MTR v12.docx

by Anish Mahapatra

Submission date: 28-Mar-2021 09:26PM (UTC+0100)
Submission ID: 148231908
File name: MTR_v12.docx (859.71K)
Word count: 13819
Character count: 76072

Prediction of Customer Attrition in the Telecom Industry using Machine Learning

Anish Mahapatra
Student ID 944563

Under the supervision of
Karthick Kaliannan Neelamohan

Mid-Thesis Report
Master of Science in Data Science
Liverpool John Moores University

MARCH 2021

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.

ABSTRACT

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

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

This research intends to build a predictive framework that can predict churn accurately and identify behaviour patterns that indicate customer churn. We 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. We shall perform extensive feature selection, processing and work with multiple hybrid models to

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

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EDA	Exploratory Data Analysis
SVM	Support Vector Machine
KNN	K Nearest Neighbour
AUC	Area under ROC Curve
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Oversampling Technique
XGBoost	Extreme Gradient Boosting
GSA	Gravitational Search Algorithm
PPforest	Projection Pursuit Random Forest
LDA	Linear Discriminant Analysis
CRM	Customer Relationship Management
AdaBoost	Adaptive Boosting

CHAPTER 1: INTRODUCTION

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

1.1 Background of the Study

With the increase in the number of options consumers have in the Digital Age, for a company to be successful, it is vital to keep costs low and profits high. One of the most effective ways to do this is to retain the existing customer base and focus the remaining budget on acquiring new customers. The retention of the existing customer base in a focused and systemic manner is to be done, or its bottom line can be affected. A targeted way to approach the end goal of customer retention is to flag customers that have a high probability of churn. Based on customer behaviour and attributes, if we can flag the customers that are likely to churn, we can run targeted campaigns 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 understand customer behaviour and mentality virtually. 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 we were to increase customer retention rates by just 5%, profits could see an increase from 25% to even 95% (Hadden et al., 2006). For a company to be profitable, it is thus essential to take pre-emptive action to retain customers that may churn. Churn is defined as customers who stop using their specific services and plans for long periods.

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

1.1.2 Flagging customers and retention policies

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

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

1.2 Struggles of the Telecom Industry

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

Fig 1.2.1: 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 we should 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, we can monitor the model's evaluation metrics, 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. We will be able to predict customers that will churn accurately. 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 we suggest balancing techniques for increasing the accuracy of the model?
- Can we trust the results obtained from interpretable models?
- 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
- We will focus on interpretable models. If time permits, we will attempt to use other models to perform customer attrition analysis

1.7 Significance of the Study

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

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

1.8 Structure of Study

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

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

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

CHAPTER 2: LITERATURE REVIEW

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

With the advent of massive investments from telecom operators in this internet age, both old and new conglomerates globally, the market is the most competitive it has been in decades. The literature review will focus on reducing customer churn; we will also focus on 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 survey, we would like to identify the different types of churners that exist. The visualisation below showcases a tree-based visualisation to showcase the same. In this literature review, we shall focus on the set of churners that churn voluntarily; it is difficult to flag whether the churn was incidental or deliberate every time.

