# Generalized Churn Classification Across Multiple Business Domains

Author
Maryam Naveed
MS-18 (CE) 00000277111

Supervisor
Dr. Arslan Shaukat

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

JAN 2022

# Generalized Churn Classification Across Multiple Business Domains

Author

Maryam Naveed

MS-18 (CE) 00000277111

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor:

Dr. Arslan Shaukat

Thesis Supervisor's Signature:- ____________________

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

JAN 2022

# DECLARATION

I certify that this research work titled *"Generalized Churn Classification Across Multiple Business Domains"* is my work under the supervision of Dr. Arslan Shaukat. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been appropriately acknowledged/referred to.

Signature of Student<br>
Maryam Naveed<br>
MS-18 (CE) 00000277111

# LANGUAGE CORRECTNESS CERTIFICATE

This dissertation has been read by an expert in English and is completely free of typing, syntax, semantic, grammatical, and spelling mistakes. It is also as per the format given by the university.

Signature of Student
Maryam Naveed
MS-18 (CE) 00000277111

Signature of Supervisor
Dr. Arslan Shaukat

# COPYRIGHT STATEMENT

- Copyright in the text of this thesis rests with the student author. Copies (by any process) either in whole or of extracts may be made only according to instructions given by the author and lodged in the Library of NUST College of E & E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the author’s permission (in writing).
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary. It may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi

# ACKNOWLEDGEMENTS

All praises are to Allah. The favors He has bestowed on me have led to the successful completion of my research work. No hurdle was complex enough not to resolve, and I have eternal gratitude for that.

I sincerely thank my supervisor, Dr. Arslan Shaukat, for his extreme patience and expert guidance. He played an integral role in keeping my direction focused and ensuring I completed all my milestones promptly. I thank my GEC members, Dr. Usman Akram and Dr. Sajid Gul Khwaja, for filling in any gaps in my methodology and offering their kind help. I am lucky to have had the support of three of the soundest academic minds of the department.

This section is incomplete without mentioning the unwavering support of my family and friends. My parents' blind trust in my capabilities and constant prayers culminated in this achievement. My husband, Usman, with his continuous encouragement and brother, Hassaan, with his silent support, pushed me to step out of my comfort zone and perform even better than I had imagined. My best friend, Momina, was my emotional strength and helped out whenever I was technically blocked.

I am indebted to every one of you, along with every other individual who may have played even a small but significant role in helping me complete my research and degree. I am grateful for being allowed to leave a small mark in academia for generations to come.

*For Family.*

# ABSTRACT

For any organization, customers are the basis for company success, so Customer Relations Management (CRM) is an integral department. CRM research shows that it is more beneficial to retain customers, as it guarantees a higher return than it is to acquire new ones at five times the cost. For this purpose, organizations target minimal churning. Churning is defined as any customer ending a subscription or stop using a service being provided by an entity. Customer churn is happening across various business domains and has quite an impact on revenue generation. For companies to retain their essentials, they must be identified well in time. In the event of their identification, they are subjected to retention strategies. It is also much easier to target a specific group of customers than all of them to ensure retention when possible churning characteristics are identified. This makes churn identification and classification very important for the growth of a business.

This research aims to provide a generalized system that includes pre-processing and feature selection that can be utilized with different parts and business rules to identify customers on the verge of churning. A centralized hybrid algorithm has been devised to identify possible at-risk customers. We have addressed the gap created when a researcher has to rely on a hit and trial method to locate the best possible algorithm to solve their problem. Telecommunications data is widely available and has been made the benchmark to test the proposed methodology. We have used available datasets IBM Watson and Cell2Cell and a locally sourced dataset. Classifiers such as Support Vector Machines with RBF kernel, GP-AdaBoost, and Random Forest are used with SMOTE-ENN sampling, RFE feature selection, and normalization techniques. A potent combination of classification evaluation metrics is employed for thorough testing and 10-fold cross-validation for further support. Experiments have been performed with varying parameters and components. We can achieve a ground-breaking accuracy of 0.984 on IBM Watson and 0.994 on Cell2Cell. The locally sourced dataset has not been used in previous research. Hence, it was used as scoring data on which we have achieved an accuracy greater than 0.990. The results achieved on the two benchmark datasets using our proposed system are competitive compared to previous literature reports.

***Keywords: Customer Churn, SVM, Random Forest, GP-AdaBoost, SMOTE-ENN***

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1

# Introduction

# CHAPTER 1: INTRODUCTION

This chapter offers a brief introduction containing the background of the study, the problem statement, the motivation behind choosing this particular topic, the academic contribution of this research, and the detailed organization of this thesis.

## 1.1. BACKGROUND

Client churn is used to depict clients' affinity to stop working with an association in a legally binding structure [1]. Client Relation Management (CRM) is where studies on churn customarily begin [2]. It is essential to forestall client churn when offering types of assistance. Associations have seen a proficient client securing to churn previously. Notwithstanding, as the market became immersed because administrations were globalized and the rivalry became savage, client securing costs quickly rose [3], [4].

Reinartz et al. showed that, for long haul business activities, expansion in the degree of consistency of clients as far as CRM is less expensive than focusing on few clients for securing [5]. Likewise, Sasser et al. recommended that held by and large clients produce better yields than an irregular focusing of new clients [6]. Furthermore, Mozer et al. suggested that advertising lobbies for having existing clients are more effective than investing energy to draw in new clients, as far as the net profit from speculation [7]. Reichheld et al. showed that a 5% increment in client consistency standard brought about 35% and 95% expansions in the net present worth produced by clients for a product house and a publicizing organization, separately [8]. Like this, churn expectation can build the consistency standard of faithful clients and at last increment the organization's worth.

Studies on churning customers have been put forth in various fields. Multiple indicators have been used to identify well in time the propensity of a customer to churn. The churn rate of customers [9] is a general indicator of analysis in customer churn. This points to the percentage of users who cancel service to the total number of subscribers during a specific period [9], [10], [11]. The churn rate is the most customarily involved KPI for working out the maintenance time of administration of supporters in many business fields. Due to its importance, churn has been presented in different areas and grown further to suit unmitigated attributes. Subsequently, the exploration on client churn investigation was divided by each examination field; in this way, the estimation standards are altogether unique.

Already, client churn was utilized to characterize the client's status. CRM is an administration technique for organizations that were first acquainted with increment effectiveness in retail, advertising, deals, client assistance, and production network and increment proficiency and the association's client esteem capacities. From that point forward, according to a structural perspective, the development of CRM has isolated it into CRM of Operations and CRM of Analysis. The CRM, in light of investigation, is fixated on making information bases and assets containing client social ascribes [12], [13], [14]. At first, the logical CRM was used to make proper promoting techniques utilizing client status and client conduct information. Specifically, it has been used to satisfy the individual and great necessities of clients [15]. Organizations began applying innovations that were explicit for procurement, maintenance, churn, and determination of clients [16], and when advances of the IT field began being carried out in the client relationship the board, different organizations started to involve these advances in regions including warehousing of information, sites, correspondence, and financing [17]. As recently referenced, with CRM studies guaranteeing that even a minor consistency standard increment of existing clients demonstrates more proficiency than the obtaining of new clients, churn examination has become one of the fundamental customized client executives procedures [17], [18], [19]. Papers and studies gathered and summed up examination of churning methods in the different fields [20], [21], [22]. The information utilized to investigate churn does exclude elements of time series, endurance and maintenance, and KPI highlights. Papers additionally exist for administrations using different profound learning models given churn investigation strategies in software engineering [23], [24]. These investigations, in any case, are restricted to deep learning calculation, and they need boundary depictions. A couple of review papers on client churn exist yet don't cover the latest procedures; however, churn in explicit modern fields [25]. The inclination of building churn forecast models is moving, and execution is quickly refining.

### 1.1.1. DEFINITION OF CHURN

Client churn is when clients of a business never again buy or connect with the organization. A high churn implies that more clients need to purchase labor and products. The pace of client churn is the numerical computation of the level of purchasers who cannot keep utilizing services from a business.

Client churn happens when clients choose not to keep buying items/ services from an association and end their affiliation. It is an essential boundary for the association since procuring another client could cost almost five times more than holding clients an association now has. For the outstanding development of an association, the churning of clients can cause a genuine issue, and a procedure to keep ought to be chosen to avoid an increment in client churn rates.

In the most straightforward structure, the client churn rate is the number of lost clients ratioed by the general complete number of clients. Division of clients should be possible in light of the recurrence of their buys to have a guess.

$$\text{Customer Churn Rate} = \frac{\text{Customers Lost}}{\text{Total Customers}} \text{ x } 100 \tag{1.1}$$

### 1.1.2. CUSTOMER RELATIONS MANAGEMENT

Client churn is a broadly utilized proportion of clients that are lost. Organizations frequently lose significant clients and, hence, incomes to the opposition. Organizations have undergone massive changes throughout the thirty years, like new administrations, innovative progressions, and expanded rivalry because of liberation [26]. Client churn expectation has, hence, become crucial for industry players to ensure a loyal client base, development of the association, and client relationship the executives (CRM) improvement [27], [28]. Retention of customers with a high risk of churning is one of today’s toughest challenges [29]. Maintenance of clients with an increased risk of churning is perhaps the most challenging test [29]. Because of a more massive number of specialist co-ops and more exceptional rivalry, clients today have different choices to churn. Subsequently, the top players are awakening to the significance of holding existing clients instead of obtaining new ones [27].

Numerous factors sway customers to change organizations. Dissimilar to post-paid customers, prepay any organization arrangements do not restrict customers, and they often blend for the shallowest reasons. In like manner, it isn't tricky to expect their churn rate. Another variable is enduring customer ness, which isn't wholly settled forever by customer help and thing quality. Issues like association consideration and get-together quality may affect customers to move to competitors with a broader reach and better assembling quality. Various components that increase the probability of customers fleeing resistance consolidate drowsy or deficient responses to grumblings and charging botches. Factors like packaging costs, harming features, and obsolete advancement might cause customers to be attracted to the resistance. Providers are as often as possible differentiated, and customers churn to whoever they feel has better as a rule worth [30].

An association can do okay by zeroing in on existing customers, whether or not it suggests acquiring no new customers. It is surveyed that the cost for convincing a conventional customer not to churn to the competitor is, on various occasions, not the expense of looking and incorporating into contact with another customer. The cost of attracting new customers is multiple times more than holding existing ones. Another investigation checks that an expert center can fabricate 25 to 85 percent benefits by diminishing customer churn rate by 5%. This shows the overall impact customer churn rate can have on business achievement.

Mechanical types of progress have helped associations with the arrangement that their merciless strategies should ensure high customer levels of consistency to get by in the business [31]. This especially applies to the media transmission industry. Thus, enormous investigation activity is presently [32]. The importance of regulating customer churn is moreover implied by various examiners who consider it imperative for CRM [33]. CRM requires the relationship to know and appreciate its business areas and customers. CRM incorporates learning the customer's display to hold the most gainful customers and separate those whose churn no longer has any impact. CRM also plans the progression of the arrangement and cutoff points: which thing to propose to which customers through which medium and which thing needs advancements.

### 1.1.3. DIFFICULTIES IN CHURN IDENTIFICATION

The capacity to foresee that a specific client is at extremely great danger of churning. Churn addresses a tremendous added potential income-producing hotspot for any business, even though there is still some ideal opportunity to accomplish something significant with regards to it.

- Drawing in new clients is expensive; however, losing clients an organization has will cost much more as existing clients usually return and rehash buys from the brand whenever fulfilled.
- Contest in any market is on the ascent, which urges organizations to zero in on gaining new clients and holding clients they now have.
- The fundamental stage towards anticipating client churn is compensating existing clients for consistent buys and backing.
- As recently referenced, a client's aim to quit utilizing a specific item/administration is intrinsically a choice framed over an extensive stretch. Different variables lead to this choice. Associations need to see every component for clients to be held and proceed with their buys.

### 1.1.4. ADVANTAGES OF CHURN IDENTIFICATION

Below are some of the advantages of identifying customer churn.

- Acquiring Information for Improvement: Dissatisfied clients are a wellspring of helpful criticism for an association's advancement. An association will receive data about angles that need improvement while executing techniques to forestall client churn.
- Diminishing Risk of Business: Customer churn straightforwardly interprets misfortune to the business. Selling another item/administration to a client presents to be a lot more straightforward than offering it to an entirely new likely client. Accordingly, to the development of the business, client churn can be destructive.
- Understanding the Target Market: Constantly running after lessening client churn will uncover layers of the market that were generally obscure. Different studies and other such roads can be led to get what the objective market needs/needs viably and assist with lessening the churning of the client.
- Building a Competitive Advantage in the Market: In a world with a steady contest to achieve new clients and hold existing ones, having the edge over the opposition. In lessening client churn, the organizations empower clients to know the lesser-known parts of an organization and structure advantage over the other comparative specialist co-ops on the lookout.

## 1.2. PROBLEM STATEMENT

The maintenance and obtaining of clients are critical worries of any business industry. The rapid development of the commercial center in each business leads to an expanded supporter base. Likewise, organizations have perceived the meaning of holding a client on hand. It has become essential for specialist co-ops to decrease the churn rate of clients since neglect may contrarily impact the organization's benefit. Churn forecast adds to recognizing those clients who will probably switch an organization over another. The identification of these clients is of the most extreme criticality.

Accordingly, the current study employs a machine learning pipeline to facilitate businesses with efficient approaches for lessening churn rate by timely identification. In identification, companies can target said customers for retention through various incentives. This saves the business marketing and customer acquisition costs and aids in generating profit.

## 1.3. MOTIVATION

First and foremost, this problem was seen as a viable solution that should be actively employed to improve the country's businesses. We are a developing nation and constantly need innovative ways to improve our national assets and financial worth. Upgrading a country's companies of any magnitude casts a direct positive reflection on the economy.

Secondly, having briefly worked in the telecommunications sector and is currently employed as a data engineer in an IT firm instilled confidence in the skill set I have acquired. It has enabled me to add relevant significance to the research for identifying and proposing a solution for predicting churn.

## 1.4. RESEARCH CONTRIBUTION

### 1.4.1. OBJECTIVES

The study followed the below objectives:

1. To understand and facilitate customer churn prediction in various businesses
2. To create a generalized algorithm capable of identifying churn
3. To achieve results superior to previously reported
4. To compare various parameters used in churn prediction to identify the most effective ones

### 1.4.2. RESEARCH QUESTIONS

1. What is the role of customer churn prediction in various businesses?
2. Can a generalized algorithm be devised to facilitate multiple business domains?
3. Which pipeline components are most effective in reducing business churn rate?

### 1.4.3. SIGNIFICANCE OF STUDY

Improved services ensure a company's economic growth, enabling new markets to be established and wealth to be generated. This causes a ripple effect that prompts job creation and increased national cash flow as living standards are enhanced. Better earnings by the people ensure a higher federal income in higher tax returns.

