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 various market segments, companies must retain customers to maximise profits. Customer retention policies can affect the annual turnover drastically depending on the rate of churn. 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 processes and efficient 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.

Datasets for customer churn are large and is saved in large data warehouses where many features are present. Not all attributes are significant for churn predictions. Hence, feature engineering requires not only excessive computation but a substantial amount of time as well.

This research intends to find the model that can predict churn most accurately and the behaviour patterns that can indicate customer churn. The aim is to predict churn accurately and showcase the variation in performance of various algorithms. The dataset to be used for this research paper is the IBM Watson Dataset on customer churn in the Telecom industry.

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 The need for Customer Churn Analysis

The ability to retain customers showcases the company's ability to run the business. With the digital age, where everything is online, any business needs to virtually understand customer behaviour and mentality. The cost of customer churn in the Telecom Industry is approximately $10 billion annually (Castanedo et al., 2014). Customer acquisition costs are higher than customer retention by 700%; if 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.

1.2 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.

1.3 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

1.4 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?

1.5 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

1.6 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.

1.7 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 the identification of 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 pre-processing, 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, especially those involved in the telecom industry, will help us understand the advantages and disadvantages of many techniques. This literature review will set a 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 leveraged by the authors range from using algorithms such as ANN, SVM, using efficient pre-processing and feature engineering, and some even included social (Oskarsdottir et al., 2016). This survey will also understand the methods that other authors have used to solve traditional churn problems (Umayaparvathi and Iyakutti, 2016).

2.1 Introduction

This chapter 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 2.2 will review data-driven techniques, implement them and the effect they can have on the customer base. Through Section 2.3 and Section 2.4, we will focus on feature engineering for the data and how we can handle class imbalance. Efficiently carrying out data pre-processing will help us obtain better results in the following stages of implementing machine learning and validation via k-fold cross-validation in Section 2.5 and Section 2.6. We will also understand the evaluation methods used to assess the models' performance through the literature review. Section 2.7, we will review the evaluation metrics used for classification (Karimi et al., 2021). Finally, we will conclude the article by discussing our learning from the related research in Section 2.9 and summarise our learnings in Section 2.10. The algorithms and methods, as implemented by the researchers, has also been detailed out for reference.

2.2 Data Analytics in the Telecom Industry

The telecom industry is the backbone of a nation. To have a higher stake in the Industrial Revolution 4.0, telecom operators need to improve their CRM infrastructure to move from an internal support system to a strong value proposition for customers and other stakeholders (Li and Li, 2018). Attracting new customers may seem like a feasible way to grow market share, but the natural growth and value drivers lie within the system; gaining new customers is 5 to 10 times more expensive than focusing on keeping existing customers loyal (Wassouf et al., n.d.; Ebrah and Elnasir, 2019). Everyone is talking about data science, changing the way industries work. Implementation is vital here, with various options that focus on the explainability and usage of models rather than a black-box approach.

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 seems to be the long-term aim, with most companies requiring 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.

2.3 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 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).

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 also been used. GSA helps reduce the dimensionality of the data and, in turn, helps improve the accuracy of the data by optimising the search for significant features (Lalwani et al., 2021).

Methods for pre-processing 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).

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.

2.4 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). A modification of the conventional method, undersampling-boost, is also used to handle class imbalance (Saonard, 2020). Some of the other methods to deal with class imbalance include Adaptive Synthetic (ADASYN) and Borderline Smote (Induja and Eswaramurthy, 2015).

2.5 Machine Learning models

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).

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).

2.5.1 Review 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).

2.7 Related Research on Customer Churn Prediction

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.

