted.

I must express my very profound gratitude to my parents, Mustafa and Maha Balfagih, and to my beloved wife Dania Kherd, for providing me with unfailing support, and continuous encouragement throughout my years of study, and through the process of researching and writing this thesis, with their patience. This accomplishment would not have been possible without them. I dedicate my success to them and I hope they are always proud of me.

Finally, I would not forget to thank King Abdullah bin AbdulAziz, the former King of Kingdom of Saudi Arabia, who lead the revolution of education and scholarships in my country, and who give the opportunity of conducting my graduate studies in Canada, with his unlimited material and moral support. May his soul rest in peace.

Author

Ahmed Mustafa Balfagih

Chapter 1 **Introduction**

1.1 Background and Problem Statement

Direct Selling (or network marketing) is a type of marketing of special products creating a social network of distributors at multiple levels of profits. This is an alternative to traditional marketing. Direct selling is mainly important for its advantage of saving the costs of advertising for the company by taking advantage of peoples' word of mouth marketing, and sharing the profits with these people [1] [2]. Figure 1.1 shows the difference between the traditional retail model and the direct selling model.

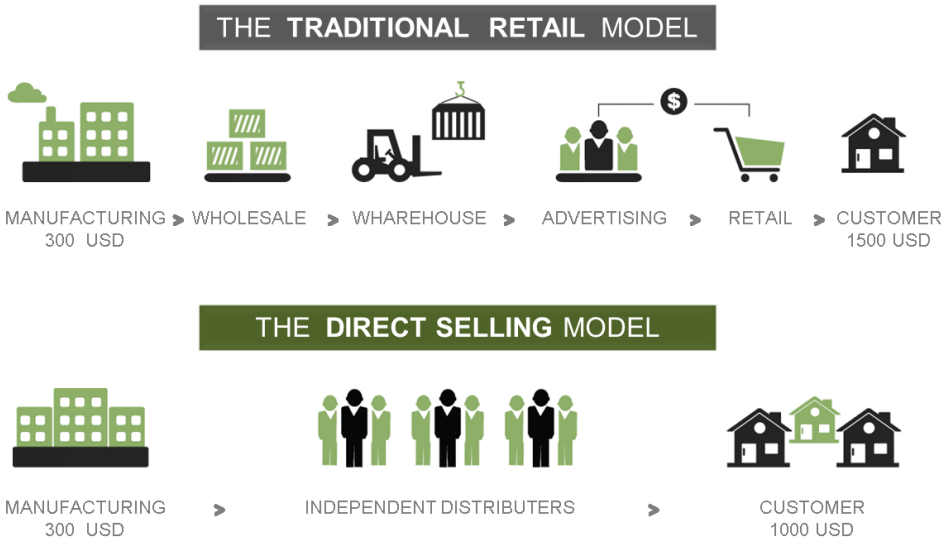


Figure 1.1 Difference between the traditional retail model and the direct selling model.

The levels of marketing in direct selling take different shapes depending on the company's policy of sharing profits. The difference between the traditional marketing plan and direct selling is that the middleman in direct selling gets paid by the manufacturer (the company) as a fixed percentage of the cost of selling the product, not from the customer as in traditional marketing. Where the customer buys from the company directly, probably more than one middleman will get paid from the company if they are qualified to earn the commission by reaching the target of the direct selling system. (Figure 1.2).

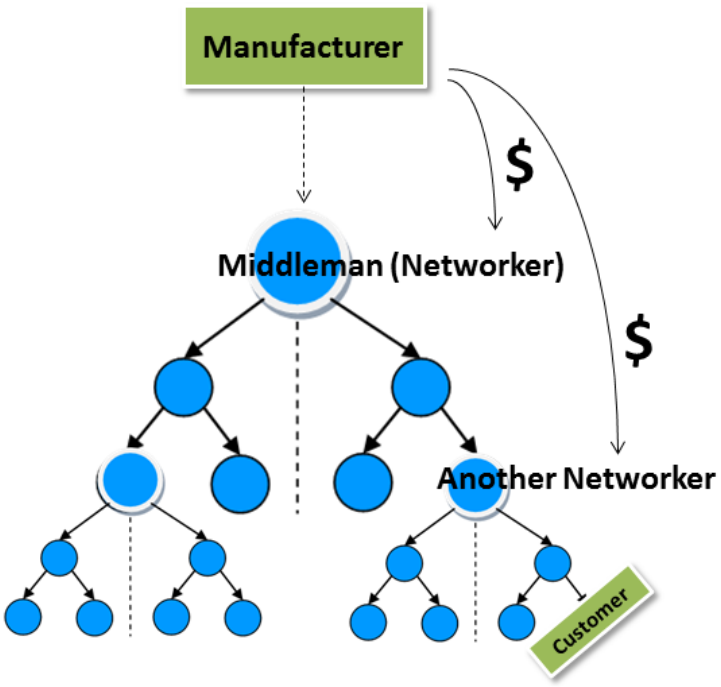


Figure 1.2 The manufacturer pays the independent agents depending on the targets they achieve that belong to the manufacturer's rules and policies in a certain pay plan, and the figure shows an example of a binary plan, which is one of the most popular plans in direct selling solutions.

In recent years, the internet medium has become more essential for the direct selling industry, and according to the E-Commerce field [3]. This type of business has two phases:

- Customer to Business (C2B): business here is represented by an independent representative (IR) of the company (the promoter or the networker), and the customer is the regular consumer of the product.
- Business to Business (B2B): The first "B" represents the independent representative (IR) and the second "B" represents the company.

The direct selling philosophy is based on word of mouth. Since a customer purchases a product from a direct selling company, (possibly through a promoter's word of mouth), he also has the choice to represent the company by promoting the same products or services to other people, and receive a share in profits. The direct selling industry is also known as *Network Marketing* because this business is based on a network of relationships and network of sales, and people who work in this field are often called *Networkers* [2][4].

Most of the entrepreneurship experts find this business opportunity is a good way to build a business with the least amount of capital and gives the networkers the ability to have financial freedom in the future [1]. Success in direct selling requires the networker to build a team of promoters in the same network organization who share similar goals, have a system of meetings and training development in order to assist each other to achieve his/her goals as part of the team. Always in direct selling, there are three parties participating in the selling process: the direct seller (or the networker) who promotes the product, the direct selling company that manufactured the product and the customer who buys the product.

The direct seller (the networker):

Networkers choose this type of business because it allows them to build a new business with a small amount of capital, because of its flexibility in managing the way of selling, and meeting the team members. It is all about freedom. When they choose which direct selling company they want to work with, they consider many things such as: the type of product that they will promote and its needs; the product quality; the product exclusivity; the efficiency of the profit plan; and the company's support to them [1] [2].

Different direct selling companies try to gain networkers' loyalty by giving them the diversity of income plans and offering the best supporting services, which include the system of training and teaching, logistics support and online support tools. The online support systems are a very important issue for networkers who care about the systems' security, its productivity and its quality.

Direct selling company's needs:

The manufacturing companies that work in the direct selling field need to remain stable, gaining more loyal networkers. The more networkers they deal with, the more sales are conducted, which means more profits, and that is what guarantees the direct selling companies' continuation [1]. The main principle that these companies stand for is that they have a unique service or unique product that they own exclusively, and they want to sell it to the world through a word of mouth network and without a budget for advertising. Thus, they are always seeking those loyal networkers they pledge to train, educate and support to

bring the company more sales and customers. One of the biggest challenges that different direct selling companies face is providing online support tools to their networkers.

The customers in direct selling:

Customers' interaction differs in direct selling depending on the company and the promoters. Some customers are purchasing from direct selling companies because of their need for the product, while other customers are attracted by the way that the promoter promotes to them [1] [2]. Some prospective customers are losing trust in this kind of business due to their belief that the purpose of the company and promoter is just to sell and earn, without considering the value of the product and the quality of the online system. Moreover, the lack of trust is a result of confusion between two concepts; the direct selling and some pyramid schemes which represent a form of illegal scam business [1].

There are significant differences between direct selling and pyramid schemes, of which the most prominent are:

- **The product:** there is a valuable and useful product for the customer in direct selling while there is no real product in the pyramid scheme because it depends only on bringing people to the scheme and sharing the profits without selling any products, and that leads to the company closing.
- **The levels of profits:** always the upline (the referrer) earns more than the downline (the referral) in the pyramid scheme, which allows the new referral people to earn by their effort, sometimes more than the early promoters.
- **The legality:** Direct selling is a legally recognized business practice, unlike the pyramid scheme which is a criminalized business in the majority of countries around the world.

Qualifying a new network recruit is crucial before he/she practices sales to the customer. Understanding the history of direct selling, the differences between it and pyramid schemes, and the product information enables promoters (networkers) to sell better and make them skilled enough to promote more products. Customers who have bought the product can become promoters later if they tried the product and liked it.

Finally, the customer who is potentially a new recruit, is looking for a good product, good service, a good business plan and a good on-line system to trust in the direct selling company, if he/she is willing later promote.

The problem to be tackled:

A prospective list of potential customers is the basic tool that the networker relies on to succeed in the direct selling business. Since the networker has as many prospects as possible and has an organized way to add more prospects to his/her name list, he/she can proceed more confidently in this business. Social media is taking an important role in peoples' communication and is considered the widest platform through which companies are looking for customers. Facebook has more than one billion members [5], who are communicating and networking naturally and sharing their interests and preferences. Twitter, moreover, has a wide popularity and can be a good area to look for new prospects (over 300 million) [6]. Any social media website can be used to find prospects, but Facebook and Twitter are more likely to be used due to their sweeping popularity. Therefore, an ideal source for the networkers to find potential customers is social media. The goal is to have a data mining solution to help networkers to predict leads to increase their name list to invite more qualified people to the business through Facebook, and to build their marketing team easily. This proposed tool solution is very beneficial, especially for new networkers and those who do not have enough experience in the field of direct selling, to target the more likely people from their acquaintances to join their business and to reduce the rejection rate from non-interested people.

1.2 Current Solutions and Challenges

Previous studies designed special data collection tools for obtaining information about consumers' purchasing and online shopping behavior for a certain product to find patterns in direct sales leads. According to my research, there are no tools for direct selling leads prediction in social media currently available. Networkers who want to take advantage of social media search manually in social network websites and look for its users' attributes. The main challenge is the huge database of the Facebook website. This issue can be handled

by reducing the database to focus primarily on the friend list. Another challenge is to have a cleaned and organized dataset from Facebook to be applied to design a data mining model that predicts leads.

1.3 Proposed Solution and Research Contribution

The primary goal of this thesis is to design and build a data mining tool that generates leads for the networker from his/her Facebook friend list to join his/her direct selling network. Facebook is the best social medium to apply the system to because of its popularity and the abundance of information on the users' profiles which leads to more accurate results. Furthermore, Facebook may in the future develop an app in the same social network media through (Facebook Apps Center). The system also could be expanded to include the whole Facebook database to find more direct selling leads from outside the friends list and acquaintances circle. Networkers are also a part of this broad global social network, and by virtue of their work they are always seeking two types of people:

- **Customers:** prospected people who are more likely to be interested in the exclusive products that the networker promotes. They purchase the product due to their need for the product and their desire to use it.
- **Team Partners:** prospected people who are more likely to work with the networker to promote the same products. They join his network, become a networker like him and receive the same support from the company. They may be customers at the same time.

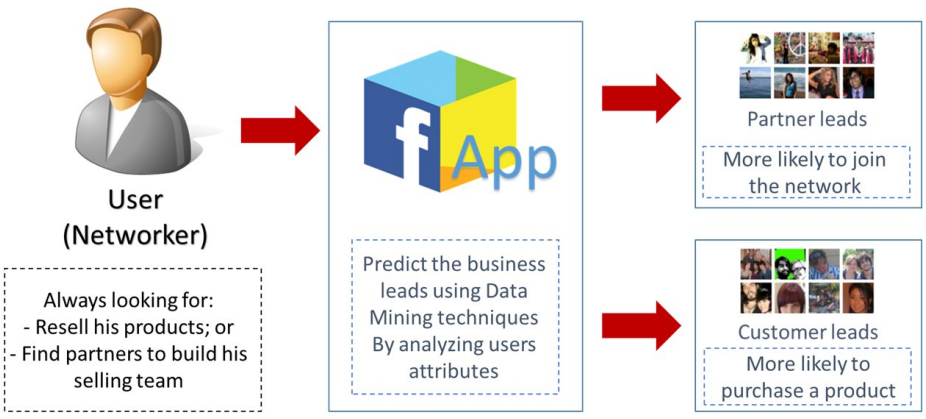


Figure 1.3 The role of the application is to predict potential customers and partners (people) to the networker by a data mining technique and analyzing their attributes.

Facebook contains these two types of people who the networker is always seeking. There can be overlap of the two types because customers can also be partners. Figure 1.3 (above) shows briefly how the proposed system can work.

My primary research contributions are: (a) to compare different classification methods and evaluate its performance; (b) to develop an ensemble classification model that aims to assess performance improvement, thus leading to consider the best classification method for this study. Other contributions are: selecting the best features of the potential customers and partners from the Facebook dataset, and evaluating the classes' modeling quality and its role in generating leads.

The expected impact of the application:

The direct selling leads generation tool on Facebook has a possible impact on the parent company and on the networker specifically, especially the new networkers or those who have a lack of experience, to target the most likely people to join a direct selling business. The networker relies on finding new prospects in order to succeed and perform well in his marketing business. The presence of a sufficient number of prospective customers ensures continuous marketing activities, and is sine qua non to ensuring the earnings possibilities the networker is looking for. Furthermore, having more prospective partners helps to expand business activity, so that it will abbreviate the effort in reaching the appropriate people who can duplicate the number of working hours and achieve faster results. The goal of this program is to help any networker to find new people and to perform better in the direct selling business.

The company can ensure its success in the direct selling business world if the networkers are doing their work easily and in a continuous way. Any direct selling company will exploit the advantage of having a leads generation system for direct selling connecting to Facebook, and the system will probably be considered one of the essential tools in the electronic direct selling concept. Companies could compete to improve this system and redesign the system in different ways to match their products and goals.

The prospective customer who is interested in joining the business cares about what features and tools that the direct selling company can provide to the new networker to succeed in his/her business.

Having this leads generation system can affect the decision of joining the business, and it would be useful because it differentiates companies that have the Facebook leads generation system from the companies that do not have it yet. It makes new networkers feel the simplicity of starting the direct selling business with a broad slice of targeted people who are qualified, prospective customers.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 presents the background studies and literature review;

Chapter 3 presents a preliminary system analysis and data preprocessing, including feature selection;

Chapter 4 presents the used data mining methods in this study for lead prediction;

Chapter 5 presents the experimental results and evaluations; and

Chapter 6 presents the conclusions and future work.

Chapter 2 **Background and Literature Review**

2.1 Direct Selling

2.1.1 Direct Selling Authenticity

The reliability and credibility of the direct selling companies are some of the most important topics of concern to marketers and customers who deal with these companies. Many illegal corporate and pyramid scams try to delude people with their legitimacy, and coverage of their activities under the name of direct selling. Whereas, the real direct selling companies accelerate in documenting their activities in different legal ways to preserve a reputation and the reputation of the industry [2]. Some company activities are monitored by authorized organizations with broad powers to ensure consumer rights such as the Federal Trade Commission (FTC) in the United States, or the Financial Conduct Authority (FCA) in the United Kingdom. Some companies seek to offer its shares for trade on different international stock exchanges, such as NASDAQ, or Over the Counter (OTC), which has high standards for accepting trading on its platforms, making these companies as public joint-stock companies, and all their information is visible to the shareholders and the consumer. Other direct selling companies record their memberships under associations in the direct selling industry that grants certification to the legal corporations which is compatible with its clauses and standards according to the international laws on the practice of direct selling, such as the Direct Selling Association (DSA), and the World Federation of the Direct Selling Association (WFDSA) [2] [7] [8].

The DSA is an example of the trade association that gathers the leading and trusted companies of direct selling, and is subject to a high standard in its code of ethics and credibility. The association contains around 200 members, including many of the well-known companies in the industry. The idea of DSA exists locally in some countries to include many companies that are local or regional in nature [7].

The mission of the association is, "To protect, serve and promote the effectiveness of member companies and the independent business people they represent. To ensure that the marketing by member companies of products and/or the direct sales opportunity is conducted with the highest level of business ethics and service to consumers" [7].

The direct selling model requires a commitment to practice ethics and customer service as a basis of its code of ethics. Every member company in the association pledges to commit to the code's standard as a condition of acceptance to the membership and to the continuation of its membership. Direct selling company membership in the DSA is considered a source of trust for customers to deal with the company, and this is a big advantage to the company for the sake of its expansion and to achieve higher targets of sales. DSA guarantees the legality of the direct selling company and protects consumers' rights [7] [8].

2.1.2 Direct Selling and Pyramid Schemes

There is frequently confusion between the direct selling systems and illegal pyramid schemes. The misconception refers to the close similarity between both in the idea of marketing through social relationships. However, many people ignore the fundamental differences between them, and that affects the general reputation of the direct selling industry [9].

One of the essential roles of the DSA and member companies is to differentiate legitimate direct selling companies and illegal pyramid promotional schemes. First, in legal direct selling companies the promoted product or service is real, used and consumed, and compensation relies on those sales and consumption by the end-user. On the other hand, pyramid schemes' product or service - if any really exists - is not used or consumed by anybody; rather, income is made from the mere act of recruiting new participants into the scheme. Therefore, the main focus of a direct selling company is product distribution. In fact, in a legitimate direct selling company, distributors are not required to recruit new distributors in order to earn the commission; they can earn money purely by selling the company's product.

The second way to recognize pyramid schemes is that you do not get income unless you have successfully added a number of new recruits into the pyramid, while direct selling systems pay money for sales and repurchasing of the consumable services and products. Basically, pyramid schemes concentrate on the money that you could earn by recruiting new individuals into the pyramid and generally ignore the advertising and selling of any products or services.

The U.S. Federal Trade Commission (FTC) has regulatory power over numerous U.S. business activities, including direct selling. That power has been used to set anti-pyramid standards and has been instrumental in deciding the business standards used by legal multilevel companies in the United States [7].

2.1.3 Different Compensation Plans

Direct selling companies are similar in their concept of shortening the chain of traditional marketing and sharing the profits with regular people. Nevertheless, there are some differences in the compensation plans and the systems of payment. There are many plan models, and each plan has some pros and cons. The most popular are three: the binary plan, the unilevel plan, and the matrix plan.

2.1.3.1 Binary

From its name it indicates two "pay groups", one on the right of the business tracking center, and the other on the left; each group is typically called a "leg", which makes up the core structure of this plan (See figure 2.1). This network is organized by bringing in at least two partners participating in the binary network. And additional direct partners joining the network will be "spilled over" on the right leg or on the left leg under the previous partners. The spillover feature is very good to support the team members on the right leg or left leg and increase their sales points in one of their sides, and that encourages the team working mood among the team members [10].

Payment to the networkers depends on the balance of the volume of sales between their two legs, either by pooling the sales point or recycling it (re-purchasing consumable products). The advantage of this plan is this pool, which pools the total group volume of each leg through infinite levels. Early binary plans struggled with paying out too much, and that may cause a company collapse. For that reason, binary plans should have "flush-out" security (in other words maximum limit to pay within the commission period). Furthermore, many innovations occurred to binary plans like combining them with a unilevel system to attract people to resell their

products on their network periodically, and this makes more sales' points on both right and left legs.

The binary became a big trend in the 1990s and had developed into a common place compensation plan. This plan is presently in high demand in Asian countries because of its high income in comparison with other plans. However, a low percentage of direct sellers are considered as successful in this type of direct selling plan due to the difficulty of balancing the two legs. That needs permanent follow-up from the networker to his key team members on both sides, motivating them, training them, evaluating their results and working on enhancing their income. It is regularly either loved or hated among industry experts (there are few who have neutral feelings about the plan). The plan has been used successfully by many of highly successful companies, demonstrating its longevity and staying power [10].

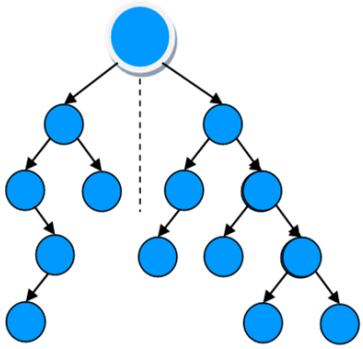


Figure 2.1 The binary compensation plan structure

2.1.3.2 Unilevel

This plan is named unilevel because all sales and enrollments are put in one level belonging to the same business tracking center. In contrast to the binary plan which has only two legs to construct the network, unilevel has infinite possible legs to form. This is the most common compensation structure, and it is also the least controversial (figure 2.2). Most of the unilevel plans are based on consumable products that the customer can purchase periodically. The networker gets paid by high percentage for his first level, and a lesser percentage for the next levels, and thus, until reaching to a specified level where the company does not pay for its sales.

Networkers who apply the unilevel plan are concerned about the limited depth of the paid levels, and concerned regarding teamwork as key problem areas for the unilevel. Some networkers like this plan because of its structural simplicity and ease of explanation, while other networkers mention paying people fairly as a reason for their love of the plan [10].

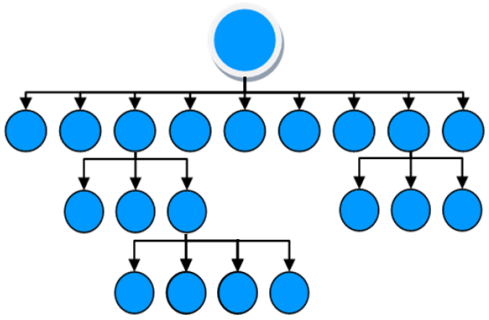


Figure 2.2 The unilevel plan firm

2.1.3.3 Matrix

The Matrix Compensation plan has a fixed structure. The key feature is that it has limited width and limited depth. The Commission given on a level by level basis - either a percentage of sale prices, or a set dollar amount given for each sale made. Proponents of the matrix plan love the 'teamwork' model. They believe it builds as sales teams are forced to work in a structured model [10]. For example, in a 3X matrix, there would only be three positions available on the first level (the front line). The fourth person enrolled would be placed on the first partner enrolled. This 'spill over' effect is believed by matrix supporters to stir up excitement and help others to start their businesses. The matrix plan is based on filling the gaps strategy. This is not recommended for non-active people in marketing, and conversely, there could be a little production of new sales or enrollments, which may be a concern for the independent marketers who make big commissions.

As a result of these concerns, developments in matrix plans have led to implementing hybrid bonuses for example generational, infinity, matching, and/or coded bonuses, to make it possible for active networkers in deeper levels of the

matrix structure to be paid. Some companies even offer "expanding matrix" formations that present new front line positions at set qualification levels (some might place this type of plan in the unilevel category). This feature was designed to eliminate concerns about the limited income of a single networker at any level. This plan is nested; as one matrix fills another matrix is started. The matrix is here to stay as a common structure in direct selling. It is also a polarizing compensation plan. Like the binary, people typically love or hate a matrix [10].

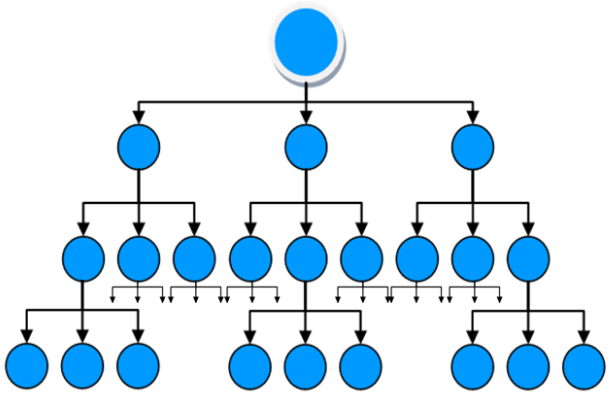


Figure 2.3 The matrix plan firm

2.1.4 Business Preference against Traditional Marketing

Direct selling offers a realistic and reasonable opportunity for starting a private business with a minimum amount of capital (and usually that includes the price of the product, or the service promoted, in addition to personal website subscription fees). In comparison with traditional business, there is a huge reduction in the costs of starting a new business such as what traditional businesses require (i.e. renting a showroom or shop, warehousing products, shipping products, advertising, hiring employees and paying for them). While in direct selling, people can manage their business anywhere, anytime, which offers them more flexible options and smarter solutions to make income more easily [2]. (Figure 2.4).

In his book, *Rich Dad and Poor Dad*, Robert Kiyosaki, the American investor and financial literacy activist defines four types of people relying on their work and income in the "Cashflow Quadrant" diagram. He explained how people are aiming to achieve their

financial freedom in a regular way and spend many years to do that, while the direct selling business is representing the shortcut path for those who desire financial freedom in a couple of years [11]. Other famous and influential public figures such as Bill Gates, Donald Trump and Warren Buffet agreed with Kiyosaki‘s mission and encouraged people to join the world of direct selling and to start their journey [12].