It is significantly more expensive to gain a new customer than retain an existing one. Statistically, companies gain revenue by identifying problem areas that cause churning by developing strategies that boost customer loyalty and relations. Churn prediction is an excellent way to identify all the problem areas in different demographics and services. Companies can devise better retention policies, which play a massive role in increasing profits and asset value. We aim to help companies identify which customers are likely to churn to identify key features that may be lacking in each profile.

Any contractual or non-contractual service provider falls under the scope of this study. Some examples are the telecommunications industry, banking sector, online video/music streaming services, social media premium subscriptions, etc.

## 1.5. PROPOSED METHODOLOGY

In this sub-category, pictorial representation of the system architecture for a generic machine learning pipeline is shown in Figure 1.1, which includes various phases, namely, data loading cleaning and pre-processing, feature selection, data sampling, and normalization, splitting of pre-processed data into a training and testing set, classifier application for model training and testing with evaluation and, finally, 10-cross validation for evaluation support. We have broken our methodology into the above phases to fit them into a machine learning pipeline.

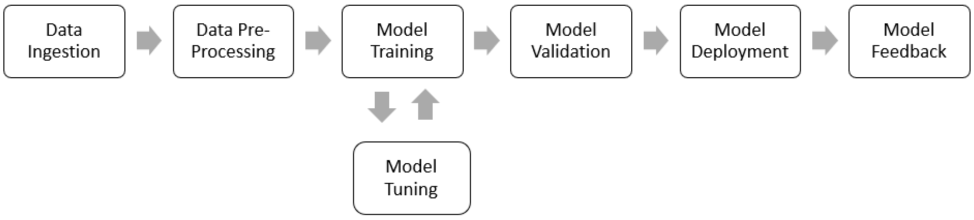


Figure 1.1: Generic Machine Learning Pipeline

Phase 1 will comprise data loading. Files will be loaded to a server. Data will be cleaned entirely by removing redundant columns and managing missing values. Phase 2 will comprise feature selection only to keep columns with maximum usability. Data will be normalized and sampled. Phase 2 will also include splitting the data into a training and testing set. Phase 3 will be model training using various algorithms to find which work most efficiently. These models will be evaluated against the testing sets to determine performance. Phase 4 will comprise further evaluation support by testing the complete data with 10-fold cross-validation. Details of which components were finally chosen for each phase after extensive experimentation are given below in Figure 1.2.

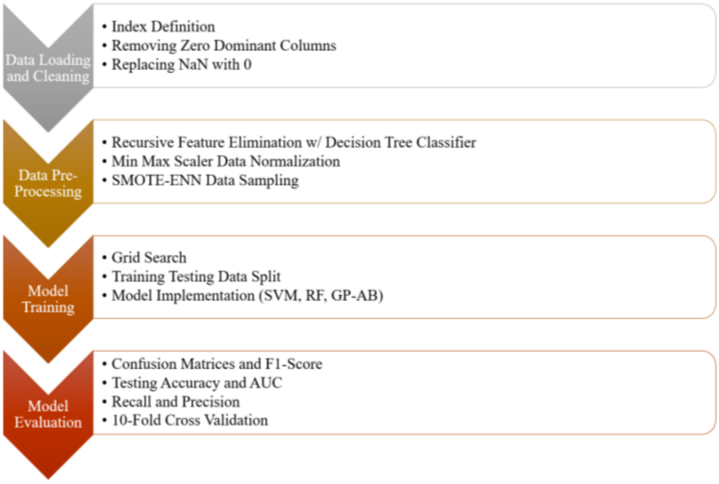

Data Loading and Cleaning
• Index Definition
• Removing Zero Dominant Columns
• Replacing NaN with 0
Data Pre-Processing
• Recursive Feature Elimination w/ Decision Tree Classifier
• Min Max Scaler Data Normalization
• SMOTE-ENN Data Sampling
Model Training
• Grid Search
• Training Testing Data Split
• Model Implementation (SVM, RF, GP-AB)
Model Evaluation
• Confusion Matrices and F1-Score
• Testing Accuracy and AUC
• Recall and Precision
• 10-Fold Cross Validation

Figure 1.2: Detailed Pipeline for Study

## 1.6. THESIS ORGANIZATION

The organization of the thesis is represented in Figure 1.3.

**CHAPTER 1: INTRODUCTION** offers a brief introduction containing the background study, problem statement, motivation, research contribution, and thesis organization. **CHAPTER 2: LITERATURE REVIEW** provides a detailed review highlighting the work done in the domain of Churn Prediction using various methodologies. The systematic literature review is composed of two main sections. The first is broken down into four sub-categories. The first is a review of all the work to understand churn, including detailed surveys and insights into CRM. The second offers details on churn classification and variant methodologies. The third is specific to the reported results of the chosen benchmark datasets. At the same time, the last category highlights the research gaps encountered. We then present a summary featuring the critical points of the review. **CHAPTER 3: EXPERIMENTAL METHODOLOGY** covers the details of the datasets, chosen classifiers, evaluation techniques, and implementation. **CHAPTER 4: RESULTS AND DISCUSSION** presents detailed findings after running our machine learning pipeline for various parameters and combinations of methods. We also give a comparative study to highlight the significance of our proposed methodology. **CHAPTER 5: CONCLUSION** concludes the research recommends ideas for future research, and highlights current limitations.

Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Experimental Methodology
Chapter 4: Results and Discussion
Chapter 5: Conclusion

Figure 1.3: Thesis Organization

# Chapter 2

# Literature Review

# CHAPTER 2: LITERATURE REVIEW

This chapter provides a detailed review highlighting the work done in the domain of Churn Prediction using various methodologies. The systematic literature review is composed of two main sections. The first is broken down into four sub-categories. The first is a review of all the work being carried out to understand churn, including detailed surveys and insights into CRM. The second offers details on churn classification and variant methodologies. The third is specific to the reported results of the chosen benchmark datasets. At the same time, the last category highlights the research gaps encountered. We then present a summary featuring the critical points of the review.

## 2.1. INTRODUCTION

Extensive research has been carried out across various industries for churn analysis. Researchers have used many machine learning algorithms and statistical techniques to predict churn. Neural Networks, Regression algorithms, Game Theory, Clustering, Genetic algorithms, Role Generation, Ensembles, and Hybrids have all been used to solve churn in specific business domains. Recent studies have highlighted all the algorithms used in different business domains and their particulars to help better researchers identify which best suits their needs without trial and error. There are two kinds of papers available on churn prediction. Documents that help determine which factors/features of a service cause churn or essays that feed certain features to a classification algorithm to identify churning/churned customers. While research has been plentiful, we are yet to see a generalized algorithm that can be used across different platforms and business domains to help identify churn.

### 2.1.1. UNDERSTANDING CHURN

Client churn is fundamental since it's costlier to gain another client than to offer more to a current customer. This metric can represent the deciding moment of a business: If you make an excellent showing keeping clients around, you should see your average client lifetime esteem increment, making each future deal significantly more critical and eventually further developing your unit edges.

The best utilization of organization assets is frequently to increment repeating membership income or dependable recurrent business instead of putting more in new client obtaining. Save steadfast clients for a long time, and you'll have a lot shorter time developing and enduring unpleasant monetary patches as opposed to spending to acquire new clients to supplant the people who left.

Churn rates should be evaluated in setting: 33% we referenced might be by the business average, or it could be vastly improved or much more terrible. On the off chance that yours is an upward industry where churn information is accessible freely or through an industry exchange association, it's worth benchmarking your outcomes against rivals.

High-performing organizations frequently set an underneath industry-normal objective churn level; track their measurements over the long run, maybe utilizing a KPI scorecard; and act immediately to address blips before they become negative patterns.

The main element to accomplishment in this plan of action is client maintenance. Clients obtaining help that is membership-based is the moderately simple aspect. Holding the clients can be profoundly challenging. Luckily, membership-based administrations can utilize information to develop their churn rate further.

### 2.1.2. CHURN CLASSIFICATION

There are four areas to construct a beat expectation model: conventional ML, statistics, graph theory, and deep learning.

There is exceptionally obscured limit between the four disciplines referenced previously. In any case, Breiman et al. separated insights and ML into an algorithmic model, impacting Matthew Stewart's way and Bzdok et al. partitioned insights and AI [34], [35], [36]. Witten et al. depicted that ML procedures had created with mining of information since the start of PCs while focal point of measurements was on theory tests in view of arithmetic [37]. In measurements, models of probablity have been predominantly utilized for agitate forecasts. Especially, likelihood models have been utilized, generally, for examination that depends on clients [38]. In research that depends on clients, stir rate is applied to the endurance time assessment while working out the Customer Lifetime Value (CLV). Expectation calculations of CLV are joined with working out the normal income produced by a client a model for stir rate forecast. A moved Beta-Geometric (sBG) model computes CLV inside the legally binding settings. The beta circulation is utilized by the sBG model to make shifts for changes in time t to fit the pace of maintenance [39].

In settings that are non-authoritative, the recurrent conduct of client buys has been recently communicated through bad binomial disseminations (NBD) [40]. Further, the conveyance of agitate utilized a gamma combination of remarkable, otherwise called the Pareto (of the subsequent kind) dispersion. The CLV can be determined by consolidating the purchaser conduct and endurance dispersion. This technique is alluded to as the NBD model. The NBD technique has been effectively utilized as a likelihood model for determining the CLV as of not long ago [41], [42], [43]. The beta-mathematical/beta-binomial (BG/BB) model is another technique [44]. In settings that are non-legally binding, agitating clients will generally have ceaseless probabilistic attributes. To decide client agitate in the non-legally binding settings, scientists involved time series procedures for altering like gathering client id to clean up information or ascertain social fluctuations [45]. Scientists start anticipating client stir utilizing the highlights delivered from this handled information. The model of measurements utilized here depends on factual derivation and speculation testing. ML methods were additionally utilized in client stir for all settings. In contrast with factual strategies, ML procedures have powerful connections that are non-straight between highlights. The diagram hypothesis characterizes agitate as a relationship of arithmetic. When a diagram is assembled, it looks for beating clients through relationship investigation. Profound learning procedures have as of late arisen to foresee agitate in clients. Profound learning is an extreme augmentation of AI with neural organization calculations.

The deep learning model is a moderately ongoing investigation technique for foreseeing stir. As per Goodfellow et al., profound learning is important for AI [46]. In any case, in light of the fact that its scholarly importance has as of late developed, it has set up a good foundation for itself as a solitary scholastic field. This is valid for building beat forecast investigation models. Lee et al. uncovered that a profound learning model anticipated client stir with a higher likelihood than a conventional AI model for game beat forecast examination [23]. They summed up the highlights utilizing the retention and speculation strategies depicted in include adjustment and expanded stir consistency by joining profound learning and conventional AI models. Zhang et al. analyzed the standard AI model and the profound learning model in client beat expectation issues in the protection business. Zhang et al. characterized highlights to be applied to the profound learning model, handled them, and joined the outcomes. The review looked at the profound learning-based stir expectation technique they created with the conventional AI based beat forecast calculation. The Deep and Shallow model they constructed showed fantastic beat forecast execution contrasted with different models. In this review, albeit profound learning is important for AI, it is utilized as an advancement calculation for agitate forecast issues.

Table 2.1 shows a rundown of strategies that are grouped by business field. We affirmed that favored displaying procedures were different relying upon the business field. Organizations with thick log information and simple admittance to client data, like the games and broadcast communications ventures, are applying somewhat many profound learning methods utilizing huge information, which is a quick pattern. With respect to the monetary and protection areas, since the log information is generally little and the data got from clients doesn't change essentially, numerous factual methodologies utilize conventional AI models or endurance examination.

| Business Field | Traditional Machine Learning | Statistics | Graph Theory | Deep Learning |
|---|---|---|---|---|
| Game | 5 | 2 | | 2 |
| Finance | 10 | 2 | | 1 |
| Insurance | 3 | | | 1 |
| Marketing | 1 | | | |
| Newspapers | 3 | | | |
| Music | 1 | | | |
| Internet | 3 | 1 | | 1 |
| Psychology | 1 | | | |
| Energy | 1 | | | |
| Human Resources | 1 | | | |
| Telecom | 26 | 3 | 3 | 1 |
| Retail | 2 | | | |

Table 2.1: Churn Prediction Models in Various Business Fields

### 2.1.3. BENCHMARK DATASETS

We obtained two benchmark datasets, IBM Watson and Cell2Cell, and concocted our tests to contrast results from both and our discoveries. The proposed proactive model [48] recognized the characteristics that exceptionally affected the churning of clients with the assistance of Machine Learning (ML) calculations like K Nearest Neighbors, Random Forest, and XG Boost. IBM Watson was utilized, and it was observed that XG Boost played out fantastic, with Fiber Optic clients having a higher affinity for churning. The concentrate likewise proposed investigating mixture methods with more profundity. [52] again, observed XG-Boost be the best performing calculation. One more review  [49] checked out Cell2Cell and IBM Watson and observed that SVM played out awesome for both datasets. The creators proposed involving records for other client reaction models, like strategically pitching, up-selling, or client obtaining.

It was noted from a review [50] that precision is undoubtedly not a reliable measure for deciding model execution while examining imbalanced/uneven datasets. Moreover, the Logistic Regression model was displayed to have a superior responsiveness rate, implying that it seldom misclassified the minority class. The inclination is bothersome in any forecast model; simultaneously, it is inevitable. The churn expectation procedure and the information calculations and information examination have a significant impact in the current computerized period as the information is assembled out from the different AI calculations. The information investigation covers a vast degree, assisting us with being familiar with the various variables influencing the churning rate. Additionally, it can draw out the prescient limit of the client's mentality empowering the different telecom specialist organizations to change the new plans conceivable. Further, it leaves us with the other imaginative uses of AI calculations and the information examination way to deal with addressing the current difficulties confronting society, likewise the different requirements for information investigators. The calculations have very much envisioned the significance of information; this additionally opens out another passage for the information investigators and the use of different inventive spellbinding, proactive, and prescriptive calculations conceivable.. [51]

The churn issue was broken down utilizing a comprehensive arrangement of characterization methods. Specifically, the Tsallis and Renyi entropy measures were applied to choose trees and contrasted with other arrangement techniques. An old-style choice tree calculation C4.5 was altered by fusing parametrized α entropies, expanding grouping prospects. That considered the viability of the trees as a component of the entropy boundary. [53] Another review utilized strategic relapse and KNN with enormous information for foreseeing shopper churn in the telecom area. [54] Logistic regression has been used broadly to gauge the likelihood of churn as a component of client factors set or highlights. Also, for churn, K-Nearest Neighbor is used to inspect whether a client churns or not given their element's vicinity to clients in each class. This study involved IBM Watson in anticipating and breaking down churn. The review results showed that the precision pace of forecast in customer churn is viewed as 0.80 percent, and the region under the bend is viewed as 0.71 percent.

The GP-AdaBoost estimation in blend in with PSO is presented as a shake figure approach (Ch-GPAB) for telecom [55]. PSO-based methodology under-models more significant part class events. It encourages a fair arrangement set that loosens up better sorting out how than the proposed shake marker, achieving further developed assumption execution. Ch-GPAB approach progresses various GP programs per class using the AdaBoost style supporting system, building up its organization capacities. Each GP program probably goes as a lone class classifier. A higher weighted measure of GP programs finishes up the last figure of a test event, expanding more specific results. Consequently, the proposed Ch-GPAB beat assumption approach is worthwhile for current use.

