**Advanced Business Analytics**

Contents

[Introduction 3](#_Toc205797576)

[Problem Statement 3](#_Toc205797577)

[Research Objective 4](#_Toc205797578)

[Research Questions 4](#_Toc205797579)

[1.0 Dataset Domain Knowledge Understanding 5](#_Toc205797580)

[1.1 Understanding Customer Churn in the Telcom Industry 7](#_Toc205797581)

[1.2 Purpose of Dataset 7](#_Toc205797582)

[1.4 Strategic Business Value & Use Cases 8](#_Toc205797583)

[2.0 Understanding of Data Analysis 10](#_Toc205797584)

[3.0 Visualization Using Any Tool 13](#_Toc205797585)

[4.0 Research Methodology 15](#_Toc205797586)

[4.1 Business Understanding 15](#_Toc205797587)

[4.2 Data Understanding 15](#_Toc205797588)

[4.3 Data Preparation 17](#_Toc205797589)

[Tools and Environment 18](#_Toc205797590)

[4.4 Evaluation 20](#_Toc205797591)

[5.0 Methods of Modelling 21](#_Toc205797592)

[6.0 Literature Review and References 25](#_Toc205797593)

[7.0 Comparison of Model 34](#_Toc205797594)

[8.0 Discussion of New Findings 38](#_Toc205797595)

[References and Annotated Bibliography 42](#_Toc205797596)

**Telecommunication Churn Prediction and the Rising Issue of Customer Retention with Business Analytics**

**Keywords**: data mining, customer churn prediction, telecommunication, business analytics, predictive analysis, customer segmentation, targeted marketing

# Introduction

Customer retention is a significant problem that exists within the telecommunication industry driving the overall profits of the company, therefore it is imperative for companies to truly utilize and understand the data behind the increasingly challenging issue of retaining valuable customers (Matellio, 2025). According to a study done by Statista (2023) during the years 2020-2023 telecommunication companies have spent an estimated total between 15-20% of their service revenue to acquire and retain customer amidst growing competition within a saturated market, in leveraging machine learning models we hope to reduce and optimize revenue spent in targeted marketing for companies to maintain a competitive advantage in the market through minimizing churn and enhancing customer loyalty.

Customer churn, also known as customer attrition, is defined as the propensity of customers to cease existing subscriptions or business with a company in a given time period (Zhang, Moro, & Ramos, 2022), which in turn reflects the end of a Business-to-Customer (B2C) relationship. Customer churn can occur for a variety of factors, including dissatisfaction with service quality, pricing, customer service, and better alternatives offered by competitors. Within a B2C a high churn rate tends to indicate a company’s financial stability, market standing, and overall performance leading to brand reputation loss and growth prospects (O’Brien & Downie, 2025).

# Problem Statement

The telecommunication industry is severely impacted by a persistent issue of high customer churn, which had significantly deteriorated the company profitability, compromises market standing, and hindered their long-term growth. Amidst the ever increasingly saturated and competitive market, companies are forced to allocate substantial portions towards customer acquisition with often inefficient retention strategies. This persistent customer churn problem, driven by problems such as service dissatisfaction, pricing concerns, inadequate customer service, and aggressive competitor offerings, has resulted in a continuous loop of financial drain and brand reputation loss. Due to this, there’s a critical need for advanced, data-driven approaches to identify customers that are at risk of churning and understand the reasons influencing their decision to leave. This can help to implement targeted, cost-effective customer retention strategies to minimize financial losses and maintain a competitive advantage.

## Research Objective

1. Can machine learning algorithms effectively predict customer churn based on customer feedback and behaviour data
2. Identify key factors that significantly influence customer churn.
3. Develop and compare multiple predictive models to evaluate their performance in forecasting customer churn

## Research Questions

1. Comparison of accuracy between decision tree model, linear regression, XGBoost and random Forest
2. Do customer attributes such as gender, senior citizen, partner and etc, affect the churn rate?
3. How do monthly charges and total charges relate to customer churn behavior?

# 1.0 Dataset Domain Knowledge Understanding

Dataset: Telecom Customer Churn

Dataset Name: Telecom Customer Churn

Domain: Telecommunications

Dataset Dimensions

Number of Rows/Instances: 100,000

Number of Columns: 100

|  |  |
| --- | --- |
| **Category** | **Column Names** |
| Customer Usage and Activity | mou\_Mean, da\_Mean, totmrc\_Mean, rev\_Mean, ovrmou\_Mean, ovrrev\_Mean, vceovr\_Mean, datovr\_Mean, roam\_Mean, change\_mou, change\_rev |
| Customer Call Quality/Network Experience | drop\_vce\_Mean, drop\_dat\_Mean, blck\_vce\_Mean, blck\_dat\_Mean, unan\_vce\_Mean, unan\_dat\_Mean, plcd\_vce\_Mean, plcd\_dat\_Mean, recv\_vce\_Mean, recv\_sms\_Mean, comp\_vce\_Mean, comp\_dat\_Mean, callfwdv\_Mean, callwait\_Mean |
| Customer Billing and Revenue | totrev, adjrev, avgrev, totmrc\_Mean, cc\_mou\_Mean, ccrndmou\_Mean |
| Customer Care Interaction | custcare\_Mean, cc\_mou\_Mean, ccrndmou\_Mean, mou\_cvce\_Mean, mou\_cdat\_Mean |
| Customer Behavorial Features | threeway\_Mean, inonemin\_Mean, mou\_rvce\_Mean, owylis\_vce\_Mean, iwylis\_vce\_Mean, peak\_vce\_Mean, peak\_dat\_Mean, mou\_peav\_Mean, mou\_pead\_Mean, opk\_vce\_Mean, opk\_dat\_Mean, mou\_opkv\_Mean, mou\_opkd\_Mean |
| Customer Subscription Information | months, uniqsubs, actvsubs, new\_cell, eqpdays, phones, models |
| Customer Device Plan | dualband, refurb\_new, hnd\_price, hnd\_webcap, crclscod, asl\_flag |
| Customer Information | area, truck, rv, ownrent, lor, marital, adults, income, numbcars, ethnic, prizm\_social\_one, HHstatin, dwllsize, dwlltype, kid0\_2, kid3\_5, kid6\_10, kid11\_15, kid16\_17, creditcd, Customer\_ID, churn |

## 1.1 Understanding Customer Churn in the Telcom Industry

In the telecommunications industry, customer churn refers to the situation where a customer discontinues their service or switches to a competitor. This can happen due to a variety of factors such as poor service quality, high prices, better alternatives in the market, or inadequate customer support.

