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. Thus, 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. Accurate predictions coupled with business understanding from interpretable machine learning can revolutionize the telecom industry.

This research intends to build a predictive framework that can predict churn accurately and identify interpretable behaviour patterns using interpretable machine learning models 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, model tuning, and interpretable machine learning can predict churn accurately.

Keywords: Machine Learning, Churn, Classification, SHAP, LIME, Ensemble Models

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
ANOVA	Analysis of Variance
AUC	Area under ROC Curve
CRM	Customer Relationship Management
CV	Cross Validation
EDA	Exploratory Data Analysis
GBM	Gradient Boosting Machine
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
LIME	Locally Interpretable Model-Agnostic Explanations
PPforest	Projection Pursuit Random Forest
ROC	Receiver Operating Characteristics
RBF	Radial Basis Function
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-NC	Synthetic Minority Over-sampling Technique for Nominal and Continuous features
SVM	Support Vector Machine
SVC	Support Vector Classification
XGBoost	Extreme Gradient Boosting

CHAPTER 1: INTRODUCTION

With the increase in the number of options consumers have in the Digital Age, a company needs to keep costs low and profits high for a company to be successful. 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, a company needs to keep costs low and profits high for a company to be successful. 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 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 try 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. There are likely to be a few significant indicators of why the customer is willing to take the active step of moving across service providers in the customer's behaviour patterns. 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 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 1.1: 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 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 must 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 in Section 2.7, and 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 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 churn has been given along with the novel methods that solve 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 reduce customer churn, 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 2.1 showcases a tree-based visualisation to showcase the same. In this literature review, the focus will be on churners that churn voluntarily; it is difficult to flag whether the churn was incidental or deliberate every time.

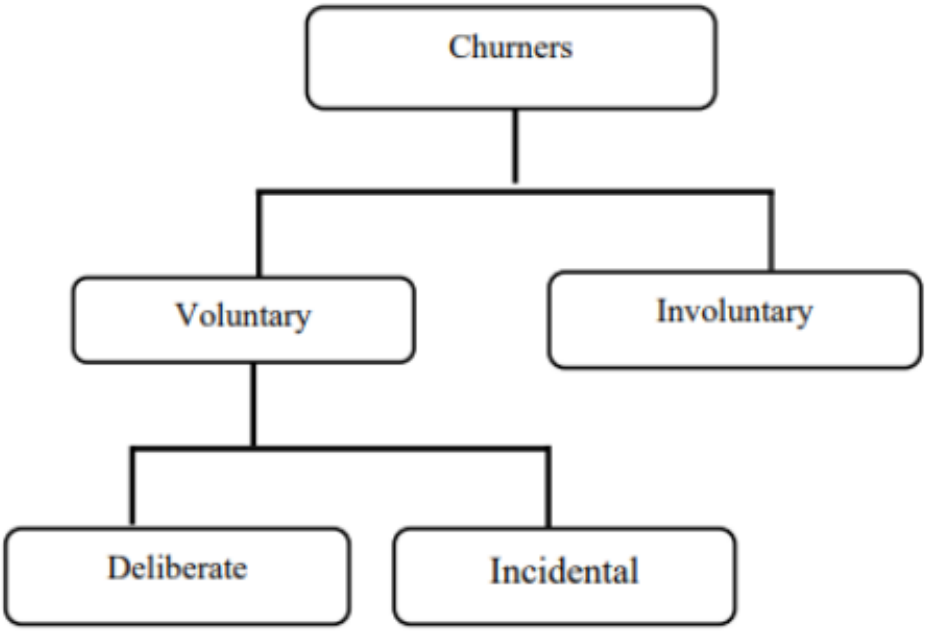


Figure 2.1: Types of Churners (Saraswat, S. & Tiwari, 2018)

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 leverage 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 visual analytics is also used to visualise large data sets. 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 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 for the end-customers, and 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 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 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 should a company 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 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 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 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 involves 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 classification problem predicts one of two things; if a customer will 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. Gravitational Search Algorithm 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 data type, 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, 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 2.2 illustrates the relationship between data, visualisation and models with the intermediary knowledge gained from visual analytics (Yuan et al., 2021).

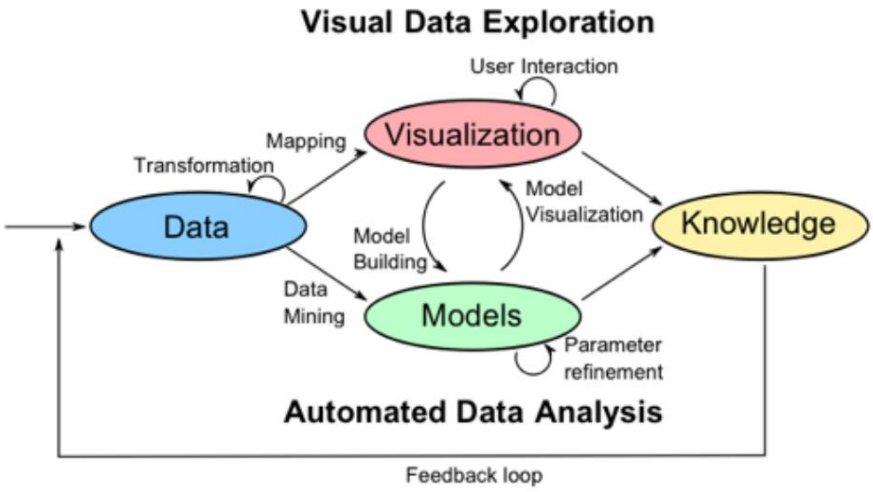


Figure 2.2: Visual Data Exploration

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 review 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, keeping the dataset and the predicted model's accuracy in mind is only done. When performing feature engineering on a dataset, another critical task is identifying 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 that 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 researchers have used other methods to tackle the class imbalance problem in telecom-based datasets: 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 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 enhance 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 in the industry where the result and 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 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 - evaluating the classifiers' performance in the below section (Halibas et al., 2019).

- 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.

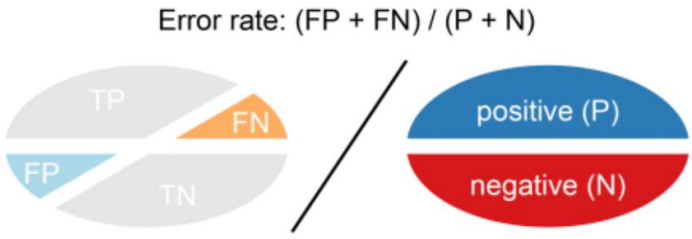


Figure 2.3: Visual Representation of Error Rate

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 assessed to understand 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. More importance can be given to feature engineering from the above section, 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, there are various ways to identify the customers at a high risk of churn. 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 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 and more advanced feature selection methods.

Table 2.1: 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 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.

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 a baseline understanding before deciding the appropriate research methodology for our use case.

CHAPTER 3: RESEARCH METHODOLOGY

This chapter is dedicated to the research methodology that will be used with the IBM Watson Telecom dataset. From the literature reviewed and the understanding of the telecom business, the customers that are at a high risk of churn will be flagged. Learning from the literature review will be applied in this research methodology in the sections of data preprocessing, feature engineering, predictive framework, evaluation metrics and interpretale machine learning. An accurate process flow to flag customers at a high risk of attrition with the underlying mechanism will be explained.

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: data selection, data preprocessing, data transformation, 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.

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: 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

The above description shows 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 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 to be implemented will ensure that the data is standardised as 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 feature engineering step. Outlier analysis is performed, and an analysis of the data's skewness to understand the feature's impact on customer churn. After understanding each features' distribution, a 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 heuristics 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.

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.

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. A baseline model will be set, post which both individual and ensemble models will be implemented to assess intitial performance. This can include monitoring p-values, the number of features, model performance, variance inflation factor scores which would differ across models. Post implementing hyperparameter tuning, the models will be assessed for overfitting or underfitting on train, test and validation data. To tackle the issue of overfitting, a cross validation strategy will be levergaed. 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. An interation will also be done after the data has been oversampled to perform class balancing and the appropriate results whill be chosen based on multiple evaluation parameters.

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 the perspective of a Data Scientist, 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. 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 customer behaviour patterns linked to churn's high likelihood. 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.

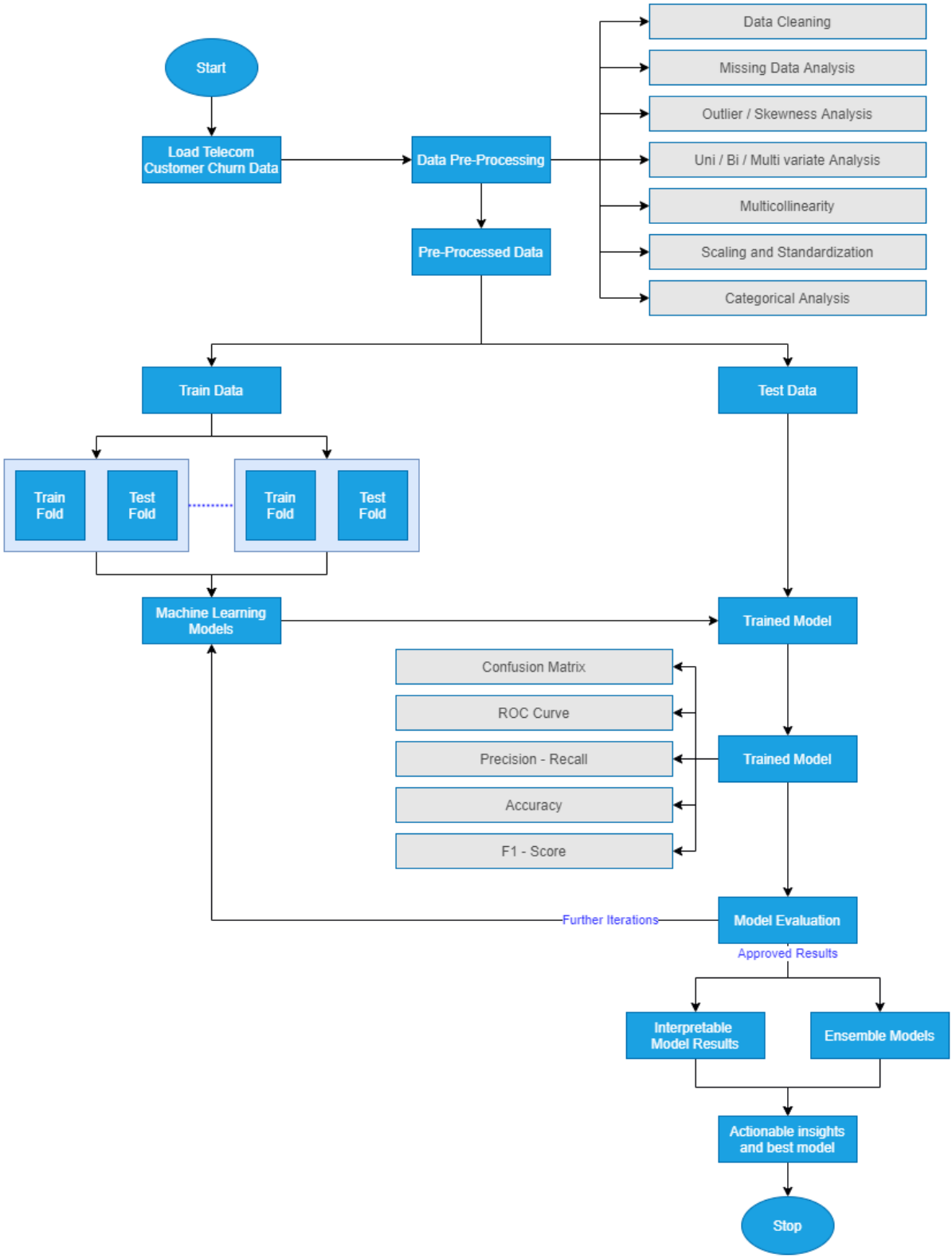


Figure 3.1: Model Building Process

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

A comparison of the model results will be made based on the metrics obtained from the literature previously surveyed. They used the same accuracy metrics, 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 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 will be decided to finish the research project. 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. 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 a business explanation and technical justification.

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 has been collected for over a month. If the customer has left in the last month, they have been flagged as a churned customer; else, they have been marked as not churned. 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 marked as churned customers were 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 indicate that 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 data details and how the data can be used as input 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. Following an approach of explaining all the details will help researchers eliminate any confusion regarding the details of the data. As the next steps, observations on the distribution of the variables will be done. The missing values in the data and the outlier analysis will help analyse the minute details of the dataset—univariate analysis on the data attributes 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 churn customers and do not churn.

4.3.1 Distribution of Variables

In this section, the distribution of variables will be understood. The percentage distribution of variables and absolute distribution will be analyzed through the visualizations in Figure 4.1. The distribution in Figure 4.1 highlights the distribution of the variables by percentage. The chart 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.