SVM, Logistic Regression, and Random Forest performed better compared to Decision Tree for client churn examination for IBM Watson in a specific report. [56] These discoveries were validated by one more arrangement of scientists who additionally chipped away at tracking down factors that superior client lifetime esteem. [57]

Another examination introduced a model that can anticipate which clients will leave the association and who will not. To do this, the creators utilized various information mining strategies. [58] As for the bunching strategy has demonstrated that DB examines grouping is reasonable and compelling in bunching the informational index into two classifications and giving a high level of non-churners over churners at the group 0. An original copy zeroed in on the churn expectation part of the media transmission industry by proposing an Artificial Bee Colony (ABC) based model for preparing ANN. [59]

A clever TL-DeepE method is proposed to anticipate possible churners, vital for the cutthroat telecom industry. The size and dimensionality of the telecom information are enormous and require substantial computational power for churn forecasts. The proposed technique can be possibly viable intending to the worries and difficulties of the business. [60]

Analysts have utilized SMOTE and Deep Belief Network (DBN) against the two expense delicate learning techniques: central misfortune and weighted misfortune in churn expectation issues. The exact outcomes show that the churn forecast issue's generally prescient execution of major trouble and weighted misfortune techniques is superior to SMOTE and DBN. [61]

Various explores have demonstrated that neglecting will forever exist. In any case, it is probably going to reduce the likelihood of devastating forgetting by the long-lasting learning idea. An analysis demonstrated that deep rooted learning offers a more adaptable approach to figuring out how to additional exploration in unique instruction. [62]

Table 2.2 summarizes the best results produced with the benchmark datasets and the classifier and evaluation metric used.

| Paper | Year | Dataset | Algorithm | Technique |
|---|---|---|---|---|
| Pamina, et al. [48] | 2019 | IBM Watson | XG-Boost | Basic Pipeline |
| **Ebrah, et al. [49]** | **2019** | **IBM Watson** | **Support Vector Machine** | **Basic Pipeline** |
| | | **Cell2Cell** | **Support Vector Machine** | **Basic Pipeline** |
| **Tamuka, et al. [50]** | **2020** | **IBM Watson** | **Logistic Regression** | **Feature Selection** |
| Apurvasree, et al. [51] | 2019 | IBM Watson | Support Vector Machine | Basic Pipeline |
| Parmar, et al. [52] | 2021 | IBM Watson | XG-Boost | Basic Pipeline |
| Gajowniczek, et al. [53] | 2016 | Cell2Cell | Decision Tree with Shannon Entropy | Basic Pipeline |
| Joolfoo, et al. [54] | 2020 | IBM Watson | Logistic Regression | Basic Pipeline |
| Idris, et al. [55] | 2019 | Cell2Cell | GP-AdaBoost | PSO Under-Sampling |
| Sundararajan, et al. [56] | 2020 | IBM Watson | Random Forest | Basic Pipeline |
| Townsend, et al. [57] | 2019 | IBM Watson | XG-Boost | Feature Selection |
| Mitkees, et al. [58] | 2017 | IBM Watson | Multilayer Perceptron | Basic Pipeline |
| Paliwal, et al. [59] | 2017 | IBM Watson | Artificial Bee Colony Neural Network | Basic Pipeline |
| Ahmed, et al. [60] | 2019 | Cell2Cell | Transfer Learning Neural Network | Basic Pipeline |
| Nguyen, et al. [61] | 2021 | Cell2Cell | Focal Loss with XG-Boost | Over-Sampling w/ SMOTE |
| Sarnen, et al. [62] | 2020 | IBM Watson | Elastic Weight Consolidation | Basic Pipeline |

Table 2.2: Summary of the Best Benchmark Results

A primary pipeline comprises data cleaning, model training, and evaluation for papers that employ a direct channel. The Technique column states that for documents that employed a different technique, mention just that. These techniques are used in addition to all the components of a primary pipeline.

## 2.2. SUMMARY

Churn Analysis is a widely used analysis on Subscription Oriented Industries to analyze customer behaviors to forecast the customers who are likely to churn from the company's service agreement. The study is based on classification methods and algorithms. It has become so crucial for companies in today's commercial conditions that gaining a new customer is higher than retaining an existing customer. The paper reviews the relevant studies on Customer Churn Analysis in various industries in literature to present information pertinent to aspiring researchers about the often-used data mining methods, results, and performances of said methods and shedding light on studies that can be done going forward.

Customer churn prediction is a binary classification problem. Still, the high data dimensionality and a usually small number of minority classes in the telecom datasets make it a big hurdle for conventional classifiers to show desired performance. The scale of the problem justifies the need for its accurate identification and proposes some retention activities in advance. Researchers emphasize that an essential role in the whole process depends on the technique used, data type, and quality.

In literature, most churn studies use data mining or statistical methods to predict churn probabilities. The standard approach is to use a set of techniques, e.g., decision trees, logistic regression, or neural networks, and compare their predictive power to determine the best method for churn classification. Some studies were focused on support vector machines performance for churn detection, e.g., how effectively support vector machines can detect churn in comparison to back-propagation neural networks for predicting on a data set from a credit card company; comparison analysis between one-class support vector machines, neural networks, decision trees, and Naive Bayes.

Some demonstrated neural network and decision trees for customer insolvency (involuntary churn) in cellular telecommunications, and the results proved that neural network models are more stable than decision trees. Some used logistic regression and decision tree model and focused on binomial logistic regression model for churn prediction and identified customer dissatisfaction, service usage, switching cost and demographic variable affects customer churn. Others evaluated their churn prediction technique based on multi classifier class-combined approach that predicts churning from contractual subscriber information and call pattern changes.

In general, many practical applications use supervised learning techniques, such as logit and probit, which extend classical regression methods explicitly adopted for classification. Some have used decision trees, which are graphical decision-support methods used in decision theory. Others, including having successfully used artificial neural networks. From this perspective, a comprehensive study of different approaches seems reasonable to profoundly analyze the underlying problem and draw general or specific conclusions on the topic.

# Chapter 3

# Experimental Methodology

# CHAPTER 3: EXPERIMENTAL METHODOLOGY

This chapter covers the details of the benchmark and scoring datasets, the components that were chosen as part of the machine learning pipeline, the employed evaluation techniques and complete pipeline implementation.

## 3.1. DATASETS

### 3.1.1. SOURCES

Three data sets have been sourced to be used in this study. IBM Watson Analytics provided a constructed dataset that mirrored the data of a telecommunications company. Cell2Cell is actual data of customers belonging to a telecommunications company. Data from a local Pakistani business was also sourced as a means of providing support to the constructed pipeline.

Details of the datasets are summarized in Table 3.1.

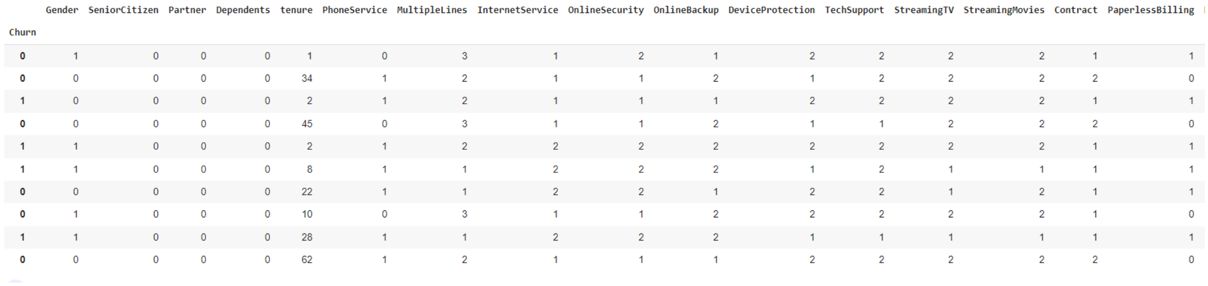
| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| **Samples** | 7043 | 51047 | 19000 |
| **Features** | 19 | 55 | 898 |
| **Non-Churn Class** | 5174 | 36336 | 10928 |
| **Churn Class** | 1869 | 14711 | 8072 |
| **Churn Class %** | 26.5% | 28.8% | 42.5% |
| **Data Source** | IBM Watson Analytics | Cell2Cell | Local Business |

Table 3.1: Dataset Specifics

### 3.1.2. DESCRIPTION OF ATTRIBUTES

#### *3.1.2.1. IBM WATSON*

The data set has been acquired from IBM Watson Analytics. The data set has 7,043 ocurrances and 19 features. The dataset is composed of 4 numerical and 15 nominal features. It has 1,869 records of the minority class and 5,174 records of the majority class that amounts to 26.5% share of the minority class in the entire dataset. A partial snapshot of the data is shown in Figure 3.1.



| | Gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract | PaperlessBilling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Churn | | | | | | | | | | | | | | | | |
| **0** | 1 | 0 | 0 | 0 | 1 | 0 | 3 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 |
| **0** | 0 | 0 | 0 | 0 | 34 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 0 |
| **1** | 0 | 0 | 0 | 0 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 |
| **0** | 0 | 0 | 0 | 0 | 45 | 0 | 3 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 0 |
| **1** | 1 | 0 | 0 | 0 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| **1** | 1 | 0 | 0 | 0 | 8 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |
| **0** | 0 | 0 | 0 | 0 | 22 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 |
| **0** | 1 | 0 | 0 | 0 | 10 | 0 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 0 |
| **1** | 1 | 0 | 0 | 0 | 28 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| **0** | 0 | 0 | 0 | 0 | 62 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 |

Figure 3.1: Partial Snapshot of IBM Watson

The dataset comprises of the below segments

1. Churn column is a flag that indicates customers who left.
2. All services that the customer has subscribed to.
3. Customer accounts and revenue generation
4. Customer Demographics

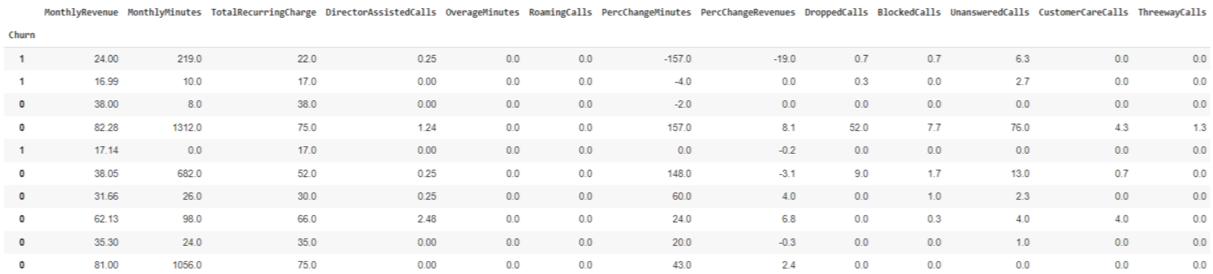
Table 3.2 presents some of the key features in the dataset along with possible values.

| Attribute | Definition | Possible Values |
|---|---|---|
| **Churn** | Customer happened to churn or not | Yes or No |
| **Tenure** | Months the user availed the services offered | Integer Values |
| **Senior Citizen** | User is a senior citizen or not | 1, 0 |
| **Phone Service** | User has a phone service or not | Yes or No |
| **Multiple Lines** | User has multiple lines | Yes or No |
| **Online Security** | User has online security | Yes or No |
| **Online Backup** | User has an online backup | Yes or No |
| **Device Protection** | User has device protection | Yes or No |
| **Payment Method** | User payment method | Electronic check or Mailed check |
| **Tech Support** | User has tech support or not | Yes or No |
| **Streaming TV** | User has streaming TV or not | Yes or No |
| **Streaming Movies** | User has streaming movies or not | Yes or No |
| **Contract** | User term of contract | One year or Two years |
| **Monthly Charges** | User Monthly Charge | Integer Values |
| **Paperless Billing** | User utilizes paperless billing or not | Yes or No |

Table 3.2: IBM Watson Features

#### *3.1.2.2. CELL2CELL*

The sixth largest company in the US for wireless telecommunication, Cell2Cell dataset comprises 51,047 samples suggesting whether the customer happened to churn two months after survey and 55 features. The dataset comprises 33 numerical and 22 nominal features. It has 14711 rows of the minority class and 36336 records of the majority class that amounts to 28.8% share of the minority class in the entire set of data. A partial snapshot of the data is shown in Figure 3.2.



| | MonthlyRevenue | MonthlyMinutes | TotalRecurringCharge | DirectorAssistedCalls | OverageMinutes | RoamingCalls | PercChangeMinutes | PercChangeRevenues | DroppedCalls | BlockedCalls | UnansweredCalls | CustomerCareCalls | ThreewayCalls |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Churn | | | | | | | | | | | | | |
| 1 | 24.00 | 219.0 | 22.0 | 0.25 | 0.0 | 0.0 | -157.0 | -19.0 | 0.7 | 0.7 | 6.3 | 0.0 | 0.0 |
| 1 | 16.99 | 10.0 | 17.0 | 0.00 | 0.0 | 0.0 | -4.0 | 0.0 | 0.3 | 0.0 | 2.7 | 0.0 | 0.0 |
| 0 | 38.00 | 8.0 | 38.0 | 0.00 | 0.0 | 0.0 | -2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0 | 82.28 | 1312.0 | 75.0 | 1.24 | 0.0 | 0.0 | 157.0 | 8.1 | 52.0 | 7.7 | 76.0 | 4.3 | 1.3 |
| 1 | 17.14 | 0.0 | 17.0 | 0.00 | 0.0 | 0.0 | 0.0 | -0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0 | 38.05 | 682.0 | 52.0 | 0.25 | 0.0 | 0.0 | 148.0 | -3.1 | 9.0 | 1.7 | 13.0 | 0.7 | 0.0 |
| 0 | 31.66 | 26.0 | 30.0 | 0.25 | 0.0 | 0.0 | 60.0 | 4.0 | 0.0 | 1.0 | 2.3 | 0.0 | 0.0 |
| 0 | 62.13 | 98.0 | 66.0 | 2.48 | 0.0 | 0.0 | 24.0 | 6.8 | 0.0 | 0.3 | 4.0 | 4.0 | 0.0 |
| 0 | 35.30 | 24.0 | 35.0 | 0.00 | 0.0 | 0.0 | 20.0 | -0.3 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 0 | 81.00 | 1056.0 | 75.0 | 0.00 | 0.0 | 0.0 | 43.0 | 2.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Figure 3.2: Partial Snapshot of Cell2Cell

Some notable observations on Cell2Cell are:

1. Customers that churn have an average of less than 530 minutes.
2. The number of months they had service was for only 11 - 15 months.
3. The numbers of days their equipment was in use were between 300 & 361 day.
4. The numbers of phone models issued were less than 2.
5. Their prizm code referred mostly to town.
6. Their handsets had the capability to connect to the internet.