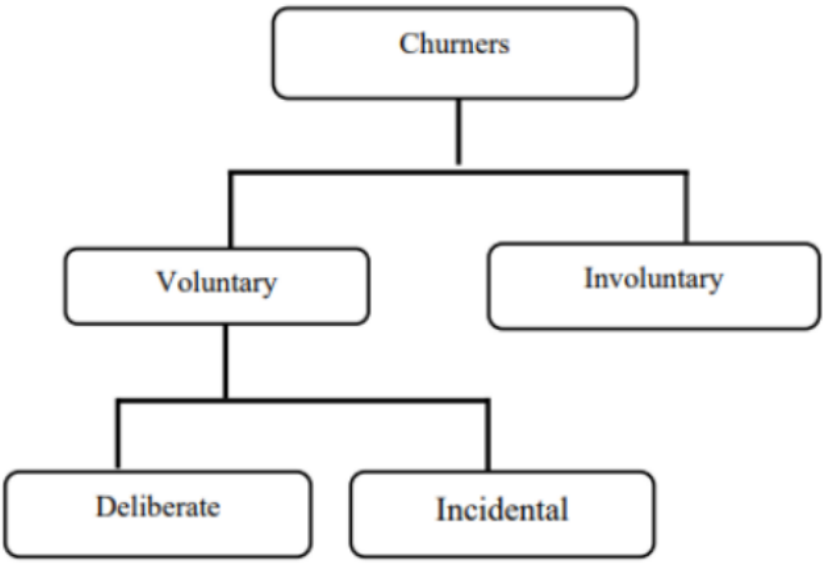


Figure 2.1: Types of Churners (Saraswat, S. & Tiwari, 2018)

With the above visualisation from the authors, we can identify that the set of customers who undergo voluntary attrition is the ones we would like to focus on in our literature survey. This understanding will help understand the customers' behavioural patterns that churn voluntarily are so that we can selectively profile and target the customers to get higher accuracy.

2.1 Introduction

As we dive into the literature review, having a proper structure for our analysis is critical when dealing with the telecom industry's churn. In section 2.2, we will focus on the telecom industry and the data-driven analytics driving the industry. This will give us an idea of 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 will deep-dive into customer attrition in the telecom industry and how this is a significant driver for the drain in finances and telecom operators' stability globally. In the next section, we will understand the way companies are leveraging predictive modelling in customer churn attrition and the models and methodologies used to keep profitability up.

Here, we will analyse in-depth the models and the methodology and ideas behind the working of predictive frameworks. We can leverage our learnings from the above sections in how visual analytics is also being used to visualise large sets of data. Before we proceed to the related research publications, we will understand more about the metrics and formulations that authors have used in the literature survey; this will help us leverage the summary table along with the steps of data preprocessing, feature engineering, models applied and results of the modelling efforts as displayed. This table will summarise all of our learnings in a quick referential format for future publications to leverage the latest in the field. In Section 2.7, we will discuss our learnings from the related work and the previous sections and 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, we will summarise all of the analysis we have done 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.

Following our learning in the sections below, we shall analyse the literature survey gaps in the authors' recent work. 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. We have learned that 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 we cannot directly just call, message or communicate with someone at a fraction of the cost paid for the same service about just a decade ago. The rate of mobile and internet penetration in third-world countries is increasing exponentially every day; this leads to a whole host of some of the largest companies in the world backing up telecom operators to be able to acquire a customer base as loyal and dedicated as possible so that this cash-burn can be leveraged to profit in the future.

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

Several low-code or no-code tools are being used to start build proof of concept projects; the reality

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

2.3 Customer Attrition in the Telecom Industry

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

There is one common element in the literature reviewed; there are always certain behavioural traits of a customer that can be identified as a customer trend that is to churn. Customers tend to move across telecom operators for several reasons, with countries enabling inter-operator portability globally. It is easier for a customer to move if they are dissatisfied with the services of a company. There are a few factors with the digital age to determine how likely a customer is to churn. If a customer has enabled auto-pay for their bills, if a customer has been associated for many years, if a customer has internet services and has opted in for a host of other services that their everyday life or family's life is dependent on, the customer is less likely to churn.

The literature review observation is that a customer should have the least amount of friction while getting into the services offered. This ease of movement and a tie-in to other services offered at multiple fronts will increase customer loyalty. When a customer is likely to move across, the company should have an open communication line with effective teams on multiple touchpoints. A surprising find is that the main reason a customer moves across telecom operators is not due to a new promotion/offer. Instead, the primary reason a customer moves across operators is due to dissatisfaction with current services. If we can identify 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.

We understand that 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 that we get from the dataset is a deciding factor in retaining customers. It is not merely identifying the set of customers that are at a high risk of churn; if timed right with the right kind of targeted campaign, there is a high chance that even if the telecom operator was to take a slight loss in the form of additional discounts offered to the high-risk customer in the short term, the cost could be recovered and a profit can be made in the long-term. Various strategies can be employed based on our learnings from the model. However, the suggestions of the personnel involved directly with the customer and customer database must be taken into account as they have more real-world context when it comes to customer behaviour and sentiment.

