CUSTOMER CHURN ANALYSIS

A PROJECT REPORT

Submitted by

DIYA NARESH RAO [Reg No: 1081310001]
CHIRAG JANI [Reg No: 1081310038]
SHIVAKUMAR CHANDRASHEKAR [Reg No: 1081310013]

Under the guidance of

Ms. Nimala K
(Assistant Professor, Department of Information Technology)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM
UNIVERSITY
(Under section 3 of UGC Act 1956)

S.R.M. Nagar, Kattankulathur, Kancheepuram District

APRIL 2017

SRM UNIVERSITY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled "**CUSTOMER CHURN ANALYSIS**" is the bonafide work of " **DIYA NARESH RAO [Reg No: 1081310001], CHIRAG JANI [Reg No: 1081310038], SHIVAKUMAR CHANDRASHEKAR [Reg No: 1081310013]** ", who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Ms. Nimala K **GUIDE** Assistant Professor Dept. of Information Technology	Dr. G Vadivu **HEAD OF THE DEPARTMENT** Dept. of Information Technology
Signature of the Internal Examiner	Signature of the External Examiner

ABSTRACT

The Banking industry is rapidly developing, with banks looking to provide the best services to their customers. The trust that customers have in their bank is found to be a strong determinant of advocacy. Customer churn refers to when a consumer ceases his or her relationship with a company. The full cost of customer churn includes both lost revenue and the marketing costs involved with replacing those customers with new ones. Reducing customer churn is a key business goal of every online business. The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered. A major cause of Customer churn is poor customer service. Our project analyses a data set of consumer complaints belonging to the banking industry.Our project proposes to analyze a series of variables such as time periods, consumer response, and bank response to acquire relevant insights using visualization methods. Gathering insights will help us understand the various factors involved in leading to customer churn. The decision tree model and the adaptive boost model are compared for error rates and efficiency. Ultimately we choose to apply the more efficient algorithm on our data. We integrate various parts of our project to form a Shiny web application using the R programming language. Analytics and banking (2017)

ACKNOWLEDGEMENTS

We would like to express my deepest gratitude to my guide, Ms. Nimala K for her valuable guidance, consistent encouragement, personal caring, timely help and providing me with an excellent atmosphere for doing research. All through the work, in spite of her busy schedule, she has extended cheerful and cordial support to us for completing this research work.

Diya Naresh Rao
Shivakumar Chandrashekar
Chirag Jani

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

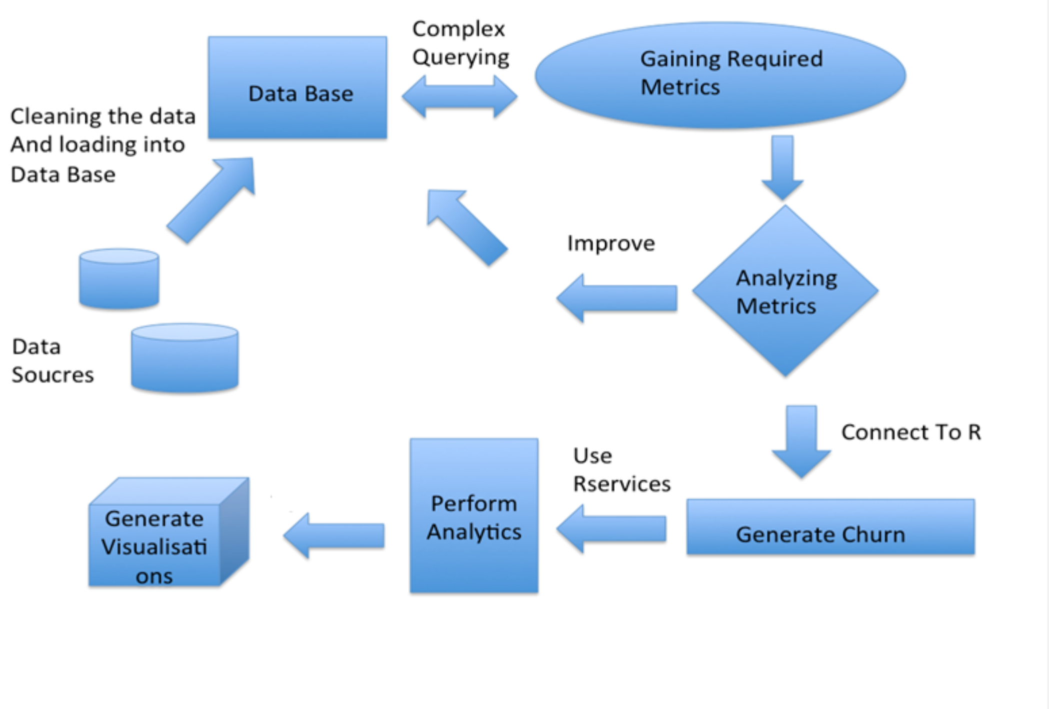


Figure 1.1: Workflow of the process

Organizations may apply analytics to business data to describe, predict, and improve business performance. Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modeling, web analytics, sales force sizing and optimization, price and promotion modeling, predictive science, credit risk analysis, and fraud analytics. Since analytics can require extensive computation (see big data), the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics. Customer churn analysis is increasingly used in multiple industries. It is of atmost importance to ensure customer satisfaction.

1.1 Customer Churn

To win and retain customers in 2017 and beyond, companies will have to assess their competencies in using data and advanced analytics to develop actionable insights. Data science and predictive analytics can help organizations synthesize data sources across multiple channels to better target the right customer with the right offer at the right time. Advanced segmentation strategies that help to identify niches based on consumer behavior will also significantly boost marketing effectiveness. Companies that deploy these techniques will accelerate past the rest of the pack in developing and deepening customer relationships.

Using techniques from data mining and text mining, predictive analytics lets executives look at historical patterns and make predictions about future behavior for specific individuals. By taking customer data that is held internally and adding what people have said and done, companies can map out what customers are likely to do next.

What is Churn Analysis?
Why we are carrying out Churn Analysis?
What are the Benefits Of Churn Analysis?

Figure 1.2: Customer churn

Customer Churn occurs when customers stop using a service or business offered by a company. Churn can occur across a number of industries from retail to banks. Customer churn is also referred to as Customer Attrition. Cultivating new customers involves using a large number of resources. Potential customers need to be cultivated and followed through till they become dependable business consumers. It is more expensive to gain

new customers than it is to retain existing customers. Thinkapps.com (2017)
Customer churn and churn rate are important metrics for measuring and controlling business growth. Businesses calculate Churn periodically.They can do so every quarter or every year. They can calculate churn in a number of ways. It depends on the type of industry, product and various other environment and market factors. Wikipedia (a)

One of the leading causes of Customer Churn is ineffective customer service and response. Consumers are of the highest priority. Therefore hearing out their problems with the product and providing effective solutions is important. There are a number of channels through which companies offer customer service such as through a call center, the web, fax, post or in person. These channels should be periodically improved. Kurt (2014)
Banks, telephone service companies, Internet service providers, pay television companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one.

Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

When companies are measuring their customer turnover, they typically make the distinction between gross attrition and net attrition. Gross attrition is the loss of existing customers and their associated recurring revenue for contracted goods or services during a particular period. Net attrition is gross attrition plus the addition or recruitment of similar customers at the original location. Financial institutions often track and measure attrition using a weighted calculation called Recurring Monthly Revenue. Earlier, there are also a number of business intelligence software programs which can mine databases of customer information and analyze the factors that are associated with customer attrition, such as dissatisfaction with service or technical support, billing disputes, or a disagreement over company policies. More sophisticated predictive analytics software use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

1.1.1 Importance of churn analysis

Customer service centers and call centers handle various customer complaints. These are subject to various cost and time metrics. Increased time spent in handling complaints leads to an increase in costs. Cost of servicing constitutes 6-7 percent of total operational costs of banks. The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source. It is more expensive to gain new customers than it is to retain existing customers. Customer churn and churn rate are important metrics for measuring and controlling business growth. Businesses calculate Churn periodically. They can do so every quarter or every year. They can calculate churn in a number of ways. It depends on the type of industry, product and various other environment and market factors. One of the leading causes of Customer Churn is ineffective customer service and response. Consumers are of the highest priority. Therefore hearing out their problems with the product and providing effective solutions is important. There are a number of channels through which companies offer customer service such as through a call center, the web, fax, and post or in person. These channels should be periodically improved.

1.1.2 Measuring customer churn over time

1. **Cohorting**: How churn changes between groups of customers segmented by the time/date that they were acquired or performed some key behavior such as sign up or first purchase; cohort analysis answers the question: How did users acquired in July compare to users acquired in June?

2. **Rolling churn**: How churn changes from one period to another for all customers that were active in the previous period regardless of the date they were acquired; rolling churn analysis answers the question: How did users active in July compare to users active in June?

3. **Fixed churn**: How churn changes over time within a fixed period for all customers that were active in the previous period regardless of the date they were acquired. Aghion (2015)

1.2 Customer lifetime value

Customer lifetime value can also be defined as the dollar value of a customer relationship, based on the present value of the projected future cash flows from the customer relationship. Customer lifetime value is an important concept in that it encourages firms to shift their focus from quarterly profits to the long-term health of their customer relationships. Customer lifetime value is an important number because it represents an upper limit on spending to acquire new customers. For this reason it is an important element in calculating payback of advertising spent in marketing mix modeling. The purpose of the customer lifetime value metric is to assess the financial value of each customer. Quantifying this is a matter of and forecasting future activity.

The present value of the future cash flows attributed to the customer during his/her entire relationship with the company. Present value is the discounted sum of future cash flows: each future cash flow is multiplied by a carefully selected number less than one, before being added together. The multiplication factor accounts for the way the value of money is discounted over time. The time-based value of money captures the intuition that everyone would prefer to get paid sooner rather than later but would prefer to pay later rather than sooner. The multiplication factors depend on the discount rate chosen (and the length of time before each cash flow occurs.

For example, money received ten years from now must be discounted more than dollars received five years in the future. Customer lifetime value applies the concept of present value to cash flows attributed to the customer relationship. Because the present value of any stream of future cash flows is designed to measure the single lump sum value today of the future stream of cash flows, it will represent the single lump sum value today of the customer relationship. Customer Lifetime value is the dollar value of the customer. Skok (2012)

1.2.1 Use of customer lifetime value

One of the major uses of Customer Lifetime Value is customer segmentation, which starts with the under- standing that not all customers are equally important. Value based segmentation model allows the company to predict the most profitable group of customers, understand those customers' common characteristics, and focus more on them rather than on less profitable customers. Value -based segmentation can be combined with a Share of Wallet customer lifetime value model to identify "high value but low wallet" customers with the assumption that the company's profit could be maximized by investing marketing resources in those customers. Han and Kamber (2011)

Customer Lifetime Value metrics are used mainly in relationship-focused businesses, especially those with customer contracts. Examples include banking and insurance services, telecommunications and most of the business-to-business sector. However, the principles may be extended to transactions-focused categories such as consumer packaged goods by incorporating stochastic purchase models of individual or aggregate behavior. In either case, retention has a decisive impact, since low retention rates result in Customer Lifetime Value.The Disadvantages do not generally stem from the modeling, but from its incorrect application. barely increasing over time.

CLV (2017) , University.custora.com (2017)

1.2.2 Advantages of life time value

1) Management of customer relationship as an asset.
2) Monitoring the impact of management strategies and marketing investments on the value of customer assets,that is Marketing Mix Modeling simulators can use a multi year lifetime value model to show the true value (versus acquisition cost) of an additional customer.
3) They can also show reduced churn rate, product up sell determination of the optimal level of investments in marketing and sales activities encourages marketers to focus on the long-term value of customers instead of investing resources in acquiring "cheap" customers with low total revenue value.
4) Implementation of sensitivity analysis in order to determinate getting impact by spending extra money on each customer. optimal allocation of limited resources for ongoing marketing activities in order to achieve a maximum return.
5) A good basis for selecting customers and for decision making regarding customer specific communication strategies.
6) A natural decision criterion to use in automation of customer relationship management systems.
7) Measurement of customer loyalty (proportion of purchase, probability of purchase and repurchase, purchase frequency and sequence etc.)

1.3 Retention Rate

Retention rate is the ratio of the number of retained customers to the number at risk". In contractual situations, it makes sense to talk about the number of customers currently under contract and the percentage retained when the contract period runs out. This term should not be confused with growth (decline) in customer counts. Retention refers only to existing customers in contractual situations. In non-contractual situations (such as catalog sales), it makes less sense to talk about the current number of customers, but instead to count the number of customers of a specified recency.

The purpose of the "retention rate" metric in a marketing atmosphere is to monitor firm performance in attracting and retaining customers. Only recently have most marketers worried about developing metrics that focus on individual customers. In order to begin to think about managing individual customer relationships, the firm must first be able to count its customers. Although consistency in counting customers is probably more important than formulating a precise definition, a definition is needed nonetheless. In particular, we think the definition of and the counting of customers will be different in contractual versus non-contractual situations. In the workplace arena, the purpose of the retention rate is to assist organizations with deciding when to take action in order to keep employees happy and motivated. Wikipedia (b)

1.4 Churn Rate

Churn rate, when applied to a customer base, refers to the proportion of contractual customers or subscribers who leave a supplier during a given time period. It is a possible indicator of customer dissatisfaction, cheaper and/or better offers from the competition, more successful sales and/or marketing by the competition, or reasons having to do with the customer life cycle. Churn is closely related to the concept of average customer life time. For example, an annual churn rate of 25 percent implies an average customer life of four years. An annual churn rate of 33 percent implies an average customer life of three years.

The churn rate can be minimized by creating barriers which discourage customers to change suppliers (contractual binding periods, use of proprietary technology, value-added services, unique business models, etc.), or through retention activities such as loyalty programs. It is possible to overstate the churn rate, as when a consumer drops the service but then restarts it within the same year. Thus, a clear distinction needs to be made between "gross churn", the total number of absolute disconnections, and "net churn", the overall loss of subscribers or members. The difference between the two measures is the number of new subscribers or members that have joined during the same period. Suppliers may find that if they offer a loss-leader "introductory special", it can lead to a higher churn rate and subscriber abuse, as some subscribers will sign on, let the service lapse, then sign on again to take continuous advantage of current specials.

When talking about subscribers or customers, sometimes the expression "survival rate" is used to mean 1 minus the churn rate. For example, for a group of subscribers, an annual churn rate of 25 percent is the same as an annual survival rate of 75 percent. Both imply a customer lifetime of four years. That is, a customer lifetime can be calculated as the inverse of that customer's predicted churn rate. For a group or segment of customers, their customer life (or tenure) is the inverse of their aggregate churn rate. Gompertz distribution models of distribution of customer life times can therefore also predict a distribution of churn rates.