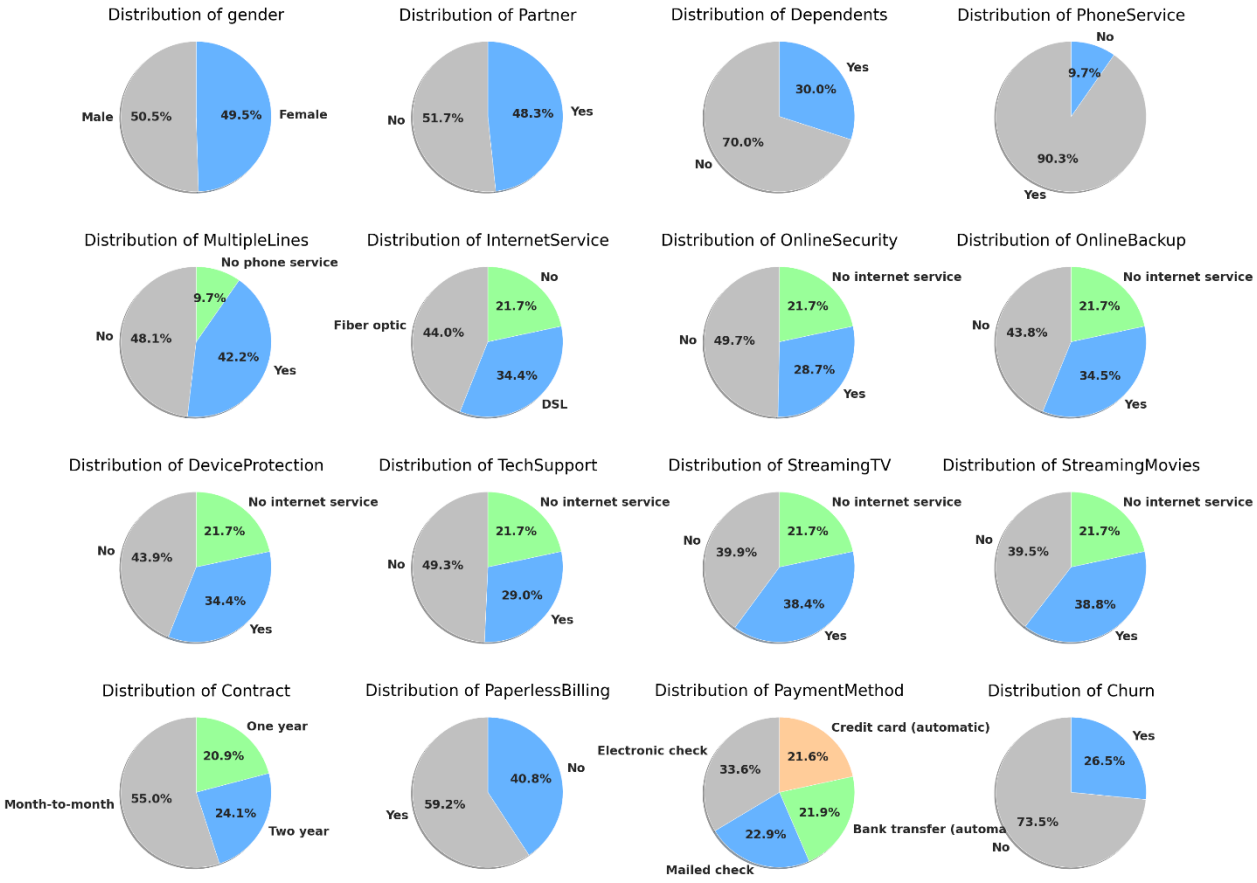


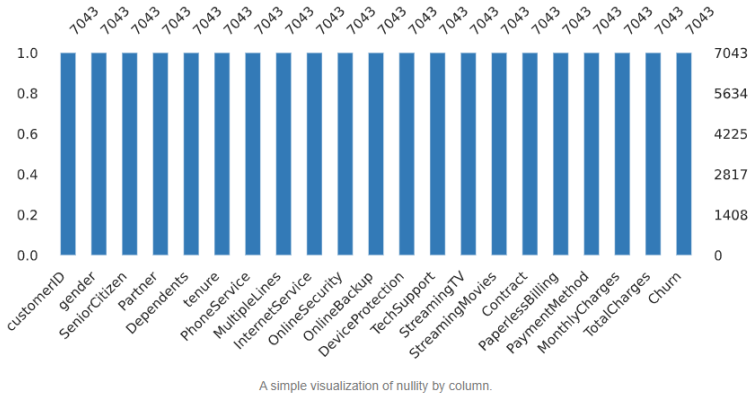
Figure 4.1: Distribution of variables (by percentage)

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.

The customer base with no internet service is 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 two-year contract. In Figure 4.1, 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. A 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 4.2: No missing values - Nullity by column for IBM Teleco Data

Post analysis from Figure 4.2 shows no missing values in the IBM Teleco Dataset from the missing value analysis using the nullity matrix. 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 (Interquartile Range) 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.

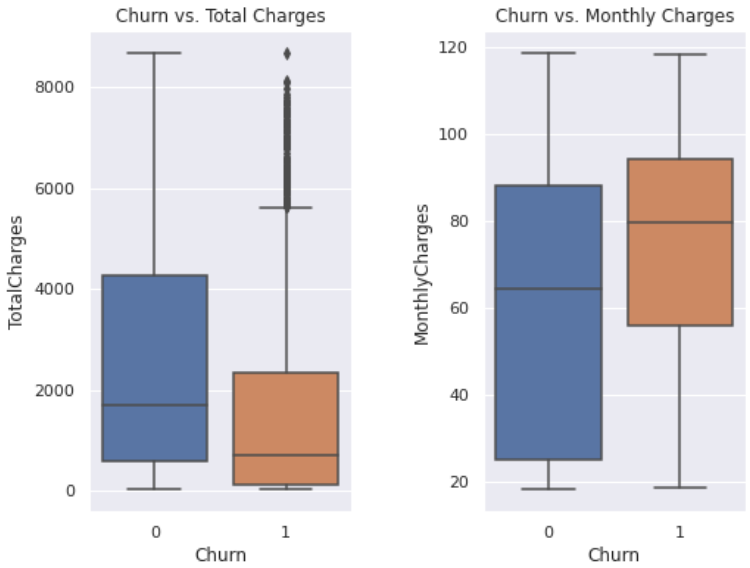


Figure 4.3 Boxplots of Churn versus Total Charges and Churn versus Monthly Charges

The distribution in Figure 4.3 shows that for Total Charges, most customers have a customer lifetime value of less than 2000, which indicates that customers who have a lower tenure with the company are likely to churn. Whereas, in the boxplot of Monthly Charges, the distribution of customers that churn is populated between 60 to 90. Customers that have lower monthly charges are less likely to churn.

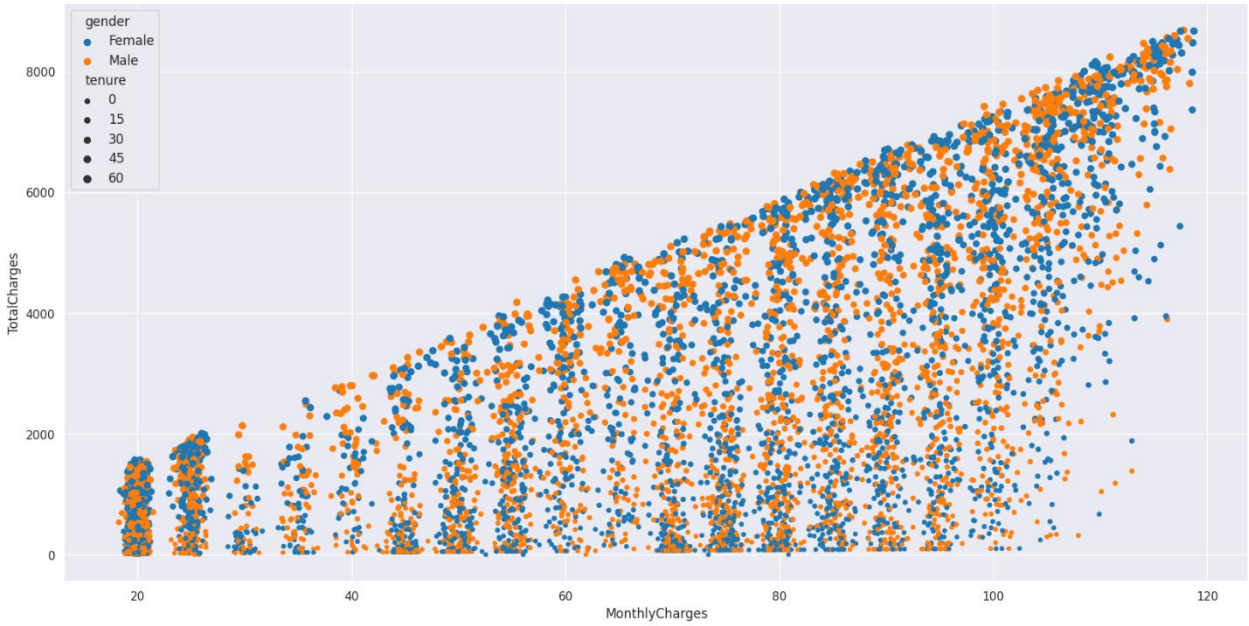


Figure 4.4 Scatter plot of Monthly Charges versus Total Charges

There is a significant correlation between Monthly Charges and Total Charges, as expected. As expected, Figure 4.4 illustrates that as the monthly charge per customer increases, the total charges or the customer lifetime value to the company increase. 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

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. 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.

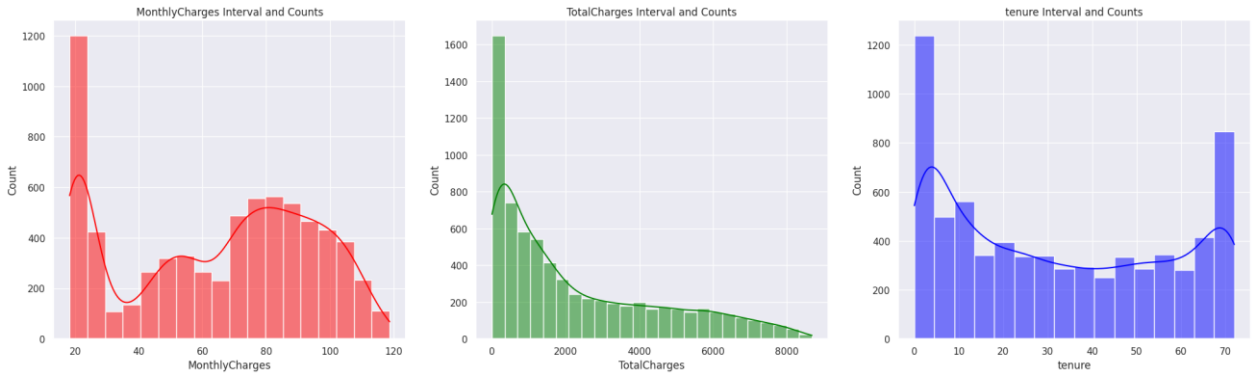


Figure 4.5: Univariate Analysis of numerical features of IBM Teleco Data

The visualization in Figure 4.5 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. A deeper understanding of the behaviour or churn is observed when visualizations are used. The relation of the demographic variables with churn can be seen in Figure 4.6. It is observed that customers that do not have a partner or dependents are more likely to churn. This churn 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 4.6 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.

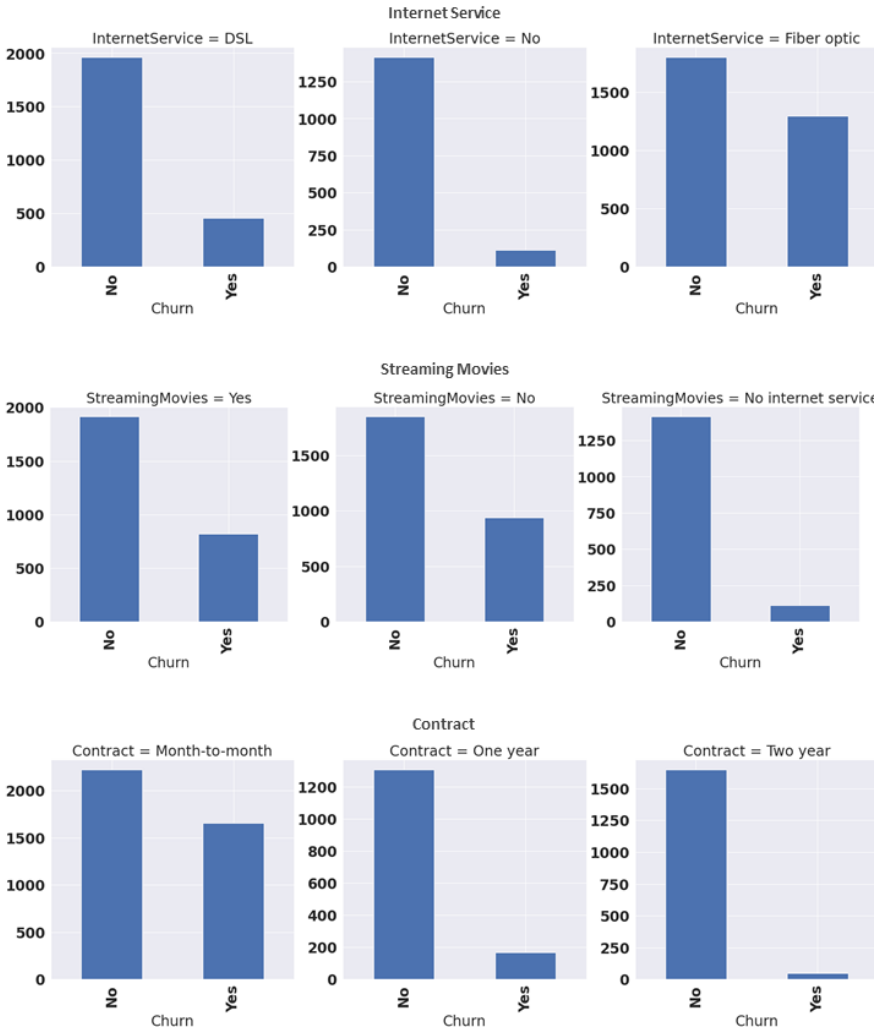


Figure 4.6: Internet Service, Streaming Movies and Contract plotted with respect to the target variable - Churn

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 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 relationship between quantitative variables and the correlation between qualitative/ categorical variables will also be plotted. The correlation between the variables in the dataset can be understood using these plots.