2.4 Predictive Modelling in Customer Churn Analysis

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

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

2.5 Visual Analytics in Telecom

For data of any form to be leveraged, we need to understand the dataset. One of the fastest ways to perform exploratory data analysis is to visualise the data. The below diagram illustrates the relationship between data, visualisation and models with the intermediary knowledge gained from visual analytics (Yuan et al., 2021).

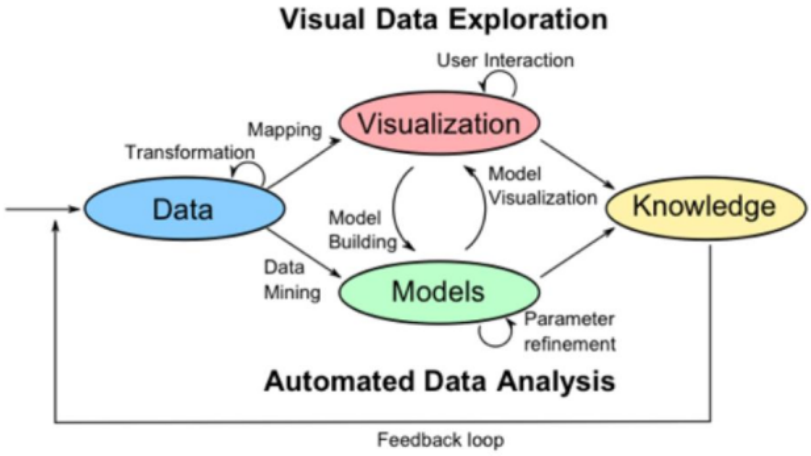


Figure 2.5.1: Visual Data Exploration

Being able to perform automated data analysis involves using visual cues is the essence of visual data exploration. Based on the visualisations formed, we will understand more about row-level data. When data transformation is performed, we can visualise the data post-processing to 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. If we were to visualise and sum up all of the scores we got for features visually, we should be able to identify the top feature using a bar chart with an indication of the top features to choose from for the next steps. Using multiple methods of visualising the features' distribution, the variance of the data points and the other analysis helps us make decisions for the next steps.

2.6 Related Research Publications

This section will provide a review of how data analytics is used in the telecom industry to identify customers at a high risk of attrition and the data-driven processes followed to set the baseline of the techniques carried out in the industry far. Section 2.6.1 and Section 2.6.2 will focus on feature engineering for the data and handle class imbalance. Efficiently carrying out data preprocessing will help us obtain better results in the following stages of implementing machine learning and validation via k-fold cross-validation. We will also understand the evaluation methods used to assess the models' performance through the literature review. 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. Here, we analyse the existing techniques implemented by authors to either pick the significant features from the dataset that can affect churn or generate new features from the existing set of attributes that can help us predict churn better. When the authors have set out to perform feature engineering, it is only done keeping the dataset and the predicted model's accuracy in mind. When we perform feature engineering on a dataset, another critical task is to identify the attributes that have the highest impact on the target variable. This can be done by leveraging rigorous algorithms or even RapidMiner and Azure ML Studio (Thontirawong and Chinchanachokchai, 2021).

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

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

2.6.2 Handling Class Imbalance in Machine Learning

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

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

2.6.3 Implementation of a predictive framework

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

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

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

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

2.6.4 Reviews of Evaluation Metrics for Classification

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

There are different ways of evaluating the performance of a classifier. The methods used are the ROC curve or derivatives of the confusion matrix, such as F-Score. Understanding the confusion matrix's derivation is vital to decipher many of the results when machine learning models are involved. We shall go over a few of the standard metrics in the below sections to understand the metrics used for evaluation. We will evaluate 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.

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

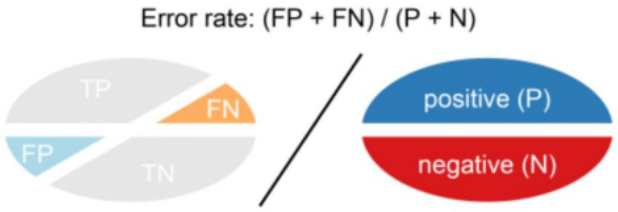
$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

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.