For companies with a fast-growing customer base, confusion can arise between the statistical analyses associated with what percentage of the whole customer base churns in a given year versus a particular customer cohort's churn rate. Examining churn for a fast-growing aggregated customer base will understate the true churn rate compared to cohort based approach to the calculation. The cohort based approach will also allow you to calculate the survival rate and the average customer life, whereas the aggregate approach cannot calculate these two metrics.

The phrase "rotational churn" is used to describe the phenomenon where a customer churns and immediately rejoins. This is common in prepaid mobile phone services, where existing customers may take up a new subscription from their current provider in order to avail of special offers only available to new customers.

1.5 Analytics in the industry

Business analytics has the capability to enable business owners, strategic marketing professionals and even business managers to analyze and simply understand business opportunities. Another thing, analysis is used for positioning of products as well into the market. In fact, the importance of data analytics cannot be compared to some business tools out there. Data analytics belong to the business intelligence family and the only one that assists a business convert heaps of gathered raw data onto useful business info that can drive business decisions. It is frequently observed that the organizations which pertains data analytics surpass their counterparts. The business owners have to be positive. Additionally, they must focus on their value chain.

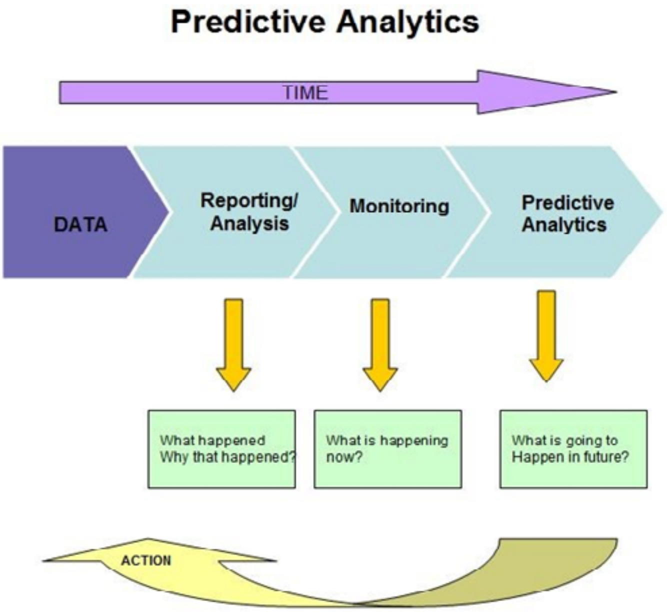


Figure 1.3: Predictive Analytics

1.5.1 Analytics in finance and banking

Customers are expecting a more personalized service from their banks. Regulators have reacted to the credit crunch with significant changes to regulation with more intrusive and granular supervision. A challenge for the industry is, therefore, how to use the breadth and depth of data available to satisfy more demanding regulators, but also improve services for customers.

The opportunity for the sector is to unlock the potential in the data through analytics and shape the strategy for business through reliable factual insight rather than intuition. Recognizing that data is a significant corporate asset, a number of organizations are appointing chief data officers following pressure to police the integrity of the numbers.

This pressure is driven by business leaders wanting more consistency in information and by regulators becoming increasingly concerned at the quality of data they receive during a time when regulatory demands are growing. This is made clear by the increasing number of references to integrity of data in each new regulatory requirement. These processes and procedures could (and usually do) involve technology, but should also include data policies, standards, roles and responsibilities to ensure data integrity is appropriately governed.

While it is crucial to ensure the integrity of data provided to executive management and regulators, unlocking the insights in the data to better understand customers, competitors and employees represents a significant opportunity to gain competitive advantage. Many financial institutions are seeing improved data quality and the use of analytics as an opportunity to fundamentally change the way decisions are made and to use the data for commercial gain.

The real strategic value in the data is the insight it can give into what will happen in the future. Predicting how customers and competitors' customers will behave and how that behavior will change is critical to tailoring and pricing products. Big data should be about changing the way you do business to harnesses the real value in your data, reshape the interaction with the market and increase the lifetime value of your customers. Therefore, which data is required to achieve these objectives, which needs it and how often are key pieces of the big data puzzle.

Big data should also involve using multiple data sources, internally and externally. Geospatial data, social media, voice, video and other unstructured data all have their part to play in knowing the customer today and their future behaviour. For example, leading firms are looking at using both internal and external data, both structured and unstructured, to develop personalized banking products. Customers are more likely to be attracted and retained with personalized products - hence, lifetime value goes up. Similarly, analytics have an increasingly important part to play in the recovery of bad debt. Recoveries functions typically target based on the delinquency status of the account. However, a better understanding of customer circumstances can improve targeting and have an immediate impact on recovery rates while also reducing cost.

Harnessing the power of analytics can enhance organizational performance. An organization derives genuine insight from their data and changes the way they interact with customers, competitors and the market through fact-driven decision-making. Those organizations that master this will set the trend in customer service, improve profitability and respond more rapidly to the evolving regulatory and competitive demands of the industry.

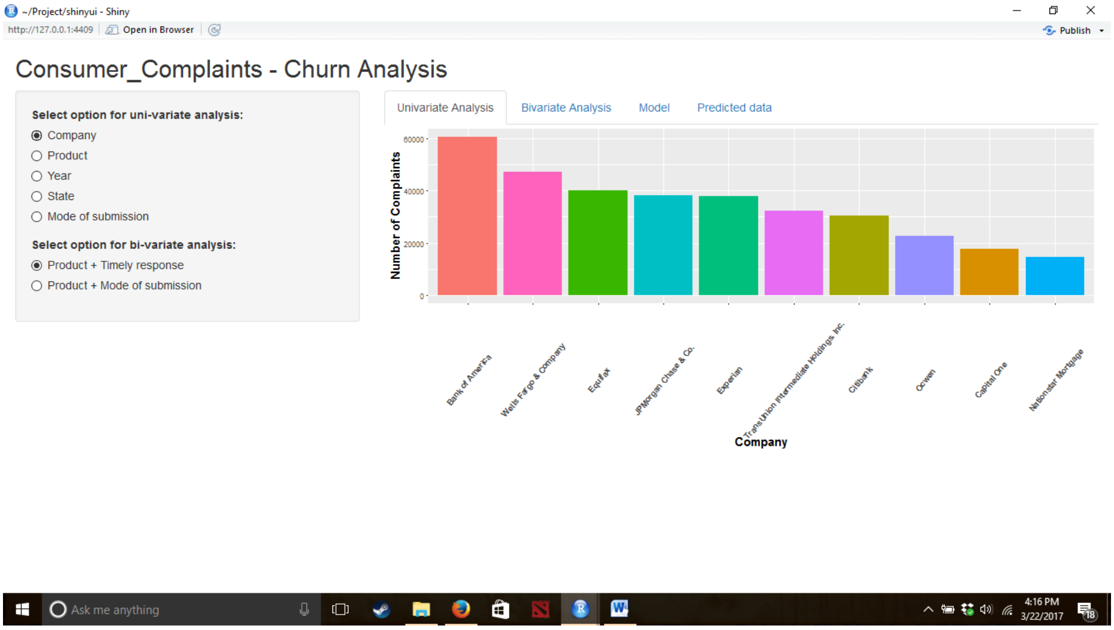
1.6 R Programming language

R is a powerful language used widely for data analysis and statistical computing. It was developed in early 90s. Since then, endless efforts have been made to improve the user interface. The journey of R language from a rudimentary text editor to interactive R Studio and more recently Jupyter Notebooks has engaged many data science communities across the world. This was possible only because of generous contributions by R users the style of coding is easy. R is the most comprehensive statistical analysis package available. It incorporates all of the standard statistical tests, models, and analyses, as well as providing a comprehensive language for managing and manipulating data. New technology and ideas often appear first in R.

The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages. It code enhancements, and new packages, and the wealth of quality packages available for R is a testament to this approach to software development and sharing. R has over 4800 packages available from multiple repositories specializing in topics like econometrics, data mining, spatial analysis, and bio-informatics. R is cross-platform. R runs on many operating systems and different hardware. R plays well with many other tools, importing data. Analyticstraining.com (2017)

1.6.1 Shiny package

With the increasing speed and popularity of the Internet, web applications are the most suited .The Shiny package in R helps to write powerful interactive web applications. It can execute R code on the backend so the app can perform any R calculation that is run on the desktop. The user types in the web address of the application in his browser. This calls upon the server, that send the relevant web page data back to the client. The client can interact with the User Interface, also designed using Shiny. The application runs within the browser. Wikipedia (2017c)



Consumer_Complaints - Churn Analysis

Select option for uni-variate analysis:

● Company
○ Product
○ Year
○ State
○ Mode of submission

Select option for bi-variate analysis:

● Product + Timely response
○ Product + Mode of submission

Figure 1.4: User Interface of Shiny Application

1.6.2 Adaptive boost model

The Ada package can be used to implement the AdaBoost algorithm. AdaBoost is a "boosting" machine learning algorithm, used to enhance and improve performance. It is best used with weak learners to provide a suitable output. A weak learner can vary, for example it can be a stump in a decision tree. In the training dataset, each element is weighted. The model is used on the training dataset and then evaluated on the test set.

1.6.3 Package for Visualization

Ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

1.6.4 RSQlite

R supports the use of SQL to retrieve data from centrally located relational databases. However, several packages in R allow you to go beyond this realm and create and query ad-hoc datasets on the fly in the midst of processing and analyzing data, regardless of the source or final destination of the data. It is extremely easy to use, and can be of great value to developers who need a database available but want to avoid the overhead often associated with installing and configuring an external database. It is available as a package in R.

1.7 Our approach to predict churn

In our project, based on the above definations we have proposed to predict customer churn in the banking industry. We have a dataset of approximately 600000 records of consumer complaints. We first clean the data and remove missing values. Next we perform an exploratory data analysis, and provide various visualisations and insghts. We then run various prediction algorithms on the training and test sets of data. We finally choose the algorithm of highest efficiency. We obtain the customers who are likely to churn. We integrate our entire project using the Shiny Package to form a web application. This provides an interactive user interface. Aghion (2015)

1.8 Pros and cons of predictive analytics

When it comes to technology management, planning, and decision making, extracting information from existing data setsâĂŤor, predictive analysisâĂŤcan be an essential business tool. Predictive models are used to examine existing data and trends to better understand customers and products while also identifying potential future opportunities and risks.
These business intelligence models create forecasts by integrating data mining, machine learning, statistical modeling, and other data technology.

1.8.1 Benefits of predictive analytics

In its multiple forms predictive modeling, decision analysis and optimization, transaction profiling, and predictive search predictive analytics can be applied to a range of business strategies and has been a key player in search advertising and recommendation engines. These techniques can provide managers and executives with decision-making tools to influence upselling, sales and revenue forecasting, manufacturing optimization, and even new product development. Though useful and beneficial, predictive analytics isn't for everyone.

1.8.2 Drawbacks of predictive analytics

A company that wishes to utilize data-driven decision-making needs to have access to substantial relevant data from a range of activities, and sometimes big data sets are hard to come by. Even if a company has sufficient data, critics argue that when anticipating human behavior computers and algorithms fail to consider variables from changing weather to moods to relationships that might influence customer purchasing patterns.
Time also plays a role in how well these techniques work. Though a model may be successful at one point in time, customer behavior changes with time and therefore a model must be updated. The financial crisis in 2008-2009 exemplifies how crucial time consideration is because invalid models were predicting the likelihood of mortgage customers repaying loans without considering the possibility that housing prices might drop.

A thorough understanding of predictive analytics can help you with business forecasting, deciding when and when not to implement predictive methods into a technology management plan, and managing data scientists.

CHAPTER 2

LITERATURE SURVEY

1. The paper "Customer Churn Analysis in Telecom Industry " by Kiran Dahiya and Surbhi Bhatiya talks about the customer churn analysis in depth. With the help of this paper we the terms involving churn prediction modeling. Also, we were able to understand the analysis and development of various data mining techniques.
There was also various techniques involved in churn prediction. Through this paper, we got insight about churn analysis in the telecom industry. It says that this industry suffers from high churn rates which can be avoided by developing and enhancing the existing methods. Various prediction models were also discussed like regression analysis, decision trees. Gursoy (2015)

2. The paper "Churn Prediction -A Review " by Navneet Kaur and Naseeb Singh is about churn prediction and various data mining techniques required in churn prediction. It also implies on how churn prediction plays a major role in the development of an organization. This paper is an enhancement of boosted tree for churn prediction with more accuracy.
It presented an approach that is an algorithm based on boosted tree. The proposed methodology shows how a dataset can be classified into loyal and churn customers. It states that companies should put more focus on churn analysis in order to retain their customers and sustain their profits for a better future.

3. The paper "Customer churn analysis - a case study" by Teemu Mutanen gives a detailed explanation on customer churn in the banking sectore. It also gives information on how the methods that are used to calculate churn. This study compares the two methods used for churn i.e. lift curves and logistic regression based on the case study discussed in it.This paper also gives an indication in the case of logistic regression model, that the user should update the model to be able to make predictions with high accuracy.

CHAPTER 3

SYSTEM ANALYSIS

3.1 Installing R studio

RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Windows, macOS, and Linux. RStudio is available in open source and commercial editions and runs on the desktop (Windows, macOS, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro.

Follow the steps below for installing R Studio:
1) Go to https://www.rstudio.com/products/rstudio/download/
2)In 'Installers for Supported Platforms' section choose and click the RStudio installer based on your operating system. The download should begin as soon as you click.
3) Click Next..Next..Finish.
4) Download Complete.
5)To Start RStudio,click on its desktop icon to access the program.

3.2 The Data Set

Our dataset consists of 18 variables and 650313 records. There are various variables describing complaint attributes such as Product, Sub Product, Issue, Sub Issue and Location. There are also various binary variables such as "Consumer disputed and Timely

Response". The primary key of the dataset in the Complaint ID. The datset has been obtained from https://catalog.data.gov/dataset/consumer-complaint-database.

3.3 Factors affecting Churn

3.3.1 Corporate Reputation

The level of scrutiny of the ethical standards and corporate social responsibility initiatives undertaken by companies has never been so searching. The attention of pressure groups and the media given to company behavior is unprecedented, continually escalating and frequently hostile. Damage to corporate reputation of a business-because of ethics violations or failure to provide societal benefits-can substantially reduce its ability to compete and can undermine the value of a company. Importantly, the strength or weakness of an organization's corporate reputation significantly impacts customer perceptions of how attractive it is to do business with that company. In addition, reputation influences investors, the capital markets, and the ability to recruit talented people, and the influential commentary of investment analysts and the media.

