

**Predicting Customer Churn in SafeBoda Ride-Hailing Application Using Machine Learning Techniques**

**By**

**Norman Angel Agong**

**147513**

A Research Proposal Submitted in Partial Fulfilment of the Requirements for Completion of Master of Science in Data Science & Analytics (MSc. DSA)

**iLab Africa**

**Institute of Mathematical Sciences (SIMS)**

**Strathmore University Nairobi, Kenya**

**Jan 2024**

# **Declaration and Approval**

## **Declaration**

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the proposal contains no material previously published or written by another person except where due reference is made in the proposal itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University

Student’s Name: **Norman Angel Agong**

**Sign.** A close up of a paper

Description automatically generated **Date: January 22, 2024**

## **Approval**

The proposal was reviewed and approved for defence by the

**Dr. Henry Muchiri**

School of Computing & Engineering Sciences

Strathmore University

Sign: .......................................................... Date: ..........................................................

# **ABSTRACT**

In the rapidly evolving landscape of the ride-hailing industry, characterized by the emergence of new players offering specialized services, competition is fierce where new firms can capture market share by providing specific services at lower prices, ride-hailing providers face a similar dynamic. Incumbent providers may retain their market position by pricing their services higher, leading to fluctuations in rates based on the level of competition.

As customer satisfaction and retention are paramount, the ability to predict and manage customer churn becomes crucial. For ride-hailing companies, this involves understanding and addressing customer attrition caused by factors like changing preferences, inadequate customer relationship strategies, or relocations. Accurate churn prediction enables organizations to identify high-risk customers and tailor services to reduce attrition.

This study addresses the critical need for early client churn prediction and recommends an optimal machine-learning strategy for this purpose in ride hailing industry. Leveraging customer historical with the goal to predict and understand existing customers' responses to implement effective retention strategies.

Various classification algorithms including Logistic Regression, K-Nearest Neighbours, Gradient Boosting Machine, Decision Tree, Random Forest, Support Vector Machine shall be used to predict customer churn. By examining these algorithms comprehensively, this study aims to provide insights into the most effective approach for early client churn prediction. The evaluation shall consider diverse perspectives to determine the algorithm that offers optimal results. This research contributes valuable recommendations for businesses seeking to enhance their client retention initiatives through advanced predictive analytics.

**Keywords**: Customer churn, ride-hailing service, machine learning, customer retention

Table of Contents

[Declaration and Approval 2](#_Toc156992679)

[Declaration 2](#_Toc156992680)

[Approval 2](#_Toc156992681)

[ABSTRACT 3](#_Toc156992682)

[Table of Figures 5](#_Toc156992683)

[CHAPTER 1: INTRODUCTION 6](#_Toc156992684)

[1.1 Background to the study 6](#_Toc156992685)

[1.2 Problem Statement 6](#_Toc156992686)

[1.3 Research Aim 7](#_Toc156992687)

[1.4 Research Objectives 7](#_Toc156992688)

[1.5 Research Questions 8](#_Toc156992689)

[1.6 Significant of the Study 8](#_Toc156992690)

[1.5 Scope of the Study 8](#_Toc156992691)

[CHAPTER 2: LITERATURE REVIEW 9](#_Toc156992692)

[2.1 Theoretical Framework 9](#_Toc156992693)

[**2.1.1 Customer Churn in Service Industries** 9](#_Toc156992694)

[**2.1.2 Technology Adoption Theory** 9](#_Toc156992695)

[**2.2 .3 Relationship Management (CRM)** 10](#_Toc156992696)

[2.2 Empirical Studies 10](#_Toc156992697)

[**2.2.2 Predictive Modelling in Churn Analysis** 10](#_Toc156992698)

[**2.2.3 Customer Behaviour in Ride-Hailing** 11](#_Toc156992699)

[**2.2.4 Machine Learning Applications in Churn Prediction** 11](#_Toc156992700)

[2.3 Integration of Theoretical and Empirical Insights 11](#_Toc156992701)

[2.4 Ride Churn in the Ride-Hailing Industry 11](#_Toc156992702)

[2.5 Industry Relevance and Contribution 12](#_Toc156992703)

[**2.5.1 Gaps in Existing Literature** 12](#_Toc156992704)