Telcom companies face fierce competition and operate in a saturated market with limited product differentiation. In such an environment, retaining existing customer is more cost-effective than acquiring new ones. Studies suggest that acquiring a new customer can cost 5 to 10 times more than retaining an existing one.

Churn is considered a critical Key Performance Indicator (KPI) because it:

* Revenue Impact: Directly affects monthly recurring revenue
* Market Position: Signals potential dissatisfaction or service issues
* Customer Lifetime Value: Every churned customer represents a loss of potential or future revenue, diminishing the overall CLV
* Operational Costs: Elevated churn necessitates higher spending on marketing and sales to acquire new customers, straining the budget

For telecom operators, predicting and reducing churn is crucial for maintaining profitability, market share, and long-term growth. This is why customer churn has become one of the top metrics monitored by analysts, executives, and investors alike.

## 1.2 Purpose of Dataset

This Telco Customer Churn dataset used in this project was sourced from Kaggle (Abhinav) and consists of 100,000 customer records and 100 variables.

The primary purpose of this dataset is to enable the development of machine learning models that can accurately predict whether a customer is likely to churn, based on usage patterns, billing details, service quality, and demographic factors.

In a Business Use Case, it is used to:

* Identify high-risk customers before they leave
* Optimize retention campaigns by focusing on the most predictive features
* Customize service offerings based on customer segments
* Improve product, pricing, and customer service strategies

By analyzing this dataset, telecom companies can adopt data-drive, cost effective, and proactive retention strategies to prevent or reduce churn rates and boost customer satisfaction.

Why is This dataset valuable for Business Analytics

* The telecom churn dataset is particularly useful because:  
  It combines usage, billing, service, and demographic data in one source
* It reflects real-world problems faced by telecom firms globally
* It supports classification modelling (Logistic Regression, Decision Tree, Random Forest), feature engineering, segmentation, and LPI analysis
* It allows companies to predict, understand, and respond to Churn behaviour

## 1.4 Strategic Business Value & Use Cases

The insights derived from this dataset are invaluable and directly translate into several strategic business applications:

* Proactive, Targeted Retention Campaigns: The most immediate use is to identify high-risk customers and intervene with targeted offers before they churn.
  + How the dataset helps: By using the churn column as a target variable, a model can be trained to recognize patterns in others features. For instance, a significant increase in drop\_vce\_Mean (dropped voice calls) or custcare\_Mean (customer care calls), or a sudden decrease in usage seen in change\_mou (change in minutes of use), can be powerful predictors of dissatisfactions. The model can assign a “churn score” to each customer, allowing retention teams to prioritize their efforts on those most likely to leave.
* Personalized Marketing and Customer Segmentation: Understanding that not all customers are the same is key to effective marketing. This dataset enables the creation of distinct customer segments for personalized communication.
  + How the dataset helps: Demographic data like marital, income, ethnic, and the presence of children (the kid\* columns) can be combined with behavioural data like mou\_Mean (mean minutes of use), ovrmou\_Mean (overage minutes), and handset price (hnd\_price). This allows for the creation of personas, such as ‘high-value, tech-savvy families’ or ‘budget-conscious individuals’, who
  + can then be targeted with relevant promotions and service bundles.

# 2.0 Understanding of Data Analysis

This section details the initial exploratory data analysis (EDA) conducted on the "Telco Customer Churn" dataset to gain preliminary insights, identify data characteristics, and detect potential quality issues. This is the foundational step to understand the data to proceed to the subsequent data preparation and modeling steps.

**2.1 Initial Data Inspection and Descriptive Statistics**

Upon loading the dataset, a general overview of its structure and content was looked up. This was achieved by:

* **Inspecting the first few rows:** Using methods like df.head() to get an idea of the data's layout and values.
* **Checking data types and non-null counts:** df.info() was used to confirm the data types of each column and identify any immediate missing values.
* **Generating descriptive statistics:** Using df.describe() to obtain summary statistics such as mean, median, standard deviation, minimum and maximum values for the numerical columns. This helped understand the distribution and range of these key numerical features.
* **Analyzing categorical feature distributions:** Frequenct counts for object-type columns were performed to understand the distribution of categories within each feature. This was to identify the class imbalance in the Churn target.

**2.2 Identification of Initial Data Inconsistencies and Challenges**

During this preliminary analysis, several data inconsistencies and potential challenges were noted, which were later addressed in the Data Preparation phase:

* **TotalCharges Data Type:** It was immediately apparent that the TotalCharges column was identified as an Object (string) type by df.info(), despite containing numerical values. This indicated the presence of non-numeric entries, such as empty strings, preventing direct numerical calculations.
* **Missing Values:** While df.info() provided an initial count, a more detailed check for missing values across all columns revealed that TotalCharges contained a small number of missing or empty entries, particularly for new customers.
* **Target Variable Imbalance:** The analysis of the Churn column's value distribution (e.g., df['Churn'].value\_counts()) clearly showed a significant imbalance, with a much larger proportion of customers identified as 'No Churn' compared to 'Churn'. This observation was critical as it directly impacts model training and evaluation strategies, necessitating techniques like SMOTE (as mentioned in your "1.0 Dataset Domain Knowledge Understanding" for the banking dataset, and which would apply here too) or adjusting evaluation metrics beyond simple accuracy.
* **Potential Outliers/Skewness:** While no severe outliers were initially evident, the descriptive statistics for MonthlyCharges and tenure indicated some skewness, suggesting that feature scaling might be beneficial for certain models.

**2.3 Preliminary Insights and Hypotheses**

Based on this initial data exploration, several preliminary insights and hypotheses were formed, which would be further investigated through visualization and modelling:

* **Usage Patterns and Churn:** It was hypothesized that significant changes in monthly usage or revenue (change\_mou, change\_rev) or consistently low/high usage (mou\_Mean, rev\_Mean) might be strong indicators of churn.
* **Customer Care Interactions:** Features like custcare\_Mean and cc\_mou\_Mean were expected to be positively correlated with churn, suggesting that customers frequently contacting customer service might be experiencing dissatisfaction.
* **Device and Subscription Longevity:** eqpdays (days since equipment purchase) and months (length of customer relationship) were anticipated to be inverse predictors of churn as newer equipment or longer tenure might indicate higher satisfaction and lower churn risk.
* **Billing and Revenue Factors:** While totmrc\_Mean (total monthly recurring charges) is present, the complex relationship between various revenue metrics (rev\_Mean, ovrrev\_Mean, datovr\_Mean, totrev, avgrev) and churn would need detailed exploration. It was initially hypothesized that unexpected increases in overage charges could drive churn.
* **Demographic and Lifestyle Factors:** Features like prizm\_social\_one, income, marital, and kid counts were considered for their potential influence on churn behaviour, often interacting with service usage or pricing perception.