Table 3.3 presents some of the key features in the dataset along with possible values.

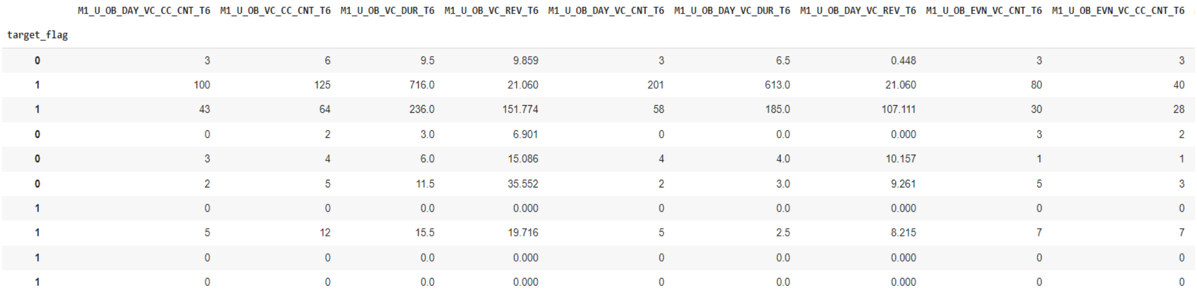
| Attribute | Definition | Possible Values |
|---|---|---|
| **Churn** | Customer happened to churn or not | Integer Values |
| **MonthlyRevenue** | How much monthly revenue the customer generates | Integer Values |
| **MonthlyMinutes** | How many minutes per month the customer uses the service | Integer Values |
| **TotalRecurringCharge** | How much recurring charge the customer pays | Integer Values |
| **DirectorAssistedCalls** | Whether the customer avails director assisted calls | Integer Values |
| **PercChangeRevenues** | How much percentage change revenue the customer generates | Integer Values |
| **MonthsInService** | How many months the customer has used the service | Integer Values |
| **TruckOwner** | Customer owns a truck | Yes or No |
| **RVOwner** | Customer owns an RV | Yes or No |
| **OwnsComputer** | Customer owns a computer or not | Yes or No |
| **HasCreditCard** | Customer owns a credit card | Yes or No |
| **IncomeGroup** | How much income the customer generates | 0 – 9 |
| **OwnsMotorcycle** | Whether the customer has a motorcycle or not | Yes or No |
| **CreditRating** | The monthly amount charged to the customer | 1 – 7 |
| **PrizmCode** | What social-economic coding scheme the customer lies in | Suburban, Town, Rural |
| **Occupation** | What the customer does for a living | Professional, Retired, Self |
| **MaritalStatus** | Whether Customer is married, single or divorced. | Yes, No, Unknown |

Table 3.3: Cell2Cell Features

### *3.1.2.3. LOCAL BUSINESS*

This a locally sourced dataset of a Pakistani service providing company consisting of 19000 samples and 898 attributes. This was completely denormalized data, hence the magnitude of the dimensions.

The dataset is composed of all numerical features. It has 8072 records of the minority class and 10928 records of the majority class that amounts to 42.5% share of the minority class in the entire dataset. A partial snapshot of the data is shown in Figure 3.3.



| | M1_U_OB_DAY_VC_CC_CNT_T6 | M1_U_OB_VC_CC_CNT_T6 | M1_U_OB_VC_DUR_T6 | M1_U_OB_VC_REV_T6 | M1_U_OB_DAY_VC_CNT_T6 | M1_U_OB_DAY_VC_DUR_T6 | M1_U_OB_DAY_VC_REV_T6 | M1_U_OB_EVN_VC_CNT_T6 | M1_U_OB_EVN_VC_CC_CNT_T6 |
|---|---|---|---|---|---|---|---|---|---|
| target_flag | | | | | | | | | |
| **0** | 3 | 6 | 9.5 | 9.859 | 3 | 6.5 | 0.448 | 3 | 3 |
| **1** | 100 | 125 | 716.0 | 21.060 | 201 | 613.0 | 21.060 | 80 | 40 |
| **1** | 43 | 64 | 236.0 | 151.774 | 58 | 185.0 | 107.111 | 30 | 28 |
| **0** | 0 | 2 | 3.0 | 6.901 | 0 | 0.0 | 0.000 | 3 | 2 |
| **0** | 3 | 4 | 6.0 | 15.086 | 4 | 4.0 | 10.157 | 1 | 1 |
| **0** | 2 | 5 | 11.5 | 35.552 | 2 | 3.0 | 9.261 | 5 | 3 |
| **1** | 0 | 0 | 0.0 | 0.000 | 0 | 0.0 | 0.000 | 0 | 0 |
| **1** | 5 | 12 | 15.5 | 19.716 | 5 | 2.5 | 8.215 | 7 | 7 |
| **1** | 0 | 0 | 0.0 | 0.000 | 0 | 0.0 | 0.000 | 0 | 0 |
| **1** | 0 | 0 | 0.0 | 0.000 | 0 | 0.0 | 0.000 | 0 | 0 |

Figure 3.3: Partial snapshot of Local Business Data

The data set is purely usage based. Value added services subscription and usage is also included in the data set. The main attributes for each activity are counts, revenue and time window for service availed. Table 3.4 presents some of the key features in the dataset along with possible values.

| Attribute | Definition | Possible Values |
|---|---|---|
| **target_flag** | Identifies churn or non-churn class | 0 or 1 |
| **OG_FAILURES** | Out Going Failures | Integers |
| **OG_CALLS** | Out Going Calls | Integers |
| **GPRS_REV** | GPRS Revenue | Integers |
| **GPRS_VOL** | GPRS Volume | Integers |
| **DATA_LOAN_AMT** | Data Loan Amount | Integers |
| **INT_BAL** | Internet Balance | Integers |
| **SMS_REV** | SMS Revenue | Integers |
| **SMS_CNT** | SMS Count | Integers |
| **OFNT_VC_CNT** | Offnet Voice Call Count | Integers |
| **OFNT_VC_REV** | Offnet Voice Call Revenue | Integers |

Table 3.4: Local Business Features

### 3.1.3. DATA PRE-PROCESSING

Data cleansing was an important part of the process for detecting and removing parts of the data considered irrelevant and then modifying coarse data [63]. This was done so because the acquired data was in the absolute raw tabular form. In a dataset of 898 columns, it was inefficient to use all them. We employed two techniques to eliminate and choose the strongest features (columns) from the data. We also found it inefficient to use the dataset with missing values.

#### *3.1.3.1. REMOVAL OF ZERO DOMINANT COLUMNS*

After some careful analysis, it was observed that some columns, either had null values or 0 in most of the cells. These columns were deemed weak features as values for comparison did not exist. This was either a result of human error or these fields were not applicable to those users. We performed an analysis of the number of zeros in every column, painting a clearer picture of the dataset. The columns that contained more than 80% zeros were removed from the dataset. This was the first technique used to reduce columns from the dataset.

#### *3.1.3.2. REMOVAL OF NULLS*

Of the columns that remained, we decided to replace the null values with 0. NaN denotes “Not A Number” and is a commonly used way of representing missing data. It is a special value (floating-point) and cannot be converted to any data type other than float. NaN values are a major problem in any kind of data analytics. It is imperative that we deal with NaN in order to get the results we desire.We replaced all NaN values with 0, as well.

#### *3.1.3.3. FEATURE SELECTION*

Recursive Feature Elimination (RFE) is a method for selecting features that fits any model and removes weak elements until a specified number of columns is reached. [64]. We do not know how many valuable features are available, but RFE requires a certain number of them to keep and the rest to discard [65].

We used a decision tree as a classifier wrapped inside RFE to test different parts and select the best ones to find optimal features. RFE is a feature selection algorithm that incorporates wrapping. This means that another machine learning algorithm is used inside an RFE to help the algorithm select the best features. This is opposite to filter-based feature selections that score each element and choose those features with the relevant score.

Technically, RFE also uses filter-based feature selection internally. RFE starts by creating subsets and then incrementally removes features by scoring them based on performance. This is mainly achieved by fitting the selected ML algorithm used in the model’s core, positioning components by rank, removing the minor significant attributes, and re-fitting the model. These iterations are repeated until the desired number of columns remains in the dataset.

#### 3.1.3.4.

#### *3.1.3.5. DATA NORMALIZATION*

After the data dimensions were decreased, the values had to be normalized. Decision making algorithms work best when the units of measurement for data are eliminated, enabling them to more easily perform extensive computations. The Min Max Scalar technique was employed to normalize the data to construct a final dataset that was fed to the classifiers.

The MinMaxScaler is quite a well-known and used technique for data normalization, and follows the following formula:

$$x_{sc} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

It essentially compacts the range of the feature values to be represented between 0 and 1 [66].

#### *3.1.3.6. DATA SAMPLING*

ML algorithms mostly work best when the number of samples of each class are approximately equal. When the samples of one class exceed the other, various problems arise.

Developed by Batista et al. [67], we chose a hybrid method for resolving the class imbalance problem that combines the SMOTE ability to generate examples for minority classes that are synthetic and the ENN ability to remove observations from both categories that are identified as outliers. The process of SMOTE-ENN can be explained as follows.

1. (Start of SMOTE) Random information selected from the minority class.
2. Distance determined between the chosen data and its k closest neighbors.
3. Difference increased with a random value somewhere in the range of 0 and 1; the result then, at that point, added to the minority class as a manufactured example.
4. Step number 2-3 rehashed until the ideal percentage of minority class is met. (End of SMOTE)
5. (Start of ENN) K is not entirely settled as the number of nearest neighbors. By default, K is 3.
6. K-closest neighbor is found of the perception among different perceptions in the dataset; then, at that point, the more significant part class from the K-closest neighbor is returned.
7. If the class of the perception and the more significant part class from the perception's K-nearest neighbor is unique, then, at that point, the perception and its K-nearest neighbor are erased from the dataset.
8. Steps 6 and 7 are rehashed until the ideal percentage of each class is satisfied. (End of ENN)

## 3.2. CHOSEN CLASSIFIERS

### 3.2.1. SUPPORT VECTOR MACHINE

‘Support Vector Machine’ (SVM) is a supervised ML technique that is be used for problems of both classification or regression. Classification problems, however, use it the most. In SVM, we plot each instance as a point in n-dimensional space with the value of each attribute being the value of a particular coordinate on a plane. n is the number of features we have in the dataset. Then a hyper-plane is found that differentiates the two classes very well and aids classification (refer to the below figure 3.4).

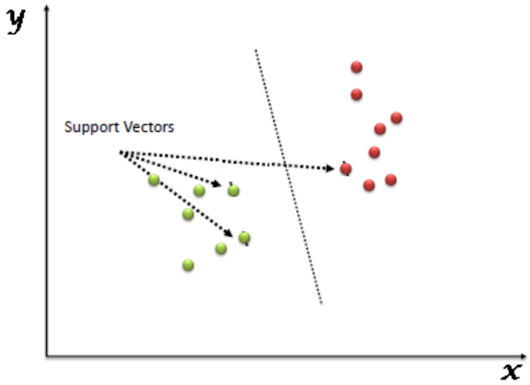


Figure 3.4: SVM Hyperplanes [68]

Support Vectors are basically the directions of individual occurrences. In SVM, the simplest is the straight hyper-plane between two classes. The SVM calculation utilizes a method called the portion stunt. The SVM bit is a capacity that takes low layered info space and changes it to a higher layered space. It is generally helpful in issues of non-direct division. It does a few incredibly complex information changes, then, at that point, discovers the cycle to isolate the information in view of the names or results that have been chosen. We have picked the Radial Basis Function, rbf bit, as our bit work.

Gamma is a kernel coefficient for ‘RBF,’ ‘poly,’ and ‘sigmoid.’ Higher the value of gamma, will try to fit them as per training data set precisely, i.e., generalization error and cause an over-fitting problem. C is the penalty limit of the error term. It also controls the trade-off between smooth decision boundaries and classifying the training points correctly.

RBF can be described with the following formula:

$$K(x,x') = e^{-\gamma||x-x'||^2} \tag{3.2}$$

where gamma can be set manually and has to be greater than 0. The default value for gamma in sklearn’s SVM classification algorithm is:

$$\gamma = \frac{1}{n\,features * \sigma^2} \tag{3.3}$$

Some advantages of using SVM are:

- It functions admirably with an unmistakable margin of detachment.
- It is efficient in high dimensional spaces.

- It is efficient in situations where the quantity of aspects is more prominent than the quantity of tests.
- It utilizes a subset of preparing points in the choice capacity (called SVs), so it is likewise memory proficient.

### 3.2.2. RANDOM FOREST

Random forest is an ML classification algorithm. The "forest" it assembles is a troupe of decision trees, normally prepared with the "bagging" technique. The overall thought of the bagging technique is that a mix of learning models expands the general outcome.

Random forest forms numerous decision trees and unions them to get a more exact and stable forecast.

One huge benefit of random forest is that it tends to be utilized for both classification and regression issues, which structure most of current AI frameworks. We should take a gander at the random forest in order since characterization is here and there viewed as the structure square of AI. In Figure 3.5, we can see how a random forest would look like with two trees:

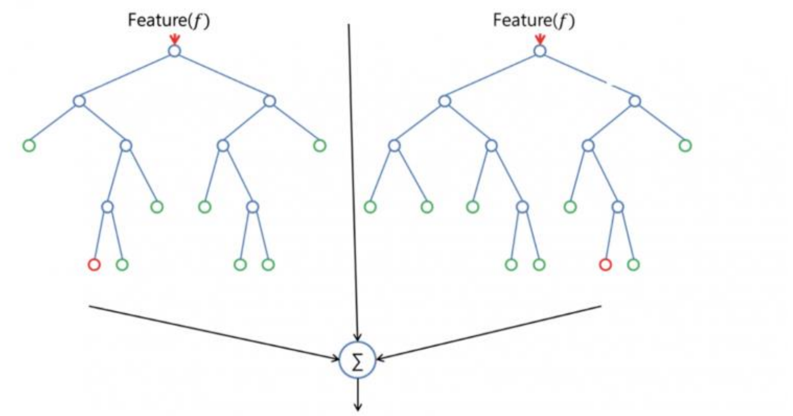


Figure 3.5: Random Tree generation for Random Forest Classifier [69]

The random forest has almost a similar hyperparameters as a decision tree or a bagging classifier. Luckily, there's no compelling reason to consolidate a decision tree with a bagging classifier since you can without much of a stretch utilize the classifier-class of random forest. With random forest, you can likewise manage regression issues utilizing the calculation's regressor.

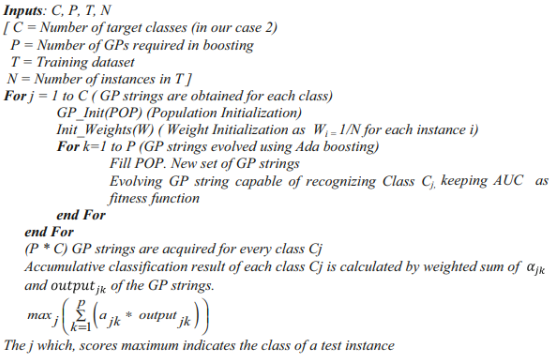
To expand the prescient power, there is the 'n_estimators' hyperparameter, which is only the quantity of trees the calculation works prior to taking the most extreme democratic or taking the midpoints of expectations. A bigger number of trees builds the presentation and makes the projections steadier, dialing back the calculation.

