**Customer Churn Prediction Using Predictive Analytics**

The Telecom Industry as Case Study

Contents

[**DECLARATION** 4](#_Toc111552077)

[**ACKNOWLEDGEMENT** 5](#_Toc111552078)

[**ABSTRACT** 6](#_Toc111552079)

[**CHAPTER ONE - INTRODUCTION** 7](#_Toc111552080)

[**1.1** **Introduction** 7](#_Toc111552081)

[**1.2** **Customer Churn in the Telecommunication Industry** 7](#_Toc111552082)

[**1.3** **Rationale for the Study** 8](#_Toc111552083)

[**1.4** **Context of the Study** 8](#_Toc111552084)

[**1.5** **Research Questions** 9](#_Toc111552085)

[**1.6** **Aim** 9](#_Toc111552086)

[**1.7** **Objectives** 10](#_Toc111552087)

[**1.8** **Delimitations of the Study** 10](#_Toc111552088)

[**1.9** **Limitations of the Study** 10](#_Toc111552089)

[**1.10** **Proposed Structure** 10](#_Toc111552090)

[**1.11** **Conclusion** 11](#_Toc111552091)

[**CHAPTER TWO - LITERATURE REVIEW** 12](#_Toc111552092)

[**2.1** **Introduction** 12](#_Toc111552093)

[**2.2** **Customer Churn** 12](#_Toc111552094)

[**2.3** **Data Mining** 13](#_Toc111552095)

[**2.3.1** **Data Mining Techniques** 14](#_Toc111552096)

[**2.3.2** **Predictive Analytics** 19](#_Toc111552097)

[***2.3.3*** **Customer Churn Prediction Modelling** 25](#_Toc111552098)

[**2.4** **Conclusion** 27](#_Toc111552099)

[**CHAPTER THREE - RESEARCH METHODOLOGY** 28](#_Toc111552100)

[**3.1** **Introduction** 28](#_Toc111552101)

[**3.2** **Research Methodology** 28](#_Toc111552102)

[**3.2.1** **Experimental Research** 28](#_Toc111552103)

[**3.2.2** **CRISP-DM Methodology** 29](#_Toc111552104)

[**3.3** **Conclusion** 55](#_Toc111552105)

[**CHAPTER FOUR – IMPLEMENTATION** 56](#_Toc111552106)

[**4.1** **Introduction** 56](#_Toc111552107)

[**4.2** **Environment Setup** 56](#_Toc111552108)

[**4.3** **Modelling** 56](#_Toc111552109)

[**4.3.1** **Model Development** 56](#_Toc111552110)

[**4.3.2** **Model Hyperparameter Tuning** 61](#_Toc111552111)

[**4.4** **Evaluation** 64](#_Toc111552112)

[**4.5** **Deployment** 69](#_Toc111552113)

[**4.6** **Conclusion** 69](#_Toc111552114)

[**CHAPTER FIVE – CONCLUSIONS AND FUTURE WORK** 70](#_Toc111552115)

[**5.1** **Introduction** 70](#_Toc111552116)

[**5.2** **Customer Churn Prediction** 70](#_Toc111552117)

[**5.3** **Future Work** 70](#_Toc111552118)

[**5.4** **Conclusion** 70](#_Toc111552119)

[**REFERENCE** 73](#_Toc111552120)

**Table of Figures**

[Figure 1: Data Mining Process 14](#_Toc111553963)

[Figure 2: Classification Technique 16](#_Toc111553964)

[Figure 3: Predictive Analytics Process 20](#_Toc111553965)

[Figure 4: Decision Tree 22](#_Toc111553966)

[Figure 5: Neural Networks 24](#_Toc111553967)

[Figure 6 - CRISP-DM Life Cycle 29](file:///C:\Users\user\Desktop\CLIENT%20PROJECTS\ACADEMIC%20CLIENTS\Digital%20Marketing%20Dissertation\Customer%20Churn%20Prediction%20Using%20Predictive%20Analytics%20Proposal.docx#_Toc111553968)

[Figure 7: Missing Values Visualization 34](#_Toc111553969)

[Figure 8: Dataset Report Function 35](#_Toc111553970)

[Figure 9: Churn Distribution 37](#_Toc111553971)

[Figure 10: GenderChurn Function 38](#_Toc111553972)

[Figure 11: Churn Gender Distribution 38](#_Toc111553973)

[Figure 12: Contract Distribution Function 39](#_Toc111553974)

[Figure 13: Contract Distribution Visualisation 39](#_Toc111553975)

[Figure 14: Senior Citizen Distribution 40](#_Toc111553976)

[Figure 15: Payment Method Distribution Function 41](#_Toc111553977)

[Figure 16: Payment Method Distribution Visualisation 41](#_Toc111553978)

[Figure 17: Monthly Charges Distribution 42](#_Toc111553979)

[Figure 18: Column Names Updating 43](#_Toc111553980)

[Figure 19: Columns Delete 44](#_Toc111553981)

[Figure 20: Variable Info 45](#_Toc111553982)

[Figure 21: Variable Summary 46](#_Toc111553983)

[Figure 22: Correlation Matrix Code 47](#_Toc111553984)

[Figure 23: Correlation Matrix 48](#_Toc111553985)

[Figure 24: Correlation Summary 49](#_Toc111553986)

[Figure 25: Bagging Ensemble Technique 50](#_Toc111553987)

[Figure 26: Random Forest Algorithm 51](#_Toc111553988)

[Figure 27: Random Forest Classifier 53](#_Toc111553989)

[Figure 28: Explainable AI 54](#_Toc111553990)

[Figure 29: Variable Extraction 57](#_Toc111553991)

[Figure 30: Final Dataset Features 57](#_Toc111553992)

[Figure 31: Label Encoding 58](#_Toc111553993)

[Figure 32: Correlation Values 59](#_Toc111553994)

[Figure 33: Data Split 59](#_Toc111553995)

[Figure 34: Model Development 60](#_Toc111553996)

[Figure 35: Model Prediction 60](#_Toc111553997)

[Figure 36: RandomForestClassifier Tuning 61](#_Toc111553998)

[Figure 37: Parameter Setting for Tuning 61](#_Toc111553999)

[Figure 38: Data Split for Tuning 62](#_Toc111554000)

[Figure 39: Tuning Training 62](#_Toc111554001)

[Figure 40: Parameter Tuning Results Display Function 63](#_Toc111554002)

[Figure 41: Parameter Tuning Output 64](file:///C:\Users\user\Desktop\CLIENT%20PROJECTS\ACADEMIC%20CLIENTS\Digital%20Marketing%20Dissertation\Customer%20Churn%20Prediction%20Using%20Predictive%20Analytics%20Proposal.docx#_Toc111554003)

[Figure 42: Confusion Matrix 65](#_Toc111554004)

[Figure 43: RandomForestClassifier Confusion Matrix 66](#_Toc111554005)

[Figure 44: Accuracy Equation 66](#_Toc111554006)

[Figure 45: Sensitivity Equation 67](#_Toc111554007)

[Figure 46: F-Score Equation 67](#_Toc111554008)

[Figure 47: Specificity Equation 67](#_Toc111554009)

[Figure 48: Precision Equation 68](#_Toc111554010)

[Figure 49: ROC Curve 69](#_Toc111554011)

[Figure 50: Classification Report 69](#_Toc111554012)

# **DECLARATION**

I declare that this report is my research work and has not been replicated in another university for academic purposes. While all relevant references and citations are made, I am solely responsible for every possible error or mistake both in ideation and implementation.

# **ACKNOWLEDGEMENT**

I want to thank my family – wife and kids for letting me focus on developing this work from start to completion. Also, sincere gratitude goes to Goodnews Daniel and Emmanuel Ikiomoye for their immense support throughout the development of this work. My hearty appreciation goes to my supervisor – Dr Pouria, for the patient and supportive supervision from the start of the project to the end. And most importantly, my boundless gratitude goes to God Almighty for the wisdom and grace to complete this work.

**Bamidele Rasaq**

# **ABSTRACT**

The Telecom industry is robust and highly competitive as modern tech advancement enables more industry entrants and inspires more competition. Thus, customers have the luxury of arbitrarily switching between service providers. This act of leaving a service provider for another is called churn. Customer churn is now a general concern for players in the telco industry. However, the dire need to tell if a customer will likely churn has become a greater common interest.

This study developed a predictive model on the publicly available and accessible IBM telco dataset using the RandomForestClassifier ensemble supervised learning technique. The model performed with an accuracy score of 80%, having a precision of 84%, a recall of 91%, and an F1-score of 87%. To improve the model’s performance, the model’s hyperparameters were tuned, and the result was an improved overall accuracy score of 81%.

The model’s performance and recall score showed to be top among supervised learning model performances in the existing literature on customer churn in the telecom sector.

**Keywords:** Customer Churn, Predictive Analytics, Model, RandomForest, Hyperparameter Tuning

# **CHAPTER ONE - INTRODUCTION**

## **Introduction**

*Chapter One* of this study reports covers the contextual background for the study, rationale for the study, research questions, research aim, research objectives, delimitations of the study, limitations of the study, and the proposed study structure.

## **Customer Churn in the Telecommunication Industry**

The telecommunications industry, saturated with fierce competition resulting from technological advancements, has been instrumental in industrialisation since the early 1990s. Through the decades, the industry has transitioned from an archetypically monopolistic ecosystem to a highly liberal and dynamically competitive (Meena & Geng, 2022).

The telecom sector has evolved through three transformational eras to its present state—the first era, based on copper networks, had a monopolistic market structure. However, in the 1980s, its partial liberalisation in the United States heralded its second era, capitalising on commercialisation, privatisation, market access, and competition. Both licensed and unregulated operators instead characterise the third era. (Noam, 2010) (Fransman, 2001).

Whereas global players such as American Telephone and Telegraph Company (AT & T), Nippon Electric Company (NEC) in Japan, France Telecom in France, Deutsche Telecom in Germany, and British Telecom (BT) in Great Britain, dominated the telecom sector in the early 1990s, the 1997 World Trade Organization (WTO) Agreement on Basic Telecommunications Services catalysed competition in 72 countries, thus, diffusing government’s influence while fanning the embers of competition among operators (Meena & Geng, 2022).

The telecom industry has transformed from what was described as a “natural monopoly” to a dynamically competitive market with customers arbitrarily switching between operators from time to time. This movement from one provider to another is referred to as customer churn. This fierce competition inevitably compels operators to develop strategies for acquiring new and retaining old customers since an increase in customer base should translate into more profits. However, given that studies have shown that acquiring a new customer is more costly than retaining an old one (Dahiya & Bhatia, 2015), many operators channel more resources towards keeping existing customers.

There are probably many strategies to build and improve customer relationships and retention. Still, given its vast customer base, it is not easy to individualise such plans manually with each customer. Thus, predictive analytics becomes vital. Predictive analytics helps to predict which customers are at the risk of churn so that such customer retention strategies can be channelled and redirected to them precisely (Jain et al., 2020)

## **Rationale for the Study**

The telecom sector is highly competitive. Customer data is growing exponentially, making it difficult to manage manually. Subscribers change providers arbitrarily because they have the leisure to retain the same Mobile Number Portability. And given that acquiring new customers is more expensive than keeping existing ones, providers deploy predictive analytics to help them detect and predict future customer behaviour based on their historical data.

The existing literature on telecom customer churn prediction shows that most of the work focused on only providing prediction using supervised and unsupervised learning techniques and often without model explainability.

Therefore, this study aims to develop a model for churn prediction in the telecom industry on the publicly available IBM telecom customer churn dataset using an ensemble learning technique, provide model performance improvement through hyperparameter tuning, and provide model interpretability to enable service providers to identify not only subscribers on the verge of churn for customer retention marketing strategies but also understand the reasoning behind the model’s performance for trust-building.

## **Context of the Study**

Customer churn is a problem generally associated with service or customer-centric sectors and uniquely defined according to identified indicators. The industry of choice for this study is the telecoms industry because it has evolved from a highly monopolistic market to a dynamically competitive ecosystem. The fierce competition and vast customer data make it difficult, if not entirely impossible, for providers to deal with each customer according to their needs.

This data which often includes customer personal and transaction (calls) data, continues to increase, thus, necessitating the use of predictive analytics to detect where a high churn risk potentially exists. This study seeks to develop a churn prediction model for customer churn in the telecom industry by the following desired requirements in a churn prediction model (Balle et al. 2013, as cited in Esteves, 2016):

1. *Precision and Recall* - A churn prediction system must have a high level of precision (low number of false positives) and high level of recall (almost all churners must be identified)
2. *Performance* - A churn prediction model must have a high execution speed on new datasets to expedite business decision-making.
3. *Flexibility* - A churn prediction model must be flexible with a prediction with new data, considering all relevant changes in patterns.
4. *Scalability* - A churn prediction model must be able to react positively if data input increases
5. *Targeting* - A churn prediction model must be able to identify customers most likely to abandon a service provider.

This study, therefore, focuses on the literature on customer churn in the telecommunications industry and the empirical development of a model that successfully predicts customer churn for providers in the telecommunications industry with model interpretability.

## **Research Questions**

1. How can predictive analytics techniques effectively predict customer churn in the telecom industry?
2. How can a machine learning model be developed for predicting customer churn using the publicly available IBM telecom dataset?
3. How can model interpretation be applied to the machine learning customer churn prediction model?
4. What are the opportunities and challenges in applying predictive analytics and machine learning techniques in the telecom industry?