This initial data analysis phase is important in confirming the dataset's suitability for churn prediction, understanding its inherent characteristic and identifying the necessary data cleaning and preprocessing steps required.

# 3.0 Visualization Using Any Tool

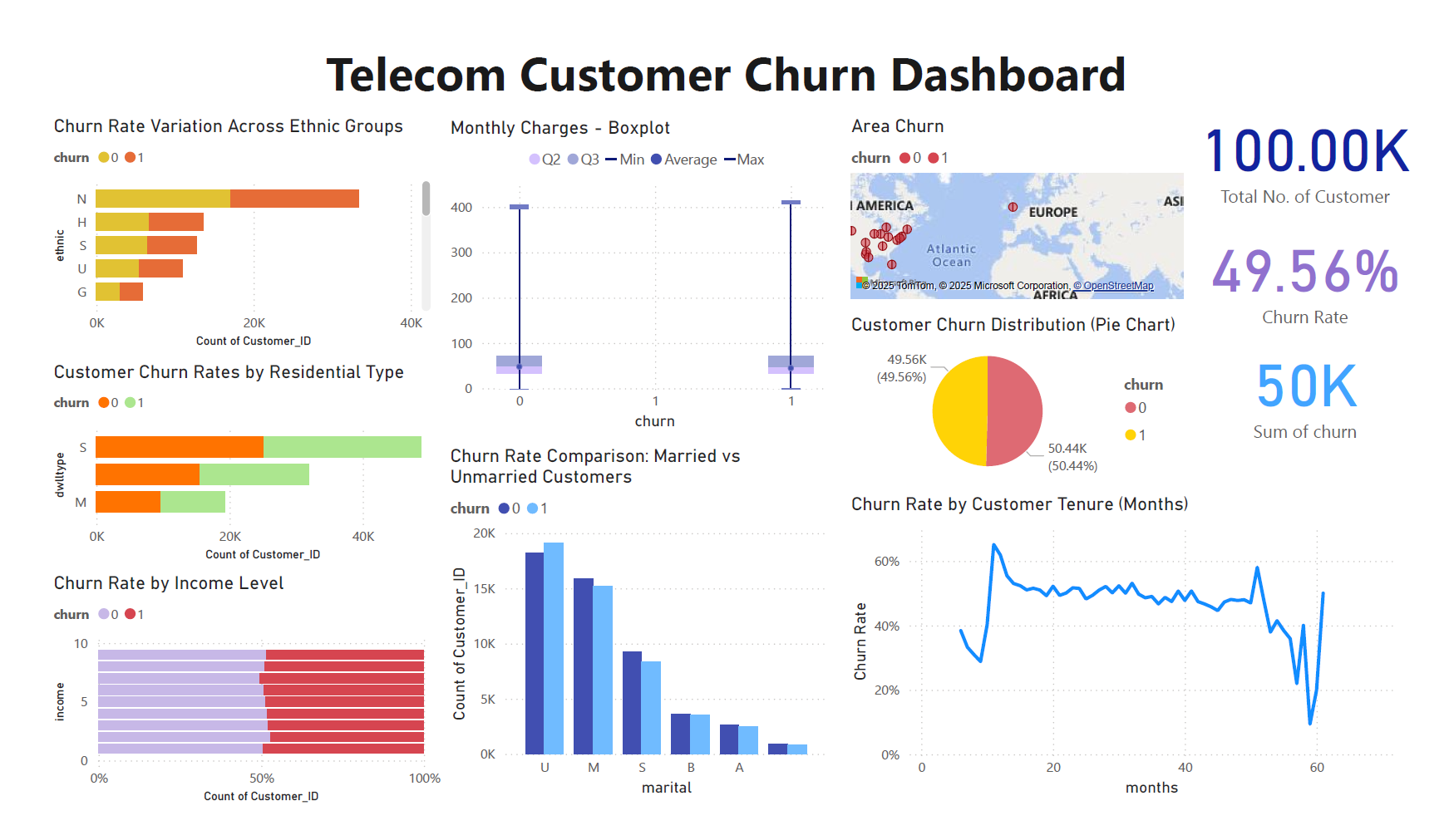


Figure 3.0.1: Telecom Customer Churn Dashboard

The dashboard above provides comprehensive customer churn patterns in the telecom dataset. It includes demographic, financial, and behavioural factors. Churn is concentrated mainly in the Americas and Europe, and the overall churn is 49.56% (nearly half of the customer base). This high rate shows that the strategies of customer retention used are highly inadequate currently.

The distribution of churn across ethnic groups shows a noticeable difference. The churn rate of the N and H groups is significantly higher than the average level. This may be related to differences in cultural preferences, communication needs, or service experiences among different groups in their residential areas.

Monthly charges are strongly linked to churn. Churned customers tend to have higher monthly payments, both in median and upper quartile values. High-paying customers may feel that the price does not match the perceived value, making them more vulnerable to competitor offers.

Income level is another important factor. Low-income customers churn more frequently. It is likely due to higher price sensitivity and a stronger tendency to seek cheaper plans. Customers in type S residences churn more than type M. This could relate to coverage differences or competitive alternatives in certain dwelling areas. Marital status is associated with churn as well. The customers who are not married churn more compared to married customers. This may represent more flexibility in switching services or a higher willingness to change providers for better offerings.

Tenure shows a non-linear relationship with churn. The risk of churning is high in the first 0-6 months. This indicates that the first impressions are important for the company. At 55-60 months, it rises again. This often matches with contract renewals. Customers may review their services or switch to another service provider.

Overall, multiple aspects affect customer churn. These trends offer valuable insights for understanding and strengthening customer retention in the telecommunications industry.

# 4.0 Research Methodology

This research follows the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. The methodology is widely adopted in data science projects. The process consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. But in this research, there will not be any deployment of any sort and will only focus on data analysis. This framework provides a structured approach to soling business problems using machine learning.

## 4.1 Business Understanding

The goal of the project is to predict customer churn in the telecommunications sector using machine learning techniques and models. When customer churn or when a customer discontinues the company’s service, it directly impacts the company revenue. Since customer retainment is easier than customer acquirement, it is crucial for telecommunication companies to detect early signs of potential churn by customers.

This research aims to:

1. Develop machine learning models that predict whether a customer will churn
2. Identify key factors that influence churn behaviour
3. Compare the performance of machine learning models chosen to determine the most effective one of the 3.

## 4.2 Data Understanding

The dataset that will be analysed is the Telecom Customer Churn Dataset. Source provided by Abhinav89 from Kaggle [18]. It includes 100000 observations and features. Below is a detailed categorized breakdown of the important features.

