Prediction of Customer Attrition in the Telecom Industry using Machine Learning

Anish Mahapatra

Student ID 944563

Under the supervision of

Karthick Kaliannan Neelamohan

Master of Science in Data Science

Liverpool John Moores University

MAY 2021

ABSTRACT

With the advent of increasing competition in the telecom industry, companies must retain customers to maximise profits. With an average rate of churn of 30%, customer retention policies affect the annual turnover drastically. The cost of customer churn to the telecom industry is about $10 billion per year globally. Studies show that customer acquisition cost is 5-10 times higher than the price of customer retention. Companies, on average, can lose 10-30% of their customer annually. Developing effective customer relationship management processes and consumer-centric policies can help reduce spend on customer relations. For this, one would need to understand and track customer behaviour to understand the indicators that make a customer likely to churn.

Harnessing valuable data for business intelligence to develop churn management strategies is a proven data-driven strategy. Machine learning models require modest computation power and can deliver high accuracy when it comes to predicting attrition.

This research intends to build a predictive framework that can predict churn accurately and identify behaviour patterns that indicate customer churn. The paper will showcase the performance of various machine learning algorithms and how the process can be optimised. The dataset to be used for this research paper is the IBM Watson Dataset on customer churn in the Telecom industry. Extensive feature selection, processing, and work with multiple hybrid models to predict churn accurately.

Keywords: Machine Learning, Churn, Telecom, Attrition, Classification, Data Science

Contents

DEDICATION

This dissertation is dedicated to my family, whose unyielding love, support and encouragement have inspired me to pursue and complete this research.

ACKNOWLEDGEMENTS

I would like to acknowledge Liverpool John Moores University for the opportunity to learn and obtain a renowned degree. I want to express my heartfelt gratitude to my thesis supervisor, Karthick Kaliannan Neelamohan, for his invaluable guidance. He has guided and encouraged me to be professional even when the going gets tough, and I am fortunate to have him as a mentor.

I would like to thank my committee members and mentors from Liverpool John Moores University for their patient advice and guidance through the research process.

Finally, I thank my family, who supported me with love and understanding. Without you, I could have never reached this current level of success. Thank you all for your unwavering support.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
AUC	Area under ROC Curve
CRM	Customer Relationship Management
EDA	Exploratory Data Analysis
GBM	Gradient Boosting Machine
GSA	Gravitational Search Algorithm
IQR	Interquartile Range
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
LIME	Locally Interpretable Model-Agnostic Explanations
PPforest	Projection Pursuit Random Forest
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-NC	Synthetic Minority Over-sampling Technique for Nominal and Continuous features
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

CHAPTER 1: INTRODUCTION

With the increase in the number of options consumers have in the Digital Age, for a company to be successful, it is vital to keep costs low and profits high. One of the most effective ways to do this is to retain the existing customer base and focus the remaining budget on acquiring new customers.

1.1 Background of the Study

With the increase in the number of options consumers have in the Digital Age, for a company to be successful, it is vital to keep costs low and profits high. One of the most effective ways to do this is to retain the existing customer base and focus the remaining budget on acquiring new customers. The retention of the existing customer base in a focused and systemic manner is to be done, or its bottom line can be affected. A targeted way to approach the end goal of customer retention is to flag customers that have a high probability of churn. Based on customer behaviour and attributes, the likelihood to churn customers can be flagged, and targeted campaigns can be run to retain customers (Jain et al., 2021).

1.1.1 Churn Analysis in the Telecom Industry

The ability to retain customers showcases the company's ability to run the business. With the digital age, where everything is online, any business needs to virtually understand customer behaviour and mentality. The cost of customer churn in the Telecom Industry is approximately $10 billion annually (Castanedo et al., 2014). Customer acquisition costs are higher than customer retention by 700%; if customer retention rates were increased by just 5%, profits could see an increase from 25% to even 95% (Hadden et al., 2006). For a company to be profitable, it is thus essential to take pre-emptive action to retain customers that may churn. Churn is defined as customers who stop using their specific services and plans for long periods. Churn can occur due to various reasons and can be broadly classified into voluntary and involuntary churn.

In this post-pandemic age, where virtual presence via calls and the internet is the top priority, customers are trying to reduce their monthly expenditure month to month. Competitors are employing low prices or value-add services to get consumers to switch telecom operators. After acquiring a significant customer base, the companies monetise their customer base and profit in the long term (Jain et al., 2021). The companies that identify the segment of customers that are likely to leave and run targeted campaigns to showcase more value in their current offerings at a minimal budget are the ones that will be successful in the long run.

1.1.2 Flagging customers and retention policies

As service providers contend for a customer's rights, customers are free to choose a service provider from an ever-increasing set of corporations. This increase in competition has led customers to expect tailor-made products at a fraction of the price (Kuo et al., 2009). Churned customers move from one service provider to another (Ahmad et al., n.d.) (Andrews, 2019). Customer churn can be due to the non-satisfaction of current services, better offerings from other service providers, new industry trends and lifestyle changes. Companies use retention strategies (Jahromi et al., 2014) to maximise customer lifetime value by increasing the associated tenure. For telecom companies to reduce churn, it is vital to analyse and predict key performance indicators to identify high-risk customers, estimated time to attrite and likelihood to churn.

The learnings from multiple such experiments have been introduced as deployable machine learning algorithms that have been iterated and refined based on the evolving need to flag prone patrons more accurately. The choice of the techniques to utilise will depend on the model's performance on the selected dataset, be it meta-heuristic, data mining, machine learning or even deep-learning techniques. In the customer's behaviour patterns, there is likely to be a few significant indicators as to why the customer is willing to take the active step of moving across service providers. Identifying attributes that indicate if a customer is likely to churn in our methodology will be made through this research. Identifying the right attributes from the model will improve interpretability and help the customer relationship management move from a reactive to a proactive approach to increase customer retention rate.

1.2 Struggles of the Telecom Industry

The telecom industry has been struggling for years now. Telecom businesses have struggled to launch 5.58 products annually. The Huthwaite study shows that telecom companies have at least a new product failure annually – costing companies millions of dollars annually. Rather than developing strategies that meet evolving customer needs, telecom operators follow the traditional cycle of setting up networks, building cross-channel presence, and offering revamped plans. The losses, as seen by the industry, highlights the fundamental flaw in the approach. A study by Capgemini showed that most companies showed a Net Promoter Score between zero and negative (Why is the telecom industry struggling with product success?, 2021). The telecom industry is rife with disruption in all areas. The pandemic has changed how everyday communication supplements and enhances discussion between customers and brands.

Figure 1: Most significant challenges faced by the industry
(Digital transformation for 2020 and beyond eight telco considerations, 2021)

Disruptive competition is the primary reason why telecom operators are struggling globally. Customer attrition is the main reason to track at-risk customers that may churn and target programs to retain them. This targeted effort will help retain customers and ultimately increase the telecom

company's profits by employing churn prediction strategies.

1.3 Problem Statement

The reduction of attrition of customers from a company is vital to a company's bottom line. To maintain a good market share in the competitive telecom industry, understand and tackle the root cause of why a customer might shift their service provider. This research will help telecom companies leverage their existing consumer database to predict and actively target campaigns to customers likely to churn. The machine learning methodology employed can be personalised to the use-case based on the operator. When a suitable set of machine learning algorithms run on a newer dataset, the model's evaluation metrics can be monitored, and high-risk customers can be appropriately targeted.

The recommended model's primary users will be telecom conglomerates that wish to reduce customer attrition and improve their profitability in the market. This needs to be done, keeping in mind overhead costs. The set cadence and the hardware resources used for the same will be optimised to keep overhead costs nominal.

1.4 Aim and Objectives

The paper aims to develop a trustworthy and interpretable model that will predict the customers that will churn from a Telecom Company based on historical customer telecom data. The identification of the customers that churn will aid telecom companies in significantly reducing expenditure on customer relations.

The objectives of the research are based on the above aim and are as follows:

- To analyse the relationship and visualise patterns of customer behaviour to indicate to the telecom company if a customer is going to churn
- To suggest suitable feature engineering steps to extract the most value from the data, including picking the most significant features
- To find appropriate balancing techniques to enhance the model performance on the dataset
- To compare the classification or predictive models to identify the most accurate model to

determine the customers that will churn

- To understand the factors and behaviour of consumers that leads to customer attrition in the telecom industry
- To evaluate the performance of the models to identify the appropriate models

1.5 Research Questions

The following research questions have been formulated based on the literature review done so far in the field of customer churn:

- Is there a clear conclusion regarding the best overall modelling approach, be it classical machine learning or more complicated algorithms?
- Does the presence of multicollinearity, outliers, or missing values in the training data impact customer churn prediction accuracy?
- Do techniques such as hyperparameter tuning result in significantly better models?
- Can balancing techniques be suggested to increase the accuracy of the model?
- Are the results obtained from interpretable models reliable?
- Do statistically significant features mean that the business can take actionable insights directly?

1.6 Scope of the Study

Due to the limitation of the time frame in this research, the scope of the study will be limited to the below points:

- The data for the study has directly been obtained from the authorised source, and data validation will not be part of this research
- The research will include the development and evaluation of various machine learning

algorithms. The latest algorithms such as Neural Networks and Deep learning will not be considered as a part of this study due to a lack of resources and time

- The study will limit the use of classification algorithms such as logistic regression, decision tree, K-nearest Neighbour as a part of interpretable models, whereas random forest, support vector machine, gradient boosting, and XGBoost will be leveraged as black-box models for this study
- The focus of the research is on interpretable models. If time permits, an attempt to use other models to perform customer attrition analysis can be made

1.7 Significance of the Study

The research contributes to explain and interpret various predictive models to support decision-making and increase the company's bottom line by flagging customers that are going to churn. This will help the telecom company allocate the optimal budget and effort directed at customers that are likely to churn by running targeted campaigns. The sales team will be able to offer value add-ons to high-risk and high-value customers. This can help the company recognise its customers' pain points and ultimately help in fundamental policy changes that can increase the overall profit. With the recent struggles of the telecom companies becoming dire where the top companies are wiping out or acquiring the competition, telecom operators have to maintain a solid customer base to remain steadfast in this fiercely competitive environment.

The conventional approach of going by mere observations of senior folks has dissipated over the years. The companies that make effective decisions concerning future-facing strategies do so with the backing of their data. Predictive frameworks that can predict the customers that are likely to churn can change the game. Adopting companies effectively, these machine learning systems give companies a headstart with churn management strategies, but they also become better and more effective with time as the database and learning can leverage exponentially.

1.8 Structure of Study

The structure of the study is as follows. Chapter 1 discusses the background of the Customer Churn Analysis in the Telecom Industry. The study's aim and objectives and the research questions are discussed in Section 1.3 and Section 1.4. The study's significance to the Telecom Industry is discussed in Section 1.6 and contributes to identifying churn as a driver for business growth.

Chapter 2 has been structured to state the telecom industry's theoretical understanding and highlight its work to identify customer attrition. Analytics and visualisation play a pivotal role in performing predictive modelling on telecom data; this has been highlighted in Section 2.4 to understand how machine learning is being used to identify customers at a high attrition risk. Feature engineering and visualisation techniques for exploratory data analysis have also been discussed in Section 2.5, followed by a detailed review of related Customer Churn and Telecom research papers in Section 2.6. Discussion on the literature survey carried out is done in Section 2.7, along with the summary of the work carried out in Chapter 2 is done in Section 2.8 to conclude.

Components of Chapter 3 discusses the research methodology and the proposed research framework for the dissertation. The study's framework is described under research design to present the proposed model's approach through the steps of data selection, data preprocessing, data transformation, data visualisation, class balancing, model building, model evaluation, and model deployment in the subsequent sub-sections under Section 3.2. Section 3.3 explains the proposed model to be employed based on the experiments carried out. Finally, the classification model to evaluate the customers at a high risk of churn in the telecom industry and the evaluation methods and subsequent steps is discussed in Section 3.4, the summary.

CHAPTER 2: LITERATURE REVIEW

A thorough survey of the research and work done in customer attrition in the telecom industry will understand more about the telecom industry's nuances. This literature review will set the baseline to understand the expected standard to implement a robust classification model to predict customers' high risk of churn in the telecom industry. The approaches used by the authors range from using single machine learning models, meta-heuristic models, hybrid models, data mining techniques and even social methods (Oskarsdottir et al., 2016). Weightage for conventional methods that solved the problem of churn has been given along with the novel methods that solve the problem of churn.

With the advent of massive investments from telecom operators in this internet age, both old and new conglomerates globally, the market is the most competitive it has been in decades. The literature review will focus on reducing customer churn and the telecom industry's ongoing trends and how data analytics affect the telecom industry. Customers have moved from expecting just the cheapest plans; the average customer now expects to have tailor-made plans and solutions at a fraction of the cost that their monthly bill used to be (Umayaparvathi and Iyakutti, 2016).

Customers no longer need to stick to a monthly commitment of a subscribed plan; they can quickly get the benefits of the company's infrastructure within minimal commitments using a prepaid plan rather than a postpaid one. There can be many reasons why a customer can churn. On average, a telecom company loses 30% of its customer base annually; of this, not all customers can be stopped from churning (Umayaparvathi and Iyakutti, 2016). There are classes of customers that leave voluntarily and involuntarily; among the churners that leave voluntarily, there is a further bifurcation of those that attrite deliberately and incidentally.

The ideology that all customers that churn are the same does not hold when it comes to real-world analysis. In our literature survey, identify the different types of churners that exist will be undertaken. The visualisation in Figure 2 showcases a tree-based visualisation to showcase the same. In this literature review, the focus will be on the set of churners that churn voluntarily; it is difficult to flag whether the churn was incidental or deliberate every time.

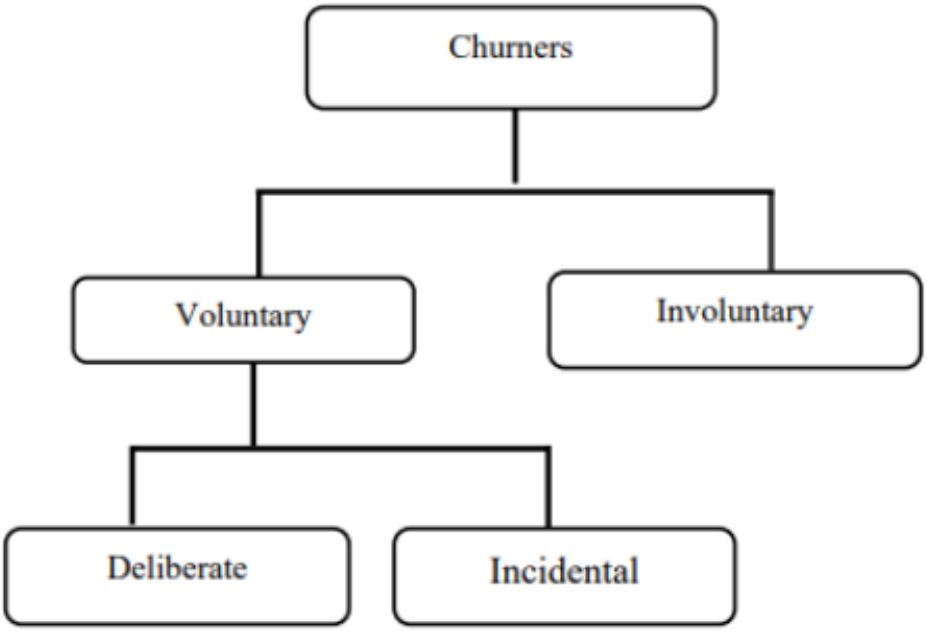


Figure 2: Types of Churners (Saraswat, S. & Tiwari, 2018)

With the above visualisation from the authors, the set of customers who undergo voluntary attrition can be identified as the ones to focus on in our literature survey. This understanding will help understand the customers' behavioural patterns that churn voluntarily are so that customers can be selectively profiled and targeted to get higher accuracy.

2.1 Introduction

For the literature review, having a proper structure for our analysis is critical when dealing with the telecom industry's churn. In section 2.2, there will be a focus on the telecom industry and the data-driven analytics driving the industry. This will explain how critical it is to flag customers and how designing custom campaigns for this segment of customers can increase certain companies' bottom line and profitability. Section 2.3 deep dives into customer attrition in the telecom industry and how this is a significant driver for the drain in finances and telecom operators' stability globally. The following section will understand how companies are leveraging predictive modelling in customer churn attrition and the models and methodologies used to keep profitability up.