## **Aim**

This study aims to develop a machine learning model that predicts customer churn in the telecommunications industry.

## **Objectives**

1. To explore the current literature on predictive analytics application in customer churn prediction in the telecom sector
2. To develop a machine learning model based on the publicly available IBM telecom dataset
3. To improve the performance accuracy of the developed model
4. To explore the opportunities and possible challenges in the application of predictive analytics and machine learning in the telecom sector

## **Delimitations of the Study**

Customer churn remains a common menace for all service and customer-centric sectors such as transport, banking, insurance, eCommerce, retail, betting, etc., albeit this research study will be limited to customer churn prediction in the telecom industry. Given that acquiring the actual data of any global telecom players proved elusive, the publicly available IBM telecom dataset will be used. This is because the focus of this study is not so much about the dataset as it is about the predictive analytics technique used. Thus, all analysis, pre-processing, and modelling will be implemented only on the IBM telecom dataset. This means that the findings of this study shall not be applied to data of service providers from other sectors.

## **Limitations of the Study**

Acquiring datasets or real-time data from an actual telecom operator is difficult, if not impossible, given the required compliances with General Data Protection Regulations (GDPR) and other related regulations. Also, given that the research duration is limited, data scraping and collection from a genuine firm and exploratory data analysis from scratch is impossible. This underscores the choice for a publicly available and accessible dataset on which many research studies have already been conducted.

## **Proposed Structure**

The structure of this study will be in chapters as follows:

1. Chapter One - Introduction
2. Chapter Two - Literature Review
3. Chapter Three - Research Methodology
4. Chapter Four - Implementation
5. Chapter Five - Conclusion

## **Conclusion**

This Chapter discussed the contextual background of the study, its research questions, aim and objectives, delimitations and limitations, and the structure of the study. The next chapter will discuss studies conducted on customer churn in the telecoms industry and predictive analytics, including data mining techniques and machine learning algorithms applicable to our research focus.

# **CHAPTER TWO - LITERATURE REVIEW**

## **Introduction**

*Chapter Two* reviews the literature on customer churn in the telecommunications industry, data mining and its techniques, predictive modelling or analytics, machine learning algorithms, and a conclusion.

## **Customer Churn**

Customer churn, or customer attrition, refers to consumer movement from one provider to another (Cunha Esteves, 2016). Simply put, it is when a customer leaves a product or service (Coussement et al., 2017). In a customer-centric and dynamically competitive industry like telecommunications, subscribers have the luxury of quickly switching from one provider to another.

As the forces of globalisation and market deregulation policies, new competitors, and technologies continue to increase the competition in most economic sectors, service-providing companies have become preoccupied with creating stronger customer bonds. Thus, customer churn prediction has become a competitive advantage (García et al., 2016). Since acquiring a new customer is more expensive than keeping an existing one, service providers in the telecom sector develop strategies to retain their loyal customers to preclude them from churning (Wei & Chiu, 2002, M. Almana et al., 2014, Hanif, 2019).

Further, churn is either voluntary or involuntary. Churn is voluntary where a customer actively discontinues using a service or product of a service provider, often based on specific reasons such as service dissatisfaction or cost; it is involuntary when a customer leaves a service or product because of circumstances outside of the customer’s and or provider’s control. Such events may include death, relocation, etc. (Wei & Chiu, 2002; M. Almana et al., 2014, Hanif, 2019).

However, authors Tsai & Lu (2009) identified two main targeted approaches to managing customer churn: reactive and proactive approaches. The former is where the service provider swings into churn prevention or customer retention actions, including incentivising the customer with promos or other advantages only after a customer request to discontinue using the service provider’s services. On the other hand, the latter is where a service provider tries to identify customers with a high propensity to churn and incentivise them to retain them before they do (Cunha Esteves, 2016). Thus, churn prediction helps to identify this situation in advance for providers to carry out preventive actions (Celik & O. Osmanoglu, 2019).

This work focuses on the proactive approach and uses predictive analytics to predict and identify which customers will likely churn beforehand.

## **Data Mining**

The amount of data generated today is enormous. Data is daily generated from Internet of Things (IoT) devices to social media, industry 4.0, retail 4.0, chatbots, artificial intelligent devices and apps, weather forecast, eCommerce, etc. This humongous data and its exponential growth are attributed to the emergence of the internet and its associated technologies such as artificial intelligence (AI), Virtual Reality (VR), Augmented Reality (AR), social network services (SNS), eCommerce, social media, etc. Often stored in databases, this data is volatile, voluminous, complex, unstructured, heterogenous, and generated speedily and rapidly, thus, making it difficult to manage manually and process. This underscores the need for data mining, which helps elicit meaningful information from the vast amount of such generated data to aid business decision-making.

According to Ming-Syan Chen et al. (1996), data mining is a nontrivial extraction of implicit, previously unknown, and potentially useful information (such as knowledge rules, constraints, and regularities) from data in databases. Similarly, Lakshmi & Raghunandhan (2011) described it as extracting interesting patterns or knowledge from a vast amount of data. Put differently, data mining is a computer-aided process that digs and analyses enormous data sets and extracts their knowledge or information. Or, more precisely put, data mining automates the detection of relevant patterns in a database. This process applies to sales, sound, weather forecasting, revenue, products, electric load prediction, product design, etc. (M Raval, 2012, Zuha & Achuthan, 2016).

Often described as knowledge discovery in databases (KDD), knowledge extraction, data/pattern analysis, data archaeology, data dredging, information harvesting, business intelligence and several others, finding a universally acceptable definition was challenging as these multiple terms failed to present an accurate definition, albeit data mining can be referred to as the analysis of data in a database using tools which look for trends or anomalies without the knowledge of the meaning of the data. (M Raval, 2012, Zuha & Achuthan, 2016). Furthermore, it discovers previously unknown relationships among the data while extracting its information. This process applies to any repository. Zhang et al. (2018) present the data mining process in the diagram below:

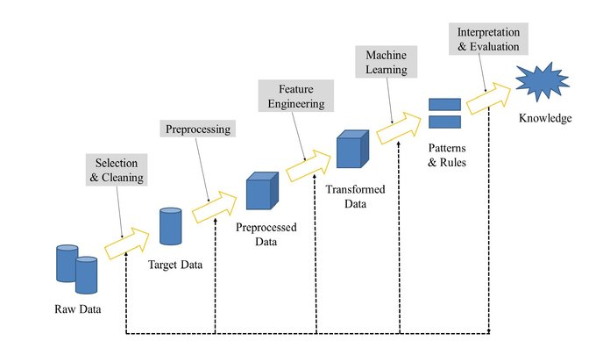


Figure 1: Data Mining Process

Figure 1. above shows the six distinct phases of data mining, but the process can be collapsed into or subsumed under three significant steps: exploration, pattern recognition, and model deployment. But first, these steps entail exploring the acquired raw dataset by selecting and cleaning the target dataset. The cleaning and preprocessing step involves removing incorrect, null, or missing values from the dataset.

Feature engineering also consists in extracting the relevant variables or columns and defining the target variable. This is done following the cleaning and preprocessing. But to make accurate predictions, the target variable or column must be converted to binary to develop a classification-based prediction model. The result of this feature extraction process or step produces what is called a Vector Space Model. Upon complete development of the prediction model, which involves feeding the vector space model to a machine learning algorithm, the model is evaluated for performance and prediction accuracy and, if good, deployed. A prediction model’s performance and accuracy are determined by the size and quality of the dataset.

### **Data Mining Techniques**

Modern organisations have the luxury of data more than ever. A massive amount of data is available at their disposal daily. However, given its volume, velocity, and variety, it is challenging to manually process such an avalanche of structured and unstructured data and successfully elicit its inherent knowledge or information for the organisation’s overall benefit. Data mining depicts a process that enables organisations and businesses to discover patterns or trends or derive meaningful insights from data for better and more effective business decision-making (Ramageri, 2010, Zuha & Achuthan, 2016, Qiao & Jiao, 2018, Jassim & Abdulwahid, 2021).

Thus, various businesses or organisations deploy different data mining techniques to convert raw data into meaningful or actionable insights, ranging from advanced artificial intelligence to fundamental data planning. Data mining techniques include the following:

* + - 1. **Classification**

Classification is the most common data mining technique (Zuha & Achuthan, 2016). As the name suggests, classification allows the user to classify into a predefined set of classes. It is a data mining task of predicting the value of a categorical variable (target or class) by building a model based on one or more numerical and categorical variables (predictors or attributes) (Zaki et al., 2014m, Cunha Esteves, 2016).

Put differently, classification assigns items in a collection to target categories or classes to accurately predict the target class for each case in the dataset. Often, a classifier is constructed to predict categorical labels (the class label attributes)(Kesavaraj & Sukumaran, 2013). Use cases of classification, among others, include loan applicants’ credit risk level identification (low, medium, or high), fraud detection, patient classification, churn prediction, cancer prediction, cyber threats recognition, etc.

Take telecom customer churn prediction, for example. The classification begins with a dataset in which the class assignments are known. These classes could range from customer personal data, including date of birth, address, level of education, job title, sector, monthly salary, subscription plans, and related transaction history. Thus, while the known variables are the predictors, the outcome (which in this case is churn or no churn) is the target variable.

Further, these known variables and targets are fed into a classification-based algorithm for the training and development of a model. During the model development (training) process, a classification algorithm finds relationships between the predictors and the target's values. Still, different classification algorithms often use different techniques for finding relationships. But at the end of the process, the identified and defined relationships are summarised in a model, which is then applied to a different dataset in which the class assignments are unknown. Kesavaraj & Sukumaran, 2013). Popular classification techniques include Decision Trees (DT), Neural Networks (NN), Naïve Bayes (NB), Support Vector Machines (SVM) and classification based on association (Kesavaraj & Sukumaran, 2013, Zuha & Achuthan, 2016).

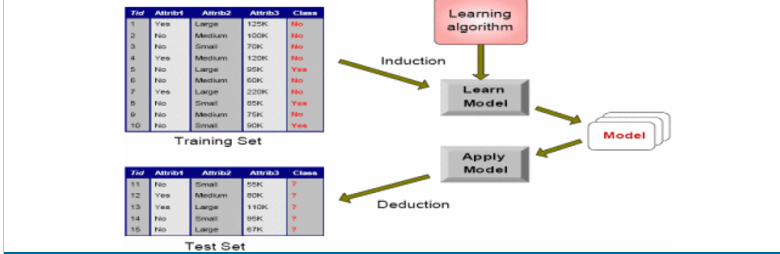


Figure 2: Classification Technique

According to Kesavaraj & Sukumaran (2013), “the commonly used methods for data mining classification tasks can be classified into the following groups.1. Decision tree induction methods, 2. Rule-based methods, 3. Memory-based learning, 4. Neural networks, 5. Bayesian network, 6. Support vector machines.”

* + - 1. **Clustering**

Diday & Simon (1976) described cluster analysis as a pattern recognition technique that characterises resemblance or dissemblance measures between the objects to be identified to obtain a symbolic description of the problem and an identification procedure. Clustering algorithms presented according to extended descriptions include hierarchies, minimum spanning trees, partitions, and their representations.

However, Lee (1981)described clustering analysis as a product of research fields such as computer science, statistics, pattern recognition, and operations research. The author further posited that clustering analysis has different names because of the diverse backgrounds of researchers. Accordingly, while it is "taxonomy" in Biology, it is "unsupervised learning” in pattern recognition. The author further argued that "classification" sometimes denotes clustering or discriminant analysis, which differs from clustering analysis.

Differentiating between both, the author further argued that while clustering analysis is pertinent to dividing a given set of samples into homogeneous groups, discriminant analysis is dividing a given set of samples into groups with the foreknowledge of which group each sample belongs to. Simply put, discriminant analysis constructs a decision rule that classifies these samples and determines to which class a new sample belongs while clustering groups samples according to similarities or dissimilarities.

Similarly, Berkhin (2006) argued that clustering divides data into groups of similar objects. Clustering can be viewed as a data modelling technique that provides concise data summaries. Agreeing with Lee (1981), Berkhin posited that clustering is related to many disciplines and plays a significant role in many applications. They often deal with large datasets and data with many attributes. Also, Ramageri (2010) described clustering as identifying similar classes of objects. The classification approach can also be used to distinguish groups of objects.

Clustering can be used as preprocessing method for attribute subset selection and classification, grouping customers based on purchasing patterns, and categorising genes with similar functionality. Types of clustering methods include partitioning methods, hierarchical agglomerative (divisive) methods, density-based methods, grid-based methods, and model-based methods. Clustering is a form of data extraction used to classify related data. This method helps to consider gaps between data and similarities (Bian et al., 2018,Jassim & Abdulwahid, 2021, Hai et al., 2022).

* + - 1. **Regression**

Chatterjee & Hadi (2006) explained Regression Analysis as a conceptually simple method for investigating functional relationships among variables. For example, regression analysis is used to determine or examine whether cigarette consumption is related to socioeconomic and demographic variables such as age, education, income, and price of cigarettes. This functional relationship is expressed in an equation depicting the connection between the response variable and one or more predictor variables.