In several ways, corporate reputation is about the ability to compete. A poor corporate reputation-regarding handling of suppliers and customers, honesty and fairness in making deals, behavior toward the environment, working standards for employees in the value chain, and similar judgments can make a company unattractive to customers. Customers may reject a potential supplier because they do not want to be contaminated by association or to face the criticisms of their own customers and investors. Conversely, a strong corporate reputation can make a company more attractive than competitors to some customers because they benefit by association,corporate reputation adds to the value of the customer relationship by providing the customer with assurance as to a supplier's good standing and that it is safe to deal with that company without risking their own reputation.

Ethical standards and corporate responsibility initiatives are an important foundation for an attractive corporate reputation, which impacts positively on relationships with consumers and business-to-business customers. Wang (2015) Ethical Imperatives Ethical behavior is concerned with good conduct. In a business context, ethics refer to the rules or standards guiding the behavior of the members of a profession or organization. It is important to understand ethics and its relevance in an organization. Moreover, management needs to be proactive in building an ethical environment and employee commitment to ethical behavior. Importantly, ethical practices help to create and sustain trust with consumers and value chain members. There have been extensive public concerns about poor ethical practices in business organizations, fueled by highly publicized scandals such as those at Enron, Siemens, and others.

3.3.2 Defining Ethical Standards

Business ethics are concerned with ethical rules and principles, moral or ethical problems, and special duties and obligations that apply to persons engaged in commerce. Marketing ethics focus on those ethical situations that fall within the scope of marketing operations in an organization. Marketing's involvement outside the organization with customers and value chain members often exposes it to situations where ethical and moral issues are particularly evident. For example, salespeople have been frequent targets for criticism regarding ethical Standards-they are exposed to more ethical pressures than people in other jobs; they work in relatively unsupervised settings; they typically face demanding sales revenue targets; and many are largely "paid by results". A survey of sales managers by Sales and Marketing Management revealed that 49 percent of managers said their sales people had lied on a sales call, 34 percent said their salespeople made unrealistic.

3.4 Graphical representation of data

Graphic representation is another way of analysing numerical data. A graph is a sort of chart through which statistical data are represented in the form of lines or curves drawn across the coordinated points plotted on its surface. They enable us in studying the cause and effect relationship between two variables. Graphs help to measure the extent of change in one variable when another variable changes by a certain amount.Graphs also enable us in studying both time series and frequency distribution as they give clear account and precise picture of problem. Graphs are also easy to understand and eye catching. (Plot.ly)

Graphical representation of data enjoys various advantages which are as follows:
1. **Acceptability**: Such report is acceptable to the busy persons because it easily highlights about the theme of the report. This helps to avoid wastage of time.
2. **Less cost**: Information if descriptive involves huge time to present properly. It involves more money to print the information but graphical presentation can be made in short but catchy view to make the report understandable. It obviously involves less cost.
3. **Decision Making**: Business executives can view the graphs at a glance and can make decision very quickly which is hardly possible through descriptive report.
4. **Logical Ideas**: If tables, design and graphs are used to represent information then a logical sequence is created to clear the idea of the audience.
5. **Helpful for less literate Audience**: Less literate or illiterate people can understand graphical representation easily because it does not involve going through line by line of any descriptive report.
6. **Less Effort and Time**: To present any table, design, image or graphs require less effort and time. Furthermore, such presentation makes quick understanding of the information.
7. **Less Error and Mistakes**: Qualitative or informative or descriptive reports involve errors or mistakes. As graphical representations are exhibited through numerical figures, tables or graphs, it usually involves less error and mistake.
8. **A complete Idea**: Such representation creates clear and complete idea in the mind of audience. Reading hundred pages may not give any scope to make decision. But an instant view or looking at a glance obviously makes an impression in the mind of audience regarding the topic or subject.
9.**Comparative Analysis**: Information can be compared interms of graphical representation.Such comparative analysis helps for quick understanding and attention.. Rasel (2013)

3.5 Common packages used in R

3.5.1 PLYR Package in R

Plyr is an R package that makes it simple to split data apart, do stu to it, and mash it back together. This is a common data-manipulation step. Importantly, plyr makes it easy to control the input and output data format from a syntactically consistent set of functions. Plyr is a set of tools that solves a common set of problems: breaking a big problem down into manageable pieces, operating on each of the pieces and then put all the pieces back together. It's already possible to do this with split and the apply functions, but plyr makes it easier.

Why use plyr over base Apply functions?
1. Plyr has a common syntax - easier to remember
2. Plyr requires less code since it takes care of the input and output format
3. Plyr can easily be run in parallel faster

3.5.2 DPLYR Package in R

When working with data one must:
* Figure out what should be done.
* Describe those tasks in the form of a computer program.
* Execute the program.

The dplyr package makes these steps fast and easy:
* By constraining options, it simplifies common data manipulation tasks.
* It provides simple verbs, functions that correspond to the most common data manipulation tasks, to help translate those thoughts into code.
* It uses efficient data storage backends, hence spending less time waiting for the computer

Dplyr aims to provide a function for each basic verb of data manipulation:

* filter() (and slice())
* arrange()
* select() (and rename())
* distinct()
* mutate() (and transmute())
* summarise()
* sample-n() (and sample-frac())

3.5.3 GGPLOT2 Package in R

Ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

There are two major functions that are used in ggplot2:

* qplot() - for quick plots
* ggplot() - for fine, granular control of everything (not going to get into this in this post)

Advantages of ggplot2:

* consistent underlying grammar of graphics
* plot specification at a high level of abstraction
* very flexible
* theme system for polishing plot appearance
* mature and complete graphics system
* many users, active mailing list

3.6 Data cleaning

Data cleaning is a valuable process that can help companies save time and increase their efficiency. Data cleaning software tools are used by various organisations to remove duplicate data, fix and amend badly-formatted, incorrect and amend incomplete data from marketing lists and databases. They can achieve in a short period of time what could take days or weeks for an administrator working manually to fix. This means that companies can save not only time but money by acquiring data cleaning tools. Data cleaning is of particular value to organisations that have vast swathes of data to deal with.These organisations can include banks or government organisations but small to medium enterprises can also find a good use for the programmes.

In fact, it's suggested by many sources that any firm that works with and hold data should invest in cleaning tools. The tools should also be used on a regular basis as inaccurate data levels can grow quickly, compromising database and decreasing business efficiency. Companies may also find that cleaning enables them to remain compliant with standards that are legally expected of them. In most territories, companies are duty-bound to ensure that their data is as accurate and current as possible. The tools can be used for everything from correcting spelling mistakes to postcodes, whilst removing unnecessary records from systems, which means that space can be preserved and that information that is no longer needed or data which companies are no longer permitted to keep can be removed simply, quickly and efficiently.

Users of data cleaning software can set their own rules to increase the efficiency of a database, making the capabilities of the cleaning software as applicable to the company's needs and requirements as possible.Some common problems with databases can also include incorrectly formatted phone numbers and e-mail addresses, rendering clients and customers unreachable. The software can be used to put things right in a matter of seconds. This makes it a perfect tool for companies that need to stay in touch with outside parties. Meanwhile, companies that employ more than one database companies that are spread across various branches or offices for example, one can use the tools to ensure that each branch of their organisation can share the same accurate information.

3.6.1 Role of missing values

Anyone who does statistical data analysis or data cleaning of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes.

For example in surveys people generally tend to leave the field of income blank or sometimes people have no information available and cannot answer the question. Also in the process of collecting data from multiple sources some data may be inadvertently lost. For all these and many other reasons,missing data is a universal problem in both social and health sciences. This is because every standard statistical method works on the fact that every problem has information on all the variables an it needs to be analyzed. The most common and simple solution to this problem is if any case has missing data for any of the attribute to be analyzed we can simply ignore it. This will give us a dataset which will not contain any missing value and we can then use any standard methods to process it further. But this method has a major drawback which is deleting missing values sometimes might lead to ignoring a large section of the original sample.

Data cleaning and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss approaches to deal with missing data in different scenarios.

3.6.2 Terms related to data cleaning

1) **Data cleaning**: Process of detecting, diagnosing, and editing faulty data.

2) **Data editing**: Changing the value of data shown to be incorrect.

3) **Data flow**: Passage of recorded information through successive information carriers.

4) **Inlier**: Data value falling within the expected range.

5) **Outlier**: Data value falling outside the expected range.

6) **Robust estimation**: Estimation of statistical parameters, using methods that are less sensitive to the effect of outliers than more conventional methods.

3.7 R and SQL

The Structured Query Language is a domain-specific language used in programming and designed for managing data held in a relational database management system. SQL consists of a data definition language, data manipulation language, and data control language. The scope of SQL includes data insert, query, update and delete, schema creation and modification, and data access control. It can be used to manipulate, store and access data.

However in performing complex calculations and using the very same data for analytics, using a programming language like R is more efficient. R provides built-in functions for data frame manipulation.Therefore the database can be loaded into a structure in R such as a table or a data frame. Once this is done various R functions and packages can be used to manipulate the data.Common packages used to connect R and SQL are RSQLite and sqldf. These packages provide an interface between SQLite memory database and R through SQL. The packages need to be called to query a data frame in the current environment.

3.8 Data Models

3.8.1 Support Vector Machine model

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, a support vector machine training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. a support vector machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.In addition to performing linear classification, a support vector machine model can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.Permutation tests based on Support vector machine weights have been suggested as a mechanism for interpretation of Support vector machine models. Support vector machine weights have also been used to interpret Support vector machine models in the past. Posthoc interpretation of support vector machine models in order to identify features used by the model to make predictions is a relatively new area of research with special significance in the biological sciences.

They can be used to solve various real world problems:
1. Support vector machines are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings.
2. Classification of images can also be performed using support vector machines.
3. Experimental results show that support vector machines achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.
4. This is also true of image segmentation systems, including those using a modified version of the model that uses the privileged approach.
5. Hand-written characters can be recognized using support vector machines.
6. The algorithm has been widely applied in the biological and other sciences.
7. They have been used to classify proteins with up to ninety percent of the compounds being classified correctly. Wikipedia (c)

3.8.2 Random Forest

The Random Forests algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. Walker (2013)

The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and regressors. Single decision trees often have high variance or high bias. Random Forests attempts to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Considering that Random Forests have few parameters to tune and can be used simply with default parameter settings, they are a simple tool to use without having a model or to produce a reasonable model fast and efficiently.

Random Forests are easy to learn and use for both professionals and lay people - with little research and programming required and may be used by folks without a strong statistical background. Simply put, you can safely make more accurate predictions without most basic mistakes common to other methods.The Random Forests algorithm was developed by Leo Breiman and Adele Cutler. Random Forests grows many classification trees. Each tree is grown as follows:
1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to

split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning. Srivastava (2014)

3.9 What is Cluster analysis?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939[1][2] and famously used by Cattell beginning in 1943 for trait theory classification in personality psychology.The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common denominator: a group of data objects. However, different researchers employ different cluster models.

Typical cluster models include:

1. **Connectivity models**: for example, hierarchical clustering builds models based on distance connectivity.

2. **Centroid models**: for example, the k-means algorithm represents each cluster by a single mean vector.

3. **Distribution models**: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.

4. **Density models**: for example, OPTICS defines clusters as connected dense regions in the data space.

Subspace models: in Biclustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.

5. **Group models**: some algorithms do not provide a refined model for their results and just provide the grouping information.

6. **Graph-based models**: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster.

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other. Clusterings can be roughly distinguished as:

1. **Hard clustering**: each object belongs to a cluster or not

2. **Soft clustering (also: fuzzy clustering)**: each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)

3. **Strict partitioning clustering**: each object belongs to exactly one cluster

4. **Strict partitioning clustering with outliers**: objects can also belong to no cluster, and are considered outliers

5.**Overlapping clustering (also: alternative clustering, multi-view clustering)**: objects may belong to more than one cluster; usually involving hard clusters

6. **Hierarchical clustering**: objects that belong to a child cluster also belong to the parent cluster

7. **Subspace clustering**: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap Wikipedia (2017a)

3.10 Shiny in R

Layout Shiny ui.R scripts use the function fluidPage to create a display that automatically adjusts to the dimensions of the user's browser window. It can be laid out on the app by placing elements in the fluidPage function. titlePanel and sidebarLayout are the two most popular elements to add to fluidPage. They create a basic Shiny app with a sidebar.

sidebarLayout always takes two arguments:
* sidebarPanel function output
* mainPanel function output

Control widgets :
Shiny comes with a family of pre-built widgets, each created with a transparently named R function. For example, Shiny provides a function named actionButton that creates an Action Button and a function named sliderInput that creates a slider bar.
The standard Shiny widgets are:

FUNCTION - WIDGET
* actionButton - Action Button
* checkboxGroupInput - A group of check boxes
* checkboxInput -A single check box
* dateInput -A calendar to aid date selection
* dateRangeInput -A pair of calendars for selecting a date range
* fileInput -A file upload control wizard
* helpText -Help text that can be added to an input form
* numericInput -A field to enter numbers
* radioButtons -A set of radio buttons

* selectInput -A box with choices to select from

* sliderInput -A slider bar

* submitButton -A submit button

* textInput -A field to enter text

Display reactive output:

Placing a function in ui.R tells Shiny where to display the object. Next, the user needs to tell Shiny how to build the object.

This is done by providing R code that builds the object in server.R. The code should go in the unnamed function that appears inside shinyServer in the server.R script.The unnamed function plays a special role in the Shiny process; it builds a list-like object named output that contains all of the code needed to update the R objects in the app. Each R object needs to have its own entry in the list.

An entry can be created by defining a new element for output within the unnamed function, like below. The element name should match the name of the reactive element that has been created in ui.R. Render function creates :-

* renderImage :images (saved as a link to a source file)

* renderPlot :plots

* renderPrint :any printed output

* renderTable :data frame, matrix, other table like structures

* renderText :character strings

* renderUI :a Shiny tag object or HTML

Each render function takes a single argument: an R expression surrounded by curly braces. The expression can be one simple line of text, or it can involve many lines of code, as if it were a complicated function call.Launch the app to see the reactive output.The next step is to update server.R and ui.R files to match those above. Then the final step is launching the Shiny app.When Shiny rebuilds an output, it highlights the code server.R script is running. This temporary highlighting showcases how Shiny generates reactive output. (shiny.rstudio.com)

CHAPTER 4

SYSTEM DESIGN

4.1 Proposed Methodology - Churn Analysis Using R

Analytics involves reading and analyzing large volumes of data for insights and predictions. For this, R is the best language. It comes with a variety of packages and functions that aid in the process of delivering the most accurate information.In order to predict customer churn we need to consider a number of factors such as issues customers have faced, untimely response and the products that they have had troubled with. R provides adequate data structures to store these variables and use them in modeling and analysis.