[**2.5.2 Conceptual Framework** 12](#_Toc156992705)

[CHAPTER 3. METHODOLOGY 14](#_Toc156992706)

[3.1. Research Design 14](#_Toc156992707)

[3.2 Data Sources and Collection 15](#_Toc156992708)

[3.3. Data Pre-processing 16](#_Toc156992709)

[**3.3.1 Data Cleaning** 16](#_Toc156992710)

[**3.3.2 Data Exploration** 16](#_Toc156992711)

[**3.3.3 Treating Missing Data** 16](#_Toc156992712)

[**3.3.4 Treating Outliers** 17](#_Toc156992713)

[**3.3.5 Data Type Conversion** 17](#_Toc156992714)

[3.4. Data Transformation 17](#_Toc156992715)

[**3.4.1 Feature Engineering** 17](#_Toc156992716)

[**3.4.2. Feature Scaling** 17](#_Toc156992717)

[**3.4.3. Feature Selection** 17](#_Toc156992718)

[**3.4.4. Data Encoding** 18](#_Toc156992719)

[3.6 Model Selection 18](#_Toc156992720)

[3.7 Model Evaluation 20](#_Toc156992721)

[3.7 Model Deployment 20](#_Toc156992722)

[REFERENCES 22](#_Toc156992723)

# **Table of Figures**

[Figure 1 The Cross Industry Standard Process for Data Mining (CRISP-DM) 14](#_Toc156990039)

# **CHAPTER 1: INTRODUCTION**

## **Background to the study**

1. In the rapidly evolving landscape of the ride-hailing industry, where the acquisition of new customers incurs higher costs compared to retaining existing ones, the importance of customer retention cannot be overstated. Recognizing that long-term customers tend to yield higher profits, Verbeke et al. (2011) underline the critical role of customer retention in increasing overall profitability. This perspective is particularly relevant in the context of the ride-hailing sector, where maintaining a loyal customer base is paramount for sustained success.
2. Today’s competitive business environment where customer churn is a fundamental concern, Tsai, and Chen (2010) argue for the necessity of effective churn management as a key survival strategy. Research by Nie et al. (2011) reveals that improvements in retention rates can lead to a substantial increase in profits, potentially up to 85% with just a 5% enhancement in retention. The growing emphasis on customer retention, as opposed to acquisition, is indicative of its pivotal role in ensuring the long-term viability of ride-hailing services.
3. Considering these industry dynamics, this study aims to discern common characteristics of churned customers within the ride-hailing sector. By identifying patterns and trends associated with customer churn, and endeavour to construct a robust customer churn prediction model. This predictive analytics approach aligns with the strategic goal of ride-hailing organizations to not only enhance customer retention but also optimize operational efficiency and maximize profitability in an increasingly competitive environment.
4. The subsequent sections of this research will expound on the problem statement, research objectives, and research questions, delving into the intricacies of ride churn prediction within the unique backdrop of SafeBoda in Kampala, Uganda. The literature review will further explore relevant publications, while the methodology will shed light on the chosen approach to address this vital research endeavour.

## **1.2 Problem Statement**

Ride-hailing services have revolutionized urban mobility, providing convenient alternatives to traditional transportation methods (Li et al., 2020; Shaheen et al., 2016). In Kampala, SafeBoda, a leading ride-hailing company, has played a transformative role in reshaping transportation dynamics. Despite its commitment to safe and reliable services, SafeBoda encounters a critical challenge—ride churn, where users discontinue their engagement with the platform (Tran & Phung, 2022).

To sustain growth and user satisfaction, SafeBoda must address ride churn, making it imperative to identify influencing factors and predict potential churn before it occurs. This necessitates a data-driven approach to understand user behaviour patterns and implement proactive strategies. As SafeBoda aims to enhance user experiences and maintain competitiveness, a comprehensive understanding of ride churn patterns becomes crucial.

This study focuses on the application of machine learning to predict ride churn tailored to SafeBoda's operations in Kampala. The subsequent sections will elaborate on the problem statement, research objectives, methodologies, and anticipated outcomes.

## **1.3 Research Aim**

The primary objective of this research is to delve into and explore diverse machine learning techniques applied to customer analytics, with a specific emphasis on predicting customer churn within the ride-hailing industry. The foundational step involves an extensive literature review to gain a profound understanding of the existing landscape of machine learning techniques in customer analytics.