The 'random_state' hyperparameter makes the model's result replicable to speed up. The model will forever create similar outcomes when it has an unequivocal worth of 'random_state' and on the off chance that it has been given similar hyperparameters and similar preparation information.

### 3.2.3. GENETIC PROGRAMMING - ADABOOST

The job of AdaBoost in GP-AdaBoost calculation is to expand support in the development of different GP programs per class. AdaBoost iteratively advances numerous GP programs, where every GP program perceives those examples which are erroneously grouped in past iterations. Boosting expands the weight of updating support over occurrence space to deal with complex cases. The AUC is utilized as the fitness function. Presenting support within GP reinforces the advancement interaction and expands power in developing numerous GP programs. Other than working on the outcomes, the support in GP-AdaBoost likewise saves time because to create a new GP program, an entirely different populace isn't made. Somewhat loads are refreshed over case space for the next cycle. The last expectation of test occurrences is based on the higher worth from a weighted amount of GP programs per class. The pseudo-code given in Figure 3.6 shows the means engaged with GP-AdaBoost calculation.

GP-AdaBoost calculation approaches churn forecast as a one-class issue by treating churners and non-churners independently. P number of GP programs is being developed for each class utilizing boosting. The acquired outcomes are summarized for each category, and the higher, most extreme result recognizes the course of a test occasion. The boundaries for GP-AdaBoost-based calculation are fixed experimentally after performing several analyses.



```
Inputs: C, P, T, N
[ C = Number of target classes (in our case 2)
  P = Number of GPs required in boosting
  T = Training dataset
 N = Number of instances in T ]
For j = 1 to C ( GP strings are obtained for each class)
        GP_Init(POP) (Population Initialization)
        Init_Weights(W) ( Weight Initialization as $W_i = 1/N$ for each instance i)
        For k=1 to P (GP strings evolved using Ada boosting)
                Fill POP. New set of GP strings
                Evolving GP string capable of recognizing Class $C_j$, keeping AUC as
                fitness function
        end For
    end For
    (P * C) GP strings are acquired for every class Cj
    Accumulative classification result of each class Cj is calculated by weighted sum of $\alpha_{jk}$
    and $output_{jk}$ of the GP strings.
    $max_j\left(\sum_{k=1}^{p}\left(a_{jk} * output_{jk}\right)\right)$
The j which, scores maximum indicates the class of a test instance
```

Figure 3.6: Pseudo Code GP-AdaBoost [55]

## 3.3. EVALUATION TECHNIQUES

Below are details of all the evaluation metrics that have been employed for testing model performance.

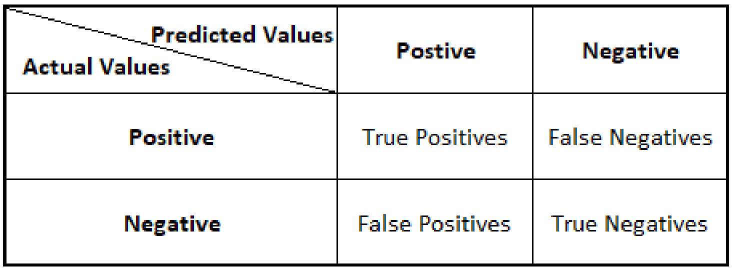
### 3.3.1. CONFUSION MATRIX

A confusion matrix is an evaluation metric used to measure the performance of a classification algorithm. Some common terms associated with it are:

- True Positive: Positive prediction is in fact positive.

- ➢ True Negative: Negative prediction is in fact negative.
- ➢ False Positive: Positive prediction is in fact negative.
- ➢ False Negative: Negative prediction is in fact positive.

A confusion matrix follows the table as shown in Figure 3.7:



| Predicted Values<br>Actual Values | Postive | Negative |
|---|---|---|
| **Positive** | True Positives | False Negatives |
| **Negative** | False Positives | True Negatives |

Figure 3.7: Representation of Confusion Matrix

### 3.3.2. F1 SCORE

A trade-off when avoiding both positives and negatives, that are false, is similarly vital for our problem. F1 Score was thus defined. An f1 score is calculated as the harmonic mean of precision and recall.

$$F_{1}= \frac{2}{\frac{1}{precision}+\frac{1}{recall}} \qquad (3.4)$$

### 3.3.3. AREA UNDER CURVE

To check model performance, we use the receiver operating curve. Below Figure 3.8 is an example:

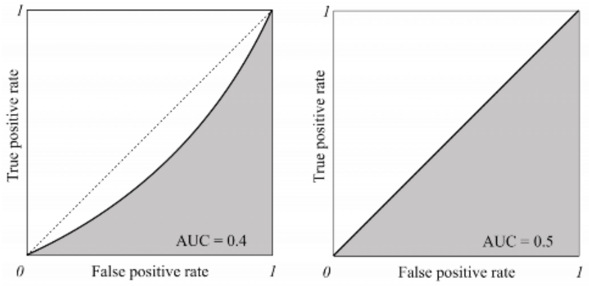


Figure 3.8: Area Under Curve Representation

The x-axis signifies the FRP and the y-axis denotes the TPR.

Recall is also known as True Positive Rate (TPR) and False Positive Rate (FPR) is the proportion of negative examples predicted incorrectly. Both these metrics have a range of 0 to 1. Below are the formulas:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN} \qquad (3.5)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP+TN} \qquad (3.6)$$

Area Under Curve (AUC) is the shaded region. In mathematics the ROC curve is the region between the origin and the coordinates (TPR, FPR).

$$AUC = \int_0^1 (TPR)d\,(FPR) \tag{3.7}$$

A higher area under curve denotes a better performance by the algorithm. AUC-ROC score can be improved by altering true and false-positive rates.

### 3.3.4. ACCURACY

Accuracy can be calculated as a ratio of predictions that are correct made by the classification algorithm.

It can be written as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Number\ of\ Rows\ in\ Data} \tag{3.8}$$

Which can also be denoted as follows:

$$Accuracy = \frac{TP+TN}{Number\ of\ Rows\ in\ Data} \tag{3.9}$$

For classes that are balanced accuracy is a good metric to use, i.e. share of occurrences of all classes are fairly equal. However, when classes are not balanced, accuracy is not a very reliable evaluation metric to be utilized, i.e. the occurrences of each class are not fairly equal to each other, quantity of one is dominant.

### 3.3.5. PRECISION

Precision specifies positive predictions out of all how many are positive in actual. It is calculated as a percentage of accurate predictions that are positive to total predictions that are positive.

$$Precision = \frac{TP}{TP+FP} \tag{3.10}$$

### 3.3.6. RECALL

Recall specifies the actual positive instances overall how many are accurately predicted as positive. It is a percentage of accurate predictions that are positive to the overall number of positive instances in the data.

$$Recall = \frac{TP}{TP+FN} \tag{3.11}$$

Mathematically, it is not possible for both precision and recall to increase simultaneously, as both are proportional inversely to one another. It also depends on the ML problem being dealt as it is decided, then, which of them is more important to us.

### 3.3.7. CROSS-VALIDATION

Cross-validation is a technique that can help gauge the expertise of ML models.

It is regularly used in ML applications to help select a model for a given forecasting issue since it is straightforward, simple to execute, and brings about expertise approximations that for the most part have a lower bias than different strategies.

Cross-validation is a resampling strategy used to assess ML models on a restricted information dataset.

The system has a solitary boundary called k that alludes to the quantity of gatherings that a given data instance is to be divided into. In that way, the method is frequently called k-fold cross-validation. At the point when a particular incentive for k is picked, it very well might be utilized instead of k in the reference to the model, for example, k=10 becoming 10-fold cross-validation.

Cross-validation is basically utilized in applied ML to appraise the expertise of an ML model on information that is not commonly known. That is, to involve a restricted example to assess how the model is relied upon to act in everyday when used to make forecasts on information not utilized during the preparation of the model.

It is a well-known technique since it is easy to comprehend and in light of the fact that it by and large outcomes in a less one-sided or less hopeful gauge of the model expertise than different strategies, like a basic train/test split.

- The overall strategy is as per the following:
- Mix the dataset arbitrarily.
- Split the dataset into k folds.
- For every interesting gathering:
    - Accept the fold as a wait or test data
    - Accept the leftover folds as a training data
    - Fit a model on the training set and assess it on the test set
    - Hold the assessment score and dispose of the model
    - Sum up the ability of the model utilizing the example of model assessment scores

Significantly, every perception in the information test is doled out to a singular gathering and stays in that gathering for the term of the technique. This implies that each example is offered the chance to be utilized in the hold out set 1 time and used to prepare the model k-fold times.

## 3.4. IMPLEMENTATION

This section explains in detail how our pipeline was implemented.

### 3.4.1. TOOLS

We used Google Colaboratory for the complete implementation. Colaboratory or Colab is a product by Google. Colab allows anybody to write and execute python code through the Google browser.

### 3.4.2. PYTHON LIBRARIES

#### *3.4.2.1. NUMPY*

NumPy is a basic package for all scientific and mathematical computing in Python. It contains functions for all kinds of array handling and manipulation. We used numpy for NaN elimination and some basic mathematical functions.

#### *3.4.2.2. PANDAS*

Pandas is an open-source package for data wrangling in Python. It includes multiple libraries pertinent for data handling of all sorts. We used pandas for all our data frame manipulations.

#### *3.4.2.3. SEABORN*

Seaborn is a data visualization library integrated with Pandas in Python. Visualizations an important part of Seaborn which help in better understanding of data. We used seaborn for generating heatmaps of our confusion matrices.

#### *3.4.2.4. IMBLEARN*

Imblearn is mainly used to cater for class imbalance problems. We have two options for solving an issue of this sort. We can either down-sample the data which means that the samples of the majority class are removed based on some criteria or we can up-sample the data which means that synthetic samples are created for the minority class. We used imblearn forn hybrid data sampling.

#### *3.4.2.5. SKLEARN*

Sklearn is a fundamental library for Python that is mostly used in ML projects. Scikit-learn focuses on ML tools including mathematics, statistics and general-purpose algorithms that form the basis for many ML technologies. We used sklearn for all our evaluation metrics, model selections, feature selections, preprocessing and ensembles.

### 3.4.3. MACHINE LEARNING PIPELINE

The below steps were performed in order to run our pipeline and achieve our results.

1. Relevant libraries were imported.
2. Data was read.
3. Index, containing the label, was defined.
4. Columns with more than 80% zeros were eliminated.
5. NaN values were replaced with 0.
6. Recursive Feature Elimination was used to reduce the dimensions of the datasets with a Decision Tree estimator.
7. Reduced data was normalized using Min Max Scaler technique.
8. Normalized data was sampled using a hybrid technique which combined SMOTE and ENN.
9. Grid Search was employed to calculate the best parameter values.
10. Data was split into 70/30 training and testing.
11. Three different classification algorithms were applied to the pre-processed training data i.e. SVM, RF and GP-AdaBoost.
12. Evaluation metrics were calculated for the processed testing data.
13. Further support for the results was obtained through 10-Fold cross-validation.
14. Average evaluation metrics were calculated, again, for the cross-validation results.

# Chapter 4

# Results and Discussion

# CHAPTER 4: RESULTS AND DISCUSSION

This chapter talks about the detailed findings after running our machine learning pipeline for various parameters and combinations of techniques. We, also, present a comparative study to highlight the significance of our proposed methodology.

## 4.1. RESULTS

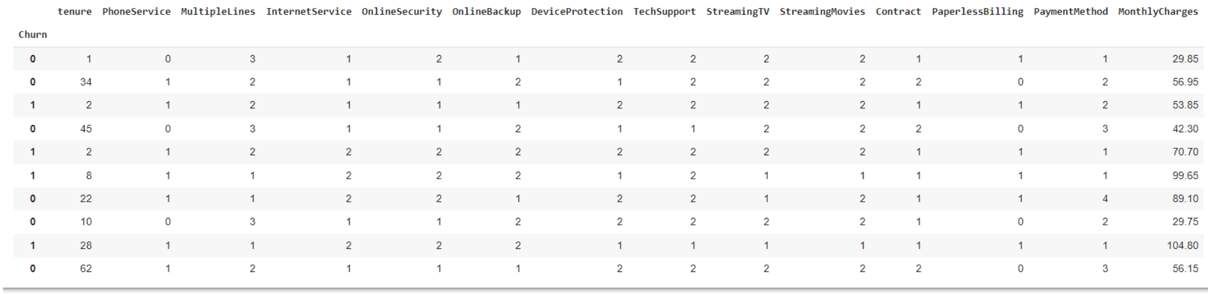
### 4.1.1. FEATURE SELECTION

We used two methods of feature elimination. We removed zero dominant columns to reduce dimensions. For IBM Watson, there weren't any columns with more than 80% zeros but for Cell2Cell and Local Business we dropped 18 and 48 columns, respectively. After removing columns with more than 80% zeros we employed Recursive Feature Elimination with the Decision Tree Classifier.

For IBM Watson, we had literature available to verify what the most efficient and useful features are. In [48], Internet Service and Monthly Charge were identified as two of the most effective features. Tenure, Usage, Monthly Charge, Contract Type and Payment Method were highlighted in [52]. [56] emphasized Tenure, Paperless Billing, Contract Type, Internet Service, Monthly Charge as being the best features.