Here, the models will be analysed in-depth, and the methodology and ideas behind the working of

predictive frameworks. Leveraging our learnings from the above sections in how visual analytics is also being used to visualise large sets of data. Before proceeding to the related research publications, understanding more about the metrics and formulations that authors have used in the literature survey; will help leverage the summary table and the steps of data preprocessing, feature engineering, models applied and results of the modelling efforts as displayed. This table will summarise all of our learnings in a quick referential format for future publications to leverage the latest in the field. In Section 2.7, the discussion of our learnings from related work and the previous sections showcases how the components of an efficient predictive framework for customer churn analysis can be set up for our use case. Finally, in the last section, a summary of all of the analysis to understand how telecom operators can leverage data science and machine learning to predict the segment of customers at a high risk of voluntary churn will be done.

Following our learning in the sections below, an analysis of the literature survey gaps in the authors' recent work will be done. Bridging the gap in terms of data preprocessing, feature selection, visual analytics, modelling and evaluation of the holistic predictive framework will help build better practices for telecom operators going ahead. It is also essential to have a good spread of recent literature and review the papers that have impacted customer attrition in the telecom industry.

2.2 Data Analytics in the Telecom Industry

The telecom industry might seem like it is booming with the internet age, but that is not the case for most telecom operators. The telecom industry has a heavy dependency on external factors riddled with serious debt complications in the industry. The investments range from building infrastructure that can carry lines across the country, investments in the latest technologies that will help enable the latest in voice and internet technology like 5G, money spent on buying bandwidth frequencies. Additionally, the cost of upkeep and maintenance of a vast network can be grossly expensive as operators have to pay rents, keep up the set infrastructure, lobby the government, provide customer service, and deal with the unexpected changes in the ecosystem.

For all of these risks that telecom operators take to run a business, various models can be followed to ensure a steady income. Since a Business to Consumer (B2C) model is high-risk and high-reward, ensuring that there are guaranteed paying customers at the end of the month can be crucial

whilst maintaining steadfast while holding up market share in the space. The telecom sector's riskiest customers are prepaid, as it is challenging to flag if they are active or not because different segments of customers have different behavioural patterns. The telecom industry has truly earned its place as the backbone of our country and even the economy. It is exceedingly difficult to imagine a world in which a call, message or communication with someone at a fraction of the cost paid for the same service about just a decade ago. The rate of mobile and internet penetration in third-world countries is increasing exponentially every day; this leads to a whole host of some of the largest companies in the world backing up telecom operators to be able to acquire a customer base as loyal and dedicated as possible so that this cash-burn can be leveraged to profit in the future.

To have a higher stake in the Industrial Revolution 4.0, telecom operators need to move away from a conventional customer retention approach. A customer is no longer associated with a company because only one service exists in the area. The telecom operators should improve their CRM infrastructure to move away from merely fulfilling an internal need to a full-fledged ecosystem with value-proposition not just for the end-customers but also for all stakeholders involved telecom pipeline. A happy customer is a loyal one. Attracting new customers might seem like an attractive way to grow market share. However, the experienced players in the market know that the secret to being profitable in the long run is two-fold, first, focusing on the retention of customers, especially the high-value customers and second, being able to leverage the existing database that is a trove of customers who are likely to come back to the company if courted aptly. Gaining new customers is 5 to 10 times more expensive than keeping existing customers loyal (Wassouf et al., n.d.; Ebrah and Elnasir, 2019). The recommended method to effectively implement a data science predictive framework is to scale and leverage it to make a robust and effective model as a custom-designed use case. A custom solution is an exciting ask in terms of strategy for leadership as one would like to invest less effort on a proof of concept and leverage the long-term benefits for the company if the project can help increase the profits in the long term. The idea of investing in the future to move from a model that reduces loss to increases profit is a game-changer.

Several low-code or no-code tools are being used to start build proof of concept projects; the reality is that implementation is vital. Models need to focus on explainability and usage of metrics rather than a black-box approach. This is critical to building a solid data science muscle within the organisation because it may be easier and even faster to build a proof of concept with a ready-made

tool or technology. However, when it comes to scaling the exact implementation at an org-wide level whilst keeping the overhead costs minimal, it can get complicated. Implementing a tool on a large scale has one of two problems. First, it may be costly to get multiple licences or pass large amounts of data in the tool. Secondly, there may be a black-box approach for the data problems, so modifying the code may not be feasible. Tools such as RapidMiner that can leverage explainable models that can be understood by senior management can be a good starting point (Halibas et al., 2019) for proof of concept implementations. Developing an in-house custom analytics solution is the long-term aim of a company and building data science competencies. Most companies require a custom setup for churn analysis on account of different datasets, technology stacks, databases and overall requirements (Fonseca Coelho, n.d.). Understanding the requirement for the cadence of forecasting based on the model selected is also a vital area of research to move from a batch-processing system to a more real-time system (Tamuka and Sibanda, 2021). Depending on the complexity of requirements and budget, a cloud-based flexible architecture can also be set up.

2.3 Customer Attrition in the Telecom Industry

Understanding the customer is an integral part of whether a customer gets to keep an existing customer or not. Deciding the budget allocation at the start of the fiscal cycle is the deciding factor in its culture. Let us look at a company where most of its cash burn will be focused on discounts to attract new customers. Is it going to be spent on marketing mix to build brand equity that can be leveraged later on in the future, or is a company going to majorly focus its budget distribution on customer service to retain a high number of high-value customers. Understanding all of a customer's nuances will help predict if a customer is looking to churn voluntarily. Here, hundreds or even thousands of attributes on the customer can be leveraged to perform churn analytics. Choosing the right set of features that can help in this prediction is an area of research in itself. The right set of features is dependent on the company's dataset as a more extensive set of data from the company can help high-risk flag customers more accurately (Fonseca Coelho, n.d.).

There is one common element in the literature reviewed; there are always certain behavioural traits of a customer that can be identified as a customer trend that is to churn. Customers tend to move across telecom operators for several reasons, with countries enabling inter-operator portability

globally. It is easier for a customer to move if they are dissatisfied with the services of a company. There are a few factors with the digital age to determine how likely a customer is to churn. If a customer has enabled auto-pay for their bills, if a customer has been associated for many years, if a customer has internet services and has opted in for a host of other services that their everyday life or family's life is dependent on, the customer is less likely to churn.

The literature review observation is that a customer should have the least amount of friction while getting into the services offered. This ease of movement and a tie-in to other services offered at multiple fronts will increase customer loyalty. When a customer is likely to move across, the company should have an open communication line with effective teams on multiple touchpoints. A surprising find is that the main reason a customer moves across telecom operators is not due to a new promotion/offer. Instead, the primary reason a customer moves across operators is dissatisfaction with current services (Wassouf et al., n.d.; Ebrah and Elnasir, 2019). Identifying the customers that are dissatisfied with the current services, via several tickets raised for a unique customer id, and the number of calls gives an account of the satisfaction to a segment of workers in the company, the satisfaction scores will increase and thus, lead to a reduced rate of churn.

The focus should not only be given to the data that is collected recently, but also to the already existing database of customers; setting up various focus groups for the different segment of users within the company will help us understand what the deciding factors for which a customer is likely to churn are. Being able to leverage this understanding from the dataset is a deciding factor in retaining customers. It is not merely identifying the set of customers that are at a high risk of churn; if timed right with the right kind of targeted campaign, there is a high chance that even if the telecom operator was to take a slight loss in the form of additional discounts offered to the high-risk customer in the short term, the cost could be recovered and a profit can be made in the long-term. Various strategies can be employed based on our learnings from the model. However, the suggestions of the personnel involved directly with the customer and customer database must be taken into account as they have more real-world context when it comes to customer behaviour and sentiment.

2.4 Predictive Modelling in Customer Churn Analysis

A predictive modelling framework for data science is an involved process with a list of tasks that can be understood through the literature survey. In this section, let us understand the details of the supervised machine learning techniques. Customer churn analytics in the telecom industry aims to flag the segment of customers likely to churn and some confidence. This is a classification problem to predict one of two things; if a customer is going to churn or not. There are different methods to do this, and in the literature review below, an understanding of supervised machine learning algorithms will be given. The fusion of multilayer features uses a framework of complementary fusion by employing feature construction and feature factorisation to improve churn prediction accuracy. This approach resolved the problem of high dimensionality and imbalance of data. Feature selection was also attempted, which led to the reappearance of imbalanced data (Ahmed and Linen, 2017). Novel methods of engineering the data was also used in the research where tokenisation was used for categorical attributes and standardisation was used to standardise numerical attributes (Momin et al., 2020).

Novel methods for feature selection, such as gravitational search algorithm (Lalwani et al., 2017), have been used. GSA helps reduce the dimensionality of the data and improves the data's accuracy by optimising the search for significant features (Lalwani et al., 2021). Methods for preprocessing data tasks such as missing value imputation have developed well over the last few years. A method used to explore and perform multiple missing value imputations to fill up quantitative variables that suffer from an uneven distribution is Predictive Mean Matching (Mahdi et al., 2020). While some methods are agnostic to the type of data, specific methods assess numeric variables' uneven distribution using a logarithmic transformation (Tamuka and Sibanda, 2021). Categorical variables used in telecom datasets are also converted to numeric variables using techniques such as label encoding or one-hot encoding (Agrawal, 2018). The popular methods used to handle categorical variables are label encoding and one-hot encoding. With larger datasets, the issue of high dimensionality is a problem – for this, some of the authors with large datasets have worked with sparse matrices or have leveraged dimensionality reduction techniques such as principal component analysis. Some of the authors have leveraged modelling techniques that work with categorical variables, continuous and discrete variables.

2.5 Visual Analytics in Telecom

For data of any form to be leveraged, understanding the dataset is fundamental. One of the fastest ways to perform exploratory data analysis is to visualise the data. Figure 3 illustrates the relationship between data, visualisation and models with the intermediary knowledge gained from visual analytics (Yuan et al., 2021).

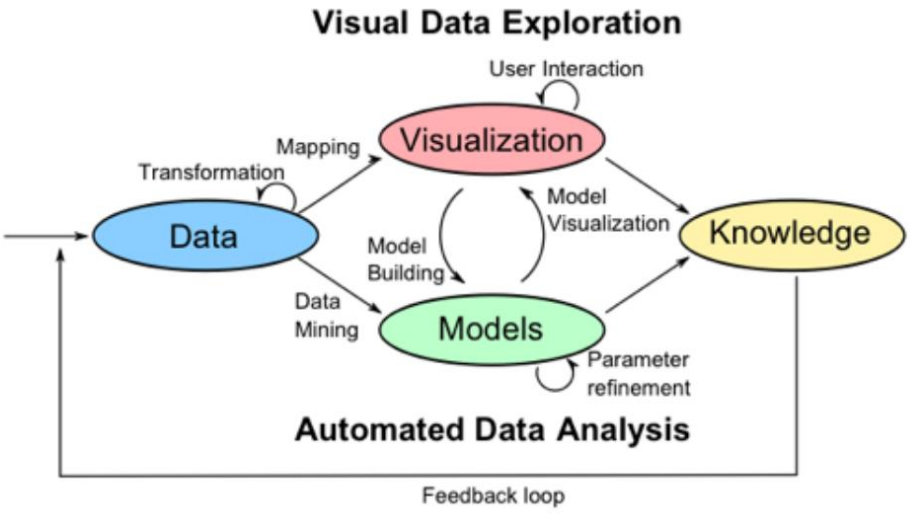


Figure 3: Visual Data Exploration

Being able to perform automated data analysis involves using visual cues is the essence of visual data exploration. Based on the visualisations formed, further understanding of row-level data is developed. When data transformation is performed, visualising the data post-processing helps understand if further data manipulation is required before the modelling phase. For instance, feature importance using a method out of advanced regression, XGBoost or random forest has been calculated. At the same time, the visualization and sum of feature importance scores obtained for features visually, the identification of the top feature using a bar chart with an indication of the top features to choose from for the next steps. Using multiple methods of visualising the features' distribution, the variance of the data points and the other analysis helps us make decisions for the next steps.

2.6 Related Research Publications

This section will provide a review of how data analytics is used in the telecom industry to identify customers at a high risk of attrition and the data-driven processes followed to set the baseline of the techniques carried out in the industry far. Section 2.6.1 and Section 2.6.2 will focus on feature engineering for the data and handle class imbalance. Efficiently carrying out data preprocessing will help us obtain better results in the following stages of implementing machine learning and validation via k-fold cross-validation. In the literature review, an understanding of the evaluation methodology used to assess the models' performance will be analysed. Section 2.6.3 will review the evaluation metrics used for classification (Karimi et al., 2021).

2.6.1 Feature Engineering for Telecom Datasets

Feature engineering is a critical step in the data science flow. Based on the analysis of the existing techniques implemented by authors, the significant features from the dataset that can affect churn are picked or generate new features from the existing set of attributes that can help predict churn better. When the authors have set out to perform feature engineering, it is only done keeping the dataset and the predicted model's accuracy in mind. When performing feature engineering on a dataset, another critical task is to identify the attributes that have the highest impact on the target variable. This can be done by leveraging rigorous algorithms or even RapidMiner and Azure ML Studio (Thontirawong and Chinchanachokchai, 2021).

Feature selection is made using attribute scoring methods such as random forest, xgboost and advanced regression, based on which the less significant values are discarded and the effect on the accuracy of churn prediction is observed. Techniques that leverage the correlation with the target variable are also used; the correlation matrix operator (Halibas et al., 2019) performs feature selection, and less significant features were discarded. The scoring of features based on their relation to the target variable indicates the variable's feature importance in consideration. Since the data has been generated from various sources and periods, standardisation of the data to compare different sets effectively helps the author decide the essential features based on the correlation matrix operator. The operator produces a pairwise table of correlation coefficients. This output was

then fed into a Gradient Boosted Tree model before and after oversampling, and the results were tested over multiple iterations and different hold-out conditions. For evaluation, F-measure, %Recall, %Precision, %Classification Error and %Accuracy were used to assess the models' performance. The experiments showed that gradient Boosted Trees outperformed the rest of the classifiers in all performance criteria. One interesting thing to note here is to all the classifiers tested resulted in an accuracy of over 70% (Halibas et al., 2019). All of the classifiers also showcased a much better performance once the oversampling technique was applied, which implies that class balancing enhances classifiers' performance in this case.

2.6.2 Handling Class Imbalance in Machine Learning

Class imbalance is a problem in machine learning, particularly classification, where there is an unequal distribution of classes in the dataset. For instance, there can be an uneven distribution of churned and non-churned customers (Thabtah et al., 2020). Synthetic Minority Over-Sampling Technique (SMOTE) is a method that some researchers have used to reduce the data imbalance (Induja and Eswaramurthy, 2015). The other methods the researchers have used to tackle the class imbalance problem in telecom based datasets are undersampling or oversampling (Ambildhuke et al., 2021). Random oversampling and undersampling are two of the more straightforward techniques that are used to train the model. Another method that used to have greater control over the class balancing process is stratified sampling. Stratified sampling lets the user select the classes which should be over or undersampled and based on the ratio. The model can be trained on a balanced set of the data. A modification of the conventional method, undersampling-boost, is also used to handle class imbalance (Saonard, 2020).

The methods that incorporate Synthetic Minority Oversampling Technique have been observed to have better results when various classifiers have been trained on the balanced dataset. Some of the other methods to deal with class imbalance include Adaptive Synthetic (ADASYN) and Borderline Smote (Induja and Eswaramurthy, 2015). ADASYN generates synthetic data and does not replicate the minority data. Instead, it generates new data based on the characteristics of the minority data. Class balancing is a method a few authors have leveraged to get enhanced model performance compared to those that do not use class balancing techniques.