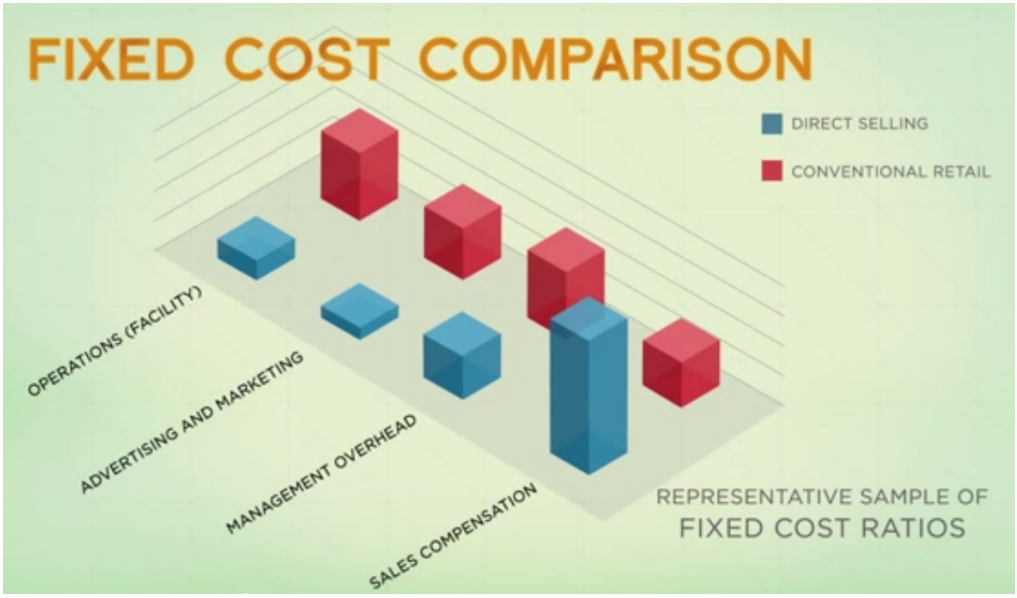


Figure 2.4 Comparison in costs between direct selling and traditional business [1].

Furthermore, the former president of the United States, Bill Clinton, claimed that direct selling is strengthening the economy and lessening the averages of unemployment, "You strengthen our country and our economy not just by striving for your own success, but by offering opportunity to others." [13].

2.1.5 Electronic Direct Selling

Internet technology has been transforming the direct selling industry to a different stage. It allows promoters to spread their network globally and to take advantage of social media. Doing business from home in direct selling becomes easier because the company is providing an integrated online direct selling system. (Figure 2.5).

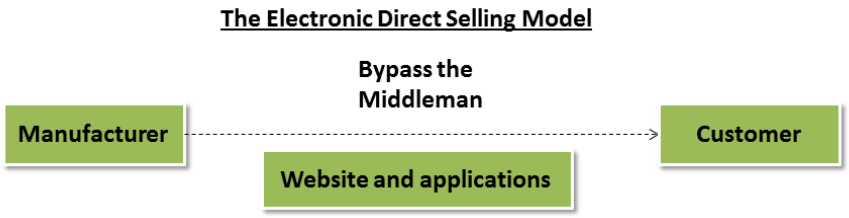


Figure 2.5 The direct selling environment with the electronic medium

Using the advantage of social media websites can provide the people who work in direct selling a wide opportunity to reach more prospective people. As an example of an electronic direct selling company model, DubLi (one of the successful direct selling companies), is a good example for many reasons:

- They were one of the first companies that started 100% online based
- Their services cover most countries
- They have a variety and wide range of products and services that can match different preferences and needs.
- They have integrated online features [14]; besides their website they have:
    - The back office: where the BA (Business Associate) can manage his/her business virtually, withdraw his commissions, review his network flow and have direct communication with the main company with high security.
    - The mobile services: Their applications cover most popular mobile devices types and tablets.
    - The landing page: a customizable interface or virtual showroom that enables the BA to use it in e-marketing, and the customer can reach directly to the online promoter through it because it's linked to the mobile phone and email. The website also is embedded in social networks including many promoters around the world so it could be a good way to share experiences. It can help in reaching potential new customers.
    - DubLi TV: is a service for online training and live or recorded webinars around the world.
    - The different social media channels that belong to the company, specially YouTube channel, Facebook page, Twitter account and Google+.
    - The Google Chrome tool bar add-on: a smart component that interact with

Google search engine heuristics to help in marketing.

- Recruiting center: an additional paid service, which is a tool to recruit, register and follow up with new prospective partners on the user behalf.

All these features enabled DubLi to be a global direct selling company and one of the fastest growing companies worldwide, and to encourage the BAs to build their networks easily and think globally. This is making this company a good example to study, in order to develop and enhance electronic direct selling. Most good direct selling companies have similar but distinctive online services [14].

Another direct selling company, WORLD GN (Global Network), specializes in the telecommunication industry and renewable energy. WORLD GN has more than four years in the industry, and they could prove their success by their highly needed exclusive services and their integrated online supporting tools. Such as DubLi, they have many online supporting tools. One of the most interesting tools they have is the video conference room for each networker, which enables them to do their long distance meetings easily and share their screen with others, with a lot of supporting tools in the same room such as the on-line whiteboard. Also, networkers can communicate to each other worldwide through their smartphones by an application called (SpaceVoIP), which makes the permanent communication easier among them and maintains the efficiency of their direct selling business [15].

2.1.6 Leads Generation

Leads generation is a term commonly used in marketing electronic marketing that means attract potential customers; it describes how to attract the interest of the consumer or his attention to the products or services for any company. It can be done to achieve a variety of purposes, including for example: create lists of sales, or listing eNewsletter list, or acquire customers. Companies seeking to attract potential clients, and usually determined by the quality depending on the degree of potential customer tendency to take the next step towards a purchase. The term also refers to the discovery of prospective sales agents who can be converted to be actual clients of the company [16].

It is possible to reach to potential customers or agents from a variety of sources and activities, including, for example, through online, calls, advertisements, and lists of purchases. Also, companies can rely on the recommendation of existing customers and telemarketing and advertising to attract potential customers.

Leads generation in direct selling is one of the most important activities while it represent the core actions of direct selling industry: attracting potential customers, and attracting potential agents. Most of the direct selling companies and societies are encouraging their leaders and partners to generate leads and giving tips and instruction on how to enrich their contact list [17].

2.2 Social Media

Social media technology is a phenomenal development on the Internet accompanied by the emergence of many Web 2.0 technologies. Generally, social media represent a big leap from the past when it was limited to communicating very small amounts of information and with greater control of data managers to the present of connecting through the web interactively. Social media is becoming more important in leads generation for direct selling business.

Social media also provides many opportunities, including information sharing among all subscribers of the network with the possibilities of free and direct interaction on social networking sites. Furthermore, it allows participants to determine what they like or dislike, or what information they are sharing on their personal profiles, which gives an effective measurement for the popularity of products, activities, opinions, and trends, and gives useful data to make statistical studies related to this data.

The term social media is the use of the Internet and mobile technologies to turn communication into more interactive dialogue. Andreas Kaplan and Michael Haenlein (2010) defined social media as “a group of Internet applications that build on the foundations of ideology and technology from Web 2.0, which allows the creation and exchange of data" [18].

Trisha Baruah (2012) listed many different forms of social media technology including: Internet forums, social blogs, microblogging, wikis, social networks, podcasts, photographs or pictures, videos, rating and social bookmarking. Technologies include activities such as blogging, wall-posting, picture-sharing, vlogs, music-sharing, crowdsourcing and voice over IP, rating, liking, and voting. Social networks also can integrate many platforms [19].

Facebook is a useful and effective social media tool. This social network can be accessed free of charge and allows users to join networks organized by geographical place or point of work, educational background or region, in order to communicate with others and interact with them. Also, users can add friends to their friends list and send them messages, and also update their profiles and define their own friends. This social network became an essential application that users use either on their PCs or their smartphones for its features of making friends, finding friends and figuring out friends' preferences. These features made Facebook one of the indispensable stations for e-commerce advocates and affiliated companies due to the huge number of participants in the website (over one billion users around the world) [5]. Alexa website indicates that Facebook is the second most popular website in the world [20].

Twitter, which is considered a microblogging social media website is another common social networking website. Twitter is a good tool to express short ideas, giving short feedback, sharing important activities such as retweets and hashtags, which are two of the most interesting tools to measure trends and the popularity of opinions. That's why companies and organizations are eager to have an account on Twitter, to know the number of followers, retweets for each tweet, and spreading news on the high trend hashtags [21].

Instagram, another important free photo-sharing social network allows users to capture an image and add a digital filter to appear like professional captures. The user can share in a variety of social networking services, as well as the Instagram network itself. In the beginning, Instagram was only supported in iOS devices: iPhone, iPad, and iPod Touch. Now, Instagram support Android smartphones. The application provides the features of comments and likes, and now has developed shooting 15-second intermittent videos [22].

YouTube is a well-known website for videos, allowing users to upload, watch, share, and rate video clips for free. YouTube has now become the third ranked website in the world by Alexa website rating. Because of the high demand on the website and the high number of users, YouTube has become a good target for companies to make their own channels, and spread their advertisements before users can stream a video, and has encouraged people to subscribe in their channels to inform subscribers about new videos uploaded. Currently, YouTube users have joined with Google+ website, and the activities that happen in YouTube can be shared in the linked Google+ page [23].

In summary, social media is the information systems treasure for this century. The richness of data is doubling rapidly, calling e-commerce advocates to take advantage of these media services to build successful applications or extract hidden information. Direct sellers require some applications that rely on social media data to help them in their promotion.

2.3 Data Mining

Data mining (DM) is the extension of multiple computer science fields like machine learning (ML), statistics, visualization, informatics, etc., and it is the process of discovering new knowledge from a dataset. The discovered knowledge is supposed to help in describing the current behavior of business or predicting new results of that business [24].

2.3.1 Data Mining Functions

Therefore, data mining functions categorized into two parts: descriptive data mining and predictive data mining, and each basic function has multiple types of models and algorithms. Descriptive data mining consists of four major functions: association rules (AR), clustering, summarization, and sequence discovery.

While predictive data mining has classification, regression, prediction and time series analysis functions. The four most common functions of data mining are association rules, clustering, classification and regression [24].

2.3.1.1 Association Rule

The association rule data mining function is focused on finding patterns of frequencies and associated regularities between different items, features, or sequences in a dataset. There are multiple algorithms in association rule methodology, such as Apriori algorithm, FP-Growth, constraint-based association rule, multilevel association rule, and multidimensional association rule [24].

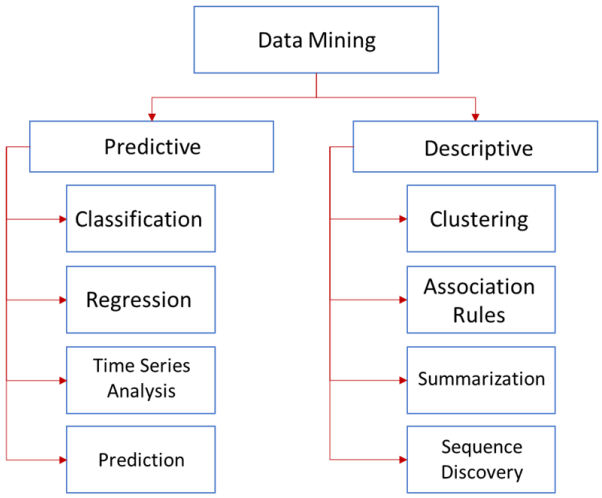


Figure 2.6 The distribution of the different types of data mining techniques.

2.3.1.2 Clustering

One more popular data mining function is clustering. The clustering concept is based on grouping data objects in different clusters. The different types of clustering methods vary by data type (interval, nominal, mixed), or by different approaches (partitioning (e.g. k-means method), hierarchical, density) [24].

2.3.1.3 Classification

Another important data mining function is classification, which is a kind of data analysis that can find classifiers to predict categories by inductive learning on the dataset. This type of data mining function has different kinds of methodologies such as decision tree (e.g. ID3 algorithm, random forest algorithms, and C4.5, C5.0 algorithms), neural networks (NN), Bayesian classification (e.g. Naïve Bayes algorithm), text

classification, rule-based classification, backpropagation classification, lazy learner methodology (e.g. k-nearest neighbor algorithm), and associative classification (which combines the functions of classification and association rule) [16].

2.3.1.4 Regression

Regression is a data mining function that predicts a number. Age, weight, distance, temperature, salary, or sales can all be anticipated using regression methods. A regression role starts with a dataset in which the objective values are known. In the model form (training) process, a regression algorithm appraises the estimation of the objective as a function of the prediction for every case in the assembled information. These relationships between predictors and target are summarized in a model, which can then be connected to an alternate dataset in which the objective value is obscure. Regression models are tried by computing different statistics that measure the contrast between the predicted values and the expected value. This type of data mining function has different kinds of algorithms; the most important of them are the Generalized Linear Model (GLM) and the Support Vector Machine (SVM) [24].

The complexities and the diversities of data mining techniques and algorithms encouraged the software development companies to develop data mining tools that simplify the designing of the data mining models and testing the accuracy of the wanted results.

2.3.2 Feature Selection

Feature selection is one of the main stages of most knowledge mining approaches that is used in exploration of predictive data. When having a set of data features more than that can be learned, or in the initial discovery phase, then some features are chosen over others. Feature selection methods forms a set of the preferred features in the first place, without considering that there is a linear relationship, and therefore this process is considered prior to treatment of the pre-processing exploration. Choosing manageable groups of the most likely of predictions that have a relationship to required fields can be used later for deeper analysis by selection methods and classification [25].

The main goal of feature selection in machine learning is to find the best set of attributes that allows one to build useful models of studied subjects and make them easier to interpret [25].

In classification, the number of features that can be used in data mining could be high. Applying all these features may cause high dimensionality which makes the processing difficult. Also the existence of a high number of variables is an obstacle to most users even when these features by themselves are relevant for the task, not to mention irrelevant or redundant variables which can hide other patterns. Sometimes, various filtered subsets can be tried and their efficiency tested, or we can use some algorithms specially designed to feature selection and reduction. These selected features add a measure of predictability.

Depending on the organization of a search process, feature selection algorithms are typically categorized as belonging to filters, wrappers, or embedded approaches. There are also constructed combinations of approaches, where for example firstly a filter is employed, then a wrapper, or when a wrapper is used as a filter. It is also possible to apply some algorithms to obtain ranking of attributes, based on which feature selection or reduction is next executed.

2.3.3.1 Filter Approach

Filters are completely separate processes to systems used for classification, working independently on their performance and other parameters. They can be treated as a pre-processing procedure. Filter type methods select variables regardless of the model. They are based only on general features like the correlation with the variable to predict. Filter methods suppress the least interesting variables. The other variables will be part of the model classification, regression used to classify or a data prediction. These methods are particularly efficient in computation time and robust to overfitting. The most efficient algorithms for the filter approach are Information Gain, Gain Ratio, Chi-Square, and ReliefF algorithm [26].

However, filter methods tend to pick redundant variables because they do not consider the relationships between variables. Therefore, they are mainly used as a pre-processing method.

2.3.3.2 Wrapper Approach

In a wrapper approach to feature selection it is argued that the best evaluation of some candidate variable subset is obtained by checking its usefulness in classification, as the estimated predictive accuracy is typically regarded to be the most important indicator of relevance for attributes. The induction algorithm can be run over the entire training set and then measured against the testing set, or a cross-validation method can be employed. Since the search and selection process is adjusted to a particular characteristic of the inducer, they can show a bias, resulting in increased performance of the chosen classifier, but worse results for another, especially when they significantly vary in properties [26].

In other words, wrappers tend to construct sets of attributes which are customized, tailored to some particular task and some particular system. Another disadvantage of this approach is in computational costs required. Execution of the learning algorithm for many subsets of features can become unfeasible, not only when there are very high numbers of attributes to consider, but also in cases when the training step is complex and time-consuming even for smaller numbers of variables. The wrapper model can be used not only for feature selection or reduction but for other purposes, to adjust better some parameters of a classification system.

2.3.3.3 Embedded Approach

Several predictors have their own, inherent mechanisms, built-in in learning algorithm, dedicated to feature selection. When such a mechanism is actively used we have an embedded solution. The primary criterion is to measure feature subset "usefulness" such as the wrapper approach, but the learning process guides the search method. The general results of this approach are similar to wrappers but less computationally expensive, and less prone to overfitting [26].

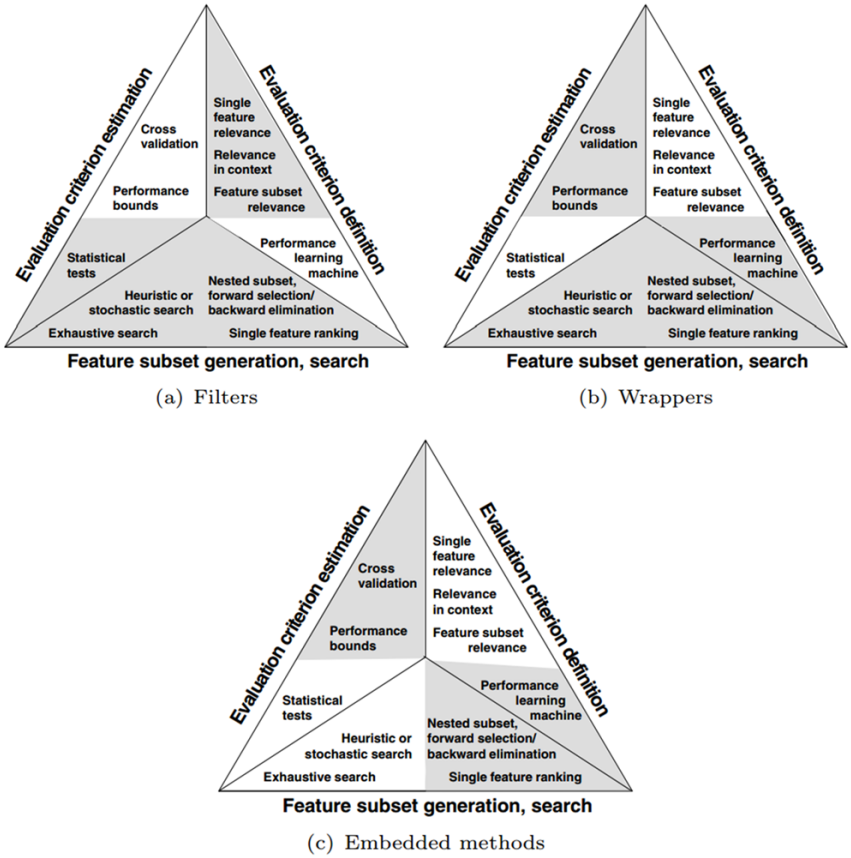


Figure 2.7 The distribution of the different types of feature selection approaches **[26]**

2.3.3.4 Search methods in feature selection

The feature selection process assesses a subset of features as a group for suitability. The selection algorithms can be separated into the wrapper, filter and embedded approaches. The wrapper uses a search algorithm to seek through the possible features and assess every subset by running a model on that subset. The wrapper can be costly and has a possible overfitting problem. The filter approach is similar to the wrapper approach in searching, but instead of evaluating the subset against a model, a simple filter is evaluated. Many popular Search approaches use strategies like greedy hill climbing, which evaluates a chosen subset of features then alters the subset if the new subset is an improvement over the old. Evaluation of the subsets requires a scoring metric that grades the subset features. Exhaustive search is impractical, so the subset of features with the most astounding score found up to that point is chosen as the attractive

feature subset. The stopping criterion varies by algorithm. Possible criteria include a subset score that exceeds a threshold, a program's maximum allowed run time that has been exceeded, and so forth. Another search-based method situated in targeted projection pursuit discovers low-dimensional projections of the information that scores highly. The features that have the biggest projections in the lower-dimensional space are then chosen. Most common search approaches include: Exhaustive, Best first, Simulated annealing, Genetic algorithm, Greedy stepwise and Scatter Search [25].

2.3.3 Data Mining Software Tools

With the huge expansions in the data mining usage and with the overall complexity of designing different data mining models regarding the rapid increase in data, the necessity of having powerful tools to analyze the large amount of data and to find useful knowledge becomes urgent. In this study, open source software was more extensively examined to find the most suitable one to apply to the study [27].

Weka is one of these well-known open source data mining software packages implemented in Java and developed by the University of Waikato, New Zealand. This tool can be linked to SQL databases, and it supports many tasks of data mining from data preprocessing to exploring techniques such as classification, clustering, regression, visualization, selection, and association rules. Weka has many advantages; firstly that it is free and licensed under GNU General Public License, secondly that it runs on any platform, thirdly that it contains a variety of supported data mining modeling algorithms, and finally it is GUI-based software, which makes it easy to use and understand [28].

R project is another interesting open source environment for statistical computing and graphics. This project includes several packages relevant for social network analysis such as igraph (generic network analysis package) and sna (sociometric analysis of networks). R is created by S programming language at the University of Auckland, and the source code is available in C, FORTRAN and R language as well [29].

A further piece of software for data mining analysis is **Orange**, which is a software package implemented in C++ and Python that was developed by the University of Ljubljana, Slovenia. It is considered component-based, provided by a visual programming front-end to visualize and analyze data. Orange has many components that enable its users to preprocess data, do filter tasks, modeling, evaluation and different exploration methods [30].

ELKI is another open source software developed by the University of Munich in Germany written in Java for research purposes, particularly focused on discovering knowledge. The software has wide options of different and advanced algorithms that allow data mining evaluation and its interaction with database index structures [31].

One more software is **KNIME**, the Java implemented open source software that analyzes data. The advantages of using KNIME is the capability of processing a large volume of data (in millions of records), and allowing integration of data mining methods through different components and plugins, as well as the integration with other open source software like Weka [32].

2.3.4 Recommender Systems

Recommender systems are a part of information systems, and one of the data mining applications with a wide range of techniques like information retrieval, statistics and machine learning that predict a preference of an information system **user** for an **item** [33]. These types of data mining became very popular for users to help those making decisions and for companies as a marketing solution. Many applications and commercial websites are using these technologies to predict the preferred movies (such as Netflix), videos (such as YouTube), music (such as Pandora Radio and Last.fm), restaurants (such as Yelp), hotels and vacations (such as Booking and TripAdvisor), books (such as Amazon), products in general (also by Amazon and eBay), people (including friends, followers and online dating relationships, such as most of social networks), services and promotions (such as Groupon). One of the most successful examples of recommender system is **Netflix**, the media streaming service that offer a wide database of movies and TV series with a monthly charge.

In 2006, Netflix developed its personalized recommendation system that based on rating and reviewing. Later, The Company offered one million dollar prize to the developer who can develop a better recommendation system algorithm that gives 10% better results. For new users, the system asking the user in the beginning about his favorites five movies out of a list of movies, then the system start to recommend many categorized lists of movies related to one of the five favorite selected. For example, if one of the selected movies belongs to animation genre, the system will create a list of suggestion of other animation movies, and so on. After watching one movie, the system will direct the user to other possible category of similar movies. For example, if the user watches an action movie that has a revenge story, the system will create a sub category list of other movies that has similar revenge story. The idea that each movie is been tagged by the company by couple of keywords that describes the movie, and characterized by the name of actors, the name of the director, the year of production, the genre, and the rate, and start to find relationships between each movies has been watched to create a new list of suggested movie [34].

Netflix was a social networking based website when it started, which means that users who are friends in the system can see each other rates and reviews. In 2010, the company disabled the friends feature and it wouldn't be considered a social network site, but there are an application on the Facebook social network called 'Netflix' that allow people access other subscribed friend reviews and recommendation. The attribute-based and the item-to-item approaches are commonly used in Netflix recommendation system [34].

Another important example of recommendation system is the **YouTube** recommender system. YouTube recommending videos for watching and channels for subscribing based on the previous video clips that the user has seen before. For example, if the user watched a video clip about the Second World War, the system later will recommend another videos that have similar title, similar content or another videos of the same channel.

The giant global e-commerce company **Amazon.com** also has one of the most well-known e-commerce recommender systems that follow this approach by recording data on the user behavior to enables the system to offer or recommend to an individual specific item for

him/her, or set of items based upon his/her preferences demonstrated through purchases or items visited. Besides that, Amazon.com launched Amazon Seller Product Suggestions in 2010 to recommend a certain products to the third-party sellers in Amazon.com through its affiliate marketing program 'Amazon Associates', the suggested products depends on the costumer's browsing history profile [35].

Social networks are highly needed to have a recommender system to suggest social relationships. While Facebook is recommending friends, Twitter recommends user's accounts to follow, and LinkedIn suggest business relationships. Facebook for example is using an important recommender tool named 'people you may know', which suggests friends to the user based on set of parameters such as mutual friends, city of living, age, educational background, occupation, and so on. This recommender tool became essential in most social networks.

In Summary, Recommender systems are becoming an essential tool for several information systems applications to learn new and hidden information. It may follow either one of the collaborative filtering approaches or the content-based approach, or combining more than one type of recommender system approaches to reach to the more logical and realistic result. Understanding the user-item relationship in the recommender system, the attributes of each of them, and the defining the wanted information lead to know the best blend of approaches to find right suggestions and the best likelihood of it for the user.