The first step in any analysis or modeling process is to load, store and access the data in a proper manner. 'RSQLite' is a package that can be used to connect to the sql server. On calling the package and objects declared, the required 'dbConnect' and 'dbDisconnect' can be used to connect and disconnect from the database respectively.'dbGetQuery' and 'dbSendQuery' can be used to to give required queries to create tables and extract data.The Complaint ID is the primary key.

R can be used to form a number of models. Packages such as rpart, randomForest and Ada can be s=used for the decision tree, random forest and Adaptive Boost algorithms respectively. Within these packages there are various functions for running and evaluation. Training and test datasets are loaded into dataframes. The model is then built. For example in the ada package the method ada() is used to build and run the model.

The variables used in model construction and the most frequent variables used are obtained. After receiving the response from the Ada Boost model, it is then evaluated.A confusion matrix is generated showing proportions. An error matrix is also obtained where by the overall error percentage and the averaged class error percentages are calculated .Finally the predict function is used to predict the results on the test set. This generates the prediction matrix with binary values for the complaint ID's in the test set.

In order to visualize the results obtained from the tables and created and populate the numbers the ggplot2 package in R can be used. Using ggplot2 various plots and graphs can be created for effective univariate, bivariate and time series analysis. The package ggplot2 also provides customization options for the appearance of the graphs besides plotting the x,y relationships.

Our system consists of three stages:
1) Analysis and cleaning
2) Modeling
3) Integration using Shiny Web Application.

The aim of our system is to provide an adequate and interactive user interface to display the outputs and visualizations at the front end. This will represent the back end code with the model and the graphs.

4.2 Analysis and cleaning

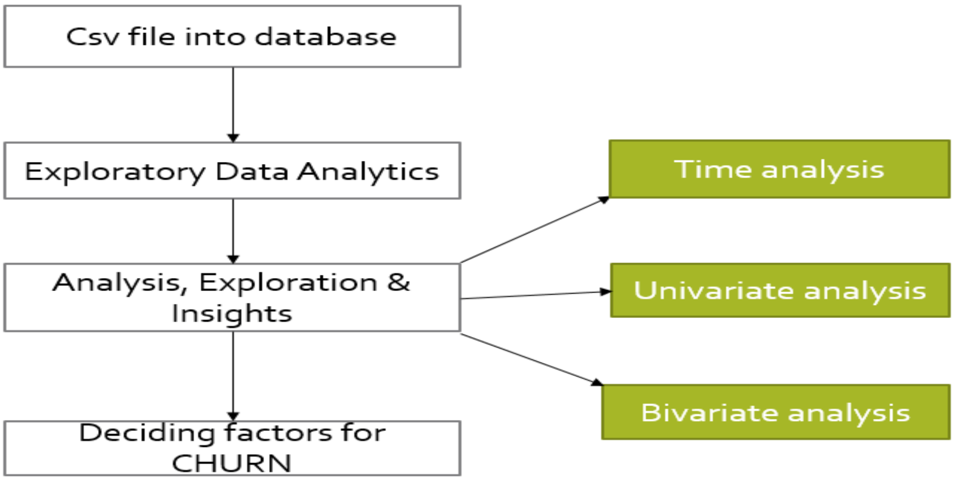

Csv file into database
Exploratory Data Analytics
Analysis, Exploration & Insights
Deciding factors for CHURN
Time analysis
Univariate analysis
Bivariate analysis

Figure 4.1: Analysis Stage

4.1 The first step to initiate our project is to clean the data. The missing values are identified and removed in order to get accurate results. After this the data is explored using various commands and functions using the RSQLite package to perform data analytics. This package helps in embedding the SQLite database engine in R. After the data is analyzed, the information derived from the analytics i.e. insights is displayed using the ggplot2 package in time analysis, univariate analysis and bivariate analysis.

4.3 Modeling

4.3.1 The Procedure

A relevant target variable is chosen for predictive modeling. Two models namely the Decision Tree model and the Adaptive Boosting Model are run on our data set and compared. The packages used in the modeling include Ada for adaptive boosting, rpart for decision tree and Rattle for various insights, model building and evaluation . We aim to use the model with the lower error percentage. In order to evaluate the decision tree and adaptive boosting models, the R data miner using rattle comes in handy. Using rattle, both the models are built and then evaluated to form error matrices and ROC curves.This mode of evaluation helps us decide the most apt model for our data. . Once we obtain the required model we run it on the training data set. We then evaluate the model on the test set to obtain desired results. We aim to predict which customers in the test set are likely to churn.

4.3.2 Choosing the target variable

We analyzed the dataset to find the target variable. We chose "Timely Response" as the target variable because it has most relevance. It is a binary variable that represents if the bank has solved the customer's problem or complaint in a time appropriate manner. It also had the least number of missing values and could be targeted properly.

There is a huge disconnect between customers and businesses when it comes to feedback. Consumers feel their feedback is not being heard. Getting honest customer feedback can be essential to businesses who are looking to improve their customer's experience. However capturing their customers' feedback isn't easy when customers don't complain when they are unhappy. This is because they think that taking the time to provide feedback isn't worth the time because the business simply don't respond on time. Therefore timely response is an important churn target variable.

4.3.3 Model Evaluation

The next step is to evaluate various predictive models based on their efficiency. The two models we have chosen to evaluate are the decision tree model and the Ada Boost Model. The entire data set is divided into Training and Test data sets. We first run the models on the Training sets and then obtain performance attributes. We predict churn using the test set.

4.3.4 Decision Tree Model

In computational complexity and communication complexity theories the decision tree model is the model of computation or communication in which an algorithm or communication process is considered to be basically a decision tree,that is, a sequence of branching operations based on comparisons of some quantities, the comparisons being assigned the unit computational cost. Han and Kamber (2011).

The branching operations are called "tests" or "queries". Several variants of decision tree models have been introduced, depending on the complexity of the operations allowed in the computation of a single comparison and the way of branching. Decision trees models are instrumental in establishing lower bounds for computational complexity for certain classes of computational problems and algorithms: the lower bound for worst-case computational complexity is proportional to the largest depth among the decision trees for all possible inputs for a given computational problem. The computation complexity of a problem or an algorithm expressed in terms of the decision tree model is called decision tree complexity or query complexity.

Classification by query computational complexity: Simple decision tree: The model in which every decision is based on the comparison of two numbers within constant time is called simply a decision tree model. It was introduced to establish computational complexity of sorting and searching.The simplest illustration of this lower bound technique is for the problem of finding the smallest number among n numbers using only comparisons. In this case the decision tree model is a binary tree. Algorithms for this searching problem may result in n different outcomes (since any of the n given numbers may turn out to be the smallest one).

However this lower bound is known to be slack, since the following simple argument shows that at least n - 1 comparisons are needed. Before the smallest number can be determined, every number except the smallest must "lose" (compare greater) in at least one comparison. Types:
1. **Linear decision tree**: Linear decision trees, just like the simple decision trees, make a branching decision based on a set of values as input. As opposed to binary decision trees, linear decision trees have three output branches. A linear function
2. **Algebraic decision tree**: Algebraic decision trees are a generalization of linear decision trees to allow test functions to be polynomials of degree d. Geometrically, the space is divided into semi-algebraic sets (a generalization of hyperplane). The evaluation of the complexity is more difficult. Wikipedia (2017c)

4.3.5 Limitations of Decison Tree

Trees do not tend to be as accurate as other approaches.Trees can be very non-robust. A small change in the training data can result in a big change in the tree, and thus a big change in final predictions.The problem of learning an optimal decision tree is known to be complete under several aspects of optimality and even for simple concepts. Consequently, practical decision tree learning algorithms are based on heuristics such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. To reduce the greedy effect of local optimality some methods such as the dual information distance) tree were proposed.

Decision-tree learners can create over-complex trees that do not generalize well from the training data. This is known as overfitting. Mechanisms such as pruning are necessary to avoid this problem (with the exception of some algorithms such as the Conditional Inference approach, that does not require pruning).For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favor of those attributes with more levels. However, the issue of biased predictor selection is avoided by the Conditional Inference approach.

4.3.6 Adaptive Boost Model

AdaBoost is a 'boosting' machine learning algorithm, used to enhance and improve performance. It is best used with weak learners to provide a suitable output. A weak learner can vary, for example it can be a stump in a decision tree. In the training dataset, each element is weighted. The model is used on the training dataset and then evaluated on the test set. Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers. Boosting is a general ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model.

Models are added until the training set is predicted perfectly or a maximum number of models are added. AdaBoost was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting. Modern boosting methods build on AdaBoost, most notably stochastic gradient boosting machines. AdaBoost is best used to boost the performance of decision trees on binary classification problems. AdaBoost was originally called AdaBoost.M1 by the authors of the technique Freund and Schapire. More recently it may be referred to as discrete AdaBoost because it is used for classification rather than regression. AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners.

These are models that achieve accuracy just above random chance on a classification problem.The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps. Weak models are added sequentially, trained using the weighted training data. (Jaakkola)

4.3.7 Adaboost ensemble

AdaBoost, short for "Adaptive Boosting", is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire. It can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing (e.g., their error rate is smaller than 0.5 for binary classification), the final model can be proven to converge to a strong learner.

While every learning algorithm will tend to suit some problem types better than others, and will typically have many different parameters and configurations to be adjusted before achieving optimal performance on a dataset, AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier. When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples.

Weak models are added sequentially, trained using the weighted training data.The process continues until a pre-set number of weak learners have been created (a user parameter) or no further improvement can be made on the training dataset.Once completed, you are left with a pool of weak learners each with a stage value. (Nickgillian.com)

This section lists some heuristics for best preparing your data for AdaBoost.
1) **Quality Data**: Because the ensemble method continues to attempt to correct misclassifications in the training data, you need to be careful that the training data is of a high-quality.
2) **Outliers**: Outliers will force the ensemble down the rabbit hole of working hard to correct for cases that are unrealistic. These could be removed from the training dataset.
3) **Noisy Data**: Noisy data, specifically noise in the output variable can be problematic. If possible, attempt to isolate and clean these from your training dataset.

4.3.8 Advantages of adaboost

AdaBoost is a powerful classification algorithm that has enjoyed practical success with applications in a wide variety of fields, such as biology, computer vision, and speech processing. Unlike other powerful classifiers, such as Support Vector Machine, AdaBoost can achieve similar classification results with much less tweaking of parameters or settings . The user only needs to choose which weak classifier might work best to solve their given classification problem.the user also needs to choose the number of boosting rounds that should be used during the training phase. The user can add several weak classifiers to the family of weak classifiers that should be used at each round of boosting. The AdaBoost algorithm will select the weak classifier that works best at that round of boosting.

4.3.9 Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another). It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table). Wikipedia (2017b)

4.3.10 Receiver operating characteristic in modeling

In statistics, a receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 - specificity). The ROC curve is thus the sensitivity as a function of fall-out. Schoonjans (2017) , Bmj.com (a)

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.

The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.The contingency table can derive several evaluation "metrics" (see infobox). To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test. Bmj.com (b)

A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 - specificity, the ROC graph is sometimes called the sensitivity vs (1 - specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100 percent sensitivity (no false negatives) and 100 percent specificity (no false positives). The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners (regardless of the positive and negative base rates). Wikipedia (2017d)

4.4 Analytics packages

4.4.1 Plotly

There are two main ways to initiate a plotly object in R. The plot_ly function transforms data into a plotly object, while the ggplotly() function transforms a ggplot object into a plotly object. Regardless of how a plotly object is created, printing it results in an interactive web-based visualization with tooltips, zooming, and panning enabled by default. The R package also has special semantics for arranging, linking, and animating plotly objects. This chapter discusses some of the philosophy behind each approach, explores some of their similarities, and explains why understanding both approaches is extremely powerful. The initial inspiration for the plot_ly() function was to support plotly.js chart types that ggplot2 doesn't support, such as 3D surface and mesh plots.

Over time, this effort snowballed into an interface to the entire plotly.js graphing library with additional abstractions inspired by the grammar of graphics. This newer "non-ggplot2" interface to plotly.js is currently not, and may never be, as fully featured as ggplot2. Since it can already translate a fairly large amount of ggplot objects to plotly objects. Plotly's R graphing library makes interactive, publication-quality graphs online. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, and 3D (WebGL based) charts.

4.4.2 Wordcloud

Text mining methods allow to highlight the most frequently used keywords in a paragraph of texts. One can create a word cloud, also referred as text cloud or tag cloud, which is a visual representation of text data using this package.The following arguments are used

1. **term.matrix**: A term frequency matrix whose rows represent words and whose columns represent documents.

2. **comonality.measure** A function taking a vector of frequencies for a single term, and returning a common frequency.

3. **max.words** Maximum number of words to be plotted.

4. **freq** : the word frequencies.

5. **scale** : A vector of length 2 indicating the range of the size of the words.

6. **min.freq** : words with frequency below min.freq will not be plotted.

7. **random.order** : plot words in random order. If false, they will be plotted in decreasing frequency.

8. **random.color** : choose colors randomly from the colors. If false, the color is chosen based on the frequency.

9. **rot.per** : proportion words with 90 degree rotation.

10. **colors** : color words from least to most frequent.

11. **ordered.colors** : if true, then colors are assigned to words in order.

12. **use.r.layout** : if false, then c++ code is used for collision detection, otherwise R is

used..
13. **fixed.asp** : if TRUE, the aspect ratio is fixed. Variable aspect ratio only supported if rot.per==0.
14. **Additional parameters** : to be passed to wordcloud, has no value. (RDocumentation)

4.4.3 RColorBrewer

This package creates nice looking color palettes especially for thematic maps. It has the following arguemnts:
1. **n**: Number of different colors in the palette, minimum 3, maximum depending on palette
2. **name**: A palette name from the lists below
3. **type**: One of the string "div", "qual", "seq", or "all"
4. **select**: A list of names of existing palettes
5. **exact.n**: If TRUE, only display palettes with a color number given by n

"Brewer.pal" makes the color palettes from ColorBrewer available as R palettes."display.brewer.pal" displays the selected palette in a graphics window."display.brewer.all" displays the a few palettes simultaneously in a graphics window."brewer.all.info" returns information about the available palettes as a dataframe.
There are 3 types of palettes, sequential, diverging, and qualitative.
1. Sequential palettes are suited to ordered data that progress from low to high. Lightness steps dominate the look of these schemes, with light colors for low data values to dark colors for high data values.
2. Diverging palettes put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues.
3. Qualitative palettes do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative

schemes are best suited to representing nominal or categorical data. (Mike)

4.4.4 Choroplethr

Choropleth maps are a useful way to visualize this kind of information. In a choropleth, regions are colored based on some metric, such as which presidential candidate a state voted for. There is a package in R to facilitate creating choropleths called choroplethr. choroplethr makes it easy to visualize data. choroplethr simplifies the creation of choropleth maps in R. Choropleths are thematic maps where geographic regions, such as states, are colored according to some metric, such as the number of people who live in that state.

choroplethr simplifies this process by:

1. Providing ready-made functions for creating choropleths using 220 different maps.
2. Providing API connections to interesting data sources for making choropleths.
3. Providing a framework for creating choropleths from arbitrary shapefiles.