Our findings were supported by the mentioned literature. We selected 14 out of a total of 19 features. Figure 4.1 shows a snippet of some of the columns. Some important features identified were:

- Tenure
- Internet Service
- Contract
- Paperless Billing
- Monthly Charges

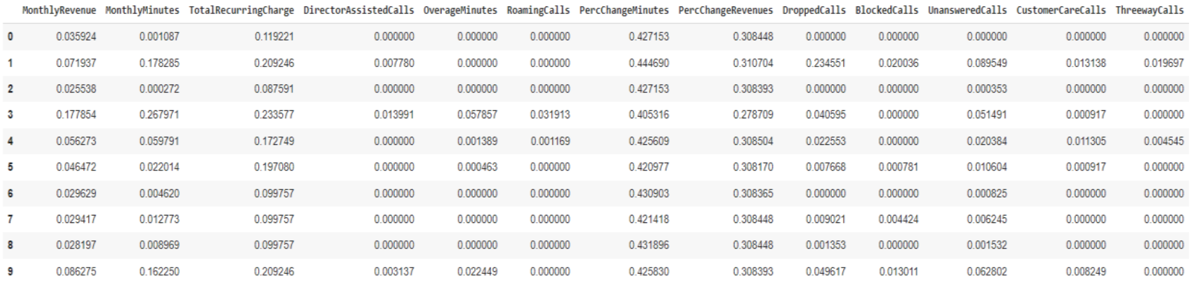


| | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract | PaperlessBilling | PaymentMethod | MonthlyCharges |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Churn | | | | | | | | | | | | | | |
| **0** | 1 | 0 | 3 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 29.85 |
| **0** | 34 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 0 | 2 | 56.95 |
| **1** | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 53.85 |
| **0** | 45 | 0 | 3 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 0 | 3 | 42.30 |
| **1** | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 70.70 |
| **1** | 8 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 99.65 |
| **0** | 22 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 4 | 89.10 |
| **0** | 10 | 0 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 2 | 29.75 |
| **1** | 28 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 104.80 |
| **0** | 62 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 3 | 56.15 |

Figure 4.1: IBM Watson Reduced Features

For Cell2Cell, we selected 30 out of a total of 55 features. Figure 4.2 shows a snippet of some of the columns. Some important features identified were:

- ➢ Monthly Revenue
- ➢ Has Credit Card
- ➢ Prizm Code
- ➢ Credit Rating
- ➢ Total Recurring Charge

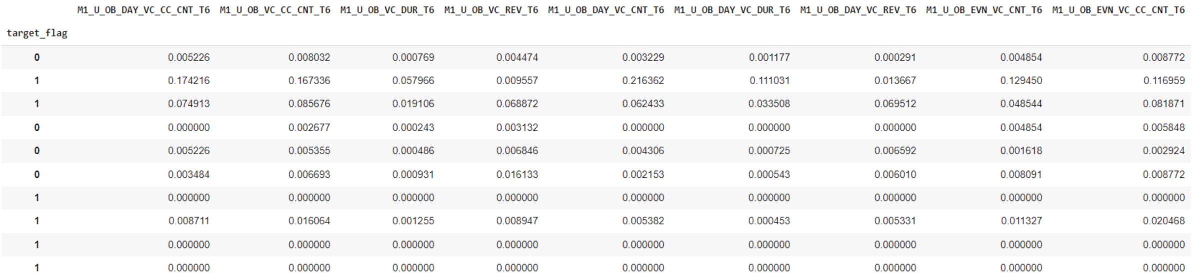


| | MonthlyRevenue | MonthlyMinutes | TotalRecurringCharge | DirectorAssistedCalls | OverageMinutes | RoamingCalls | PercChangeMinutes | PercChangeRevenues | DroppedCalls | BlockedCalls | UnansweredCalls | CustomerCareCalls | ThreewayCalls |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.035924 | 0.001087 | 0.119221 | 0.000000 | 0.000000 | 0.000000 | 0.427153 | 0.308448 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.071937 | 0.178285 | 0.209246 | 0.007780 | 0.000000 | 0.000000 | 0.444690 | 0.310704 | 0.234551 | 0.020036 | 0.089549 | 0.013138 | 0.019697 |
| 2 | 0.025538 | 0.000272 | 0.087591 | 0.000000 | 0.000000 | 0.000000 | 0.427153 | 0.308393 | 0.000000 | 0.000000 | 0.000353 | 0.000000 | 0.000000 |
| 3 | 0.177854 | 0.267971 | 0.233577 | 0.013991 | 0.057857 | 0.031913 | 0.405316 | 0.278709 | 0.040595 | 0.000000 | 0.051491 | 0.000917 | 0.000000 |
| 4 | 0.056273 | 0.059791 | 0.172749 | 0.000000 | 0.001389 | 0.001169 | 0.425609 | 0.308504 | 0.022553 | 0.000000 | 0.020384 | 0.011305 | 0.004545 |
| 5 | 0.046472 | 0.022014 | 0.197080 | 0.000000 | 0.000463 | 0.000000 | 0.420977 | 0.308170 | 0.007668 | 0.000781 | 0.010604 | 0.000917 | 0.000000 |
| 6 | 0.029629 | 0.004620 | 0.099757 | 0.000000 | 0.000000 | 0.000000 | 0.430903 | 0.308365 | 0.000000 | 0.000000 | 0.000825 | 0.000000 | 0.000000 |
| 7 | 0.029417 | 0.012773 | 0.099757 | 0.000000 | 0.000000 | 0.000000 | 0.421418 | 0.308448 | 0.009021 | 0.004424 | 0.006245 | 0.000000 | 0.000000 |
| 8 | 0.028197 | 0.008969 | 0.099757 | 0.000000 | 0.000000 | 0.000000 | 0.431896 | 0.308448 | 0.001353 | 0.000000 | 0.001532 | 0.000000 | 0.000000 |
| 9 | 0.086275 | 0.162250 | 0.209246 | 0.003137 | 0.022449 | 0.000000 | 0.425830 | 0.308393 | 0.049617 | 0.013011 | 0.062802 | 0.008249 | 0.000000 |

Figure 4.2: Cell2Cell Reduced Features

For the Local Business, we selected 60 out of a total of 898 features. Figure 4.3 shows a snippet of some of the columns. Some important features identified were:

- ➢ GPRS Volume
- ➢ GPRS Revenue
- ➢ Loan Amount
- ➢ Voice Call Revenue
- ➢ SMS Revenue



| | M1_U_OB_DAY_VC_CC_CNT_T6 | M1_U_OB_VC_CC_CNT_T6 | M1_U_OB_VC_DUR_T6 | M1_U_OB_VC_REV_T6 | M1_U_OB_DAY_VC_CNT_T6 | M1_U_OB_DAY_VC_DUR_T6 | M1_U_OB_DAY_VC_REV_T6 | M1_U_OB_EVN_VC_CNT_T6 | M1_U_OB_EVN_VC_CC_CNT_T6 |
|---|---|---|---|---|---|---|---|---|---|
| target_flag | | | | | | | | | |
| 0 | 0.005226 | 0.008032 | 0.000769 | 0.004474 | 0.003229 | 0.001177 | 0.000291 | 0.004854 | 0.008772 |
| 1 | 0.174216 | 0.167336 | 0.057966 | 0.009557 | 0.216362 | 0.111031 | 0.013667 | 0.129450 | 0.116959 |
| 1 | 0.074913 | 0.085676 | 0.019106 | 0.068872 | 0.062433 | 0.033508 | 0.069512 | 0.048544 | 0.081871 |
| 0 | 0.000000 | 0.002677 | 0.000243 | 0.003132 | 0.000000 | 0.000000 | 0.000000 | 0.004854 | 0.005848 |
| 0 | 0.005226 | 0.005355 | 0.000486 | 0.006846 | 0.004306 | 0.000725 | 0.006592 | 0.001618 | 0.002924 |
| 0 | 0.003484 | 0.006693 | 0.000931 | 0.016133 | 0.002153 | 0.000543 | 0.006010 | 0.008091 | 0.008772 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.008711 | 0.016064 | 0.001255 | 0.008947 | 0.005382 | 0.000453 | 0.005331 | 0.011327 | 0.020468 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |

Figure 4.3: Local Business Reduced Features

It is to be noted that the features selected for the benchmark datasets, IBM Watson and Cell2Cell, were features pertaining to demographic attributes and revenue-based facts. These features can be made generic for other business domains, as well, hence can be considered as generalized features.

Usually, applications of mining data use datasets with a very high number of attributes. It is quite unfortunate, however, that a large number of features affect the time it takes for the processing to complete and the classification result. A very common and effective solution is to take only a limited number of the most relevant features. Accuracy can be increased by selecting the right attributes. Results of the evaluation metrics suggest that selecting limited features does improve the accuracy percentage. Based on our results, however, selection of feature isn't always the solution to the classification problem, it heavily depends on the kind of features and the method used. In this study, we present the least dimensions that can be used to gain an increase in accuracy, however minimal.

### 4.1.2. CLASSIFIER RESULTS

The experimental setup comprised of four categories.

1. Without RFE and SMOTE-ENN
2. With RFE, Without SMOTE-ENN
3. Without RFE, With SMOTE-ENN
4. With RFE and SMOTE-ENN

All of these setups included data cleansing, handling of missing values, data normalization and extensive evaluation. The evaluation consisted of 70/30 training/testing, 10-fold cross validation, AUC, F1 score, recall and precision along with accuracy.

Tables 4.1 to 4.6 give a very detailed account of the values that were acquired as a result of running all the pipelines in each category. These tables also present a detailed comparison of all three algorithms.

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.772 | 0.675 | 0.518 | 0.470 | 0.577 |
| | Result w/o Sampling | 0.768 | 0.680 | 0.533 | 0.488 | 0.588 |
| | Result w/o RFE | 0.970 | 0.971 | 0.971 | 0.960 | 0.982 |
| | Result w/ RFE, Sampling | 0.984 | 0.984 | 0.984 | 0.986 | 0.982 |
| RF | Result w/o RFE, Sampling | 0.788 | 0.690 | 0.544 | 0.486 | 0.618 |
| | Result w/o Sampling | 0.784 | 0.690 | 0.549 | 0.484 | 0.633 |
| | Result w/o RFE | 0.969 | 0.969 | 0.970 | 0.964 | 0.976 |
| | Result w/ RFE, Sampling | 0.977 | 0.977 | 0.977 | 0.988 | 0.967 |
| GP-AB | Result w/o RFE, Sampling | 0.781 | 0.709 | 0.584 | 0.542 | 0.631 |
| | Result w/o Sampling | 0.798 | 0.711 | 0.583 | 0.521 | 0.663 |
| | Result w/o RFE | 0.921 | 0.921 | 0.924 | 0.933 | 0.916 |
| | Result w/ RFE, Sampling | 0.937 | 0.937 | 0.938 | 0.960 | 0.918 |

Table 4.1: IBM Watson 70/30 Results

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.765 | 0.679 | 0.529 | 0.497 | 0.566 |
| | Result w/o Sampling | 0.770 | 0.686 | 0.54 | 0.508 | 0.575 |
| | Result w/o RFE | 0.984 | 0.984 | 0.984 | 0.977 | 0.991 |
| | Result w/ RFE, Sampling | 0.983 | 0.983 | 0.983 | 0.978 | 0.988 |
| RF | Result w/o RFE, Sampling | 0.784 | 0.692 | 0.549 | 0.495 | 0.616 |
| | Result w/o Sampling | 0.799 | 0.711 | 0.58 | 0.525 | 0.649 |
| | Result w/o RFE | 0.976 | 0.976 | 0.976 | 0.975 | 0.977 |
| | Result w/ RFE, Sampling | 0.975 | 0.975 | 0.975 | 0.975 | 0.976 |
| GP-AB | Result w/o RFE, Sampling | 0.801 | 0.714 | 0.586 | 0.53 | 0.654 |
| | Result w/o Sampling | 0.802 | 0.719 | 0.592 | 0.54 | 0.655 |
| | Result w/o RFE | 0.92 | 0.919 | 0.923 | 0.931 | 0.916 |
| | Result w/ RFE, Sampling | 0.925 | 0.925 | 0.928 | 0.935 | 0.920 |

Table 4.2: IBM Watson 10-Fold CV Results

Table 4.1 and 4.2 are results obtained after running the pipeline on IBM Watson. The best results obtained are for SVM.

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.948 | 0.956 | 0.952 | 0.916 | 0.99 |
| | Result w/o Sampling | 0.973 | 0.915 | 0.903 | 0.839 | 0.977 |
| | Result w/o RFE | 0.994 | 0.991 | 0.995 | 0.998 | 0.992 |
| | Result w/ RFE, Sampling | 0.992 | 0.989 | 0.994 | 0.999 | 0.99 |
| RF | Result w/o RFE, Sampling | 0.723 | 0.532 | 0.159 | 0.089 | 0.578 |
| | Result w/o Sampling | 0.723 | 0.532 | 0.157 | 0.091 | 0.573 |
| | Result w/o RFE | 0.838 | 0.797 | 0.884 | 0.919 | 0.851 |
| | Result w/ RFE, Sampling | 0.839 | 0.794 | 0.885 | 0.927 | 0.846 |
| GP-AB | Result w/o RFE, Sampling | 0.718 | 0.526 | 0.141 | 0.082 | 0.515 |
| | Result w/o Sampling | 0.719 | 0.524 | 0.132 | 0.075 | 0.528 |
| | Result w/o RFE | 0.782 | 0.73 | 0.845 | 0.886 | 0.809 |
| | Result w/ RFE, Sampling | 0.791 | 0.737 | 0.852 | 0.894 | 0.814 |

Table 4.3: Cell2Cell 70/30 Results

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.939 | 0.950 | 0.944 | 0.903 | 0.989 |
| | Result w/o Sampling | 0.969 | 0.900 | 0.885 | 0.809 | 0.975 |
| | Result w/o RFE | 0.993 | 0.991 | 0.995 | 0.998 | 0.992 |
| | Result w/ RFE, Sampling | 0.993 | 0.990 | 0.995 | 0.999 | 0.991 |
| RF | Result w/o RFE, Sampling | 0.719 | 0.532 | 0.160 | 0.925 | 0.578 |
| | Result w/o Sampling | 0.719 | 0.532 | 0.160 | 0.093 | 0.577 |
| | Result w/o RFE | 0.844 | 0.805 | 0.888 | 0.923 | 0.855 |
| | Result w/ RFE, Sampling | 0.843 | 0.804 | 0.887 | 0.921 | 0.856 |
| GP-AB | Result w/o RFE, Sampling | 0.716 | 0.528 | 0.143 | 0.082 | 0.552 |
| | Result w/o Sampling | 0.715 | 0.527 | 0.141 | 0.081 | 0.542 |
| | Result w/o RFE | 0.784 | 0.733 | 0.846 | 0.886 | 0.809 |
| | Result w/ RFE, Sampling | 0.793 | 0.739 | 0.854 | 0.895 | 0.813 |

Table 4.4: Cell2Cell 10-Fold CV Results

Table 4.3 and 4.4 are results obtained after running the pipeline on Cell2Cell. The best results obtained are, also, for SVM.

It is to be observed that these two are benchmark datasets. The results obtained are superior to the previously reported results. A detailed comparison has been done in a later section of this chapter.