The proposed models will be designed to analyse historical customer data, aiming to identify patterns and indicators indicative of potential churn. Additionally, the research endeavours to deploy these machine learning models using Application Programming Interfaces (APIs), facilitating real-time integration and accessibility. This deployment strategy aims to streamline and automate customer churn prediction processes. By seamlessly incorporating the models into existing systems through APIs, businesses can harness generated insights for prompt decision-making.

Conclusively, the overarching goal of this research is to elevate customer analytics in the ride-hailing industry through the strategic application of machine learning techniques. The study aims to empower businesses to proactively identify potential churners and implement targeted strategies for enhanced customer retention.

## **1.4 Research Objectives**

1. To review the existing literature on factors influencing customer churn
2. To identify the key factors that influence the success of the machine learning model in predicting churn in ride-hailing service.
3. To develop a machine learning model that can effectively identify and predict accurately whether a user is likely to churn.
4. To evaluate the performance of the machine learning model in predicting ride churn in ride-hailing services

## **1.5 Research Questions**

1. What are the key factors identified in existing literature that influence customer churn in the context of ride-hailing services?
2. What critical factors significantly impact the success of machine learning models in predicting churn within ride-hailing services?
3. What features and parameters contribute most to the accuracy of the churn prediction model?
4. How well does the developed machine learning model perform in accurately predicting ride churn in ride-hailing services?

## **1.6 Significant of the Study**

This research is significant as its objective is to enhance businesses' comprehension of machine learning applications in customer analytics, specifically focusing on customer churn predictions and customer segmentation. By delivering actionable insights, the study aims to refine customer retention strategies and enhance overall business performance.

## **1.5 Scope of the Study**

The primary emphasis of this study is to explore machine learning methods for predicting customer churn in the ride-hailing industry. The research encompasses an in-depth literature review, the creation and assessment of machine learning models, the deployment of these models using APIs, and the presentation of insights and recommendations.

# **CHAPTER 2: LITERATURE REVIEW**

In recent years, the ride-hailing industry has witnessed exponential growth, transforming urban transportation and consumer mobility patterns. Ride-hailing services such as SafeBoda, provide users with convenient, on-demand transportation options using mobile applications. However, alongside the industry's rapid expansion, the challenge of ride churn has emerged as a significant concern for service providers seeking to maintain user engagement and sustain growth. This chapter provides theoretical review and empirical review of customer churn.

## **2.1 Theoretical Framework**

Customer churn According to Sharma and Panigrahi (2011), churning refers to a customer who leaves one company to go to another company. Customer churn introduces not only some loss in income but also other negative effects on the operation of companies (Chen et al. 2014). As Hadden et al. (2005) stipulated, “Churn management is the concept of identifying those customers who are intending to move their custom to a competing service provider.” Risselada et al. (2010) stated that churn management is becoming part of customer relationship management. It is important for companies to consider it as they try to establish long-term relationships with customers and maximize the value of their customer base.

### **2.1.1 Customer Churn in Service Industries**

In the digital era, where customer behaviours are evolving rapidly, understanding the factors contributing to churn is paramount for businesses. Smith et al (2021) employs a mixed-methods approach, combining quantitative data analysis with qualitative insights from customer feedback and social media sentiments. Findings reveal nuanced aspects of customer loyalty and churn, shedding light on the impact of digital interfaces, personalized experiences, and shifting consumer expectations. The study provides actionable recommendations for service-oriented businesses aiming to navigate customer churn challenges in the current digital landscape.

### **2.1.2 Technology Adoption Theory**

Rogers' Diffusion of Innovations theory (2003) provides a foundational framework for understanding the adoption of innovations, particularly in the context of ride-hailing. This theoretical perspective illuminates the dynamics of how users incorporate new technologies, like ride-hailing services, and elucidates their influence on the occurrence of customer churn. Our research leverages Rogers' theory to gain insights into the adoption patterns of ride-hailing services, aiming to discern its correlation with customer churn in this dynamic technological landscape.