2.6.3 Implementation of a predictive framework

Through this literature survey, various machine learning models have been assessed. Models range from individual machine learning classification models like logistic regression, decision tree, random forest, Naïve Bayes, k-nearest neighbour. The algorithm support vector machine gives better results as compared to the other machine learning models. Hybrid models using boosting and bagging models such as AdaBoost, Gradient Boosted Trees, CatBoost, and XGBoost provide incremental accuracy improvements (Labhsetwar, n.d.; Sharma et al., 2020; Lalwani et al., 2021). Churn prediction is better with hybrid algorithms than single algorithms (Ahmed and Maheswari, 2017). All of the classifiers were able to achieve accuracy greater than 70%.

Oversampling is observed to be an accuracy booster (Halibas et al., 2019). Papers that implemented deep learning in artificial neural networks were seen to have accuracy similar to that of the other machine learning algorithms (Agrawal, 2018; Oka and Arifin, 2020). Algorithms such as Artificial Bee Colony Neural Networks has also been implemented to predict churn in the telecommunication sector (Priyanka Paliwal and Divya Kumar, 2017). Interpretable models via RapidMiner using the SHapely Additive exPlanations (SHAP) and Local Interpretable Model-agnostic explanations (LIME) (Kriti, 2019). Model explainability is a fundamental skill that has been getting popular in the industry where the result and the logic should be explained.

A factor that has been considered keeping in purview the task to run the real-world models is the processing time comparison. In this paper (Oka and Arifin, 2020), the author showcases through visualisation the processing time that different models take on the IBM Watson customer churn dataset. The visualisation showcases that deep neural networks take the least processing time with just 68 seconds, whereas the more frequently models, such as XGBoost with 175 seconds and the highest with random forest taking 529 seconds., where random forest have an accuracy of about 80.6%. Another author worked on a survival analysis of the telecom industry based on critical total losses. It depended on the survival probability that the company defined and depended on its strategy, position, and situation in the market. The models used were the semi-parametric cox model proportional model, parametric Weibull and log-normal survival models (Havrylovych and Nataliia Kuznietsova, 2019). Per the analysis, the log-normal model was found to be the best model in this scenario. Projection Pursuit Random Forest (PPforest) based on Linear Discriminant

Analysis, Support Vector Machine provided good accuracy and AUC values. This was done with six sets of data with the IBM Telecom dataset giving the best results for the PPforest based on LDA (Mahdi et al., 2020).

2.6.4 Reviews of Evaluation Metrics for Classification

Various evaluation metrics can be used for the classification. Deciding on the right metrics to use is a part of effectively assessing classification machine, learning models. Some of the evaluation metrics used through the literature review are AUC, Accuracy and F-Score. Another way to deep-dive into the model's performance is to leverage the confusion matrix to understand more evaluation metrics such as precision, recall, type 1 error and type 2 error. A standardised evaluation method across machine learning algorithms will help decide customer churn's recommended model (Mukhopadhyay et al., 2021).

There are different ways of evaluating the performance of a classifier. The methods used are the ROC curve or derivatives of the confusion matrix, such as F-Score. Understanding the confusion matrix's derivation is vital to decipher many of the results when machine learning models are involved. Let us go over a few of the standard metrics in the below sections to understand the metrics used for evaluation. Evaluation of the classifiers' performance in the below section (Halibas et al., 2019) will occur.

- True Negative (TN): This is an indication that the model successfully predicted the expected outcome – predicted 0
- False Negative (FN): This is an indication that the model has failed to predict the expected outcome – predicted 0 instead of 1
- False Positive (FP): This is an indication that the model predicted the opposite of the expected outcome – predicted 1 instead of 0
- True Positive (TP): This is an indication that the model successfully predicted the outcome as expected – predicted 1

Accuracy is defined as the ratio of all correct predictions made to total predictions made. It is obtained by dividing the correct cases predicted by the total number of cases present

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision is defined as the ratio of correct positive predictions out of all the model's positive predictions. It is computed by dividing the number of true positives by the number of true positives and false positives.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The recall is defined as the number of correct positive predictions made from all positive predictions possible in the overall setting. It is defined as the number of true positives by the number of true positives and false positives.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

F-Measure is defined as a combination via a harmonic mean of precision and recall. The F1 score is a way to express both precision and recall scores as one metric.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The AUC or area under the curve is a recommended metric for binary classification. The recommendation is because AUC is not sensitive to imbalanced classes. Many papers have leveraged AUC in place of accuracy due to this reason. The AUC score varies from 0 to 1, and a score of 1 is considered a perfect score. The curve is plotted as a true positive versus a false positive.

Another factor some papers have bought up is that many authors focus only on improving the model's accuracy. That is, authors are focused more on being able to get as many churned

customers. The author (Tuck et al., 2020) proposes that just as much effort needs to reduce the machine learning algorithms' error or misclassification rate. The error rate can be viewed as an additional method to be able to evaluate a model effectively.

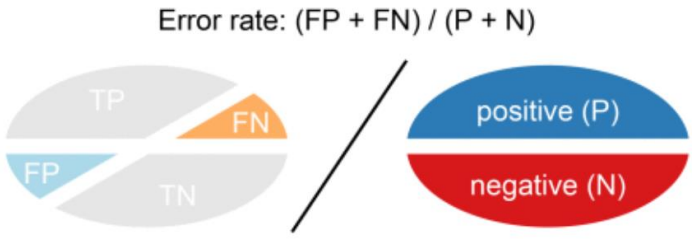


Figure 4: Visual Representation of Error Rate

A combination of the evaluation metrics is the ones that were used in the literature review for the evaluation of predictive models for classifiers. The model is to be evaluated, keeping the metrics in mind to understand the performance based on the rest of the metrics, such as specificity and sensitivity.

2.6.5 Summary of Literature Review

The telecom industry is a competitive space, and authors have been trying to solve customer attrition for years. There are multiple ways to tackle churn and as machine learning advances, so do the methods by which a customer that may leave is flagged. The data present within a company is a golden opportunity to build a robust model that can be leveraged to increase profitability. There have been some stellar research in classification, from single machine learning models to hybrid models (Induja and Eswaramurthy, 2015). Recent literature has a significant impact on the modelling of customer attrition in the telecom industry. Being able to view all of the work in the form of the below table gives us an overview of the significant work that has been done to support the same. From the above section, more importance can be given to feature engineering, as most papers have used more conventional methods. Similarly, for class balancing, instead of opting for simple random oversampling techniques, other structured oversampling techniques can be leveraged for the next steps.

2.7 Discussion

From the above literature review carried out, there are various ways to identify the customers at a high risk of churn through machine learning. The problem's approach varies from focusing on data mining techniques to select the right set of attributes, valuable data preprocessing and efficient feature selection. This effort to obtain the right set of data to feed results in choosing a simpler model to perform classification; thus, saving computation time and keeping the overall computational requirements minimal, saving companies' overhead costs.

The other approach followed is to rely on the machine learning model to flag the customers that are likely to churn effectively. The data size plays a considerable role; if the data's size is limited, focusing on the machine learning algorithm is more sensible, whereas a hybrid approach can be experimented with for larger datasets. The literature on deep learning suggests that even though a neural network approach works for some cases, the model's performance is not significantly better to opt-in for deep learning models exclusively. It is a common misconception that deep learning models perform better than machine learning models in all use-cases. From the literature review, the understanding for telecom use cases studied where a predictive framework based on a machine learning or deep learning framework has been made, hybrid machine learning models and a balancing technique have given the best results.

Different feature selection techniques, in turn, have resulted in a different set of features being selected for different algorithms. Exploring more feature engineering techniques and summarising our results, so the observed and latent relationships of the features with the target variables are considered will aid future implementation. Imputation of the data is also a step where some authors have taken advanced methods such as logarithmic transformations and predictive mean matching for imputing missing data rather than the conventional methods to impute the missing values with mean, median or mode (Tamuka and Sibanda, 2021). This approach, along with oversampling techniques, has given some of the best results per the literature survey. In Chapter 3, this is the approach to take inspiration for, along with more advanced feature selection methods

Table 2.7.1: Literature Review for IBM Watson Telecom Dataset

Authors	Year	Feature Engineering	Model
(Tamuka and Sibanda, 2021)	2021	Feature Importance, Logarithmic Transformation	**Accuracy**: Logistic Regression - 97.8%, Decision Tree - 78.3%, Random Forest - 79.2% **F1-Measure**: Logistic Regression - 97.8, Decision Tree - 77.9, Random Forest - 77.8
(Lalwani et al., 2021)	2021	*Phase 1*: Variance Analysis, Correlation Matrix, Outliers Removed *Phase 2*: Cleaning & Filtering *Phase 3*: Feature Selection using Gravitational Search Algorithm, Feature Importance	**AUC**: Logistic regression - 0.82, Logistic Regression (AdaBoost) - 0.78, Decision Tree - 0.83, Adaboost classifier - 0.84, Adaboost Classifier (Extra Tree) - 0.72, KNN classifier - 0.80, Random Forest - 0.82, Random Forest (AdaBoost) - 0.82, Naive Bayes (Gaussian) - 0.80, SVM Classifier Linear - 0.79, SVM Classifier Poly - 0.80, SVM (Adaboost) - 0.80, XGBoost - 0.84, CatBoost - 0.82
(Momin et al., 2020)	2020	Tokenisation, Standardisation	**Accuracy**: Logistic Regression - 78.87%, Naïve Bayes - 76.45%, Random Forest - 77.87%, Decision Trees - 73.05%, K-Nearest Neighbor - 79.86%, Artificial Neural Network - 82.83%

(Oka and Arifin, 2020)	2020	Label Encoding Binary Columns, Scaling Numerical Columns, Feature Importance	**Accuracy**: Random Forest - 77.87%, XGBoost - 76.45%, Deep Neural Network - 80.62% **AUC**: Random Forest 0.83, XGBoost 0.84, Deep Neural Network - 0.84
(Mahdi et al., 2020)	2020	PMM - Predictive Mean Matching for imputation	**Accuracy**: PPForest with LDA - 72%, PPForest with SVM - 75% **AUC**: PPForest with LDA - 0.67, PPForest with SVM - 0.73
(Ebrah and Elnasir, 2019)	2019	K-Cross Validation with hold-out (30%) method (k=10)	**Accuracy**: Naïve Bayes - 76%, SVM - 80%, Decision Tree - 76.3% **AUC**: Naïve Bayes - 0.82, SVM - 0.83, Decision Trees - 0.76
(Havrylovych and Nataliia Kuznietsova, 2019)	2019		Semiparametric Cox Proportional Model, Parametric Weibull, Log-normal survival model Best model: log-normal model
(Halibas et al., 2019)	2019	Feature Selection using Correlation Matrix Operator RapidMiner is used to perform feature selection	**AUC**: Gradient Boosted Trees (*before oversampling*) - 0.834, Gradient Boosted Trees (*after oversampling*) - 0.865, Generalised Linear Model - 0.841, Logistic Regression - 0.841

(Kriti, 2019)	2019	Feature Selection using XGBoost	**AUC**: XGBoost - 0.85, Random forest - 0.84, Decision Tree - 0.81 SHAP, LIME is used for Local interpretable model agnostic
(Hargreaves, 2019)	2019	Top 5 Significant features using Feature Selection XGBoost	Logistic Regression: Accuracy - 76.7% AUC - 0.767
(Pamina et al., 2019)	2019	Feature Selection - XGBoost Classifier	Accuracy: K-Nearest Neighbour - 0.754, Random Forest - 0.775, XGBoost - 0.798
(Induja and Eswaramurthy, 2015)	2019	Feature Selection	AUC: Random Forest *with RFE* - 0.96, ANN *with RFE* - 0.77
(Agrawal, 2018)	2018	One-Hot Encoding	Accuracy: ANN - 80.03%

From the above papers, it is understood that the focus is on either data processing or modelling. With novel preprocessing methods, such as predictive mean matching or gravitational search algorithm for processing to single, hybrid or advanced methods of forecasting for the predictive framework, the gap in the research is a paper that implements both. Trying novel methods of multiple feature selection on the telecom data, coupled with a robust predictive framework, seems to give the highest returns in model performance. Having observed a few scenarios when the data is over-engineered or refined beyond a point, overfitting of the data occurs on the training set, and the performance on the hold-out or test dataset is not as expected.

Many of the papers reviewed introduced feature engineering, but there is a gap in one way or the other. For instance, a lot of the papers have not carried out k-fold cross-validation on the data, even though the data that they are using is a small dataset, thus, risking the fact that the model might have a bias and may not be robust when the predictive framework is applied in other scenarios. The focus of some papers has been to try new algorithms to be able to increase accuracy. For the use-cases that have attempted to focus on the framework, a gap can be filled through our research methodology.

2.8 Summary

A whole host of machine learning models can be used for the use case of solving for the classification of high-risk customers. An excellent approach to try would be to focus on the machine learning approach and the data preprocessing. A few authors implemented class balancing techniques, and better accuracy was observed. Our approach will be made on all of the steps mentioned above of data preprocessing, missing value analysis, outlier analysis, variance analysis, k-fold cross-validation and class balancing techniques for phase 1. This will be followed by single machine learning algorithms and hybrid machine learning models in phase 2. Once the best models can be found for our use case, k-fold cross-validation will be performed to get the best generalised and robust model. This thorough literature review of the best the academic community offers has provided us with the baseline understanding before deciding the appropriate research methodology for our use case.

CHAPTER 3: RESEARCH METHODOLOGY

This chapter is dedicated to the research methodology to work with the IBM Watson Telecom dataset. From our learnings from the literature review and our understanding of the telecom business, flagging the segment of customers at a high risk of churn will occur. This chapter is dedicated to taking our learnings from the related research in data preprocessing, feature engineering, predictive framework and evaluation metrics and applying it to provide an accurate process flow to flag customers at a high risk of attrition.

3.1 Introduction

A baseline understanding of how to tackle the customer churn problem in the telecom industry from the literature review will help us decide the improvements that can be made. This section will set up the research methodology for tackling the use-case for our study. Section 3.1.1 and section 3.1.2 focuses on business understanding and data understanding. The research methodology follows this in section 3.2 that consists of data selection, data preprocessing, data transformation, data visualisation, class balancing, model building, model evaluation and model monitoring. This modelling will be proceeded by the proposed model in Section 3.3, ultimately followed by the summary in Section 3.4.

3.1.1 Business Understanding

The telecom industry is a highly competitive industry where customers can choose to move across operators if they believe they are getting more value with another service provider. Based on the customer's behaviour patterns, there are indicators to report if a customer might churn or not. Since the retention cost is much higher than customer acquisition, it is vital to identify the customers likely to churn and run targeted campaigns to retain the existing customer base. It was also observed that a reduction of customer attrition of 5% could lead to profit margins increasing from 25% to 95% (Hadden et al., 2006). In the telecom industry, where the approximated annual cost of customer attrition is $ 10 billion annually (Castanedo et al., 2014), and 30% of customers churn on

average, there is a substantial need to perform active targeting to retain the customer base.

3.1.2 Data Understanding

There are various data sources used to predict customer churn in the telecom industry through the literature survey. This research shall be using the IBM Watson Telecom churn data found on the Kaggle website derived from the IBM Cognos Analytics Community (Cognos Analytics - IBM Business Analytics Community, 2021). The telecom churn data consists of 7043 rows and 21 attributes at a customer-id level. The data combines numerical and categorical variables that can be used as feature variables to predict the target variable churn. Churn is indicated within the dataset as a "Yes" or a "No", indicating if a customer has churned or not churned respectively. This data presented is for the last month based on which predictions are to be made.

Each row in the telecom churn represents customer attributes used to describe the customer's behaviour. The data is unique at a Customer ID level with a high cardinality of 7043. The Total Charges column is uniquely distributed. There is an equal 50-50 distribution of male and female customers. As one would expect in the Churn column, there is an imbalance, with 27% of customers churning and 73% retention. This dataset has been collected over a month with a Kaggle Usability Score of 8.8 based on the provided metadata and various other factors, as mentioned in the website (Kaggle, 2018).