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.879 | 0.866 | 0.845 | 0.778 | 0.926 |
| | Result w/o Sampling | 0.905 | 0.911 | 0.895 | 0.949 | 0.847 |
| | Result w/o RFE | 0.953 | 0.957 | 0.956 | 0.925 | 0.989 |
| | Result w/ RFE, Sampling | 0.962 | 0.964 | 0.964 | 0.937 | 0.992 |
| RF | Result w/o RFE, Sampling | 0.994 | 0.994 | 0.993 | 0.991 | 0.994 |
| | Result w/o Sampling | 0.995 | 0.995 | 0.995 | 0.994 | 0.995 |
| | Result w/o RFE | 0.998 | 0.998 | 0.998 | 0.997 | 0.999 |
| | Result w/ RFE, Sampling | 0.998 | 0.998 | 0.998 | 0.997 | 0.999 |
| GP-AB | Result w/o RFE, Sampling | 0.997 | 0.998 | 0.997 | 1.000 | 0.995 |
| | Result w/o Sampling | 0.998 | 0.998 | 0.997 | 0.999 | 0.995 |
| | Result w/o RFE | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| | Result w/ RFE, Sampling | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 4.5: Local Business 70/30 Results

| Algorithm | Methodology | Accuracy | Area Under Curve | F1 Score | Recall | Precision |
|---|---|---|---|---|---|---|
| SVM | Result w/o RFE, Sampling | 0.890 | 0.877 | 0.860 | 0.793 | 0.900 |
| | Result w/o Sampling | 0.911 | 0.916 | 0.900 | 0.950 | 0.938 |
| | Result w/o RFE | 0.962 | 0.964 | 0.964 | 0.940 | 0.990 |
| | Result w/ RFE, Sampling | 0.963 | 0.965 | 0.964 | 0.939 | 0.991 |
| RF | Result w/o RFE, Sampling | 0.993 | 0.992 | 0.991 | 0.990 | 0.993 |
| | Result w/o Sampling | 0.996 | 0.995 | 0.995 | 0.995 | 0.995 |
| | Result w/o RFE | 0.998 | 0.998 | 0.998 | 0.997 | 0.999 |
| | Result w/ RFE, Sampling | 0.998 | 0.998 | 0.998 | 0.998 | 0.999 |
| GP-AB | Result w/o RFE, Sampling | 0.997 | 0.997 | 0.997 | 1.000 | 0.993 |
| | Result w/o Sampling | 0.997 | 0.997 | 0.997 | 1.000 | 0.994 |
| | Result w/o RFE | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | Result w/ RFE, Sampling | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 4.6: Local Business 10-Fold CV Results

Table 4.5 and 4.6 are results obtained after running the pipeline on the Local Business data. The best results obtained are for GP-AdaBoost. SVM, however, has also, produced promising results.

It is to be noted for the Local Business, the accuracy was far better on the run without RFE and Sampling for the data. A major reason for this is the data being less imbalanced than the other two.

Table 4.7 to 4.9 are summarized tables of the best results obtained for every dataset and algorithm. We picked up the experimental setup with RFE and Sampling for a fair comparison between datasets.

| IBM WATSON | | | |
|---|---|---|---|
| Classifier | Accuracy | AUC | F1 Score |
| **SVM** | **0.984** | **0.984** | **0.984** |
| **RF** | 0.977 | 0.977 | 0.977 |
| **GP-AB** | 0.937 | 0.937 | 0.938 |

Table 4.7: IBM Watson Results w/ RFE and Sampling

| CELL2CELL | | | |
|---|---|---|---|
| Classifier | Accuracy | AUC | F1 Score |
| **SVM** | **0.993** | **0.990** | **0.995** |
| **RF** | 0.843 | 0.804 | 0.887 |
| **GP-AB** | 0.791 | 0.737 | 0.852 |

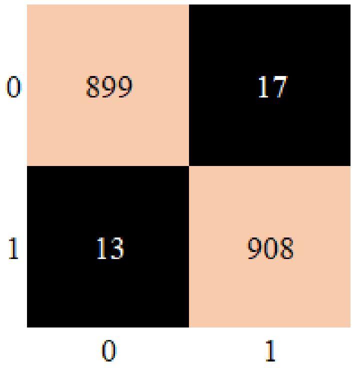
Table 4.8: Cell2Cell Results w/ RFE and Sampling

| LOCAL BUSINESS | | | |
|---|---|---|---|
| Classifier | Accuracy | AUC | F1 Score |
| **SVM** | 0.963 | 0.965 | 0.964 |
| **RF** | 0.998 | 0.998 | 0.998 |
| **GP-AB** | **1.000** | **1.000** | **1.000** |

Table 4.9: Local Business Results w/ RFE and Sampling

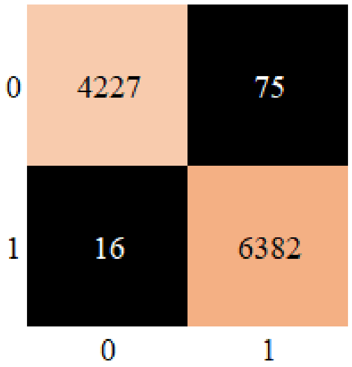
SVM produced all around best results for all the datasets. For the two benchmark datasets, it produced the best results, which were better than any reported result, thus far. For the Local Business, even though SVM had promising results, GP-AdaBoost gave a near perfect accuracy.

Figures 4.4 to 4.6 are confusion matrices created for all three datasets for SVM for the 70/30 training/testing evaluation. Figures 4.7 to 4.9 are confusion matrices created for all three datasets for SVM for the 10-fold cross validation.



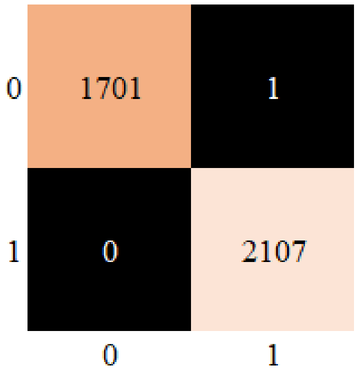
|   |   |   |
|---|---|---|
| 0 | 899 | 17 |
| 1 | 13 | 908 |
|   | 0 | 1 |

Figure 4.4: IBM Watson 70/30 Confusion Matrix



|   |   |   |
|---|---|---|
| 0 | 4227 | 75 |
| 1 | 16 | 6382 |
|   | 0 | 1 |

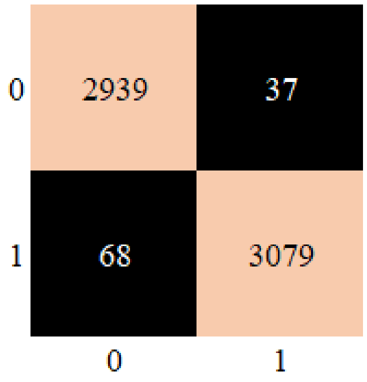
Figure 4.5: Cell2Cell 70/30 Confusion Matrix



|   |   |   |
|---|---|---|
| 0 | 1701 | 1 |
| 1 | 0 | 2107 |
|   | 0 | 1 |

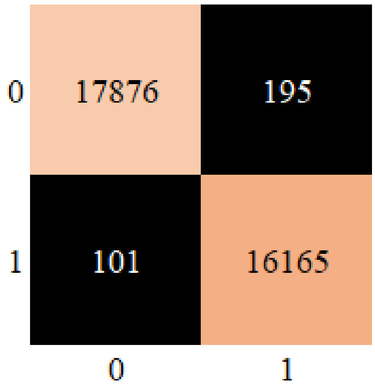
Figure 4.6: Local Business 70/30 Confusion Matrix

It can clearly be seen that miss classifications did exist in both categories but numbers were small enough to create an almost negligible effect.



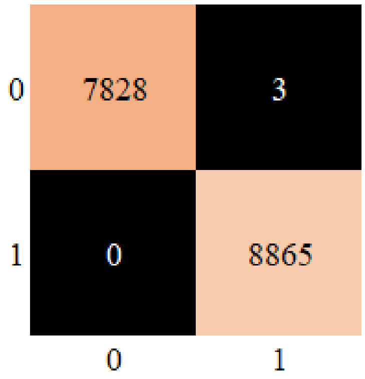
|  |  |  |
|---|---|---|
| 0 | 2939 | 37 |
| 1 | 68 | 3079 |
|  | 0 | 1 |

Figure 4.7: IBM Watson 10-Fold CV Confusion Matrix



|  |  |  |
|---|---|---|
| 0 | 17876 | 195 |
| 1 | 101 | 16165 |
|  | 0 | 1 |

Figure 4.8: Cell2Cell 10-Fold CV Confusion Matrix



|  |  |  |
|---|---|---|
| 0 | 7828 | 3 |
| 1 | 0 | 8865 |
|  | 0 | 1 |

Figure 4.9: Local Business 10-Fold CV Confusion Matrix

These confusion matrices just corroborated the findings of the 70/30 evaluation run. 10-fold is most beneficial in cases like these because data in its entirety is tested.

The problem we are catering for causes the organization to lose money in case of either false negatives or false positives. In case of false negatives, a customer that is likely to be lost or churned is not identified in time and in case of a false positive and already satisfied customer is identified as an unhappy one. In the first case the company would have to acquire a completely new customer to compensate for the lots one and in the second case the company would have spent money trying to retain a customer that was already satisfied with the services as opposed to having spent that money on retaining a customer at risk of churning.

These confusion matrices show a very promising result for the different categories of our classification. The accuracy presented in these matrices is quite good and presents a sound analysis of how well the algorithms performed.

## 4.2. COMPARATIVE STUDY

The comparative study was done in two phases. The first was an easy comparison between the four experimental setups. The second phase comprised of benchmark results compared with previously reported results.

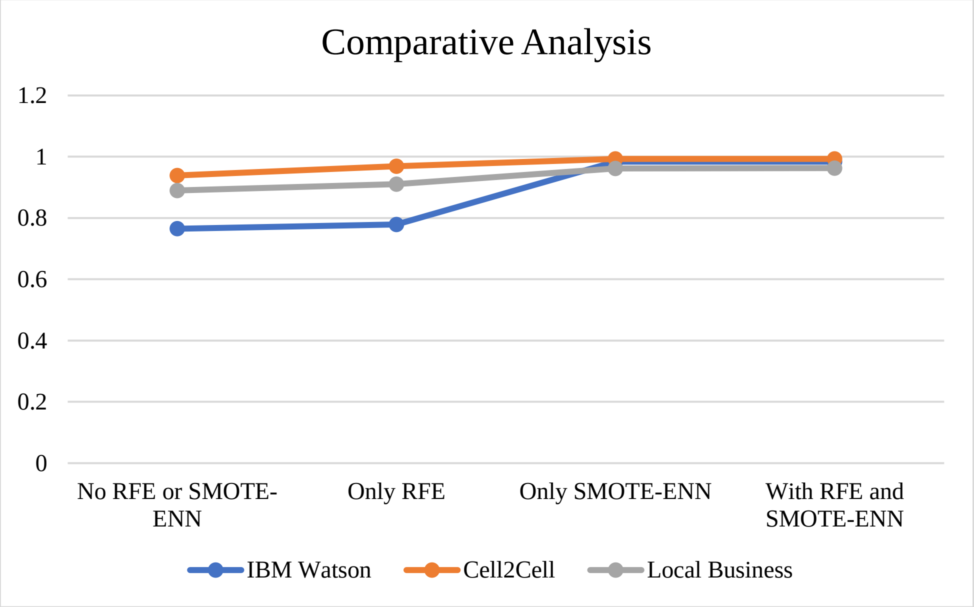


Figure 4.10: Comparative Analysis b/w Different Pipeline Components

Figure 4.10 is a clear representation of how the accuracy increased with each upgrade to the experimental setup. Even though feature extraction made a positive difference to the accuracy, what caused the most significant change was sampling. Sampling was the star performer of the entire pipeline.

We started by running the algorithm on normalized and cleaned data without RFE and sampling. We then introduced RFE to see how the accuracy was affected. We saw a minimal change but it was on the positive side. We then removed feature selection and introduced a hybrid sampling technique into the pipeline to test results. Quite pleasantly, we noticed a significant change in accuracy and pipeline performance. Combining the two techniques, increased the accuracy marginally. Overall, it is advised to use both techniques together because feature selection helps reduce processing times.

| Paper | Algorithm | Technique | Evaluation Metric | Result |
|---|---|---|---|---|
| **[50]** | Naïve Bayes | Basic Pipeline | Area Under Curve | 0.820 |
| | **Support Vector Machine** | | | **0.870** |
| | Decision Trees | | | 0.770 |
| **[51]** | Logistic Regression | Feature Selection | F1 Score | 0.978 |
| | Decision Trees | | | 0.779 |
| | **Random Forest** | | | **0.778** |
| **[56]** | **Random Forest** | Basic Pipeline | Accuracy | **0.936** |
| | **Support Vector Machine** | | | **0.819** |
| | Decision Tree | | | 0.762 |
| | Logistic Regression | | | 0.789 |
| **[57]** | XG-Boost | Feature Selection | Area Under Curve | 0.850 |
| | **Random Forest** | | | **0.840** |
| | Decision Tree | | | 0.810 |
| **Proposed Model** | **Random Forest** | RFE and SMOTE-ENN | **Accuracy, Area Under Curve, F1 Score** | **0.977** |
| | **Support Vector Machine** | | | **0.984** |

Table 4.10: IBM Watson Comparison

| Paper | Algorithm | Technique | Evaluation Metric | Result |
|---|---|---|---|---|
| **[50]** | Naïve Bayes | Feature Selection | Area Under Curve | 0.980 |
| | **Support Vector Machine** | | | **0.990** |
| | Decision Trees | | | 0.980 |
| **[55]** | **Random Forest** | PSO Under-Sampling | Area Under Curve | **0.592** |
| | Rotation Forest | | | 0.610 |
| | **GP-AdaBoost w/ DTC** | | | **0.910** |
| **Proposed Model** | **Support Vector Machine** | **RFE and SMOTE-ENN** | **Area Under Curve** | **0.994** |
| | **Random Forest** | | | **0.991** |
| | **GP-AdaBoost** | | | **0.737** |

Table 4.11: Cell2Cell Comparison

Table 4.10 shows that all papers that used IBM Watson as a benchmark either used a basic machine learning pipeline as the proposed method or added feature selection to it. As shown in Tables 13, we achieved a far better accuracy after incorporation of sampling with SMOTE-ENN.

Table 4.11 shows a comparison with papers using Cell2Cell as a benchmark. The best previously reported results were either with Feature Selection or with PSO Under-Sampling. There are two very notable observations to be made, explanations for which are given, below.

1. Random Forest performed significantly better with sampling for our pipeline.
2. GP-AdaBoost did not perform as well as previously reported.

Cell2Cell is a heavily imbalanced dataset, as is stated in Table 3.1. We found that under-sampling alone does not produce good results but when combined with over sampling, it produces magnificent results. Our pipeline used a different approach to sampling than used in the research work. We used a hybrid approach to sample our data.

For the second observation, we found that using a superior sampling technique did not help us achieve the desired results. Decision Tree is being used as a wrapped classifier in addition to GP-AdaBoost in the paper. We are, however, using GP-AdaBoost as a stand-alone classifier. The DTC and GP-AdaBoost hybrid is helping them achieve a better accuracy than we were able to achieve. GP-AdaBoost, as constructed in this study, proved to be an effective classifier for datasets other than Cell2Cell.

Overall, superior results were achieved as compared to previous studies.

# Chapter 5

# Conclusion

# CHAPTER 5: CONCLUSION

This chapter concludes the research, recommends ideas for future research and highlights some limitations.

## 5.1. CONCLUSION

This work is valuable to business owners looking to make sure that they maintain an upper hand against their competition. Customer retention and satisfaction is a challenging aspect and this method adds depth and understanding to it making it simple to achieve. Churning of customers in the various industries is indeed an issue that needs prompt detection/resolution so that retention strategies are taken for remaining competitive. Customer data needs to be managed quite cautiously because of the likely class imbalance, large magnitudes and a structure that is multi-dimensional. Feature selection/elimination is preferred for reducing dimensions while sampling is mostly chosen for balancing churn and non-churn classes.