2.3.5 Data Mining for Lead Generation in Direct Selling Industry

Using data mining technology for the lead generation purposes is one of the most interesting research studies for big data companies. The real value behind data mining lies in defining more concentrated methodologies related to lead management, and lead generation is the first step. Using customers data, which could be combined with behavioral analysis, predictive models are designed that support marketers to qualify leads [36].

Examining data from various perspectives is used to focus on client offers and services, building revenue-generating opportunities, and decreasing promotional expenses. As stated by a study by Econsultancy and Signal (June 2015), both B2B and B2C marketers would use the data sorts the vast majority frequently: transaction historical data by 87%, client data by 80%, and behavioral data from the web and marketing campaigns by 74%. The study motivates marketers also to access more real-time data to understand the leads very well. Another study made by MarketingProfs (2013), states that leads generation is the biggest challenge in B2B marketing [37] [38].

Predictive analytics gives scientifically determined equations that figure how your prospects may respond to promoting efforts, empowering B2B organizations to serve them at each pivotal point. Mining data is the foundation of predictive analytics [39]. Data mining technologies vary in lead generation applications, depending on the type of data, type of lead, nature of business, and the goals of the leads generation. What makes data mining for leads generation in the direct selling industry different from other types of businesses that it combines B2C and B2B e-commerce business styles, which needs a proper modeling different kinds of leads, partner leads, and customer leads. These two kinds also can have multiple situations, such as whether this lead is an only partner, only customer, or both partner-customer lead.

2.4 Related Researches

2.4.1 Mining Customer Knowledge for Direct Selling and Marketing

In this study, researchers designed a data mining model that consists of the k-means clustering algorithm then an association rule (Apriori), but they did not choose any open source software to apply that. The researcher analyzed consumer adumbration, lifestyle habits and purchasing behavior. Finally, this study figures out some models including cluster consumer purchase preference and demand in order to generate different marketing alternatives for decisions. The outcomes of this study can help attract more direct marketing firms to use data mining to improve markets and earn higher profits for direct selling [4].

2.4.2 The Impact of Preprocessing on Data Mining: An Evaluation of Classifier Sensitivity in Direct Marketing

The paper applies different types of classification techniques like decision trees, neural networks and support vector machines (SVM). The study focuses on choosing the appropriate data mining methodology model that represents direct marketing. The researchers used Weka software to prove their results [40].

2.4.3 Knowledge management and data mining for marketing

Many marketing solutions focus on the customer characteristics and find their purchase patterns. In this study, researchers define three major areas of application of data mining for knowledge-based marketing; firstly, customer profiling, secondly, deviation analysis (clustering), and thirdly, trend analysis to support the marketing decision [41].

2.4.4 Predicting User Personality by Mining Social Interactions in Facebook

This study investigates a way of analyzing data about the users of social media to infer their personality, by collecting specific data from a questionnaire made as a Facebook application. The study depends on multiple classifiers to discover users' personalities by finding interactions between the parameters such as users' friend number and number of timeline posts. The researchers applied Naïve Bayes, K-nearest neighbor's, decision trees, and association rules techniques to analyze the dataset. They used Weka software to test their theory accuracy. Other algorithms such as C4.5 algorithm, and the lineal rule of Fisher were used by the R software [42].

2.4.5 An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis

This study looked at the airline industry to measure the feedback of customers in social media about the airline companies' services and analyze the customer tweets on the famous social media website Twitter.

The researcher used the R data mining tool to import data from Twitter then worked on preprocessing it through several steps: (1) he classified his instances using R language, (2) used the Weka data mining tool for feature selection, (3) then tested several chosen data mining classifiers to build an ensemble classifier to improve the overall accuracy and to predict more accurate customer feedback [43].

2.4.6 An Integrated CRM Data Mining Method for Predicting Best Next Offer

The researcher studied the systems of customer relationship management (CRM) to help marketing solutions to find leads and predict the best next offer. A data mining system has been designed to group the business leads from the datasets. A feature selection filter is developed to reduce the dimensionality of the data, and improve results accuracy. The steps of the preprocessing phase and the feature selection part require further investigation [44].

2.4.7 A Data Mining Framework for Automatic Online Customer Lead Generation

This study has a good example of data mining solutions for lead generation in the real estate industry. The researcher developed a framework that consists of several machine learning steps, includes data modeling, data integration, online web data streams, data mining, and system evaluation for pattern discovery and lead prediction [45].

Chapter 3 **Preliminary System Analysis and Data Preprocessing**

3.1 Social Media Data Selection and Business Scenario

In this study, the user is supposed to be the networker, and he/she looks to generate leads from the social media. The dataset for this study is from a social network, specifically Facebook, is selected based on the following reasons:

1. The availability of huge amount of training data considering that Facebook users had already passed one billion, which attracts business companies to target potential customers.
2. Facebook's users profile data may contain better features that can be employed for lead prediction.
3. The variety of applications in Facebook could potentially provide additional information available for conducting other data mining research.

Due to the problem of not having complete access to the Facebook database to apply to the study, a training data set sample was applied to it, to design a data mining model and check the results accuracy.

The user should be a member of the direct selling company, as well as being a member in Facebook. As a user, he/she is supposed to use a lead generation tool to find those new partner to join his/her network, or customers to buy a product he/she promote. The user has the option to input some attribute to narrow down and close in to the wanted results, and reduce the unwanted results.

On the other hand, the lead in this case is a regular Facebook member who the networker is seeking. As a Facebook member, he already may have a detailed profile or may not (because not all members necessarily fill in all their personal information, due to their indifference or privacy). The system, whether it has some inputs from the user or not, will figure out the appropriate leads for the user depending on their attributes. The item’s attributes can lead to find whether it represents a potential customer or not. Returning to the topic of the nature of direct selling and its work environment, anyone can work in this industry and anyone can succeed, but there are a type of people that most of networkers

prefer to do business with more than others who consume more time and effort to train and build their personal development. Most networkers target those people who have the following attributes (main attributes):

- People who have an educational background in the business and marketing field.
- People who work in marketing or for marketing companies.
- People who are familiar with the direct selling business and have experience.
- People who have specific skills that represent daring and self-confidence.
- People who are social enough and have many acquaintances.

In addition, there are other specific attributes that filter the networker preferences such as the age range that the networker is targeting, gender, customer's hometown, etc. The networker should also pick some customers' main attributes, such as if he/she wants to target customers from a specific city, or to target a specific gender, and this can be called "given attributes".

The proposed system should have a questionnaire that enables the networker to input the given attributes, as parameters, to narrow to his targeted people, while the main attributes should be discovered by mining in the database. If the prospective customer has the wanted attributes he/she will be highly recommended to the user. The results should be listed in descending order from the highest rated item according to the system to the lowest.

Consequently, the prospected customer's attributes are:

- Main attributes more related to personalities such as current work, work experience, educational background, activities, interests, number of friends, type of friendship, groups he has enrolled.
- Given attributes more related to demographical characteristics such as hometown, current location, age, gender, language

Another attributes can be discovered by a recommender system that use an association rule algorithms to find similarities and frequent item sets such as mutual friends and similar customers' attributes, and behavior.

Both the main and given attributes can be imported, prepared and learned to predict leads generation in direct selling industry. After the learning process, some database inquiry commands can be applied in order to show the suitable list of leads to the networker. To clarify the business terms, Table 3.1 describes business terms meanings.

Term	Definition
Networker	The person who is conducting networking. In this model this word is used to describe the user of the system.
Networking	The activity of building a network of sales within a network of people in the society for a direct selling company favor.
Direct Selling Company	The company that sell a product with offering the advantage to the customer to promote the company's product to other people and get a commission.
Customer	The person who buy from the direct selling company. The customer can become a networker with the direct selling company.
Partner	Another term for the networker. In this model this word is used to describe the prospected person to become a networker.
Prospect	The word that describes the people who are maybe become a customer or partners in the future.

Table 3.1 Business glossary

3.2 System Architecture

The system's role is to collect a direct selling company's information, the given attributes the networker is seeking, and the data set information. A list of potential customers who would buy the networker's products that they promote is sorted from the expected results from the system, and the list of potential customers who are more likely to become business partners to the networker is generated.

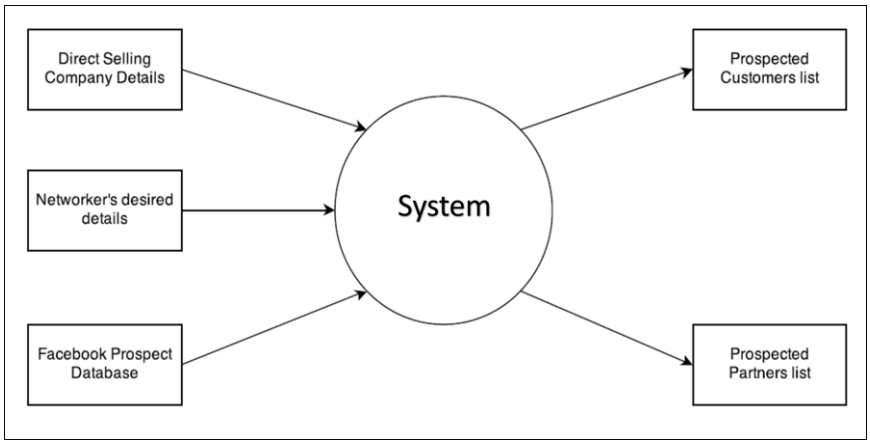


Figure 3.1 System scoop model

In this study, one direct selling company's characteristics will be considered and tested to find interested customers for its products. The company is (DubLi) which is an electronic shopping mall that includes hundreds of well-known brands in many consumer industries. They reward customers by a cashback percentage from all their purchases through the company's electronic mall. The reason for choosing this company in this study is the wide variety of products they are linked to, to match many dataset instances' interests.

With reference to what has been discussed above in this chapter, and to what has been discussed in chapter 2, the general form of the system consists of three major parts:

1. Data pre-processing part which includes data importing from Facebook, data cleaning, data modeling process and feature selection processes.
2. Data mining modeling part which includes multiple classification experiments and tests to choose the best classifiers, and design an ensemble classification method.
3. Leads Prediction. See (figure 3.2**)**.

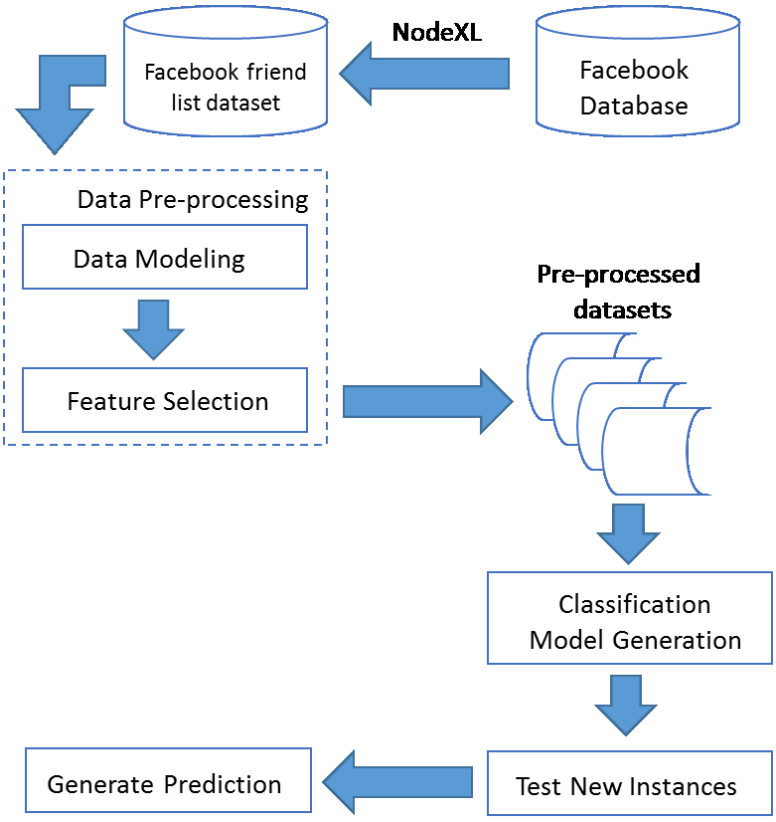


Figure 3.2 The proposed system architecture.

The data mining tool which has been chosen is Weka, due to the abundance of research studies that are similar to this study and used the same tool. Weka will be used to analyze the dataset information and to design the appropriate data mining model.

3.3 Data Preparation

3.3.1 Data Collection

The dataset that will be used is imported from a Facebook friends' list using the NodeXL importing add-on tool with the Microsoft Excel application. NodeXL Network Graph tool is an open source project that belongs to the Social Media Research Foundation. This tool allows users to get access to the allowed Facebook data using Facebook Graph API, and sorts friends' information into a spreadsheet with the desired dataset [46] **(**figure 3.3**)**. The dataset that has been imported from Facebook using the NodeXL tool consists of 1710 instances and 43 features. The imported features are shown in **(**Table 3.2**)**.

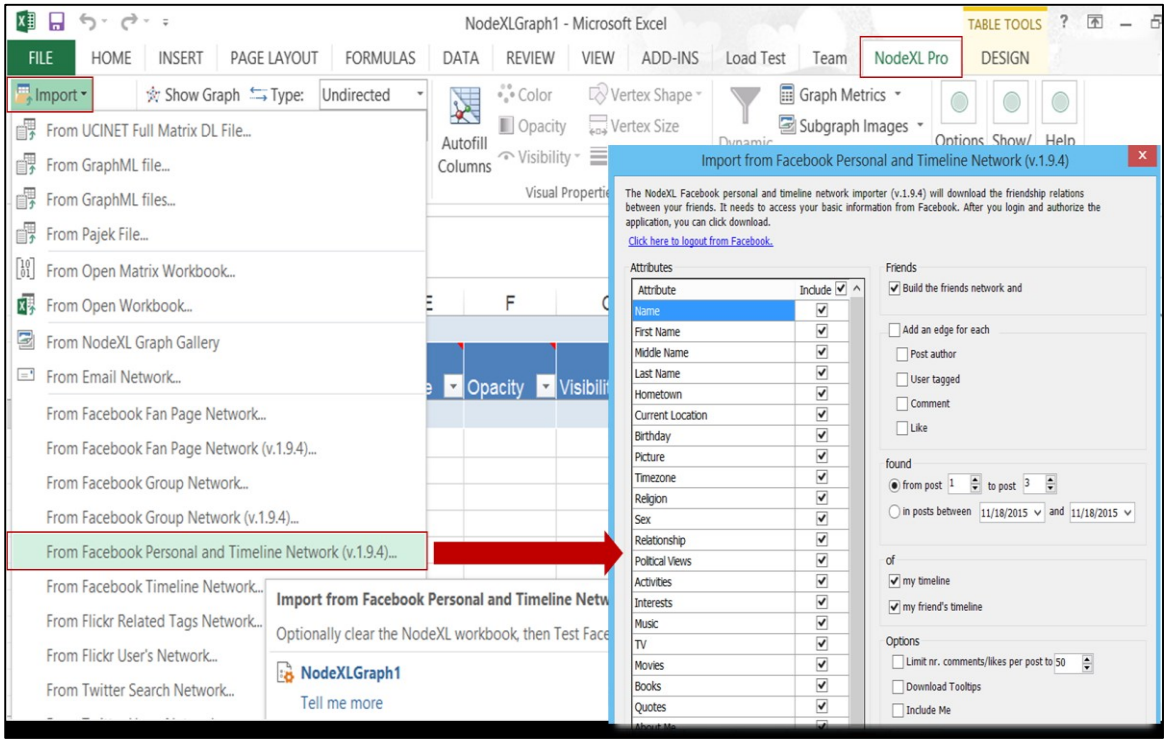


Figure 3.3 The interface of NodeXL tool

3.3.2 Data Preprocessing

Some of these features contain a set of data and split data at the same time. For example, the feature of name has the full name of the instance, followed by three different sub features (First name, middle name and last name). Some other features need to be divided into sub features. The imported features value is partitioned into different sub features to make it easier to deal with one single value as shown in figure 3.4. The final number of features expanded to be 95 features.

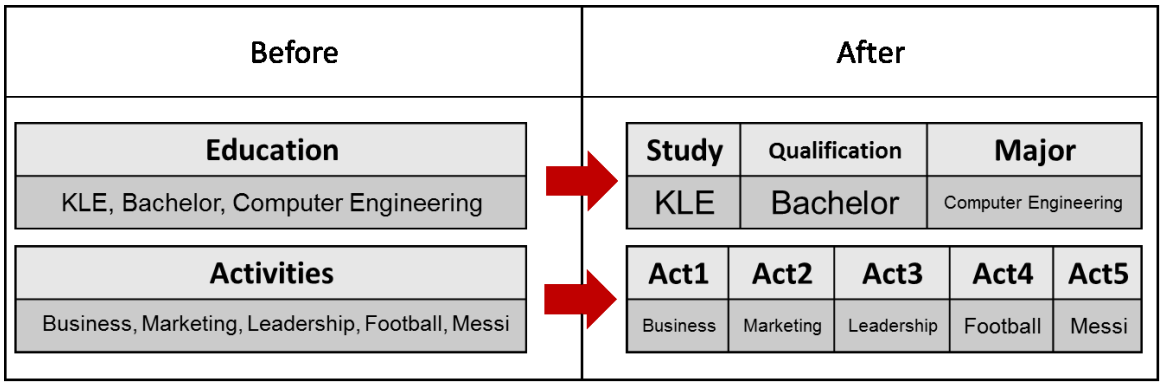


Figure 3.4 Shows two examples of how features looks like before and after division and expansion.

The imported dataset needs to be cleaned and some spreadsheet cells reformatted, like changing the non-English languages data to the English alphabet, removing bugs and symbols that prevent the data mining tools from dealing with it.

#	Imported Features	Description	Status	#	Modified Features
1	Vertex	Field with the name of the user	deleted	1	index
2	Name	Full name of the user	same	2	Name
3	First Name		same	3	First Name
4	Middle Name		same	4	Middle Name
5	Last Name		same	5	Last Name
6	Hometown	Full location at birth	deleted		
7			same	6	Hometown City
8			same	7	Hometown State
9			same	8	Hometown Country
10	Current Location	Full location at present	deleted		
11	Current Location City		same	9	Current City
12	Current Location State		same	10	Current State
13	Current Location Country		same	11	Current Country

14	Birthday		split	12 13 14 15	Birthday BDay BMonth BYear
15	Picture	Profile in Facebook	deleted		
16	Profile Update Time		split	16 17 18 19 20	Profile Update Time PDay PMonth PYear PTime
17	Timezone		deleted		
18	Religion		deleted		
19	Sex	Gender	same	21	Gender
20	Relationship		same	22	Relationship
21	Political Views		deleted		
22	Activities		split	23 24 25 26 27	Act1 Act2 Act3 Act4 Act5
23	Interests		split	28 29 30 31 32	Int1 Int2 Int3 Int4 Int5
24	Music		split	33 34 35 36 37	Music1 Music2 Music3 Music4 Music5
25	TV		split	38 39 40 41 42	TV1 TV2 TV3 TV4 TV5
26	Movies		split	43 44 45 46 47	Movie1 Movie2 Movie3 Movie4 Movie5
27	Books		split	48 49 50 51 52	Book1 Book2 Book3 Book4 Book5
28	Quotes		split	53 54 55 56 57	Quote1 Quote2 Quote3 Quote4 Quote5

29	About Me		split	58 59 60 61 62	About1 About2 About3 About4 About5
30	Online Presence		same	63	Online Presence
31	Locale	Interface Language	split	64 65 66	Locale First Language Second Language
32	Website	Personal website	deleted		
33	Image File	Profile in facebook	deleted		
34	Custom Menu Item Text	Open Facebook Page for This User	deleted		
35	Custom Menu Item Action	Facebook serial number	deleted		
36	Type	Friend	deleted		
37	Work		split	67 68 69	Work Position Field
38	Education		split	70 71 72	Study Qualification Major
39	Professional Skills		split	73 74 75 76 77	Skill1 Skill2 Skill3 Skill4 Skill5
40	Contact info		split	78 79 80 81 82 83	Email Mobile Skype BBM Twitter Facebook
41	Friends	Friends number and mutual friends	split	84 85	Friends Mutual Friends
42	Fan		split	86 87 88 89 90	Fan1 Fan2 Fan3 Fan4 Fan5
43	Other Likes		split	91 92 93 94 95	Like1 Like2 Like3 Like4 Like5

Table 3.2 Dataset attributes before and after preprocessing phase

3.3.3 Survey

The study is focused on the direct selling field which includes the direct sellers. A questionnaire form was sent to over five hundred direct sellers from the researcher Facebook contact list. The purpose of this survey is to ensure what are the most important features for the direct sellers that they consider more when approaching new customers on social media. This information is important to support the feature selection phase and taking into account the common preferences of direct sellers. Two hundred-two responses were received. Figure 3.5 shows the results of this survey; most direct sellers are considering the users' interests, then his/her activities (the users' likes). After that, they consider his/her work background and his/her number of friends. The first may indicate his/her field of experience, and the second may measure his/her social interactions on Facebook and how many people he/she knows.

Figure 3.5 Survey results

3.3.4 Data Labeling

Using the previous survey to determine the most significant preferences for potential partners and customers from the direct seller's point of view, the instances can be labeled into two different classes: business leads class which includes two types of leads (partner business lead, customer business lead), and non-lead class.

A simple program was implemented in R language to calculate the frequency of wanted keywords in partner business lead class and customer business lead class. The top three preferences from the previous survey were considered as the preferences used as the keywords for the partner business lead in. These preferences are the instance interests (which includes interest, music, TV, movies, books, fans, and other likes features), the instance activities (which includes activities, quotes, about me, and professional skills features) and the work experience features.

The program builds three dictionaries of words: a dictionary from interests' features, a dictionary from activities features and a dictionary from work features. The three dictionaries were reviewed to reform three new lexicons with chosen keywords: keywords for partner lead (partner dictionary and work dictionary), and keywords for customer lead (customer dictionary). The program uses the partner lead dictionary to search in all interests and activities features (including the books, music, quotes, etc.) and matches keywords in instances, same as customer lead dictionary, while the work dictionary searches for the keywords in the work features only. After that, the program builds a counter to calculate the frequency for how many keywords match in each instance record. For example, if the root "market" from the wanted keyword "marketing" (market, marketing, network marketing, multi-level marketing, marketing skills, etc.) matched in the work field feature, or frequent in activities or interest features, then the instance is labeled as a business lead (partner type). See figure 3.6 for more clarification.

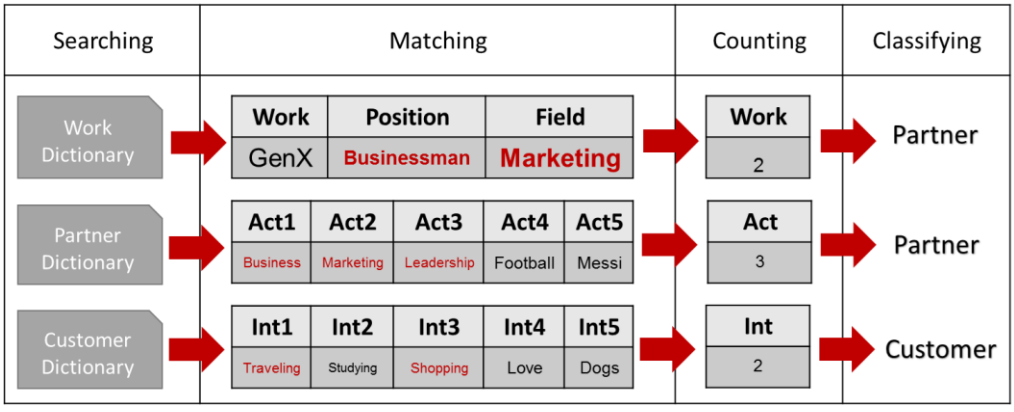


Figure 3.6 Counting frequencies of features and labeling each instance by matching keyword from dictionary

The next process is to search for frequent keywords in work features to add labels to more business leads who do not have any information in interests or activities features. The program calculates the frequency of keywords of the customer business lead in the same features from the rest of the dataset that is not labeled in the first class. For example, if the root keyword "shop" from the wanted keyword "shopping" (shop, shopping, shopping mall, the body shop, etc.) is frequent in activities or interest features, the instance is labeled as a business lead (customer type). Figure 3.7 shows the classifying method.