Let us understand the descriptive dataset statistics in detail. Here, let us analyse and understand the dataset better by deep driving into the statistics of each column:

- Customer ID: Unique Customer Id assigned to each customer (7043 unique values)
- Gender: Indicative of whether a customer is male or female
- Senior Citizen: Binary of whether the customer is a senior citizen or not
- Partner: Information on whether the customer has a partner or not
- Dependents: Indicative of whether the customer has dependents or not
- Tenure: Number of months the customer has stayed with the company

- Phone Service: Indicative of whether the customer uses the phone service or not
- Multiple Lines: Whether the customer has multiple lines or not
- Internet Service: Information regarding the internet service provider (DSL, Fiber optic, No)
- Online Security: Whether the customer has online security or not
- Online Backup: Whether the customer has opted in for Online Backup
- Device Protection: Whether the customer has open in for Device Protection Plan
- Technical Support: Whether the customer has requested Technical Support
- Streaming T.V.: Whether the customer has opted in for T.V. Streaming services
- Streaming Movies: Whether the customer has opted in for Streaming Movies services
- Contract: Whether the customer has opted for a monthly, annual or two-year plan
- Paperless Billing: Whether the customer has opted in for paperless billing
- Payment Method: Method of payment of the customer: Electronic check, Mailed check, Bank Transfer or Credit Card
- Monthly Charges: Monthly Charges of the customer
- Total Charges: The total charges of the customer
- Churn: Whether the customer has churned or not

From the above description, there is a deep understanding of the IBM Telecom Churn dataset's descriptive statistics used in this study. Eighteen features are categorical, two integer features and one feature of type float. The dataset has 7043 rows and 21 columns that describe customer behaviour. The dataset is taken over one month and will be used for analysis and predictive modelling in this study. The dataset range is also essential, including the summary statistics, to get a brief of the dataset.

3.2 Research Methodology

The following section contains the steps to perform predictive modelling to predict the customers with a high attrition risk. The steps followed are data selection, data preprocessing, data transformation, data visualisation, class balancing, model building, model evaluation and model deployment.

3.2.1 Data Selection

There were a few datasets to choose from when it comes to telecom data. The data selected is the IBM Watson Telco Customer Churn Data. The dataset is at an employee level with a usability score of 8.8. The dataset has information that can be leveraged at a customer level to identify customers likely to churn effectively.

The information obtained from the data can be broken down into four broad categories and is as follows (Ebrah and Elnasir, 2019):

- Services that the customer may be using such as streaming movies and tv, technical support, device protection, online backup and service, broadband services
- Account Information of the customer such as customer tenure, total costing, monthly charges, paperless billing, payment method
- Demographic information such as age, gender, information about dependents and partners
- The given data consists of multiple factors about the customers regarding lifestyle, behaviour in a Yes or No format that can be leveraged post-processing. It is presented in a .csv format with customer attributes information as metadata

Understanding the different segments of the data available will help us profile the various customer segments and their behaviour, which will, in turn, be able to accurately flag the set of behaviours that are indicative of customer churn for telecom operators.

3.2.2 Data Preprocessing

Now that the dataset is selected, let us proceed to understand the domain. Discussion on the Data Pre-processing steps that are to be implemented will ensure that the data is standardised as it is used in the following steps. A sense check of the telecom churn dataset is performed to understand if the import of the data and the dataset's encoding are per expectations. Once the data types of the features are noted, the shape of the data is checked to ensure the number of rows and columns is consistent per expectations. Focus is then directed on the columns that have at least one missing value. Once the attributes to consider are accounted for, the percentage of missing values column-wise is analysed. This will help us to decide the strategies to take for the next steps. Post missing value analysis will determine if all the columns or selected columns will be carried forward to the next step if columns must be dropped based on absent value percentage or employ methods such as mean imputation, mode imputation, deletion of rows and iterative imputation.

The percentage of missing values for each attribute after the missing-value analysis will help us understand the base dataset used before the next step of feature engineering. Outlier analysis is performed along with an analysis of the data's skewness to understand the feature's impact on customer churn. After understanding each features' distribution, univariate analysis is performed. This will help us understand and map out the inherent properties and distributions of each attribute. The bivariate analysis will then be performed on the data, ultimately followed by multivariate analysis to understand the features' direct and latent impact on the customer churn's target variable.

3.2.3 Data Transformation

The following successive steps to extract the most value from the dataset will be carried out based on the cleaned dataset. Steps such as one-hot encoding are applied to the categorical features. Besides this, features are derived from the existing dataset and feature engineer newer attributes. Based on the understanding of telecom's business, business rules and heuristical methods are applied to the business and derive new features. Performing efficient feature engineering will save us the hassle of running complicated models to get an accurate prediction.

This will make the machine learning pipeline easier to deploy, thus reducing the business

expenditure on hardware. Data visualisation here will play a crucial part here to be able to draw insights that might help to be able to derive more from the data. Using advanced Exploratory Data Analysis packages such as pandas profiling, Sweetviz and data prep to perform visualisation of the data; will give us a complete overview of the data. Mapping out and understanding the relationship of each numerical and categorical variable with churn will help us start identifying the attributes that might have a direct or latent impact on customer churn. After performing multicollinearity and variance inflation factor tests to understand the data's inherent properties, an analysis of the significant features will be selected for modelling. Additionally, the correlation scores for the numerical variables will be analysed to identify the features with a high positive or negative correlation with the target variable. A categorical analysis will also perform type object variables to deep-drive into implicit and latent connections within the data.

3.2.4 Data Visualization

Data visualisation is an integral part of exploratory data analysis to be able to understand the data. Visualisation packages to analyse and understand the data such as pandas profiling, sweetviz and data prep can be leveraged. This will help us understand the distribution of the columns, the variance, and the data profile. Comparing the data visually before and after processing will also help us understand datasets that will serve as inputs to the machine learning models in the model building steps in Section 3.2.7. Let us visualise a few of the features and the target variables to understand the distribution of the data points.

Figure 5: Distribution of Churn (Target variable)

There are 21 features and 7043 data points. Let us analyse the distribution of a few of the dependent variables.

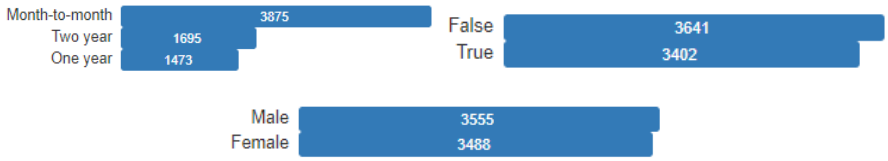


Figure 6: Distribution of Contract, Partner, Gender

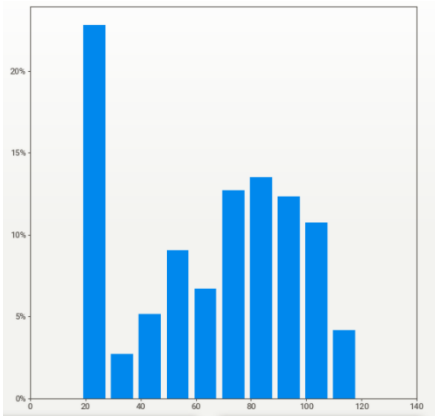


Figure 7: Distribution of Monthly Charges

3.2.5 Class Balancing

Oversampling and SMOTE are the techniques that will be leveraged to perform class balancing. The classification models had improved performance when class balancing was performed. Class balancing is performed in this section by using the recommended class balancing techniques of oversampling and SMOTE.

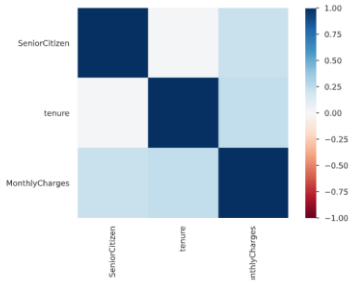


Figure 8: Correlation Matrix using Pearson's correlation coefficient (r)

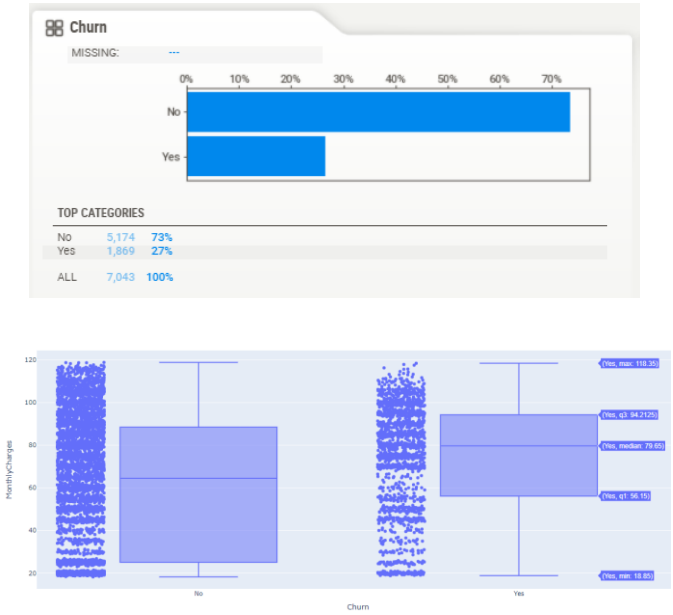


Figure 9: Distribution of Monthly Charges based on Churn

3.2.6 Model Building

Model Building is one of the more crucial components of this study. The following steps will help identify the right set of models and appropriate techniques to leverage to get optimal results. This model building is followed by choosing the models to implement after the data cleaning, feature engineering, and data formatting steps.

3.2.6.1 Model Selection Techniques

The best performing models are selected based on multiple factors ranging from accuracy to interpretability. From the literature review, it has been observed that the supervised classifier models have given good results. Single algorithm models are implemented to pick out the models that have the best performance. The models used Logistic Regression, decision trees, Naïve Bayes, random forest, support vector machine, and how the algorithms perform.

Based on the unique algorithms' analysis, bagging and boosting techniques are also attempted to have multiple weak classifiers combine to form a robust classifier using ensemble models such as XGBoost and Light GBM. To ensure that the model training is done right, the model is trained with two datasets – one with the original data and one on which class balancing techniques have been applied.

3.2.6.2 Test Designing

Another vital step to model building is to decide the train and test split strategically. If there were a larger dataset, a validation dataset could have also been leveraged. An 80-20 train-test split is leveraged for the models. For the top-performing models, a 90-10 split is attempted as well. This aspect of model building is also vital as having the right split will result in better results when cross-validation is carried out in the model validation phase for the models that are performing well, not only in a controlled but also in a robust setting in the long term.

3.2.6.3 Model Iterations

After the above model building steps, as mentioned earlier, are performed, more iterations will be performed, correspondingly to assessing model performance with each iteration. This can include monitoring p-values, the number of features, model performance, variance inflation factor scores which would differ across models. The top selected models will now be the challenger models based on which the best model will be decided. Hyperparameter tuning is done on the given models using previous learnings and methods such as Grid Search, Random Search, and Bayesian optimisation depending on the model considered.

3.2.6.4 Model Assessment

For any models to be used by the business, model assessment is a critical part of the process. As models are developed from a Data Scientist's eyes up until this point, the following steps will also ensure that the predictions are as expected for the company to leverage the model. There are multiple metrics one can use to perform the model assessment in this stage.

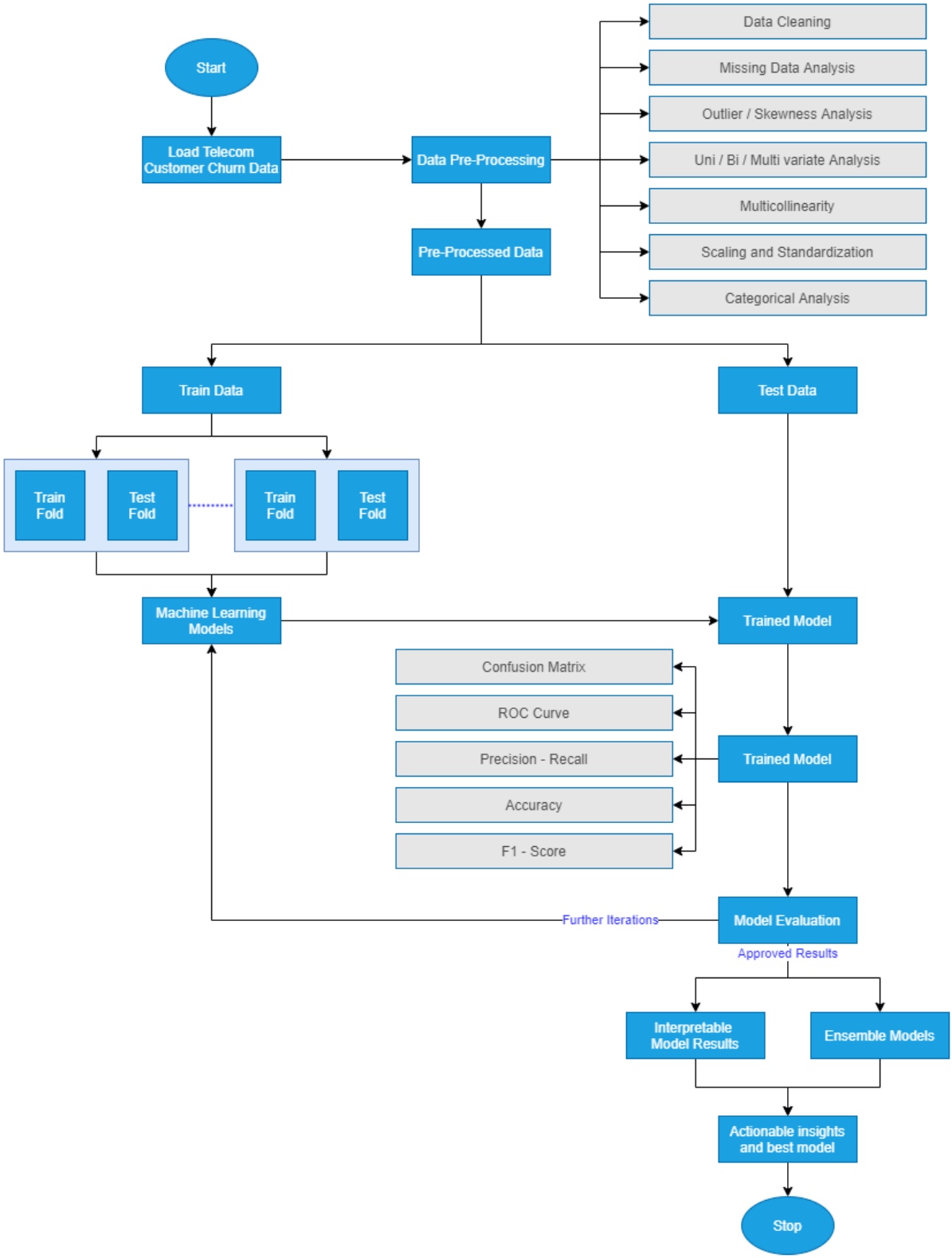


Figure 10: Model Building Process

The accuracy and AUC were used to assess models across the board from our literature review. Focus is also on model sensitivity and specificity curves to make a generalised model that can be leveraged. Model interpretability is vital to the business's functioning as they would like to understand the customers that are likely to churn and gain insights as to why. Therefore, in the model assessment stage, the focus needs to be on actionable insights and provide the business with the customer behaviour patterns linked with the high likelihood of churn. The diagram above highlights the stages to use for the model building process, from the data loading to the final model output. This step-by-step process has been drawn out in detail based on the extensive literature review carried out.

3.2.7 Model Evaluation

The best model is now chosen for the showcase. This is the model on which extensive feature engineering has been carried out, and from a wide range of models, the best model is now chosen. The below-mentioned steps are followed to perform the model evaluation. It is vital to perform a holistic evaluation of the model to assess our use-case's most appropriate model. The evaluation of the model will be done using all of the metrics mentioned in the literature review, including F-Measure, AUC, and accuracy.

3.2.7.1 Metrics for Evaluation

Comparison of the model results will be made based on the metrics obtained with the literature previously surveyed. They were using the same metrics of accuracy, F-Score, the area under the curve, and the new ensemble or individual models' performance to the models' performance in the field's reviewed literature. Once the results are evaluated and are satisfactory, the following steps will be carried out. Else, if they are not adequate, the approach will be re-evaluated to improve iteratively. This process is iterative as the final model selection should be as accurate as possible. Based on the literature review, the predictive framework's misclassification rate is also decreased. There are standard metrics that can be used and can be visually compared to select a model that can excel in all or most of the evaluation metrics chosen for classification.