### **2.2 .3 Relationship Management (CRM)**

Building on the foundational work of Payne and Frow (2005) in Customer Relationship Management (CRM) theories, this research underscores the critical role of CRM principles in managing customer relationships within the ride-hailing industry. Payne and Frow's comprehensive exploration of CRM emphasizes the enhancement of customer loyalty through effective relationship management. By applying these CRM principles to the unique dynamics of ride-hailing services, this study aims to unravel valuable insights into customer interactions, satisfaction, and the intricate factors influencing churn. The integration of CRM theories within the ride-hailing context offers a theoretical framework to inform practical strategies for customer retention.

## **2.2 Empirical Studies**

2**.2.1 Factors Influencing Customer Churn**

The empirical study conducted by Tuzovic and Kabadayi in 2018 offers valuable insights into the factors influencing customer churn within the realm of mobile apps. Their rigorous examination of these factors provides a foundation for understanding the dynamics of customer churn in technologically driven service industries. In the context of our research on predicting customer churn in the ride-hailing industry, the findings of Tuzovic and Kabadayi serve as a crucial reference point. By extrapolating their empirical evidence, we aim to elucidate the applicability of similar factors and mechanisms that drive customer churn in the specific context of ride-hailing applications.

### **2.2.2 Predictive Modelling in Churn Analysis**

Nguyen et al.'s (2015) comprehensive study on predictive modelling in churn analysis serves as a pivotal reference for our research on predicting customer churn in the ride-hailing industry. Their empirical work delves into the application of machine learning algorithms, providing valuable insights into the intricacies of churn prediction. By leveraging the methodologies and findings presented by Nguyen et al., our study aims to adapt and apply these predictive modelling techniques to the specific context of ride-hailing platforms. This reference provides a robust foundation for integrating machine learning algorithms into the exploration of customer churn within the ride-hailing landscape.

### **2.2.3 Customer Behaviour in Ride-Hailing**

Xu et al.'s (2018) empirical study on customer behaviour in the ride-hailing industry provides valuable insights into the dynamics influencing customer churn. By exploring user preferences and their impact on continued engagement, this study offers a comprehensive understanding of the factors contributing to customer attrition. Our research draws inspiration from Xu et al.'s work to investigate similar dynamics within the context of ride-hailing platforms. The study emphasizes the significance of user preferences in shaping customer behaviour, offering a foundation for our exploration of the ride-hailing industry's customer churn patterns.

### **2.2.4 Machine Learning Applications in Churn Prediction**

Verbeke et al.'s (2013) empirical study explores the application of machine learning for customer churn prediction, offering valuable insights into its practical implementation. Within the ride-hailing sector, where customer churn is a critical concern, this work serves as a guide for leveraging machine learning models. Our research draws inspiration from Verbeke et al.'s expertise, aiming to apply similar methodologies to predict customer churn in the ride-hailing industry. By incorporating their findings, we intend to enhance the understanding and implementation of machine learning techniques for effective churn prediction.

## **2.3 Integration of Theoretical and Empirical Insights**

The theoretical framework emphasizes the economic implications of customer loyalty, the diffusion of innovations, and CRM principles. Integrating these theories with empirical findings on factors influencing churn, predictive modelling, and customer behaviour in the ride-hailing context forms a robust foundation for the application of machine learning to predict customer churn.

## **2.4 Ride Churn in the Ride-Hailing Industry**

Ride churn, commonly referred to as user churn or attrition, pertains to the phenomenon where users cease utilizing ride-hailing services after an initial period of engagement. Ride churn poses a substantial threat to the long-term viability of ride-hailing platforms, as it can lead to reduced revenue, decreased user satisfaction, and hindered market expansion. A study by Zhao et al. (2019) emphasized that understanding ride churn dynamics is crucial for improving customer retention and optimizing service offerings in the ride-hailing sector.

## **2.5 Industry Relevance and Contribution**

This literature review underlines the significance of predicting ride churn for the ride-hailing industry's sustainability and growth. While various models have been proposed, the application of data science techniques to churn prediction in ride-hailing applications remains relatively unexplored. This study aims to bridge this gap by offering insights into user behaviour and retention strategies unique to SafeBoda's context.

### **2.5.1 Gaps in Existing Literature**

Limited Application of Data Science Techniques to predict customer churn in ride-hailing applications. While there is a substantial body of literature on customer churn prediction, there is a noticeable gap in the application of data science techniques specifically to in the ride-hailing industry. Existing studies might lack insights into the unique user behaviour patterns and retention challenges that ride-hailing applications like faces.