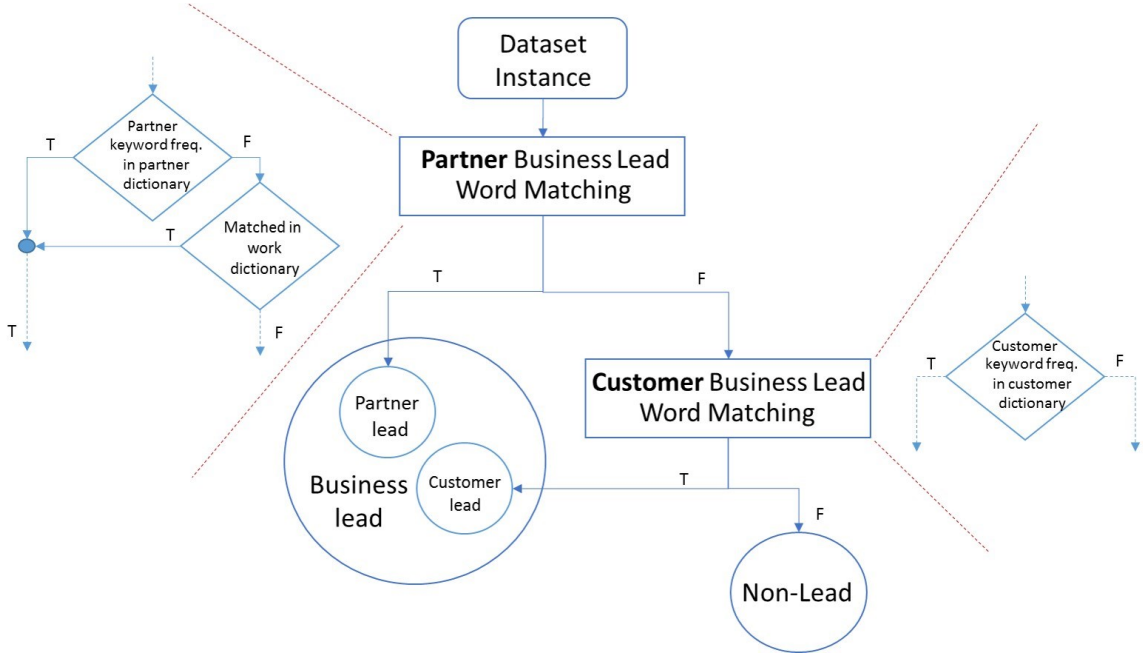


Figure 3.7 Labeling method

Therefore, we have four types of leads generation:

1. A lead who is a potential partner.
2. A lead who is a potential customer.
3. A lead who is a potential partner or potential customer.
4. Leads who are a group of potential partners and a group of potential customers.

To explain the third type, in more detail, the dataset is supposed to be used regarding to the nature of the classified data that is more suitable for users who do not care whether the lead is a potential partner or potential customer, he/she is just looking for leads, and finds as

many contacts as possible to approach them all. In the fourth type, the user is looking for a group of potential partners and a group of potential customers to approach each class in specific way to avoid people rejections or misunderstandings.

The result of this process understanding brings the need of having two different classification methods:

A) On three different classes, or the (three-classes based dataset), which consist of:
   - Partner lead class
   - Customer lead class
   - Non-lead class

   And this targeted data named (3C),

B) On Two different classes, or the (two-classes based datasets), and this has three different sub-classification ways:
   - First: Lead (both partner, and customer), and non-lead, named (2C)
   - Second: Partner lead, and non-lead (including customer), named (2CP)
   - Third: Customer leads, and non-lead (including partners), named (2CC)

The benefit of these different approaches is to visualize more specific patterns and results for each classification way. In part 5.3.3 an evaluation to choose the most suitable dataset to the task type, to use it in lead prediction. Finally, I generated factor tables that contain the frequencies of each frequent keyword in each instance. The reason for building this factor table is to use it in some classification methods that work more efficiently with transformed data, and compare its accuracy together.

3.4 Feature Selection

Feature selection is one of the important steps in the pre-processing phase which is based on mathematical methods to reduce the dimensionality data representation by removing irrelevant and redundant features, or subsets of features. The lower the dimensional representation of data, the more accurate the classification results. This plays a crucial role

in determining separating properties of pattern classes. Selecting the more related features has an important influence on time and cost of applying the classification.

3.4.1 Manual feature selection

Problems to the accuracy of the study results occur from the presence of non-correlated features. Features such as contact information or phone number for example will not be useful to find leads or mine data, so they should be eliminated. In addition, some features should be removed manually according their dependency, because dependent features are similar to have a duplicate in features and that affects the result. For example, if we have a feature of hometown that has the value "city, country" and another feature of hometown country, it is necessary to have the value of the same country for each instance, so one of them should be removed. Building classifiers with features not properly selected will cause problems in training phase, and will not score the best classification accuracy result.

In order evaluate the efficiency of feature selection phase, different versions of the original dataset has been created:

1. Four different MFS datasets: the datasets after the manual feature selection with different classes (Three classes version, and the other three with two classes different versions: 3C MFS, 2C MFS, 2CP MFS, and 2CC MFS)
2. Four different transformed data, which include the new frequency values of some features instead of multiple values for one feature type: T3C, T2C T2CP, and T2CC).
3. Four different factor tables that only include the frequency of the features that we used in classifying the instances: FT3C, FT2C, FT2CP, FT2CC. Table 3.3 shows an example of how instances in FT3C dataset look like.

Id	About	Quotes	Books	TV	Mus	Mov	Act	Int	Work	Study	Lead
1	7	2	0	0	0	2	6	2	3	0	Partner
2	0	0	2	0	0	0	4	1	0	1	Partner
3	0	0	0	1	0	1	0	0	0	0	Customer
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
1708	0	1	0	0	0	0	0	0	0	0	None
1709	0	0	1	0	0	0	1	0	0	0	Customer
1710	0	0	2	1	4	0	0	0	0	0	Customer

Table 3.3 Example of how instances of FT3C dataset look like

3.4.2 Choosing the feature selection algorithm

After the manual feature selection, many feature selection algorithms have been applied on the training dataset to figure out what is the more suitable feature selection algorithm to select the most relevant features. Weka data mining tool has been used to find the best feature selection method. The most suitable feature selection approach for this study that selects the more collaborated features is the wrapper method.

The primary goal of using wrapper feature selection strategy, is constructing and selecting subsets of features that are useful to build an accurate classifier. This contrasts with other feature selection algorithms such as chi-square or gain information, where the goal is ranking all potentially relevant variables. Wrappers generally result in better performance with the dataset of this study.

Another reason is that the wrapper approach uses different searching methods that specify the features that should be selected, unlike the filter approach that uses a ranking method that only ranks features from most important to least important, without eliminating the redundant features, making it difficult to a specify the suitable number to retain.

Two feature selection algorithms were selected for this study, applied and compared to each other. The first one was WrapperSubsetEval, and the second one was ClassifierSubsetEval. Both of these algorithms use a classifier to evaluate the important features. Naïve Bayes classifier and J48 classifier are well-known classification algorithms that can be used to evaluate the wrapper methods. Furthermore, three different searching methods were applied and evaluated to discover the best features.

Another side of the study compared different datasets passed through manual feature selection, and then the transformed dataset was compared to the automatic feature selection for accuracies. The experiment and the evaluation of the feature selection phase will be discussed later and evaluated in chapter 5.

3.4.3 The applied feature selection algorithms:

3.4.4.1 WrapperSubsetEval

It is a wrapper feature selection method which test a set of features by using a learning scheme. It uses a cross-validation to rate the accuracy of the data mining algorithm for a group of features [28].

3.4.4.2 ClassifierSubsetEval

It is a wrapper feature selection method as well as the WrapperSubsetEval. Unlike WrapperSubsetEval, it separates the dataset into training and testing datasets, instead of cross validation [28].

3.4.4 The applied searching strategies in feature selection

Three different searching methods that works with wrapper feature selection approach are has been chosen. And those search methods are:

3.4.4.1 Best First (BF)

Best-first search is a search algorithm which investigates a diagram by extending the most encouraging node picked by indicated principle.

Judea Pearl explained best-first search as assessing the guarantee of node n by a "heuristic evaluation function f(n) which, all in all, might rely on upon the description of n, the description of the objective, the data accumulated by the search up to that point, and on any additional information about the issue domain." [27] [47].

3.4.4.2 Genetic Algorithm (GA)

In artificial intelligence, a genetic algorithm (GA) is a search heuristic that impersonates the procedure of regular determination. This heuristic is routinely used to produce helpful answers for improvement and search problems [27]. Genetic algorithm have a place with the bigger class of developmental calculations (EA), which create answers for optimization issues used methods enlivened by regular evolution, for example, inheritance, mutation, selection, and crossover.

3.4.4.3 Greedy Stepwise

This method performs a greedy forward or in backward search through the space of feature subsets. It might begin with no/all characteristics or from a subjective point in space. It stops when the expansion/cancelation of any remaining attributes results in a lessening in the evaluation. It can likewise deliver a positioned rundown of attributes by traversing the space from one side to the next and recording the requested attributes that are chosen. [27]

3.4.5 Feature Selection Results

A number of features were selected after many experimental steps on different types of the original datasets. Two feature selection wrapper algorithms, two classifiers, three search methods, were applied to 12 different datasets to build up a new four datasets with the selected features. Table 3.4 shows the selected features in each dataset. The experiments show that the best feature selection algorithm is WrapperSubsetEval by C4.5 classifier and genetic algorithm method. More explanation and evaluation is in chapter 5.

Datasets	Number of features	Selected Features Subset
3C	8	Work, Study, Interests, Activities, Movies, Music, Books, Quotes
2C	7	Work, Interests, Activities, Movies, Music, TV, Books
2CP	4	Work, Interests, Activities, Quotes
2CC	5	Interests, Activities, Movies, Books, About

Table 3.4 Summery of 3-classes based datasets after feature selection with different feature selection algorithm and different search methods

3.5 Data Mining and Prediction

After finishing the preprocessing phase and having feature selected datasets labeled with two, and three classes, it is obvious that we need to apply a supervised data mining techniques to predict leads. The plan is to test the most common classification algorithms and compare the accuracy of prediction to each other.

The algorithms were chosen belong to various classification methods' concepts, such as: kernel functions method, lazy learner method, Bayesian method, and decision tree method. Five different classification algorithms have been chosen:

1. Support Vector Machine (SVM)
2. K-Nearest Neighbor (KNN)
3. Naïve Bayes (NB)
4. C4.5 Decision Tree Algorithm (J48)
5. Random Forest Decision Tree (RF)

Moreover, an ensemble classification method, or Meta learning method, was applied to combine different classification algorithms with a classifier in order to enhance the prediction accuracy, then compared to accuracy with the previous mentioned algorithms. The ensemble classification has been applied by stacking technique. The details about these data mining methods will be discussed in chapter 4.

Finally, after testing all the data mining classifiers, a prediction step retrieved the scored instances as a lead. In chapter 4, there will be a coverage for this phase of study.

Chapter 4 **Data Mining Methods for Lead Prediction**

4.1 Introduction

Lead generation system for direct selling industry has three main phases: the pre-processing, which is includes data modeling and feature selection; the data mining; and prediction. In this chapter, data mining and prediction will be introduced. The study will focus on supervised learning data mining method to build a model for predicting leads. Most of chosen data mining methods are classified as a classification methods.

The steps of classification and prediction are: testing the applied common classification methods, designing an ensemble method (stacking) and compare it with other classifiers, build prediction tables and decision trees, querying the leads, and finally proposing a framework for the system will be designed by the end of chapter 4.

4.2 The Applied Classification methods

4.2.1 Support Vector Machine (SVM)

Support vector machine (SVMs) is a supervised learning model that analyzes data for regression and classification. Given a set of training samples, every set is separated to fit in with one of two groupings, the SVM preparing calculation creates a model that names new cases into one class or another, to make it a non-probabilistic binary linear classifier. The SVM model is a representation of the illustrations as focused in space, mapped out so that the different classes are separated by a clear gap that is as wide as could be allowed. New cases are then mapped into that same space and anticipated as having a place with a classification given to which side of the gap they fall [48].

In addition to linear classification, SVMs can efficiently execute a non-linear classification by what is named the kernel trick, by mapping their inputs into high-dimensional feature spaces. For using SVM for linearly separable binary sets, let's suppose that we have two features (x1, x2) for example. The goal is to create a "hyperplane" that classifies all training vectors into two classes, as shown in figure 4.1. The best choice will be the hyperplane that

leaves the maximum margin from both categories. The next equation defines the best hyperplane:

$$g(\vec{x}) = \vec{w}^T \vec{x} + \omega_0 \text{ , while}$$

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class 1}$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class 2}$$

And the distance z between the closest elements to the best hyperplane is calculated by the next equation:

$$z = \frac{|g(\vec{x})|}{\|\vec{\omega}\|} = \frac{1}{\|\vec{\omega}\|}$$

And the total margin will be calculated by the next equation:

$$\frac{1}{\|\vec{\omega}\|} + \frac{1}{\|\vec{\omega}\|} = \frac{2}{\|\vec{\omega}\|}$$

With considering a significant factor that minimizes the $\vec{\omega}$ term in the right side of the equation will maximize the reparability.

Minimizing $\vec{\omega}$ is a nonlinear optimization task, solved by the karush-Kuhn-Tucker (KKT) conditions, using multipliers λ_i.

$$\vec{\omega} = \sum_{i=0}^{N} \lambda_i y_i \vec{x}_i$$

$$\sum_{i=0}^{N} \lambda_i y_i = 0$$

SVMs can be a helpful tool for insolvency analysis, in the case of non-regularity in the data. It can help evaluate information, i.e. financial ratios which should be transformed before entering the score of classical classification techniques. SVM has regularization parameter, which makes the user think about avoiding overfitting. Also, it uses the kernel trick so that the expert knowledge could be built in about the problem via engineering kernel [24]. On the other hand, the SVM's limitations are that the theory only covers the determination of the parameters for a given value of the regularization, kernel parameters, and selection of the kernel. Another common disadvantage of SVMs is the lack of

transparency of outcomes. SVMs cannot represent all scores as a simple parametric function since its dimension may be very high [49].

The `svm` function in R data mining tool in the package {e1071} was used to achieve this task, as well as to examine the algorithm in Weka data mining tool by using the `libSVM` function.

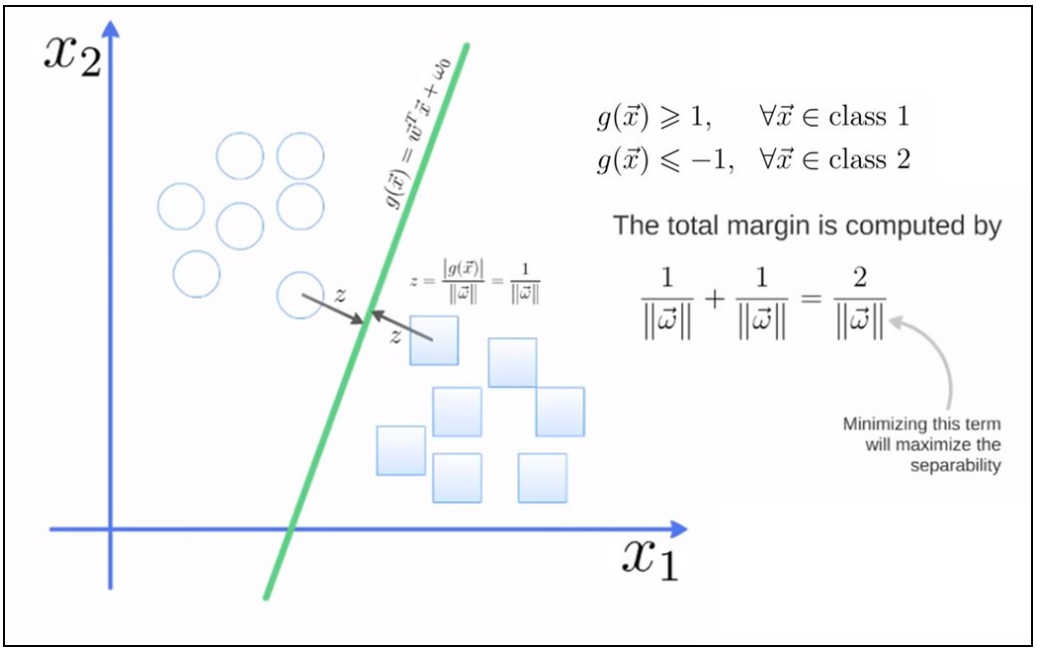


Figure 4.1 Support Vector Machine

.

4.2.2 K-Nearest Neighbours (KNN)

In pattern recognition, the k-Nearest Neighbours algorithm is a non-parametric technique used for classification and regression [50]. As a part of both cases, the input comprises of the k nearest training samples in the attribute space. The result relies on upon whether KNN is used for classification or regression.

The concept of KNN classification algorithm is that the object is classified by the majority votes of its neighbours. Given N training vectors, KNN algorithm identifies the k nearest neighbours of 'c' to estimate its class, regardless of labels of the class. For example, if we have k=3, and we have two classes 'a' and 'b', and the algorithm should find the class if unclassified object 'c'. Because k=3, we have to find the nearest three neighbours of object

'c', if more 'b's are nearest neighbours to 'c' more than 'a', then 'c' become 'b'. Figure (4.2) explain the algorithm of KNN.

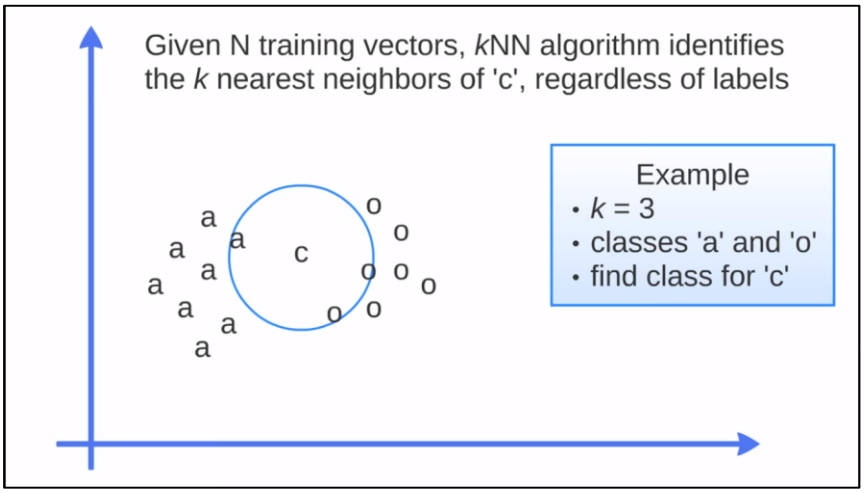


Figure 4.2 K-Nearest Neighbour methodology

For another example, if k=1, each training vector defines a region in space, defining a Voronoi partition of space. Figure (4.3) explains the partitioning method in KNN algorithm.

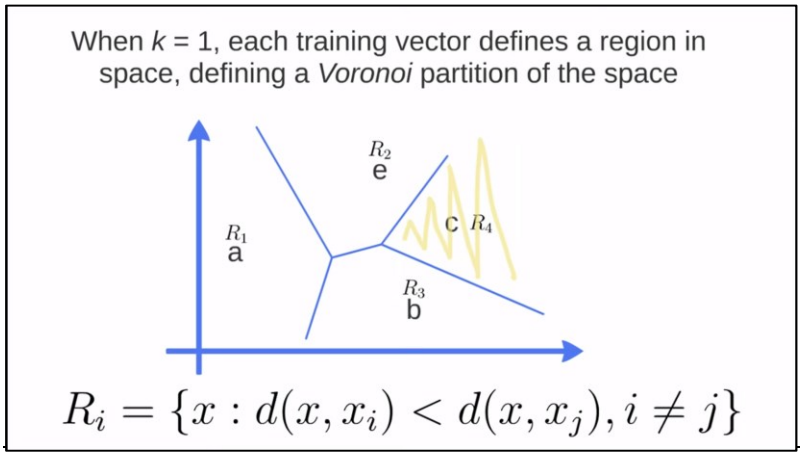


Figure 4.3 K-Nearest Neighbor area partitioning

To perform this algorithm in the best way, k value should be an odd value for two classes' problems, and it must not be a multiple of the numbers of classes. The main negativity of KNN is the complexity in searching the nearest neighbors for each sample [50].

Another negativity is that KNN is a lazy learner. It learns nothing from the training dataset and just uses the training dataset itself for classification. For new instances label prediction, the KNN algorithm will find the K closest neighbours to the new instance from the training data, the predicted class label will then be set as the most common label among the K closest neighbouring points. Furthermore, it is slow method for large number of data [50]. The `knn` function in R data mining tool was used to implement the experiments and to evaluate the algorithm for this study, as well as examining the algorithm `IBK` in the Weka data mining tool, which is the same as the KNN algorithm.

4.2.3 Naïve Bayes (NB)

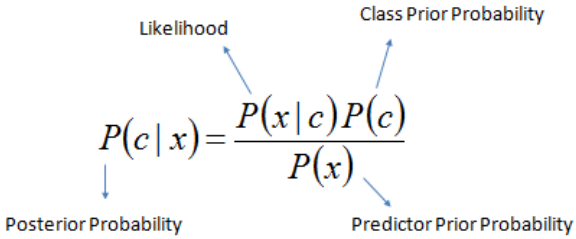
In machine learning, Naïve Bayes classifiers are a group of basic probabilistic classifiers use to apply Bayes' hypothesis with powerful independence assumptions between the attributes. A Naïve Bayesian model is easy to build with no complicated iterative parameter estimation, which makes it particularly useful for large datasets. Furthermore, Naive Bayesian classifier often works well and is commonly used because it often performs better that other complicated methods.

Naive Bayes has been employed widely since the 1950s. It was brought in as alternate name into the text retrieval challenges in the mid-1960s [51], and remains a popular method for text categorization, judging documents as fitting in with one classification or the other, (for example, spam or legitimate, sports or political issues, etc.), and with word frequencies as the components. With proper preprocessing, it is used in this space with added strategies including support vector machines.

Bayes theorem provides a method for calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naïve Bayes classifier which assumes that the result of the value of the predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [24].

One of the notable limitations of NB is that it has strong feature independence assumptions, while, on the other hand, it is fast to train and to classify, and it is not sensitive to irrelevant features. Another important feature of Naïve Bayes algorithm is that handles both real and discrete data, as well as handling streaming data [24]. (see explanation below).

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

The `naiveBayes` function in R data mining tool in the package {e1071} was used to implement the experiments and tests to evaluate the algorithm for this study, as well as examining the algorithm in Weka data mining tool.

4.2.4 Decision Tree Methods (DT)

A decision tree is a decision support instrument that uses a tree-like chart or model of choices and their possible results, including chance occasion results, asset expenses, and utility. It is one approach to show an algorithm.

Decision trees are generally used as a part of research, examination, particularly in decision analysis, to recognize a technique destined to achieve an objective, but on the other hand they are a common tool in machine learning.

This method has many algorithms, and in this research, two common algorithms were chosen to apply to the dataset and to evaluate its accuracy.

3.2.4.1 C4.5 algorithm (J48)

C4.5 is an algorithm used to build a decision tree created by Ross Quinlan [52]. C4.5 is an extension of ID3 decision tree algorithm. The decision tree produced by C4.5 can be used for classification, and hence, C4.5 is frequently referred to as a statistical classifier.

C4.5 assembles decision trees from an arrangement of training data similarly as ID3, utilizing the idea of data entropy. The training dataset, S = {s_1, s_2, ...}, is a set of effectively grouped examples. Every specimen s_i comprises of a p-dimensional vector

$(x_{\{1,i\}}, x_{\{2,i\}}, ..., x_{\{p,i\}})$, where the x_j represent features values or feature of the sample, as well as the class in which s_i falls.

At every hub of the tree, C4.5 picks the quality of the information that most successfully parts its arrangement of tests into subsets improved in one class or the other. The part basis is the standardized data pick up (distinction in entropy). The characteristic with the most noteworthy standardized data addition is settled on the choice. The C4.5 calculation then repeats on the littler sub lists [52].

This algorithm has a couple base cases. All the samples in the list fit in with the same class. At the point when this happens, it essentially makes a leaf node for the decision tree saying to pick that class. None of the features give any information gain. For this situation, C4.5 makes a decision node up the tree using the normal estimation of the class. Occasion of beforehand concealed class experienced. Once more, C4.5 makes a decision node higher up the tree using the normal quality. C4.5 can build models that can interpreted easily. Another features of c4.5 is that it can use categorical and continuous values, in addition to dealing with noise properly. This method cannot work properly with small training datasets, and small data can generate different decision trees [24]. The `J48` decision tree function in Weka data mining tool was used to test and to evaluate the algorithm for this study. J48 is an open source decision tree classifier written in Java. It represent the C4.5 algorithm in Weka.

4.2.4.2 Random Forest (RF)

Random forest is a reflection of the general method of arbitrary decision forests [53] that are a group learning approaches for classification, regression and other different roles. This method works on building a large number of choice trees at training time, and resulting in the class that is the chosen method of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests is right for a decision trees' propensity for the overfitting issue to a training set. The algorithm was developed by Leo Breiman [54] and Adele Cutler [55]. The method merges Breiman's "bagging" idea and the random selection of features.

Random forest runs efficiently on large databases and handles thousands of input variables without variable deletion. It has an efficient method for estimating missing data and maintains accuracy when a large proportion of the data are missing. At same time it has some limitations, overfitting could occur in some datasets with noise, and it could biased for some dataset features [24].

The `randomForest` function in R data mining tool in the package {randomForest} was used to implement the experiments and tests to evaluate the algorithm for this study, as well as to examine the algorithm in the Weka data mining tool.