3.2.7.2 Process Review

The final process lists the different iterations carried out and carefully reviewed the process compared to the other research done in this field, analysing any potential misses, flaws in approaches, and addressing them. Based on the process review carried out in the above step, the following steps to finish the research project will be decided. If not, further iterations will be initiated, and the model will be refined. This is an essential step and will be based on the comparative analysis performed to benchmark our model.

3.2.8 Model Review

The following steps for the business users will be to decide if the model evaluation is satisfactorily completed. This is critical so that a machine learning operationalisation pipeline can be set up within the environment to execute robust models to identify customers at a high risk of churn. The model is to be utilised by telecom companies to reduce the churn rate by targeting customers at a high likelihood of churn. There are certain factors to consider here based on which the company's return on investment can be maximised. 80% of revenue is generated by 20% of the customer base (Rajagopal, 2011). Based on the allocated budget for customer retention, high-value customers must be filtered with a high customer lifetime value and target those most likely to churn.

Allocating too much time to customers who are not generating as much revenue can be prioritised lower. A cost-benefit analysis will be carried out to understand the actual cost of running the model in real-time. There might be potential data anomalies while new data comes in. Robust machine learning pipelines along with teams to monitor the same will be deployed. This will help monitor the results and understand how to make the deployment more efficient.

For a machine learning model to improve with time, it is essential to create a feedback loop. Documentation of the research carried out, the results, and loopholes must be carefully documented to improve the model in the next iteration. If a similar accuracy can be obtained with lesser processing, this will also help the company save operationalisation expenditure costs. This is essential as reporting the research results and providing a list of assumptions so that the model's performance on future data will be based on an end-to-end understanding of the data and its

characteristics. In the final review, contemplation of the things done right and what went wrong will be done. There will be learnings from the entire process that can be documented and used in our next steps. Additionally, one can learn what was done well and what could have been avoided.

3.3 Summary

Once all of the above steps have executed, the proposed model is ready for the telecom company to use. The proposed model will be a hybrid tree-based classifier whose accuracy will be improved by SMOTE to select the class balancing technique. The model evaluation metrics are AUC and accuracy. The misclassification rate will be minimal to reduce overhead expenses by targeting customers who are likely to churn. It is advisable to opt-in for an accurate model and computationally sensible for this use case for operationalisation. The research methodology highlights all of the steps that can be taken to get the best predictive performance from the attrition model. The steps include data cleaning, data preprocessing, data transformation, data visualisation, class balancing, model building, model evaluation and model deployment Post the literature review carried out in the previous sections, the most appropriate model is now chosen for the dataset in consideration. All steps have been carried out per industry best practices.

CHAPTER 4: ANALYSIS

This chapter is to detail the process of building and implementing machine learning models in Python. The evaluation of the model will be done with the various evaluation metrics analyzed previously, such as AUC, Accuracy, precision and recall. By the end of this chapter, the various models will be implemented, and Chapter 5 will be used to analyze the results obtained from the analysis in Chapter 4. In this chapter, an in-depth analysis of the steps that can be taken to perform customer churn analysis will be explained with business explanation and justifications for each step.

4.1 Introduction

Chapter 4 in this research is to explore the selected telecom dataset in depth. The dataset will be described along with subtle details in Section 4.2. In Section 4.3, the steps covered for data preparation are noted. The distribution of variables, transformation of categorical variables, univariate analysis, missing value analysis and outlier analysis is done. The following Section 4.4 covers the extensive methodology that has been carried out on the telecom dataset. Section 4.5 covers the analysis of the models followed by model interpretability.

4.2 Dataset Description

The dataset used is sourced from IBM (Kaggle, 2018). The dataset will be analyzed to understand customer behaviour to predict the likelihood to churn customers. The data is at a customer level, where each row indicates a unique customer. The dataset is collected over a month, where if a customer has left in the last month, they have been flagged as a churned customer. Each column in the dataset is an indication of the customer's characteristics as captured by the system. Data points for all of the customers help analyze the various metadata associated with a customer within the database of the telecom company. The customers that have been marked as a churned customer are those customers that had churned in the month before the data was collected. When a customer is to be marked as a churned customer, it is an indication that the customer will churn.

The data has information about the following about the customers:

- Services that the customer has signed up for: movies, streaming tv, tech support, device protection, online backup, online security, internet, multiple lines, phone
- Information about the customer account: the tenure of the customer, payment methods, total charges, monthly charges, paperless billing, type of contract
- Demographic information about the customer: information about partners or dependents, gender, age-range

The target variable is the attribute Churn. There are 21 attributes, and the Churn column is the variable that is being predicted. 7043 data points capture customer level data along with their metadata in the form of attributes. This data has been sourced from the Cognos Analytics Team at IBM. It contains information about a telecom company that provided telecom and internet services to 7043 customers. The data indicates the customers that have stayed left or signed up for the service. It contains 18 categorical attributes and three numerical attributes, including the target variable. The dataset does not contain any missing values and can be used for churn analysis.

4.3 Exploratory Data Analysis

In this section, the details of the IBM Telecom dataset will be understood. The focus will be on the details of the data and how the data can be prepared to be used as inputs for the various models. Analysis of the data in the form of analysis, both univariate and bivariate, will be presented. The distribution of the variables will also be analysed along with missing value analysis and outlier analysis. The methods followed is to present details about the dataset, even if it is implied. This will help researchers in the future eliminate any confusion with regards to the details of the data. As the next steps, observations on the distribution of the variables will be done. This, along with analysis of the missing values in the data and the outlier analysis, will help analyse the minute details of the dataset—univariate analysis on the attributes of the data and mention the notable distribution. Finally, the relation of specific attributes with the target variable is also observed to understand how the frequency or distribution of attributes changes for customers that churn and do not churn.

4.3.1 Distribution of Variables

In this section, the distribution of variables will be understood. The focus will be given to understand both a percentage distribution of the variables and the absolute distribution of the data points within the dataset.

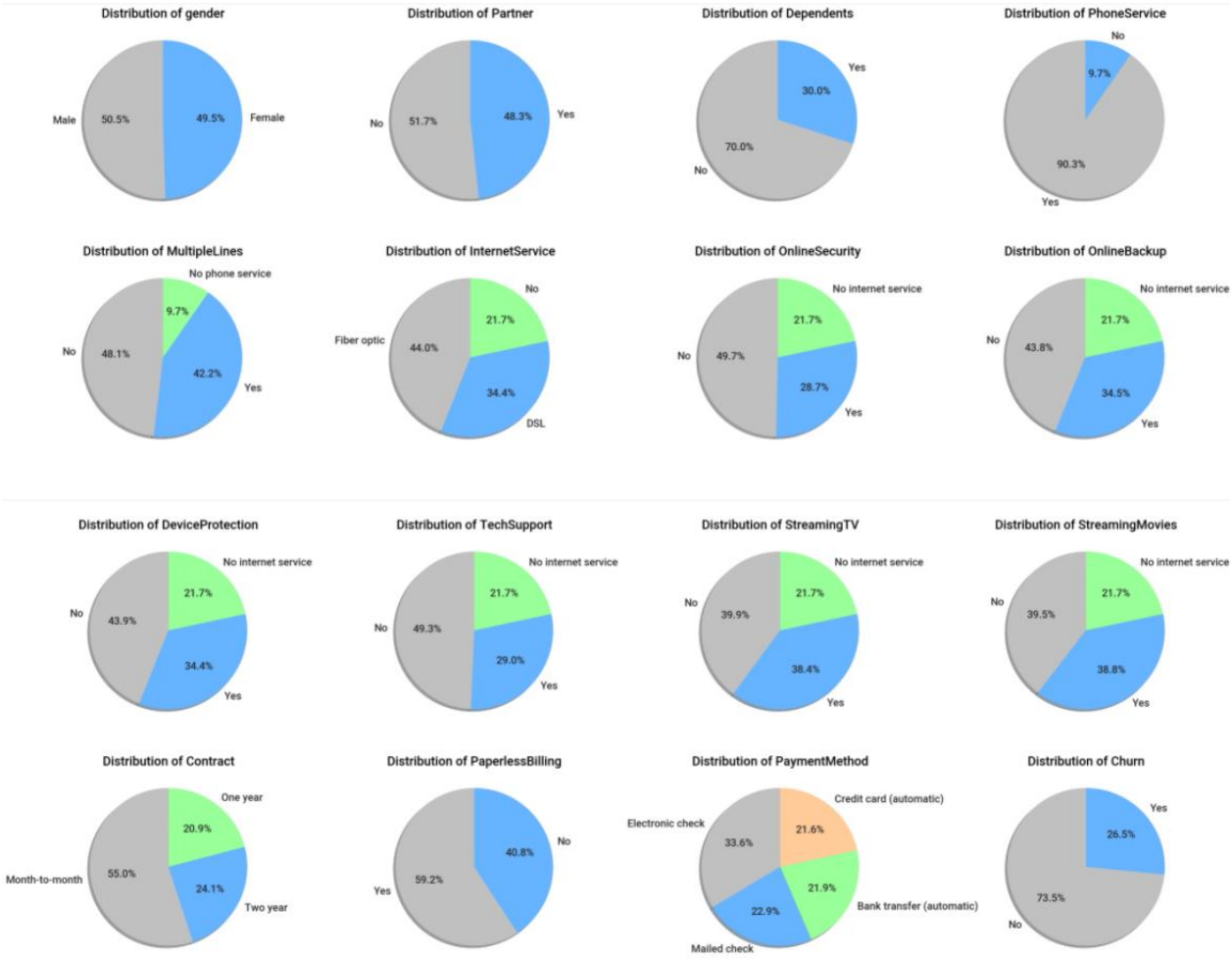


Figure 11: Distribution of variables (by percentage)

The distribution in *Figure 11* highlights the distribution of the variables by percentage. This helps us get an overall understanding of the distribution of each of the variables being considered. The distribution of gender suggests that the distribution of males and females is almost equal in the customer base of the telecom data.

Based on the distribution of Phone Service, it is also understood that 90.3% of customers use the phone service and 9.7% of customers use other services such as the internet from the company. It is noteworthy that 21.7% of the customer base has opted not to take internet services or has internet services from an alternative provider. This can lead to a potential market to tap in the future to cross-sell products of the telecom company. 40.8% of customer have opted not to go for paperless billing; this might be a place to reduce the amount spent to send the paper bill every month by defaulting the customer to paperless billing. The percentage of customers on a month to month contract is 55% can be increased to a one year or a two-year contract. This can help in customer retention policies.

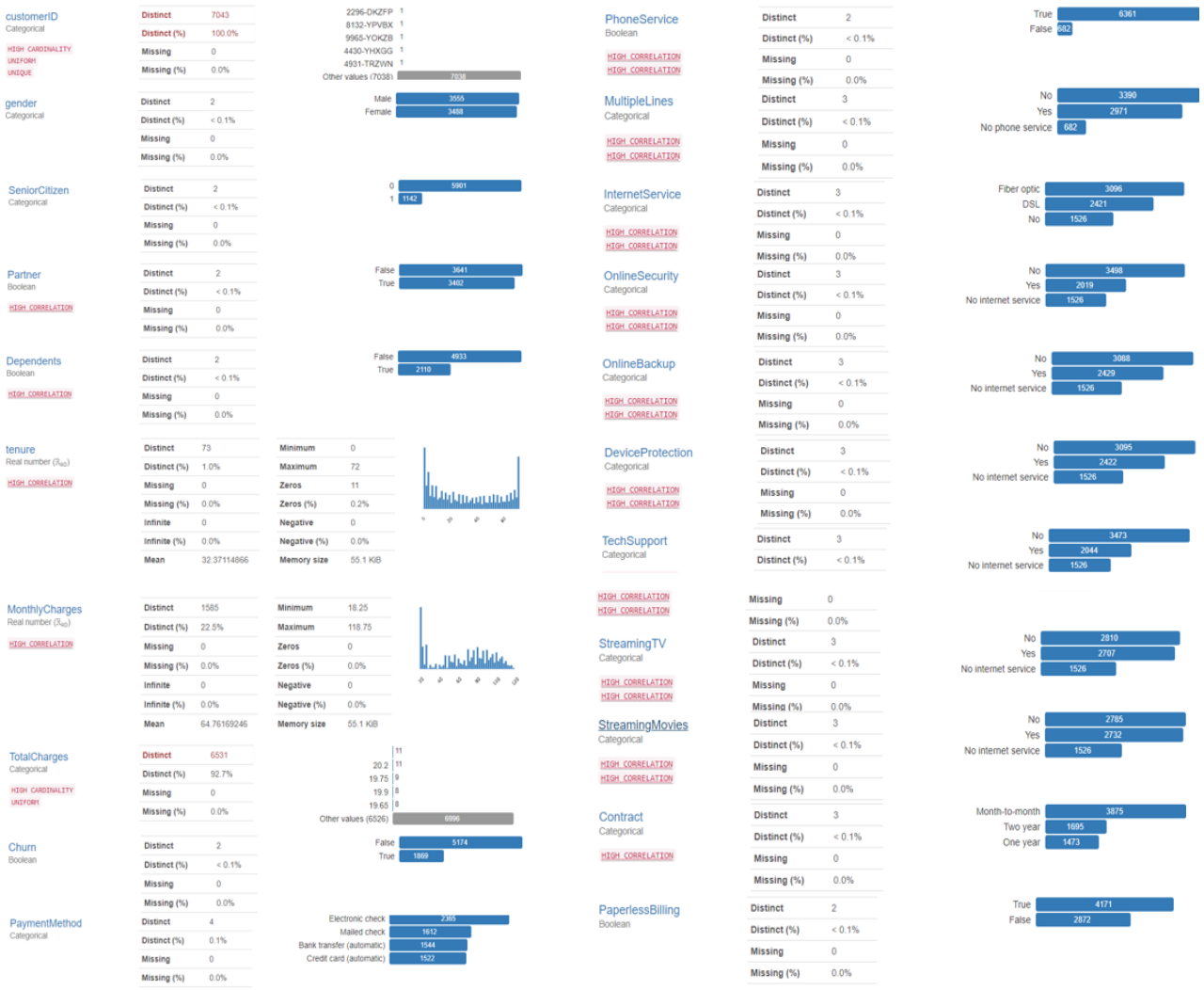
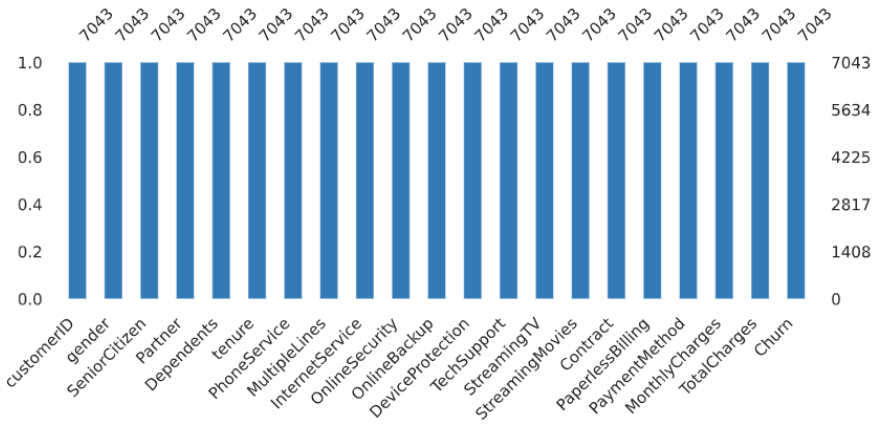


Figure 12: Distribution of variables (by absolute values)

In Figure 12, the details of the distribution of the variables have been mentioned. The visualizations also have tags attached to them to indicate the warnings and unique characteristics of the attributes. For instance, the Total Charges column has been flagged as a column with high Cardinality with 6531 distinct values. Additionally, there is information regarding the correlation and distribution of the variables mentioned in the below sections.

4.3.2 Missing Values Analysis

Identification of missing values is a crucial process during Data Understanding. The dataset in consideration does not have missing values. As noted earlier, there are 7043 rows at a customer level, and the below visualizations will showcase the visualization of nullity by column. Nullity matrix is a data-dense display that helps to pick out patterns in data completion visually – this helps define patterns visually to quantify missing data, especially for larger datasets.