Contextual Understanding of ride-hailing Operations: Previous literature may not adequately address the distinct operational context of SafeBoda. The company operates in Kampala, Uganda, and its commitment to safe and reliable transportation services may introduce unique factors influencing ride churn that have not been comprehensively explored in the literature.

Lack of Focus on User Retention Strategies: Some studies may primarily concentrate on predicting ride churn without delving into effective user retention strategies. Understanding how SafeBoda can employ proactive measures to retain users is crucial for the company's sustained growth, but this aspect might be underrepresented in the existing literature.

### **2.5.2 Conceptual Framework**

The conceptual framework for this study revolves around the following key components.

User Behaviour Analysis: The framework will involve a detailed analysis of user behaviour patterns within the SafeBoda platform. This includes examining factors leading to ride churn, such as pricing, service quality, and user satisfaction.

Machine Learning Models for Predicting Churn: The study will employ various machine learning models, including Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Trees, Support Vector Machine, Naive Bayes, and Gradient Boosting, to predict ride churn. Each model will be assessed for its effectiveness in the SafeBoda context.

User Retention Strategies: Strategies to enhance user retention will be an integral part of the conceptual framework. This involves understanding how insights from user behavior analysis and churn prediction models can inform the development of targeted retention initiatives.

Performance Evaluation Metrics: To determine the success of the machine learning models, the framework will incorporate performance evaluation metrics such as accuracy, precision, recall, and F1 score.

Contextual Considerations: Recognizing the contextual factors unique to SafeBoda, such as the cultural and economic landscape of Kampala, will be essential. This ensures that the findings and recommendations are relevant and applicable to the company's specific operational environment.

# **CHAPTER 3. METHODOLOGY**

This section outlines the methodologies employed to address the research objectives and questions presented in Chapter One.

## **3.1. Research Design**

The research will adopt the well-established Cross Industry Standard Process for Data Mining (CRISP-DM) model, a structured methodology for executing data science projects. This model facilitates a systematic approach to various data science tasks, ensuring a thorough understanding of the underlying data encompassing six sequential phases:

Business Understanding is the first phase with the focus of comprehending the project's objectives and requirements with the task of determining business objectives, assessing the situation, defining project goals and plan hence establishing a robust understanding of business needs is crucial for laying the foundation for successful projects. The second phase is Data Understanding which concentrates on identifying, collecting, and analysing datasets essential for achieving project goals. Tasks include collecting initial data, describing data, exploring data, and verifying data quality.

Data Preparation is the third phase involves readying the final dataset(s) for modelling. With tasks including selecting data, cleaning data, constructing data, integrating data, and formatting data. This will be followed by the Modelling phase which involves building and assessing various models based on diverse modelling techniques. The study will encourage Iteration in model building and assessment until a good enough model is achieved.

Evaluation phase takes a broader perspective, evaluating which model aligns best with business objectives. Tasks involve evaluating results, reviewing the process, and determining next steps. This will be followed by the last and final phase of Deployment to ensure that the model is accessible to users. Tasks will include planning deployment, planning monitoring and maintenance, producing a final report, and reviewing the project. The Phases of CRISP-DM are highlighted in the figure below.