4.3 The Ensemble Method (By Stacking)

An ensemble classification methods model was proposed. It allowed many classifiers to work together as components in order to enhance machine learning prediction results. The ensemble method works by adding multiple different algorithms prepared on the training data and a Meta classifier to learn how to take the predictions of each classifier and to make accurate predictions on unseen data. The Meta classifier that evaluates the stack is usually a regression function, such as a logistic regression or SVM algorithm [56].

The stacking algorithm includes training a learning algorithm to combine the predictions of several other learning algorithms. To begin with, the majority of other algorithms are trained using the available data, then a combiner algorithm is trained to make a last prediction using all the predictions of the other algorithms as additional inputs.

Stacking is expected to have better performance, superior to any single algorithm. Figure 4.4 shows the proposed ensemble model for this study, which is included all the tested classifiers in the study; KNN, NB, C4.5 and RF as stacked algorithms, and SVM as a Meta combiner.

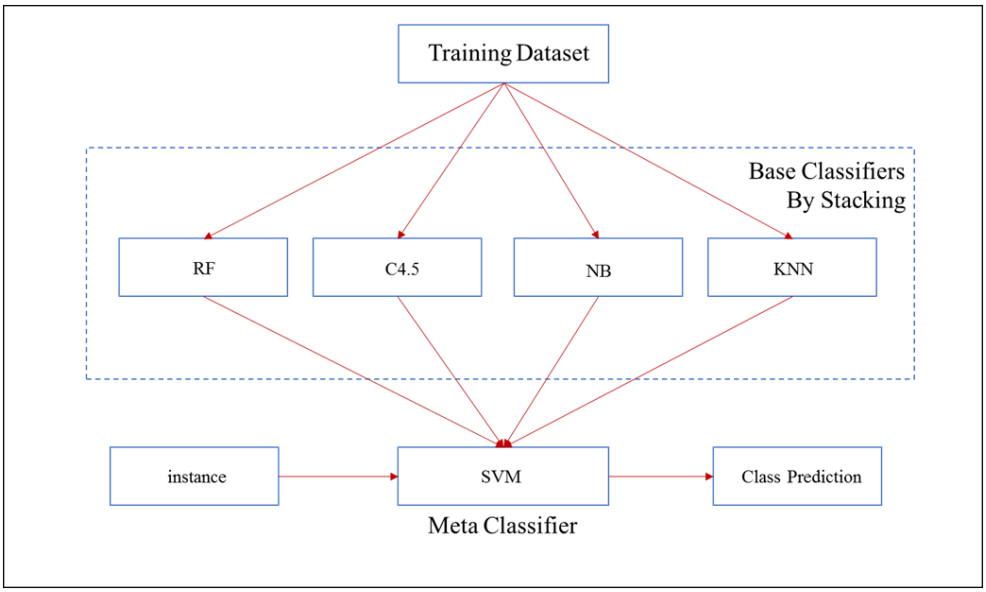


Figure 4.4 The proposed ensemble modeling by stacking method

4.3 Classification results

A number of single classifiers were trained and tested to evaluate their accuracies, in addition to an ensemble classification on four different datasets. Table 4.12 shows the results of accuracies of each classifier on the trained data. The results shows converge on classifiers results, except Naïve Bayes algorithm, which had the lowest results in three different datasets. On the other hand, the proposed ensemble method was the best, and the other classifiers, C4.5, KNN, SVM and RF performed well and had close results to the ensemble method.

Classifier	3C Accuracy	2C Accuracy	2CP Accuracy	2CC Accuracy
SVM	72.1%	86.5%	83.4%	72.1%
KNN	69.8%	86.4%	83.5%	71.1%
NB	63.3%	79.8%	80.5%	71.6%
C4.5	72.6%	85.1%	84.0%	72.0%
RF	70.8%	86.4%	83.4%	71.4%
Ensemble	72.7%	86.5%	83.9%	72.2%

Table 4.4 The overall Accuracies of each final datasets by each classifier

4.4 Prediction

An implementation in R data mining tool was developed to build a prediction results file for the used classifiers, and these results were combined with the leads information table that was exported from main dataset. The classifiers with higher accuracies were considered to build prediction results and used in the query process. This simple query command shows the list of leads generation for direct selling.

4.5 Proposed Framework

The proposed framework is shown in figure. The framework diagram is divided into three phases which are the main study phases: the pre-processing phase, the data mining or classification phase, and the prediction phase. See figure 4.4.

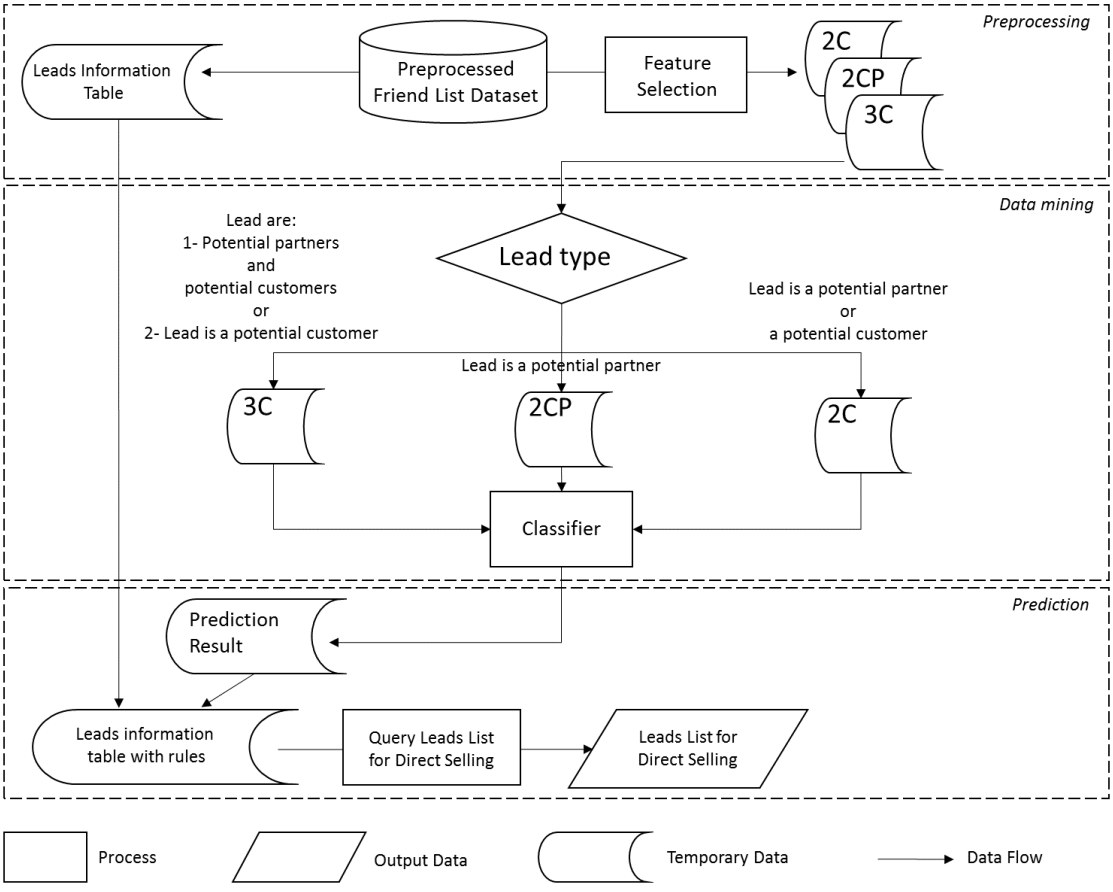


Figure 4.5 The proposed framework for lead generation system for direct selling

In the diagram, several temporarily data files were considered: the 3C (three-classes based dataset), the 2C (two-classes based dataset), and the 2CP (two-classes based dataset). The 2CC file were discarded due to low class sensitivity and accuracy. A further studies and results are conducted and discussed in Chapter 5 to evaluate the class's strength in the dataset.

Chapter 5 **Experimental Results and Evaluation**

5.1 Evaluation Plan

5.2.1 Cross Validation

Cross-validation is a mathematical calculation and statistical division of data and samples into sub-groups [57]. The analysis process is conducted initially on one subgroup while other sub-groups are reserved for later use to verify the initial analysis accuracy. The first sub-group is called the training set, and other sub-groups are called the testing and validation kits. The problem with the use of supervised data mining methods is that they are not given an indication of how the learner, who when asked to predict the dataset not ready yet, was accused of generalizing the process to all future data (which cannot be generalized). To overcome this problem we have to delete some of the data before the training process, and after conducting the training process (by not using the full data set in the training process). The data that has been removed is used in the testing process to calculate the performance, and this is the idea of the (cross-validation).

One of the simplest methods to use cross-validation is the test set method, where the data is split into two groups: the first group used for training and the second for testing. Following that, a regression or classification method is determined according to data used in the training process. Guessing the performance of this approach and by defining the percentage of the error through the used data for testing, accuracy can be improved.

There is a difference and a significant variation in the assessment of the ways of cross-validation depending on the type of points that have been selected in the training and testing process and the way in which the data is divided. The advantages of this approach are its ease of execution, and does not need to a long time for processing the calculation. Its main disadvantage is the loss of part of the training data, up to 30% of the data used in the testing process. The data that used in testing is usually chosen randomly and comprises a third of the whole dataset. The rest used as training data. We can then calculate the error rate (mean square error) to see which of the best algorithm to be used in data mining.

Another method for cross validation is (K-fold Cross Validation). Through this method the data is randomly divided into groups of k, then trained for a number of k time's processes by using all the points in the group with the exception of k.
Action steps algorithm:

- *Data divided randomly into K groups.*
- *In each group we train using all the points that are not belong to this group.*
- *Calculate the total error rate in this group.*
- *Repeat steps 2-3 until we finish all groups.*
- *Calculate the mean to the total error of all groups.*

Some advantages of K-fold cross validation is that if we select the correct value of the variable k, it reduces the use of data in comparison with the process of testing set method, There is no theory to choose the values of k, but common is the choice of k = 10, and sometimes k = 5, depending on the dataset size.

5.2.2 Accuracy for Different Classes

In classifying data, there are two ways of classification for the instances. The first is the three classes approach, and the second is the two classes approach.

In the three classes approach, the classes are: partner, customer, and non-lead. Therefore, there will be six different types of errors: partners being predicted as customers or non-lead, customers being predicted as partner or non-lead, and non-lead being predicted as partners or customers. Also in the two classes approach, the classes are: lead (both partners and customers), non-lead, or partners and non-lead, or customer and non-lead. Therefore, there will be six different types of errors generated from this approach.

Interpreting different algorithms' errors identifies important information about the algorithm, to which extent it is useful for the prediction or not, the results of different classes' accuracies, and the overall accuracy of each method. All these results will be compared in order to find the most suitable methods for this study.

5.2.3 Other Evaluating methods

Other measures were used to evaluate the overall accuracy of the algorithm, and the most important measures included in this study were: Recall, Precision and F-measure.
Recall and precision are concepts in mathematics in computer science that specialize in the field of information retrieval. Recall is measured by calculating the number of results related to research on the number of overall results. While the precision is a standard measured by calculating the number of results related to research on the total recovered results [58]. They can be expressed as the following equations:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

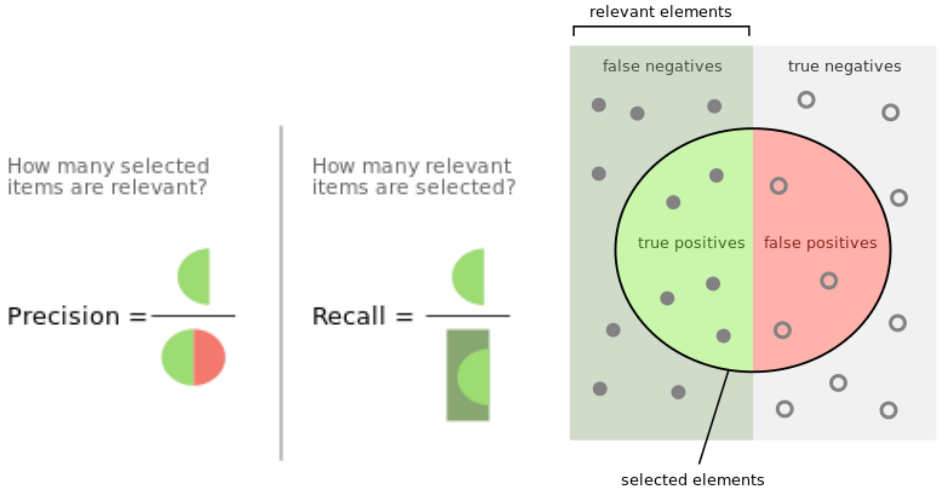


Figure 5.1 Precision and recall

Another important measure is F-measure to evaluate accuracy. It considers both the precision, and the recall, to calculate its score as the following equation:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

It can be described as a weighted of both of precision and recall.

Receiver Operating Characteristic (ROC) curve is a graph that describe the performance of a classifier and examines how good a classifier distinguishes between two things, such as which instance is lead and which instance is not. Better classifiers can correctly say that an instance is a lead or not, whereas the worst classifiers have difficulty in distinguishing between them.

The curve is created by mapping two results, the true positive rate (TPR), and the false positive rate (FPR). TPR is also known as sensitivity or recall in machine learning. FPR known as fall-out which is equal (1- specificity), and can be represented as the following equation:

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

The result of a classifier can be a value that must be determined by a threshold value, or by a discrete class label. A ROC space is represented by FPR and TPR as x and y respectively, which draw relative trade-offs between FPR (cost) and TPR (benefit). Figure shows how the ROC space should appear.

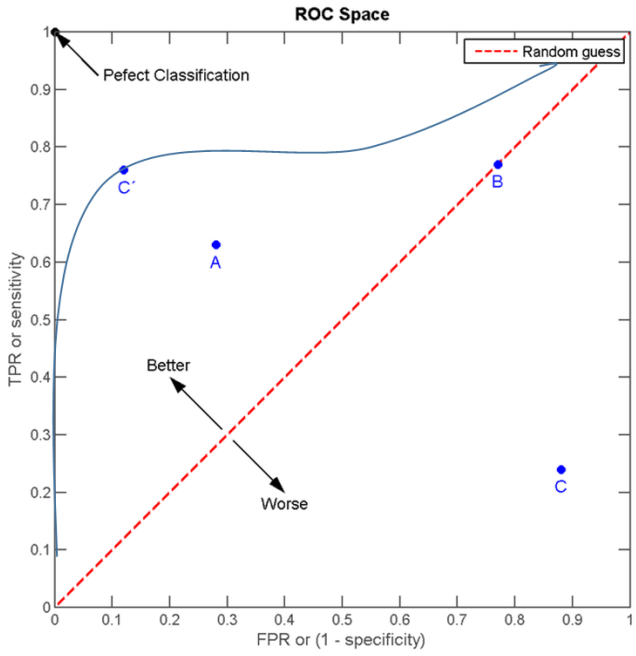


Figure 5.2 ROC Space

The ROC curve is graphed as a direct relationship in the space, while the better classifier's curve is the one that can fill the space more than the other. The area below the curve called Area under Curve (AUC or AUROC). The closest AUC score to value 1 represents the highest accuracy of the classifier.

5.2 Feature Selection Experiment and Evaluation

Two feature selection algorithms were selected, applied and compared to each other; the WrapperSubsetEval, and the ClassifierSubsetEval algorithms. Because they are wrapper feature selection algorithms, the candidate features needed to be checked by a classifier to evaluate subset of features validity.

Naïve Bayes classifier and C4.5 decision tree classifier were chosen for both wrapper methods in order to test preferences. 10-folds cross validation were set up to test the training dataset to the mentioned classifiers in WrapperSubsetEval algorithm. The ClassifierSubsetEval algorithm separates the dataset to training and testing dataset instead to cross validation.

To expand the research and to find the best search method in wrapper feature selection, three different search methods were chosen, the best first (BF), genetic algorithm (GA), and greedy stepwise (GS) methods. Weka data mining tool was used to test the efficiency of both feature selection algorithms with different chosen classifiers and search methods.

5.2.1 Three-Classes Based Datasets Feature Selection

Three versions of three-classes based datasets were applied into experiments. Table 5.1 shows the number of selected features in each experiment.

We can observe that BF and GS search methods selected the same features in most experiments in this study, more in WrapperSubsetEval than ClassifierSubsetEval algorithm.

Dataset	Number of features	Feature Selection Algorithm	Classifier	Search Method	Number of Selected features
MFS 3C	58	WrapperSubsetEval	Naïve Bayes	BF	11
				GA	27
				GS	11
			C4.5	BF	5
				GA	26
				GS	5
		ClassifierSubsetEval	Naïve Bayes	BF	3
				GA	13
				GS	2
			C4.5	BF	12
				GA	25
				GS	10
Transformed T3C	22	WrapperSubsetEval	Naïve Bayes	BF	10
				GA	16
				GS	10
			C4.5	BF	8
				GA	11
				GS	8
		ClassifierSubsetEval	Naïve Bayes	BF	2
				GA	2
				GS	2
			C4.5	BF	15
				GA	14
				GS	11
Factor Table FT3C	12	WrapperSubsetEval	Naïve Bayes	BF	7
				GA	6
				GS	7
			C4.5	BF	8
				GA	8
				GS	8
		ClassifierSubsetEval	Naïve Bayes	BF	5
				GA	6
				GS	5
			C4.5	BF	8
				GA	8
				GS	8

Table Keys		
		Same features subset in one algorithm by one classifier
		Same features subset in all search methods by same classifier

Table 5.1 Summery of 3-classes based datasets after feature selection with different feature selection algorithm and different search methods

Another observation is that selected features in different search methods are the same when used on FT3C dataset and C4.5 classifiers, when the dataset is narrowed down to the most important features, gives more the accurate results. Table 5.2 interprets:

Dataset	Accuracy before Feature Selection		Feature Selection Algorithm	Classifier	Search Method	Accuracy After Feature selection	
						Train/ Test	5 folds CV
MFS 3C	Naïve Bayes		WrapperSubsetEval	Naïve Bayes	BF	55.26%	60.76%
	Train/Test	5 folds CV			GA	54.39%	58.71%
	52.82%	51.16%		C4.5	BF	63.94%	63.27%
					GA	66.67%	65.03%
	J48		ClassifierSubsetEval	Naïve Bayes	BF	38.59%	38.36%
	Train/Test	5 folds CV			GA	41.13%	39.70%
	66.86%	65.20%		C4.5	BF	66.47%	62.92%
					GA	66.67%	63.09%
					GS	65.89%	62.45%
T3C	Naïve Bayes		WrapperSubsetEval	Naïve Bayes	BF	66.86%	67.54%
	Train/Test	5 folds CV			GA	66.27%	67.66%
	67.44%	66.96%		C4.5	BF	71.34%	72.16%
					GA	65.11%	62.45%
	J48		ClassifierSubsetEval	Naïve Bayes	BF	38.40%	42.92%
	Train/Test	5 folds CV			GA	38.40%	42.92%
	67.44%	67.72%		C4.5	BF	70.56%	70.40%
					GA	72.12%	70.40%
					GS	72.12%	69.82%
FT3C	Naïve Bayes		WrapperSubsetEval	Naïve Bayes	BF	63.74%	64.79%
	Train/Test	5 folds CV			GA	63.15%	64.26%
	64.71%	64.67%		C4.5	BF	70.95%	72.11%
					GA	70.95%	72.11%
					GS	70.95%	72.11%
	J48		ClassifierSubsetEval	Naïve Bayes	BF	61.79%	64.50%
	Train/Test	5 folds CV			GA	61.79%	65.09%
	70.37%	72.04%		C4.5	BF	70.56%	70.11%
					GA	70.56%	70.11%
					GS	70.56%	70.11%

Table Keys		
		Bad accuracy after feature selection with big difference from original accuracy
		Better accuracy with considerable difference or improvement.

Table 5.2 Accuracy of Naïve Bayes and C4.5 on selected features

Due to the similarity in the selected features between BF and GS search methods, the accuracies related GS in some experiments are not shown. We can generally observe the accuracy of the classifiers on datasets before using the feature selection algorithms increase when we narrow down to the FT3C dataset, in both 70 to 30 training/testing test, and 5-folds cross validation test.

Another observation is that WrapperSubsetEval performs better than ClassifierSubsetEval in selecting features in all searching methods by Naïve Bayes classifier, and most of the experiments conducted in C4.5 classifier. The selected features in the green-marked experiments were compared to each other, the most frequent selected features were already chosen by WrapperSubsetEval algorithm using C4.5 classifier.

5.2.2 Two-Classes Based Datasets Feature Selection

Three versions of two-classes based datasets, defined by three different picked classes, were applied into the experiments. Table 5.3 shows the number of selected features in each experiment. As shown in the table, many of experiments tested on the two-classes based datasets did not elect any features, especially when the GS search method was used, and less so in BF experiments. This is an indication that the GS searching method is not suitable to use in this part. That problem is more obvious in 2Cs and 2CCs datasets.

This can be interpreted by the data in 2CPs datasets, which is more distinguishable than the other type, and allows the classifier to learn and to select the features better. Likewise in the three-classes based feature selection, similarity of selected features among different searching methods increases when we narrow down the numbers of original features. The number of failed experiments, with no selected features, were done by ClassifierSubsetEval algorithm, and by C4.5 classifier more than Naïve Bayes. Therefore, ClassifierSubsetEval is a weaker algorithm for this study.

Table 5.4 shows that experimental accuracies improved more than before by feature selection in the experiments using WrapperSubsetEval algorithm more than ClassifierSubsetEval, and experiments that used C4.5 more than Naïve Bayes. Another interesting finding is that some experiments has poor results in precision, recall and f-measure when classification was conducted on dataset before feature selection, but provides more accuracy after features are selected by both algorithms.

Dataset	# features	Feature Selection Algorithm	Classifier	Search Method	# Selected features From 2Cs	# Selected features From 2CPs	# Selected features From 2CCs
MFS Datasets	58	WrapperSubsetEval	Naïve Bayes	BF	3	15	12
				GA	5	30	20
				GS	3	11	7
			C4.5	BF	N/A	3	N/A
				GA	3	15	3
				GS	N/A	3	N/A
MFS 2C		ClassifierSubsetEval	Naïve Bayes	BF	10	2	1
				GA	26	18	22
MFS 2CP				GS	10	2	1
			C4.5	BF	N/A	6	N/A
MFS 2CC				GA	3	12	3
				GS	N/A	4	N/A
Transformed Datasets	22	WrapperSubsetEval	Naïve Bayes	BF	4	4	1
				GA	12	14	8
				GS	4	2	1
			C4.5	BF	N/A	6	6
T2C				GA	12	6	10
				GS	N/A	6	N/A
T2CP		ClassifierSubsetEval	Naïve Bayes	BF	5	1	1
				GA	2	1	1
T2CC				GS	5	1	1
			C4.5	BF	10	15	11
				GA	12	14	10
				GS	N/A	5	N/A
Factor Table Datasets	12	WrapperSubsetEval	Naïve Bayes	BF	N/A	2	8
				GA	9	2	6
				GS	N/A	2	N/A
			C4.5	BF	8	5	N/A
				GA	9	4	6
FT2C				GS	N/A	4	N/A
		ClassifierSubsetEval	Naïve Bayes	BF	N/A	3	N/A
FT2CP				GA	4	3	5
				GS	N/A	3	N/A
FT2CC			C4.5	BF	8	10	9
				GA	7	10	9
				GS	N/A	10	N/A

Table Keys		
		Bad feature Selection
		Same features subset in all search methods by same classifier
		Same features subset in one algorithm by one classifier
	N/A	No features selected by this method

Table 5.3 Summery of 2-classes based datasets after feature selection with different feature selection algorithm and different search methods

Dataset	Accuracy before Feature Selection			Feature Selection Algorithm	Classifier	Search Method	Accuracy After Feature selection		
							2Cs	2CPs	2CCs
MFS 2C MFS 2CP MFS 2CC	Naïve Bayes			WrapperSubsetEval	Naïve Bayes	BF	75.09%	84.79%	76.08%
	2C	2CP	2CC			GA	70.87%	81.57%	73.68%
	47.95%	78.07%	63.27%		C4.5	BF	75.08%	67.66%	67.66%
	J48					GA	75.08%	66.31%	67.66%
	2C	2CP	2CC	ClassifierSubsetEval	Naïve Bayes	BF	48.83%	57.06%	N/A
	75.08%	73.91%	67.66%			GA	43.15%	75.14%	66.37%
					C4.5	BF	75.08%	67.60%	67.66%
						GA	75.08%	65.09%	67.66%
T2C T2CP T2CC	Naïve Bayes			WrapperSubsetEval	Naïve Bayes	BF	75.26%	82.69%	N/A
	2C	2CP	2CC			GA	79.64%	81.87%	72.39%
	78.36%	80.81%	59.53%		C4.5	BF	78.36%	84.15%	75.03%
	J48					GA	86.14%	84.09%	75.14%
	2C	2CP	2CC	ClassifierSubsetEval	Naïve Bayes	BF	55.20%	N/A	N/A
	84.97%	82.16%	67.66%			GA	75.20%	N/A	N/A
					C4.5	BF	79.47%	82.33%	73.85%
						GA	81.70%	82.63%	74.27%
FT2C FT2CP FT2CC	Naïve Bayes			WrapperSubsetEval	Naïve Bayes	BF	79.53%	82.45%	71.28%
	2C	2CP	2CC			GA	79.64%	82.45%	71.69%
	79.53%	80.00%	57.25%		C4.5	BF	85.03%	84.56%	75.02%
	J48					GA	85.09%	84.09%	75.02%
	2C	2CP	2CC	ClassifierSubsetEval	Naïve Bayes	BF	79.53%	82.28%	57.25%
	85.03%	84.15%	73.91%			GA	75.56%	82.28%	71.22%
					C4.5	BF	84.85%	83.56%	73.57%
						GA	84.91%	83.56%	73.57%

Table Keys		
		Classifier not works on dataset and cannot predict / Bad F-measure result
		Good accuracy after feature selection, while same classifier not works on dataset before feature selection
		Good accuracy and considerable improvement in comparison with accuracy before feature selection
	N/A	No features were selected

Table 5.4 Accuracy of Naïve Bayes and C4.5 on selected features

The selected features in all the green-marked experiments were compared to each other, and the most frequent selected features already chosen by WrapperSubsetEval algorithm using C4.5 classifier and GA search method.