Customer Usage and Activity

* mou\_Mean: Average monthly minutes of use (voice).
* da\_Mean: Average number of data calls per month.
* totmrc\_Mean: Total monthly recurring charges.
* rev\_Mean: Average monthly revenue per customer.
* ovrmou\_Mean: Mean overage minutes used.
* ovrrev\_Mean: Mean overage revenue.
* vceovr\_Mean: Mean voice overage revenue.
* datovr\_Mean: Mean data overage revenue.
* roam\_Mean: Mean number of roaming minutes.

Customer Call Quality / Network Experience

* drop\_vce\_Mean, drop\_dat\_Mean: Average number of dropped voice/data calls.
* blck\_vce\_Mean, blck\_dat\_Mean: Average number of blocked calls.
* unan\_vce\_Mean, unan\_dat\_Mean: Unanswered voice/data calls.
* plcd\_vce\_Mean, plcd\_dat\_Mean: Outgoing calls placed (voice/data).
* recv\_vce\_Mean, recv\_sms\_Mean: Incoming voice calls/SMS received.
* comp\_vce\_Mean, comp\_dat\_Mean: Successfully completed voice/data calls.
* callfwdv\_Mean, callwait\_Mean: Call forwarding and waiting metrics.

Customer Billing and Revenue

* totrev, adjrev, avgrev: Total, adjusted, and average revenue.
* cc\_mou\_Mean, ccrndmou\_Mean: Mean time spent on customer care calls.
* change\_mou, change\_rev: Change in usage and revenue from prior period.
* totmrc\_Mean: Total monthly recurring charges (before overages).

Customer Care Interaction

* custcare\_Mean: Number of customer care calls.
* cc\_mou\_Mean, ccrndmou\_Mean: Time spent with customer service.

Customer Behavioural Features

* threeway\_Mean: Three-way calling usage.
* inonemin\_Mean: Proportion of calls under one minute.
* mou\_cvce\_Mean, mou\_cdat\_Mean: Minutes of use to customer service (voice/data).
* owylis\_vce\_Mean, iwylis\_vce\_Mean: Calls to/from wireless numbers.
* peak\_vce\_Mean, peak\_dat\_Mean: Peak usage minutes.
* mou\_peav\_Mean, mou\_pead\_Mean: Voice and data minutes during peak hours.

Customer Subscription Information

* months: Length of the customer relationship.
* uniqsubs, actvsubs: Number of unique and active subscriptions.
* new\_cell: Indicator if the customer recently got a new cellular plan.
* eqpdays: Days since equipment purchase.
* phones, models: Number of phones and device models owned.

Customer Device Plan

* dualband, refurb\_new, hnd\_price, hnd\_webcap: Device capabilities and pricing.
* crclscod, asl\_flag: Credit class and account-level flags.

Customer Information

* area: Geographic area code.
* truck, rv, ownrent: Ownership of vehicle/property.
* lor: Length of residence.
* marital, adults, income, numbcars, infobase: Family and economic indicators.
* ethnic, prizm\_social\_one: Ethnicity and social classification.
* HHstatin, dwllsize, dwlltype: Household size and dwelling type.
* kid0\_2, kid3\_5, kid6\_10, kid11\_15, kid16\_17: Number of children in different age brackets.
* creditcd: Whether the customer has a credit card.

The meanings of the columns used in this dataset were interpreted based on the definitions provided in the Telecom Churn Case Study Data Dictionary available on Scribd [20].

## 4.3 Data Preparation

Before any machine learning models can be applied, the dataset should be preprocessed first. The dataset is required to undergo several preprocessing steps to ensure quality and consistency of the data as well as the models itself.

The dataset was first loaded from Kaggle using GPU dataframes with cuDF, and only the relevant columns were selected for analysis. During the initial inspection, it was discovered that some numeric features contained empty or missing values. To avoid losing valuable data, these missing values were replaced with the median of each respective numeric column. For categorical columns, missing entries were filled with the mode, which ensured that no rows were discarded unnecessarily.

Once missing values were addressed, categorical variables were converted to numerical form using one‑hot encoding. This transformation allowed the models to interpret categorical information without introducing any artificial ordering. The target column churn was converted into an integer format (0 for non‑churn, 1 for churn) to serve as the label for classification.

The dataset was then split into features (X) and target (y), followed by a stratified train‑test split to preserve the original class distribution. To tackle the imbalance in churn data, the training set was further balanced to a 50/50 ratio by randomly undersampling the majority (non‑churn) class to match the minority (churn) class.

Finally, the numeric features were standardized using Z‑score normalization with StandardScaler, fitting the scaler on the balanced training set and applying the same transformation to the test set. This ensured that all features contributed on a comparable scale, improving model performance and convergence during GPU‑accelerated training.

### Tools and Environment

This research utilized a combination of tools to support data preprocessing, analysis, visualization, and modeling. Python was selected as the primary programming language due to its flexibility and extensive library support for machine learning and data manipulation. Development was carried out using Google Colab, which provided a cloud-based environment with the necessary computational resources and library integrations.

Several Python libraries were used:

* Pandas and NumPy for data cleaning and transformation
* Scikit-learn for model training, evaluation, and preprocessing
* Matplotlib and Seaborn for visualizing feature relationships and distributions

In addition, Power BI was employed for advanced visual analytics and dashboard creation. It enabled interactive exploration of customer churn patterns, helping to support data understanding and stakeholder presentation.

Together, these tools ensured a robust, scalable, and interpretable data science workflow from data acquisition to insights generation.

## 4.4 Evaluation

To measure the performance of the 3 predictive models developed for customer churn analysis, several classification evaluation metrics were utilized. Because of the binary nature of the Churn column, accuracy alone was not sufficient to measure the performance of each model. And so, a combination of performance metrics was employed to gain a more comprehensive evaluation.

The selected evaluation metrics include accuracy, precision, recall, F1-score, and the confusion matrix. These metrics allow for detailed performance analysis by examining how well the models correctly predict churn versus non-churn customers. Precision measures the proportion of true churn predictions that were actually correct, while recall assesses the model’s ability to identify all actual churn cases. The F1-score, being the harmonic mean of precision and recall, offers a balance between the two. The confusion matrix was used to visualize the true positives, true negatives, false positives, and false negatives for each model.

Additionally, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were incorporated to evaluate the model’s discriminatory ability. A higher AUC indicates better performance in distinguishing between churn and non-churn customers across all classification thresholds.

These evaluation metrics were consistently applied across all three machine learning models: Decision Tree, Logistic Regression and Random Forest to ensure fair and objective comparison. The use of multiple evaluation criteria enabled a more robust model selection process aligned with the study's goal of maximizing predictive accuracy and practical relevance in churn forecasting.