The above equation shows as the response variable and *x1, x2, xp* as the predictor variables. The *xp* denotes the number of predictor variables, and theis assumed a random error depicting the discrepancy in the approximation.

Also, Jassim & Abdulwahid (2021) defined regression analysis as a data extraction process describing and analysing the relationship between variables.In the same vein, Ramageri (2010) explained that regression analysis could be used to model the relationship between one or more independent variables (attributes already known) and dependent or response variables (expected prediction outcome).

In other words, regression is a data mining function that predicts a numeric value which can be used to model the relationship between one or more independent variables and dependent variables. For example, it could predict a customer churn given their income level, education, and past transactions. To apply regression to a dataset, the target values must be known. Using the previously mentioned example, historic data capturing customer personal data, educational qualification, job, monthly income, monthly subscriptions, and churn must exist to predict customer churn. Churn is considered the response variable, while all others are explanatory or predictor variables (Zaki et al., 2014, Cunha Esteves, 2016, Sarstedt & Mooi, 2018).

During model training, a regression algorithm determines the value of the response variable, assessed as a function of the explanatory variables. Upon complete development of the model, it can be applied to a different dataset with unknown target values. Four main groups of Regression algorithms are (1) Frequency Table (e.g. Decision Tree), (2) Covariance Matix (e.g. Multiple Linear Regression), (3) Similarity Functions (e.g. K – Nearest Neighbours), and (4) Others (e.g. Artificial Neural Networks, Support Vector Machines).

* + - 1. **Association Rules**

Introduced in the early 1990s, Association rules is a data extraction technique that helps to find the connection between two or more objects or detects a pattern in a dataset. (Bian et al., 2018,Jassim & Abdulwahid, 2021, Hai et al., 2022). Simply put, the association rule is an expression X→YX→Y, where *X* and *Y* are sets of items. The rule X→YX→Y means transactions containing a set X of items tend to have a set Y of items.

An expression {bread, cheese} →→ {wine} means that customers who buy bread and cheese often purchase wine. Confidence and support are two primary measures of association rules (Rauch, 2018).Association rules are designed to extract correlations, patterns, associations, or simple structures among items and satisfy the predefined minimum support and confidence in transaction databases. This data mining technique is widely applied in telecommunication networks, market and risk management, inventory control, etc. (Ünvan, 2020, Kotsiantis & Kanellopoulos, 2006, Yin & Han, 2003).

Association rules usually find frequent item sets among large data sets. This finding helps businesses make certain decisions, such as catalogue design, cross-marketing, and customer shopping behaviour analysis. The number of possible Association Rules for a given dataset is generally massive, but many of these rules are usually of little value (Ramageri, 2010).

Types of association rules include multilevel association rule, multidimensional association rule, and quantitative association rule.

* + - 1. **Outer Detection**

This data extraction technique refers to observing data points that do not fit the predicted behaviour pattern in the data collection. This technique is applied in areas like intrusion detection, fraud, etc. (Bian et al., 2018,Jassim & Abdulwahid, 2021, Hai et al., 2022).

* + - 1. **Sequential Patterns**

This data mining technique helps to detect or discover similar patterns or trends in transaction data over a given timeframe. (Bian et al., 2018,Jassim & Abdulwahid, 2021, Hai et al., 2022). According to Fournier-Viger et al. (2017),sequential pattern mining consists of finding interesting sub-sequences in a set of sequences, where the interestingness of a subsequence can be measured via criteria such as its occurrence frequency, length, and profit. Sequential pattern mining is applied in bioinformatics, e-learning, market basket analysis, text analysis, and webpage click-stream analysis.

* + - 1. **Prediction**

Prediction uses other techniques for data mining, such as patterns, sequences, clustering, grouping, etc., to analyse historical data to predict a future occurrence. (Bian et al., 2018,Jassim & Abdulwahid, 2021, Hai et al., 2022).

### **Predictive Analytics**

While most authors identify Data mining as a step in predictive analytics, the reverse is the case for others. This work sees predictive analytics as a broad term describing various statistical and analytical techniques to develop models that predict future events or behaviour (Nyce, 2007). In other words, predictive modelling, as it is otherwise known, refers to the development of a prediction model using any machine learning algorithm(s) to predict a future occurrence. It analyses current and historical data to make predictions using statistics, data mining, machine learning, and artificial intelligence techniques. (Kumar & L., 2018, Eckerson, 2007).

The predictive analytics model is defined precisely as a model which predicts at a detailed level of granularity. It generates a predictive score for each individual by learning from experience to predict an individual's future behaviour. This helps in making better decisions. The accuracy of results by the model depends on the level of data analysis.

The growth of data today not only necessitates but also compels business organisations to look in the direction of predictive analytics for meaningful insights extraction. This is incredibly helpful for critical business decisions and profitability. Also, its use helps organisations to be both forward-thinking and proactive. Sectors heavily deploying predictive analytics include eCommerce, marketing, pharmaceuticals, stock forecasting, insurance, health, transport, law & justice, banking, telecommunications, etc. (Kumar & L., 2018, Eckerson, 2007).

The predictive modelling process involves data collection, preprocessing, modelling, deployment, and validation. Predictive models are built on historical data points with the target or response variable and the predictors. Take the telecoms sector, for instance. Customers have the luxury of changing service providers at will and sometimes arbitrarily. Since telecom providers desire to acquire more customers for more profits, it becomes imperative to know or predict customer behaviour to make targeted offers to keep them from churning strategically. But to successfully predict a customer’s future behaviour (churn), the customer’s historical data, such as calls, number of dropped calls, number of calls for operator support, and several others, must exist as the explanatory variables. These variables are capable of influencing future behaviour.

Predictive analytics techniques include classification algorithms, regression algorithms, clustering, neural networks, and decision trees. These techniques can be grouped into classification models and regression models. The following section will discuss some of these machine learning algorithms used for predictive analytics.

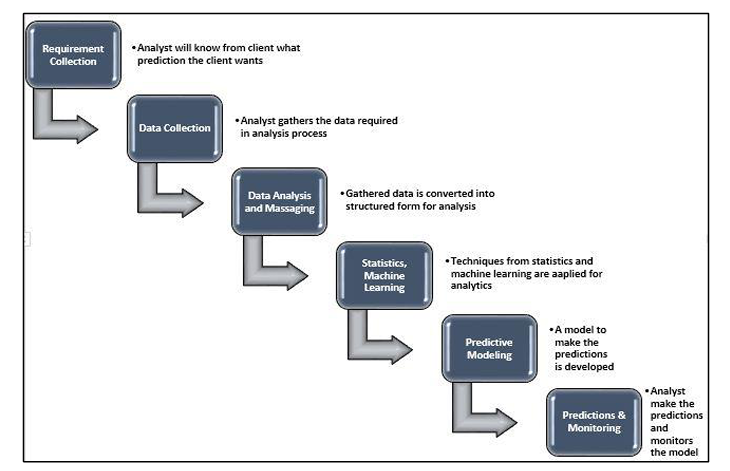


Figure 3: Predictive Analytics Process

#### **Classification Algorithms**

Classification recognises, understands, and groups objects into predefined classes. Classification algorithms use training datasets to predict the likelihood that future datasets will fall into one of the pre-categorized classes. Email classification, which involves classifying emails as either spam or no spam, is one typical example (F.Y et al., 2017, Wolff, 2020). Examples of classification algorithms include the following:

##### **Logistic Regression**

Logistic regression allows the analysis of dichotomous or binary outcomes with two (2) mutually exclusive levels. However, logistic regression permits continuous or categorical predictors and provides the ability to adjust for multiple predictors. (LaValley, 2008, F.Y et al., 2017). Logistic regression is a mathematical model or calculation used to predict a binary outcome based on current and historical data. It works with binary data, where either the event happens (1) or the event does not happen (0). So given some feature x, it tries to find out whether some event y happens or not. So y can either be 0 or 1 (Wolff, 2020).

P(Y=1|X) or P(Y=0|X)

It calculates the probability of dependent variable Y, given independent variable X. Logistic regression uses the logistic function to find a model that fits the data points. Logistic regression is used in machine learning to predict input values from previous test data (Maalouf, 2011).

##### **Naive Bayes**

This algorithm is based on Bayes’ theorem but is called “naïve” because of the strong assumption of independence between the features. Bayes’ theorem, given a class variable y and dependent feature 𝑥1 through 𝑥𝑛, states that:

𝑷( 𝒚 ∣ 𝒙𝟏, … , 𝒙𝒏) = 𝑷(𝒚)𝑷( 𝒙𝟏, … 𝒙𝒏 ∣ 𝒚 )

𝑷(𝒙𝟏, … , 𝒙𝒏)

Using the Naïve condition of independence:

𝑷(𝒙𝒊|𝒚, 𝒙𝟏, … , 𝒙𝒊−𝟏, 𝒙𝒊+𝟏, … , 𝒙𝒏) = 𝑷(𝒙𝒊|𝒚),

Since 𝑷 (𝒙𝟏, …, 𝒙𝒏) is constant because of the assumption of independence, we can use the classification rule. Despite this assumption of independence, the Naïve Bayes classifier still works well, and its use cases include the likes of spam filtering and document classification (Yang, 2018, Hanif, 2019).

##### **K-Nearest Neighbours**

K-nearest neighbours (k-NN) is a pattern recognition algorithm that uses training datasets to find the k closest relatives in future examples. When k-NN is used in classification, you calculate to place data within the category of its nearest neighbour. If k = 1, then it would be set in the class nearest 1. A plurality poll of its neighbours classifies K. (Harrison, 2019, Wolff, 2020),

The k-nearest neighbours (KNN) use proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another (Sun & Huang, 2010).

K-nearest neighbours (KNN) is a supervised learning algorithm for regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points closest to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of ‘K’ training data, and the class with the highest probability will be selected. In the regression case, the value is the mean of the ‘K’ selected training points. (Christopher, 2021)

##### **Decision Tree**

A decision tree is one of the most popular classification algorithms in data mining and machine learning. As the name suggests, in this algorithm, a tree-shaped structure represents a set of related decisions or choices (Hanif, 2019). It works like a flow chart, separating data points into two similar categories at a time from the “tree trunk” to “branches” to “leaves,” where the categories become more finitely similar. This creates categories within categories, allowing for organic classification with limited human supervision. (Wolff, 2020).

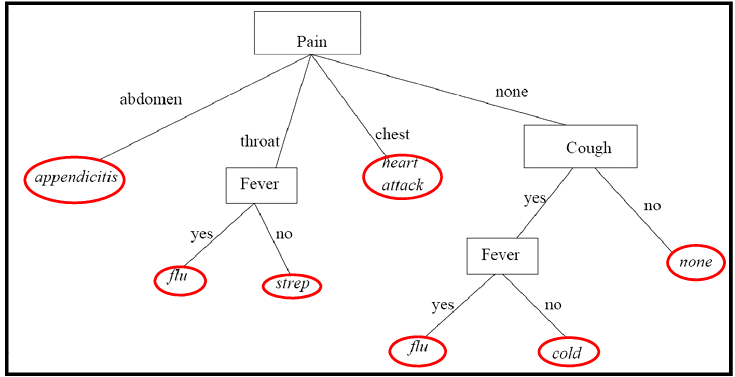


Figure 4: Decision Tree

##### **Support Vector Machines**

A support vector machine (SVM) is used for two-group classification problems. After giving an SVM model sets of labelled training data for each category, they can categorise new text. Compared to more unique algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it’s common to have access to a dataset of at most a couple of thousands of tagged samples. (Zhang, 2012) (Stecanella, 2018).

This algorithm works on the principle of assigning a boundary to a set of training examples that belong to one of two categories. There are several hyperplanes, and the algorithm’s objective is to find a hyperplane in n-dimensional space that distinctly classifies the data points. These hyperplanes act as decision boundaries that help classify the data points. The model in SVM is usually robust, meaning small changes do not lead to expensive remodelling and is relatively resistant to overfitting. However, it requires fully labelled data and has high computational costs. Its applications range from text categorisation to image classification (Hanif, 2019).

#### **Regression Algorithms**

Regression is learning relationships between inputs and continuous outputs from example data, enabling predictions for novel inputs (Stulp & Sigaud, 2015). Regression is necessary for many applications since many real-life problems can be modelled as regression problems. Types of regression include Locally Weighted Regression (LWR), rule-based regression, Projection Pursuit Regression (PPR), instance-based regression, Multivariate Adaptive Regression Splines (MARS) and recursive partitioning regression methods that induce regression trees (CART, RETIS and M5). (Uysal & Güvenir, 1999).

#### **Clustering**

According to Xu & WunschII (2005), clustering aims to separate a finite unlabelled data set into a finite and discrete set of “natural,” hidden data structures. Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories). There is no universally agreed-upon definition, but most researchers describe a cluster by considering the internal homogeneity and the external separation, i.e., patterns in the same cluster should be similar, while patterns in different clusters should not. Both the similarity and the dissimilarity should be examinable in a clear and meaningful way (Xu & WunschII, 2005).

There are different types of clustering algorithms that handle other datasets. These include the Density-based algorithm, Distribution-based algorithm, Centroid-based algorithm, Hierarchical-based algorithm, K-means clustering algorithm, DBSCAN clustering algorithm, Gaussian Mixture Model algorithm, and BIRCH algorithm (McGregor, 2020).