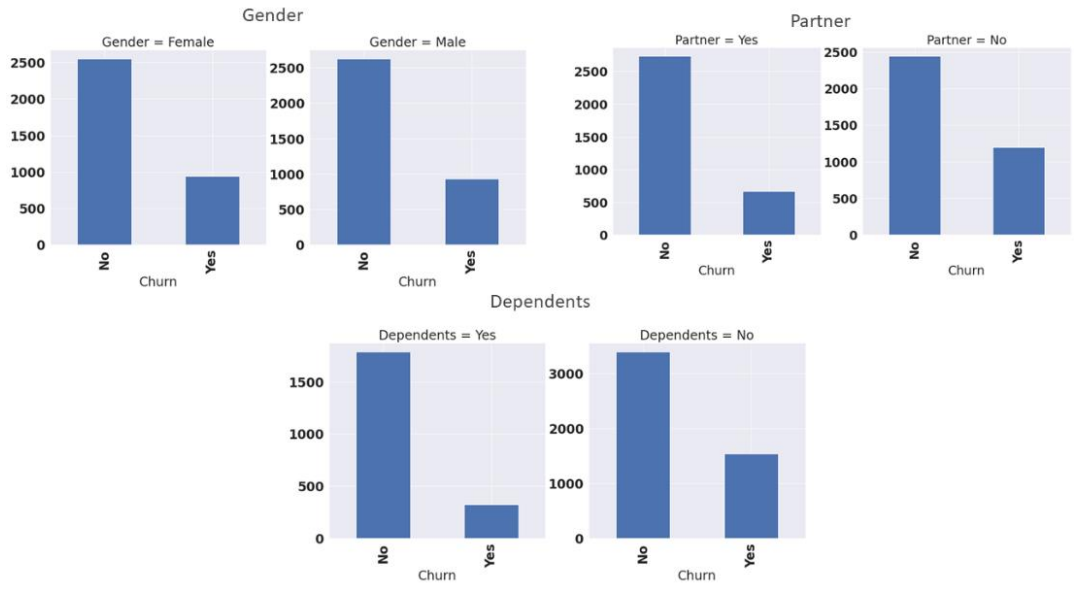


Figure 4.7: Distribution of Demographic Attributes with respect to Churn

4.3.6 Distribution of variables with respect to Churn

In Figure 4.8, 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 likely churn patterns. Some of the observations made from the visualizations can be confirmed in the feature selection techniques. The number of males and females churning is equal. Most people that churn do not have dependents. Customers with dependents are less likely to churn as they have settled into the ecosystem and do no change it unless there is something significant. Customers who use tech support are less likely to churn as they are willing to work with the telecom operator to fix the issues they may face. Electronic check is the most used payment method by the customers that churn – this signifies the method that has the maximum friction compared to an automatic deduction. Customers that do not use online security are also more likely to churn. A combination of these behaviour patterns can be observed from the charts that have been plotted in Figure 4.8, where the distribution of churned and not churned customer gives us insights into customer behaviour. Analyzing customer behaviour gives us insights that the models in the next phase can

use.

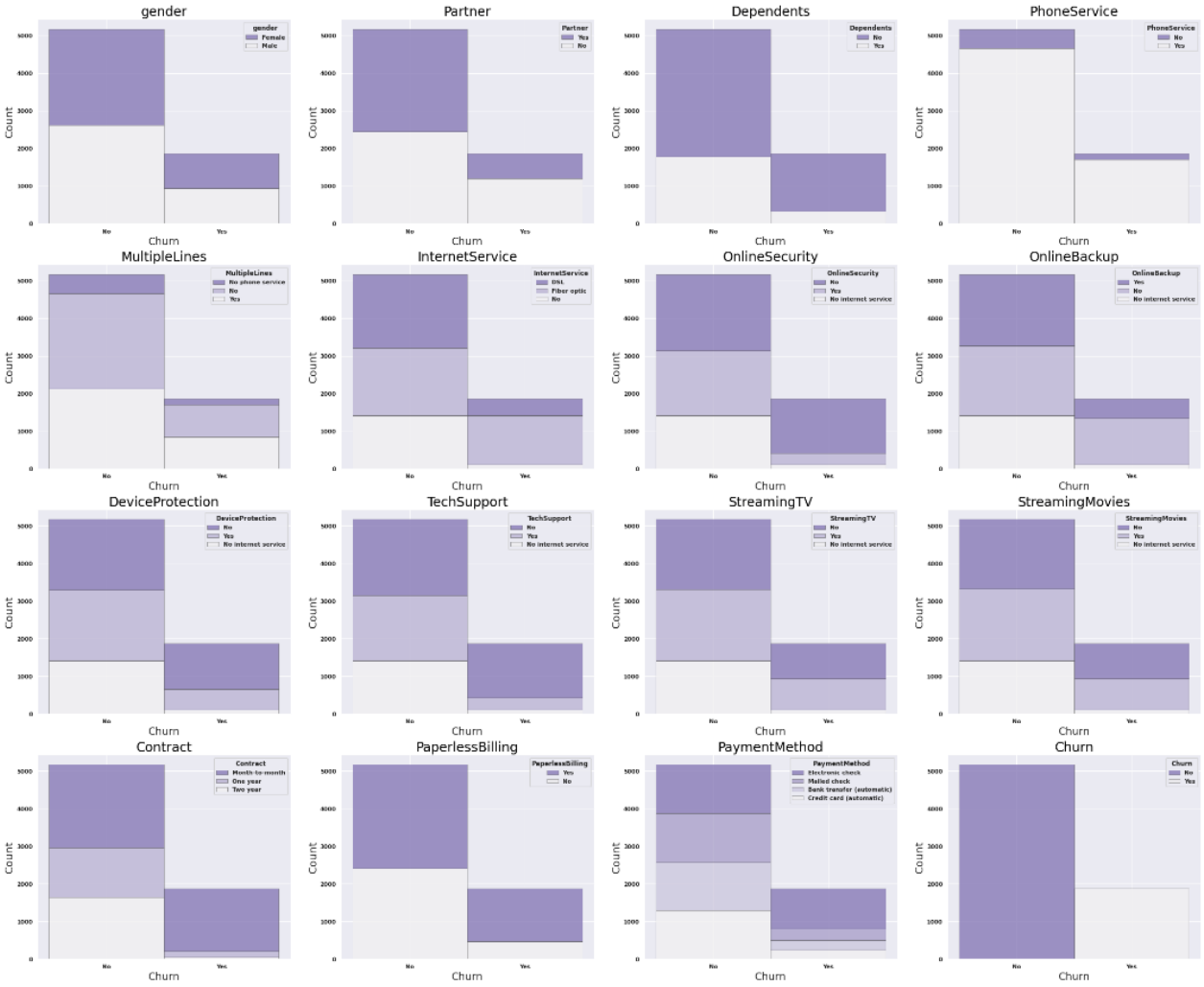


Figure 4.8: Distribution of all features with respect to Churn

Clients on a monthly contract are more likely to churn as it is easier to move out when there is no long-term commitment with the telecom operator. Customers that opt-in for multiple lines are more likely to churn. If the customer has a higher monthly billing, the customer is more likely loyal to the telecom operator. Instead, if a customer has multiple lines, the issues that may be faced with one connection gets amplified over multiple lines. Customers that do not have a phone service, which implies that only the internet facility is subscribed to, are less likely to churn. Telecom operators face stiff competition in phone services, which is showcased in the chart. The observations from the chart will be leveraged in the model building phase to drive better insights

for the business using interpretable machine learning.

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 4.9, 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. 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.

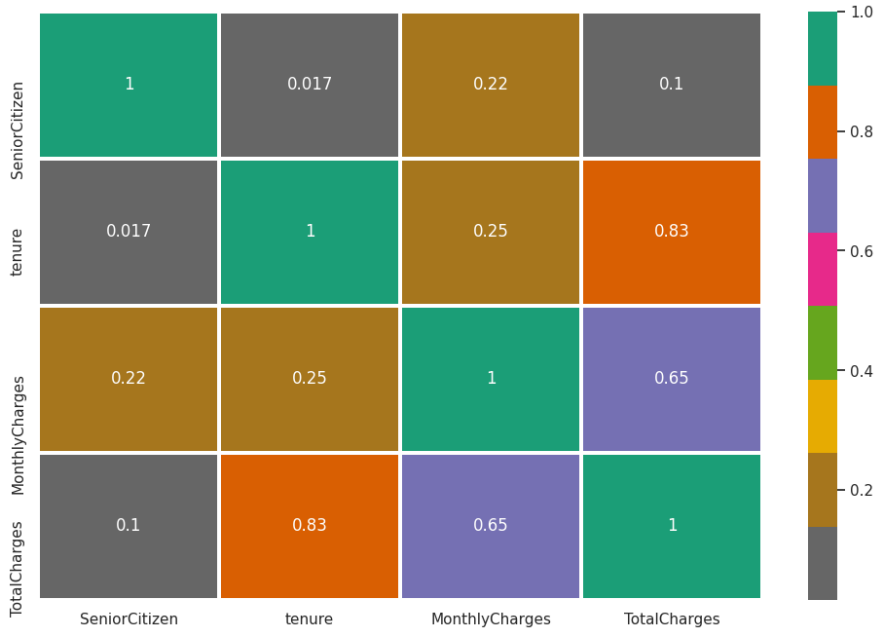


Figure 4.9: Correlation between quantitative variables

In Figure 4.10 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 4.9.

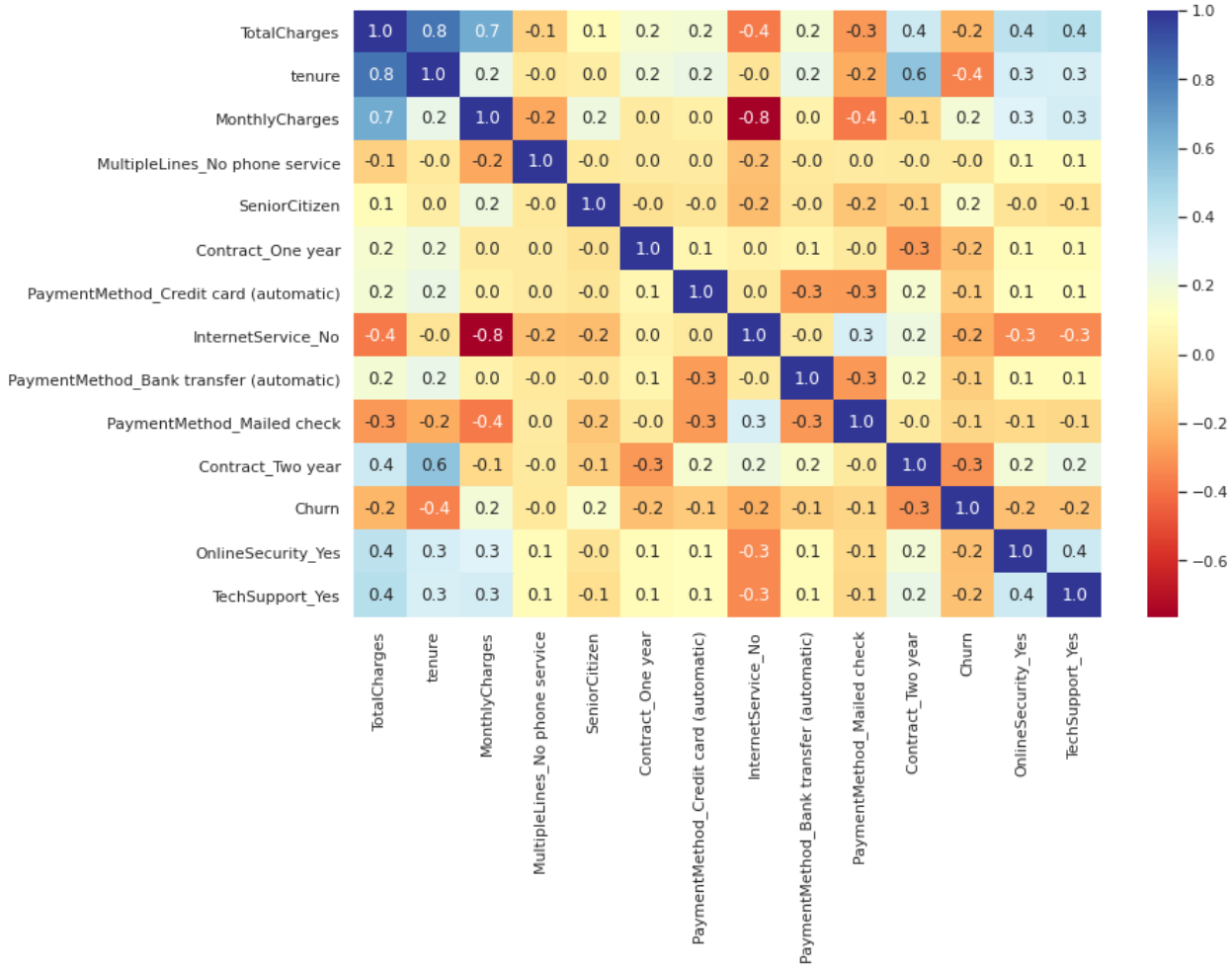


Figure 4.10: Correlation between qualitative variables

4.3.8 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 concerning the class attribute is calculated. It can only be used on categorical features. Based on the chi-squared weights assigned, the top 20 features and the weights will be showcased in Figure 4.11. Chi-Square tests are used to determine if the attributes are dependent or independent of the target variable. If the variable is independent, the feature may be removed from the dataset depending on further analysis. The Pearson's Chi-square test is used as a statistical test to test the independence of categorical variables. The alpha value to be chosen for the attributes is 0.05.