# 5.0 Methods of Modelling

5.1 Understanding the Internal Mechanics of Decision Tree Model

Decision tree model is a supervised learning method that classifies data by recursively splitting based on the feature values, thus forming the tree-like structure. It stimulates human decision making by asking a series of questions that narrows down the outcome and making it highly interpretable and useful for binary classification tasks like customer churn prediction [6].

The decision tree model has 4 components that makes up the decision tree model:

* Root Node: Represents the entire dataset and is the first feature selected for splitting.
* Decision Nodes: Intermediate nodes that split the data further based on feature values.
* Leaf Nodes: Terminal nodes that represent the final class label (e.g, churn or no churn).
* Branches: Represent the outcome of a decision and connect nodes in the tree.

Splitting Criteria

To determine the optimal feature for splitting, the algorithm uses impurity metrics like Gini index, entropy and information gain to determine how well a feature separates the target variable. Gini index quantifies the likelihood of having incorrect classification and is computationally simpler. Entropy measures the amount of uncertainty or impurity in the dataset while information gain is the reduction in entropy after the dataset has been split on the feature. [6]

Tree construction and Pruning

The tree model is built recursively or infinitely until a defined criteria has been met. Such criteria can be maximum depth, minimum samples per node or pure class labels. It is a known fact that decision tree model easily overfits the raining data if allowed to grow unchecked, and to prevent that from happening, pruning is used.

* Pre-pruning halts tree growth early using hyperparameters like max\_depth or min\_samples\_split.
* Post-pruning (or reduced error pruning) removes branches after full tree construction by evaluating their impact on validation accuracy [6].

Strength & Limitations

One of the advantages of decision tree is the transparency it has. It is relatively visible and can be easily explained to stakeholders. This makes them suitable for domains like telecom churn, where model interpretability is essential. However, Decision Trees can become unstable with small changes in data and are prone to overfitting, particularly in noisy datasets or those with many irrelevant features. These limitations are often addressed through ensemble methods like Random Forest.

5.2 Understanding the Internal Mechanics of Logistic Regression Model

Logistic Regression is a supervised learning algorithm used primarily for binary classification problems, such as predicting whether a customer will churn (1) or not (0). Despite the name, logistic regression is not used for regression tasks and instead, it models the probability that a given input belongs to a particular class, making it highly applicable for churn prediction tasks in telecommunications [7]

A black background with white text

Description automatically generated

Figure 1: Logistic (sigmoid) function

Source: Adapted from [8]

The model calculates a linear combination of input features and applies a logistic (sigmoid) function to transform the output into a value between 0 and 1. This probability is then compared to a threshold (commonly 0.5) to determine customer churn or not. [7]

To train the model, logistic regression uses a log-loss function at its cost function. The training process involves minimizing this loss using optimization algorithms such as gradient descent, which adjusts the feature weights (β) iteratively to reduce the prediction error [7].

Strength & Limitations

One of the key strengths of Logistic Regression is its interpretability. Each coefficient represents the change in the log-odds of the outcome for a one-unit change in the feature, assuming all other variables are constant. This makes the model especially useful in telecom churn prediction, where stakeholders need clear explanations for why certain customers are at risk. [7]

The limitations of logistic regression are that it assumes a linear relationship between the variables and the log-odds of the outcome and is sensitive to outliers which might skew the predicted probabilities. And despite these limitations, Logistic Regression remains a reliable, interpretable, and computationally efficient model for churn prediction and other binary classification tasks. [7]

5.3 Understanding the Internal Mechanics of Random Forest

Random Forest is a powerful supervised ensemble learning algorithm known for its impressive accuracy and resistance to overfitting. It operates by constructing numerous decision trees during training and then aggregating their predictions to produce a final output using majority vote for classification tasks and averaging for regression. This aggregation enhances both stability and robustness compared to single-tree models like the decision tree model. [9]

How Random Forest works

* Bagging (Bootstrap Aggregation): Each decision tree is trained on a random bootstrap sample (a random sampling with replacement) of the original dataset. This technique introduces variance reduction and ensures each tree learns different patterns.
* Feature Randomness: For each node split, the algorithm considers a random subset of features, rather than evaluating all features. This additional randomness helps reduce the correlation between trees, leading to improved generalization.

Hyperparameters and Tuning

Random forest builds upon the foundation of decision trees, using the same splitting logic, tree structure and classification mechanisms. But random forest creates multiple decision trees in parallel and aggregates their results for improved stability and accuracy, compared to decision tree only having 1 tree at a time. Due to this shared foundation, many of the hyperparameters are similar to those of decision model. [9]

Key hyperparameters include:

1. *n\_estimators*: number of trees in the forest.
2. *max\_features*: number of features considered at each split (commonly √p).
3. *max\_depth, min\_samples\_split, min\_samples\_leaf*: control tree complexity.
4. *n\_jobs*: allows parallel processing.

Feature importance

Random Forest computes feature importance by measuring how much each feature reduces impurity across the ensemble. This provides valuable insight into which customer attributes (e.g., tenure, monthly charges) most influence churn predictions, aiding both model performance and business decision-making.

Strength & Limitations

Random Forest is able to handle very large and high dimensional datasets efficiently and provides built in regularization to prevent situations like overfitting. It can also produce interpretable feature importance scores. But in contrast, it is computationally expensive and slow to predict with many trees. Comparing to decision tree model where there is only 1 tree, random forest is less interpretable.

# 6.0 Literature Review and References

The purpose of this literature is to explore and synthesize prior research related to customer churn prediction in the telecommunications industry using machine learning models like logistic regression mode, decision tree model and random forest. With the increasing availability of customer behaviour and service usage data, machine learning techniques have become essential tools for identifying patterns that indicate potential churn. Building an accurate and interpretable predictive model is a critical business objective as customer churning will affect the profitability and customer retention strategies.

This research paper will focus on 3 widely adopted classification models, Decision tree model, logistic regression model and random forest. These models were selected due to their effectiveness in churn prediction tasks and their balance between transparency, scalability and predictive capabilities. Decision trees and logistic regression are valued for their interpretability and transparency, making them suitable for explaining churn behaviour to stakeholders. On the other side, random forest utilizes an ensemble method to enhance performance by reducing overfitting and improving generalization through the aggregation of multiple decision trees.

To assess and compare the model performances, this research will be deploying a comprehensive set of classification evaluation metrics like precision, recall, F-1 score, support, confusion matrix, accuracy and ROC curve. These metrics provide a multi-dimensional view of the model’s effectiveness for customer churning. Being able to have a dimensional view of the model’s effectiveness will be important in imbalanced datasets where accuracy alone may not be able to truly reflect its performance. In addition, the ROC curve assist in visualizing the model’s ability to distinguish between customer who churned and those that did not across different different threshold settings.