#### **Neural Networks**

According to Bishop (1994), neural networks provide a range of powerful new techniques for solving problems in pattern recognition, data analysis, and control with several notable features, including high processing speeds and the ability to learn the solution to a problem from a set of examples. Similarly, Tsai & Lu (2009) described neural computing as a pattern recognition methodology for machine learning. Artificial neural network (ANN) is mainly used for pattern recognition, forecasting, prediction, and classification. (Cunha Esteves, 2016).

Essentially, neural networks are inspired by the biological neural networks that constitute the brain and are, strictly speaking, not an algorithm but a framework for different machine learning algorithms to work together and process data inputs (Deep AI, 2018, Hanif, 2019). This works by learning and processing examples with input, learning the characteristics of the input and using this information to construct the output correctly. Once the algorithm has processed sufficient samples, the neural network can start processing unseen inputs and successfully return the correct results (Deep AI, 2018, Hanif, 2019). The following example explains better.

Chart

Description automatically generated

Figure 5: Neural Networks

Figure 5 above depicts the network structure of an Artificial Neural network (ANN). It comprises multiple neurons grouped in three layers: input, hidden and output. The input layer represents the values of the attributes from a dataset. The hidden layer which converts inputs from the input layer into results for further processing is often referred to as a feature extraction mechanism, albeit an ANN can have multiple hidden layers. The output layer contains the solution to a problem. The critical element in ANN is connection weights. They represent the relative importance of each input to a processing element. These weights are continuously adjusted, allowing the ANN’s learning process (Cunha Esteves, 2016).

#### **Decision Trees**

Rokach & Maimon (2005) described a Decision Tree as a classifier expressed as a recursive partition of the instance space, which consists of nodes that form a rooted tree. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes).

Decision trees are supervised machine learning often applied to solve regression and classification problems. Thus, the two main decision trees are classification and regression trees. Overall, classification trees are the primary application of decision trees in machine learning, albeit they can also solve regression problems. Their difference typically lies in the problem and dataset. While Classification trees are used for decisions such as yes or no, with a categorical decision variable, Regression trees are used for a continuous outcome variable such as a number (Seldon, 2021). Decision Trees (DTs) include C4.5, ID3, RULES-3, etc.

### **Customer Churn Prediction Modelling**

Churn prediction modelling refers to building a model for the target variable (churn or no-churn) as a function of the explanatory variables. A customer churn prediction model can accurately identify potential churners so that a retention solution may be provided. Accordingly, Wei & Chiu (2002) proposed a churn prediction model for customer retention based on Decision Tree C4.5 algorithm, while Dahiya & Bhatia (2015) proposed and implemented a prediction model based on Decision Trees (DT) and Logistic Regression (LR) using WEKA software.

Analysing the trends and causes of customer churn through data mining algorithms, Zhao et al. (2021) addressed questions about how customer churn occurs, its influencing factors, and how enterprises win back churned customers. Meanwhile, Jain et al. (2020) proposed a prediction model to predict customer churn using logistic regression and logit boost. Wael Fujo et al. (2022) proposed a prediction model based on deep learning techniques with results that outperformed Machine Learning techniques such as XG\_Boost, Logistic Regression, Naïve Bayes, and K-Nearest Neighbour (KNN).

Also, Ullah et al. (2019) proposed a churn prediction model built using Random Forest. Its accuracy performance was 88.63% while also identifying the root factors essential to determining the causes of churn. Siddika et al. (2021), in the same vein, carried out a comparative analysis of churn prediction models in which Random Forest achieved supremacy over the others, followed by deep learning (DL CNN) and MLP. For Kassem et al. (2020), however, the research focused on how user-generated content (UGC) could be used to determine customer churn and increase customer retention. This social media approach to developing a churn prediction model involved the analysis of UGC such as comments, posts, messages, and product/service reviews using sentiment analysis.

Kraljević & Gotovac (2010) suggested a novel methodology for developing data mining applications. According to the authors, the proposed framework has seven phases: business goals definition, data mining goals definition, data preparation, data modelling, analysis of results, deployment, and monitoring. The authors argue that this novel methodology is better than CRISP-DM and SEMMA. In recent work, Wael et al. (2022) proposed a prediction model using Deep-BP-ANN, which was implemented using two feature selection methods: variance thresholding and lasso regression. According to the authors, this model outperformed XG\_Boost, Logistic Regression, Naïve Bayes, and KNN.

The existing literature reveals that most customer churn prediction models are based on single statistical classification and regression algorithms. Still, Fakhar Bilal et al. (2022) proposed a hybrid model based on a combination of clustering and classification algorithms using an ensemble. The analysis of results indicated that the proposed model attained the highest prediction accuracy of 94.7% on the GitHub dataset and 92.43% on the Bigml dataset. Similarly, Xu et al. (2021) also proposed a customer churn prediction model using ensemble learning techniques that combined XG\_Boost, Logistic Regression, Decision Trees, and Naïve Bayes machine learning (ML) algorithms with experimental result accuracies of 96.12% and 98.09 % for the original and new datasets, respectively.

However, current literature on customer churn in the telecom industry shows little research work on inductive learning covering algorithms and their application in customer churn in telecom (M. Almana et al., 2014). This study seeks to fill this gap by applying RULES-3 and Random Forest to the IBM telecom dataset.

## **Conclusion**

This Chapter discussed the customer churn in the telecom industry, data mining and predictive analytics techniques, and the various existing customer churn prediction modelling techniques in the literature. The next chapter will focus on the selected dataset (IBM telco), its analysis, and its preprocessing techniques.

# **CHAPTER THREE - RESEARCH METHODOLOGY**

## **Introduction**

*Chapter* Three (3) discusses the chosen research methodology for telecom customer churn prediction, the selected dataset, its analysis, preprocessing and transformation, and the conclusion.

## **Research Methodology**

According to Igwenagu (2016), research methodology is a set of systematic techniques used in guiding how research is conducted, describing analysis methods, throwing more light on their limitations and resources, clarifying their pre-suppositions and consequences, and relating their potentialities to the twilight zone at the frontiers of knowledge. There are different types of research methodology. These include Action Research, Creative Research, Descriptive Research, Experimental Research, Expost-Facto Research, Expository Research, and Historical Research (Igwenagu, 2016). This study adopted the Experimental Research methodology.

### **Experimental Research**

The cornerstone of science is generally experimental and creative research. Thus, this study adopted the experimental research type. (Jenkins, n.d., Warfield, 2010). Empirical research is primarily concerned with cause and effect in which the variables of interest (i.e. the dependent and independent variables) are identified, and the impact of changes in the independent variables on the dependent variable is determined.

According to Harland (2010), experimental research is a study that strictly adheres to a scientific research design which includes a hypothesis, a variable that the researcher can manipulate, and variables that can be measured, calculated, and compared. Most importantly, experimental research is completed in a controlled environment. The researcher collects data, and the results either support or reject the hypothesis. In the same vein, Kandel (2011) sees experimental research as quantitative and empirical research carried out on the philosophical basis of positivism which is based on the experimental manipulation of the research variables during the study.

Also,experimental research seeks to determine a relationship between dependent and independent variables. Upon completion of an empirical study, a correlation between the independent and dependent variables is either supported or rejected. As such,experimental research data must be quantifiable. Thus, given that this study seeks to predict the behaviour of a dependent variable (whether a telecom customer will churn or not) based on the behaviour of independent variables (such as monthly customer charges, age, location, tenure duration, etc.), the adopted methodology for this study is the popular Cross Industry Standard Process for Data Mining (CRISP-DM) methodology (Wirth & Hipp, 2000).

### **CRISP-DM Methodology**

CRISP-DM is the de-facto standard and an industry-independent process model for applying data mining projects (Schröer et al., 2021). CRISP-DM, the most popular framework teams use to execute data science projects, provides an easy-to-understand description of the data science project workflow (i.e., the data science life cycle) (Saltz, 2021).

Published in 1999 to standardise data mining processes across industries, CRISP-DM has since become the most common methodology for data mining, analytics, and data science projects. Simply put, the Cross Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process (Hotz, 2021). Its six sequential phases include Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. These steps of CRISP-DM are as represented in the diagram below:

Diagram

Description automatically generated

Figure - CRISP-DM Life Cycle

#### **Business Understanding**

The first phase of CRISP-DM is business understanding. Data Science projects usually require understanding the business context, which involves setting project objectives, developing a project plan, defining business success criteria, and determining data mining goals. Thus, for this study, the business context is the telecom industry which has been adequately analysed in previous chapters. However, the focus of the study is Customer Churn in the telecom industry and how to predict the same using predictive analytics techniques.

#### **Data Understanding**

The second phase of CRISP-DM involves acquiring the dataset for the project and understanding it. It also consists in describing the dataset and assessing its quality. Consequently, the IBM telecom dataset publicly available and accessible on *kaggle.com* is accessed, described, and evaluated in this section.

##### **Data Overview**

The IBM telco dataset consists of seven thousand and forty-three (7043) entries with a total of thirty-three (33) rows, as shown below. Each row represents a customer; each column contains the customer’s attributes described on the column metadata. This analysis used Visual Studio Code (as the Code Editor), Python Jupyter Notebook, and several python machine learning libraries such as Pandas, NumPy, etc.

Table 1: Telco Customer Churn Dataset

A screenshot of a computer

Description automatically generated with medium confidence

**Dataset Description**

The dataset has four (4) categories of information:

1. Churn

Churn Label, Churn Score, Churn Value, and Churn Reason are the columns representing or describing customer churn status within the last month.

1. Customer Service Subscription

The columns showing customer service subscriptions are Phone, Multiple Lines, Internet, Online Security, Online Backup, Device Protection, Tech Support, and Streaming TV and Movies.

1. Customer Account Information

The columns showing customer account information are Tenure, Contract, Payment Method, Paperless Billing, Monthly Charges, and Total Charges.

1. Customer Demographics

The customer demographic information columns are Country, State, Gender, Senior Citizen, Zip Code, Lat Long, Latitude, Longitude, Partner, and Dependents.

**Dataset Columns Description**

* CustomerID: The customer’s Customer ID
* Country: The customer’s country of origin
* State: The customer’s State of Origin
* Zip Code: The customer’s zip code
* Lat Long: The customer’s location
* Latitude: Showing the customer’s location
* Longitude: Showing the customer’s location
* Gender: Customer gender (female, male)
* Senior Citizen: Whether the customer is a senior citizen or not (1, 0)
* Partner: Whether the customer has a partner or not (Yes, No)
* Dependents: Whether the customer has dependents or not (Yes, No)
* Tenure: Number of months the customer has stayed with the company
* Phone Service: Whether the customer has a phone service or not (Yes, No)
* Multiple Lines: Whether the customer has multiple lines or not (Yes, No, No phone service)
* Internet Service: Customer’s internet service provider (DSL, Fibre Optic, No)
* Online Security: Whether the customer has online security or not (Yes, No, No internet service)
* Online Backup: Whether the customer has online backup or not (Yes, No, No internet service)
* Device Protection: Whether the customer has device protection or not (Yes, No, No internet service)
* Tech Support: Whether the customer has tech support or not (Yes, No, No internet service)
* Streaming TV: Whether the customer has streaming TV or not (Yes, No, No internet service)
* Streaming Movies: Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract: The contract term of the customer (Month-to-month, One year, Two years)
* Paperless Billing: Whether the customer has paperless billing or not (Yes, No)
* Payment Method: The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* Monthly Charges: The amount charged to the customer monthly
* Total Charges: The total amount charged to the customer
* Churn Score: Churn score given according to rating
* Churn Value: Whether customer churn value is 1 or 0
* Churn Reason: Reason for the customer’s churn (Moved, Product dissatisfaction, Price too high)
* Churn: Whether the customer churned or not (Yes or No)

***Dataset Information***

To find out more information about the dataset, the *info () function* is used. This function shows the datatypes of the columns, counts the number of entries on each column, and displays the number of empty or null values in each column. The output below shows that the Churn Reason column has many null values.

Table 2: Dataset Info

Table

Description automatically generated

The matrix () function from the missingno library helps to visualise the missing values from the IBM telco dataset. The result is as below:

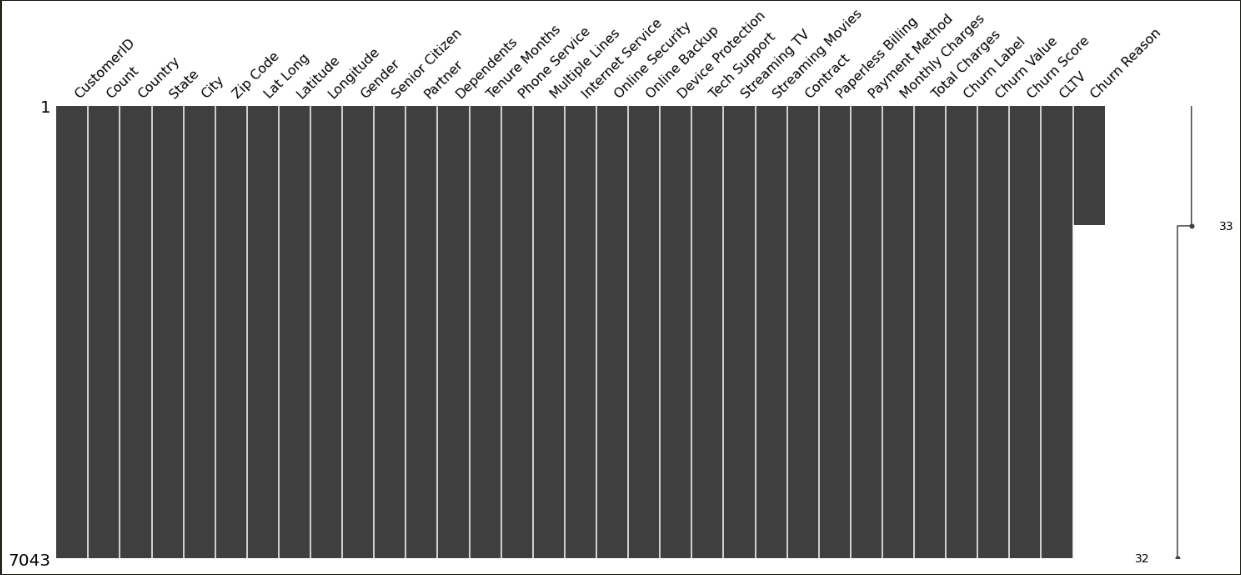


Figure 7: Missing Values Visualization

The figure above shows Churn Reason has many missing values, which makes this dataset unfit for modelling. It must first be prepared before it is fed to a machine learning model for prediction.