Missing Statistics

Missing Cells	0
Missing Cells (%)	0.0%
Missing Columns	0
Missing Rows	0
Avg Missing Cells per Column	0.0
Avg Missing Cells per Row	0.0

Figure 13: No missing values - Nullity by column for IBM Teleco Data

Post analysis from Figure 13 shows that from the missing value analysis using the nullity matrix, there are no missing values in the IBM Teleco Dataset. All of the columns have 7043 values. No further steps need to be taken post the missing value analysis.

4.3.3 Outlier Analysis

The dataset has categorical variables as metadata for each customer. There are two attributes – Monthly Charges and Total Charges-numerical values on which outlier analysis can be performed. The study will be using a boxplot with an inter-quartile range of 1.5 x (IQR) as the upper and lower whiskers for the two attributes. The attributes will be plotted against churn, where 0 indicates that the customer did not churn and one indicates that the customer did churn.

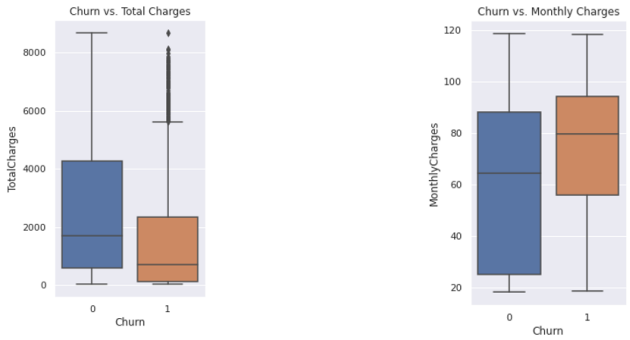


Figure 14 Boxplots of Churn versus Total Charges and Churn versus Monthly Charges

In the distribution in Figure 15, it is observed that for Total Charges, the majority of the customers have a customer lifetime value of less than 2000, which is an indication that customers that have a lower tenure with the company are likely to churn. Whereas, in the boxplot of Monthly Charges, it is observed that the distribution of customers that churn is populated between 60 to 90. This is an indication that customers that have lower monthly charges are less likely to churn.

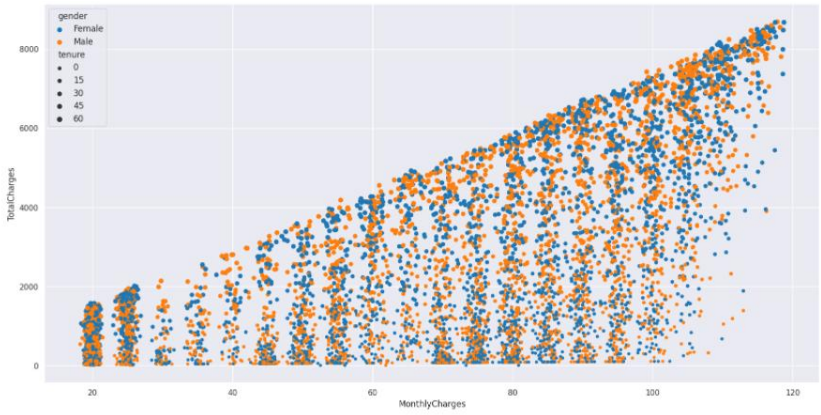


Figure 15 Scatter plot of Monthly Charges versus Total Charges

There is a significant correlation between Monthly Charges and Total Charges, as expected. This is expected as Figure 15 illustrated that as the monthly charge per customer increases, the total charges or the customer lifetime value to the company increase. This is further noted when the heatmap is generated, where it is observed that the Pearsons's coefficient for monthly charges and total charges is 0.7, which indicates a strong positive correlation between the attributes in the telecom churn dataset.

4.3.4 Univariate Analysis

From Figure 12, where all the attributes have been plotted, initial univariate analysis has already been performed. In this section, the numerical attributes of the dataset will be analyzed in greater depth. Understanding the distribution of the three numerical features – Monthly Charges, Total Charges, and Tenure is how univariate analysis can be performed.

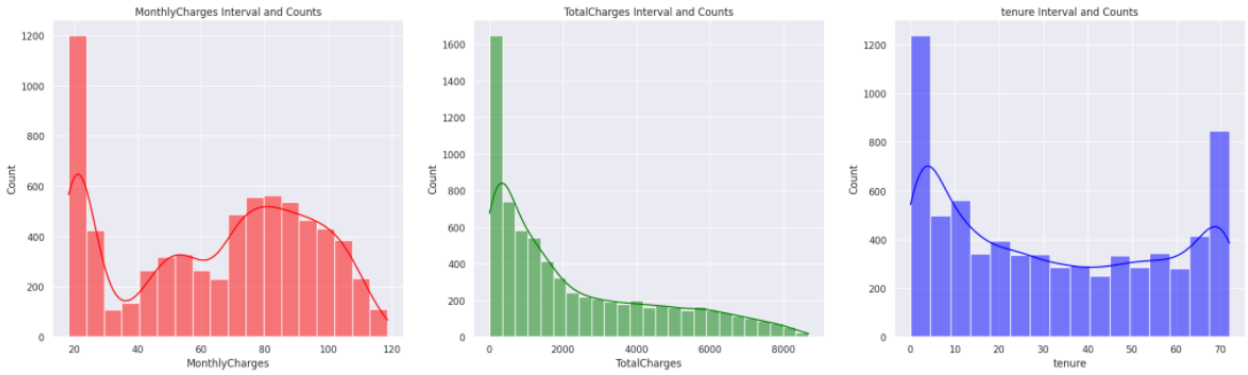


Figure 16: Univariate Analysis of numerical features of IBM Teleco Data

Most customers have a monthly charge of around 20. The histogram of Monthly Charges suggests that high-value customers peak around 80 and gradually taper off around 120. The frequency of customers based on tenure suggests that after spending around 15 months with the Telecom company, the number of customers with a high tenure decreases. The visualization in Figure 16 showcases the distribution of the numerical values, where Monthly charges seem to have an uneven distribution. As part of the next steps, transformations will be applied closer to a normal distribution.

4.3.5 Relation with Target Variable

Understanding the distribution of features in Section 4.3.4 helps understand how the distribution of attributes occurs. In this section, the relation of multiple attributes concerning churn is observed. Based on whether the churn is marked as Yes or No, the distribution of multiple features is observed. This helps gauge an intuition of churn for the attributes under consideration. A deeper understanding of the behaviour or churn is observed when visualizations are used.

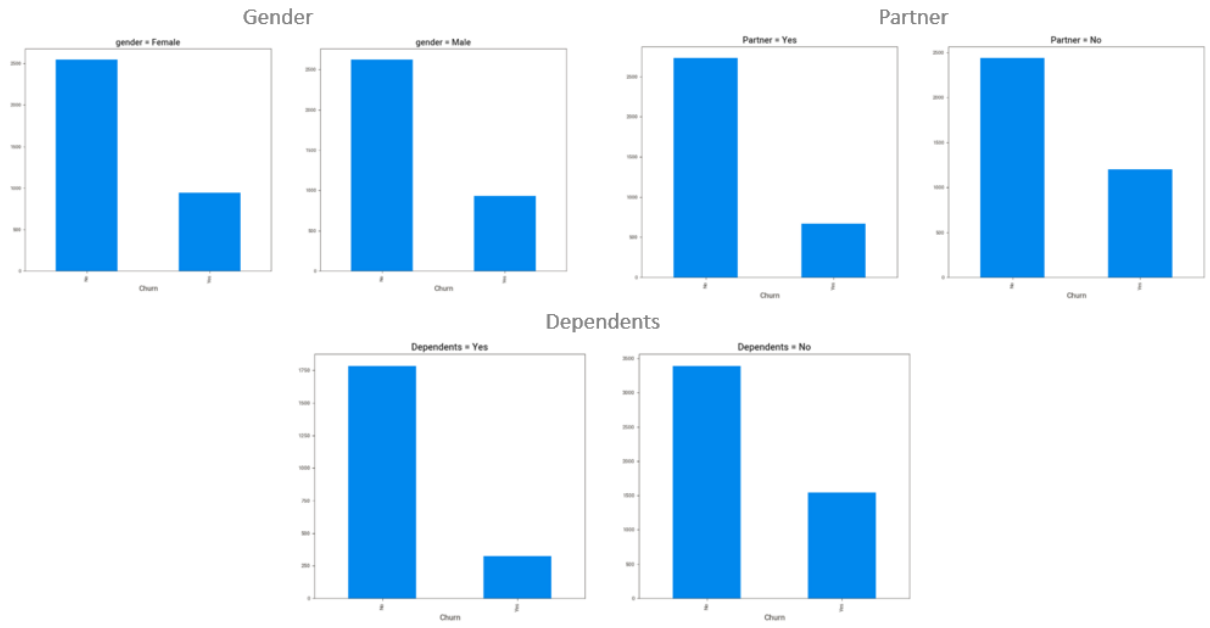


Figure 17: Distribution of Demographic Attributes with respect to Churn

The relation of the demographic variables with churn can be seen in Figure 17. It is observed that customers that do not have a partner or dependents are more likely to churn. This indicates that customers who have a family might take more services from the company and are more likely to stick to them. For the next set of visualizations, the focus will be on the attributes in the dataset that have observable trends concerning the target variable. Internet Service for Digital Subscriber Line (DSL), Fiber Optic, and not having an internet line is showcased in Figure 18 that customers with a Fiber Optic line are more likely to churn. Customers that do not stream movies are also more likely to churn. One of the most prominent observations from plotting all the features is that customers who opt-in for a contract that is charged monthly are the most likely to churn. Customers that have a one year contract or two-year contract are much less likely to churn.

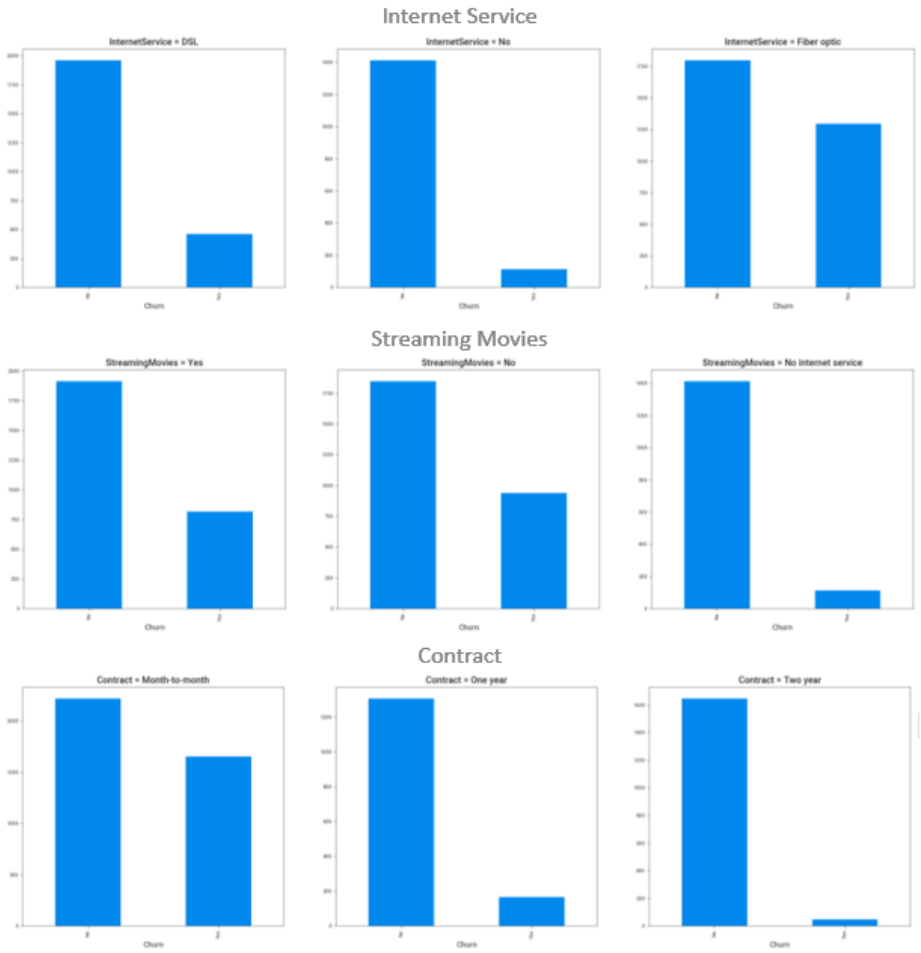


Figure 18: Internet Service, Streaming Movies and Contract plotted with respect to the target variable - Churn

The visualizations in Section 4.3 helped us understand the distribution of the data with the help of visualizations. No matter how large the data, the easiest way to gain an intuition is to perform exploratory data analysis. In this section, the analysis will understand how the attributes can be observed with respect to churn. Since visualizations operate in a two-dimensional space, there is a limitation on the number of features that can be showcased. Certain aspects of the visualization such as x-axis, y-axis, colours, shape and size can be leveraged to add more dimensions in the limited space provided. Correlation of the various attributes will also be noted, where the correlation between quantitative variables and correlation between qualitative/ categorical variables will also be plotted. This will help us understand the correlation between the variables in the dataset as well.

4.3.6 Distribution of variables with respect to Churn

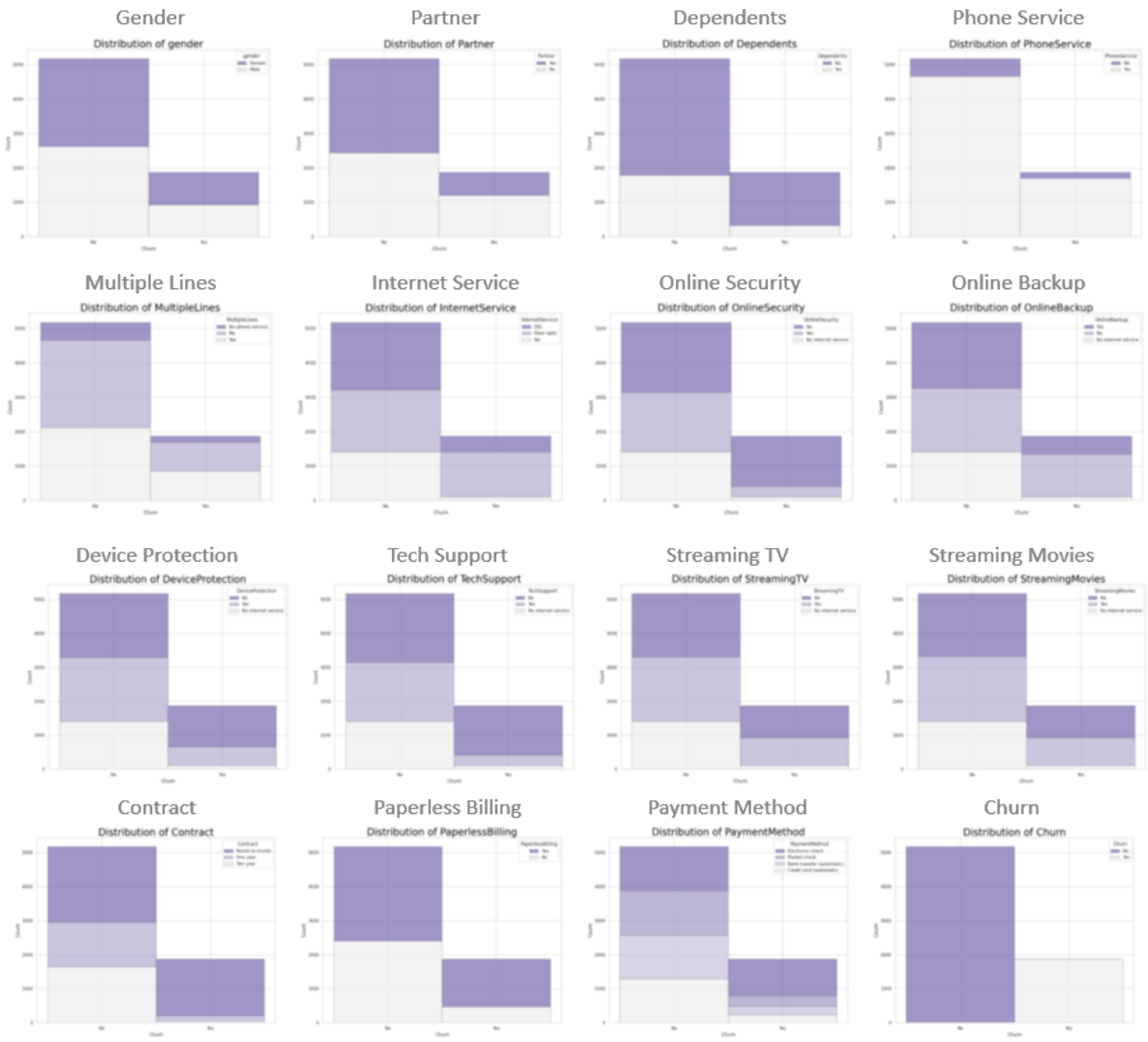


Figure 19: Distribution of all features with respect to Churn

In Figure 19, the distribution of features with respect to churn can be observed with the help of a stacked histogram. Directing focus on the customers who have churned will help us identify the patterns that are likely to churn. In the feature selection techniques, some of the observations that can be made from the visualizations can be confirmed. The number of males and females churning is equal. Most people that churn do not have dependents. Customers that do not use tech support are more likely to churn. Electronic check is the most used method of payment by the customers

that churn. Customers that do not use online security are also more likely to churn. These were a few of the notable observations that can be made when looking at the set of histograms in Figure 19. In the below section, the correlation between the variables will be analyzed.