Functions in choroplethr include the following:

1. **choroplethr_wdi**: Create a country level choropleth using data from the World Bank's World Development Indicators (WDI)

2. **Choropleth**: The base Choropleth object.

3. **Admin1Choropleth**: An R6 object for creating Administration Level 1 choropleths.

4. **admin1_choropleth**: Create an admin1-level choropleth for a specified country

4. admin1_region_choropleth: Create a map of Administrative Level 1 regions

5. **Admin1RegionChoropleth**: An R6 object for creating Administration Level 1 choropleths based on regions.

6. **df_japan_census**: A data.frame containing basic demographic information about Japan.

7. **county_choropleth_acs**: Create a US County choropleth from ACS data

8. **CountryChoropleth**: An R6 object for creating country-level choropleths.

9. **country_choropleth**: Create a country-level choropleth

10. **county_choropleth**: Create a choropleth of US Counties

11. **CountyZoomChoropleth**: Create a county-level choropleth that zooms on counties, not states.

12. **df_county_demographics**: A data.frame containing demographic statistics for each county in the United States.

13. **county_zoom_choropleth**: Create a choropleth of USA Counties, with sensible defaults, that zooms on counties.

14. **CountyChoropleth**: Create a county-level choropleth

15. **get_acs_df**: Returns a data.frame representing US Census American Community Survey (ACS) estimates.

16. **et_county_demographics**: Get a handful of demographic variables on US. 17. **state_choropleth**: Create a choropleth of US States

18. **df_pop_country**: A data.frame containing population estimates for Countries in 2012.

19. **zip_map** : Create a map visualizing US ZIP codes with sensible defaults.

4.4.5 Lubridicate

Lubridate provides tools that make it easier to parse and manipulate dates. These tools are grouped below by common purpose:

1. **Parsing dates**: Lubridate's parsing functions read strings into R as POSIXct date-time objects. Users should choose the function whose name models the order in which the year ('y'), month ('m') and day ('d') elements appear the string to be parsed. A very flexible and user friendly parser is provided by parse_date_time. Lubridate can also parse partial dates from strings into Period-class objects with the functions hm, hms and ms. Lubridate has an inbuilt very fast POSIX parser. This functionality is as yet optional and could be activated with options(lubridate.fasttime = TRUE). Lubridate will automatically detect POSIX strings and use fast parser instead of the default strptime utility.

2. **Manipulating dates**: Lubridate distinguishes between moments in time (known as instants) and spans of time (known as time spans, see Timespan-class). Time spans are further separated into Duration-class, Period-class and Interval-class objects. Instants are specific moments of time. Date, POSIXct, and POSIXlt are the three object classes

Base. R recognizes as instants. is.Date tests whether an object inherits from the Date class.

3. **Rounding dates**: Instants can be rounded to a convenient unit using the functions ceiling_date, floor_date and round_date.

4. **Time zones**: Lubridate provides two helper functions for working with time zones. with_tz changes the time zone in which an instant is displayed. The clock time displayed for the instant changes, but the moment of time described remains the same. force_tz changes only the time zone element of an instant. The clock time displayed remains the same, but the resulting instant describes a new moment of time.

5. **Timespans**: A timespan is a length of time that may or may not be connected to a particular instant. For example, three months is a timespan. So is an hour and a half. Base R uses difftime class objects to record timespans. However, people are not always consistent in how they expect time to behave. Sometimes the passage of time is a monotone progression of instants that should be as mathematically reliable as the number line. On other occasions time must follow complex conventions and rules so that the clock times reflect is observed in terms of daylight, season, and congruence with the atomic clock. Lubridate creates three additional timespan classes, each with its own specific and consistent behavior: Interval-class, Period-class and Duration-class.

6. **Durations** measure the exact amount of time that occurs between two instants. This can create unexpected results in relation to clock times if a leap second, leap year, or change in daylight savings time occurs in the interval. Functions for working with durations include is.duration, as.duration and duration.dseconds, dminutes, dhours, ddays, dweeks and dyears convenient lengths.

7. **Periods**: Periods measure the change in clock time that occurs between two instants. Periods provide robust predictions of clock time in the presence of leap seconds, leap year. Functions for working with periods include is.period, as.period and period. seconds, minutes,hours, days, weeks, months and years quickly create periods of convenient lengths.

8. **Intervals**: Intervals are timespans that begin at a specific instant and end at a specific instant. Intervals retain complete information about a timespan. They provide the only reliable way to convert between periods and durations. Functions for working with intervals include is.interval, as.interval, interval, int_shift,nt_flip, int_aligns, int_overlaps.

Intervals can also be manipulated with intersect, union, and setdiff(). Spinu (2016)

4.5 Integration using Shiny Web Application

The final stage was to integrate our results into a Shiny Web application. We used three scripts namely ui.R, server.R and global.R to run our application.The global.R file contains the backend code for the data visualizations and the Adaptive Boost model.
The user-interface (ui) script is used to design the layout and appearance of the application. It is present in the source file ui.R.The User Interface in Shiny may include buttons, tabs, labels, boxes, lists, etc. to get the input from the user.

The Server script contains the instructions to build and run the functions of the app. It is present in the source file server.R. Server is used to define the action that has to be performed for the corresponding changes in the UI. The server side script makes use of functions such as renderPlot, renderPrint, renderTable, etc. to dynamically present the output based on the varying inputs as per the user's requirements.

The global file is used to declare environment variables, to run the code for the visualizations and the back end model. It is defined in global.R. When the app is run, the global.R file is first executed and then the Shiny application is launched. The variables declared in global.R have global scope and can be used from any other script.

The layout consists of four tabs. The radio buttons can be used by the user to choose the variables to represent in the Univariate analysis. The "Univariate" tab is used to represent the dynamic graphs of the Univariate analysis. The "Bivariate" tab is used to represent the dynamic graphs of the bivariate analysis. The "Model" tab is used to represent the results from the Ada Boost model and our consolidated prediction matrix. The "Prediction" tab is used to represent the prediction outcomes and corresponding test set.

The working of our application starts when we run the application from RStudio. The global.R file is first executed. This file contains the source code for the Ada Boost

Model and for plotting the visualizations. Once the global.R file is run , the shiny application is launched through the browser. Ui.R generates the user interface for the client. The inputs given by the user are stored in input variables. An output function from the main panel which specifies the output variables. In server.R the input variables are used for performing various operations on the user's input. The output is presented using the output variables and the corresponding output functions.

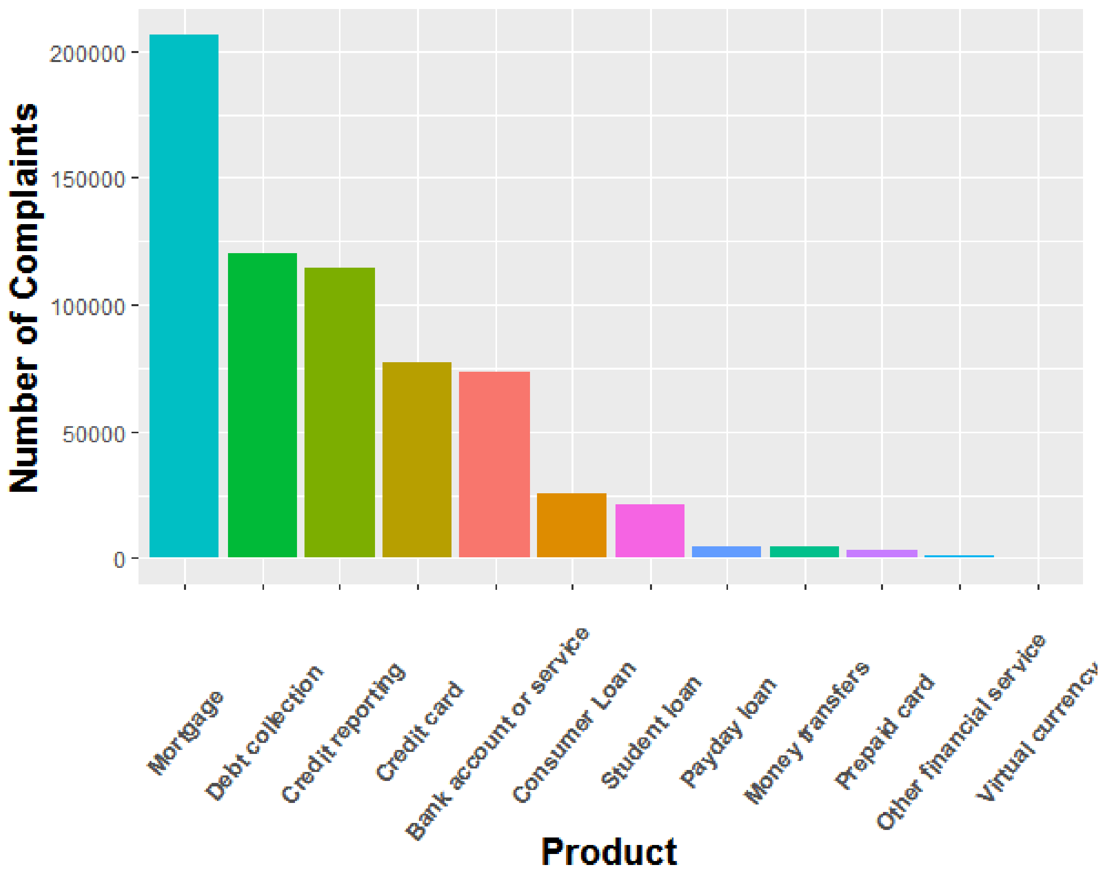


Figure 4.2: Univariate Graph of Product vs Number of Complaints

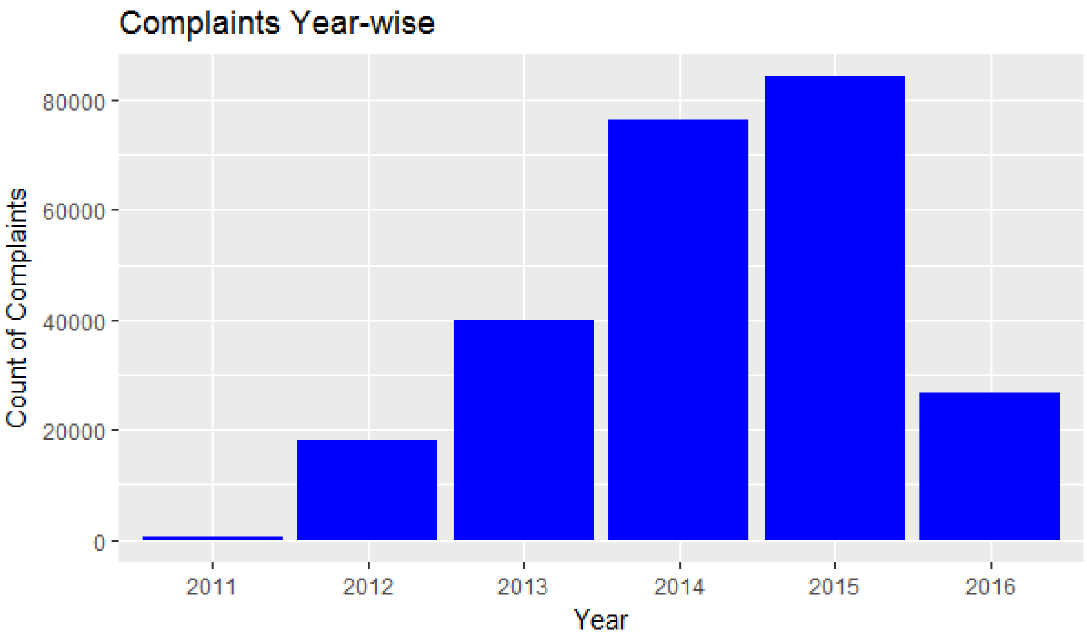
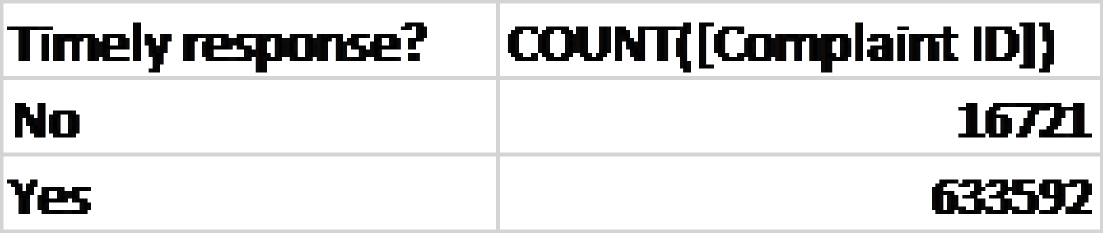
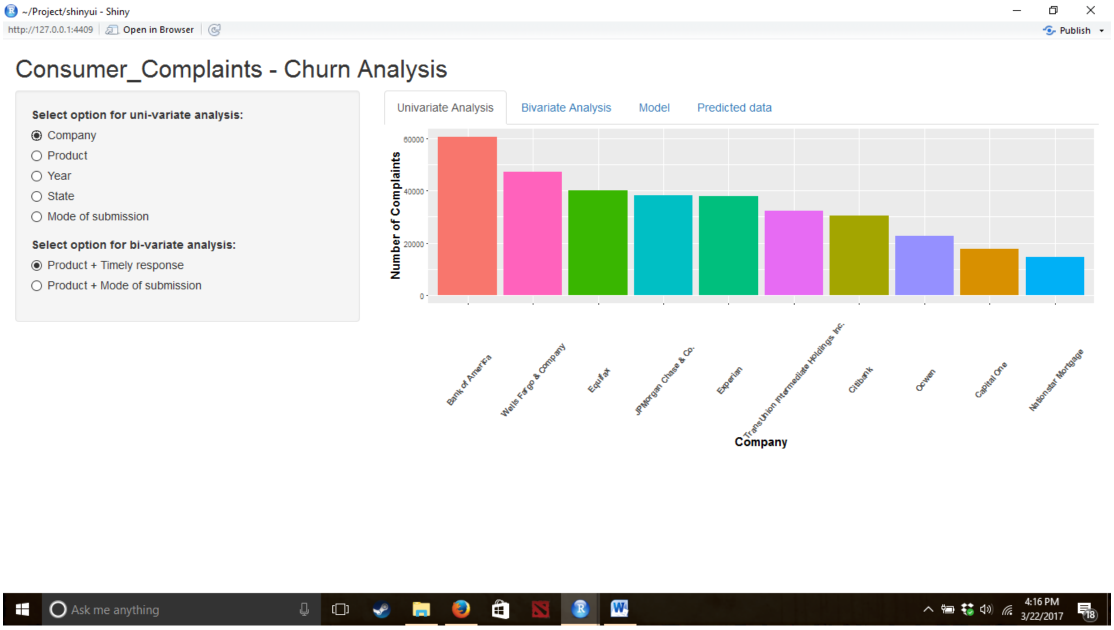