**Dataset Statistical Summary**

The *describe ()* function provides more precise information and insights from the IBM Telco Customer Churn dataset. The table below shows the statistical summary of the dataset, including count, mean, standard deviation, minimum, maximum, etc., details of each column. It reveals the oldest customer - the one with the longest tenure; it shows the highest monthly charges, the average tenure, and the Churn values.

This preliminary insight makes it possible to make accurate assumptions or conjecture on which customers are likely to churn even before feeding this dataset to a model for a more precise prediction.

Table 3: Dataset Summary

Table

Description automatically generated

The figure below is a function to help us determine the various unique values in every column and their number. This helps to determine the possible impact of the columns with fewer unique samples on the churn column and, as such, know which columns to delete.

Graphical user interface, text, application

Description automatically generated

Figure 8: Dataset Report Function

The output of the invoked *report ()* function shows that columns such as [Count, Country, State] have only one (1) unique value.

Table 4: Report Output

Graphical user interface

Description automatically generated

From the above report, columns Count, Country, and State, which have one unique value, will be deleted. Also, CustomerID will be deleted in that it does not determine churn probability. Other columns to be deleted include Zip code, Lat Long, Latitude, and Longitude. During model development, Churn Score and Churn Reason will be deleted to avoid information leaks, allowing insight at the analysis phase.

##### **Exploratory Data Analysis (EDA)**

The primary aim of the exploratory analysis is to examine the data for distribution, outliers, and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualising and understanding the data through graphical representation. It also assists with natural pattern recognition in the dataset. Finally, feature selection techniques often fall into EDA (Komorowski et al., 2016).

***Customer Churn in Data***

Chart, pie chart

Description automatically generated

Figure 9: Churn Distribution

The above figure shows the churn distribution plus indicating the exact number of female and male churners. It shows there are more female customers churn than their male counterparts.

***Variable Distributions***

The *genderChurn ()* function below helps to visualise both gender and churn distributions, revealing what percentage of customers are male and female and what percentage of customers are churners or not.

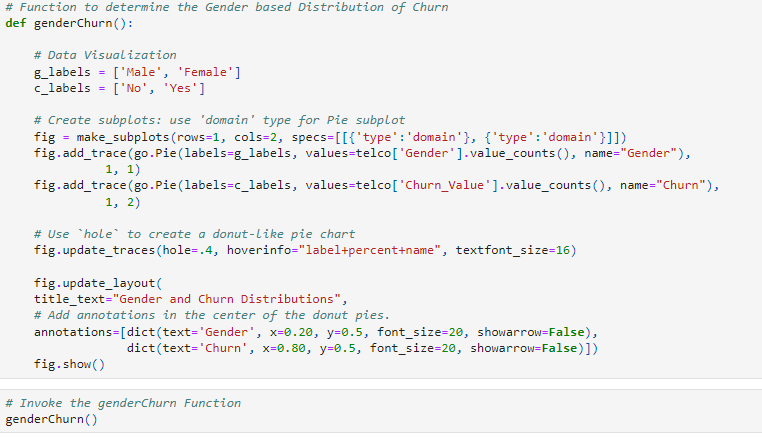


Figure 10: GenderChurn Function



Figure 11: Churn Gender Distribution

The visualisation figure above shows that 50.5% of customers are male while 49.5 are female, indicating more male customers than their female counterparts. Also, 26.5% of customers are churners, while 73.5% are not.

***Customer Contract Distribution***

**A picture containing graphical user interface

Description automatically generated**

Figure 12: Contract Distribution Function

**Chart, bar chart

Description automatically generated**

Figure 13: Contract Distribution Visualisation

The above distribution shows that customers with a month-to-month subscription are most likely to churn than the others, followed by the ones on a one-year contract and then the ones on a two-year contract.

**Senior Citizen Distribution**

**Chart, bar chart

Description automatically generated**

Figure 14: Senior Citizen Distribution

The above distribution figure shows the number of customers who are Senior Citizens from the IBM Telco dataset to be One Thousand One Hundred and Forty-Two (1142) out of Seven Thousand and Forty-Two (7042) total customers. Also, the distribution shows that the percentage of male Senior Citizen customers who are churners is almost the same as the female Senior Citizen customers.

**Payment Method Distribution**

**Graphical user interface, text, application

Description automatically generated**

Figure 15: Payment Method Distribution Function

Chart

Description automatically generated

Figure 16: Payment Method Distribution Visualisation

**Monthly Charges Distribution**

**Chart, histogram

Description automatically generated**

Figure 17: Monthly Charges Distribution

The monthly charges distribution shows that Customers with higher Monthly Charges are also more likely to churn.

#### **Data Preparation**

This phase of CRISP-DM involves preparing the dataset previously analysed and assessed as unfit for modelling. The process is also known as data preprocessing. It involves sampling, cleaning, identifying missing values and errors, and ensuring it is ready for the modelling phase. Given that the IBM telecom dataset is large and has missing values and possible errors, a sample set will be extracted and used, and that sample set will be cleansed and made ready for modelling.

Consequently, column names with spaces are replaced with underscores using the *columns. str. replace ()* function as shown below.

1. ***Column Name Space Removal***

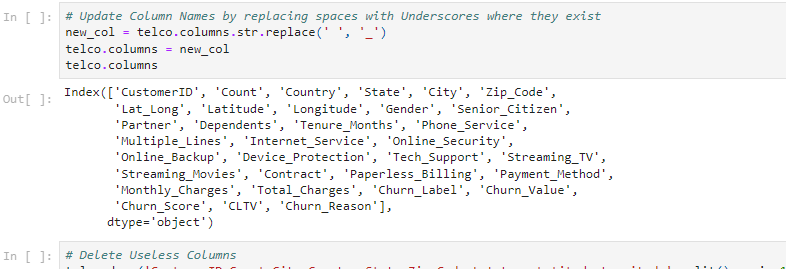
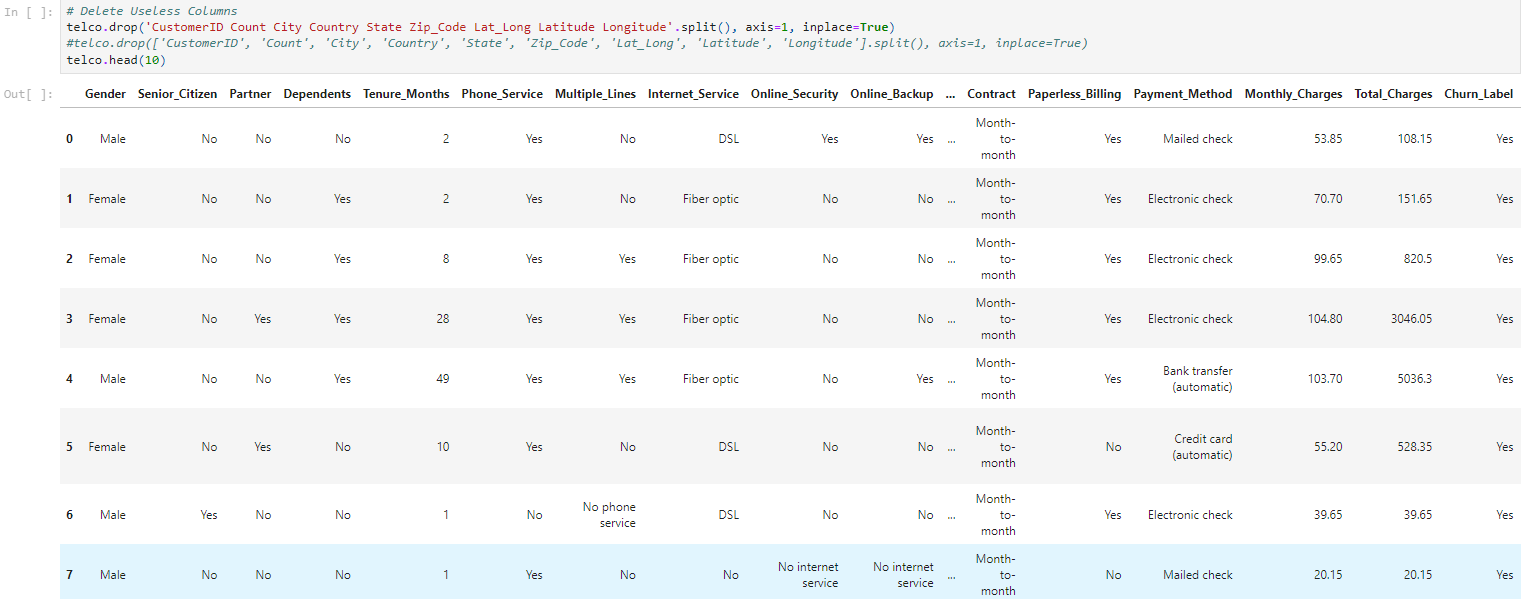


Figure 18: Column Names Updating

Further, columns with little or no influence on churn probability or with lots of missing values are deleted using the *drop ()* function as shown below.

Table 5: Dataset After Columns Deletion



1. ***Columns Deletion***

Columns Churn\_Label, Churn\_Score, CLTV, and Churn\_Reason will also be deleted to have a final data fit for modelling.

A picture containing graphical user interface

Description automatically generated

Figure 19: Columns Delete

1. **Variable Summary**

The *info ()* function output figure below shows no variable has any missing values, and the total number of variables is now twenty (20) only out of the original thirty-three columns.

Table

Description automatically generated

Figure 20: Variable Info

The info output above has three types of data: Object, int64, and float64.

Table

Description automatically generated with medium confidence

Figure 21: Variable Summary

1. ***Correlation Matrix***

A correlation matrix is a table that displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool for summarising a large dataset and identifying and visualising patterns in the given data.

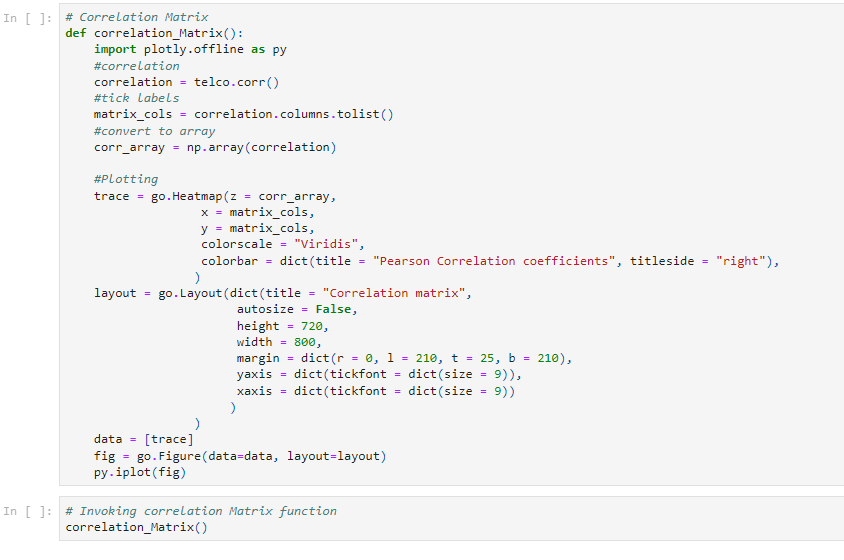


Figure 22: Correlation Matrix Code

A picture containing background pattern

Description automatically generated

Figure 23: Correlation Matrix

1. ***Data Split into Train and Test Sets***

All object datatype columns are converted to numeric datatype using this *object\_to\_int ()* defined function. The goal is to produce a vector space model with which a machine learning model can work, requiring all variables to be converted to one or zero (binary).

**Table

Description automatically generated**

1. **Correlation of Independent Variables with Target Variable (Churn\_Value)**

The figure below shows that Monthly\_Charges has the highest value and correlates more and influences the value of the target variable (Churn\_Value) than all the others.