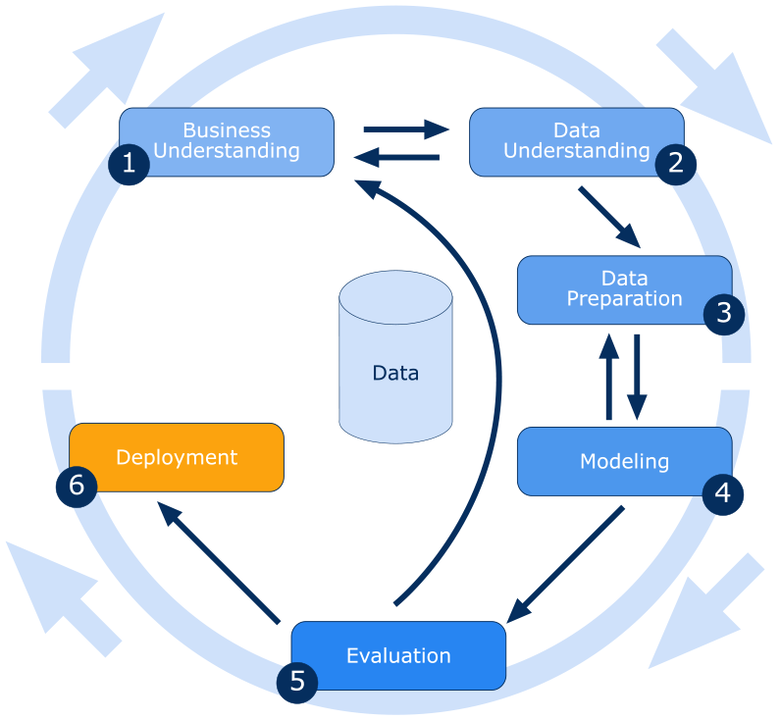


Figure 1 The Cross Industry Standard Process for Data Mining (CRISP-DM)

## **3.2 Data Sources and Collection**

The study will use historical customer data extracted from SafeBoda database which is hosted in the cloud on Amazon Relational Database Service (RDS). This historical customer data where each row represents one customer offers straightforward insights. This dataset comprises of customer’s ride history, demographic information, spending among others and will provide key indicators of customer churn allowing us to anticipate the behaviours that contribute to customer retention and predict what behaviour will help us to retain customers.

To select relevant information from different database tables, Structured Query Language (SQL) queries and Python modules and packages will be used during the data extraction process. The information relevant to the research will be contained in the queried views and tables. Ledger is a database table that contains all customer transactions and revenue brought in by each customer while the rides table contains the rides information taken by customers. The Isabel database will be queried to analyse customers interactions. The research will aim to establish a solid foundation for subsequent data analysis by consolidating these datasets. This would enable a more profound understanding of different business aspects, such as customer behaviour, product analysis, and revenue generation.

## **3.3. Data Pre-processing**

Data pre-processing will be a crucial phase in the data analysis. This step will focus on cleaning and transforming raw data into a format suitable for further analysis and machine learning modelling.

### **3.3.1 Data Cleaning**

Data cleaning is a fundamental step in the data pre-processing phase, aiming to enhance the quality and reliability of the dataset. It involves the identification and rectification of errors, inconsistencies, and inaccuracies within the data. Errors may manifest as typos, inaccurately recorded values, or inconsistencies in formatting. Through a systematic approach, data cleaning activities include imputing missing values, correcting errors, and ensuring uniformity in data representation.

### **3.3.2 Data Exploration**

Through data exploration, researcher will gain valuable insights into the distribution and structure of the data, informing decisions about appropriate analytical approaches. This phase involves descriptive statistical analyses, visualizations, and summary statistics to unveil patterns, trends, and potential relationships within the data. Exploratory Data Analysis (EDA) techniques, such as histograms, scatter plots, and correlation matrices, contribute to uncovering insights that guide subsequent analysis strategies.

### **3.3.3 Treating Missing Data**

Addressing missing data is essential for ensuring the completeness of the dataset. The treatment of missing data will involve strategic decisions on whether to impute or remove incomplete entries. Imputation methods may include mean imputation, regression imputation, or leveraging advanced machine learning techniques. The chosen approach depends on the nature of the missing data and the potential impact on subsequent analyses.

### **3.3.4 Treating Outliers**

Outliers can distort analysis results and compromise model performance. Treating outliers is a crucial step in data pre-processing to mitigate their influence. Identification methods, such as statistical measures or visualization tools, pinpoint data points exhibiting extreme behaviour.

### **3.3.5 Data Type Conversion**

Data type conversion is an essential to standardizing the format of variables for consistent analysis. Converting data types will uniformity in the representation of variables, facilitating seamless integration into analysis tools or models.

## **3.4. Data Transformation**

### **3.4.1 Feature Engineering**

Raw dataset will be refined to enhance the performance of machine learning models. This process will involve the creation of new features or modification of existing ones to extract meaningful information and patterns. Different techniques will contribute to the generation of features that better capture the underlying complexities of the data.

### **3.4.2. Feature Scaling**

This process will involve standardizing or normalizing the range of numerical features to ensure equitable influence on model training. Standardization methods, such as z-score normalization, and techniques like Min-Max scaling, bring features within a comparable scale, preventing dominant features from unduly influencing model outcomes.