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

2.6.5 Summary of Literature Review

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

2.7 Discussion

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

The other approach followed is to rely on the machine learning model to flag the customers that are likely to churn effectively. The data size plays a considerable role; if the data's size is limited, focusing on the machine learning algorithm is more sensible, whereas a hybrid approach can be experimented with for larger datasets. The literature on deep learning suggests that even though a neural network approach works for some cases, the model's performance is not significantly better to opt-in for deep learning models exclusively. It is a common misconception that deep learning models perform better than machine learning models in all use-cases. From the literature review, we understand that for the 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. We can explore where we can try more feature engineering techniques and summarise our results in a manner where we consider the observed and latent relationships of the features with the target variables. 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 we will take inspiration for, along with more advanced feature selection methods

Table 2.7.1: Literature Review for IBM Watson Telecom Dataset

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

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

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

We have understood from the above papers 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. We shall try novel methods of multiple feature selection on the telecom data couple with a robust predictive framework. We have observed with a few scenarios that when the data is over-engineered or refined beyond a point, overfitting of the data occurs on the training set, and the performance on the hold-out or test dataset is not as expected.

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

2.8 Summary

A whole host of machine learning models can be used for the use case of solving for the classification of high-risk customers. An excellent approach to try would be to focus on the machine learning approach and the data preprocessing. A few authors implemented class balancing techniques, and better accuracy was observed. For our approach, we will work 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 we can find the best models for our use-case, we will perform k-fold cross-validation to get the best generalised and robust model. This thorough literature review of the best the academic community offers has provided us with the baseline understanding we were looking for before deciding the appropriate research methodology for our use-case.

CHAPTER 3: RESEARCH METHODOLOGY

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

3.1 Introduction

We understood the set-level on how to tackle a customer churn problem in the telecom industry from the literature review. 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. We follow this up by the research methodology in section 3.2 that consists of data selection, data preprocessing, data transformation, data visualisation, class balancing, model building, model evaluation and model monitoring. We will then go through the proposed model in Section 3.3, ultimately followed by the summary in Section 3.4.

3.1.1 Business Understanding

The telecom industry is a highly competitive industry where customers can choose to move across operators if they believe they are getting more value with another service provider. Based on the customer's behaviour patterns, we have 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. We also note that 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, we will analyse and understand the dataset better by deep driving into the statistics of each column:

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

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

From the above description, we have now understood the descriptive statistics of the IBM Telecom Churn dataset that is going to be used in this study. We have 18 features that are categorical, two integer features and one feature of type float. The dataset has 7043 rows and 21 columns that describe customer behaviour. The dataset is taken over one month and will be used for analysis and predictive modelling in this study. The dataset range is also essential, including the summary statistics, to get a brief of the dataset that we are out to analyse.

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 we can choose from when it comes to telecom data. The data we have 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 we have selected the dataset, we would like to proceed within this domain. We shall now discuss the Data Pre-processing steps we will be implementing to ensure that the data is standardised as we use it in our next steps. We will perform a sense check of the telecom churn dataset to understand if the import of the data and the dataset's encoding are per expectations. Once we view the data types of the features, we will check on the shape of the data to ensure the number of rows and columns is consistent per expectations. We will then focus on the columns that have at least one missing value. Once we understand the attributes to consider, we will understand the percentage of missing values column-wise. This will help us to decide the strategies to take for the next steps. Post missing value analysis, we will determine if we can proceed with all the columns to the next step if we must drop columns based on absent value percentage or employ methods such as mean imputation, mode imputation, deletion of rows and iterative imputation.

Looking at the percentage of missing values for each attribute after the missing-value analysis will help us understand the base dataset that we will be using when we go to the next step of feature engineering. We will also perform outlier analysis and understand the skewness of the data to understand the feature's impact on customer churn. After understanding each features' distribution, we will proceed to perform a univariate analysis. 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

Based on the cleaned dataset, we will now decide the following steps to extract the most value from the dataset. We can perform steps such as one-hot encoding on the categorical features. Besides this, we shall also derive features from the existing dataset and feature engineer newer attributes. Based on the understanding of telecom's business, we will also apply business rules that make sense 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. We will use advanced Exploratory Data Analysis packages such as pandas profiling, Sweetviz and data prep to perform visualisation of the data; this 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. We shall perform multicollinearity and variance inflation factor tests to understand the data's inherent properties to understand the significant features to select for modelling. We will also look at the correlation scores for the numerical variables to identify the features with a high positive or negative correlation with the target variable. We will also perform a categorical analysis of 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. We can use the packages to analyse and understand the data are pandas profiling, sweetviz and data prep. 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. We shall visualise a few of the features and the target variables to understand the distribution of the data points.