Text

Description automatically generated with medium confidence

Figure 24: Correlation Summary

#### **Modelling**

In this phase, predictive analytics technique(s), Machine Learning algorithm(s), and data mining tools are selected for the study. It also involves generating a test design, applying the chosen techniques and tool(s) to the prepared IBM telecom dataset, and assessing the performance and quality of the model. Thus, this study uses an ensemble learning approach involving one statistical and predictive technique (Random Forest).

##### **Modelling Algorithm (Random Forest)**

According to Liu et al. (2012), Random Forest is a new Machine Learning (ML) algorithm which combines a series of tree structure classifiers widely used in classification, prediction, and regression. In the same vein, Kulkarni & Sinha (2012) described it as an ensemble supervised machine learning technique based on bagging and random feature selection, which generates several decision trees (base classifiers) with classification done via majority voting. Simply put, the “forest” Random Forest builds is an ensemble of decision trees, usually trained with the “bagging” method premised on the idea that combining learning models increases the overall result.

Random Forest uses the Bagging (Bootstrap Aggregation) ensemble technique. Bagging chooses a random sample from the dataset where each model is generated from the samples provided by the original dataset with a replacement known as row sampling. This step is called bootstrap. While each model is trained and generates results independently, the final output is based on majority voting after combining the results of all models. This step is known as aggregation (E R, 2021, Donges, 2021).

Graphical user interface, diagram

Description automatically generated with medium confidence

Figure 25: Bagging Ensemble Technique

***Random Forest Algorithm Steps***

Step 1: In Random Forest, *n* number of random records are taken from the dataset having k number of records. Step 2: Individual decision trees are constructed for each sample. Step 3: Each decision tree will generate an output. Step 4: Final result is based on Majority Voting or Averagingfor classification and regression, respectively.

Diagram

Description automatically generated

Figure 26: Random Forest Algorithm

***Essential Features of Random Forest***

* *Diversity* – Random Forest does not consider all features while making an individual tree as each tree is different.
* *Immune to Dimensionality* - Given that individual trees do not consider all the features, the feature space is reduced.
* *Parallelization* - As each tree is created independently from different data and attributes, they can fully use the CPU to build random forests.
* *Train-Test Split* - In a Random Forest, data segregation for train and test is unnecessary as there will always be 30% of the data which is not seen by the decision tree.
* *Stability* - Stability arises because the result is based on majority voting/averaging.

Random Forest has many good virtues and extensive application scope. Still, for effective learning and classification, there is a need to prune the trees (reduce the number of trees) in Random Forest. This is because the more trees it has, the better the result, but the slower the performance. This pruning is done manually or dynamically (Lin et al., 2017). Also, the hyperparameters can be tuned to improve Classifier accuracy.

***Random Forest Hyperparameters***

Hyperparameters enhance models' performance and predictive power or make the model faster. The following hyperparameters increase the predictive power of Random Forest:

* *n\_estimators* – This refers to the number of trees Random Forest builds before averaging the predictions.
* *max\_features* – This refers to the maximum number of features Random Forest considers before splitting a node.
* *mini\_sample\_leaf* – This hyperparameter determines the minimum number of leaves required to split an internal node.

But to increase the performance speed of Random Forest, the following hyperparameters are used:

* *n\_jobs* – This hyperparameter tells the engine how many processors to use. Thus, if the value is 1, it can only use one processor, but if the value is -1, there is no limit.
* *random\_state* – This hyperparameter is used to control the randomness of the sample. Thus, the model always produces the same results if given a definite random state value, the same hyperparameters, and the same training data.
* *oob\_score* – OOB, which stands for “Out of the Bag,” is a Random Forest cross-validation method which uses one-third of the sample to evaluate model performance. These samples are referred to as “out-of-bag” samples.

This study adopted the hyperparameter tuning approach to improve the accuracy of the Random Forest Classifier.

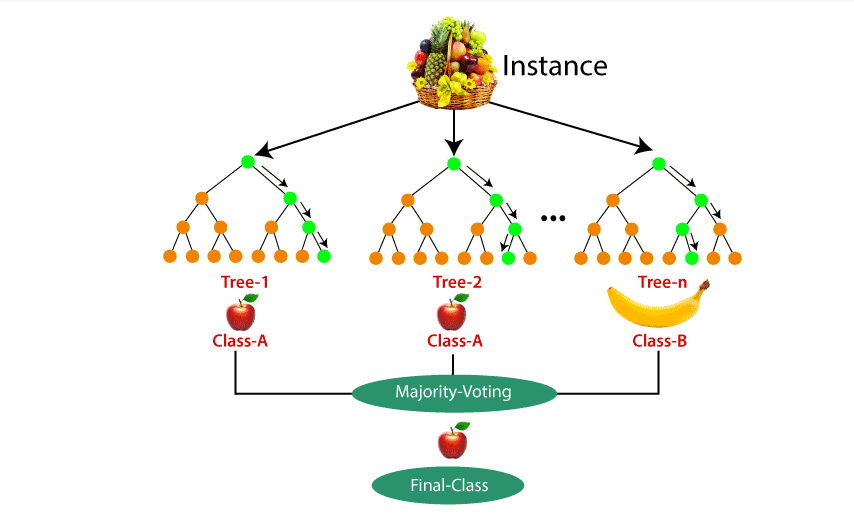


Figure 27: Random Forest Classifier

The Random Forest diagram above shows that the model works on a basket full of different fruits (representing the dataset), which splits into three other decision trees called Tree-1, Tree-2, and Tree-n. Each tree predicts with Tree-1 and Tree-2 voting Class-A (apple fruit) as the result, while Tree-n has Class-B (banana) as the result. Going by the majority voting Random Forest uses, the final class or output is apple.

##### **Model Interpretability**

Machine learning algorithms or models are usually referred to as black-box models because they often provide prediction output without explaining or interpreting the results for the user to understand. Typical examples are models which make detection, recommendation, or prediction, such as the presence of a disease or absence, determining which applicant is qualified for loan application based on credit score, which customer is most likely to churn, etc. These models produce results without adequately explaining why the output is recommended or interpreting the output for end users to understand. Thus, Explainable AI (XAI) helps to reveal what is happening within models and the reasoning behind the mathematical decisions. (Hulstaert, 2018, Dieber & Kirrane, 2020, Dataman, 2021).

This underscores the development of frameworks that enable developers to develop machine learning models and provide model interpretability upon deployment. The goal is to make machine learning models trustworthy. Thus, this study used the LIME (Local Interpretable Model-agnostic Explanations) framework to explain the model’s prediction. LIME is a technique that approximates any black box machine learning model with a local, interpretable model to explain each prediction. LIME offers a generic framework to uncover black boxes and provides the “why” behind AI-generated predictions or recommendations (Dieber & Kirrane, 2020).

The diagram below shows how a machine learning model works. It receives input data, processes it using its unseen mathematical equation (algorithm), and produces a result (output). The user sees and understands only the input and output, but the decision-making process and reasoning remain hidden. The work of Explainable AI (XAI) is to help demystify the reasoning behind the output from the model. LIME is one XAI implementation that explains the reasoning behind model decisions or recommendations.

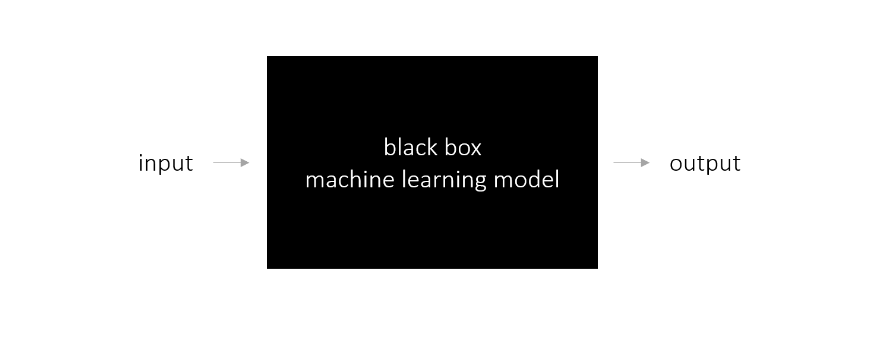


Figure 28: Explainable AI

#### **Evaluation**

This step involves evaluating the modelling phase and assessing the degree to which the model results meet the business (research) objective(s). Essentially, the entire process is reviewed to make the results satisfy the business needs. This study used the following metrics to evaluate the performance and accuracy of the developed prediction model: Precision, Accuracy, Recall/Sensitivity, F1-Call, Specificity, ROC Curve, and Confusion Matrix. The next Chapter will discuss these in more detail.

#### **Deployment**

This phase is essentially about making the model available for real-time use. It often requires a plan determining post-project steps. For this study, however, a final report with a summary of the evaluation results will be submitted. The next Chapter will discuss this phase of CRISP-DM.

## **Conclusion**

This chapter focused on the IBM Telco dataset. It described the research methodology for this study and analysed some of the steps involved in CRISP-DM as required and applied in the study. The next Chapter will focus on the study’s implementation, including the modelling, evaluation, and deployment phases of CRISP-DM.

# **CHAPTER FOUR – IMPLEMENTATION**

## **Introduction**

*Chapter* Four (4) discusses the environmental setup for developing a prediction model for telecom customer churn. It also focuses on developing the machine learning multi-algorithm model for the prediction, providing evaluation metrics for assessing the performance and accuracy of the developed prediction model, its deployment, and a conclusion.

## **Environment Setup**

Model development requires a particular development environment. This environment involves using development tools such as Visual Studio Integrated Development Environment (IDE), Jupyter Notebook, Python programming language with libraries such as Pandas, NumPy, SK-Learn, Matplotlib, Plotly, etc.

## **Modelling**

This is the fourth (4th) phase of the CRISP-DM methodology. It focuses on developing the prediction model using the tools above. This section will discuss the model development, its tuning, and model comparison.

### **Model Development**

In Chapter Three (3), the IBM telco\_customer\_churn dataset, this study showed some of the implementations of the first three phases of the CRISP-DM data science project development methodology. Chapter Three demonstrates dataset understanding, which provides an overview of information on the chosen dataset and its preparation. This section provides model development processes.

***Step 1: Feature Extraction***

To have a dataset perfectly prepared and fit for modelling, some more variables which do not correlate with the target variable will be deleted from the dataset. These variables include Churn\_Score, Churn\_Label, CLTV, and Churn\_Reason. Thus, using the *drop ()* function will achieve this result.

A picture containing graphical user interface

Description automatically generated

Figure 29: Variable Extraction

The above code snippet reduces the variables from twenty-three (23) to twenty (20). The *info* *()* function shows this more evidently.

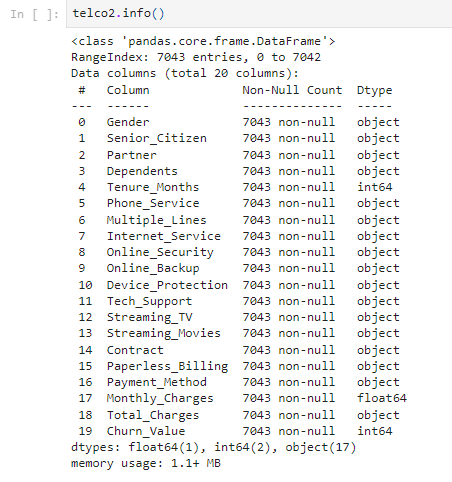


Figure 30: Final Dataset Features

The above figure shows that only three variables are numeric. The others are categorical (object data types). This form will produce biased predictions. For more accuracy, all object variables will be converted to numeric. This process is called Label Encoding. This makes the Vector Space Model that machine learning algorithms work with.

***Step 2: Label Encoding***

The figure below shows a function (object\_to\_int) that converts object datatype to the int data type with the *head ()* function output.

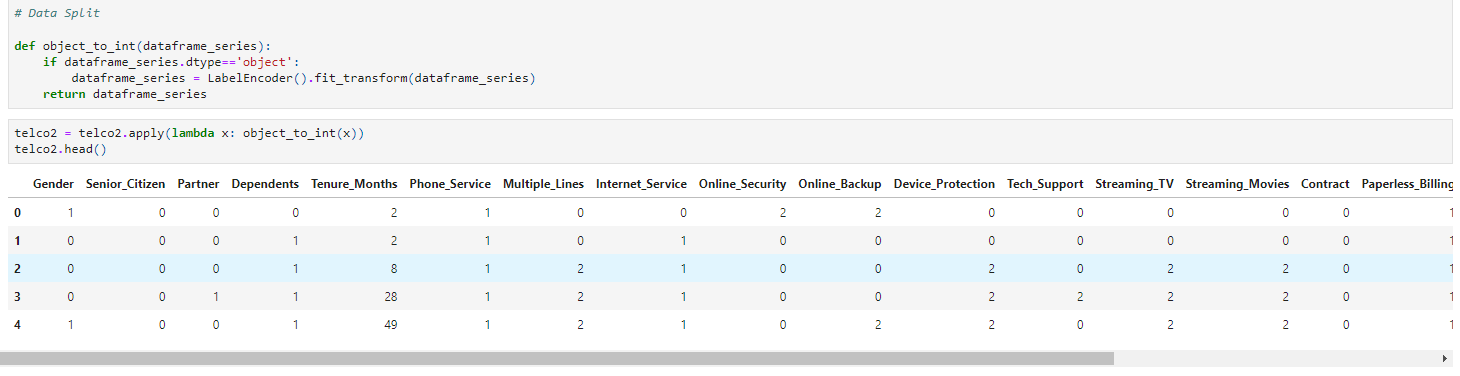


Figure 31: Label Encoding

***Step 3: Explanatory Variable Correlation***

Given the 20 label-encoded variables, the *corr ()* can help determine their correlations to the response variable (Churn\_Value). This figure below displays their respective correlation values sorted in descending order, showing that Monthly\_Charges correlates the most with the target variable. In other words, Monthly\_Charges is the highest factor determining customer churn.