Figure 4.3: Graph of Year vs Number of Complaints



Timely response?	COUNT([Complaint ID])
No	16721
Yes	633592

Figure 4.4: Table of Consumers receiving Timely Response



Consumer_Complaints - Churn Analysis

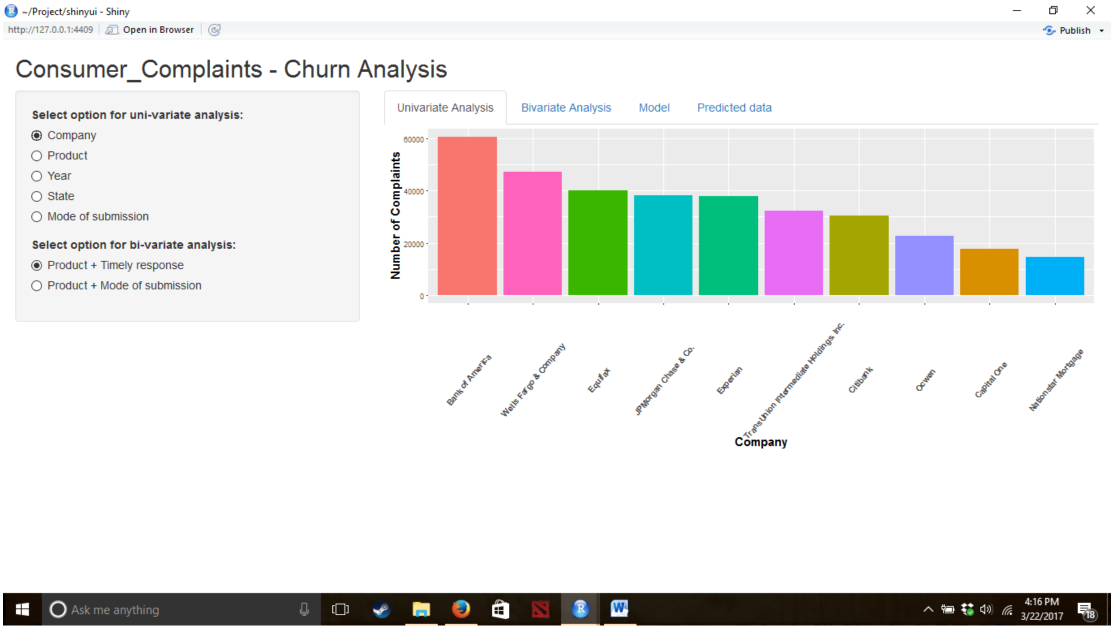
Select option for uni-variate analysis:

- ● Company
- ○ Product
- ○ Year
- ○ State
- ○ Mode of submission

Select option for bi-variate analysis:

- ● Product + Timely response
- ○ Product + Mode of submission

Figure 4.5: User Interface of Shiny Application



Consumer_Complaints - Churn Analysis

Select option for uni-variate analysis:

- ● Company
- ○ Product
- ○ Year
- ○ State
- ○ Mode of submission

Select option for bi-variate analysis:

- ● Product + Timely response
- ○ Product + Mode of submission

Figure 4.6: Prediction tab for SHINY UI

CHAPTER 5

CODING AND TESTING

Here we mention all the code for the interface and analytics

5.1 Years vs complaints

```
complaints$Date_received <-
   as.Date(complaints$Date_received, "%m/%d/%Y")
complaints$Year <-
as.factor(year(complaints$Date_received))
ggplot(complaints, aes(Year)) + geom_bar(fill = "blue")
+ ggtitle("Complaints␣Year-wise") +
   xlab("Year") + ylab("Count␣of␣Complaints")
```

5.2 Submitted via vs complaints

```
table(complaints$Submitted_via)
prop.table(table(complaints$Submitted_via))
ggplot(complaints, aes(reorder(
   Submitted_via,-table(complaints$Submitted_via)
   [Submitted_via]),
   fill = Submitted_via)) + geom_bar() +
   xlab("Mode␣of␣Complaints␣Submission") +
   ylab("Number␣of␣Complaints") +
   scale_y_continuous(breaks = seq(0, 350000, 50000)) +
   theme(
   axis.text.x = element_text(
     angle = 50,
     size = 10,
     vjust = 0.4,
     face = "bold"
   ),
   plot.title = element_text(size = 20, face = "bold",
   vjust = 2),
   axis.title.x = element_text(
     size = 15,
     vjust = -0.35
   ),
   axis.title.y = element_text(face = "bold",
   vjust = 0.35,
   size = 15)) + theme(legend.position = "none")
```

5.3 State vs complaints

```
head(by_state, 10)
tail(by_state, 10)
ggplot(head(by_state, 10), aes(reorder(State,-Count),
        Count, fill = State)) +
        geom_bar(stat = "identity") +
  xlab("State") + ylab("Number␣of␣Complaints") + theme(
  axis.text.x = element_text(
    angle = 50,
    size = 10,
    vjust = 0.4,
    face = "bold"
  ),
  plot.title = element_text(size = 20, face = "bold",
                                        vjust = 2),
  axis.title.x = element_text(
    face = "bold",
    size = 15,
    vjust = -0.35
  ),
  axis.title.y = element_text(face = "bold",
  vjust = 0.35, size = 15)
) + theme(legend.position = "none")
```

5.4 Product vs complaints

```
table(complaints$Product)
ggplot(complaints, aes(reorder(Product,-table(
  complaints$Product)[Product]), fill = Product)) +
  geom_bar() + xlab("Product") +
  ylab("Number␣of␣Complaints") +
  theme(
  axis.text.x = element_text(
    angle = 50,
    size = 10,
    vjust = 0.4,
    face = "bold"
  ),
  plot.title = element_text(size = 20, face = "bold",
  vjust = 2),
  axis.title.x = element_text(
    face = "bold",
    size = 15,
    vjust = -0.35
  ),
  axis.title.y = element_text(face = "bold",
  vjust = 0.35,size = 15))
  + theme(legend.position = "none")
```

5.5 Month vs complaints

```
complaints$Month <-
as.factor(month(complaints$Date_received))
ggplot(complaints, aes(Month))
  + geom_bar(fill = "brown")
  + ggtitle("Complaints Month-wise") +
  xlab("Month") + ylab("Count of Complaints")
```

5.6 Day vs complaints

```
complaints$Day <-
as.factor(day(complaints$Date_received))
ggplot(complaints, aes(Day)) +
  geom_bar(fill = "maroon") +
  ggtitle("Complaints Day-wise") +
  xlab("Day Of the Month") +
  ylab("Count of Complaints")
```

5.7 company vs complaints

```
ggplot(head(by_company, 10), aes(reorder(Company,-Count),
  Count, fill = Company)) + geom_bar(stat = "identity") +
  xlab("Company") + ylab("Number␣of␣Complaints") +
  theme(
  axis.text.x = element_text(
    angle = 50,
    size = 10,
    vjust = 0.4,
    face = "bold"
  ),
  plot.title = element_text(size = 20, face = "bold",
                             vjust = 2),
  axis.title.x = element_text(
    face = "bold",
    size = 15,
    vjust = -0.35
  ),
  axis.title.y = element_text(face = "bold",
  vjust = 0.35, size = 15)
) + theme(legend.position = "none")
```

5.8 Bivariate product timely complaints

```
table(complaints$Product, complaints$Timely_response)
ggplot(complaints, aes(reorder(Product,
            table(complaints$Product)[Product]),
            fill =  Timely_response)) + geom_bar() +
  xlab("Product") + ylab("Number of Complaints") +
  coord_flip() +
  theme(
  axis.text.x = element_text
  (
    angle = 50,
    size = 10,
    vjust = 0.4,
    face = "bold"
  ),
  plot.title = element_text(size = 20, face = "bold",
  vjust = 2),
  axis.title.x = element_text(
    face = "bold",
    size = 15,
    vjust = -0.35
  ),
  axis.title.y = element_text(face = "bold",
  vjust = 0.35, size = 15)
)
```

5.9 Bivariate product submitted complaints

```
table(complaints$Product,
complaints$Submitted_via)
ggplot(complaints, aes(reorder(Product,
table(complaints$Product)
[Product]),
fill = Submitted_via)) + geom_bar() +
  xlab("Product") + ylab("Number of Complaints") +
  coord_flip() +
  theme(
    axis.text.x = element_text(
    angle = 50,
    size = 10,
    vjust = 0.4,
    face = "bold"
  ),
  plot.title = element_text(size = 20, face = "bold",
  vjust = 2),
  axis.title.x = element_text(
    face = "bold",
    size = 15,
    vjust = -0.35
  ),
  axis.title.y = element_text(face = "bold",
  vjust = 0.35,
  size = 15)
)
```

5.10 Exploratory Data Analytics

```
numNA <- function(x) { return(sum(is.na(x)))}
propNA <- function(x) { return(numNA(x) / length(x))}
numUnique <- function(x) { return(length(unique(x))) }

numRows = length(colnames(complaints))
exportableTable = data.frame(numMissingVal =
                              integer(numRows),
                              propMissingVal =
                              double(numRows))
exportableTable["numMissingVal"]=
          apply(complaints,2,numNA)
exportableTable["propMissingVal"]=
          apply(complaints,2,propNA)
exportableTable["numUnique"]=
          apply(complaints,2,numUnique)

colnames(exportableTable)=c("Number␣of␣Missing␣Values",
                           "Proportion␣of␣Misisng␣Values",
                           "Number␣of␣Unique␣Values")
rownames(exportableTable) = colnames(complaints)
kable(exportableTable)
```