Table 2.7.1: Literature Review

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 result: Contract month-to-month, tenure, Internet Service Fiber Optic	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	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 Total Charges is discarded RapidMiner is used to perform feature selection: Contract, Online Security, Tech Support, Tenure & Device Protection	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 explanations

(Hargreaves, 2019)	2019	Top 5 Significant features using Feature Selection: Fiber Optic, Month To Month Contract, DSL, One Year Contract, Streaming Movies	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: total charges, monthly contract and fiber optic Internet service, senior citizen	AUC: Random Forest *with RFE* - 0.96 ANN *with RFE* - 0.77
(Agrawal, 2018)	2018	One Hot Encoding	Accuracy: ANN - 80.03%

2.8 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, useful data pre-processing and efficient feature selection. This effort to obtain the right set of data to feed results in choosing a simpler model to perform classification; this results in saving computation time and keeping the overall computational requirements minimal, saving overhead costs for companies.

The other approach followed is to relay 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.

2.9 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 pre-processing. 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 pre-processing, 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. We will be using our learning from the literature review and our understanding of the telecom business to predict the customers' high risk of churn effectively.

3.1 Introduction

We understood the set-level on how to tackle a customer churn problem in the telecom industry from the literature review. With this understanding, we will set up the research methodology on how to most effectively tackle the use-case for our study. Section 3.2 focuses on the research methodology and includes the steps of data selection, data pre-processing, data transformation, data visualisation, class balancing, model building, model evaluation and model deployment. We then discuss 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. We also noted that based on the customer's behaviour patterns, we would have indicators to report if a customer might churn or not. Since the cost of retention is much higher than customer acquisition, it is vital to the company's survival 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.

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

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.

3.2 Research Methodology

The following section contains the steps to perform predictive modelling to predict the customers with a high risk of attrition. The steps followed are data selection, data pre-processing, 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.

3.2.2 Data Pre-Processing

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.

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.

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 shall now 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 shall use logistic regression, decision trees, Naïve Bayes, random forest, support vector machine and understand how the algorithms perform. Post analysis of the individual algorithms, we shall also attempt ensemble models with boosting, such as XGBoost and Light GBM.

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.

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.
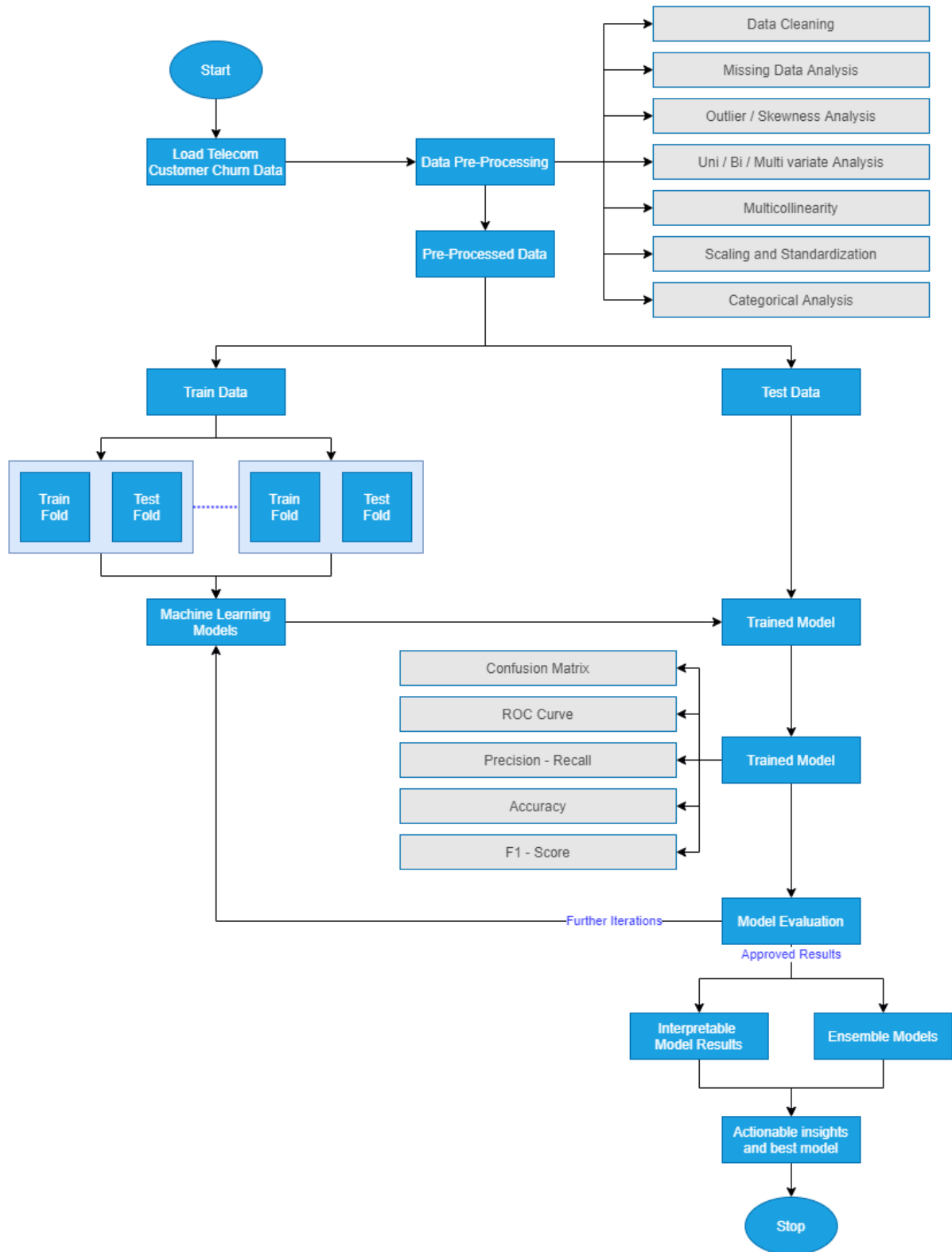
Figure 1: Model Building Process by Author via draw.io