Features	Chi2Weights
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
PaymentMethod_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 4.11: Top 20 features based on chi-squared weights

Based on the weights obtained from the initial statistical test, the chi-square analysis, the dependent features will be considered relevant for the next steps. The categorical variables that have been marked as independent variables based on the decided alpha value 0.05 will be analyzed further in the following steps.

4.3.9 ANOVA Test

The ANOVA test is used to find out if the relationship between a numerical and an absolute value is statistically significant or not. The null hypothesis is that two groups have the same variance, and the alternate hypothesis is that at least one in the group has a different variance. If the value of the variance between both groups is the same, it indicates that the feature is not essential.

score	columns
291.625610	TotalCharges
273.463959	MonthlyCharges

Figure 4.12: ANOVA Test to determine significant features

If the p-value generated from the analysis is less than 0.05, it indicates that the confidence for the variable is greater than 95%. The variables belong to the same population and hence, are correlated. The top 2 features are selected and test for statistical significance, where the p-value is less than 0.05. Total charges and monthly charges are the essential features.

4.3.10 Probability Distribution using KDE

The probability distribution is plotted using the kernel distribution estimator. The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. The probability distribution of a continuous random variable is known as a probability distribution function.

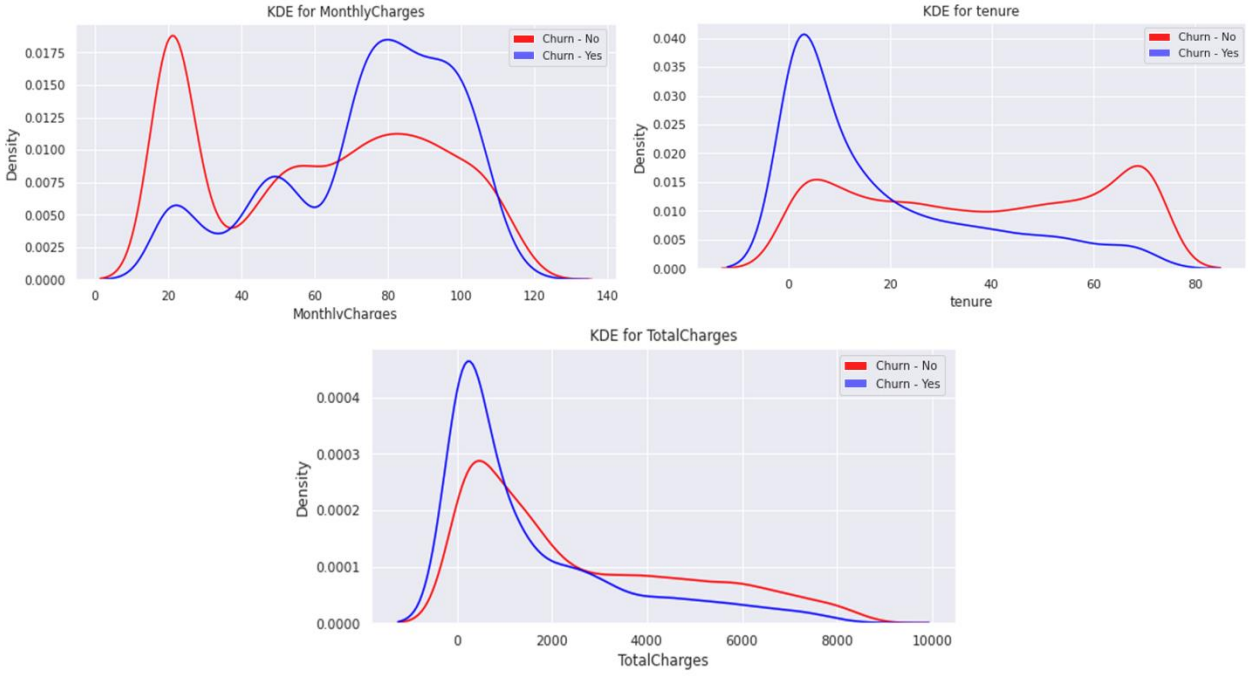


Figure 4.13: Probability Distribution using KDE for numeric attributes

As depicted in Figure 4.13, the KDE method is a non-parametric density estimator function used to fit a model to the arbitrary distribution of the data, like the kernel density estimator. KDE uses a tree-based algorithm and trades off computation time for accuracy using absolute tolerance and relative tolerance measures. The kernel bandwidth is determined using cross-validation.

4.5 Methods

In this section, the discussion will be around the methods and standards that will be leveraged in 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 to predict the customers 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. The split will be done in a stratified manner by the train-test package 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. Due to the high cardinality of certain features such as Customer id, the column was discarded as it is computationally expensive, increases the data size, and does not add any additional value.

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.

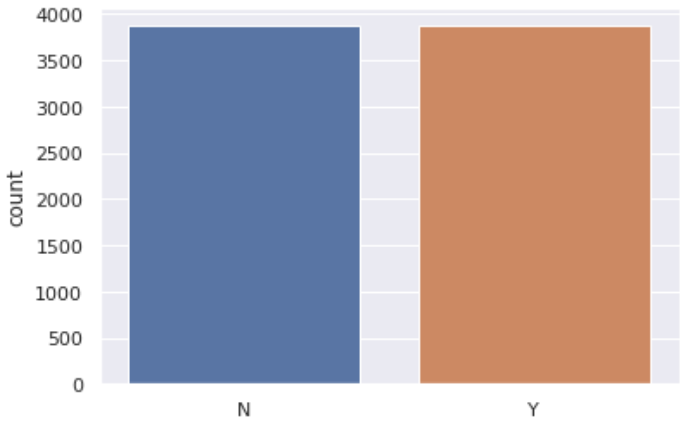


Figure 4.14: Plot of train data after SMOTE-NC is applied

4.5.6 Hyperparameter tuning

Once individual and ensemble modes have been trained on the data, it is advisable to proceed with hyperparameter tuning. One of the methods that are used is random search for hyperparameter optimization. The search space is defined as a bounded domain of hyperparameter values, and the random search algorithm randomly samples and evaluates values within the defined domain. The randomized search cross-validation function is used to run through a defined sub-space with 1000 iterations. One notable run was when randomized search cross-validation was run on the Light GBM model, and the AUC improved to 0.83. Different algorithms used appropriate methods of hyperparameter tuning. After every iteration, the top three models were chosen for hyperparameter tuning – this was done using grid search and random search through the predefined space. The optimization and results for the iterations carried out after hyperparameter tuning can be seen in Chapter 5.

4.5.7 Implementation

All of the analysis and implementation has been done on Google Colab. The virtual machine's configuration 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. 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 and implementation of research models will be decided. The models that can be implemented in Chapter 5 in the results and discussions section will also be discussed. The results of the methods and models implemented will be detailed out in the next section. It 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.

4.6.1 Baselines

In order to evaluate the models effectively, a baseline model will be set. For this study, the models selected as baselines are logistic regression and decision tree model on the dataset where one-hot encoding has been performed. The models 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 and understand if the model's performance increases when different techniques are applied to the dataset.

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. The models 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, and 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 4.15 and Figure 4.16.

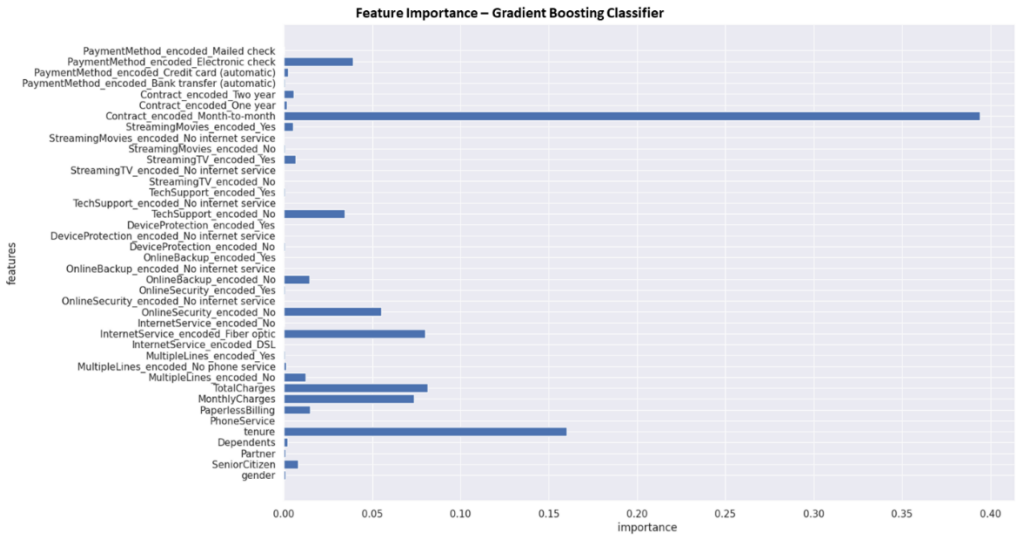


Figure 4.15: Feature Selection using Gradient Boosting Classifier

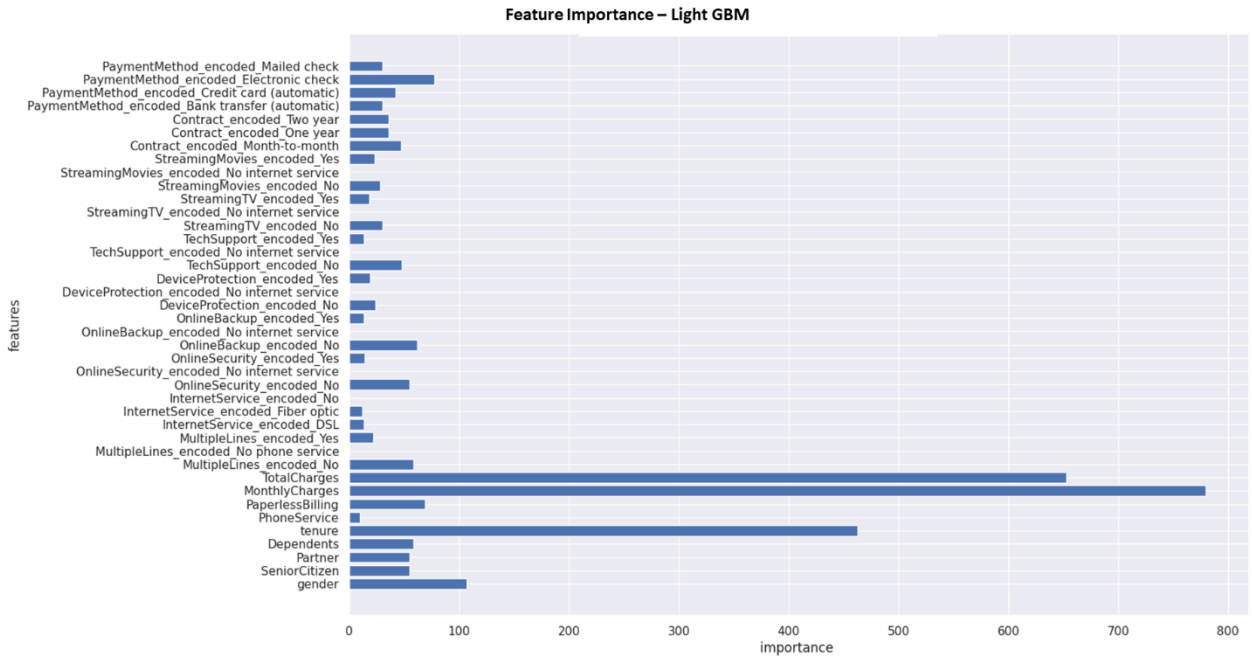


Figure 4.16: Feature Selection using Gradient Boosting Classifier and Light GBM

From the charts that showcase feature importance in Figure 4.15 and 4.16, 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. The 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 and evaluates 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 to the dataset – depending on the algorithm implemented, minute yet significant improvements in some accuracy scores are observed.