Other than evaluating model’s effectiveness, the literature review aims to identify the key factors for influencing telecommunication churn, for instance contract type, tenure, monthly charges, and service combinations. It will also address the gaps and limitation found in previous researches. These insights will serve as the foundation for comparing models and contribute to building a more actionable churn prediction framework in the telecommunication industry.

6.1 Overview of Models Used for Churn Analysis

In the telecommunication industry, customer churn prediction has become the central of focus where competition is fierce and customer retention is vital to the business and its profitability. As the cost of acquiring new customers are rising, telecom providers utilizing machine learning increases to proactively identify customers who are likely to discontinue their service. According to Forbes, acquiring a new customer can now cost up to 5 to 7 times more than retaining an existing one. Thus, customer retention is now significantly more efficient and profitable strategy for businesses [5].

Using structured data such as service usage, contract type, tenure and billing information, machine learning can uncover hidden patterns in the data that can signal customer dissatisfaction or disengagement. These models were applied used based on research papers:

1. Decision Tree

Decision Tree is a highly interpretable classification algorithm that builds rule-based logic by splitting datasets into subsets based on feature values that yield the highest information gain or Gini reduction. This structure allows it to naturally segment customers into churn and non-churn categories using attributes like tenure, contract type, monthly charges, and service features. Studies by Zhao et al. [14], Huseyinov & Okocha [15], and Orina et al. [16] recognized Decision Tree as an effective baseline due to its simplicity, transparency, and ability to handle both categorical and continuous data. However, it is susceptible to overfitting, especially with noisy or imbalanced datasets, unless proper pruning techniques are applied.

1. Logistic Regression

Logistic Regression is a statistical method tailored for binary classification problems, making it an ideal candidate for churn prediction tasks where outcomes are either churned (1) or not churned (0). It calculates the probability of customer churn based on a weighted combination of input features through a logistic (sigmoid) function. Orina et al. [16] included Logistic Regression in a comparative study of supervised learning models and found it to perform reliably while offering strong interpretability. Its coefficients directly reflect the influence of each independent variable, making it especially useful in identifying and explaining key churn factors. However, it assumes linearity between inputs and the log-odds of the outcome, which may limit its ability to capture complex relationships.

1. Random forest

Random Forest is an ensemble-based classification model that constructs a multitude of decision trees during training and aggregates their predictions to produce a final output. By introducing randomness in both feature selection and data sampling, Random Forest mitigates overfitting and enhances model generalization. It has been one of the most frequently used and top-performing models in customer churn research, with studies by Boozary et al. [13], Zhao et al. [14], Huseyinov & Okocha [15], Orina et al. [16], and Jothi Kumar et al. [17] confirming its robustness and high accuracy. Additionally, Random Forest provides feature importance rankings, making it useful not only for prediction but also for identifying the most influential churn drivers. Its main limitations are increased computational requirements and reduced interpretability compared to single-tree models.

1. XGBoost (Extreme Gradient Boosting)

XGBoost is an advanced boosting algorithm, it has been shown to offer superior accuracy and handling of imbalanced data, making it suitable for telecommunication churn. Boozary et al. [13] and Huseyinov & Okocha [15] both demonstrated the capabilities of XGBoost. But XGBoost requires careful tuning and is computationally expensive, making it less practical for situations where simplicity or interpretability is needed.

6.2 Justification for Model Selection

The selection of machine learning models for this study is grounded in both theoretical relevance and empirical performance evidence from existing literature. In the context of customer churn prediction, models must not only deliver high accuracy but also be interpretable, scalable, and adaptable to real-world data constraints. After an extensive review of academic sources and industry applications, Decision Tree, Logistic Regression, and Random Forest have been chosen for detailed evaluation in this research.

Decision Tree was selected due to its simplicity and interpretability, making it ideal for exploratory analysis and stakeholder communication. Zhao et al. [14], Huseyinove & Okocha [15] and Orina et al. [16] highlighted that decision tree model serves as a strong baseline model that performs well with categorical and numerical data while offering transparent decision rules. But decision tree often overfits, it is crucial to either prune it meticulously or integrate it into ensemble methods.

Logistic regression is included due to its widespread popularity in binary classification and the ability to provide insights into how individual predictors will influence the probability of customer churning. In the business world, knowing when a customer may churn is important. Understanding customer behaviour is as important as accurate prediction. Throughout the multiple studies, its performance has proven to be stable and reliable, particularly in well pre-processed datasets.

Random Forest is selected for its robustness and strong predictive performance, as consistently demonstrated in studies by Boozary et al., Jothi Kumar et al., and others. Its ensemble nature reduces overfitting, enhances generalization, and allows for estimation of feature importance. These characteristics make it well-suited for handling complex, high-dimensional telecom datasets.

While more advanced models like XGBoost showed high accuracy in some scenarios, they come with trade-offs such as computational cost, parameter sensitivity, and lack of interpretability. These models were considered during literature review and may be explored in future work or used as benchmarks.

In summary the decision tree model, logistic regression and random forest offers strong foundation for developing a predictive framework that balances performance, interpretability and practical applicability in the telecommunication industry where customer churn has a direct impact on profitability.

6.3 Summary of Models Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Type** | **Interpretability** | **Overfitting Tendency** | **Scalability** | **Strengths** | **Limitations** |
| **Decision Tree** | Tree-based | High | High (if unpruned) | Moderate | Simple, transparent, easy to visualize | Prone to overfitting, unstable to noise |
| **Logistic Regression** | Statistical classifier | High | Low | High | Efficient, interpretable coefficients | Assumes linearity, sensitive to outliers |
| **Random Forest** | Ensemble (bagging) | Moderate | Low | High | High accuracy, robust to overfitting | Slower and less interpretable |
| **XGBoost** | Ensemble (boosting) | Low | Low | High | Superior accuracy, handles imbalance well | Requires careful tuning, less explainable |
| **ANN (Neural Network)** | Deep learning | Low | High (if not tuned) | High (with GPU) | Learns complex patterns in large datasets | Requires large data, hard to interpret |

6.4 Accuracy Metrics

Evaluating the performance of classification models in churn prediction requires the use of appropriate and reliable metrics. Since customer churn datasets often contain class imbalances (i.e, more non-churners than churners), relying solely on accuracy can be misleading. To ensure a fair and comprehensive assessment, our study uses a combination of evaluation metrics:

|  |  |
| --- | --- |
| **Symbols** | **Meaning** |
| **TP** | True Positives (Correctly predicted churn) |
| **TN** | True Negatives (Correctly predicted non-churn) |
| **FP** | False Positives (Non-churn predicted as churn) |
| **FN** | False Negatives (Churn predicted as non-churn) |

Accuracy

Accuracy measures the overall proportion of correct predictions among all the observations. [10]

Precision

Precision measures out of all the customers predicted to churn, how many actually did churn. When there is high precision then it means there is fewer customers that was predicted to churn but actually didn’t churn. [10]

Recall (Sensitivity)

Measures out of all the actual churners, how many did the model correctly identify. When there is high recall then it means there is less missed churners. [10]

F1-Score

F1-score measures the balance between precision and recall. It is particularly useful when considering both customers who was predicted to churn but actually didn’t and those predicted to not churn but churned. In other words, it is more useful then considering false positives and false negatives. [12]

Support

Measures how many customers in the dataset belongs to each class. For example, churn (1) and no churn (0). It provides context to the other measuring metrics.