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.

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.

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.

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 if there are any potential misses, flaws in approaches an address them.

3.2.7.3 Determine Next Steps

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 Deployment

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.

3.2.8.1 Plan for Deployment

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.

3.2.8.2 Monitoring and Maintenance

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.

3.2.8.3 Reporting Results

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.

3.2.8.4 Final Review

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.

3.4 Summary

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 pre-processing, data transformation, data visualisation, class balancing, model building, model evaluation and model deployment Post the literature review carried out in the previous sections, we have now chosen the most appropriate model for the chosen dataset. All steps have been carried out per industry best practices.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jahromi, A.T., Stakhovych, S. and Ewing, M., (2014) Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, [online] 437, pp.1258–1268. Available at: https://research.monash.edu/en/publications/managing-b2b-customer-churn-retention-and-profitability [Accessed 16 Jan. 2021].

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

2021].

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

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

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

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

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

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

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., (2021) Customer churn prediction system: a machine learning approach. *Computing*.

Li, Z. and Li, Z., (2018) Telecommunication 4.0: Opportunities and Challenges. In: *Telecommunication 4.0*. [online] Springer Singapore, pp.107–133. Available at: https://link.springer.com/chapter/10.1007/978-981-10-6301-5_6 [Accessed 20 Mar. 2021].

Mahdi, A., Alzubaidi, N. and Al-Shamery, E.S., (2020) Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry discriminant random forest Linear discriminant analysis oblique tree Project pursuit index Support vector

machines. *International Journal of Electrical and Computer Engineering (IJECE)*, 102, pp.1406–1421.

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

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

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

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

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

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

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

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

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

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

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

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

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

Wassouf, W.N., Alkhatib, R., Salloum, K. and Balloul, S., (n.d.) Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. [online] Available at: https://doi.org/10.1186/s40537-020-00290-0 [Accessed 21 Mar. 2021].

APPENDIX A: RESEARCH PLAN

APPENDIX B: RESEARCH PROPOSAL

APPENDIX C: ETHICS FORMS