The final selected feature subsets for 3C, 2C, 2CP, and 2CC datasets are presented in table 3.4 in chapter 3. To test the feature selection good impact on machine learning, the chosen data mining classification algorithms were applied to the datasets with the selected features and compared to the factor tables' datasets before feature selection was conducted and the percentage of improvement calculated.

Table 5.5 shows the improvement of accuracy in 3-classes based datasets, while table 5.6 shows the improvement of accuracy in all 2-classes based datasets.

Classifier	Accuracy of FT3C	Accuracy after FS	Improvement %
SVM	49.1%	72.1%	+22.9%
KNN	63.2%	69.8%	+6.59%
NB	64.2%	63.3%	-0.9%
J48	72.5%	72.6%	+0.1%
RF	66.9%	70.8%	+3.9%
Ensemble	72.2%	72.7%	+0.5%

Table 5.5 Improvement percentage of accuracy on 3-classes based datasets before and after feature selection using six different classification algorithms.

Classifier	Improvement % on 2C	Improvement % on 2CP	Improvement % on 2CC
SVM	+12.9%	+18.8%	+5.4%
KNN	+5.5%	+7.9%	+3.4%
NB	+0.4%	+1.0%	+14.3%
J48	-0.7%	+1.2%	-2.0%
RF	+4.1%	+4.6%	+0.5%
Ensemble	+0.4%	+1.1%	-1.8%

Table 5.6 Improvement percentage of accuracy on different 2-classes based datasets before and after feature selection using six different classification algorithms.

5.3 Classification Experiment and Evaluation

Six classification algorithms were selected, applied and compared to each other; Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), C4.5 Decision Tree Algorithm (J48), Random Forest Decision Tree (RF) and ensemble data mining algorithms. The Weka data mining tool was used to test the efficiency of all mentioned algorithms by using the knowledge flow interface to design a model to deal with different datasets in the study.

5.3.1 Three-Classes Based Dataset Classification

The classification accuracies of the six classifiers on three-classes based dataset, using 10-folds cross validation test, are shown in table 5.7. The table also shows the accuracy of each class in the dataset. It is noticeable that Naïve Bayes algorithm got the lower overall accuracy, which is 63.3%, and highest overall accuracy done by the ensemble classification method.

C4.5 and SVM algorithms are considered fair in comparison with other tested classifiers. It is significantly obvious that the accuracies of the non-lead class are higher than other classes' accuracies, then partner class, and customer class is the lower. The interpretation of these results is that non-lead class supposed to have the rest of instances in the dataset that had not classified as partner or customer in labeling phase in the dataset preprocessing that means that non-lead class has the instances with least valued features to be classified as a lead.

On the other hand, the partner class has fair accuracy, better than customer class, because it had been defined by an additional feature to the interests and activities features, identified as a work feature. The work feature carries valuable information to define partners, but cannot be used to define customer class. The features of interests, activities, movies, music, books, and quotes has various values making the customer definition more difficult.

Furthermore, the concept of classifying the instances in the preprocessing phase relied on a fact in direct selling industry that every partner is a customer in the first place, but not every customer wants to work in this field. This fact makes defining customers less important than partners, and that’s why the classifying process classifies partner instances firstly, then classify customers secondly.

The goal of the ensemble method was achieved and enhanced slightly more than other classifiers, since it scores the highest accuracy, but still it is not confident to say that it is the best classifier to rely on in building the leads generation prediction model.

Classifier	Partner Accuracy			Customer Accuracy			Non Lead Accuracy			Overall Accuracy
Classified Result	Correct	Incorrect		Correct	Incorrect		Correct	Incorrect		Correct
		Customer	None		Partner	None		Partner	Customer	
SVM	82.9%	11.2%	5.9%	46.3%	27.7%	26.0%	87.1%	3.1%	9.8%	72.1%
KNN	73.3%	20.7%	5.9%	52.3%	22.2%	25.5%	86.4%	2.3%	11.3%	69.8%
NB	58%	27.9%	14.1%	46.8%	11.4%	41.7%	93.7%	0.5%	5.9%	63.3%
C.45	78%	16%	6%	54.1%	20.1%	25.8%	87.3%	0.7%	12%	72.5%
RF	74.1%	19.6%	6.3%	54.1%	20.6%	25.3%	86.6%	1.6%	11.7%	70.8%
Ensemble	78.5%	15.6%	5.9%	54.2%	20.4%	25.3%	86.6%	1.2%	12.2%	72.7%

Table 5.7 The overall accuracy and accuracies of each class in 3-classes based dataset

C4.5 performs well in comparison with other classifiers, and performs the best in relying on single classifiers in leads prediction. It surpassed the other decision tree (RF) classifier's results. The interpretation is that C4.5 adopts post pruning algorithm, unlike RF which needs to set up a value of `ntree` to prune the decision tree. Figure 5.3 show the error rate on of the classifiers that were used to predict leads generation in the 3-classes based dataset.

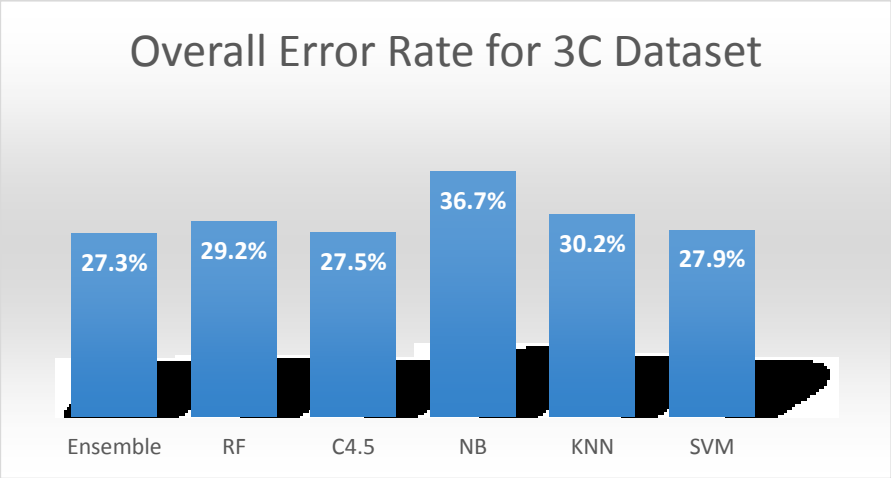


Figure 5.3 The rate of overall error that made by different classifier.

Table 5.8 shows the results of precision, recall and F-measure for each classification method. The ensemble method, C4.5 and RF, and SVM have better balanced distribution of precision, recall and F-measure.

Classifier	Partner			Customer			Non Lead			Overall		
Evaluation measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM	0.785	0.829	0.806	0.674	0.463	0.549	0.665	0.871	0.754	0.719	0.721	0.710
KNN	0.801	0.733	0.766	0.591	0.523	0.555	0.667	0.864	0.753	0.7	0.698	0.694
NB	0.867	0.580	0.695	0.531	0.468	0.498	0.544	0.937	0.689	0.678	0.633	0.630
C4.5	0.833	0.780	0.806	0.640	0.541	0.586	0.665	0.873	0.755	0.729	0.726	0.722
RF	0.817	0.741	0.778	0.608	0.541	0.572	0.665	0.866	0.752	0.712	0.708	0.705
Ensemble	0.829	0.785	0.807	0.644	0.542	0.589	0.668	0.866	0.755	0.729	0.727	0.723

Table 5.8 The precision, recall and F-measure accuracies of each class in 3-classes based datasets

On the other hand, Naïve Bayes had imbalanced distribution of accuracy in the precision, recall and F-measure, and that refers to the probabilistic nature of the algorithm, which is not best suited for the way this data has been modeled.

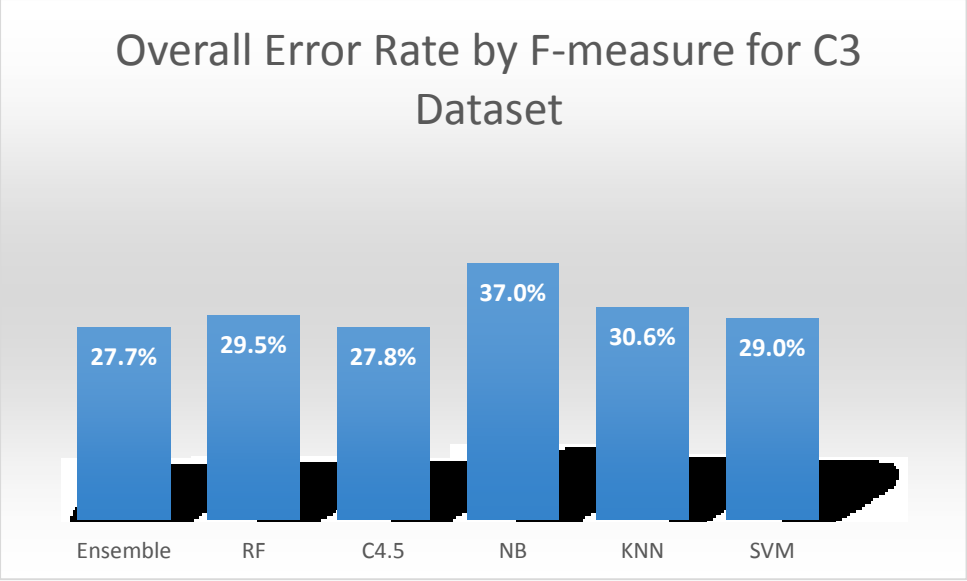


Figure 5.4 The rate of overall error by f-measure

In order to evaluate the targeted classes in the 3C dataset by visualization, the AUROC of partner class and customer class has been calculated. Table 5.9 shows the results of AUROC for these classes in for each classifier. The results shows that partner class is much better and has finer sensitivity to predict partners, but the results of customer has less sensitivity, and that can be interpreted by ROC curves in figure 5.5.

Classifier	AUROC Value Partner	AUROC Value Customer
SVM	0.830	0.678
KNN	0.813	0.727
NB	0.887	0.761
C4.5	0.882	0.749
RF	0.877	0.756
Ensemble	0.830	0.694

Table 5.9 AUROC values of partner and customer classes in 3C dataset

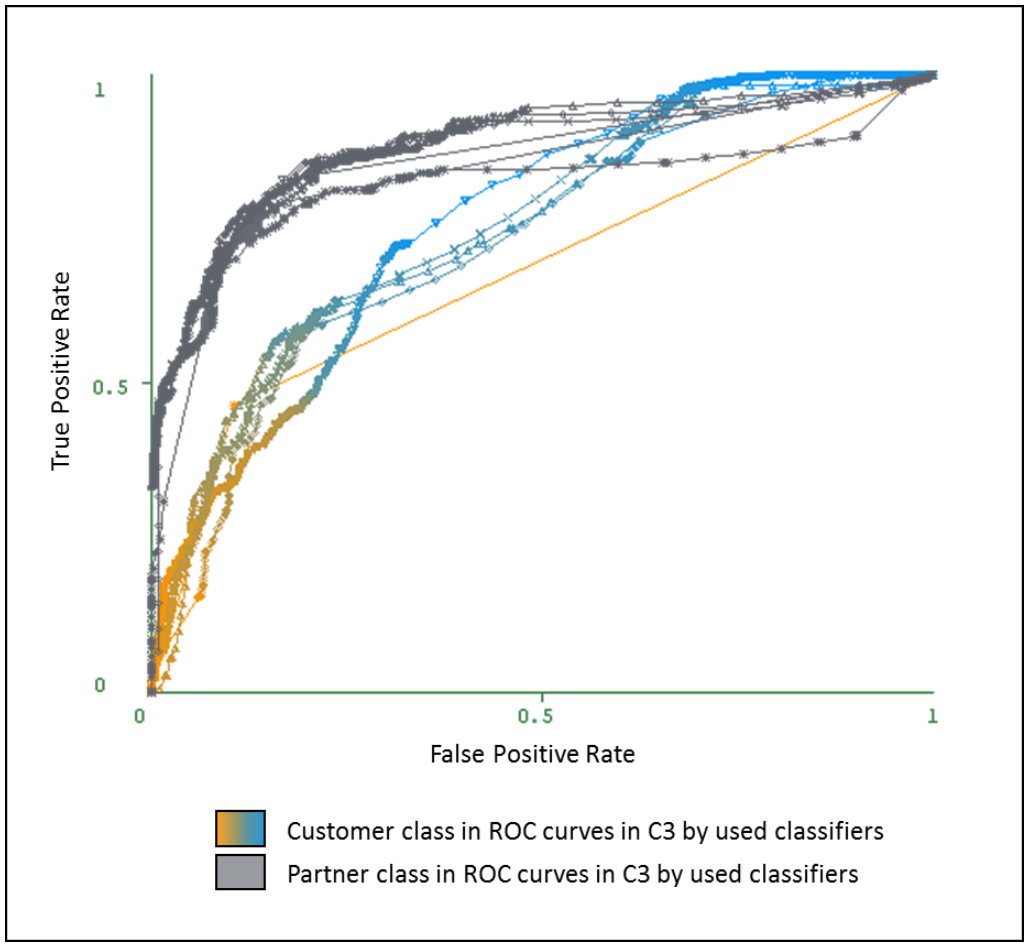


Figure 5.5 Group of ROC curves for partner class is closer to perfect classification than the group of curves for customer class in 3C dataset, which means the partner class has a better sensitivity to predict partner leads.

5.3.2 Two-Classes Based Dataset Classification

The classification accuracies of the six classifiers on two-classes based dataset, using 5-folds cross validation test, are shown in table 5.10. The table also shows the accuracy of

each class in the dataset. In 2C dataset, the one that treats both potential partners and potential customers as a one lead, it is noticeable that Naïve Bayes algorithm got the lower overall accuracy, 79.3%, while all other classifiers are close to each other in accuracy, and considered reliable classifiers. The same happens with 2CP dataset, that considers leads who are those the potential partners only.

	Classifier	Lead Accuracy	Non-lead Accuracy	Overall Accuracy
2C	SVM	86.4%	86.6%	86.5%
	KNN	86.3%	86.6%	86.4%
	NB	75.0%	94.1%	79.8%
	C4.5	87.6%	84.3%	85.1%
	RF	86.3%	86.6%	86.4%
	Ensemble	86.4%	86.6%	86.5%
	Classifier	Lead Accuracy	Non-lead Accuracy	Overall Accuracy
2CP	SVM	86.4%	86.6%	86.5%
	KNN	86.3%	86.6%	86.4%
	NB	75.0%	94.1%	79.8%
	C4.5	87.6%	84.3%	85.1%
	RF	86.3%	86.6%	86.4%
	Ensemble	86.4%	86.6%	86.5%
	Classifier	Customer Accuracy	Non-lead Accuracy	Overall Accuracy
2CC	SVM	30.4%	92.0%	72.1%
	KNN	32.5%	89.5%	71.1%
	NB	28.8%	92.1%	71.6%
	C4.5	32.7%	90.8%	72.0%
	RF	35.1%	88.8%	71.4%
	Ensemble	31.1%	91.9%	72.2%

Table 5.10 The overall accuracy and accuracies of each class in 2-classes based datasets

An interesting outcome in the 2CC dataset, is that it has the customer class in the lead, with very low percentage for all classifiers. That strongly suggest the idea that the customer class needs to be remodeled and redefined in a better way, especially because most partners are customers by the nature of direct selling at the first place, and they are not included in 2CC customer class.

Defining people who are just interested in shopping and purchasing needs further analysis to be conducted, especially on the imported data from social media, and more specifically, posts by text classification. Figure 5.6 shows the error rates that calculated by F-measure for the two-classes based datasets for all tested classifiers. It confirms also a weakness in Naïve Bayes algorithms in all classifying styles for this study as the three-classes based dataset.

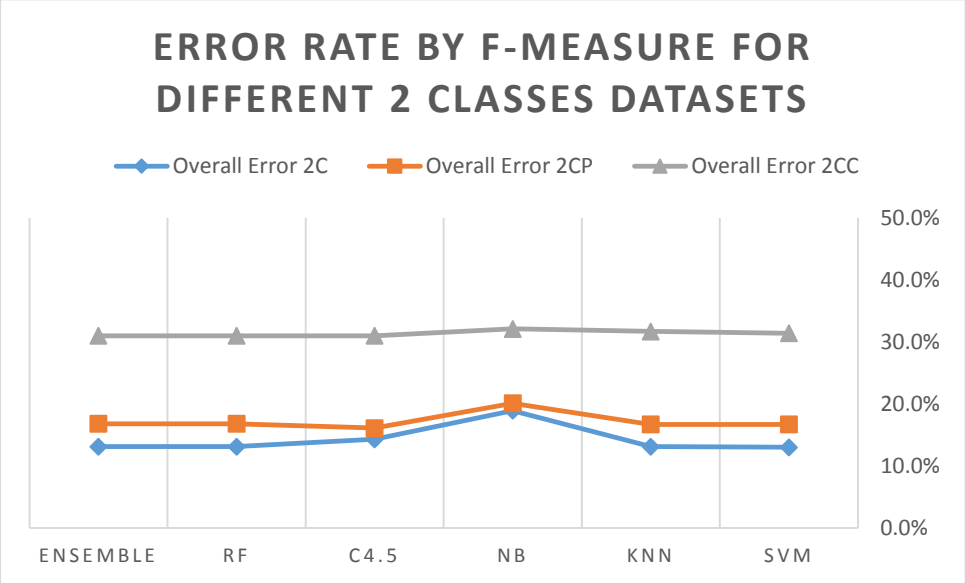


Figure 5.6 Error rate by F-measure foe two-classes based dataset for different classifiers.

5.3.3 Other Findings

As presented in section 3.3.4, there are four types of datasets for lead generation has been created. By applying ROC curves plots for each class in all datasets in the study and compare a particular class's AUROC value to itself in different dataset, a finding of which the perfect dataset should be used in predicting a particular lead.

Table 5.11 shows AUROC values of partner class in two different datasets: 3C, and 2CP dataset. The table shows strong class sensitivity in both datasets. The difference is not too big between results, and that leads to the suitability of both datasets to be learned in order to find partners leads. In this study, 2CP dataset is the suitable one.

Classifier	AUROC Value Partner 3C	AUROC Value Partner 2CP
SVM	0.830	0.825
KNN	0.813	0.870
NB	0.887	0.882
J48	0.882	0.877
RF	0.877	0.881
Ensemble	0.830	0.831

Table 5.11 The AUROC values of partner class in 3C and 2CP dataset

On the other hand, table 5.12 shows AUROC values of customer class in two different datasets: 3C, and 2CC datasets. It shows how customer class in 2CC dataset is poor, so the dataset that made for customer leads is useless. The better way to generate only customer is to use 3C dataset for classification learning.

Classifier	AUROC Value Customer 3C	AUROC Value Customer 2CC
SVM	0.678	0.612
KNN	0.727	0.654
NB	0.761	0.648
J48	0.749	0.656
RF	0.756	0.644
Ensemble	0.694	0.615

Table 5.12 The AUROC values of customer class in 3C and 2CC dataset

Chapter 6 **Conclusions and Future Work**

6. 1 Conclusions

Leads generation is an active field for studies and researchers. It creates interest for many people who work in the marketing field, and more specifically, in the direct selling field. Currently, every field of business embraces information systems, and works with information technologies to develop and enhance that field. Using social media in leads generation for direct selling is becoming an important step for most of marketers. The Facebook website is the widest social media application that has the largest number of users, and an interesting source for many businesses opportunities, including the type of businesses that takes the casual impress, such as: part-time, or business with friends, and that what applies to direct selling. Therefore, a classification technique on social media data that can predict the best leads is needed. Classifying a set of features of the Facebook friends can be a useful tools for direct sellers to help them in the industry.

Some contributions in this study have been provided in order to activate this research area by comparing different classification methods to and evaluating its accuracies, in addition to developing and assembling a classification method that aims to improve a classifiers performance. A deeper study in feature selection approach was conducted in order to build different datasets with selected features for different lead generation tasks, and that resulted in closer accuracy values among most of chosen classifiers for each task. Finally, the poor performance of customer class calls for the necessity of enhancing the class model by enriching the data by with more features that define an interested customer.

It is suggested for the developers or companies who are interested in this study to strengthen the leads classes, by adding new features such as analyzing posts and news feed by text classification, or inspecting uses heuristics and clicks, or probably evaluate pictures by image processing, to gain better modeling of classes, to get better accuracy. Building recommender system is a future goal of this study.

Since the evaluation experiment of this research is based on the dataset generated from my own personal Facebook friend list, it might have a limitation on the representativeness for people's age group, demographical characteristic and shared interests etc. It is recommended that the developers to expand the data to include more possible instances and evaluate the results also based on different sources and zones in Facebook.

6. 2 Achieved Objectives

The research supplied information system research in the industry of direct selling to help improve the industry and the performance of the direct sellers. Data from Facebook preprocessed, and remodeled it in a way that makes it ready for machine learning. Deep experiments has been made on feature selection to obtain the most important features of the dataset. This study improved the classification prediction results. Furthermore, the research applied five different classification algorithms on the feature selected datasets and evaluated its accuracies, as well as designed a stacked ensemble classification model to enhance algorithms performance, and built the classification prediction to use in leads generation enquiry.

6. 3 Study Difficulties

The field of direct selling has been lacking of data mining research, in particular on business lead prediction by applying suitable classification methods. This needs a careful understanding of different classification studies, in different business fields and goals.

Another challenge is the lack of access to the Facebook API Graph which allows developers to use Facebook data for analysis purposes. Due to the recent change in policies of some open sourced importing tools such as R Studio, or the one that was chosen, NodeXL, I imported my personal Facebook friend list dataset which reached to 1800 friends. Later, I tried to increase the imported data from Facebook because it will enhance the study, but Facebook denied the accessibility. Other social media websites still allow imported data, but it was my assumption that it would not help the study.

Finally, Some dataset component features have been lost, and because the same difficulty that was mentioned in the above, I could not re-import the dataset from Facebook, so I was forced to re-enter it manually by visiting each friend page to add these features, which took much time and effort.

6. 4 Future Work

Conduct the research on larger amount of data by having more accessibility to Facebook database. Better delineation of the customer class is required, by having more features that express the attitude of purchasing on-line, or consuming a specific type of products, in order to enhance this prediction results. Different features from Facebook or other integrated sources can be generated, such as customer posts which can be analyzed by text classification, and by employing customer visiting page heuristics. A recommender system can be built for direct sellers that use association rule, and implement it in the Facebook App Center as a component of Facebook itself.

References

[1] “Direct Selling 411” *Direct Selling 411*, 2016. [Online]. Available: http://www.directselling411.com/. [Accessed: 23- Mar- 2016].

[2] C.King, J.Robinson, The New Professionals, *Prima publishing*, 2000.

[3] K.Laudon and C.Traver, E-Commerce 2014: Business. Technology. Society*,* *Prentice-Hall*, Inc. by edition 10

[4] S. Liao, Y. Chen and H. Hsieh, "Mining customer knowledge for direct selling and marketing", *Expert Systems with Applications*, vol. 38, no. 5, pp. 6059-6069, 2011.

[5] "Facebook", *Facebook.com*, 2016. [Online]. Available: https://www.facebook.com/facebook/info. [Accessed: 23- Mar- 2016].

[6] “Statista”, *Statista.com,* [Online] Available: http://www./statistics/282087/number-of-monthly-active-twitter-users/ [Accessed: 23- Mar- 2016].

[7] “Direct Selling Association | Ethics, Trust, Confidence” *Dsa.org*, 2016. [Online]. Available: http://www.dsa.org/. [Accessed: 23- Mar- 2016].

[8] "World Federation of Direct Selling Associations", *Wfdsa.org*, 2016. [Online]. Available: http://www.wfdsa.org/. [Accessed: 23- Mar- 2016].