Table

Description automatically generated with medium confidence

Figure 32: Correlation Values

***Step 4: Dataset Split into Train and Test***

First, this step allows to separate features from the target variable – *x* represents the independent variables (19 in total), and *y* represents the dependent variable (1). Second, using the *train\_test\_split ()* function, the Vector Space Model is divided into 70% for training and 30% for testing. The figure below demonstrates this.

Graphical user interface, application

Description automatically generated

Figure 33: Data Split

***Step 5: Model Development***

This step involves developing the customer churn prediction model using the Random Forest ensemble technique. This phase is probably the easiest in the entire methodology. The bulk of the work is handled in the data preparation phase. The figure below shows that the *RandomForestClassifier* class is instantiated, and its hyperparameters are specified. As earlier explained, the *n\_estimator* defines the number of trees to use; the out-of-bag bag score is set to true, the random\_state is 50, etc.

Then the defined instance, *model\_rf*, invoked the *fit ()* function of the RandomForestClassifier class and received *X\_train* (which has the features) and the *y\_train* (which has the target) for model training.

Text

Description automatically generated

Figure 34: Model Development

***Step 6: Model Prediction***

This step tests the trained model with the separated 30% of the dataset to see how it performs. The prediction model gives an accuracy score of **0.808329389493611**, **80%**. However, an accuracy score alone is not enough to evaluate the model. Thus, further evaluation using metrics such as Area Under Curve, Sensitivity, and Specificity, are provided in section 4.4.

Graphical user interface, text, application

Description automatically generated

Figure 35: Model Prediction

### **Model Hyperparameter Tuning**

Graphical user interface, text, application, email

Description automatically generated

Figure 36: RandomForestClassifier Tuning

As earlier explained, one of the several ways to improve the model’s prediction accuracy is by tuning the hyperparameters of the *RandomForestClassifier ()* constructor method. This method accepts parameters such as n\_estimates, random\_state, max\_features, n-jobs, etc. Tuning these parameters by giving higher or lower values and different combinations can potentially improve the prediction accuracy of the developed model. The following steps tune these hyperparameters.

***Step 1: Define Parameters***

Graphical user interface, text, application

Description automatically generated

Figure 37: Parameter Setting for Tuning

Notice the values set for the n\_estimators and max\_depth parameters. The original *RandomForestClassifier ()* constructor method has its n\_estimators set to 500, but this tuning sets five different values. Each value will be tested to see the output compared to each other. The goal is to know which parameters and combination with the max\_depth provide the best accuracy improvement.

***Step 2: Split Data***

The data is split into x and y - where x represents the feature variables and y represents the target variable. Data is further split into X\_train, X\_test, y\_train, and y\_test using the train\_test\_split () function.

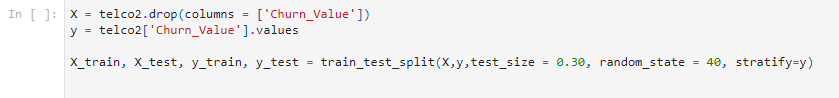


Figure 38: Data Split for Tuning

***Step 3: Train Model***

To train the model using the given tuned parameters, the GridSearchCV library has been used. The GridSearchCV () constructor method takes three arguments – rfc (RandomForestClassifier), parameters (defined in the previous steps) and cv. The model is then trained when the instance of the GridSearchCv invokes its fit () method, which also receives X\_train, and y\_train variables.

***Text, letter

Description automatically generated***

Figure 39: Tuning Training

***Step 4: Display Parameter Results***

***Text

Description automatically generated***

Figure 40: Parameter Tuning Results Display Function

The figure below shows the results of the parameter tuning. Every accuracy value is a combination of max\_depth value and n\_estimators value. The display function above iterates through all the param values for n\_estimators and max\_depth to display their respective performance accuracies. The results show a tuning combination of max\_depth = 8 and n\_estimators = 250 as the best parameters. This combination improved the Model’s accuracy score from 80% to 81%.

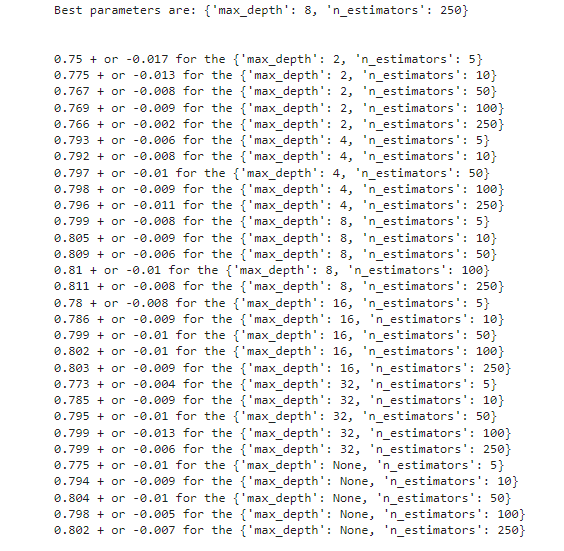


Figure : Parameter Tuning Output

## **Evaluation**

In this phase, the developed and trained model is evaluated. Though the model’s accuracy has been evaluated, it is not enough. If other metrics are not applied, the model might perform poorly when tested on new and unseen datasets. Thus, metrics such as explained below are used (Vujovic, 2021).

* **Confusion Matrix**

A confusion matrix is a matrix representation of the prediction results of any binary testing that is often used to describe the performance of a classification model (or “classifier”) on a test dataset for which the actual values are known (Chauhan, 2020).

Diagram, table

Description automatically generated

Figure 42: Confusion Matrix

Each prediction can be one of the four outcomes below based on its alignment with the actual value:

* True Positive (TP): Predicted True and True in reality.
* True Negative (TN): Predicted False and False in reality.
* False Positive (FP): Predicted True and False in reality.
* False Negative (FN): Predicted False and True in reality. (Chauhan, 2020)

The Confusion Matrix shows the actual values of the response variable (Churn\_Value). In binary classification problems, these values are usually positive (1) and negative (0). It also shows the correctly classified churners (True Positives) and the correctly classified non-churners (True Negatives); incorrectly classified churners (False Positives), and incorrectly classified non-churners (False Negatives).

The following figure shows the RandomForrestClassifier confusion matrix based on this study’s IBM customer churn dataset.

Chart, treemap chart

Description automatically generated

Figure 43: RandomForestClassifier Confusion Matrix

The above figure shows that 294 churners (TP) are correctly classified; 1414 non-churners are correctly classified (TN); 267 churners are incorrectly classified (FP), and 138 non-churners are incorrectly classified (FN).

* ***Accuracy***

Accuracy is a standard evaluation metric for classification problems. It is the number of correct predictions made as a ratio of all predictions (Chauhan, 2020). Accuracy is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00 (Vujovic, 2021).

***Diagram

Description automatically generated***

Figure 44: Accuracy Equation

* ***Recall/Sensitivity***

Sensitivity or Recall or True Positive Rate (TPR) is calculated as the number of accurate positive predictions (TP) divided by the total number of positive (P). Also called Sensitivity or Recall (REC). The best TP Rate is 1.0 and the worst 0.0 (Vujovic, 2021).

Diagram

Description automatically generated

Figure 45: Sensitivity Equation

* ***F1-Score***

F-score is a measure of the accuracy of the test. It is calculated, based on precision and reminders, by the formula: (Vujovic, 2021)

***A picture containing text

Description automatically generated***

Figure 46: F-Score Equation

* ***Specificity***

Specificity or True Negative Rate (TNR) - is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N). The best specificity is 1.0 and the worst 0.0 (Vujovic, 2021)

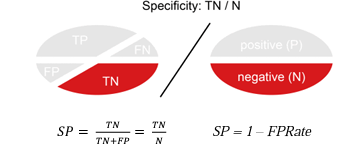


Figure 47: Specificity Equation

* ***Precision***

Precision is calculated as the number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP). The best accuracy is 1.0 and the worst 0.0. (Vujovic, 2021).

Diagram, pie chart

Description automatically generated

Figure 48: Precision Equation

* ***Receiver Operating Characteristic Area (ROC) Curve***

The ROC curve is a graph that visualises the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). True Positive and False Positive rates for each threshold are calculated and plotted on one graph. The area below the ROC curve is called the ROC AUC (Area Under Curve) score, a number that determines how good the ROC curve is (Vujovic, 2021). The ROC curve for this study is shown below. From the graph, the higher the TPR and the lower the FPR for each threshold, the better the model’s prediction or performance. Thus, when the TPR is 0.8, the FPR is 0.2.

Chart

Description automatically generated

Figure 49: ROC Curve

The figure below also shows the accuracy, precision, F1-score, and the other evaluation metrics of the RandomForestClassifier in this study.

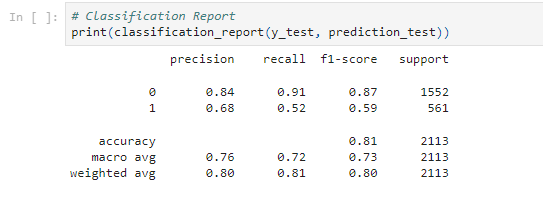


Figure 50: Classification Report

## **Deployment**

The deployment phase of CRISP-DM in this project involves the development of a report that shows the various development processes and results of this study.

## **Conclusion**

This chapter focused on the development of the customer churn prediction model. It involved setting up the development environment, the actual prediction model development and training, its performance evaluation using the machine learning traditional metrics, and the deployment phase of CRISP-DM. The next Chapter will conclude the study and suggest a direction (future work) for further research.

# **CHAPTER FIVE – CONCLUSIONS AND FUTURE WORK**

## **Introduction**

*Chapter Five (5)* provides a concise overview of the study. It also summarises the study’s outputs, suggests further research directions, and concludes the entire investigation with a conclusion.

## **Future Work**

The most significant risk in customer churn prediction is when a non-churner is predicted as a churner. That means resources that should have been used to retain real churners will probably be wasted on the wrong customer. That is why it is vital to have a model that predicts customer churn with a hundred per cent accuracy score. While that may be difficult to achieve with a supervised learning technique, this study’s model prediction accuracy score can be further improved beyond the current result.

The parameters were tuned to achieve a better result, but this study did not use other approaches to improving a model’s performance. Thus, further research on customer churn prediction will go toward performance improvement using different methods. Also, further feature extraction engineering needs to be done regarding the IBM telco dataset. Although 20 variables were extracted from the original 33 based on their distribution and correlation, further feature engineering work might help improve the vector space model and the overall performance of the model.

## **Customer Churn Prediction**

The relevance and significance of this study cannot be overemphasised as the global telecom sector continues to grow exponentially and competitively. Customer churn is a general concern for all players in the telecom sector. When customers churn, revenue is attenuated. Thus, every service provider within the telecom industry desires and works towards ensuring customers are retained. The goal is to maximise profit and stay afloat in the dynamic competition.

Customer churn is usually either voluntary or involuntary. The former is where a customer decides to leave a service provider for another for some reason(s), including service dissatisfaction; the latter is usually occasioned by circumstances beyond the control of the customer or the service provider. These circumstances may include death, relocation, etc. This study attempted to solve the voluntary churn problem by developing a predictive model that predicts whether a customer will likely churn or not, based on previous or historical data. While some ways to retain customers include upselling customers, the preferred way to retain customers is by predicting when customers are likely going to churn and focusing marketing campaigns on such customers to keep them from churning.

Research on customer churn in the telecom industry is enormous. Several machine learning algorithms have been used to develop models for churn prediction. These models range from supervised learning to unsupervised learning algorithms. But the identified gap in the existing literature is that research on ensemble learning with hyperparameter tuning for performance improvement is insufficient. Thus, this study attempted to fill this gap by developing an ensemble learning model based on RandomForestClassier by tuning its parameters for improved performance.

This study used the publicly available and accessible IBM customer churn dataset (available on Kaggle.com) for the analysis. As with every dataset, the IBM telco churn dataset required some data preprocessing and transformation. This was achieved by eliminating explanatory or independent variables with little correlation and possible impact on the target or response variable. Thus, upon complete preprocessing, the number of variables reduced from thirty-three (33) to twenty (20).

Also, the various variables needed to be encoded to have a dataset sufficiently prepared for modelling. That means the variables were converted from categorical to numeric. This was to have a Vector Space Model for modelling. A thorough Exploratory Data Analysis was conducted to determine various variable distributions on the dataset.

For the development and training of the prediction model, the Vector Space Model (that is, the finally cleaned, preprocessed, and transformed dataset) was split into sets of 70% and 30% for training and testing, respectively. Thus, the model was developed on the ensemble learning RandomForest classifier and trained on 70% of the dataset, while the other 30% was for its testing. The model evaluation was based on the model’s performance on the testing dataset.