4.7 Model Interpretability

Businesses in the real world want to understand the reasoning behind model predictions. It not always possible with machine learning models, especially as models get relatively complex, the interpretability of the model decreases. Locally Interpretable Model-Agnostic Explanations (LIME) comes in here to help make machine learning more approachable for non-experts. 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 4, the analysis and techniques used to run classification models on the telecom data to predict if customers will churn were discussed in detail. 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

In this chapter, the results from work done in Chapter 4 will be discussed in depth. The discussion and the interpretation of the results are made in detail. By the end of the chapter, the work done to keep up with industry best practices will be showcased and possible results when the focus is put on following best practices. The analysis will be on the cleaned telecom data from IBM Watson, and the flow of the work will be explained in Section 5.1.

5.1 Introduction

The below sections highlight the results and discussion of the analysis performed in Chapter 4. In Section 5.2, the baseline results of the model will be discussed. The interpretation of visualizations, where the charts discussed in Chapter 4 will be discussed in greater depth. The results post cross-validation in Section 5.3, the individual model results, and the results after using SMOTE NC for class balancing are discussed. In Section 5.4, a model agnostic technique to showcase the interpretability of customers is showcased. Using LIME to improve real-world model interpretability is something that we have not seen in all of the papers surveyed. Let us now proceed to discuss the baseline results below.

5.2 Baseline Results

This section discusses the results of the baseline models without the application of cross-validation and class balancing techniques. The model results discussed have been provided after data cleaning, and one-hot encoding is done as discussed in Section 4.5.2. The models that have been chosen as baselines are logistic regression and a decision tree classifier. The decision tree classifier on the test data has an accuracy of 70.02% and an AUC of 0.65. The logistic regression model has an accuracy of 78.16% and an AUC of 0.71. These base metrics will help us evaluate different methods that have been highlighted in the sections below.

Suppose a model is performing better than the baseline. In that case, the methodology can be considered a good model, or else if the model's performance is noticeably worse than the baseline, the methodology can be rejected for this particular use case. The focus is the methodology and not necessarily the result, so each of the results will be discussed in detail to gain intuition.

5.3 Interpretation of Visualisations

In Section 4.3, where exploratory data analysis was done on the telecom dataset, multiple visualizations were plotted to get a more profound intuition of the dataset and the relation of the attributes with the target variable. The most vital indicators of customer churn are total charges and monthly charges. The customers that have the highest monthly charges generally have internet service, as we noticed from the heatmap, as they have a high correlation of -0.8 for no internet service as seen in The observations from the chart will be leveraged in the model building phase to drive better insights for the business using interpretable machine learning.

4.3.7 Correlation; this indicates that when a customer does have internet service, the monthly charges are higher, and hence, the chance of churn is lower. The customer is less likely to churn because they are deep in the ecosystem of the telecom operator, and hence, there is high friction to move to another telecom operator.

Another notable attribute that is an indicator of churn is when the mode of the contract is month to month. Customers who do not have a one-year or two-year contract have a higher tendency to churn. The distribution of monthly charges based on churn shows that the customers who do not have dependents are more likely to churn. Customers who have partners or dependents tend to be more stable in their choices and do not have time to look out for offerings from other telecom operators. The interpretations are primarily from feature importance graphics. A high positive correlation between monthly charges and total charges is observed with an equal distribution of male and female customers in Figure 4.4 Scatter plot of Monthly Charges versus Total Charges. This correlation indicates that as monthly charges increase, the customer is more invested in the subscription they have opted for, thus indicating a significantly higher customer lifetime value. Customers who opt-in for more services such as streaming movies and online security are less likely to churn. One of the primary interpretations is that the more services the customers use from the telecom operator, the more likely they will be loyal. From the visualization, it is understood that gender is not a good indicator of churn.

Customers that do not have partners or dependents are less likely to churn. Customers with a monthly contract, customers with internet available, opt for paperless billing, and automatic payment services are more likely to churn due to customers being more technologically adept and updated on the latest market trends. Customers who leverage premium streaming services are more likely to leave, maybe due to competitors offering better quality and better price competitiveness. Additionally, customers that had multiple lines were more likely to leave. Repetitive failure from the company in satisfying customer needs can lead to customers with multiple lines churning. A few more observations were made from the correlation plots, such as senior citizens have a higher chance of having dependents and people without partners tend not to have any dependents. Having multiple lines from the telecom operator is not a strong indicator that the customer is loyal. Customers who opt-in for paperless billing tend to utilize internet service, and those that have an internet service prefer automatic transfers, especially those with a fibre optic subscription. The

customers that do not have internet services preferred to used mailed check services instead. People prefer to use manual transfer, perhaps due to safety concerns and the lower cost, compared to automatic transfers, regardless of age.

From the scatterplots, it was understood that clients with a lower tenure are more likely to churn. Customers with a higher monthly charge are more likely to churn. Tenure and monthly charges are the most significant features for predicting the customers who are likely to churn. In Section 5.4, the model results will be discussed in detail.

5.4.1 Model Results

In this section, the model results of all of the methods have been discussed in detail. The models will be compared with the baseline models discussed in Section 5.2, where they will be compared on various evaluation metrics. Wherever possible, a visual analysis of the model results has been provided to get an intuition of model performance. In the following sections, the results from individual models as well as ensemble models will be analysed. In Section 5.4.2, the results of the model post-cross-validation will be analyzed. Finally, a deep dive on models post-class balancing using SMOTE-NC will also be done.

5.4.1 Individual Models and Ensemble Models

In the following section, individual model results will be analyzed. Along with the individual model results, the ensemble model results have also been analyzed. After exploratory data analysis was performed on the dataset, missing value analysis and outlier analysis was performed. One-hot encoding was performed on the categorical attributes present in the dataset. Box cox transformation was done on the skewed variables, and the best results were taken into consideration. The train and test scores of the models were analyzed and plotted in Figure 5.1.

Table 5.1: Model Results of Individual and Ensemble Models

Model	Train (%)	Validation (%)	Test (%)	AUC
Gaussian Naïve Bayes	70.38	70.03	67.52	0.73
Bernoulli Naïve Bayes	72.93	72.87	69.36	0.74
Logistic Regression	80.8	81.39	78.3	0.71
Random Forest	99.75	79.83	75.04	0.67
Support Vector Machine	81.82	81.25	77.45	0.68
Decision Tree	99.75	75.85	70.21	0.65
K Nearest Neighbour	83.51	74.86	70.21	0.65
Gradient Boosting	83.3	79.83	77.16	0.69
Stochastic Gradient Descent	76.98	79.83	73.62	0.72
Light Gradient Boosting Machine	88.34	79.12	76.31	0.68

Table 2.1 is a table showcasing the different models leveraged to perform classification on the telecom dataset. The accuracy values on the test data, as well as the AUC values, are highlighted. It is observed that the model with the highest accuracy is logistic regression, with an accuracy of 78.3% with an AUC score of 0.71. The model with the highest AUC score of 0.74 is Bernoulli, which has an accuracy of 69.36% on the test data. The results of every model used, both the individual model and the ensemble models such as Gradient Boosting, Stochastic, and Light GBM, have been used for this study. The visual representation of the scores can be seen in Figure 5.1, where it is noticed that the decision tree and random forest classifier may be overfitting the data. Overall, all of the individual and ensemble models have an accuracy score above 67%, and this is an indication that models are getting trained satisfactorily for a preliminary run on the IBM Telecom Churn dataset.

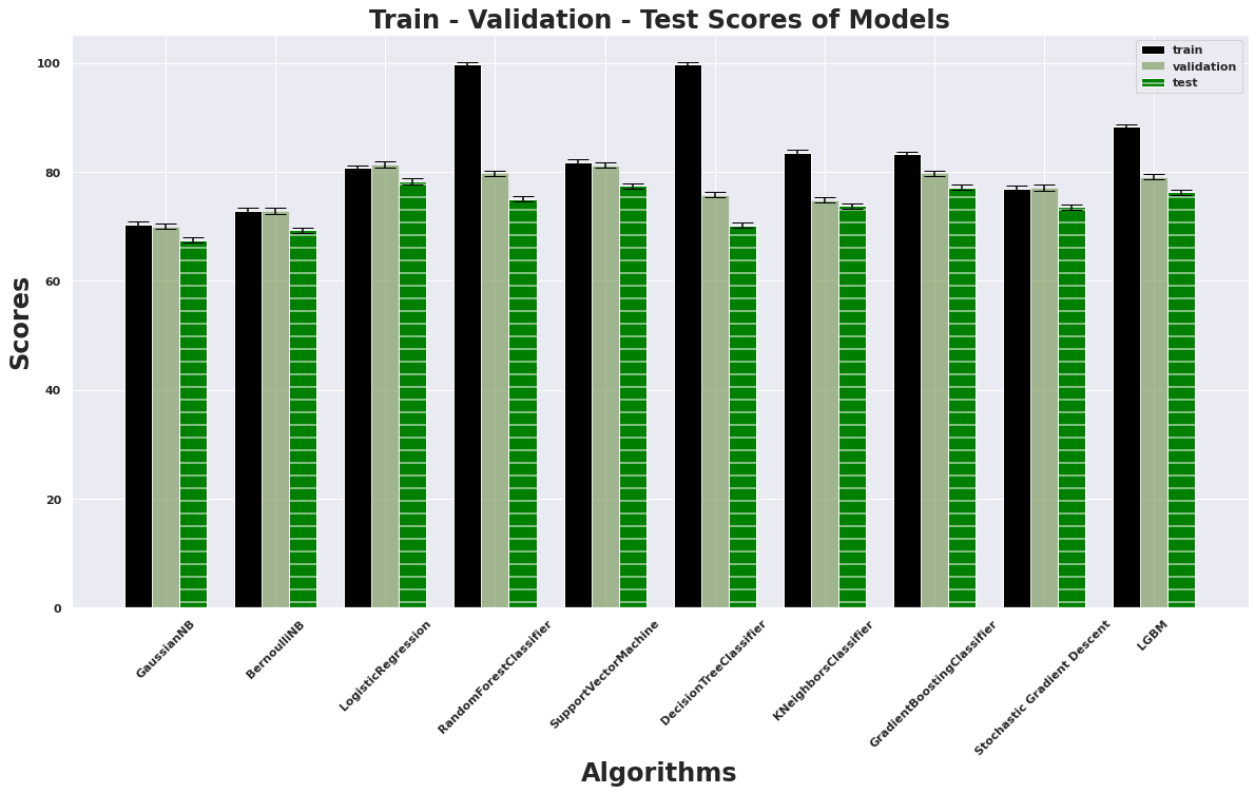


Figure 5.1: Train-Test Scores of Models

Since the data size is not large enough, the results of the breakdown of the validation dataset have not been included. It attempted to break down and perform the same exercise with a train, test, and validation dataset. To overcome the problem of overfitting for specific classifiers, in Section 5.4.2, the results of the models after cross-validation has been performed will be discussed. This section summarizes the model results from assessing individual models and ensemble models, where it has been observed that ensemble models have similar performance.

5.4.2 Cross-Validation

In this section, the results of the models after cross-validation has been done will be discussed. For this, the cross-validation package from sklearn has been leveraged, where the cross-validation strategy has been specified as ten. Compared to the accuracy scores on the test data by using only the individual models, we see an improvement in score for every model after cross-validation has been performed.

Table 5.2: Model results with Cross Validation

Model	Cross Validation Scores
Gaussian Naïve Bayes	70.45
Bernoulli Naïve Bayes	72.95
Logistic Regression	80.39
Random Forest	79.59
Support Vector Machine	79.77
Decision Tree	72.77
K Nearest Neighbor	75.67
Gradient Boosting	80.56
Stochastic Gradient Descent	76.54
Light Gradient Boosting Machine	79.85

The above table shows that the model with the highest accuracy scores is now an ensemble model, gradient descent, instead of the logistic regression model from the last model iteration. As expected, the most significant improvement is seen for the Random Forest model at 2.52% as overfitting is overcome, and the model is generalized now. There are two ensemble models, Gradient Boosting and LGBM, with CV scores of 80.56% and 79.85% in the top three models. The scores indicate that ensemble models perform better than the individual models for the use case of the IBM Watson Telecom datasets. It is further noticed that the model performance is noticeably better than the baselines as set in Section 5.2. The mean of the cross-validation scores is more than 70%, indicating that some models were overfitting due to the data size, which can be seen in Figure 5.1, the black bar.