5.11 Global.R

```
complaints <- read.csv
  ("C:/Users/Chirag/Desktop/complaints.csv")
p1 <-
  ggplot(head(by_company, 10),
  aes(reorder(Company,-Count),
  Count, fill = Company)) +
  geom_bar(stat = "identity")
  + xlab("Company")+ ylab("Number␣of␣Complaints")
  + theme(
    axis.text.x = element_text(
      angle = 50,
      size = 10,
      vjust = 0.4,
      face = "bold"
    ),
    plot.title = element_text(size = 20, face = "bold",
    vjust = 2),
    axis.title.x = element_text(
      face = "bold",
      size = 15,
      vjust = -0.35
    ),
    axis.title.y = element_text(face = "bold",
    vjust = 0.35, size = 15)
  ) + theme(legend.position = "none")
p2 <-
  ggplot(complaints,
  aes(reorder(Product,-table(complaints$Product)
  [Product]), fill = Product)) + geom_bar()
  + xlab("Product") + ylab("Number␣of␣Complaints")
  + theme(
    axis.text.x = element_text(
      angle = 50,
      size = 10,
      vjust = 0.4,
      face = "bold"
    ),
    plot.title = element_text(size = 20, face = "bold",
    vjust = 2),
    axis.title.x = element_text(
      face = "bold",
      size = 15,
      vjust = -0.35
    ),
    axis.title.y = element_text(face = "bold",
    vjust = 0.35, size =15)
  ) + theme(legend.position = "none")
complaints$Date_received <-
  as.Date(complaints$Date_received, "%m/%d/%Y")
complaints$Year<-
      as.factor(year(complaints$Date_received))
p3 <-
  ggplot(complaints, aes(Year)) + geom_bar(fill = "blue")
  +ggtitle("Complaints␣Year-wise")
  + xlab("Year") + ylab("Count␣of␣Complaints")
p4 <-
  ggplot(head(by_state, 10), aes(reorder(State,-Count),
  Count, fill = State)) + geom_bar(stat = "identity")
  + xlab("State") + ylab("Number␣of␣Complaints")
  + theme(
    axis.text.x = element_text(
      angle = 50,
      size = 10,
      vjust = 0.4,
      face = "bold"
    ),
    plot.title = element_text(size = 20, face = "bold",
    vjust = 2),
      axis.title.x = element_text(
        face = "bold",
        size = 15,
        vjust = −0.35
      ),
      axis.title.y = element_text(face = "bold",
      vjust = 0.35, size =15)
    ) + theme(legend.position = "none")
p5 <−
    ggplot(complaints, aes(reorder(
    Submitted_via,−table(complaints$Submitted_via)
    [Submitted_via]),
    fill = Submitted_via)) + geom_bar()
    + xlab("Mode␣of␣Complaints␣Submission")
    + ylab("Number␣of␣Complaints")
    + scale_y_continuous(breaks = seq(0, 350000, 50000))
    + theme(
    axis.text.x = element_text(
        angle = 50,
        size = 10,
        vjust = 0.4,
        face = "bold"
      ),
      plot.title = element_text(size = 20, face = "bold",
      vjust = 2),
      axis.title.x = element_text(
        face = "bold",
        size = 15,
        vjust = −0.35
      ),
      axis.title.y = element_text(face = "bold",
      vjust = 0.35, size =15)) +
      theme(legend.position = "none")
p6 <-
    ggplot(complaints, aes(reorder(Product,
    table(complaints$Product)[Product]),
    fill =  Timely_response)) + geom_bar() +
    xlab("Product") + ylab("Number␣of␣Complaints")
    + coord_flip() + theme(
      axis.text.x = element_text(
        angle = 50,
        size = 10,
        vjust = 0.4,
        face = "bold"
      ),
      plot.title = element_text(size = 20, face = "bold",
      vjust = 2),
      axis.title.x = element_text(
        face = "bold",
        size = 15,
        vjust = -0.35
      ),
      axis.title.y = element_text(face = "bold",
      vjust = 0.35, size=15)
    )
p7 <-ggplot(complaints, aes(reorder(Product,
    table(complaints$Product)[Product]),
    fill = Submitted_via)) + geom_bar()
    + xlab("Product") + ylab("Number␣of␣Complaints")
    + coord_flip() + theme(
      axis.text.x = element_text(
        angle = 50,
        size = 10,
        vjust = 0.4,
        face = "bold"
      ),
      plot.title = element_text(size = 20, face = "bold",
      vjust = 2),
      axis.title.x = element_text(
        face = "bold",
        size = 15,
        vjust = -0.35
      ),
      axis.title.y = element_text(face = "bold",
      vjust = 0.35,size = 15)
    )
library(ada)
library(rattle)
library(magrittr) # For the %>% and %<>% operators.

building <- TRUE
scoring  <- !building



# A pre-defined value is used to
reset the random seed so that results are repeatable.

crv$seed <- 42

#=====================================================



crs$dataset <-
    read.csv(
      "file:///C:/Users/Chirag/Desktop/complaints.csv",
      na.strings = c(".", "NA", "", "?"),
      strip.white = TRUE,
      encoding = "UTF-8"
    )

#=======================================================



# Build the training/validate/test datasets.

set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset)
   # 245709 observations
crs$sample <-
    crs$train <-
    sample(nrow(crs$dataset), 0.7 * crs$nobs)
    # 171996 observations
crs$validate <-
    sample(setdiff(seq_len(nrow(crs$dataset)), crs$train),
    0.15 *crs$nobs)
    # 36856 observations
crs$test <-
    setdiff(setdiff(seq_len(nrow(crs$dataset)), crs$train),
    crs$validate) # 36857 observations
op10 <- crs$validate
# The following variable selections have been noted.

crs$input <- c(
    "Date_received",
    "Product",
    "Issue",
    "Company",
    "State",
    "ZIP_code",
    "Consumer_consent_provided.",
    "Submitted_via",
    "Date_sent_to_company",
    "Company_response_to_consumer",
    "Consumer_disputed",
    "Complaint.ID"
)

crs$numeric <- "Complaint.ID"

crs$categoric <- c(
    "Date_received",
    "Product",
    "Issue",
    "Company",
    "State",
    "ZIP_code",
    "Consumer_consent_provided.",
    "Submitted_via",
    "Date_sent_to_company",
    "Company_response_to_consumer",
    "Consumer_disputed"
)

crs$target  <- "Timely_response"
crs$risk    <- NULL
crs$ident   <- NULL
crs$ignore  <-
    c(
    "Sub.product",
    "Sub.issue",
    "Consumer_complaint_narrative",
    "Company.public.response",
    "Tags"
  )
crs$weights <- NULL

#=========================================================


# Ada Boost

# The ‘ada’ package implements the boost algorithm.

# Build the Ada Boost model.

set.seed(crv$seed)
crs$ada <- ada::ada(
  Timely_response ~ .,
  data = crs$dataset[crs$train ,
  c(crs$input, crs$target)],
  control = rpart::rpart.control(
    maxdepth = 30,
    cp = 0.010000,
    minsplit = 20,
    xval = 10
  ),
  iter = 50
)

# Print the results of the modelling.
```

```
MYada <- crs$ada
op1 <- MYada

#print(crs$ada)
op2 <- round(crs$ada$model$errs[crs$ada$iter,], 2)

op3 <- sort(names(listAdaVarsUsed(crs$ada)))

#print(listAdaVarsUsed(crs$ada))
op4 <- listAdaVarsUsed(crs$ada)
# Time taken: 47.66 mins



# Evaluate model performance.

# Obtain the response from the Ada Boost model.

crs$pr <-
  predict(crs$ada, newdata = crs$dataset[crs$validate,
  c(crs$input,crs$target)])
op5 <- crs$pr
# Generate the confusion matrix showing counts.

op6 <-
  table(
    crs$dataset[crs$validate, c(crs$input, crs$target)]
    $Timely_response,
    crs$pr,
    useNA = "ifany",
    dnn = c("Actual", "Predicted")
  )
```

```
# Generate the confusion matrix showing proportions.

pcme <- function(actual, cl)
{
  x <- table(actual, cl)
  nc <- nrow(x) # Number of classes.
  nv <-
    length(actual) - sum(is.na(actual) |
                                   is.na(cl))
             # Number of values.
  tbl <- cbind(x / nv,
             Error =sapply(1:nc,
                    function(r)
                    round(sum(x[r,-r]) /sum(x[r,]),2)))
  names(attr(tbl, "dimnames")) <-
       c("Actual", "Predicted")
  return(tbl)
}
per <-
  pcme(crs$dataset[crs$validate,c(crs$input,crs$target)]
  $Timely_response, crs$pr)
round(per, 2)
op7 <- per
# Calculate the overall error percentage.
#cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))
op8 <-
    (100 * round(1 - sum(diag(per), na.rm = TRUE), 2))

# Calculate the averaged class error percentage.

#cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))
```

```
op9<-
  (100 * round(mean(per[, "Error"], na.rm = TRUE), 2))
```

5.12 Shiny

5.12.1 UI.R

```
library(shiny)

shinyUI(fluidPage(
  titlePanel("Consumer_Complaints␣-␣Churn␣Analysis"),
  sidebarLayout(
    sidebarPanel(
      radioButtons(
        "var",
        "Select␣option␣for␣uni-variate␣analysis:",
        c("Company", "Product", "Year", "State", "Mode␣of
␣␣␣␣␣␣␣␣submission"),
        selected = "Company"
      ),
      radioButtons(
        "var2",
        "Select␣option␣for␣bi-variate␣analysis:",
        c("Product␣+␣Timely␣response",
        "Product␣+␣Mode␣of␣submission")
      )
    ),
    mainPanel(tabsetPanel(
      tabPanel("Univariate␣Analysis",
      plotOutput(outputId = "x")),
      tabPanel("Bivariate␣Analysis",
      plotOutput(outputId = "y")),
      tabPanel("Model",
      verbatimTextOutput(outputId = "a")),
      tabPanel("Predicted␣data",
      verbatimTextOutput(outputId = "b"))
    ))
  )
))
```

5.12.2 Server.R

```
library(shiny)
library(ggplot2)



shinyServer(function(input, output) {
  output$x <- renderPlot({
    p <- switch(
      input$var,
      "Company" = p1,
      "Product" = p2,
      "Year" = p3,
      "State" = p4,
      "Mode␣of␣submission" = p5
    )
    print(p)

  })
  output$y <- renderPlot({
    q <- switch(
      input$var2,
      "Product␣+␣Timely␣response" = p6,
      "Product␣+␣Mode␣of␣submission" = p7
    )
    print(q)
  })
  output$a <- renderPrint({
   print("Results␣of␣the␣modeling")
   print(op1)
   print(op2)
   print("Variables␣actually␣used
␣␣␣␣␣␣␣␣␣␣in␣tree␣construction")
   print(op3)
   print("Frequency␣of␣variables␣actually␣used:")
   print(op4)
   print("Evaluate␣model␣performance")
   print("Generate␣confusion␣matrix␣showing␣counts:")
   print(op6)
   print("Generate␣the␣confusion␣matrix
␣␣␣␣␣␣␣␣␣␣showing␣proportions")
   print(op7)
   print("Calculate␣the␣overall
␣␣␣␣␣␣␣␣␣␣error␣percentage:")
   print(op8)
   print("Calculate␣the␣averaged
␣␣␣␣␣␣␣␣␣␣class␣error␣percentage:")
   print(op9)
  })
  output$b <- renderPrint({
    print("␣Prediction␣set
␣␣␣␣(Obtain␣the␣response␣from␣the␣Ada␣Boost␣model)")
    print(op5)
    print("crs$validate:␣Complaint␣IDs")
    print(op10)
  })
})
```

5.13 Outputs

1. **Days vs complaints graph :**
Based on this exploration, it is found that the 23rd day of the month has the highest number of complaints.

2. **Months vs complaints graph :**
Based on this exploration, it is found that the 3rd month has the highest number of complaints.

3. **Product vs complaints graph :**
Based on this exploration, it is found that the product - Mortgage has the highest number of complaints.

4. **State vs complaints graph :**
Based on this exploration, it is found that the state - California has the highest number of complaints.

5. **Sub-product vs complaints graph :**
Based on this exploration, it is found that the yea - Bank of America has the highest number of complaints.

6. **Years vs complaints graph :**
Based on this exploration, it is found that the year - 2015 has the highest number of complaints.

7. **Bivariate - Product + Mode of submission vs Complaints graph :**
Based on this exploration, it is found that the product - Mortgage has the highest number of complaints and the most commonly used mode of submission is Web

8. **Bivariate - Product + Timely response vs Complaints graph :**
Based on this exploration, it is found that the product - Mortgage has the highest num-

ber of complaints and most of the consumers were serviced on time

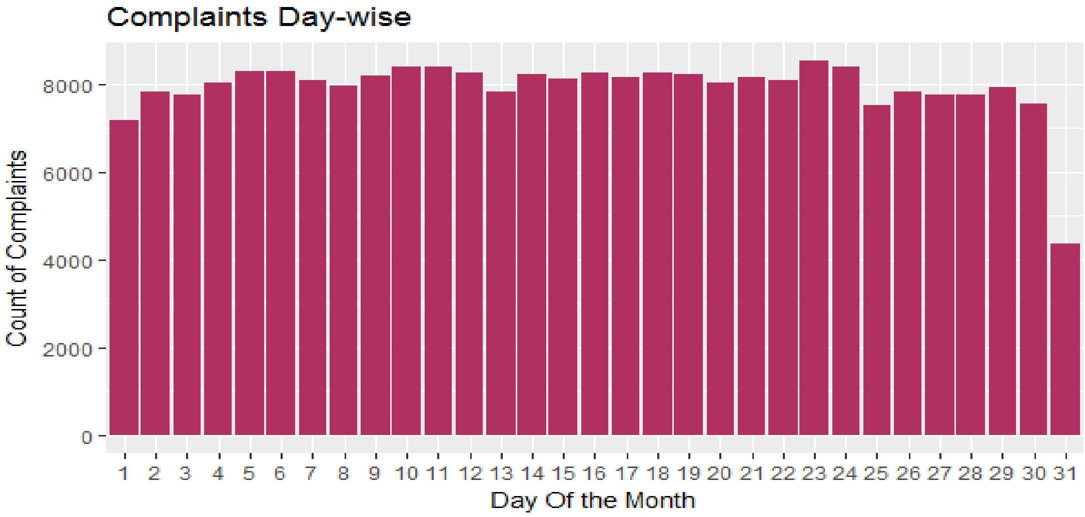


Figure 5.1: Days vs complaints

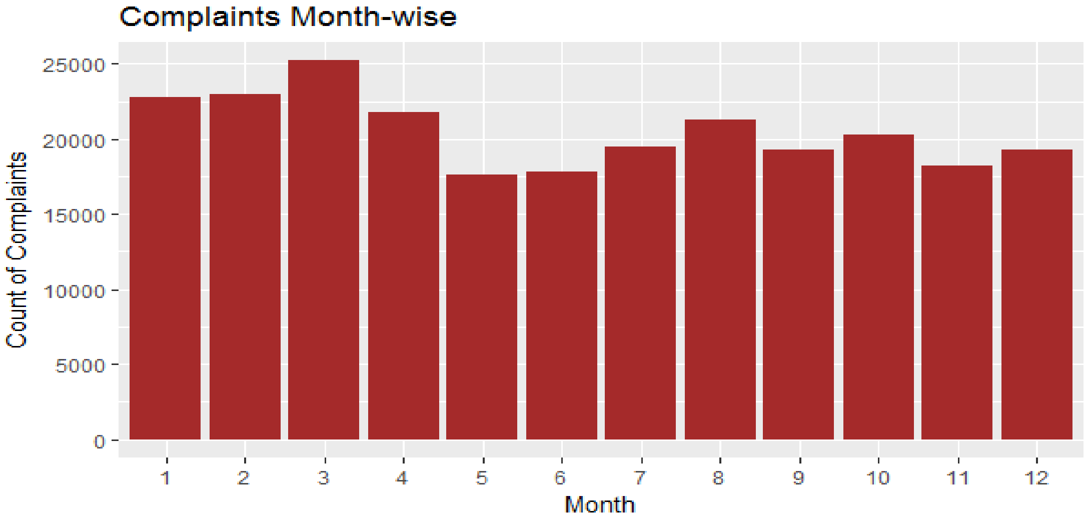


Figure 5.2: Months vs complaints

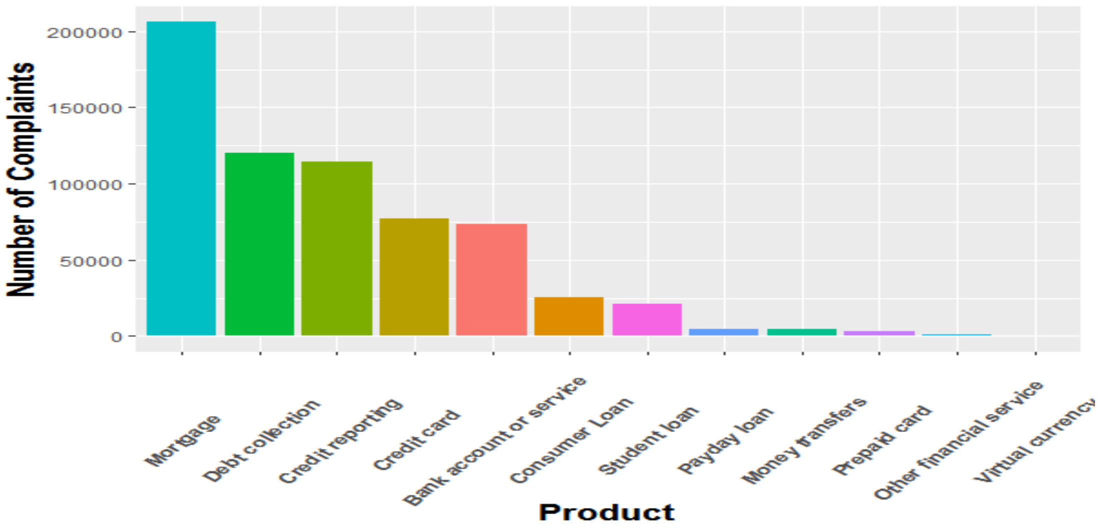


Figure 5.3: Products vs complaints

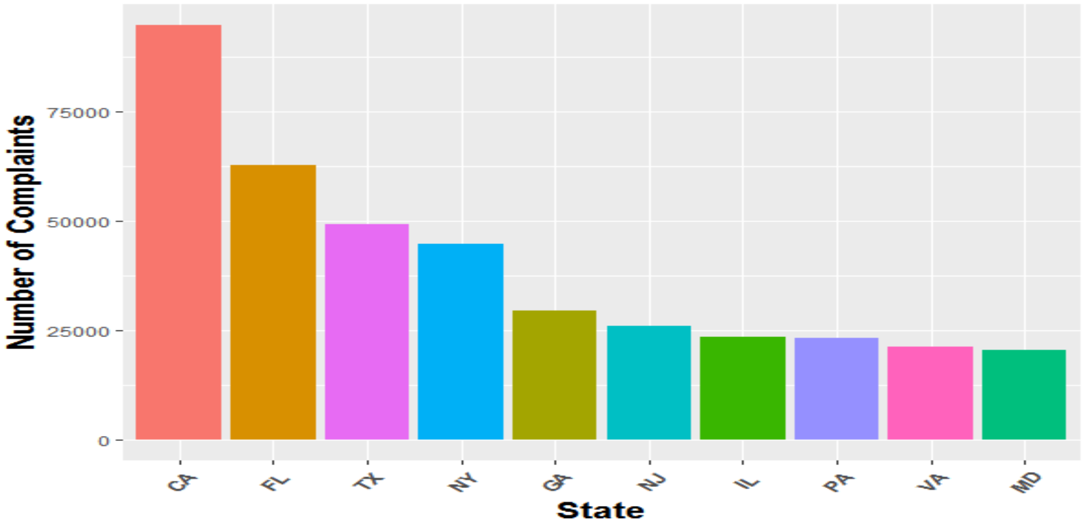


Figure 5.4: States vs complaints

5.14 Inference

We ran the Decision Tree Model and the Adaptive Boost Model and evaluated the error rates. We had to choose the one, which had a lower error rate and was compatible with our data. During the evaluation of the models, both the decision tree and adaptive boosting showed similar results for the error rate using error matrices.Hence, we made use of the ROC curve to find the area under curve for both the models. All these steps were made simpler using rattle (). Based on the results of the ROC curve, it was found that adaptive boosting had a higher area under curve compared to the decision tree model.