Further, this study adopted a research methodology called Cross Industry Standard Process for Data Mining (CRISP-DM). It is a data science-based project development methodology that allows the development team to walk through data science projects successfully. Made up of six different and iterative phases, CRISP-DM enables project managers and developers alike to provide adequate supervision and complete and competent delivery of such projects. This study, thus, followed the six phases in the customer churn prediction model development processes.

The model was developed using the RandomForest ensemble learning technique, and its prediction accuracy score was eighty (80%). The machine learning traditional model evaluation metrics were used to assess the performance of the developed model. The model had a sensitivity or recall score of 91%, a precision score of 84%, and an F1-Score of 87%. A confusion matrix showing the various details for True Positives, True Negatives, False Positive, and False Negatives is provided.

To improve the model’s overall performance, the hyperparameters of the RandomForestClassifer were tuned. These parameters include the n\_estimators, random\_state, max\_depth, etc. The results showed that the best parameters are max\_depth of 8 and n\_estimators of 250, which produced an improved accuracy score of 81%.

## **Conclusion**

This study was able to complete the research objectives and meet its overall aim: to develop a predictive model for customer churn prediction in the telecom sector using the IBM telco dataset and an ensemble learning technique with improved performance based on hyperparameter tuning.

# **REFERENCE**

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data*, 25–71. https://doi.org/10.1007/3-540-28349-8\_2

Bian, J., Tian, D., Tang, Y., & Tao, D. (2018). A survey on trajectory clustering analysis. *ArXiv:1802.06971 [Cs]*. https://arxiv.org/abs/1802.06971

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments*, *65*(6), 1803–1832. https://doi.org/10.1063/1.1144830

Celik, O., & O. Osmanoglu, O. (2019). Comparing to techniques used in Customer Churn analysis. *Journal of Multidisciplinary Developments*, *4(1)*(30-38, 2019). https://www.researchgate.net/profile/Oezer-Celik-2/publication/337103029\_Comparing\_to\_Techniques\_Used\_in\_Customer\_Churn\_Analysis/links/5dc52a29a6fdcc2d2ffc1a24/Comparing-to-Techniques-Used-in-Customer-Churn-Analysis.pdf

Chatterjee, S., & Hadi, A. S. (2006). Regression Analysis by Example. In *Google Books*. John Wiley & Sons.

Chauhan, N. S. (2020). *Model Evaluation Metrics in Machine Learning*. KDnuggets. https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html

Christopher, A. (2021, February 3). *K-Nearest Neighbor*. Medium. https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

Cunha Esteves, G. (2016). *Churn prediction in the telecom business*.

Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*. https://doi.org/10.1109/icrito.2015.7359318

Dataman, D. (2021, July 19). *Explain Your Model with LIME*. Medium. https://medium.com/dataman-in-ai/explain-your-model-with-lime-5a1a5867b423

Diday, E., & Simon, J. C. (1976). Clustering Analysis. *Digital Pattern Recognition*, 47–94. https://doi.org/10.1007/978-3-642-96303-2\_3

Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *ArXiv:2012.00093 [Cs]*. https://arxiv.org/abs/2012.00093

Donges, N. (2021, July 22). *A Complete Guide to the Random Forest Algorithm*. Built In. https://builtin.com/data-science/random-forest-algorithm

E R, S. (2021, June 17). *Random Forest | Introduction to Random Forest Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Eckerson, W. (2007). *TDWI BEST PRACTICES REPORT FIRST QUARTER 2007 Extending the Value of Your Data Warehousing Investment*. http://download.101com.com/pub/tdwi/Files/PA\_Report\_Q107\_F.pdf

F.Y, O., J.E.T, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Fakhar Bilal, S., Ali Almazroi, A., Bashir, S., Hassan Khan, F., & Ali Almazroi, A. (2022). An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Computer Science*, *8*, e854. https://doi.org/10.7717/peerj-cs.854

Fournier-Viger, P., Chun, J., Lin, W., Kiran, R., Koh, Y., & Thomas, R. (2017). A Survey of Sequential Pattern Mining. *XXXX-XXXX Ubiquitous International*, *1*(1). https://www.philippe-fournier-viger.com/dspr-paper5.pdf

Fransman, M. (2001). Analysing The Evolution Of Industry: The Relevance Of The Telecommunications Industry. *Economics of Innovation and New Technology*, *10*(2-3), 109–140. https://doi.org/10.1080/10438590100000006

García, D. L., Nebot, À., & Vellido, A. (2016). Intelligent data analysis approaches to churn as a business problem: a survey. *Knowledge and Information Systems*, *51*(3), 719–774. https://doi.org/10.1007/s10115-016-0995-z

Hai, V. V., Nhung, H. L. T. K., Prokopova, Z., Silhavy, R., & Silhavy, P. (2022). Towards Improving the Efficiency of Software Development Effort Estimation Via Clustering Analysis. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2022.3185393

Hanif, E. (2019). *Application of data mining techniques for churn prediction and cross-selling in the telecommunications industry*. https://esource.dbs.ie/bitstream/handle/10788/3709/msc\_hanif\_e\_2019.pdf?sequence=1&isAllowed=y

Harland, D. (2010). *An Introduction to Experimental Research*. https://cemast.illinoisstate.edu/downloads/hsrs/types\_of\_research.pdf

Harrison, O. (2019, July 14). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Medium. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761#:~:text=Summary

Igwenagu, C. (2016). *Fundamentals of research methodology and data collection*. ResearchGate. https://www.researchgate.net/publication/303381524\_Fundamentals\_of\_research\_methodology\_and\_data\_collection

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Jassim, M. A., & Abdulwahid, S. N. (2021). Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. *IOP Conference Series: Materials Science and Engineering*, *1090*(1), 012053. https://doi.org/10.1088/1757-899x/1090/1/012053

Jenkins, A. (n.d.). *RESEARCH METHODOLOGIES AND MIS RESEARCH*. Retrieved April 6, 2020, from https://ifipwg82.org/sites/ifipwg82.org/files/Jenkins.pdf

Kandel, R. K. (2011). Experimental Research Design: A Play of Variables. *Journal of NELTA Surkhet Vol. 2*. https://www.academia.edu/41275899/Experimental\_Research\_Design\_A\_Play\_of\_Variables

Kassem, E. A. el, Ali, S., Mostafa, A., & Kamal, F. (2020). Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. *International Journal of Advanced Computer Science and Applications*, *11*(5). https://doi.org/10.14569/ijacsa.2020.0110567

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt.2013.6726842

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. *Secondary Analysis of Electronic Health Records*, 185–203. https://doi.org/10.1007/978-3-319-43742-2\_15

Kotsiantis, S., & Kanellopoulos, D. (2006). *Association Rules Mining: A Recent Overview*. Citeseerx.ist.psu.edu. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.6295&rep=rep1&type=pdf

Kraljević, G., & Gotovac, S. (2010). Modelling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services. *Automatika*, *51*(3), 275–283. https://doi.org/10.1080/00051144.2010.11828381

Kulkarni, V. Y., & Sinha, P. K. (2012, July 1). *Pruning of Random Forest classifiers: A survey and future directions*. IEEE Xplore. https://doi.org/10.1109/ICDSE.2012.6282329

Kumar, V., & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, *182*(1), 31–37. https://doi.org/10.5120/ijca2018917434

Lakshmi, B. N., & Raghunandhan, G. H. (2011, February 1). *A conceptual overview of data mining*. IEEE Xplore. https://doi.org/10.1109/NCOIET.2011.5738828

Lars Hulstaert. (2018, July 11). *Understanding model predictions with LIME*. Medium; Towards Data Science. https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b

LaValley, M. P. (2008). Logistic Regression. *Circulation*, *117*(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Lee, R. C. T. (1981). Clustering Analysis and Its Applications. *Advances in Information Systems Science*, 169–292. https://doi.org/10.1007/978-1-4613-9883-7\_4

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access*, *5*, 16568–16575. https://doi.org/10.1109/access.2017.2738069

Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. *Information Computing and Applications*, 246–252. https://doi.org/10.1007/978-3-642-34062-8\_32

M Raval, K. (2012). *Data Mining Techniques*. Citeseerx.ist.psu.edu. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.898.6657&rep=rep1&type=pdf

M. Almana, A., Sabih Aksoy, M., & Alzaran, R. (2014). A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. *Int. Journal of Engineering Research and Applications*, *4*(5), 167–171.

Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

McGregor, M. (2020, September 21). *8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know*. FreeCodeCamp.org. https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

Meena, M. E., & Geng, J. (2022). Dynamic Competition in Telecommunications: A Systematic Literature Review. *SAGE Open*, *12*(2), 215824402210946. https://doi.org/10.1177/21582440221094609

Ming-Syan Chen, Jiawei Han, & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883. https://doi.org/10.1109/69.553155

NICK HOTZ. (2021). *CRISP-DM*. Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/

Noam, E. M. (2010). Regulation 3.0 for Telecom 3.0. *Telecommunications Policy*, *34*(1-2), 4–10. https://doi.org/10.1016/j.telpol.2009.11.004

Nyce, C. (2007). *Predictive Analytics White Paper*. http://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf

Qiao, X., & Jiao, H. (2018). Data Mining Techniques in Analyzing Process Data: A Didactic. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02231

Ramageri, B. M. (2010). *(PDF) Data mining techniques and applications*. ResearchGate. https://www.researchgate.net/publication/49616224\_Data\_mining\_techniques\_and\_applications

Rauch, J. (2018). Expert deduction rules in data mining with association rules: a case study. *Knowledge and Information Systems*, *59*(1), 167–195. https://doi.org/10.1007/s10115-018-1206-x

Rokach, L., & Maimon, O. (2005). Decision Trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer New York, NY. https://doi.org/10.1007/0-387-25465-x\_9

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata52589.2021.9671634

Sarstedt, M., & Mooi, E. (2018). Regression Analysis. *Springer Texts in Business and Economics*, 209–256. https://doi.org/10.1007/978-3-662-56707-4\_7

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Seldon. (2021, November 13). *Decision Trees in Machine Learning Explained*. Seldon. https://www.seldon.io/decision-trees-in-machine-learning#:~:text=The%20two%20main%20types%20of

Siddika, A., Faruque, A., & Masum, A. K. M. (2021). Comparative Analysis of Churn Predictive Models and Factor Identification in Telecom Industry. *2021 24th International Conference on Computer and Information Technology (ICCIT)*. https://doi.org/10.1109/iccit54785.2021.9689881

Stecanella, B. (2018, November 19). *An introduction to Support Vector Machines (SVM)*. MonkeyLearn Blog. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

Stulp, F., & Sigaud, O. (2015). Many regression algorithms, one unified model: A review. *Neural Networks*, *69*, 60–79. https://doi.org/10.1016/j.neunet.2015.05.005

Sun, S., & Huang, R. (2010, August 1). *An adaptive k-nearest neighbour algorithm*. IEEE Xplore. https://doi.org/10.1109/FSKD.2010.5569740

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553. https://doi.org/10.1016/j.eswa.2009.05.032

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/access.2019.2914999

Ünvan, Y. A. (2020). Market basket analysis with association rules. *Communications in Statistics - Theory and Methods*, 1–14. https://doi.org/10.1080/03610926.2020.1716255

UYSAL, İ., & GÜVENIR, H. A. (1999). An overview of regression techniques for knowledge discovery. *The Knowledge Engineering Review*, *14*(4), 319–340. https://doi.org/10.1017/s026988899900404x

Vujovic, Ž. Ð. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/ijacsa.2021.0120670

Wael Fujo, S., Subramanian, S., & Ahmad Khder, M. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 185–198. https://doi.org/10.18576/isl/110120

WARFIELD, D. (2010). *IS/IT RESEARCH: A RESEARCH METHODOLOGIES REVIEW*. Citeseerx.ist.psu.edu. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.818.8919&rep=rep1&type=pdf

Wei, C.-P., & Chiu, I-Tang. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, *23*(2), 103–112. https://doi.org/10.1016/s0957-4174(02)00030-1

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Wolff, R. (2020, August 26). *Classification Algorithms in Machine Learning: How They Work*. MonkeyLearn Blog. https://monkeylearn.com/blog/classification-algorithms/

Xu, R., & WunschII, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/tnn.2005.845141

Xu, T., Ma, Y., & Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. *Applied Sciences*, *11*(11), 4742. https://doi.org/10.3390/app11114742

Yang, F.-J. (2018, December 1). *An Implementation of Naive Bayes Classifier*. IEEE Xplore. https://doi.org/10.1109/CSCI46756.2018.00065

Yin, X., & Han, J. (2003). CPAR: Classification based on Predictive Association Rules. *Proceedings of the 2003 SIAM International Conference on Data Mining*. https://doi.org/10.1137/1.9781611972733.40

Zaki, M. J., Jr, W. M., & Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. In *Google Books*. Cambridge University Press.

Zhang, C., Cao, L., & Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable Cities and Society*, *39*, 508–518. https://doi.org/10.1016/j.scs.2018.02.016

Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. *Communications in Computer and Information Science*, 179–186. https://doi.org/10.1007/978-3-642-34041-3\_27

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society*, *2021*, 1–12. https://doi.org/10.1155/2021/7160527

Zuha, F., & Achuthan, G. (2016). Analysis of Data Mining Techniques and its Applications. *International Journal of Computer Applications*, *140*(3), 6–14. https://doi.org/10.5120/ijca2016909249