When trained on the telecom data, all models have a test accuracy score higher than the previous iteration with no cross-validation in the pipeline. Higher accuracy scores maybe because the preprocessing and analysis performed on the data helped us understand the minute details of the dataset. Figure 5.2 depicts the mean of the cross-validation scores in a chart to visualize the test accuracy scores for all the algorithms implemented. As shown in Figure 5.3, the best performing models are Gradient Boosting, Logistic Regression, Light GBM, Support Vector Machine, and Random Forest. The general trend being observed is that the ensemble model where the pipeline includes cross-validation and hyperparameter tuning performs better on the data than other pipelines on the IBM Telecom dataset.

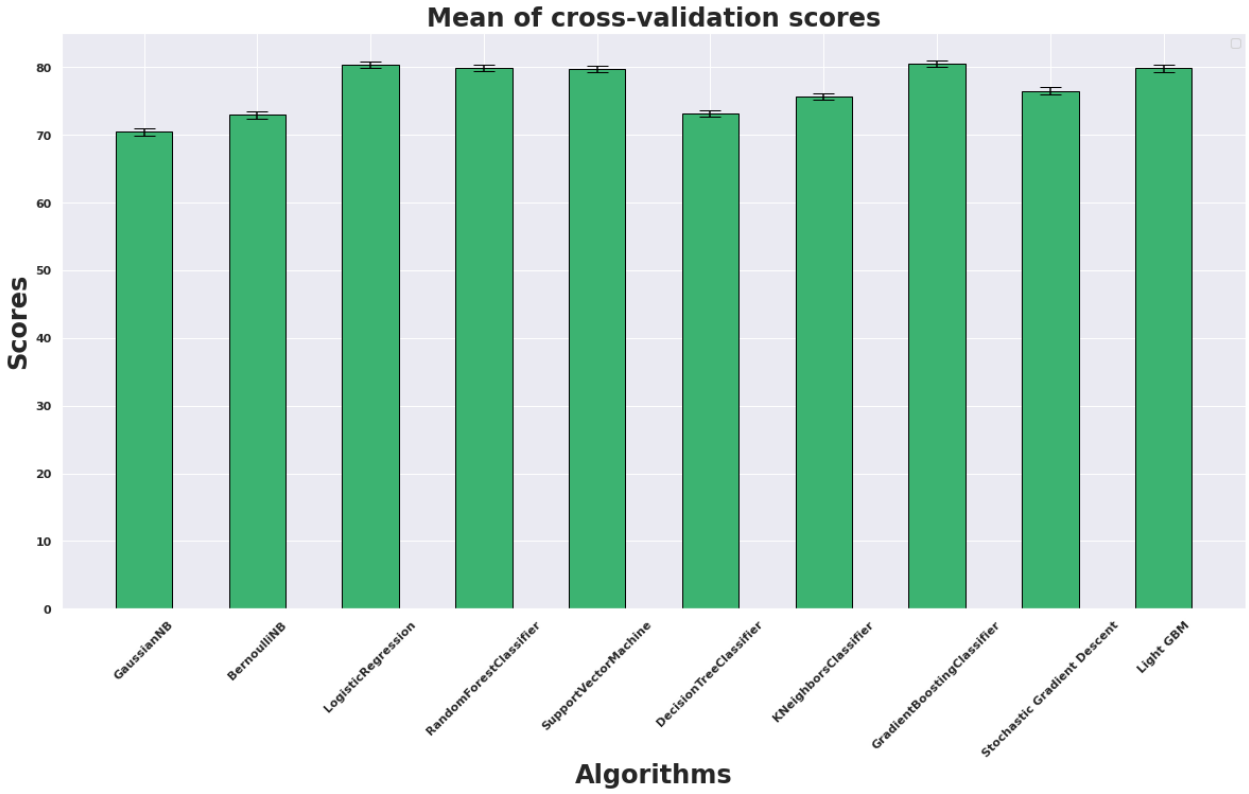


Figure 5.2: Chart depicting the mean of Cross-Validation Scores

Now that the accuracy scores of the model have improved after cross-validation, it will be attempted to see if we can get better scores. In the literature survey from Chapter 2, it was observed that better results were obtained when oversampling was done. The following section will analyze the models after class balancing has been done on the data.

5.4.3 Results after Class Balancing

In Section 5.4.1, the results were analyzed when individual and ensemble models were trained on the data. In Section 5.4.2, the models' results were observed after cross-validation was performed on the dataset. The data has an imbalance of churned and non-churned customers, where the percentage of churned customers is 26.5%, and the rest of the customers have not churned. The IBM Watson telecom data has many categorical variables, so the traditional methods were less effective. SMOTE-NC does not work with datasets that only contain categorical variables. Instead, SMOTE-NC is used with datasets that contain numerical as well as categorical variables.

The column transformer used is one-hot encoding, with standardization of the standard scaler's relevant attributes. From our analysis of model results in Section 4.5.2, it was noticed that ensemble models performed noticeably better than individual models. The literature survey in Chapter 2 observed that models performed better when oversampling was performed on the dataset. The models are now trained on cleaned data oversampled using SMOTE-NC, and the results can be found in the table below.

Table 5.3: Model Results after oversampling using SMOTE-NC

Model	Accuracy (%)	AUC
Decision Trees with Bagging	73.27	0.84
Decision Tree with AdaBoost	75.65	0.84
Logistic Regression	75.6	0.83
K-Nearest Neighbour	70.02	0.83
Random Forests	71.62	0.83
Linear SVC	75.77	0.83
CatBoost	76.45	0.83
Naïve Bayes	73.38	0.82
XGBoost	76.96	0.82
SVM with RBF Kernel	76.96	0.81
Decision Trees	74.23	0.79

The above table shows that the top three models with the best accuracy scores on the test data are Support Vector Machine with Radial Basis function, XGBoost and CatBoost, with accuracy scores 76.96%, 76.96% and 76.45%, respectively. It is also inferred that the top three models with the best AUC scores are Decision Tree with AdaBoost, Decision Tree with Bagging and CatBoost with values of 0.84, 0.84 and 0.83, respectively. From the literature survey, it was observed that ensemble models tend to perform better when compared with individual models.

It was also noted that ensemble models performed better than deep learning models that leveraged neural networks. From the results above, one of the standard models that have consistent performance across various evaluation metrics is CatBoost. An intuition of all the models leveraged has been showcased in Figure 5.3 for all the models. In Figure 5.3, the plots used to interpret CatBoost have been highlighted, including Feature Importance Measures to showcase the essential features, the Area under ROC curve to showcase AUC scores, and the precision versus recall plot.

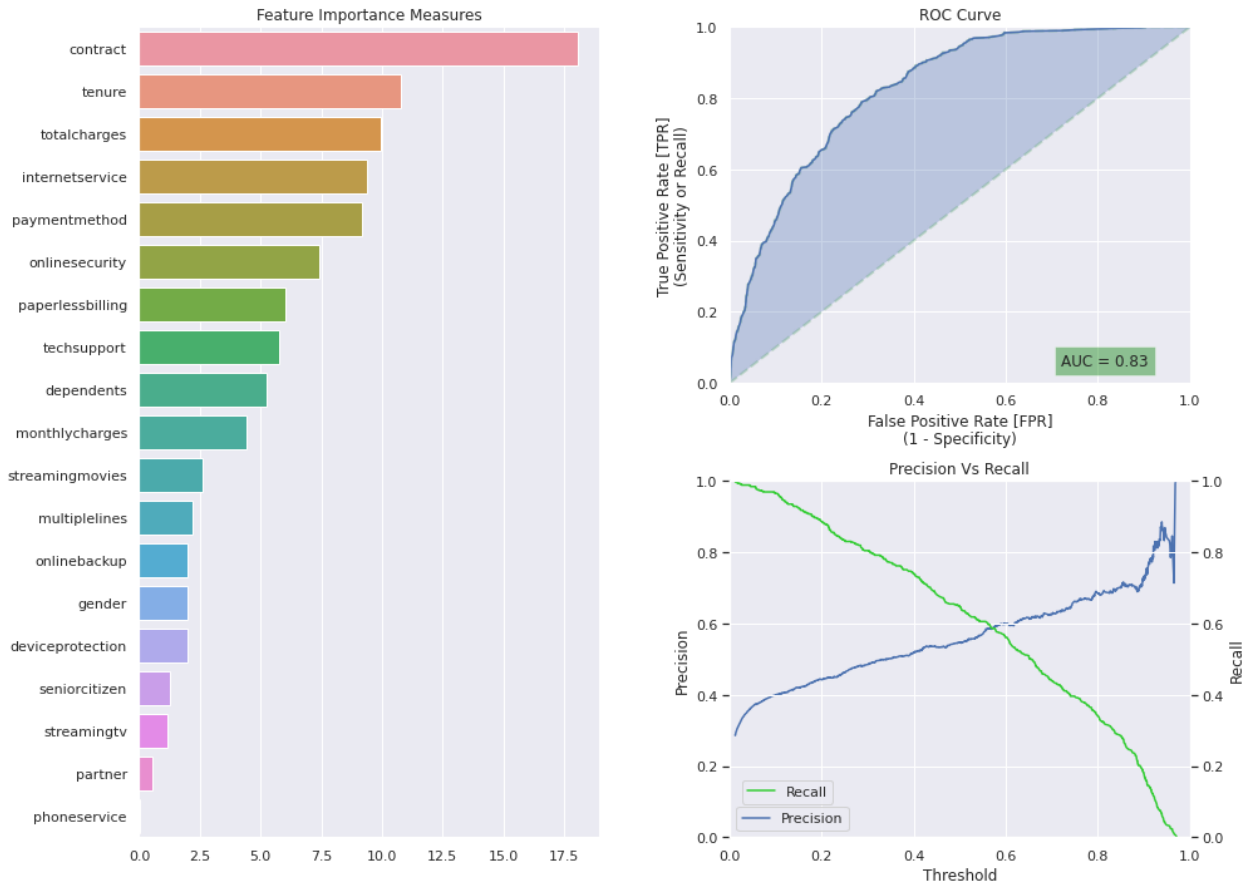


Figure 5.3: CatBoost: Feature Importance Measures, ROC Curve, Precision vs Recall chart

Plotting charts to analyze results for all models helps explain why the model performance excels or is not as expected. For instance, in Figure 5.3, the Feature Importance has been plotted – the plot indicates that the essential attributes the model has considered to get a high accuracy score and AUC score are contract, tenure, internet charges and payment method. From the Area under the ROC Curve plot, it is observed that the model has performed reasonably well with an AUC score of 0.83. The same set of model analysis charts have been plotted for all the models, as mentioned in Table 5.3. Many of the features that have been marked as necessary are familiar across Feature importance charts among the chosen models. In Figure 5.4, the ROC curve for multiple models has been plotted to gain better intuition on multiple model performance analysis. The same has been done for the precision-recall curve. Almost all the models, except Decision Trees at AUC at 0.79, have an AUC value greater than 0.80 - oversampling the dependent variable results in better accuracy and AUC values for the IBM Watson Telecom dataset.

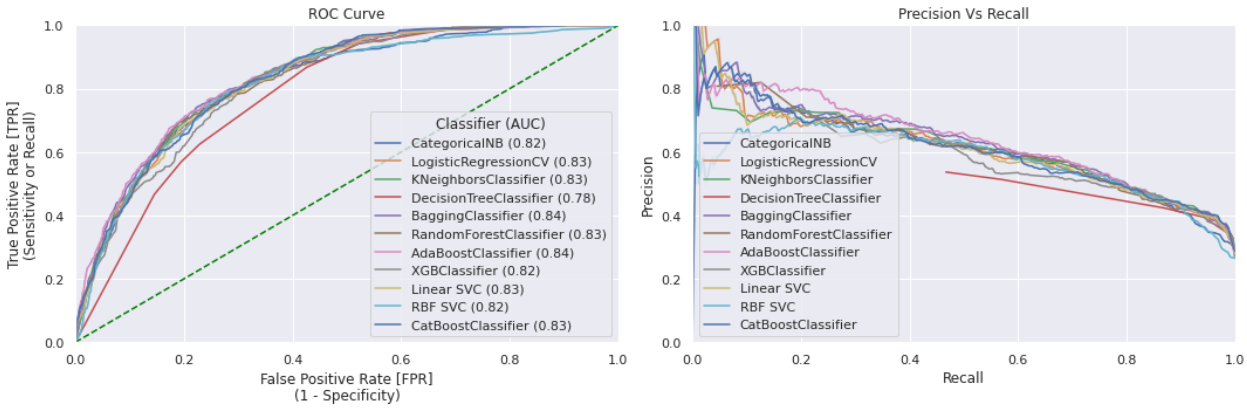


Figure 5.4: Multiple Classifiers - ROC Curve and Precision versus Recall Plots

The results are significantly better when the models with oversampled data are compared with the individual and ensemble models from Section 5.4.1. The results from multiple models have been plotted in Figure 5.4. Except for the red line, which indicates the Decision Tree model, it is observed that the rest of the models, as highlighted in Table 5.3, have good results that are significantly better than the baselines that were set in Section 5.2, without oversampling and cross-validation. Now that satisfactory model results have been achieved, the next section will focus on interpreting models in a model-agnostic manner best.