4.3.7 Correlation

In this section, the correlation between the variables will be analyzed. First, the correlation between quantitative variables will be analyzed. In Figure 20, the Pearson's coefficient has been used to calculate the correlation between numerical variables. It is observed that there is a high positive correlation of 0.83 between tenure and total charges. This is in line with the data's understanding, as this means that customers who have spent more time with the telecom operator have a higher customer lifetime value calculated using total charges. The other high positive correlation is observed between monthly charges and total charges of 0.65, where it is indicated that if a customer has high monthly charges, it is more likely that the customer has high total charges over their time with the telecom operator. The rest of the correlation coefficients are insignificant as they are neither greater than 0.5 nor less than -0.5, which would have indicated a high positive correlation or a negative correlation, respectively.

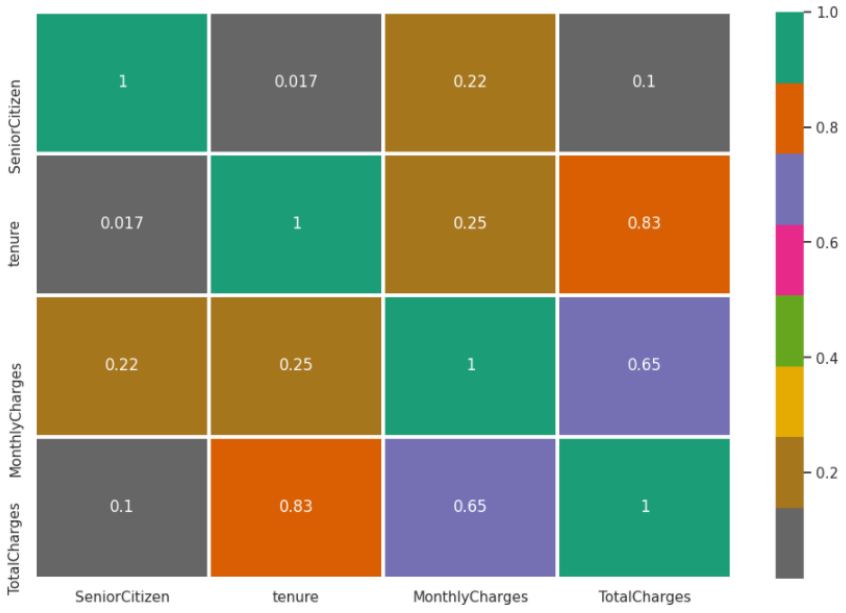


Figure 20: Correlation between quantitative variables

In Figure 19 below, all of the qualitative features have been plotted. If the customer does not have internet service, the monthly charges are lower – this can be inferred from the correlation coefficient being -0.8. There are no significant inferences that can be made from the below heatmap. The significant relationships have already been captured in Figure 18.

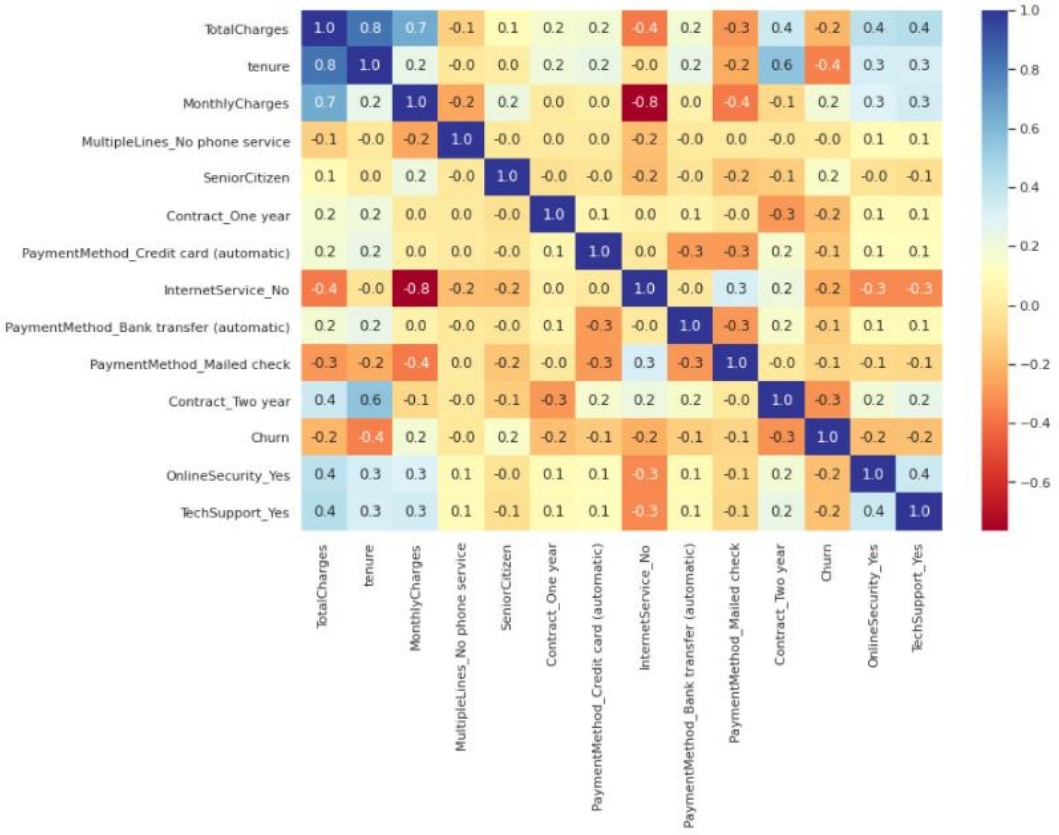


Figure 21: Correlation between qualitative variables

4.3.7 Chi-Square

The Pearson's chi-squared test is used to test how likely it is that an observed distribution is due to chance. The chi-squared statistic operator is used to assign weights to the attribute with respect to the class attribute is calculated. It can only be used on categorical features. Based on the chi-squared weights assigned, the top 20 features along with the weights will be showcased in Figure 22.

Features	Chi2Weights
TotalCharges	1.582229e+07
tenure	1.312128e+05
MonthlyCharges	9.845178e+04
MultipleLines_No phone service	6.361000e+03
SeniorCitizen	5.901000e+03
Contract_One year	5.570000e+03
PaymentMethod_Credit card (automatic)	5.521000e+03
OnlineBackup_No internet service	5.517000e+03
InternetService_No	5.517000e+03
StreamingMovies_No internet service	5.517000e+03
StreamingTV_No internet service	5.517000e+03
TechSupport_No internet service	5.517000e+03
DeviceProtection_No internet service	5.517000e+03
OnlineSecurity_No internet service	5.517000e+03
'aymentMethod_Bank transfer (automatic)	5.499000e+03
PaymentMethod_Mailed check	5.431000e+03
Contract_Two year	5.348000e+03
Churn	5.174000e+03
OnlineSecurity_Yes	5.024000e+03
TechSupport_Yes	4.999000e+03

Figure 22: Top 20 features based on chi-squared weights

Based on the weights obtained from the chi-square analysis, a few of the more essential features are total charges, tenure, monthly charges, not having a phone service and whether the customer is a senior citizen,

4.5 Methods

In this section, the discussion will be around the methods and standards that will be followed as a part of this study. The conventions followed through the study will be highlighted in the form of the data split, the encoding used and the feature engineering employed for the task of the prediction of the customers that are at a high risk of churn.

4.5.1 Data Split

The dataset will be split at a train-test ratio of 80% train data and 20% test data using the sklearn model selection library. This will be done in a stratified manner by the train-test package that will be leveraged in python. The main objective of the stratified train-test split is to keep the same proportion of train and test class samples as the original data.

4.5.2 Encoding

Label encoding was performed on the data, where each point was assigned a unique value. Keeping the size of the data in mind and the functionality, label encoding was deprioritized. One-hot encoding was used to account for categorical features as inputs in the models used to predict churn. On account of the high cardinality of certain features such as Customer id, the column was discarded as it is computationally expensive, increases the size of the data and does not add any additional value to the model.

4.5.3 Feature Engineering

Feature engineering is the process of creating new features by transforming existing features into a new feature space. Feature engineering does have the potential to improve model performance (Khurana et al., 2017). However, in our use-case where there are two numerical attributes, monthly charges and total charges, feature engineering will not make sense here as generating a new feature will bring about high multicollinearity in the data. Box-Cox transformation was also applied on the dataset for specific columns, such as monthly charges.

4.5.4 Class Imbalance

Oversampling or a few other methods offered better accuracy at times based on the model being used. For the sake of our research, the SMOTE-NC method from the imbalanced-learn was leveraged. SMOTE-NC creates synthetic data for categorical as well as quantitative data based on the k-nearest neighbour algorithm. It is only applied to the training dataset to avoid contamination.

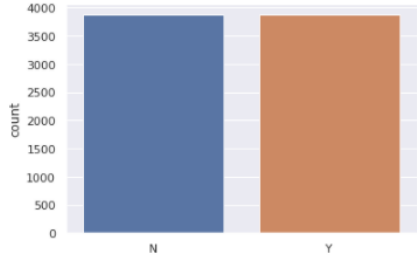


Figure 23: Plot of train data after SMOTE-NC is applied

4.5.6 Implementation

All of the analysis and implementation has been done on Google Colab. The configuration of the virtual machine is two CPU cores of the Haswell CPU family at 2.30GHz with RAM of 16 GB and disk space of 25 GB. All of the packages that have been leveraged are open source python packages. For instance, for importing the data and working with data frames, NumPy and pandas have been used. For the visualization, packages like matplotlib, seaborn, pandas-profiling and sweetviz have been used. Machine learning models have been implemented leveraging packages such as sklearn, xgboost and catboost. For data-level solutions, data balancing libraries such as imblearn have been used. All of the code was developed on the Colab platform using the native inbuild CPU and compute power on the Edge Browser. The data was sourced from Kaggle and pulled in situ.

4.6 Analysis

In this section, the baselines for the research will be decided along with the models that can be implemented to be analysed in Chapter 5 in the results and discussions section. The results of the methods and models implemented will be detailed out in the next section. This will include pre-processing, feature selection, class balancing, ensemble models, cross-validation and model interpretability. Individual models will also be compared, and the results will be showcased individually. The learnings from the literature survey will be implemented in the sections below.

4.6.1 Baselines

In order to evaluate the models effectively, a baseline model will be set. For this study, the model that will be selected as a baseline is a decision tree model on a dataset where one-hot encoding has been performed. The model will be evaluated using two main metrics of accuracy and ROC-AUC as evaluation metrics on the test data. Additionally, for the models that perform better,the f1-scores will be analyzed as well. Setting up a baseline helps us eliminate models that are not at par with the baseline

4.6.2 Models

Multiple models will be leveraged to predict customer churn from the data. Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, K-Nearest Neighbour, Gradient Boosting Classifier, Stochastic Gradient Descent, Light Gradient Boosted Machine were leveraged to predict churn and individual model performance was analyzed. This will be tested on the train, validation and test data. The initial analysis will be done to opt for the models that are above the baseline. Then, ensemble models such as Decision Trees with Bagging, Decision Trees with AdaBoost, Linear Support Vector Classification, Support Vector Machine with radial basis kernel function, XGBoost and CatBoost along with hyperparameter tuning. All of the results from the models will be discussed in Chapter 5 in visualizations that will help compare the models.

4.6.3 Feature selection

Feature selection techniques help us understand the features that are of higher importance for the prediction of churn. The feature selection techniques leveraged are Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier and Light GBM. The features that were showcased as necessary is shown in Figure 24 and Figure 25.

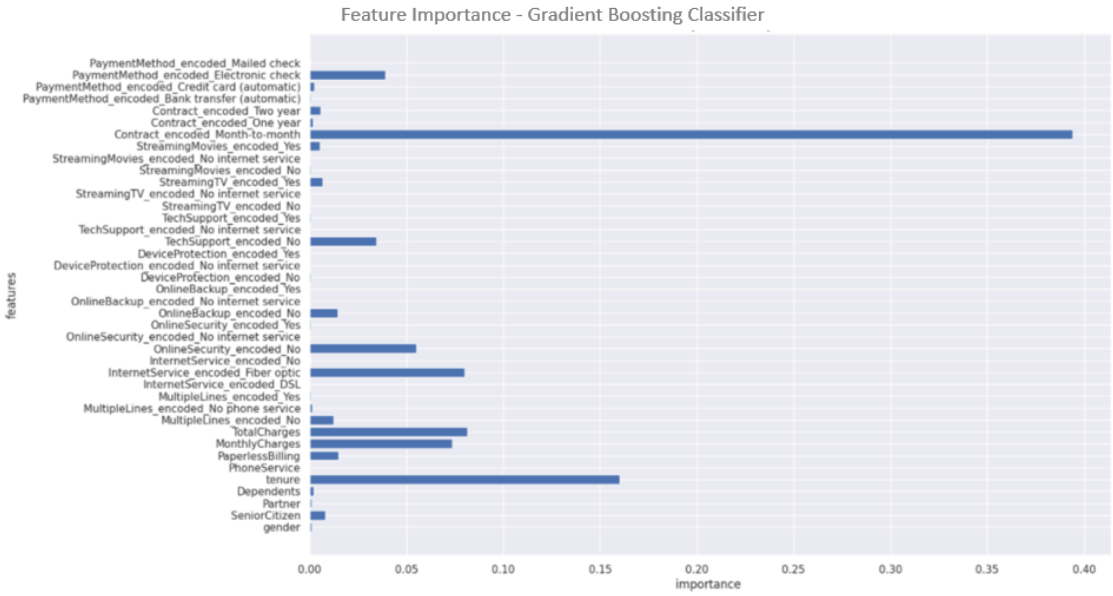


Figure 24: Feature Selection using Gradient Boosting Classifier

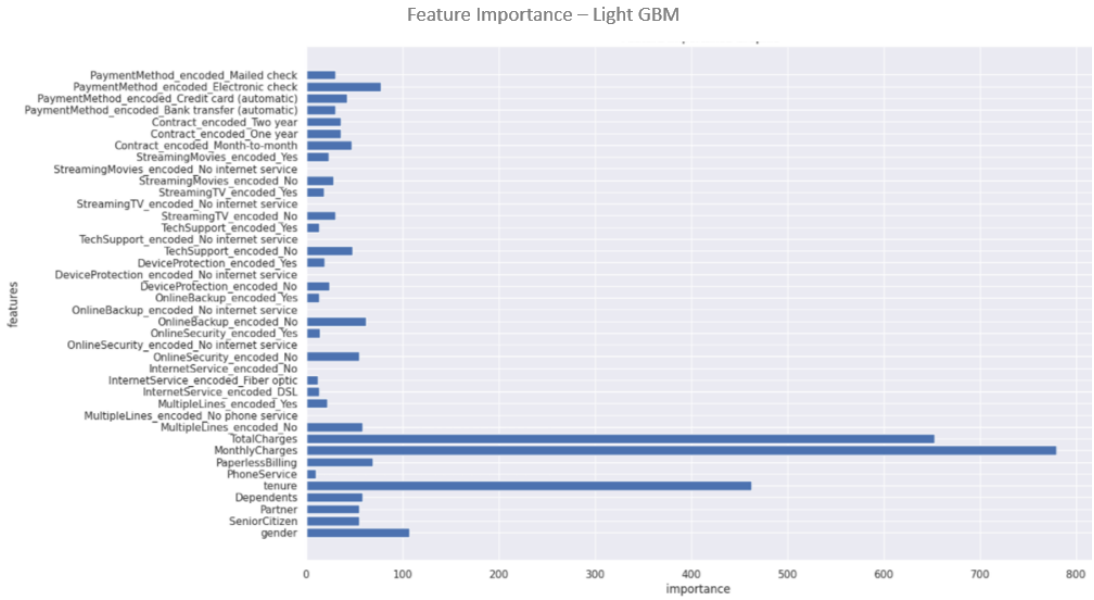


Figure 25: Feature Selection using Gradient Boosting Classifier and Light GBM

From the charts that showcase feature importance in Figure 24 and 25, it is noted that the crucial features are the month to month contracts, the tenure of the customer, the total charges and the monthly charges. This may be because these attributes have the highest variance compared to the categorical values in the other attributes that signify customer behaviour. The Random Forest and Decision Tree Classifiers had similar results, where higher weights were assigned to the variables that had a higher variance.