### **3.4.3. Feature Selection**

This step will involve a careful curation of relevant features to improve model performance and mitigate issues related to the curse of dimensionality. This step aims to identify and retain the most informative features while discarding redundant or irrelevant ones. Effective feature selection will not only enhance model interpretability but also accelerates the training process by reducing computational complexity.

### **3.4.4. Data Encoding**

This will address the challenge of incorporating categorical variables into machine learning models that typically require numerical inputs. This process will involve converting categorical data into a numerical format, allowing algorithms to interpret and utilize this information effectively. Common encoding methods include one-hot encoding, label encoding, and binary encoding, each offering distinct advantages based on the nature of the categorical variables.

## **3.6 Model Selection**

Addressing customer churn involves a classification problem, aiming to categorize instances into predefined classes based on their features. In this paper, seven widely used classifiers will be employed, and their performance is compared. Subsequently, the most suitable model is selected for predicting customer churn.

The research will consider the following classifiers. Logistic Regression (LR) will be applied as a linear model for binary classification, estimating the probability of an instance belonging to a particular class. The model will utilize the logistic function to convert the linear combination of features into probabilities, subsequently to produce the final classification. LR is acknowledged for its simplicity, speed, and ease of interpretation.

The K-Nearest Neighbour's algorithm will be explored as a non-parametric, instance-based classifier assigning an instance to the class most common among its k nearest neighbours. KNN is recognized for its simplicity, robustness to noisy data, and effectiveness in problems with well-separated classes, though challenges such as high computational complexity, sensitivity to the choice of k, and poor performance in high-dimensional spaces may be encountered.

For the Random Forest (RF) ensemble learning method, a collection of decision trees will be constructed, and their predictions will be combined via majority voting to enhance overall performance and reduce overfitting. RF is chosen for its ability to handle large datasets, robustness to noise, and natural measure of feature importance. However, it is noted that RF models can be computationally expensive and may lack interpretability compared to single decision trees.

Decision Tree (DT) will be considered as a tree-based algorithm that recursively splits the feature space to create a hierarchical structure. Each node will represent a decision rule based on a single feature, and each leaf node will correspond to a class label. DTs are valued for their interpretability, robustness to noise, and capability to handle both continuous and categorical features.

Support Vector Machine (SVM) SVM is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points belonging to different classes. SVM is effective in high-dimensional spaces and is versatile for various types of datasets.

Naive Bayes (NB) is a probabilistic classification algorithm grounded in Bayes' theorem. It operates under the assumption of feature independence, implying that the presence or absence of a particular feature does not influence the presence or absence of another feature. While this assumption might not hold true in all cases, it greatly simplifies the calculation of probabilities. Naive Bayes is particularly well-suited for high-dimensional datasets where the simplicity and efficiency of the algorithm come into play. Despite its naive assumption, Naive Bayes often performs remarkably well, especially in text classification and spam filtering.

Gradient Boosting (GB) is an ensemble learning method that constructs a series of weak learners, typically decision trees, sequentially. Each tree in the series corrects errors made by the previous one, with a focus on instances that were misclassified. It builds a strong predictive model by combining the strengths of multiple weak models. Gradient Boosting is renowned for its ability to deliver high predictive accuracy, making it a popular choice in various machine learning competitions. However, it demands careful tuning of hyperparameters to achieve optimal performance and prevent overfitting. The sequential nature of the learning process allows the algorithm to continuously improve and adapt to complex patterns in the data.

**Why the seven classifiers above**

The selection of these seven classifiers for predicting customer churn involves a thoughtful consideration of their individual strengths, suitability for the task, and trade-offs in various aspects. Here are the reasons for narrowing down to the above classifiers.

Logistic Regression (LR) is chosen for its simplicity, speed, and ease of interpretation. As a linear model for binary classification, it estimates the probability of an instance belonging to a particular class.

K-Nearest Neighbors (KNN) is included for its simplicity, robustness to noisy data, and effectiveness in problems with well-separated classes. Despite challenges like high computational complexity and sensitivity to the choice of k, KNN's suitability in certain scenarios justifies its inclusion.