5.5 Model Interpretation

In this section, the focus will be on understanding the interpretability of a model. From the literature survey done, it was noticed that many papers focus on data analysis and feature engineering or on optimizing the evaluation metrics. This approach works in theory but is a recommended solution when the dataset is static. Businesses in the real world focus on the forecast's accuracy and the underlying mechanics of the predictions made. If the wrong reasons are perceived for churn, the underlying business strategies are based on false assumptions. Methods of making models interpretable have been included in this study to assist the research in having a more significant impact on real-world impact. As correlation does not equal causality, a solid model understanding is needed to make decisions and explain them.

5.5.1 Model Interpretation using LIME

Despite widespread adoption of machine learning models, they have remained mainly as black boxes. If the business has to make decisions based on predictions made by machine learning models, it is essential to understand the underlying reasoning behind the forecasts made. LIME is a novel explanation technique that explains the predictions of any classifier by learning about the model's behaviour locally around the prediction. One of the main benefits of LIME is that it can be applied to any machine learning model as it is model-agnostic, where prediction of the data is made locally. When LIME is run, the output is a set of explanations that explain each feature's contribution to predicting a data sample. This approach provides local interpretability, and it also showcases the attributes that will have maximum impact on the decision-making

LIME stands for local interpretable model agnostic explanations and can help interpret the features that contribute to the decision-making model. The interpretation is model-agnostic, enabling the user to analyze the reason why a point may be classified as a customer at the risk of churn or not. Given that a single data point and the machine learning algorithm is used, LIME will build an understandable explanation for the specific data point. The output of LIME is a list of reasonings that explain how much each attribute contributes to the prediction of a chosen data point. This approach helps with local interpretability and allows the user to gain a data-backed intuition about which feature modification will have the maximum impact on the prediction. The interpretable models are trained on chunks of perturbations of the original instance and provide only an excellent local interpretation.

In Figure 5.5, the interpretation of the machine learning model applied has been interpreted using LIME. In the Feature selection and bivariate analysis, the information on the crucial features can be reconfirmed through these plots. A month-to-month contract is considered to contribute approximately 19% to the chance that a customer is not likely to churn, whereas having total charges greater than 3782.40 is considered a 13% chance that the customer might churn. The interpretation for any model can be made using LIME as it is model agnostic. The output of LIME is the explanation of the model, and this is done in a model agnostic manner. In Figure 5.5, the forecast of the customer and the interpretation has been made.

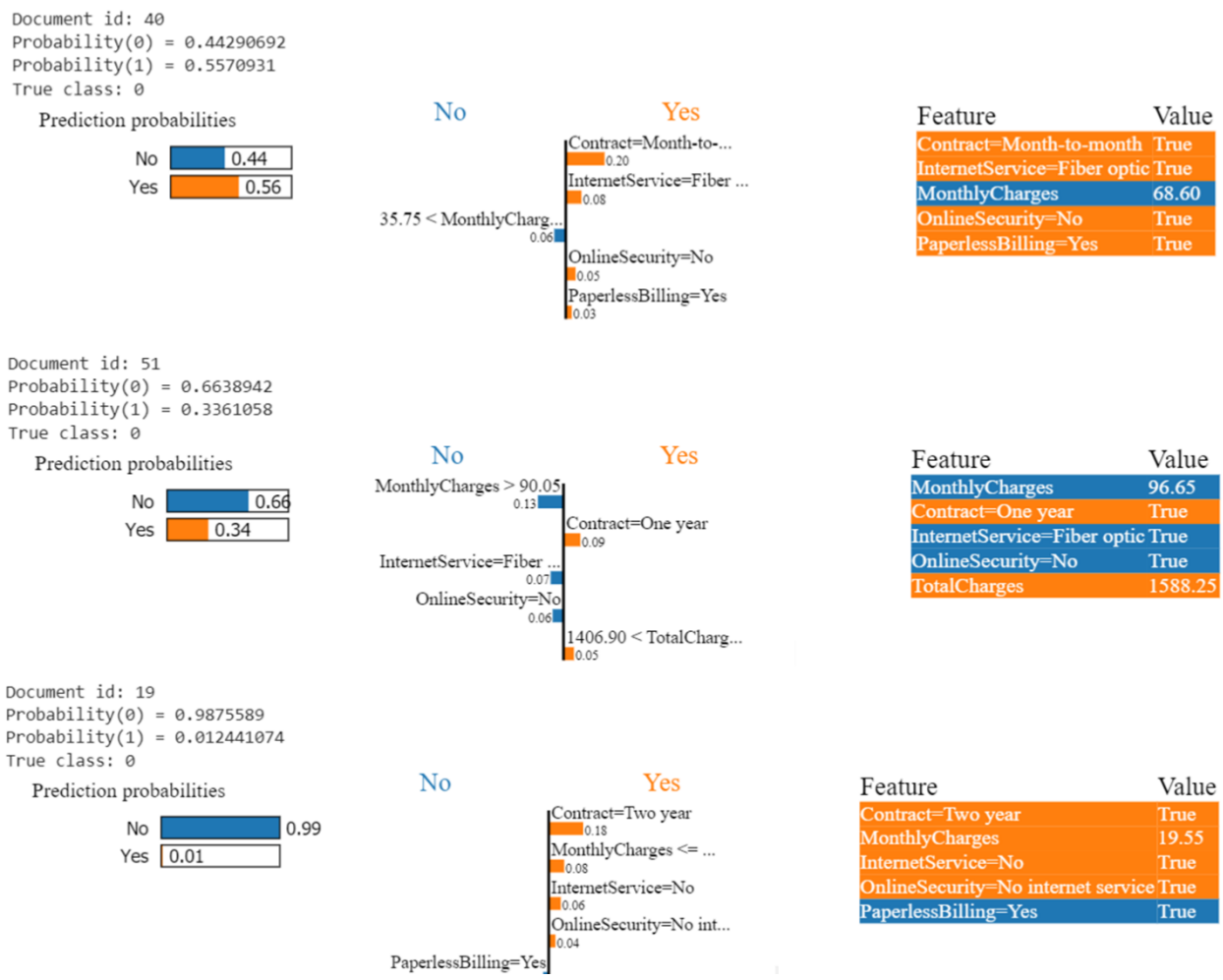


Figure 5.5: Model Interpretability with LIME

A data point is chosen, which has been represented in the form of a document id. In Figure 5.5, three documents have been chosen and for which the probabilities of both the classes have been defined. The class defined as zero indicates that the customer has not churned, and the class that indicates the probability of one indicates that the customer has churned. In the visual representation, if the probability of a class is more significant than 0.5, it is considered the dominant class. The visual representation indicates the top five features that contribute to the prediction. For instance, the document id of 40 indicates that having a monthly contract, a fiber option internet service, not having online security and having paperless billing are the top local indicators that the customer will churn with a probability of 0.55. Each data point is interpreted with a model-agnostic approach to enable the business to understand the underlying mechanism behind each forecast.

Ensemble models give good results when presented to a business stakeholder, understanding why the model has made specific predictions is more important than prediction. This approach can help the business make better-informed decisions. LIME ensures that the model behaves as expected while replicating human logic - it provides a representative set that provides an intuitive global understanding of the model. Using LIME would help non-experts compare and improve on an untrustworthy model via feature engineering. For effective business action based on machine learning, trust in the model is crucial and explaining individual predictions is an effective method.

5.5.2 Model Interpretation using SHAP

SHAP explains the predictions associated with a data point by computing the contribution of each feature to the prediction. SHAP is based on coalitional game theory, where the features of a data instance act like players in a game of coalition. SHAP stands for Shapely Additive Explanations – it is represented as an additive feature to the attribution methods. This view connects SHAP and LIME on model interpretability.

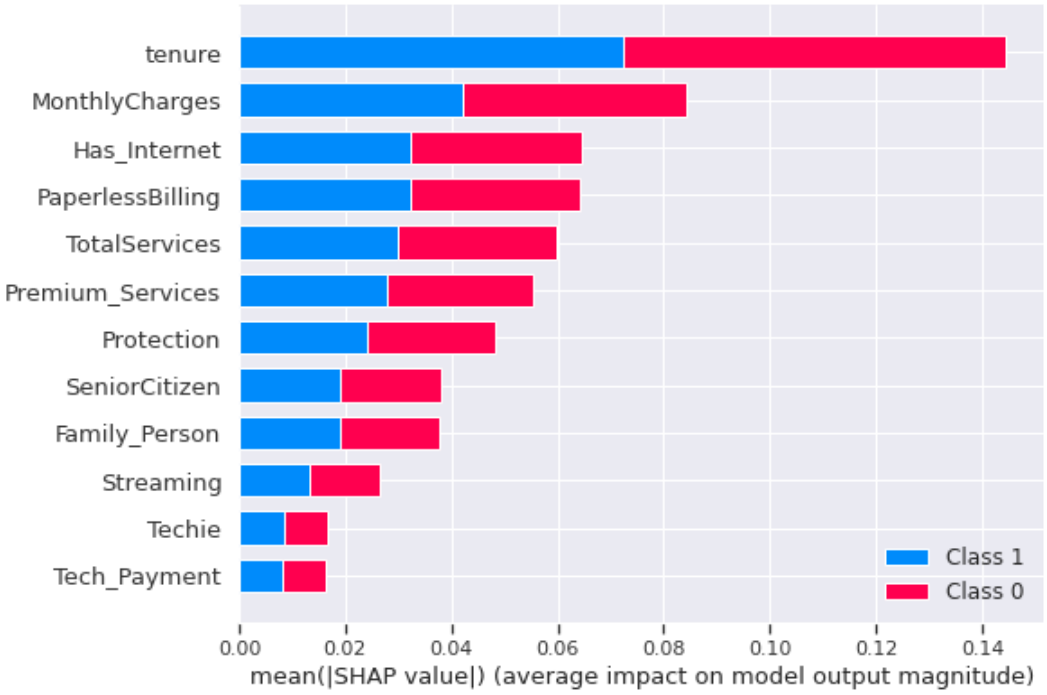


Figure 5.6 SHAP Feature Importance

In Figure 5.5, the SHAP Feature importance plot can be seen that is based on shapely. The chart signifies a global interpretation of importance per feature represented with Shapely values. It is measured as mean absolute Shapely values and is defined in terms of percentage points. The feature importance plot from SHAP showcases that the most critical features are tenure and monthly charges, which is in line with our previous observations. Data can be clustered using Shapely values, where the goal is to find similar groups for the various instances. The clustered SHAP charts are showcased in Figure 5.7, where tenure and monthly charges are highlighted.

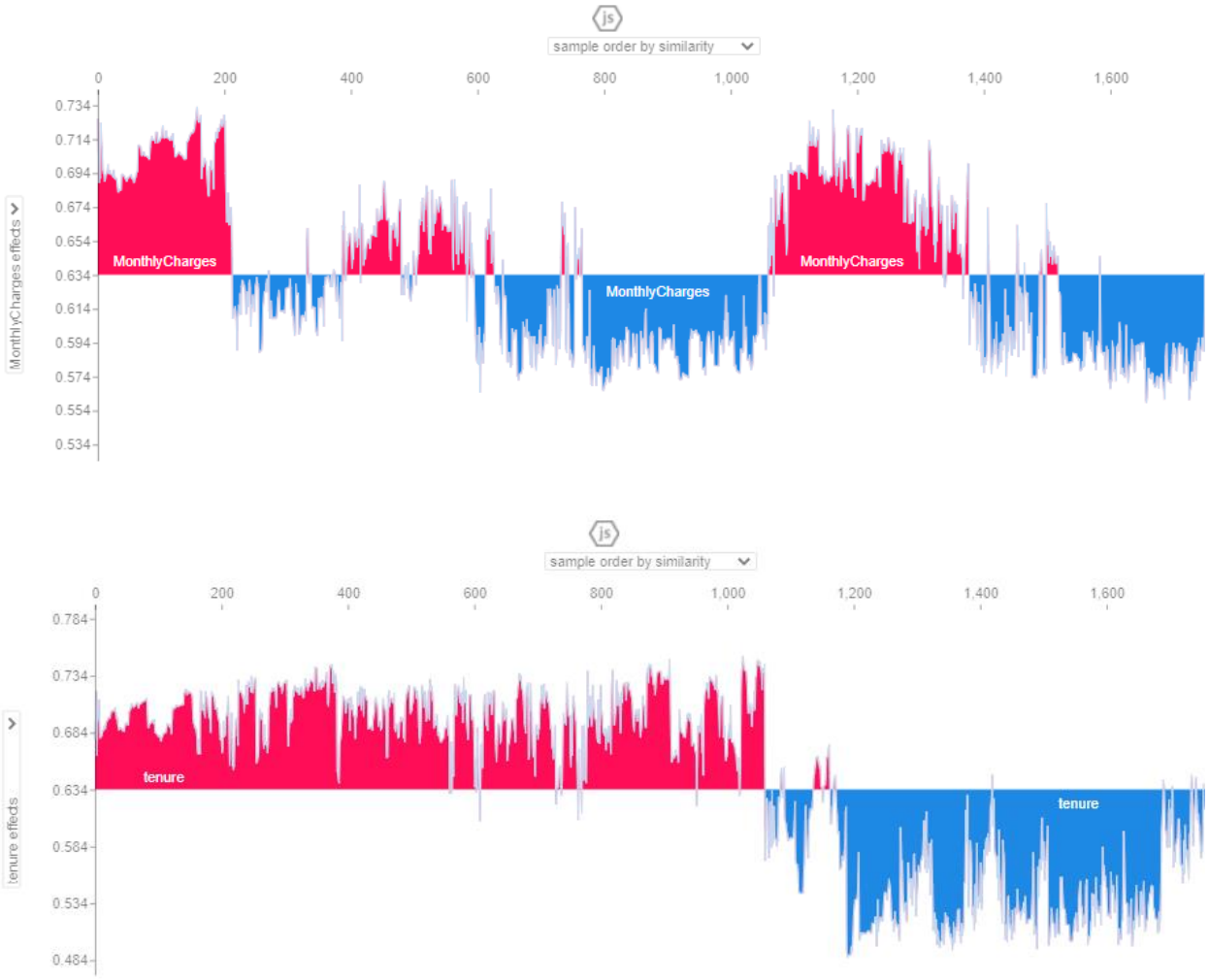


Figure 5.7: Clustering SHAP Values by explanation similarity

The unit of all SHAP values is the same across various features; it is the unit of the prediction space. Hierarchical agglomerative clustering is used to order the instances. The plot consists of multiple force plots, where the prediction of the instance is explained.