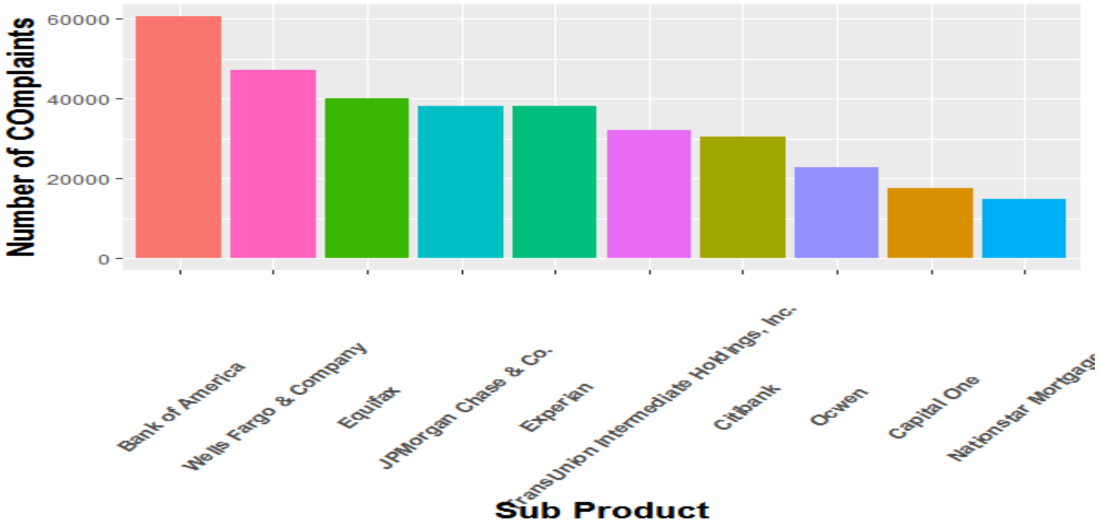


Figure 5.5: Sub-products vs complaints

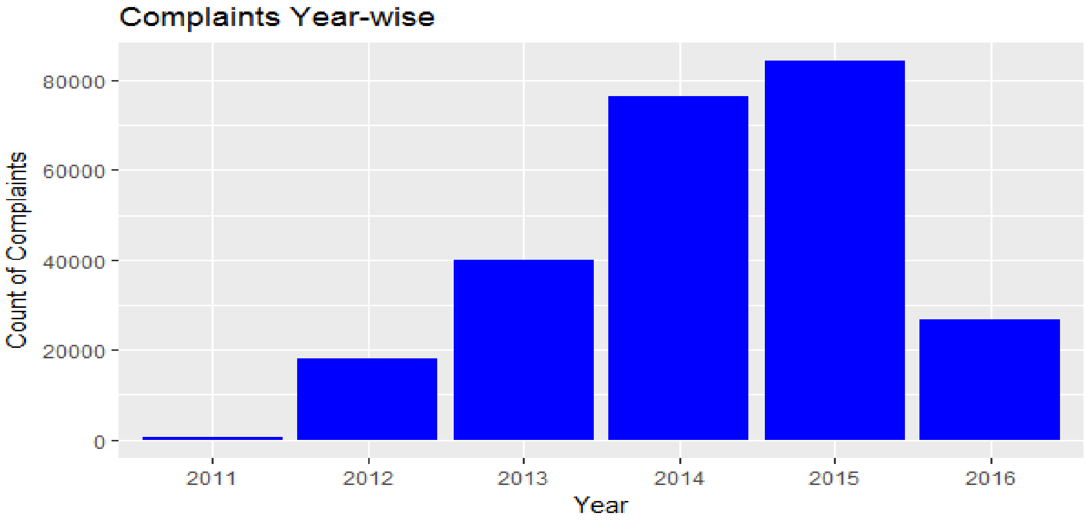


Figure 5.6: Year vs complaints

We found that the decision tree model had an overall area under curve (AUC) of 0.75. The Adaptive Boost Model had an AUC of 0.85. Due to its efficiency we chose the Adaptive boost model and ran it on our training data set. We performed a final evaluation and ran the model on the test set. We obtained a consolidated prediction matrix. We also obtained individual predictions for each of the complaint Ids in our test set.

The final stage was to integrate our results into a Shiny Web application. We used three scripts namely ui.R, server.R and global.R to run our application. The global.R file contains the backend code for the data visualizations and the Adaptive Boost model.

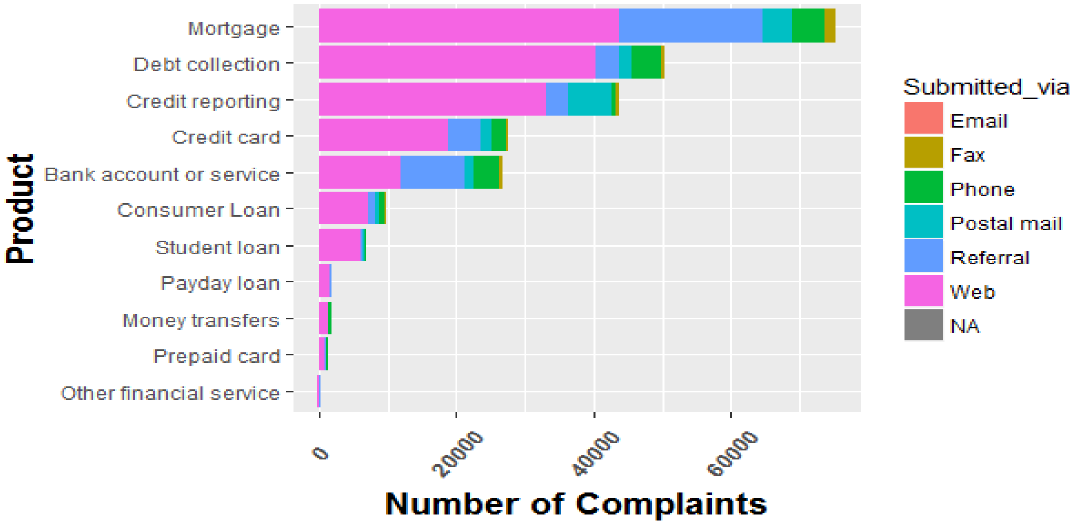


Figure 5.7: Product + Mode of submission vs complaints

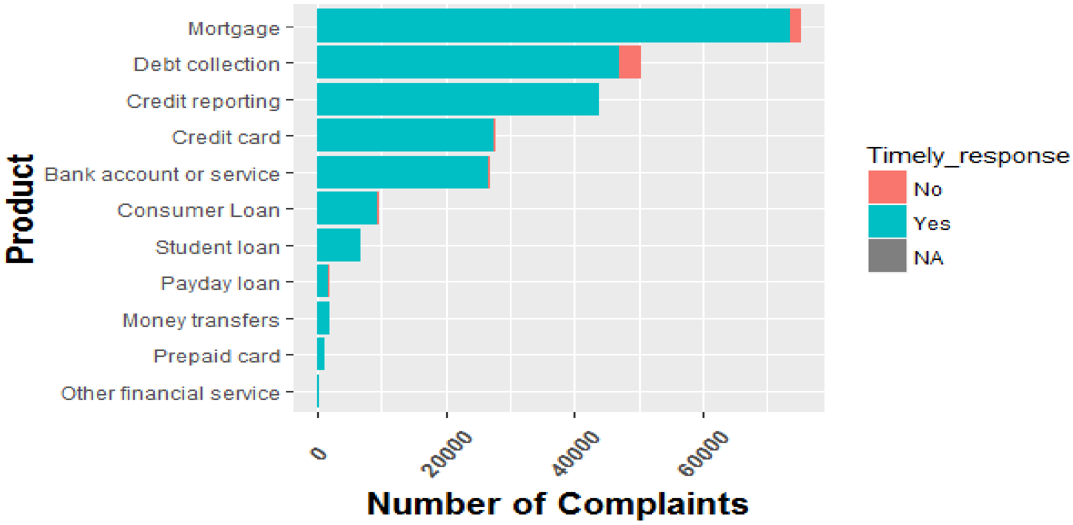


Figure 5.8: Product + Timely response vs complaints

CHAPTER 6

CONCLUSION

Based on the Prediction Matrix, the number of people who have been predicted to get an untimely response is 215. So we can conclude that 215 people are predicted to be churned. Thus we can use this model over different datasets to identify number of customers who are likely to churn. We can conclude that these people are churned due to untimely response from the company. Along with the results from the analysis we can postulate that these might have belonged to the bank product "Mortgage".

We can also conclude that the Adaptive boost model is a better fit for our data compared to the Decision Tree model. Therefore before choosing an appropriate model to run on the data it is important to compare all possible models and estimate their efficiency. This helps in choosing the best compatible model.In order to prevent future churning, the bank needs to provide special services to the churned customers in order to retain them. They can enhance their complaint service especially through the Web, as that's the most used channel based on the analysis done.

Customer lifetime value has intuitive appeal as a marketing concept, because in theory it represents exactly how much each customer is worth in monetary terms, and therefore exactly how much a marketing department should be willing to spend to acquire each customer, especially in direct response marketing.Lifetime value is typically used to judge the appropriateness of the costs of acquisition of a customer. Customer churn analysis is the stepping stone to preserving customer loyalty and meeting the standards of customer satisfaction.

CHAPTER 7

FUTURE ENHANCEMENT

Information Technology systems transformed virtually every single bank process. Today, banks have that rare opportunity to reinvent themselves again-with data and analytics. Banks can use Churn analytics to reduce risk and drive revenue. This project as covered various aspects of churn analysis.Other factors can be taken into account in the prediction of churn. Geographical and location based data can be used to study the regions of the country churning. This can give rise to new problem statements such as opening of new bank branches in popular locations. A heat map can be used to denote popular regions.

The complaint text can be analysed by performing a sentiment analysis. A database of complaints can be taken for this. Once the data is cleaned, stop words and other words of high frequency can be removed. The wordcloud package in R can be used for easy identification. Dictionaries of positive and negative words with a degree of sentiment can be defined. Different prediction models can also be applied along with or independent or along with each other. Various other modules of R can be explored. Churn analytics in the banking sector plays an important role and has a wide scope for further exploration.

REFERENCES

1. Aghion, Z. (2015). “Calculating churn", <https://www.quora.com/What-is-a-customer-churn-rate-and-how-can-it-affect-an-engine-of-growth>.

2. Analytics and banking (2017). “How advanced analytics are redefining banking".

3. Analyticstraining.com (2017). “Advantages and disadvantages of r".

4. Bmj.com. “Advantages of the roc curve".

5. Bmj.com. “Receiver operating characteristic curves".

6. CLV (2017). “Customer lifetime value", <http://www.customerlifetimevalue.co/>.

7. Gursoy, S. (2015). “Churn analysis of the telecom industry".

8. Han, J. and Kamber, M. (2011). “Decision tree algorithm".

9. Jaakkola, T. “Ada boost".

10. Kurt, W. (2014). “Churn modelling difficulties".

11. Mike. “Rcolorbrewer".

12. Nickgillian.com. “Ada boost".

13. Plot.ly. “Plotly package".

14. Rasel (2013). “Advantages of graphical representation of data, <http://www.businesscommunicationarticles.com/advantages-and-disadvantages-of-graphical-representation-of-data/> (July).

15. RDocumentation. “Wordcloud".

16. Schoonjans, F. (2017). “Roc curves", <https://www.medcalc.org/manual/roc-curves.php>.

17. shiny.rstudio.com. *Shiny in R"*, <https://shiny.rstudio.com/>.

18. Skok, D. (2012). “Negative churn".

19. Spinu, V. (2016). *Lubridate package*, <https://cran.r-project.org/web/packages/lubridate/lubridate.pdf>.

20. Srivastava, T. (2014). “Introduction to random forest - simplified | business case study".

21. Thinkapps.com (2017). “Customer churn: Why it should be the most important metric you measure and analyze".

22. University.custora.com (2017). “Customer lifetime value blog".

23. Walker, M. (2013). “Random forest algorithm".

24. Wang, Y. (2015). “Impact of churn on a company", <https://www.slideshare.net/GainsightHQ/the-incredibleimpactofchurnonthevalueofyourcompany> (March).

25. Wikipedia. “Customer attrition".

26. Wikipedia. “Retention rate".

27. Wikipedia. “Support vector machine".

28. Wikipedia (2017a). “Cluster analysis".

29. Wikipedia (2017b). “Confusion matrix".

30. Wikipedia (2017c). “Decision tree model".

31. Wikipedia (2017d). “Receiver operating characetristic".