Random Forest (RF) an ensemble learning method, is selected for its ability to handle large datasets, robustness to noise, and natural measure of feature importance. While computationally expensive, its performance benefits, especially in reducing overfitting, make it a valuable candidate.

Decision Tree (DT) is chosen for its interpretability, robustness to noise, and capability to handle both continuous and categorical features. The hierarchical structure created by recursively splitting the feature space aligns well with the interpretability requirement.

Support Vector Machine (SVM) SVM is employed for its effectiveness in high-dimensional spaces and versatility for various types of datasets. Its ability to find the optimal hyperplane for separating different classes in the feature space is crucial for addressing the classification problem.

Naive Bayes (NB) a probabilistic classification algorithm, is selected due to its simplicity and efficiency grounded in Bayes' theorem. Its assumption of feature independence makes it well-suited for high-dimensional datasets where simplicity and efficiency are advantageous.

Gradient Boosting (GB) is included as an ensemble learning method that delivers high predictive accuracy. Its sequential construction of weak learners, typically decision trees, allows it to adapt to complex patterns in the data. While demanding careful tuning, GB's performance benefits justify its inclusion.

## **3.7 Model Evaluation**

The developed predictive model will be evaluated using established metrics, Precision & Recall, such as Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), and F1 Score. These metrics gauge the model's ability to differentiate between churn and non-churn instances. The model's performance will be assessed to ensure it aligns with the research objectives.

## **3.7 Model Deployment**

The research will deploy the model using an API. The decision to deploy the machine learning models through an API stem from various considerations, including the pursuit of a versatile, modular, and easily maintainable solution that seamlessly integrates with diverse client applications. Opting for an API deployment allows the implementation of the models to be separated from that of the client, facilitating independent updates and improvements for both components. This approach not only enhances maintainability but also supports scalability, as the API can be efficiently designed to handle multiple simultaneous requests.

For the implementation of the API, FastAPI will be selected as the web framework. This choice is driven by its emphasis on speed, user-friendliness, and robustness. FastAPI provides automatic data validation, serialization, and documentation features, along with an asynchronous programming model to enhance overall performance. These attributes position FastAPI as an ideal selection for deploying the machine learning models in a production-ready, scalable, and maintainable manner.

# **REFERENCES**

1. Tran, T. & Phung, Thai. (2022). Rider Churn Prediction Model for Ride-Hailing Service: A Machine Learning Approach. 10.1007/978-3-030-97610-1\_24.
2. Zhao, Y., Liu, X., & Ruan, J. (2019). Customer churn prediction in ride-hailing platforms: Insights from a multi-wave panel dataset. Transportation Research Part C: Emerging Technologies, 99, 35-52.
3. Sharma A, Panigrahi PK (2011) A neural network-based approach for predicting customer churn in cellular network services. Int J Comput Appl 27(11):26–31
4. Chen K, Hu Y-H, Hsieh Y-C (2014) Predicting customer churn from valuable B2B customers in the logistics industry: a case study. IseB 13:475–494. doi:10.1007/s10257-014-0264-1
5. Hadden J, Tiwaria A, Roy R, Ruta D (2005) Computer assisted customer churn management: State-of-the-art and future trends. Comput Oper Res 34:2902–2917
6. Risselada H, Verhoef PC, Bijmolt THA (2010) Staying power of churn prediction models. J Interact Mark 24:198–208
7. Rogers, E. M. (2003). Diffusion of Innovations (5th ed.). Free Press.Payne, A. F., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. Journal of Marketing, 69(4), 167–176
8. Tuzovic, S., & Kabadayi, S. (2018). What Types of Mobile Apps Are Shown to Be Successful? A Data-Driven Analysis of the Characteristics of Top-Ranked Apps. Journal of Business Research, 82, 131–140.
9. Nguyen, T., Hoang, D., Pathirana, P. N., & Nguyen, T. T. (2015). A Data Mining Framework for Predicting Customer Churn in Online Businesses. Expert Systems with Applications, 42(4), 1778–1788.
10. Xu, Y., Yu, Z., & Wang, Y. (2018). What Influenced User Satisfaction and Continued Use of Mobile Apps? Information Development, 34(3), 270–284.
11. Verbeke, W., Dejaeger, K., Martens, D., & Baesens, B. (2013). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. European Journal of Operational Research, 218(1), 211–229.