The red SHAP values increase the prediction, whereas the blue values decrease it. A cluster in this setting will stand out. The second graph in Figure 5.5 of tenure is a group with a prediction of the high likelihood of churn. Since SHAP computes Shapely values, the prediction is distributed fairly among the feature set. One of the benefits of SHAP is that it connects LIME and Shapely values; this helps unify interpretable machine learning. The global interpretation methods include feature importance, feature dependence, interactions, clustering and summary plots. With SHAP, global interpretations are consistent with local explanations, as the Shapely values are standardized throughout the interpretations.

5.6 Summary

In the sections above, a complete discussion was done on the results obtained through this study. In Section 5.2, the baseline results of the logistic regression and decision tree classifier was highlighted. The interpretations of the visualizations from the exploratory data analysis carried out in Chapter 4. In Section 5.4, analysis and interpretation of model results were made. First, individual model results and ensemble model results were analyzed. Next, the mean of the cross-validation scores was analyzed, and finally, after class balancing was performed using oversampling by SMOTE-NC, the model results were compared. Finally, in Section 5.5, model interpretability using LIME was discussed. Hence, with all of the above inputs, Chapter 5 can be summarized. In Chapter 6, the conclusions of the paper, along with the recommendations, will be discussed.

CHAPTER 6:
CONCLUSIONS AND RECOMMENDATIONS

This chapter will discuss the conclusion based on Chapter 4 and the discussion of results in Chapter 5. In this study, the classification of telecom customers that will churn has been done with the help of machine learning models. Multiple papers related to customer churn in the telecom industry was analyzed to perform a preliminary analysis to ensure that best practices were implemented. Some studies focused on data processing, and there were research papers that focused on finding the best model that would give us the best results. While there were papers that focused on bringing about the best results, the focus of this study has been to bring about the best possible results along with the focus on model interpretability.

Surveys were carried out to understand the reason that few models go into production. This research has been carried out on the IBM Watson Telecom dataset, and new practices were bought about in the feature engineering and dataset preparation, such as the inclusion of a validation dataset. 3This was done to ensure that the hold out dataset does not leak into the trained model to ensure real-world applicability of the chosen model. Model interpretability has been the focus of this paper. Advanced model interpretation techniques such as SHAP and LIME have been implemented to ensure that when a new data point is bought into the dataset, the business can interpret if a customer is likely to churn or not. Along with the churn classification, the proposed machine learning pipeline will also outline the features that are likely the cause of churn.

6.1 Introduction

In this section, the conclusions and recommendations based on the study will be explained. The discussion and conclusion of the study will be explained in Section 6.2. Understanding if the study's objectives were met, the discussions to help support the arguments, and the business's conclusions will be explained. In Section 6.3, the impact of the study and the contribution to the overall community will be analyzed. This section will help highlight the study's novelty, along with the improvements made compared to the work done previously.

Based on the literature surveyed, some models had satisfactory results regarding the accuracy of the test data and the AUC scores. In Section 6.4, future work recommendations will be made. This section will help researchers understand the possibilities to continue the research.

6.2 Discussion and Conclusion

In this study, the performance of individual and ensemble models was carried out to classify churned customers. A baseline was set using the logistic regression and decision tree classifier, where the test accuracy was noted. The preliminary analysis of the data was done by looking at the fundamental statistics of the data. Then, the distribution of the variables was analyzed, followed by missing value analysis and outlier analysis. Univariate and bivariate analysis with respect to the target variable churn was done. The distribution of the variables with respect to churn was analyzed to deep dive into the latent relationships within the dataset. The was followed by analysing the Pearson's correlation coefficient by plotting heatmaps for categorical variables and the numerical attributes.

In any analysis, before more advanced techniques are implemented on the data, it is essential to understand the data in depth. The probability distribution of the numerical variables is analyzed using a non-parametric kernel density estimation. For variables that are skewed, the box-cox transformation was applied to normalize the distribution. Before more feature selection are applied to the dataset, a statistical analysis among the categorical variables is done using Chi-square analysis. An ANOVA test was performed among the numerical and categorical variables, and the statistically significant variables were chosen. Following this, the details of the data split were showcased, followed by the baselines set up using the logistic regression and decision tree classifier. The individual models and ensemble models were discussed briefly in Chapter 4, followed by cross-validation to improve the bias-variance trade-off. The various feature selections methods that can be employed were also discussed to understand the crucial features that will help understand if a customer is likely to churn or not. This approach will help in model results is as optimized as possible and interpretation of real-world applications. This study aimed to build a model that can be deployed in real-world scenarios.

In Chapter 5, the results of the models were analyzed, and further details were discussed. The baseline results were discussed in detail, followed by interpreting the visualizations that were charted out in the previous analysis. The model results of various models were discussed in detail. Individual models such as Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree and K Nearest neighbour were trained on the data. Ensemble models such as Gradient Boosting and Light GBM were also used. The data was split into train, validation and test datasets to ensure no data leak to ensure real-world applicability. It was noticed that some of the tree models were overfitting the data.

To further improve, cross-validation was carried out on the models, and the mean of the cross-validation scores was analyzed. Hyperparameter tuning using grid search and randomized grid search was used to optimize the models over multiple iterations. In the literature survey, it was observed that the models that had oversampling tend to perform better. Class balancing was done using SMOTE-NC, and the models were now trained on the oversampled data. It was observed that ensemble models tend to have better performance as compared to individual models. Decision Tree with AdaBoost, Decision Tree with Bagging, CatBoost, Linear Support Vector Classification, Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbour, Naïve Bayes, Decision Tree and SVM with radial basis function kernel were implemented. The highest AUC scores were by the decision tree with AdaBoost and decision tree with bagging with scores of 0.84 for both.

As previously stated, the main aim of this research was to aid actual work implementation of customer churn models. A gap observed in the overall methodology was that more focus was given to mode results than how the business interprets results. The study relied on model interpretability techniques such as SHAP and LIME to tackle real-world application issues. These model interpretability techniques shift focus from just whether a customer churns or not to what factors contribute while classifying a customer. A detailed analysis of local model interpretability and the global implications of why the classification occurs helps us gain deeper insights into the forecasts made by the models. The more the complexity of the model, the less interpretable it becomes. The study has aimed to bridge the gap of model understanding and has done so by focusing on state-of-the-art results and interpretation.

6.3 Contribution to Knowledge

The work done in the field of classification of customer churn is extensive. However, a gap in a central paper brings together all of the knowledge to an approved pipeline. This study highlights state-of-the-art data analysis, model building and tuning, and model interpretability. This paper brings a new perspective to the data science community regarding how incoming data points can be classified and model interpretability using SHAP and LIME. If a modelling technique is trusted in its predictions, it is crucial to understand the underlying mechanics and any underlying pitfalls. Interpretability techniques reinforce confidence that with a good understanding of the method, the likelihood to base assumptions on false understanding decreases. These model interpretability techniques can be further extrapolated to gain deeper insights from SHAP and LIME data points.

6.4 Future Recommendations

There are various areas of research that one can take going ahead. The model has now been performed on a static dataset. Implementing a similar pipeline at an enterprise level at a fixed cadence can help track customer behaviour and reinforce the model. Natural Language Processing can be leveraged on feedback gathered from focus groups as to why customers are churning from ratings, reviews, social media and calls from customer service agents. Some patterns can be analyzed, such as geography, demographic information, and other factors analyzed further. This model can be improved to calculate the percentage of revenue saved by the company based on the evaluation metrics. Based on this information, the lift in sales can be analyzed.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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*.
5. 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].
6. Andrews, R., (2019) Churn Prediction in Telecom Sector Using Machine Learning. *International Journal of Information Systems and Computer Sciences*, 82, pp.132–134.
7. 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].
8. 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].
9. 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].
10. 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.

11. 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].
12. Fonseca Coelho, A., (n.d.) *Churn Prediction in Telecom Sector: A completed data engineering Framework*.
13. Hadden, J., Tiwari, A., Roy, R. and Ruta, D., (2006) Churn Prediction: Does Technology Matter. *International Journal of Intelligent Technology*, 1, pp.104–110.
14. 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*.
15. 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.
16. Havrylovych, M. and Nataliia Kuznietsova, ©, (2019) *Survival analysis methods for churn prevention in telecommunications industry*.
17. 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].
18. 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].
19. 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].
20. Kaggle, (2018) *Telco Customer Churn*. *Kaggle.com*. Available at:

https://www.kaggle.com/blastchar/telco-customer-churn [Accessed 9 Jan. 2021].

21. 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].
22. 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].
23. 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].
24. 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.
25. Labhsetwar, S.R., (n.d.) Predictive Analysis Of Customer Churn in Telecom Industry using Supervised Learning.
26. 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].
27. Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., (2021) Customer churn prediction system: a machine learning approach. *Computing*.
28. 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.
29. Momin, S., Bohra, T. and Raut, P., (2020) *Prediction of Customer Churn Using Machine Learning*. *EAI/Springer Innovations in Communication and Computing*.

30. 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].
31. 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.
32. 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.
33. 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.
34. 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.
35. 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].
36. 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].
37. 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].

38. Tamuka, N. and Sibanda, K., (2021) Real Time Customer Churn Scoring Model for the Telecommunications Industry. *IEEE*, pp.1–9.
39. Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A., (2020) Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, pp.429–441.
40. Thontirawong, P. and Chinchanachokchai, S., (2021) TEACHING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN MARKETING. *Marketing Education Review*.
41. 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].
42. 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].
43. 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].
44. 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.

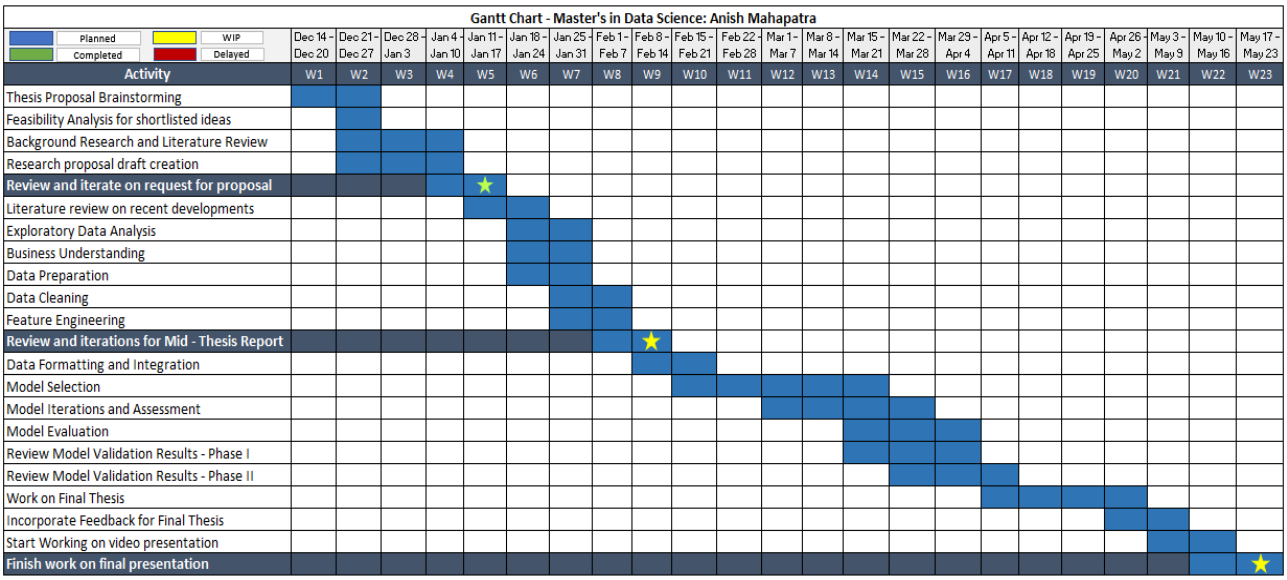


Figure 7.1: Research Plan and Timelines

APPENDIX B: RESEARCH PROPOSAL