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted Churn | Predicted Not Churn |
| Actually Churn | TP | FN |
| Actual Not Churn | FP | TN |

Confusion matrix provides complete overview of how the model performs across the prediction outcomes. [10]

ROC Curve and Area-under-the-curve (AUC)

A graph plotting the True Positive Rate (TP / [TP + FN]) against the False Positive Rate (FP / [FP + TN]) at various thresholds. [11]

AUC is a single value summary of the model’s ability to distinguish churners form non churners. When the value is closer to 1, it means its better. [11]

6.5 Key Factors Influencing Churn

Understanding the factors that drive customer churn is essential for designing effective retention strategies and building accurate prediction models. Based on the review of prior literature and the features available in the dataset used in this study (Telecom Customer Dataset by Abhinav89 [18]), several variables have consistently emerged as strong predictors of customer churn in the telecommunications and banking industries.

One frequently cited factor is contract type. Customers on short-term or month-to-month contracts tend to have higher churn rates due to lower switching costs. This was reflected in patterns observed in Boozary et al. [13], where customer segmentation and plan structures influenced churn behavior.

Customer tenure is another strong indicator. Shorter tenures often signal recent dissatisfaction or unmet expectations. Zhao et al. [14] highlight how tenure contributes to the predictive strength of churn models, especially when combined with payment and contract data.

Billing-related factors such as high monthly charges or sudden changes in fees are also crucial. Customers with higher bills tend to expect more, and dissatisfaction with service quality can trigger churn. This is discussed in Huseyinov and Okocha [15], where billing frequency and values were part of the SMOTE-balanced feature set used for classification.

Additionally, payment method has been flagged as a churn indicator. Customers using electronic check payments were more likely to churn in several ensemble model evaluations [14].

The type of service, such as internet or bundle offerings, also plays a role. Ribeiro et al. [19] emphasize that dissatisfaction with bundled telecommunications services increases churn risk, particularly when expectations on performance or customer support are unmet.

Lastly, while demographics such as gender and senior status were included in several models, their influence varied. Huseyinov and Okocha [15] found minimal standalone impact, though such factors may interact with other features in complex ways.

6.9 Limitations of Previous Literatures

While recent research has shown the effectiveness of machine learning models in predicting customer churn, several limitations remain evident across the studies reviewed.

Firstly, a strong emphasis is placed on model accuracy and performance, with comparatively less focus on feature-level insights and business interpretation. For example, Boozary et al. [13] and Zhao et al. [14] provided detailed model comparisons but offered minimal discussion on how individual variables contribute to churn decisions, limiting the actionable value of their findings.

Secondly, the scope of feature selection is often narrow or unexplored. Most studies, including Huseyinov and Okocha [15], focus on structured customer data such as demographics or contract type, while behavioral attributes like customer engagement, complaint history, or service interaction quality are rarely incorporated.

Thirdly, generalizability remains a challenge. Models trained on specific datasets such as bank or telecom churn, they are rarely validated across other industries or customer contexts, reducing their applicability. Additionally, class imbalance techniques like SMOTE were used selectively, but not all papers clearly addressed how this affected model fairness or stability.

Lastly, explainability and transparency are underrepresented. Though ensemble models like Random Forest and XGBoost show strong performance, studies such as Boozary et al. [13] and Ribeiro et al. [19] do not explore post-hoc explainability methods like SHAP or LIME, leaving decision-makers without a clear understanding of why certain predictions are made.

These limitations highlight the need for churn prediction research that balances predictive power with interpretability, addresses feature diversity, and supports cross-domain applicability goals which this study aims to address.

# 7.0 Comparison of Model

When comparing the main models used within the research to as part of the study’s main goal utilizing the telco dataset aims in predicting customer churn, we can observe that in general Gradient Boosting (XGBoost) performs in overall much better than the other models presented as seen in the tables below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 60\_40 | 70\_30 | 80\_20 | 90\_10 |
| Decision Tree | 0.5868 | 0.5758 | 0.5826 | 0.5839 |
| Gradient Boosting (XGB) | 0.6182 | 0.6199 | 0.6257 | 0.6314 |
| Logistic Regression | 0.5784 | 0.5774 | 0.5809 | 0.5834 |
| Random Forest | 0.6050 | 0.6071 | 0.6129 | 0.6121 |

Table 7.1 Original Model Accuracy Score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 60\_40 | 70\_30 | 80\_20 | 90\_10 |
| Decision Tree | 0.6088 | 0.6008 | 0.6122 | 0.6080 |
| Gradient Boosting (XGB) | 0.6683 | 0.6722 | 0.6738 | 0.6837 |
| Logistic Regression | 0.6062 | 0.6056 | 0.6088 | 0.6094 |
| Random Forest | 0.6486 | 0.6500 | 0.6562 | 0.6596 |

Table 7.2: Original Model AUC Scores

When factoring accuracy scores and AUC scores we can observe that Gradient Boosting (XGBoost) performs the best, which indicates a high level of understanding in its classifying abilities to classify outputs that are not random in nature.

A close-up of a graph

AI-generated content may be incorrect.

Figure 7.1 Original Model Confusion Matrix

A graph of a curve

AI-generated content may be incorrect.

Figure 7.2 Original Model ROC Curve

However when using the confusion matrix and ROC Curve in Figure 7.1 and Figure 7.2 we can observe that the Gradient Boosting (XGBoost) Model fails to perform in predicting class 1 values (customer does not churn) with 0 successful predictions while class 0 values (customer churn occurs) is the only predictor that is successfully predicted leading to a precision score of 0%, recall score of 0% and a F1-score of 0% indicating the model’s unreliability to truly predict customer churn [4]. When looking into the ROC Curve we thus, can see that the results shown for the Gradient Boosting (XGBoost) Model is an inaccurate representation of the true story of the data indicating it performs the “best” despite results from the confusion matrix stating otherwise which is caused by the ROC Curve only factoring the accuracy of the models [21].