[9] P. Vander Nat and W. Keep, "Marketing Fraud: An Approach for Differentiating Multilevel Marketing from Pyramid Schemes", *Journal of Public Policy & Marketing,* vol. 21, no. 1, pp. 139-151, 2002.

[10] M. Morley, The Big Book of Network Marketing Compensation Plans. *Nookbook*, 2012.

[11] Robert T. Kiyosaki, Rich Dad Poor Dad, *Plata Publishing*, LLC, 2000.

[12] “ProBlogTricks”, *Problogtricks.com*, [Online]. Available: http://www.problogtricks.com/15/network-marketing-quotes.html [Accessed: 23- Mar- 2016].

[13] “NetworkMarketing”, *Networkmarketing.se*, [Online] Available: http://www.networkmarketing.se/bill-clinton/ [Accessed: 23- Mar- 2016].

[14] “DubLi”, *DubliNetwork.net.* 2016 [Online]. Available: http://www.Dublinetwork.net/ [Accessed: 23- Mar- 2016].

[15] "WOR(l)D Global Network", *WOR(l)D*, 2016. [Online]. Available: http://www.worldgn.com. [Accessed: 23- Mar- 2016].

[16] “WhatIs”, *Whatis.techtarget.com*, [Online] Available: http://whatis.techtarget.com/definition/lead-generation [Accessed: 23- Mar- 2016].

[17] “Direct Selling Education Foundation”, *Dsef.org*, [Online] Available: http://www.dsef.org/2012/01/20/5-lead-generating-ideas-for-small-business/ [Accessed: 23- Mar- 2016].

[18] A. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media", *Business Horizons*, vol. 53, no. 1, pp. 59-68, 2010.

[19] T.Baruah, "Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study" *International Journal of Scientific and Research Publications*, Volume 2, Issue 5, May 2012

[20]"Alexa - Actionable Analytics for the Web", Alexa.com, 2016. [Online]. Available: http://www.alexa.com/. [Accessed: 23- Mar- 2016].

[21] "Twitter", *Twitter.com*, 2016. [Online]. Available: http://www.twitter.com. [Accessed: 23- Mar- 2016].

[22] "Instagram", *Instagram.com*, 2016. [Online]. Available: http://www.instagram.com. [Accessed: 23- Mar- 2016].

[23] "YouTube", *Youtube.com*, 2016. [Online]. Available: http://www.youtube.com. [Accessed: 23- Mar- 2016].

[24] J. Pei, J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Third Edition ed., *Morgan Kaufmann*, 2011.

[25] U. Stańczyk, L.Jain, Feature Selection for data and pattern recognition, *Springer*, 2015.

[26] I.Guyon (et al.), Feature extraction, Berlin: *Springer*. 2005.

[27] B. Zupan and J. Demsar, "Open-Source Tools for Data Mining", *Clinics in Laboratory Medicine*, vol. 28, no. 1, pp. 37-54, 2008.

[28]"Weka 3 - Data Mining with Open Source Machine Learning Software in Java", *Cs.waikato.ac.nz*, 2016. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 23- Mar- 2016].

[29]"R: The R Project for Statistical Computing", *R-project.org*, 2016. [Online]. Available: https://www.r-project.org/. [Accessed: 23- Mar- 2016].

[30] U. Bioinformatics Laboratory, "Orange Data Mining", *Orange.biolab.si*, 2016. [Online]. Available: http://orange.biolab.si/. [Accessed: 23- Mar- 2016].

[31] "ELKI Data Mining Framework", *Elki.dbs.ifi.lmu.de*, 2016. [Online]. Available: http://elki.dbs.ifi.lmu.de/. [Accessed: 23- Mar- 2016].

[32] "KNIME | Open for Innovation", *Knime.org*, 2016. [Online]. Available: https://www.knime.org/. [Accessed: 23- Mar- 2016].

[33] F.Ricci, L.Rokach and B.Shapira, "Recommender Systems Handbook, New York, *Springer*, 2011.

[34] "Netflix" *Netflix.com*, 2016. [Online]. Available: http://www.netflix.com. [Accessed: 23- Mar- 2016].

[35] "Amazon Associates" *Associates.Amazon.ca, 2016.* [Online]. Available: https://affiliate-program.amazon.com/gp/associates/join/landing/main.html [Accessed: 23- Mar- 2016].

[36] "CallidusCloud | Marketing Automation", *Calliduscloud.com*, [Online] Available: http://www.leadformix.com/data-mining.html [Accessed: 23- Mar- 2016].

[37] "eMarketer | Digital Marketing Research and Insight", *eMarketer.com,* [Online] Available: http://www.emarketer.com/Article/Marketers-Put-First-Party-Data-First/1012663 [Accessed: 23- Mar- 2016].

[38] "B2B Lead Generation Trends for 2014: What's Hot and What's Not [Infographic]", *MarketingProfs*, 2016. [Online]. Available: http://www.marketingprofs.com/chirp/2013/12161/b2b-lead-generation-trends-for-2014-whats-hot-and-whats-not. [Accessed: 23- Mar- 2016].

[39] Koletas T., Maximizing Lead Scoring & Analytics: How to Use Big Data in B2B. "Marketing Land". Internet: http://marketingland.com/maximizing-lead-scoring-analytics-use-big-data-b2b-101956, 2015 [Accessed: 23- Mar- 2016].

[40] S. Crone, S. Lessmann and R. Stahlbock, "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing", *European Journal of Operational Research,* vol. 173, no. 3, pp. 781-800, 2006.

[41] M. Shaw, C. Subramaniam, G. Tan and M. Welge, "Knowledge management and data mining for marketing", *Decision Support Systems*, vol. 31, no. 1, pp. 127-137, 2001.

[42] A. Ortigosa, R. Carro and J. Quiroga, "Predicting user personality by mining social interactions in Facebook", *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 57-71, 2014.

[43] Y. Wan, Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis", IEEE, 2015

[44] W. Wu," An Integrated CRM Data Mining Method for Predicting Best Next Offer", Thesis, *Dalhousie University*, 2006

[45] Md. Abdur Rahman," A Data Mining Framework for Automatic Online Customer Lead Generation ", Thesis, *Dalhousie University*, 2012

[46] "NodeXL", *Nodexl.codeplex.com*, [Online] Available: https://nodexl.codeplex.com/ [Accessed: 23- Mar- 2016].

[47] H. Farreny and H. Prade, "Heuristics—intelligent search strategies for computer problem solving, by Judea Pearl. (Reading, Ma: Addison-Wesley, 1984)", *International Journal of Intelligent Systems*, vol. 1, no. 1, pp. 69-70, 1986.

[48] C.Cortes, V.Vapnik, Support-vector networks, *Machine Learning*, 1995, Volume 20, Number 3, Page 273

[49] L. Auria and R. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis", *SSRN Electronic Journal*. 2008.

[50] N. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression ". *The American Statistician*, vol. 46, no. 3, p. 175, 1992.

[51] S. Russell and P. Norvig, "A modern, agent-oriented approach to introductory artificial intelligence", *ACM SIGART Bulletin*, vol. 6, no. 2, pp. 24-26, 1995.

[52] S. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan *Kaufmann Publishers, Inc.*, 1993", Mach Learn, vol. 16, no. 3, pp. 235-240, 1994.

[53] H. Kam. Random Decision Forests Proceedings, *the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995. pp. 278–282.

[54] L. Breiman, "Random Forests". *Machine Learning* 45 (1): 5–32. 2001.

[55] Andy L., "Documentation for R package randomForest" (16 October 2012)

[56] Opitz, D.; Maclin, R, " Popular ensemble methods: An empirical study " Journal of Artificial Intelligence Research 11: 169–198. (1999)

[57] Seymour G., "Predictive Inference" *Chapman and Hall* New York, NY: Research (1993).

[58] C. Van Rijsbergen, Information retrieval. London: *Butterworths*, 1979.

[59] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[60] L. Torgo, Data Mining with R, learning with case studies, Boca Raton, FL, *Champan & hall/CRC*, 2010.

Appendix A: Social Sciences & Humanities Research Ethics - Letter of Approval

June 18, 2015

Mr Ahmed Balfagih
Computer Science\Computer Science

Dear Ahmed,

REB #: 2014-3287
Project Title: Direct Selling Business Lead Prediction by Data Mining on Social Media Data

Expiry Date: June 27, 2016

The Social Sciences & Humanities Research Ethics Board has reviewed your annual report and has approved continuing approval of this project up to the expiry date (above).

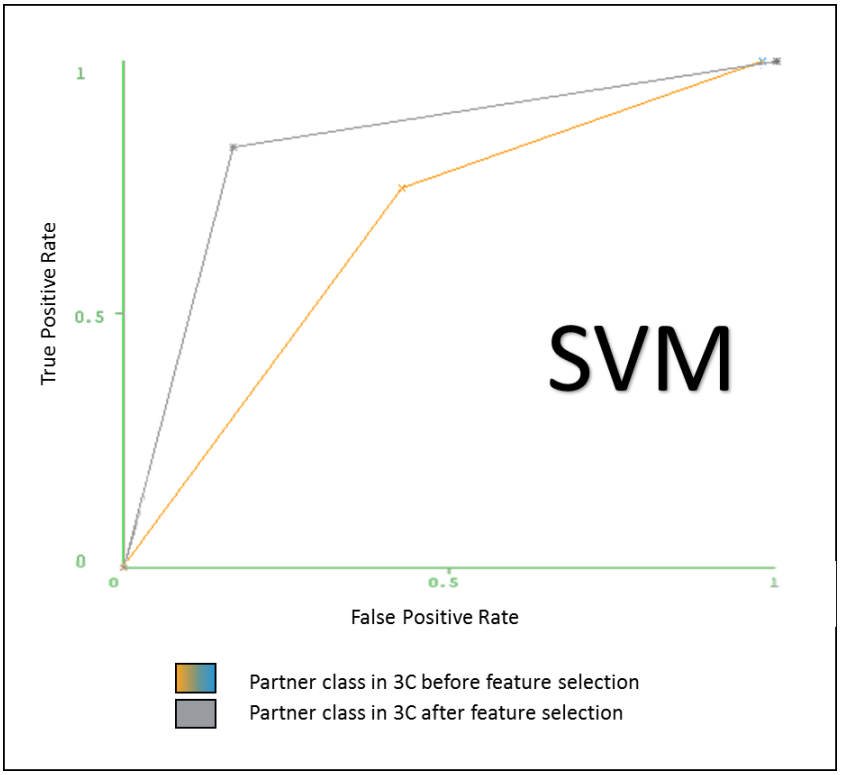
REB approval is only effective for up to 12 months (as per TCPS article 6.14) after which the research requires additional review and approval for a subsequent period of up to 12 months. Prior to the expiry of this approval, you are responsible for submitting an annual report to further renew REB approval. Forms are available on the Research Ethics website.

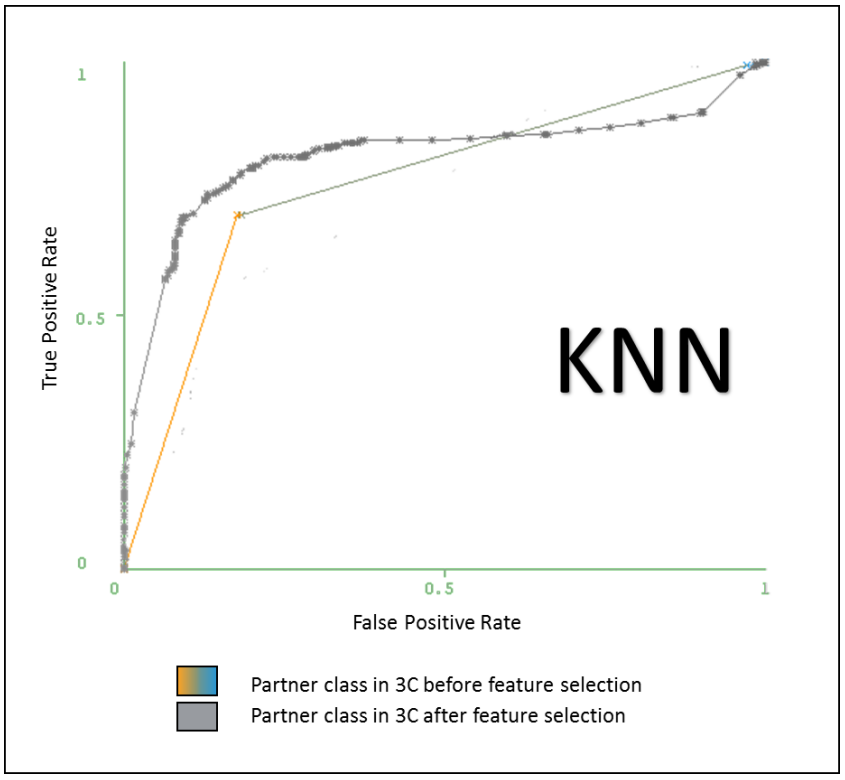
I am also including a reminder (below) of your other on-going research ethics responsibilities with respect to this research.

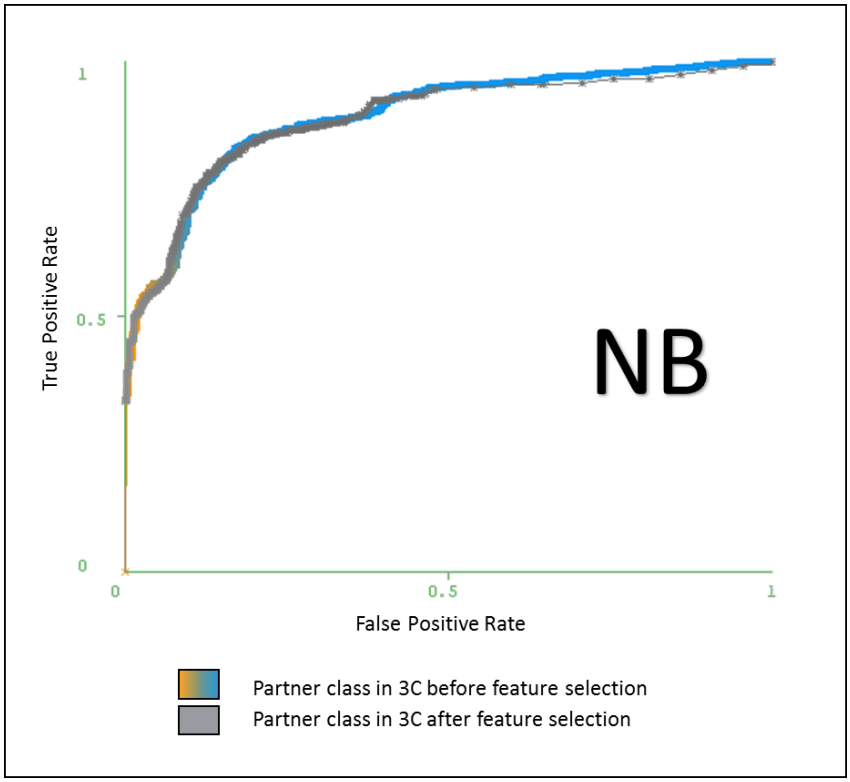
Sincerely,

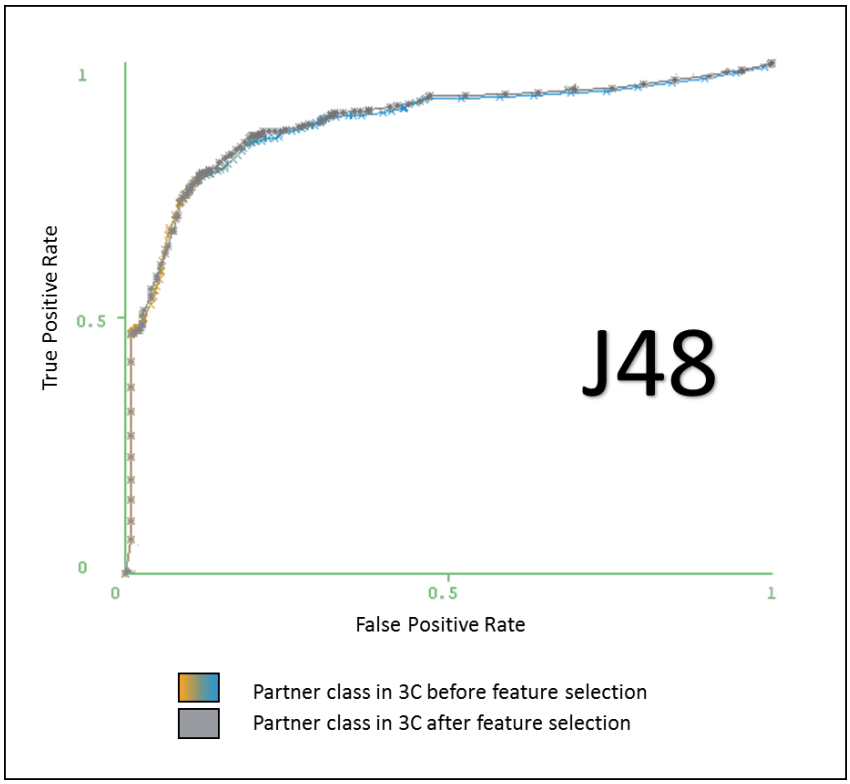
Dr. Valerie Trifts, Chair

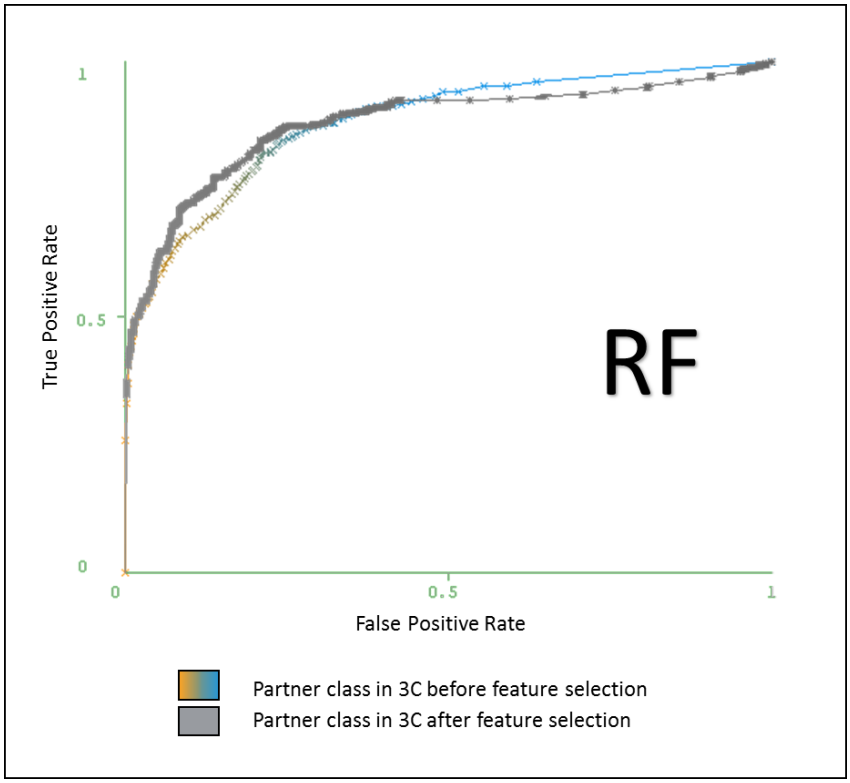
Appendix B: Feature Selection Improvement by AUROC for Each Classifier

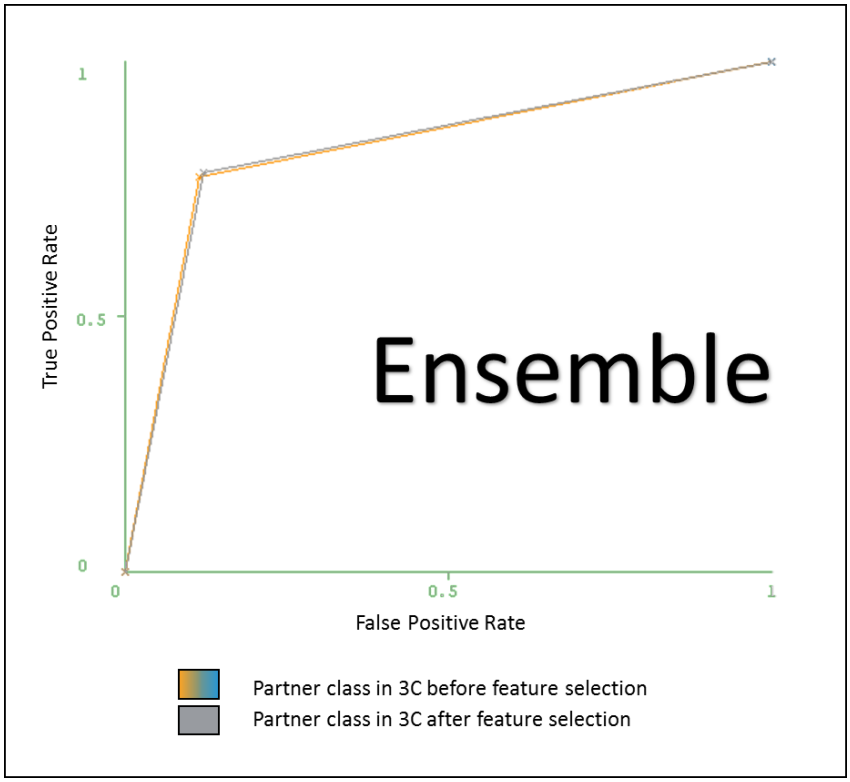


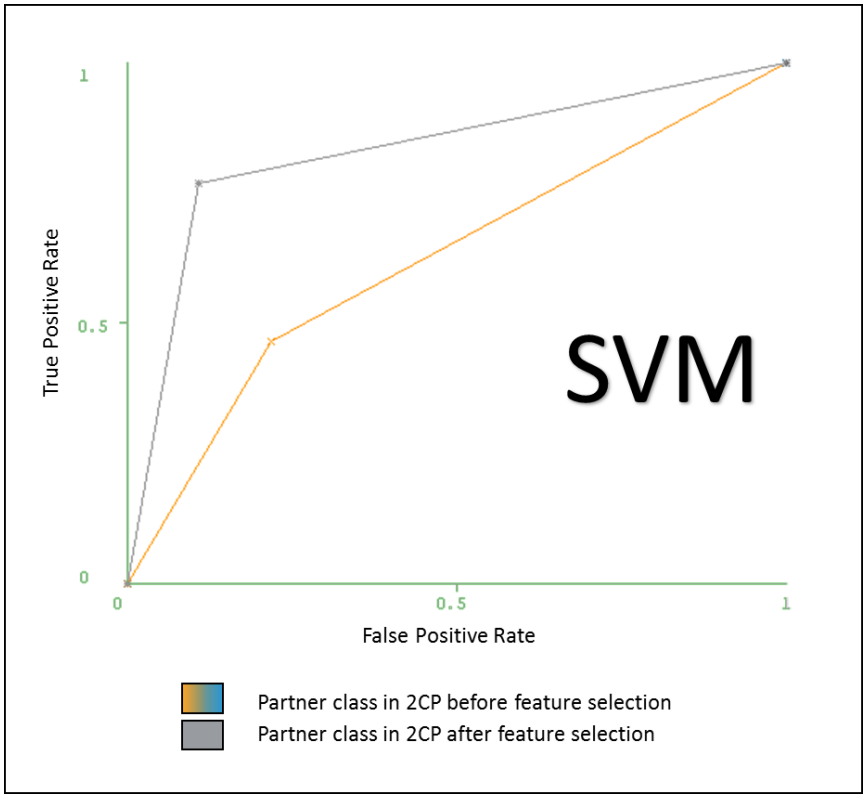


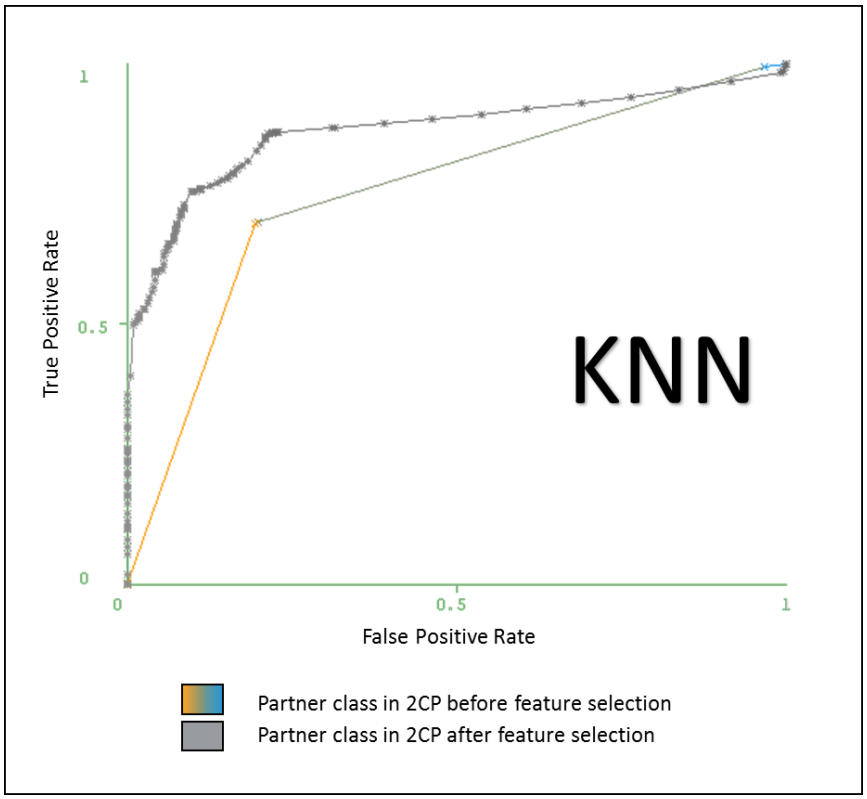

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
NB
Partner class in 3C before feature selection
Partner class in 3C after feature selection