Figure 3.2.4.1: Distribution of Churn (Target variable)

We have 21 features and 7043 data points.

We shall now analyse the distribution of a few of the dependent variables.

Figure 3.2.4.2: Distribution of Contract

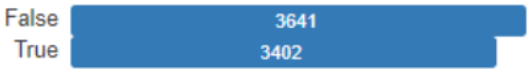


Figure 3.2.4.3: Distribution of Partner

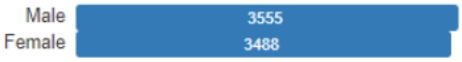


Figure 3.2.4.4: Distribution of Gender

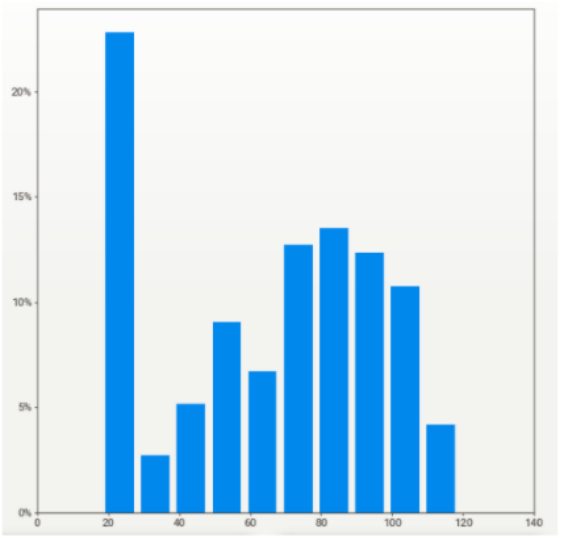


Figure 3.2.4.5: Distribution of Monthly Charges

3.2.5 Class Balancing

Oversampling and SMOTE are the techniques we will be leveraging to perform class balancing. We observed that the classification models had improved performance from the literature survey when class balancing was performed. We can perform class balancing in this section by using the recommended class balancing techniques of oversampling and SMOTE.

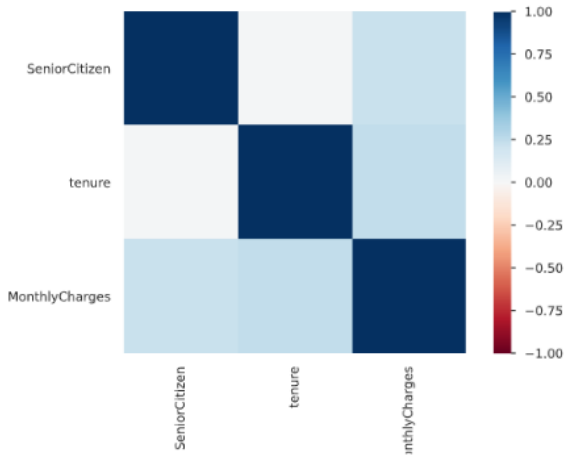


Figure 3.2.5.1: Correlation Matrix using Pearson's correlation coefficient (*r*)