4.6.4 Cross-Validation

Cross-validation for the models with k = 10 is done to improve accuracy based on the iterations. A resampling procedure provides information about how well a classifier generalizes, used to evaluate the score by cross-validating the train and test datasets. Cross-validation techniques are generally more effective on smaller datasets, and for our use case, the cross-validation strategy employed is taken as ten. Not all models have better scores on cross-validation strategies; when applied on the dataset – depending on the algorithm implemented, minute yet significant improvements in some accuracy scores is observed.

4.7 Model Interpretability

Businesses in the real world want to understand the reasoning behind model predictions. This is not always possible with machine learning models, especially as models get relatively complex, the interpretability of the model decreases. This is where Locally Interpretable Model-Agnostic Explanations (LIME) comes in. LIME is a model agnostic technique that can be applied to any machine learning model by perturbing the input of data samples to understand how the predictions change. For this study, LIME will be used on a few data samples to observe the reasons why a customer might or might not churn in a model-agnostic manner. The results of this will be analyzed further in Chapter 5.

4.8 Summary

In Chapter 14, the analysis and the techniques that were used for the research to run classification models on the telecom data to predict if customers will churn. The dataset was analysed by leveraging the distribution, missing values and outliers to understand the nuances of the data. Univariate and bivariate analysis was also performed, where the relationship with the target variable, churn, was analyzed. The correlation for quantitative as well as qualitative variables was analyzed. In Section 4.5, the standards that have been followed throughout the study have been highlighted, where the data split used, the encoding for categorical variables, the class imbalance techniques used were highlighted as well. The model baseline was also declared, and the cross-validation methods and parameters were explained. The issue of model interpretability was also taken up, and a novel solution to using LIME was showcased.

CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Introduction

Sample

5.2 Interpretation of Visualisations

Sample

5.3 Evaluation of Sampling Methods

Sample

5.4 Testing on Validation Dataset

Sample

5.6 Summary

Sample

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

Sample

6.1 Introduction

Sample

6.2 Discussion and Conclusion

Sample

6.3 Contribution to Knowledge

Sample

6.4 Future Recommendations

REFERENCES

Agrawal, S., (2018) Customer Churn Prediction Modelling Based on Behavioural patterns Analysis using Deep Learning. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pp.1–6.

Ahmad, A.K., Jafar, A. and Aljoumaa, K., (n.d.) Customer churn prediction in telecom using machine learning in big data platform. [online] Available at: https://doi.org/10.1186/s40537-019-0191-6.

Ahmed, A. and Linen, D.M., (2017) A review and analysis of churn prediction methods for customer retention in telecom industries. In: *2017 4th International Conference on Advanced Computing and Communication Systems, ICACCS 2017*. Institute of Electrical and Electronics Engineers Inc.

Ahmed, A.A. and Maheswari, D., (2017) A Review And Analysis Of Churn Prediction Methods For Customer Retention In Telecom Industries. *2017 International Conference on Advanced Computing and Communication Systems*.

Ambildhuke, G.M., Rekha, G. and Tyagi, A.K., (2021) Performance Analysis of Undersampling Approaches for Solving Customer Churn Prediction. [online] Springer, Singapore, pp.341–347. Available at: https://link.springer.com/chapter/10.1007/978-981-15-9689-6_37 [Accessed 21 Mar. 2021].

Andrews, R., (2019) Churn Prediction in Telecom Sector Using Machine Learning. *International Journal of Information Systems and Computer Sciences*, 82, pp.132–134.

Anon (2021) *Cognos Analytics - IBM Business Analytics Community*. [online] Available at: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 [Accessed 14 Mar. 2021].

Anon (2021) *Digital transformation for 2020 and beyond eight telco considerations*. [online] Available at: https://www.ey.com/en_in/tmt/digital-transformation-for-2020-and-beyond-eight-telco-considera [Accessed 25 Mar. 2021].

Anon (2021) *Why is the telecom industry struggling with product success?* [online] Available at: https://internationalfinance.com/why-telecom-industry-struggling-product-success/ [Accessed 25 Mar. 2021].

Castanedo, F., Valverde, G., Zaratiegui, J. and Vazquez, A., (2014) Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network Federico. pp.1–8.

Ebrah, K. and Elnasir, S., (2019) Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *11Journal of Computer and Communications*, [online] ``23df, pp.33–53. Available at: https://doi.org/10.4236/jcc.2019.711003 [Accessed 10 Jan. 2021].

Fonseca Coelho, A., (n.d.) *Churn Prediction in Telecom Sector: A completed data engineering Framework*.

Hadden, J., Tiwari, A., Roy, R. and Ruta, D., (2006) Churn Prediction: Does Technology Matter. *International Journal of Intelligent Technology*, 1, pp.104–110.

Halibas, A.S., Cherian Matthew, A., Pillai, I.G., Harold Reazol, J., Delvo, E.G. and Bonachita Reazol, L., (2019) Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling. *2019 4th MEC International Conference on Big Data and Smart City, ICBDSC 2019*.

Hargreaves, C.A., (2019) A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry. *International Journal of Future Computer and Communication*, 84, pp.109–113.

Havrylovych, M. and Nataliia Kuznietsova, ©, (2019) *Survival analysis methods for churn prevention in telecommunications industry*.

Induja, S. and Eswaramurthy, V.P., (2015) *Customers Churn Prediction and Attribute Selection in Telecom Industry Using Kernelized Extreme Learning Machine and Bat Algorithms*. [online] *International Journal of Science and Research (IJSR) ISSN*, Available at: www.ijsr.net [Accessed 18 Feb. 2021].

Jahromi, A.T., Stakhovych, S. and Ewing, M., (2014) Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, [online] 437, pp.1258–1268. Available at:

https://research.monash.edu/en/publications/managing-b2b-customer-churn-retention-and-profitability [Accessed 16 Jan. 2021].

Jain, H., Yadav, G. and Manoov, R., (2021) Churn Prediction and Retention in Banking, Telecom and IT Sectors Using Machine Learning Techniques. [online] Springer, Singapore, pp.137–156. Available at: https://link.springer.com/chapter/10.1007/978-981-15-5243-4_12 [Accessed 21 Mar. 2021].

Kaggle, (2018) *Telco Customer Churn*. *Kaggle.com*. Available at: https://www.kaggle.com/blastchar/telco-customer-churn [Accessed 9 Jan. 2021].

Karimi, N., Dash, A., Rautaray, S.S. and Pandey, M., (2021) A Proposed Model for Customer Churn Prediction and Factor Identification Behind Customer Churn in Telecom Industry. [online] Springer, Singapore, pp.359–369. Available at: https://link.springer.com/chapter/10.1007/978-981-15-7511-2_34 [Accessed 21 Mar. 2021].

Khurana, U., Nargesian, F., Samulowitz, H., Khalil, E.B. and Turaga, D., (2017) Learning Feature Engineering for Classification. [online] Available at: https://www.researchgate.net/publication/318829821 [Accessed 19 May 2021].

Kriti, (2019) *Customer churn: A study of factors affecting customer churn using machine learning*. [online] Available at: https://lib.dr.iastate.edu/creativecomponents [Accessed 14 Mar. 2021].

Kuo, Y.-F., Wu, C.-M. and Deng, W.-J., (2009) The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior*, 25, pp.887–896.

Labhsetwar, S.R., (n.d.) Predictive Analysis Of Customer Churn in Telecom Industry using Supervised Learning.

Lalwani, P., Banka, H. and Kumar, C., (2017) GSA-CHSR: Gravitational Search Algorithm for Cluster Head Selection and Routing in Wireless Sensor Networks. In: *Applications of Soft Computing for the Web*. [online] Springer Singapore, pp.225–252. Available at: https://link.springer.com/chapter/10.1007/978-981-10-7098-3_13 [Accessed 20 Mar. 2021].

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., (2021) Customer churn prediction system:

a machine learning approach. *Computing*.

Mahdi, A., Alzubaidi, N. and Al-Shamery, E.S., (2020) Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry discriminant random forest Linear discriminant analysis oblique tree Project pursuit index Support vector machines. *International Journal of Electrical and Computer Engineering (IJECE)*, 102, pp.1406–1421.

Momin, S., Bohra, T. and Raut, P., (2020) *Prediction of Customer Churn Using Machine Learning*. *EAI/Springer Innovations in Communication and Computing*.

Mukhopadhyay, D., Malusare, A., Nandanwar, A. and Sakshi, S., (2021) An Approach to Mitigate the Risk of Customer Churn Using Machine Learning Algorithms. In: *Lecture Notes in Networks and Systems*. [online] Springer Science and Business Media Deutschland GmbH, pp.133–142. Available at: https://link.springer.com/chapter/10.1007/978-981-15-7106-0_13 [Accessed 21 Mar. 2021].

Oka, N.P.H. and Arifin, A.S., (2020) Telecommunication Service Subscriber Churn Likelihood Prediction Analysis Using Diverse Machine Learning Model. *MECnIT 2020 - International Conference on Mechanical, Electronics, Computer, and Industrial Technology*, pp.24–29.

Oskarsdottir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B. and Vanthienen, J., (2016) A comparative study of social network classifiers for predicting churn in the telecommunication industry. In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. Institute of Electrical and Electronics Engineers Inc., pp.1151–1158.

Pamina, J., Beschi Raja, J., Sathya Bama, S., Soundarya, S., Sruthi, M.S., Kiruthika, S., Aiswaryadevi, V.J. and Priyanka, G., (2019) An effective classifier for predicting churn in telecommunication. *Journal of Advanced Research in Dynamical and Control Systems*, 111 Special Issue, pp.221–229.

Priyanka Paliwal and Divya Kumar, (2017) *ABC based neural network approach for churn prediction in telecommunication sector*. [online] *(Ictis 2017)*, Available at: http://dx.doi.org/10.1007/978-981-13-1747-7_65.

Rajagopal, D.S., (2011) Customer Data Clustering using Data Mining Technique. *International Journal of Database Management Systems*, [online] 34. Available at: http://arxiv.org/abs/1112.2663 [Accessed 17 Jan. 2021].

Saonard, A., (2020) Modified Ensemble Undersampling-Boost to Handling Imbalanced Data in Churn Prediction. [online] Available at: https://core.ac.uk/download/pdf/326763412.pdf [Accessed 21 Mar. 2021].

Sharma, T., Gupta, P., Nigam, V. and Goel, M., (2020) Customer Churn Prediction in Telecommunications Using Gradient Boosted Trees. In: *Advances in Intelligent Systems and Computing*. [online] Springer, pp.235–246. Available at: https://link.springer.com/chapter/10.1007/978-981-15-0324-5_20 [Accessed 21 Mar. 2021].

Tamuka, N. and Sibanda, K., (2021) Real Time Customer Churn Scoring Model for the Telecommunications Industry. *IEEE*, pp.1–9.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A., (2020) Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, pp.429–441.

Thontirawong, P. and Chinchanachokchai, S., (2021) TEACHING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN MARKETING. *Marketing Education Review*.

Tuck, W.K., Chien-Le, G. and Hu, N., (2020) A False Negative Cost Minimization Ensemble Methods for Customer Churn Analysis. In: *ACM International Conference Proceeding Series*. [online] New York, NY, USA: Association for Computing Machinery, pp.276–280. Available at: https://dl.acm.org/doi/10.1145/3384544.3384551 [Accessed 14 Mar. 2021].

Umayaparvathi, V. and Iyakutti, K., (2016) *A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics*. [online] *International Research Journal of Engineering and Technology*. Available at: http://www.fuqua.duke.edu/centers/ccrm/index.html [Accessed 20 Mar. 2021].

Wassouf, W.N., Alkhatib, R., Salloum, K. and Balloul, S., (n.d.) Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. [online] Available at:

https://doi.org/10.1186/s40537-020-00290-0 [Accessed 21 Mar. 2021].

Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., Liu, S. and Author, T., (2021) Computational Visual Media A survey of visual analytics techniques for machine learning. [online] 71, pp.3–36. Available at: https://doi.org/10.1007/s41095-020-0191-7 [Accessed 28 Mar. 2021].

APPENDIX A: RESEARCH PLAN

The following GANTT chart proposes the timeline for the research and implementation of the project.

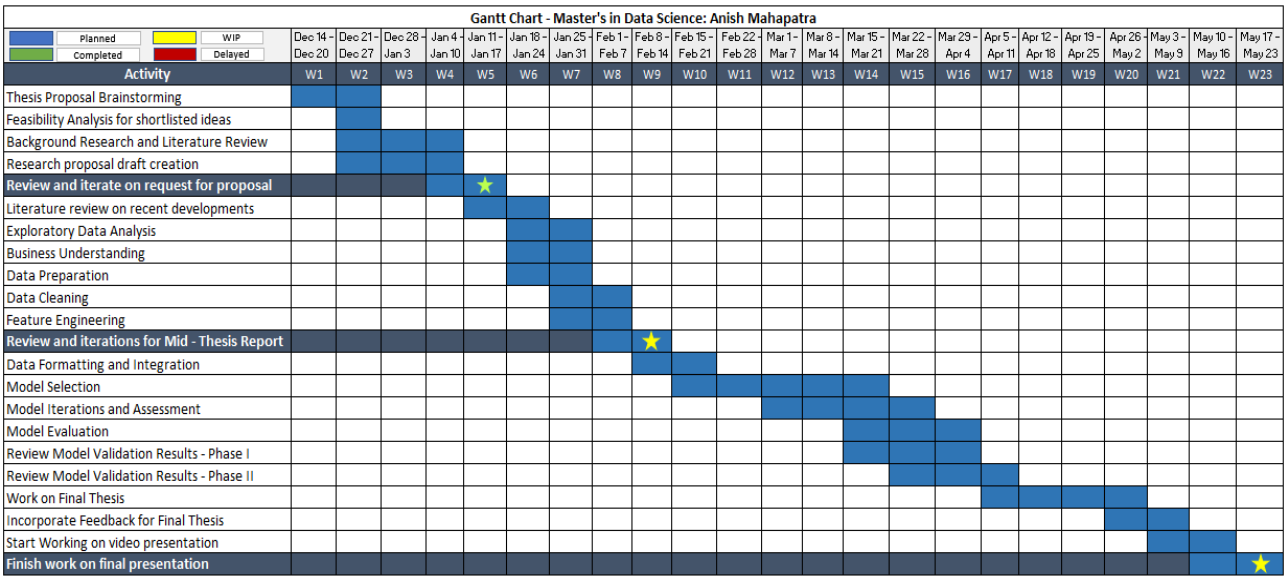


Figure 2: Research Plan and Timelines

APPENDIX B: RESEARCH PROPOSAL