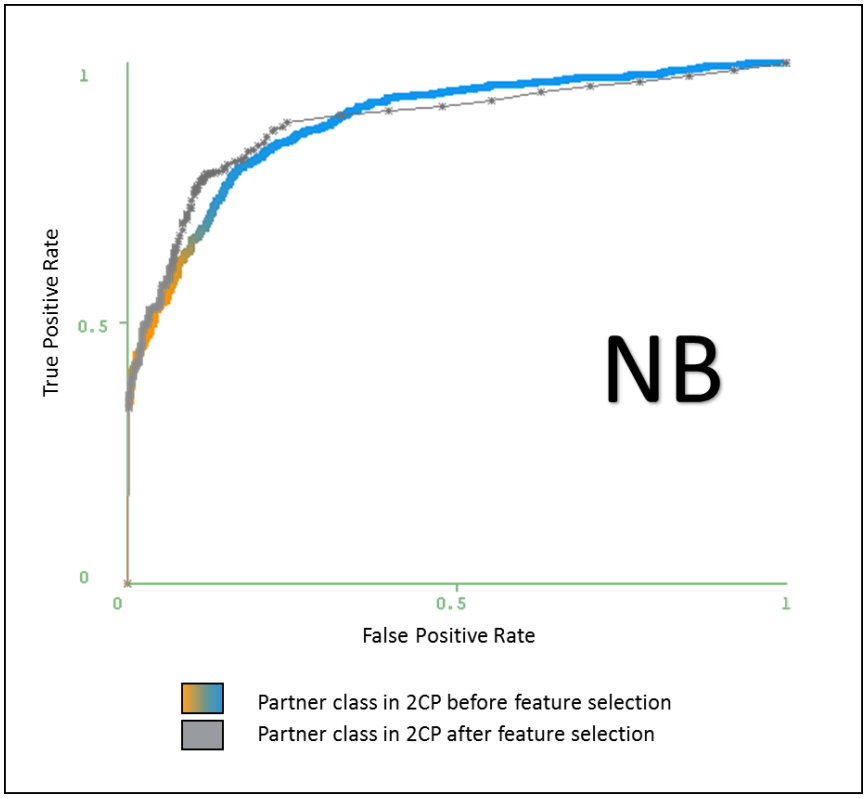

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
J48
Partner class in 3C before feature selection
Partner class in 3C after feature selection

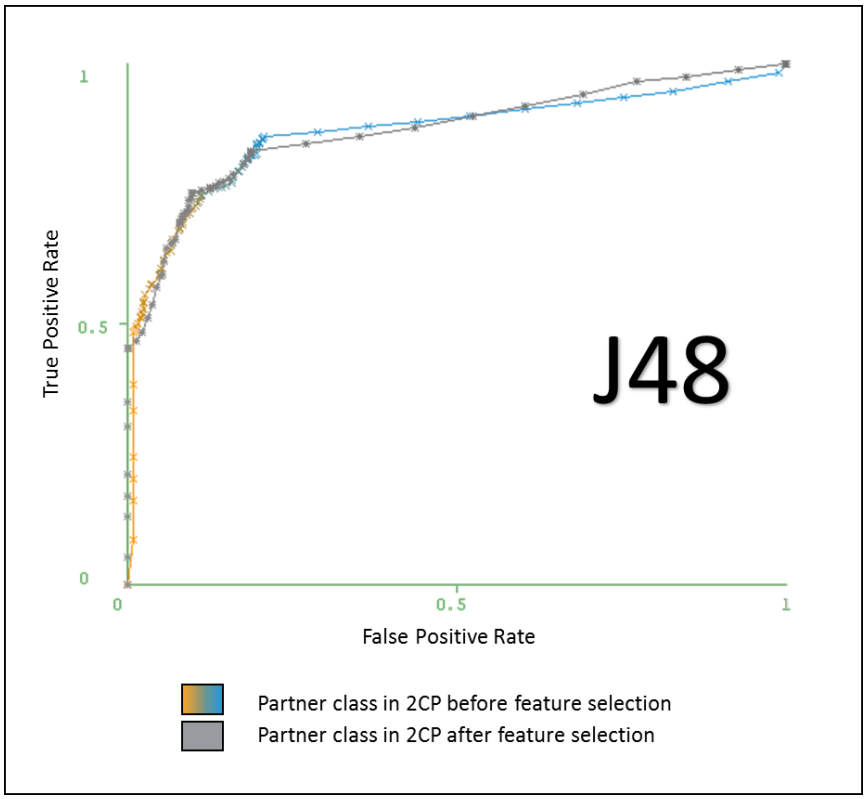

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
RF
Partner class in 3C before feature selection
Partner class in 3C after feature selection

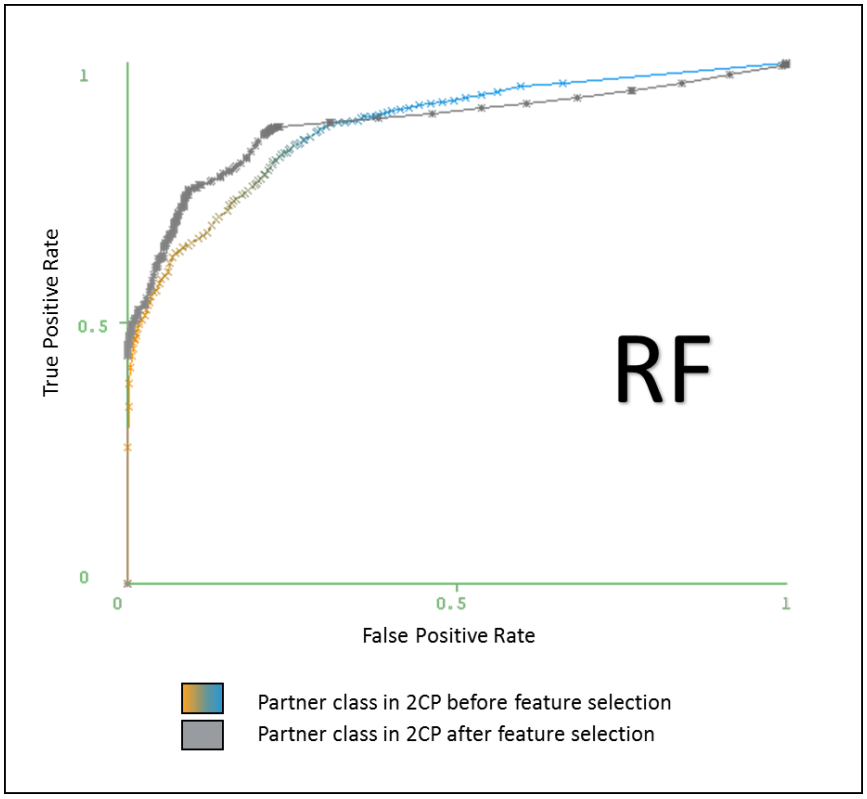

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
Ensemble
Partner class in 3C before feature selection
Partner class in 3C after feature selection

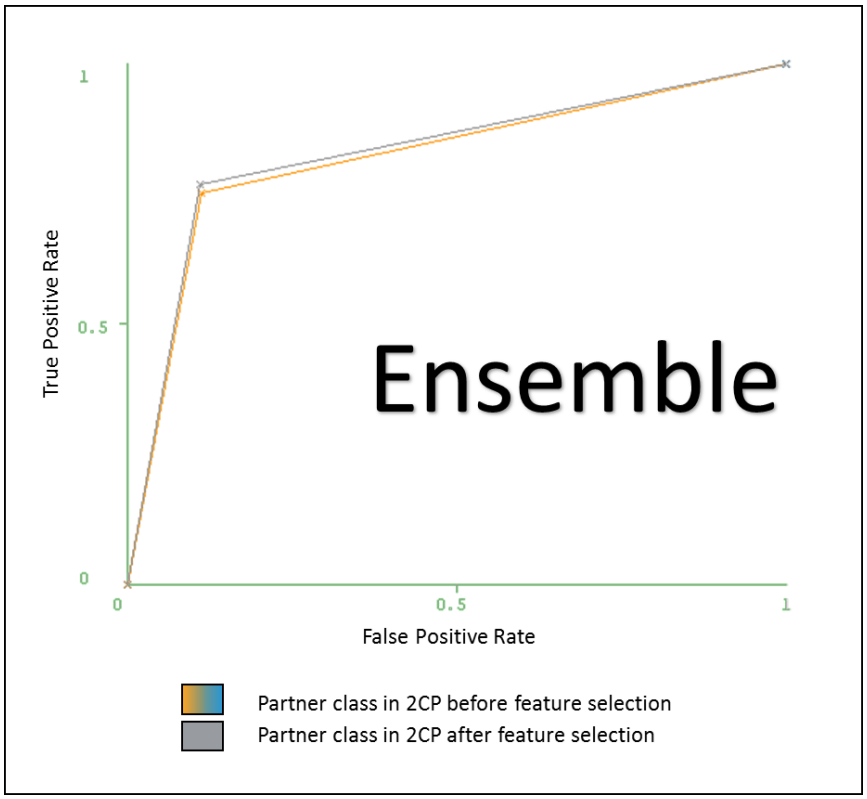

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
SVM
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

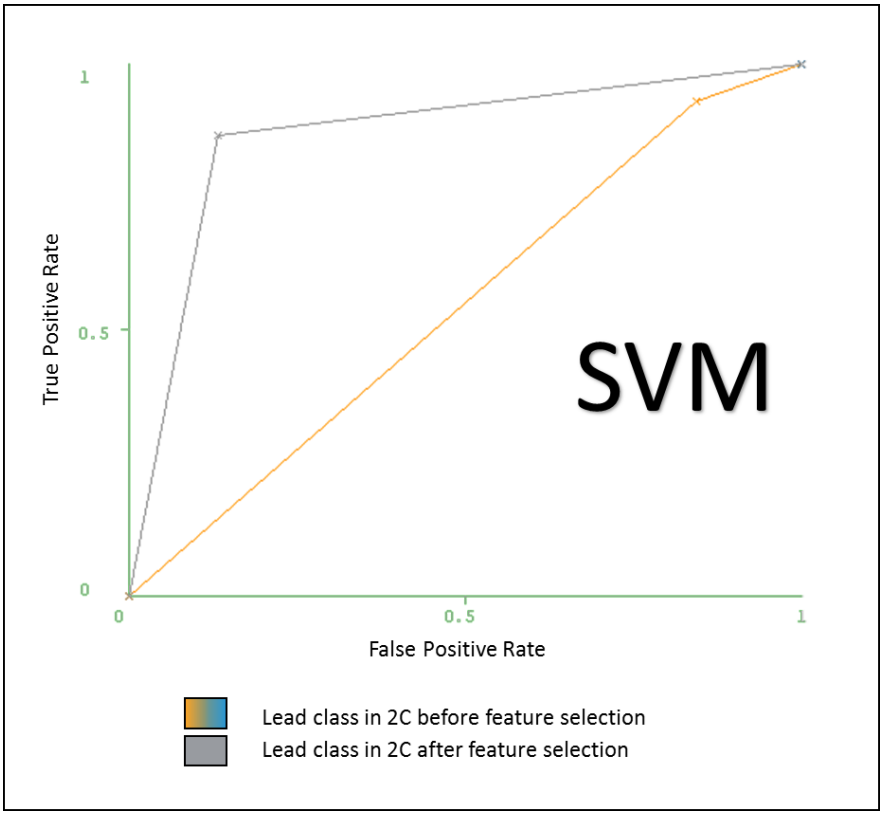

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
KNN
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

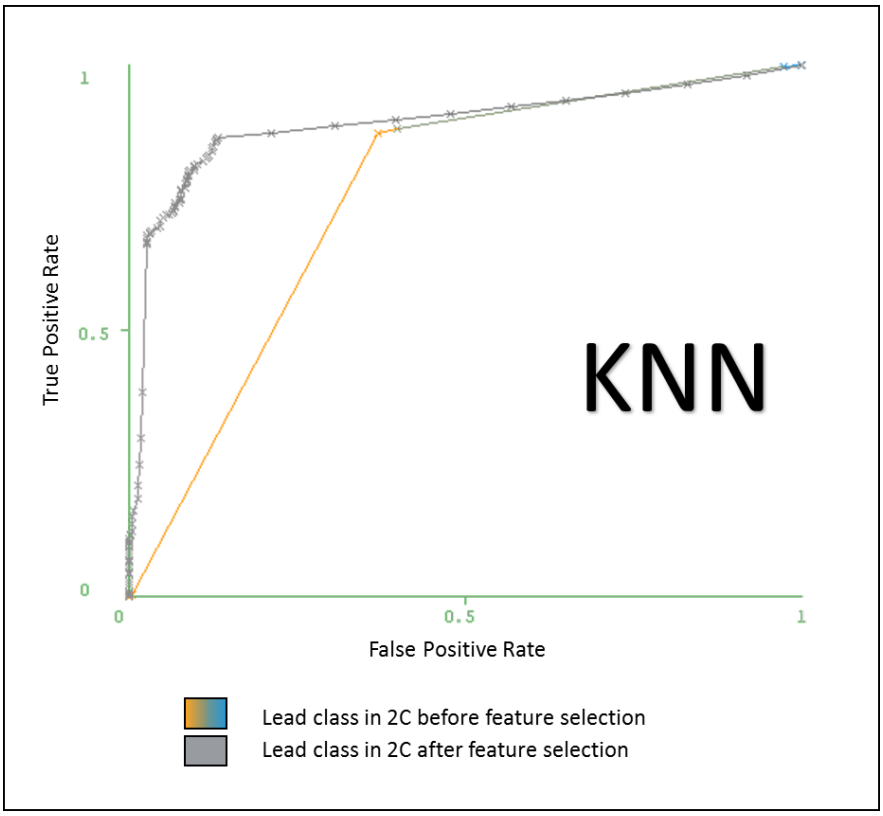

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
NB
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

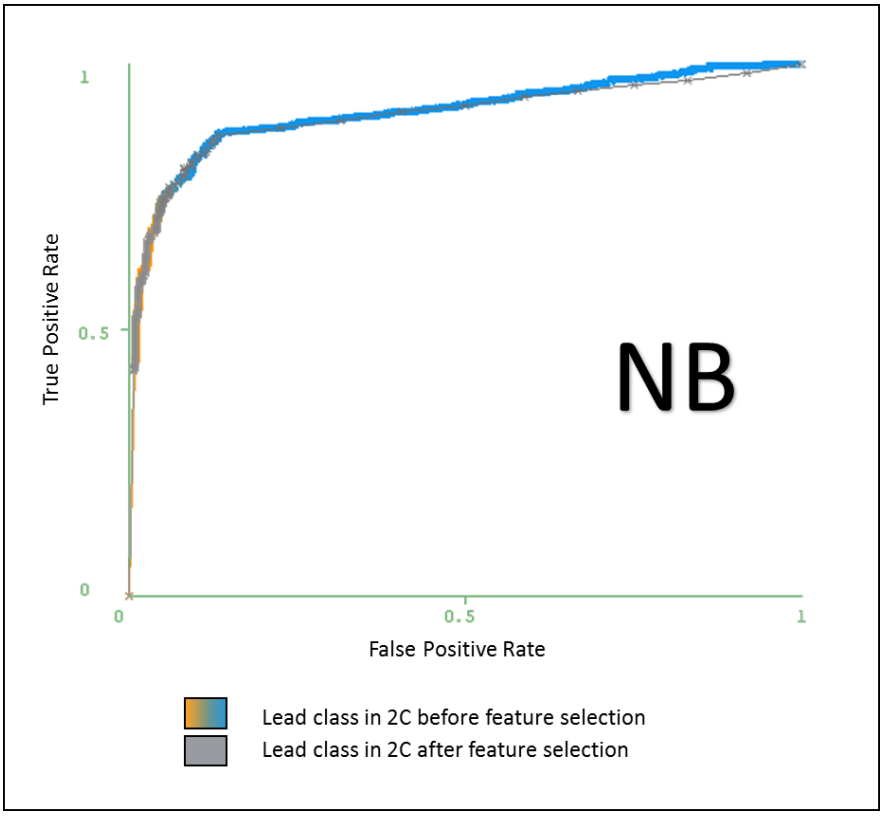

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
J48
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

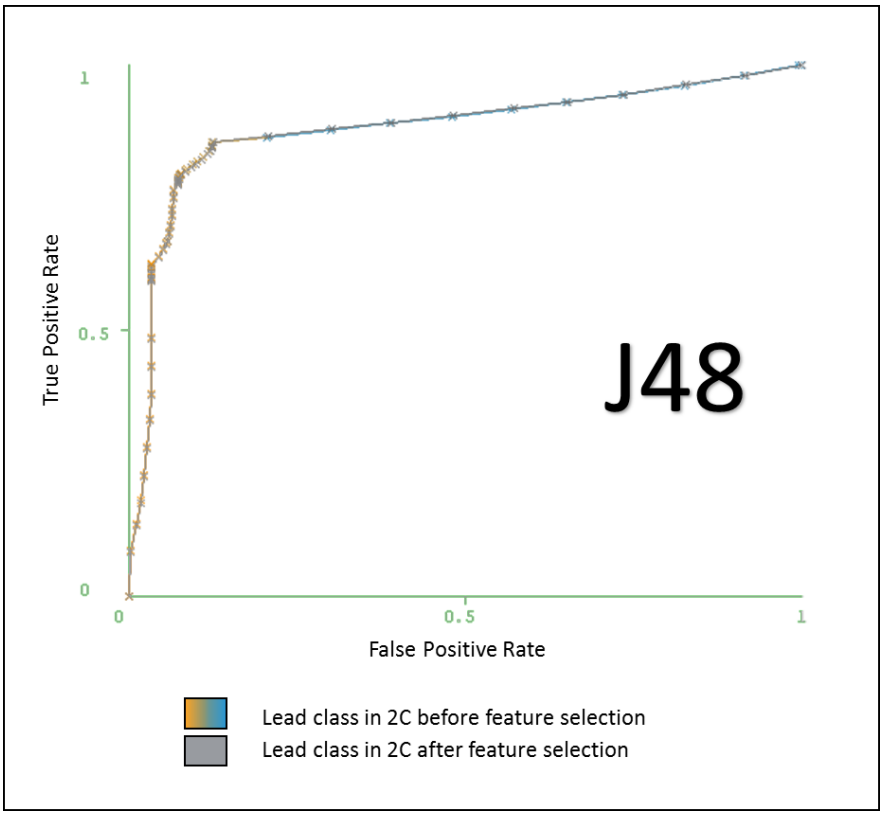

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
RF
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

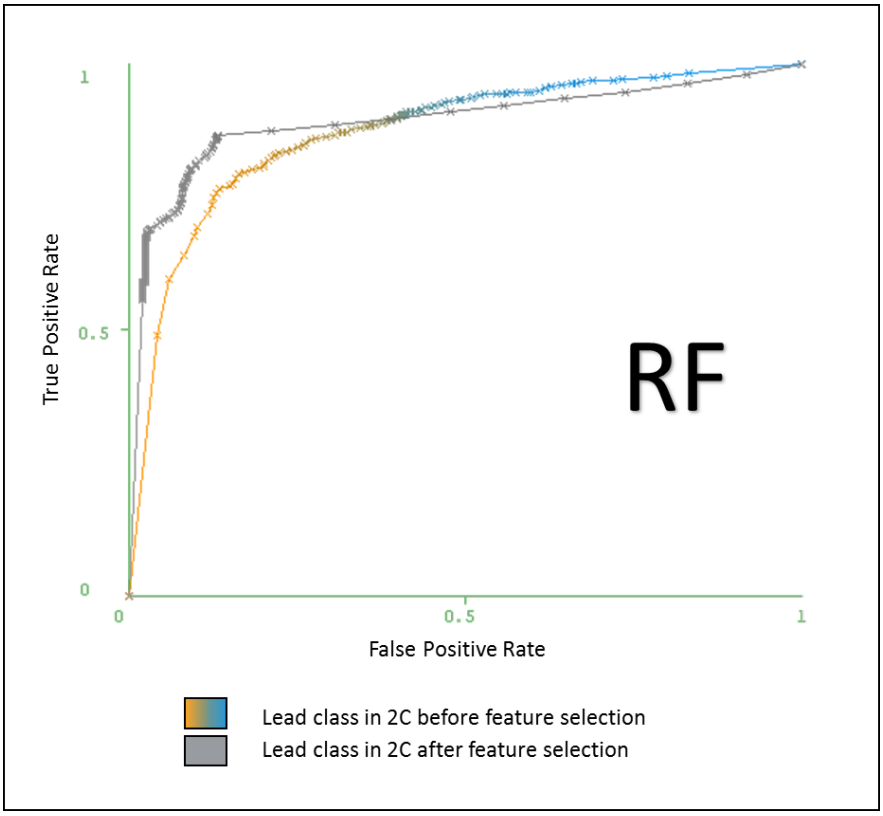

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
Ensemble
Partner class in 2CP before feature selection
Partner class in 2CP after feature selection

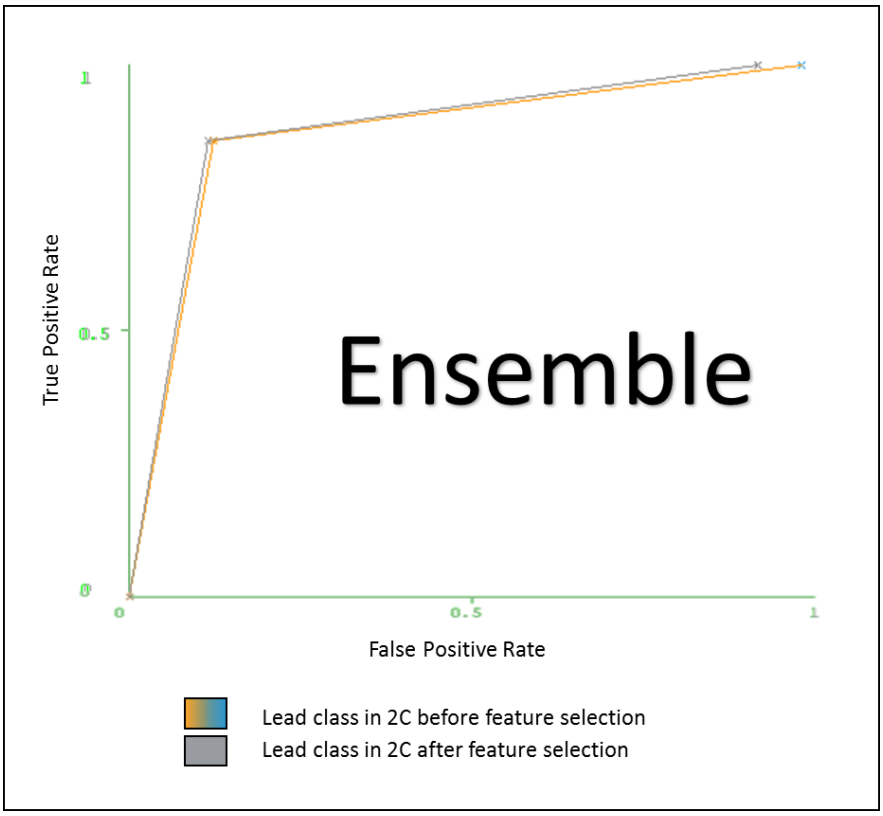

1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
SVM
Lead class in 2C before feature selection
Lead class in 2C after feature selection


1
0.5
0
True Positive Rate
0
0.5
1
False Positive Rate
KNN
Lead class in 2C before feature selection
Lead class in 2C after feature selection


True Positive Rate
1
0.5
0
0
0.5
1
False Positive Rate
NB
Lead class in 2C before feature selection
Lead class in 2C after feature selection


True Positive Rate
1
0.5
0
0
0.5
1
False Positive Rate
J48
Lead class in 2C before feature selection
Lead class in 2C after feature selection


True Positive Rate
1
0.5
0
0
0.5
1
False Positive Rate
RF
Lead class in 2C before feature selection
Lead class in 2C after feature selection


True Positive Rate
1
0.5
0
0
0.5
1
False Positive Rate
Ensemble
Lead class in 2C before feature selection
Lead class in 2C after feature selection