We first looked at an support vector machine, then built a random forest to see how we can overcome a lower accuracy of a single decision tree by combining multiple trees in an ensemble known as a random forest. The random forest makes use of the concepts of random sampling of instances, random sampling of attributes, and averaging of predictions. A separate approach was implemented as the GP-AdaBoost Classifier to, also, test results.

In terms of robustness and accuracy, all three algorithms emerged victorious in different scenarios and presented results better than the reported benchmarks. However, we found that SVM in terms of time wasn't always the best fit. We saw that for IBM Watson SVM and RF presented with a 98.4% and 97.7% accuracy, respectively. For Cell2Cell, we achieved a 99.4% accuracy from SVM. Unfortunately, RF and GP-AdaBoost did not perform up to the mark. For our local business' data, RF and GP-AdaBoost gave us a phenomenal accuracy of 99.8% and 99.9%, respectively. The AUC metric performed equally well for these runs, as well.

## 5.2. LIMITATIONS

This study is centered on a specific set of attributes with include customer demographics, service particulars and revenue fields. These fields were found to exist in both the benchmark datasets of the Telecommunications companies, as well as, the local business. However, we feel an even more robust pipeline could have been created had we been privy to more diversification of data.

## 5.3. FUTURE WORK AND RECOMMENDATIONS

Future research should concentrate on the profitability associated with customer attributes when reducing dimensions. Real-time churn prediction also presents a research gap and can be achieved effectively with the big data technologies. More versatile approaches are also needed in which likely effects of national or regional economic changes on the behaviour of customers is also considered. Natural Language Processing (NLP) could also be used to detect customers who are likely to churn using specific words and voice recognition. Further study is also needed to create proper visualizations of the huge customer churn data. Visualizations make the process of discovering data patterns faster to reveal insights action can be taken upon. Work on optimizing this current study to improve training times can also be extensively looked into.

# REFERENCES

[1] M. Chandar and P. A. L. Krishna, "Modeling churn behavior of bank customers using predictive data mining techniques," in *Proc. Nat. Conf. Soft Comput. Techn. Eng. Appl. (SCT),* 2006.

[2] A. Parvatiyar and J. N. Sheth, "Customer relationship management: Emerging practice, process, and discipline," in *J. Econ. Social Res, vol. 3, no. 2*, 2001.

[3] T. Fields, Mobile & Social Game Design: Monetization Methods and Mechanics, Boca Raton, FL, USA: CRC Press, 2014.

[4] T. Verbraken, W. Verbeke and B. Baesens, "Profit optimizing customer churn prediction with Bayesian network classifiers," in *Intell. Data Anal., vol. 18, no. 1, pp. 3–24,* Jan. 2014.

[5] W. Reinart, J. S. Thomas and V. Kumar, "Balancing acquisition and retention resources to maximize customer profitability," in *J. Marketing, vol. 69, no. 1, pp. 63–79,* Jan. 2005.

[6] W. E. Sasser, "Zero defections: Quality comes to services," in *Harvard Bus. Rev., vol. 68, no. 5, pp. 105–111*, 1990.

[7] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson and H. Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," in *IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 690–696*, May 2000.

[8] F. F. Reichheld, T. Teal and D. K. Smith, The loyalty effect, Boston, MA, USA: Harvard Bus. School Press, 1996.

[9] "Investopedia," [Online]. Available: https://www.investopedia.com/terms/c/churnrate.asp . [Accessed 20 Dec 2021].

[10] S.-Y. Hung, D. C. Yen and H.-Y. Wang, "Applying data mining to telecom churn management," in *Expert Syst. Appl., vol. 31, no. 3, pp. 515–524,* Oct. 2006.

[11] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to churn prediction: A data mining approach," in *Expert Syst. Appl., vol. 23, no. 2, pp. 103–112,* Aug. 2002.

[12] Z. He, X. Xu, J. Z. Huang and S. Deng, "Mining class outliers: Concepts, algorithms and applications in CRM," in *Expert Syst. Appl., vol. 27, no. 4, pp. 681–697,* Nov. 2004.

[13] T. S. H. Teo, P. Devadoss and S. L. Pan, "Towards a holistic perspective of customer relationship management (CRM) implementation: A case study of the housing and development board, Singapore," in *Decis. Support Syst., vol. 42, no. 3, pp. 1613–1627*, Dec. 2006.

[14] E. W. T. Ngai, L. Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," in *Expert Syst. Appl., vol. 36, no. 2, pp. 2592–2602*, Mar. 2009..

[15] M. J. Shaw, C. Subramaniam, G. W. Tan and M. E. Welge, "Knowledge management and data mining for marketing," in *Decis. Support Syst., vol. 31, no. 1, pp. 127–137,* 2001.

[16] M. Komenar, Electronic Marketing., Hoboken, NJ, USA: Wiley, 1996.

[17] R. Bose, "Customer relationship management: Key components for IT success," in *Ind. Manage. Data Syst., vol. 102, no. 2, pp. 89–97,* Mar. 2002.

[18] A. T. Jahromi, M. M. Sepehri, B. Teimourpour and S. Choobdar, "Modeling customer churn in a non-contractual setting: The case of telecommunications service providers," in *J. Strategic Marketing, vol. 18, no. 7, pp. 587–598*, Dec. 2010.

[19] E. W. T. Ngai, "Customer relationship management research (1992– 2002): An academic literature review and classification," in *Marketing Intell. Planning, vol. 23, no. 6, pp. 582–605,* 2005.

[20] A. M. Almana and M. S. A. R. Alzahrani, ‘‘A survey on data mining techniques in customer churn analysis for telecom industry,’’ Int. J. Eng. Res. Appl., vol. 4, no. 5, pp. 165–171, 2014.

[21] A. Ahmed and D. M. Linen, ‘‘A review and analysis of churn prediction methods for customer retention in telecom industries,’’ in Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Jan. 2017, pp. 1–7.

[22] M. Ahmed, H. Afzal, A. Majeed, and B. Khan, ‘‘A survey of evolution in predictive models and impacting factors in customer churn,’’ Adv. Data Sci. Adapt. Anal., vol. 9, no. 3, Jul. 2017, Art. no. 175007

[23] E. Lee, Y. Jang, D.-M. Yoon, J. Jeon, S.-I. Yang, S.-K. Lee, D.-W. Kim, P. P. Chen, A. Guitart, P. Bertens, A. Perianez, F. Hadiji, M. Müller, Y. Joo, J. Lee, I. Hwang and K.-J. Kim, ", ‘‘Game data mining competition on churn prediction and survival analysis using commercial game log data,’’ IEEE Trans. Games, vol. 11, no. 3, pp. 215–226, Sep. 2019.".

[24] R. Zhang, W. Li, W. Tan, and T. Mo, ‘‘Deep and shallow model for insurance churn prediction service,’’ in Proc. IEEE Int. Conf. Services Comput. (SCC), Jun. 2017, pp. 346–353.

[25] D. L. García, À. Nebot, and A. Vellido, ‘‘Intelligent data analysis approaches to churn as a business problem: A survey,’’ Knowl. Inf. Syst., vol. 51, no. 3, pp. 719–774, Jun. 2017.

[26] "Y. Huang, B. Huang, and M. T. Kechadi, “A rule-based method for customer churn prediction in telecommunication services,” in Advances in Knowledge Discovery and Data Mining, Springer, 2011, pp. 411–422.".

[27] A. Idris and A. Khan, “Customer churn prediction for telecommunication: Employing various various features selection techniques and tree-based ensemble classifiers,” in Multitopic Conference (INMIC), 2012 15th International, 2012, pp. 23–27.

[28] M. Kaur, K. Singh, and N. Sharma, “Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers,” Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 9, pp. 720–725, 2013.

[29] V. L. Miguéis, D. Van den Poel, a. S. Camanho, and J. Falcão e Cunha, “Modeling partial customer churn: On the value of first product-category purchase sequences,” Expert Syst. Appl., vol. 39, no. 12, pp. 11250–11256, Sep. 2012.

[30] B. C. A. C. R. G. a. D. M.-M. B. Balle, "“The Architecture of a Churn Prediction System Based on Stream Mining,” in Artificial Intelligence Research and Development: Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence, 2013, vol. 256, p. 157.".

[31] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, “Computer assisted customer churn management: State-of-the-art and future trends,” Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.

[32] Z. Kasiran, Z. Ibrahim, and M. S. M. Ribuan, “Mobile phone customers churn prediction using elman and Jordan Recurrent Neural Network,” in Computing and Convergence Technology (ICCCT), 2012 7th International Conference on, 2012, pp. 673–678.

[33] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker, and N. R. S. Sriram, “Modeling Customer Lifetime Value,” J. Serv. Res., vol. 9, no. 2, pp. 139– 155, Nov. 2006.

[34] L. Breiman, ‘‘Statistical modeling: The two cultures (with comments and a rejoinder by the author),’’ Stat. Sci., vol. 16, no. 3, pp. 199–231, Aug. 2001.

[35] M. Stewart. The Actual Difference Between Statistics and Machine Learning. Towards Data Science. Accessed: May 25, 2019. [Online]. Available: https://towardsdatascience.com/the-actual-difference between-statistics-and-machine-learning-64b49f07ea3

[36] D. Bzdok, N. Altman, and M. Krzywinski, ‘‘Statistics versus machine learning,’’ Nature Methods, vol. 15, no. 4, pp. 233–234, Apr. 2018, doi: 10.1038/nmeth.4642.

[37] I. H. Witten and E. Frank, ‘‘Data mining: Practical machine learning tools and techniques with java implementations,’’ ACM SIGMOD Rec., vol. 31, no. 1, pp. 76–77, Mar. 2002.

[38] P. S. Fader and B. G. S. Hardie, ‘‘Probability models for customer-base analysis,’’ J. Interact. Marketing, vol. 23, no. 1, pp. 61–69, Feb. 2009.

[39] P. S. Fader and B. G. S. Hardie, ‘‘How to project customer retention,’’ J. Interact. Marketing, vol. 21, no. 1, pp. 76–90, Jan. 2007.

[40] D. G. Morrison and D. C. Schmittlein, ‘‘Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort?’’ J. Bus. Econ. Statist., vol. 6, no. 2, pp. 145–159, Apr. 1988.

[41] S. Ma, ‘‘On optimal time for customer retention in non-contractual setting,’’ School Econ. Manage., Jiangsu Univ. Sci. Technol., Jiangsu, China, Tech. Rep., Dec. 2009. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1529284

[42] P. S. Fader, B. G. S. Hardie, and J. Shang, ‘‘Customer-base analysis in a discrete-time noncontractual setting,’’ Marketing Sci., vol. 29, no. 6, pp. 1086–1108, Nov. 2010.

[43] P. S. Fader and B. G. H. K. L. Lee, ‘‘‘Counting your customers’ the easy way: An alternative to the Pareto/NBD model,’’ Marketing Sci., vol. 24, no. 2, pp. 275–284, 2005.

[44] P. S. Fader and B. G. S. Hardie, ‘‘Probability models for customer-base analysis,’’ J. Interact. Marketing, vol. 23, no. 1, pp. 61–69, Feb. 2009.

[45] Z. Qian, W. Jiang, and K.-L. Tsui, ‘‘Churn detection via customer profile modelling,’’ Int. J. Prod. Res., vol. 44, no. 14, pp. 2913–2933, Jul. 2006. VOLUME 8, 2020 220837 J. Ahn et al.: Survey on Churn Analysis in Various Business Domains

[46] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[47] R. Zhang, W. Li, W. Tan, and T. Mo, ‘‘Deep and shallow model for insurance churn prediction service,’’ in Proc. IEEE Int. Conf. Services Comput. (SCC), Jun. 2017, pp. 346–353.

[48] Pamina, Jeyakumar, et al. "An effective classifier for predicting churn in telecommunication." Jour of Adv Research in Dynamical & Control Systems 11 (2019).

[49] Parmar, Miss Priyanka, and Mrs Shilpa Serasiya. "Telecom Churn Prediction Model using XgBoost Classifier and Logistic Regression Algorithm." (2021).

[50] Ebrah, Khulood, and Selma Elnasir. "Churn prediction using machine learning and recommendations plans for telecoms." Journal of Computer and Communications 7.11 (2019): 33-53.

[51] Tamuka, Nyashadzashe, and Khulumani Sibanda. "Real time customer churn scoring model for the telecommunications industry." 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC). IEEE, 2020.

[52] ApurvaSree, G., et al. "Churn prediction in telecom using classification algorithms." International Journal of Scientific Research and Engineering Development 5 (2019): 19-28.

[53] Gajowniczek, K., A. Orłowski, and T. Ząbkowski. "Entropy Based Trees to Support Decision Making for Customer Churn Management." Acta Physica Polonica, A. 129.5 (2016).

[54] Joolfoo, Muhammad BA, Rameshwar A. Jugurnauth, and Khalid MBA Joolfoo. "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform." Journal of Critical Reviews 7.11 (2020): 1991-2001.

[55] Idris, Adnan, Aksam Iftikhar, and Zia ur Rehman. "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling." Cluster Computing 22.3 (2019): 7241-7255.

[56] Sundararajan, Abhijit, and Kemal Gursoy. "Telecom customer churn prediction." (2020).," in *Rutgers University Library*.

[57] Townsend, Anthony, and Sree Nilakanta. "Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning." (2019).

[58] Mitkees, Ibrahim MM, Sherif M. Badr, and Ahmed Ibrahim Bahgat ElSeddawy. "Customer churn prediction model using data mining techniques." 2017 13th International Computer Engineering Conference (ICENCO). IEEE, 2017.

[59] Paliwal, Priyanka, and Divya Kumar. "ABC based neural network approach for churn prediction in telecommunication sector." International Conference on Information and Communication Technology for Intelligent Systems. Springer, Cham, 2017.

[60] Ahmed, Uzair, et al. "Transfer learning and meta classification based deep churn prediction system for telecom industry." arXiv preprint arXiv:1901.06091 (2019).

[61] Nguyen, Nam N., and Anh T. Duong. "Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem." J. Adv. Inf. Technol. Vol 12 (2021): 1-7.

[62] Sarnen, Fioni, Suyanto Suyanto, and Rita Rismala. "Lifelong Learning for Dynamic Churn Prediction." 2020 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2020.

[63] Wu, S. (2013). A review on coarse warranty data and analysis. Reliability Engineering & System Safety, 114, 1-11.

[64] Hansel, A. Jordan, R. Holzinger, P. Prazeller, W. Vogel Int. J. Mass Spectrom. Ion Process.,149/150 (1995), pp. 609-619

[65] W. Lindinger, A. Hansel, A. Jordan Int. J. Mass. Spectrom. Ion Procs., 173 (1998), pp. 191-241

[66] Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83(2), 83-90.

[67] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter, vol. 6, no.1, pp. 20–29.

[68] S. Ray, "Understanding Support Vector Machine(SVM) algorithm from examples (along with code)," AnalyticsVidhya.com, Sep. 2017.

[69] N. Donges, "A Complete Guide to the Random Forest Algorithm," BuiltIn.com, Sep. 2021.