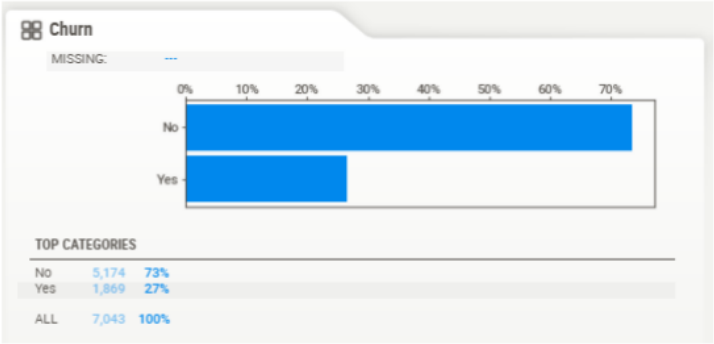


Figure 3.2.5.2: Distribution of Churn visualised using the package SweetViz

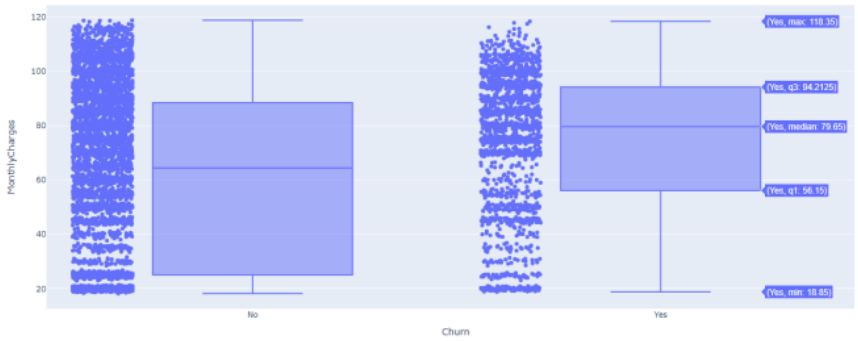


Figure 3.2.5.3: Distribution of Monthly Charges based on Churn

3.2.6 Model Building

Model Building is one of the more crucial components of this study. The following steps will help us identify the right set of models and appropriate techniques we can leverage to get optimal results. We shall now choose the models we would implement after the data cleaning, feature engineering, and data formatting steps.

3.2.6.1 Model Selection Techniques

We select the models we will be working with to predict customer churn efficiently and accurately for the model selection. From the literature review, it has been observed that the supervised classifier models have given us good results. We will implement single algorithm models to pick out the models that have the best performance. We shall use Logistic Regression, decision trees, Naïve Bayes, random forest, support vector machine and understand how the algorithms perform.

Based on the unique algorithms' analysis, we shall also attempt bagging and boosting techniques 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 happening in the best way possible, we shall also train the model 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, we could have opted to go for a validation dataset as well. We will go for an 80-20 train-test split for the models. For the top-performing models with this design, we shall also attempt a 90-10 split, as this was recommended in the literature review for a few research papers. 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, we shall perform more iterations, correspondingly assessing model performance with each iteration. This can include monitoring p-values, the number of features, model performance, variance inflation factor scores which would differ across models. The top selected models will now be the challenger models based on which the best model will be decided. We will perform hyperparameter tuning on the given models using previous learnings and methods such as Grid Search, Random Search, and Bayesian optimisation depending on the model considered.

3.2.6.4 Model Assessment

For any models to be used by the business, model assessment is a critical part of the process. As we develop models from a Data Scientist's eyes up until this point, we will need to take steps to 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. We have noticed that accuracy and AUC were used to assess models across the board from our literature review. We will also focus on model sensitivity and specificity curves to make a generalised model that can be leveraged.

