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 4](#_Toc66554492)

[ACKNOWLEDGEMENTS 5](#_Toc66554493)

[ABSTRACT 6](#_Toc66554494)

[LIST OF TABLES 7](#_Toc66554495)

[LIST OF FIGURES 7](#_Toc66554496)

[LIST OF ABBREVIATIONS 7](#_Toc66554497)

[CHAPTER 1: INTRODUCTION 8](#_Toc66554498)

[1.1 Background of the Study 8](#_Toc66554499)

[1.1.1 The need for Customer Churn Analysis 8](#_Toc66554500)

[1.1.2 Flagging customers and retention policies 9](#_Toc66554501)

[1.2 Problem Statement 10](#_Toc66554502)

[1.3 Aim and Objectives 11](#_Toc66554503)

[1.4 Research Questions 12](#_Toc66554504)

[1.5 Scope of Study 13](#_Toc66554505)

[1.6 Significance of Study 13](#_Toc66554506)

[1.7 Structure of Study 15](#_Toc66554507)

[CHAPTER 2: LITERATURE REVIEW 16](#_Toc66554508)

[2.1 Introduction 16](#_Toc66554509)

[2.2 Data Analytics in the Telecom Industry 16](#_Toc66554510)

[2.3 Customer Attrition Analytics 16](#_Toc66554511)

[2.4 Predictive Modeling in Telecom Churn on different datasets 16](#_Toc66554512)

[2.5 Visual Analytics in the Telecom Industry 16](#_Toc66554513)

[2.6 Related Research Publications 16](#_Toc66554514)

[2.7 Discussion 16](#_Toc66554515)

[2.8 Summary 16](#_Toc66554516)

[CHAPTER 3: RESEARCH METHODOLOGY 17](#_Toc66554517)

[3.1 Introduction 17](#_Toc66554518)

[3.1.1 Business Understanding 17](#_Toc66554519)

[3.1.2 Data Understanding 18](#_Toc66554520)

[3.2 Research Methodology 19](#_Toc66554521)

[3.2.1 Data Selection 19](#_Toc66554522)

[3.2.2 Data Pre-Processing 19](#_Toc66554523)

[3.2.3 Data Transformation (Feature Engineering) 20](#_Toc66554524)

[3.2.4 Data Visualization 20](#_Toc66554525)

[3.2.5 Class Balancing 20](#_Toc66554526)

[3.2.6 Model Building 21](#_Toc66554527)

[3.2.7 Model Evaluation 24](#_Toc66554528)

[3.2.8 Model Deployment 25](#_Toc66554529)

[3.3 Proposed Model 26](#_Toc66554530)

[3.4 Summary 26](#_Toc66554531)

[REFERENCES 27](#_Toc66554532)

[APPENDIX A: RESEARCH PLAN 27](#_Toc66554533)

[APPENDIX B: RESEARCH PROPOSAL 27](#_Toc66554534)

[APPENDIX C: EHICS FORMS 27](#_Toc66554535)

# 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, Mr 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 pretty 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.

# 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 |
| XGBoost | Extreme Gradient Boosting |
| 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 to 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.

### 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 the customers who stop using their specific services and plans for long periods.

In this post-pandemic age, where virtual presence via calls and mobile data is the top priority, customers are trying to reduce their monthly expenditure. Competitors are employing strategies such as offering low prices or value-add services to get consumers to switch. After acquiring a significant customer base, the companies monetise their customer base and turn a quick profit. Companies that can identify the bracket of people 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 based on customer need. 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). Churn can be due to the non-satisfaction of current services, better offerings from other service providers and even 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 predicting specific metrics such as the high-risk customers, estimated time to attrite and likelihood to churn.

The learnings from multiple such exercises have been introduced as deployable machine learning algorithms that have been iterated over and refined based on the evolving need to flag consumers more accurately. The selection of techniques to employ 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 churn in our methodology through this research.

## 1.2 Problem Statement

With the customer data acquired from the telecom company, we will accurately flag customers' bracket likely to churn. This research will help telecom companies leverage their database to predict and actively target campaigns to customers that might churn. The methodology can be a set standard in the industry where multiple machine learning algorithms can run on a newer dataset, we can monitor the accuracy of the model, and customers can be appropriately targeted.

The recommended model's primary users will be Telecom companies that wish to reduce customer attrition by leveraging what Data Science offers. Given that the model predicts customers that will churn accurately, this can be done with limited hardware and regular cadence.

## 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 explaining and interpreting the prediction of 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 customer allocate budget and time to the customers that are likely to churn by running targeted campaigns. The sales team will be able to offer value-adds to high-risk and high-value customers. This can help the company document its customers' pain points and ultimately help aid in fundamental policy changes that can increase the overall profit.

## 1.7 Structure of Study

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

## 2.2 Data Analytics in the Telecom Industry

## 2.3 Customer Attrition Analytics

## 2.4 Predictive Modeling in Telecom Churn on different datasets

## 2.5 Visual Analytics in the Telecom Industry

## 2.6 Related Research Publications

## 2.7 Discussion

## 2.8 Summary

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.1 Introduction

Write something here

### 3.1.1 Business Understanding

In this paper, we were able to identify that the telecom industry is a highly competitive industry where customers have the free will to move across companies 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 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 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. The telecom churn data consists of 8043 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.

## 3.2 Research Methodology

### 3.2.1 Data Selection

### 3.2.2 Data Pre-Processing

Data cleaning for the telecom churn dataset will occur by first doing a sense check if the data. Once it is verified that the data types of the data are expected, we will check on the shape of the data to ensure the number of rows and columns is consistent with our 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 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 target variable.

### 3.2.3 Data Transformation (Feature Engineering)

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. 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 high 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

### 3.2.5 Class Balancing

### 3.2.6 Model Building

We shall now proceed to model building to 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. From the literature, it has been seen 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.

#### 3.2.6.3 Model Iterations

After the model, as mentioned earlier, building steps are performed, we shall now perform more iterations on the models, correspondingly analysing 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.

  
Figure 1: Model Building Process by Author via [draw.io](https://app.diagrams.net/)

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 Model 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 reevaluate 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 mosses, 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.

#### 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

## 3.4 Summary

# REFERENCES

# APPENDIX A: RESEARCH PLAN

# APPENDIX B: RESEARCH PROPOSAL

# APPENDIX C: ETHICS FORMS

## 