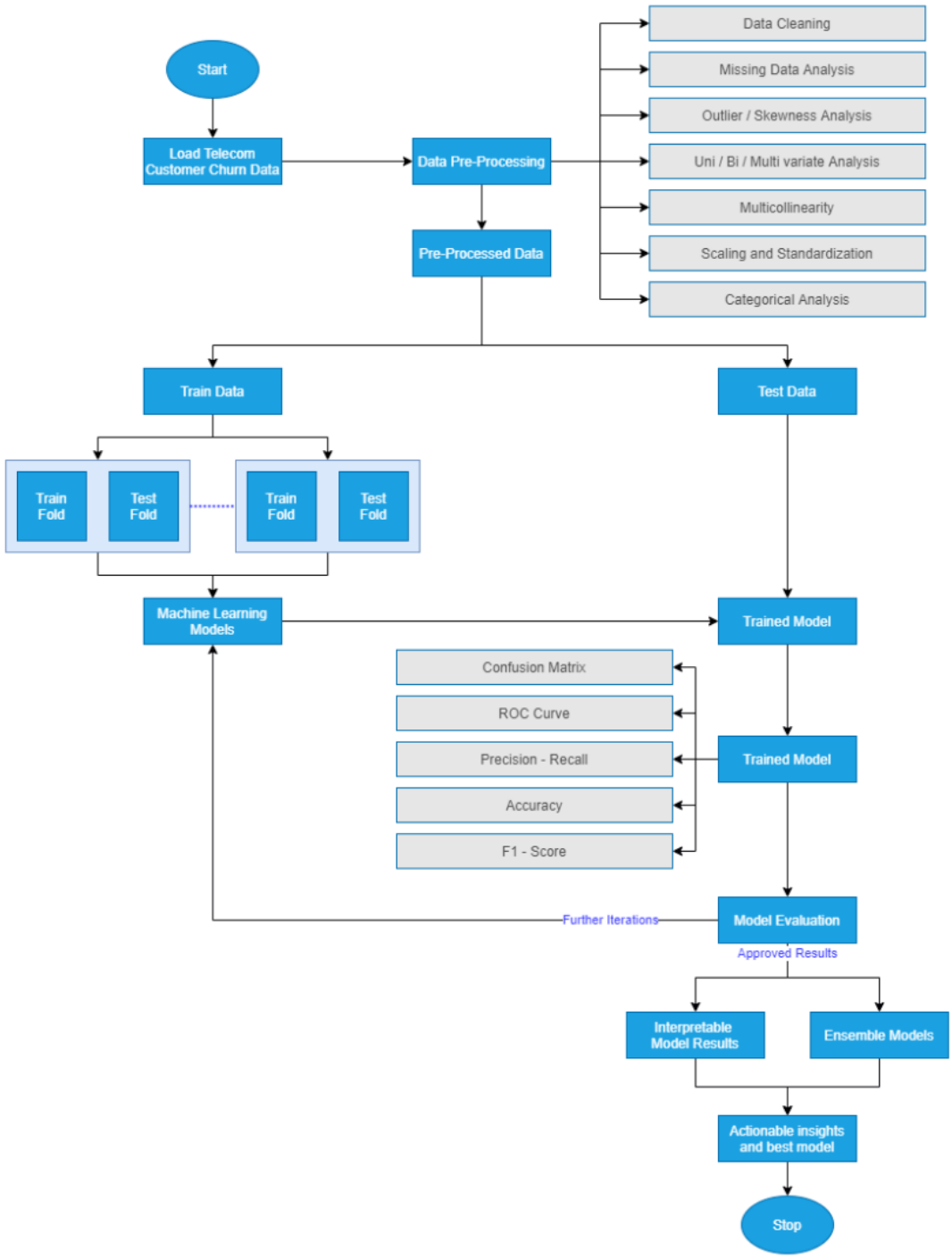


Figure 3.2.7.1: Model Building Process

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, we are in the model assessment stage; we will need to focus on actionable insights and provide the business with the customer behaviour patterns linked with the high likelihood of churn. The diagram above highlights the stages that we will be using for the model building process, from the data loading to the final model output. This step-by-step process has been drawn out in detail based on the extensive literature review carried out.

3.2.7 Model Evaluation

We have now settled on the best model that we would like to showcase. This is the model on which extensive feature engineering has been carried out, and from a wide range of models, we have chosen the best. We will follow the below-mentioned steps 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 as some of the metrics.

3.2.7.1 Metrics for Evaluation

We will now proceed to compare the model results obtained with the other literature we have previously surveyed. Using the same metrics of accuracy, F-Score, the area under the curve, we will compare the new ensemble or individual models' performance to the models' performance in the field's reviewed literature. Once we evaluate the results and see if they are satisfactory, we will proceed to the next steps. Else, if they are not adequate, we will move to re-evaluate our approach to improve iteratively. This process is iterative as the final model we would select should be as accurate as possible. Based on the literature review, we will also decrease the predictive framework's misclassification rate. There are standard metrics that w can use, and we can visually compare the metrics to select a model that can excel in all or most of the evaluation metrics chosen for classification.

3.2.7.2 Process Review

We will list the final process post the different iterations we have carried out and carefully review the process. As compared to the other research done in this field, we will analyse any potential misses, flaws in approaches and address them. Based on the process review carried out in the above step, we will decide if we would like to finish our research project and move on to the next steps. If not, we shall initiate further iterations and refine the model. This is an essential step and will be based on the comparative analysis we will perform to benchmark our model.

3.2.8 Model Review

We will now decide the next steps for the business use that our 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, we should filter out high-value customers 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 us monitor the results and understand how we can make the deployment more efficient.

For a machine learning model to improve with time, it is essential to create a feedback loop. Documentation of the research carried out, the results, and loopholes must be carefully documented to improve the model in the next iteration. If a similar accuracy can be obtained with lesser processing, this will also help the company save operationalisation expenditure costs. This is essential as we report the research results and provide 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. We will contemplate in the final review what are the things done right and what went wrong. There will be learnings from the entire process that we shall document and use in our next steps. We should also learn what was done well and what could have been avoided.

3.3 Proposed Model

Once all of the above steps have executed, we will have the proposed model 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 as we would like to reduce overhead expenses by targeting customers who are likely to churn. For operationalisation, it is advisable to opt-in for an accurate model and computationally sensible for this use-case. 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, and we have now chosen the most appropriate model for the chosen dataset. All steps have been carried out per industry best practices.

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

30. 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.
31. 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.
32. 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.
33. 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.
34. 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].
35. 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].
36. Saraswat, S. & Tiwari, A., (2018) A New Approach for Customer Churn Prediction in Telecom Industry. *International Journal of Computer Applications*, [online] Vol. 181(1, pp.40–46. Available at: https://scholar.google.com/scholar?as_ylo=2017&q=%22A+New+Approach+for+Customer+Churn+Prediction+in+Telecom+Industry%22&hl=en&as_sdt=0,5 [Accessed 23 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 Minimisation 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.

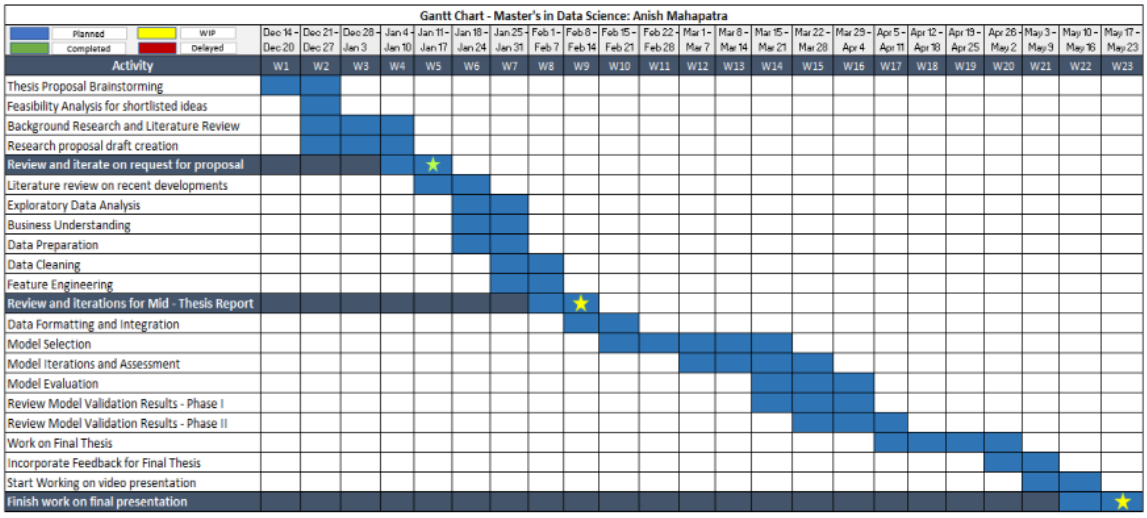


Figure 2: Research Plan and Timelines

APPENDIX B: RESEARCH PROPOSAL

MTR v12.docx

ORIGINALITY REPORT

3%	1%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	1%
2	Submitted to The College of Saint Rose Student Paper	1%
3	Alrence Santiago Halibas, Anju Cherian Matthew, Indu Govinda Pillai, Jay Harold Reazol et al. "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 Publication	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		