This issue could have been caused by a few issues, but we have successfully managed to narrow it down to a severe class-imbalance and no positive-class weighting within the dataset. By default, Gradient Boosting (XGBoost) Model treats classes equally and if positives are much rarer it can learn to minimize loss by predicting the majority class. Gradient Boosting (XGBoost) Model exposes scale\_pos\_weight exactly for this case [3]. As such this may have led to the poor prediction abilities of the Gradient Boosting (XGBoost) Model for class 1 values leading us to determine that the next best model, Random Forest, may serve as a better model for predicting customer churn.

Within the experiments that we ran Random Forest has the best F1 on churn (0.614) and the best balance of precision and recall among the models that we have used. This indicates that the model is able to catch the positive churners while keeping false alarms reasonable as opposed to the other models. When looking at how Random Forest deals with the dataset as well it is much more robust in handling noisy features by averaging many trees so the model is stable and less sensitive to a single bad split [2]. However, if given the right tuning hyperparameters it is noteworthy that Gradient Boosting (XGBoost) Model may sometimes outperform Random Forest as well, this is due to the fact that boosting builds trees that focus on the residuals of previous trees, which often produces more accurate models on structured data like the one that we’re working on allowing sequential error-correction [1]. According to previous studies, empirical benchmarks often show tree-based boosting frequently beats Random Forest on medium-sized tabular datasets when hyperparameters are well-searched leading to the ML community often using boosting as the high-performance option after a Random Forest baseline [1].

# 8.0 Discussion of New Findings

Based on the results of the research we have succinctly managed to achieve the objectives that we’ve set out to achieve as previously defined. Leveraging classification models such as Decision Tree, Logistic Regression, Gradient Boosting (XGBoost) and Random Forest we’ve managed to gather up to an all-time high accuracy of 65% on the Random Forest model with a 63.38% precision on a 70-30 train-test split, indicating the model’s overall ability to effectively predict customer churn from telco services while ensuring that the results for a customer to churn are correct within an acceptable margin. When conducting a test on the importance of the features used to train the dataset, we are also able to identify the top 6 key factors that influence a customer’s decision to discontinue their services with telco companies with top 10 features from over 100+ features within our best performing model (Random Foret) which are simplified as:

1. certain ethnic groups coded as Z, O and J are more likely to churn
2. the age of their current equipment (mobile phone/device, etc.)
3. the total number of months in service
4. the current handset prices offered by the company
5. the range of minutes the call service is being used
6. the mean of monthly charges being recurringly charged to the customer
7. the percentage change in the monthly minute usages against the last 3 months

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 1: Most Prominent Features Used For Predicting Customer Churn

These identified factors represent a higher top-down view of the cumulative points of dissatisfaction with a customer’s journey with the company. These features suggest that churn is heavily influenced by demographics and the customer's lifecycle with the company as seen in the analysis below:

* **Ethnicity**: The features **ethnic\_Z**, **ethnic\_O**, and **ethnic\_J** are significant churn predictors. This does not indicate a customer's ethnicity directly causes them to churn, but instead that there may be unaddressed needs or service issues specific to these demographic groups resulting in them churning
* **Customer Tenure and Equipment**: **eqpdays** and **months** are the time-based features that indicate their usage and activity with the service. **eqpdays** (days since equipment purchase) suggests that customers are more likely to churn a certain period after getting new equipment. This can be caused due to a new device not meeting expectations, defects in the device or a promotion ending. The importance of **months** (length of the customer relationship) highlighted here shows that customers can be at a higher risk of churning during specific points of their lifecycle with the company, such as after their initial contract ends as well.

As such to help improve the customer retention rate of the company certain actions can be taken to avoid customer churn as seen below:

**1. Targeted Customer Engagement**

As ethnicity is one of the key factors identified for customers churning, the company should look into analysing specific needs and points of contention for the identified ethnic groups.

* Launch targeted surveys or focus groups to understand the specific frustrations and preferences of customers in the **ethnic\_Z**, **ethnic\_O**, and **ethnic\_J** groups.
* Create and offer promotions, plans, and services that are tailored to the usage patterns and communication needs of these demographics by further analysing the gathered data. This can include multilingual support or culturally relevant marketing campaigns to help promote relatability and awareness to these lesser-known ethical groups.

**2. Proactive Retention for Long-Term Customers**

The importance of **months** and **eqpdays** shows a need for proactive engagement at critical points during the customer’s lifecycle journey with the company.

* Implement a system to identify customers nearing the end of their contract and offer them renewal incentives, device upgrades, bonus packages or plan adjustments before they start looking at competitors.
* Based on the **eqpdays** feature, the company can look at creating a retention campaign for customers with older devices by offering them a discounted upgrade path to a newer device to keep them engaged with the latest technology whilst being invested with the company.
* Use the **months** data that has been collected to create a "customer lifecycle map." At key milestones (this can be studied further by analysing the months data), which triggers a message for customers to check in on their service review, or provide a loyalty reward and provide incentives.

**3. Monitoring Usage and Billing Changes**

Although these features are not as prominent as demographic or lifecycle features, **hnd\_price**, **mou\_Mean**, **totmrc\_Mean**, and **change\_mou** are also important features to be brought into attention. These features who indicate that the company should monitor customer usage and billing for sudden changes that may occue.

* Implement a system to show significant drops in **mou\_Mean** (average minutes of use) or increases in **change\_mou** via a dashboard for the company. A sudden drop in usage can be a strong predictor of a customer preparing to churn which allows us to indicate how best to optimise service deals.
* Ensure that the **hnd\_price** (handset price) and **totmrc\_Mean** (total monthly recurring charges) are clearly communicated and justified to the customer potential perceived unfairness or confusion regarding the billing can lead to churn. These features may also indicate that a adjustment to price may be necessary as well by either revaluating packages offered or even contract deals to ensure that we remain competitive with competitors.

# References and Annotated Bibliography

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, pp. 785–794. doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).

[2] Y. Kong and T. Yu, "A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification," *Scientific Reports*, vol. 8, no. 1, p. 16477, 2018. doi: [10.1038/s41598-018-34833-6](https://doi.org/10.1038/s41598-018-34833-6).

[3]“XGBoost Parameters,” *XGBoost documentation (release\_0.90)*, XGBoost, 2019. [Online]. Available: <https://xgboost.readthedocs.io/en/release_0.90/parameter.html>

[4] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," Technologies, vol. 12, no. 7, p. 113, 2024. doi: 10.3390/technologies12070113.

[5] S. Kumar, “Customer Retention Versus Customer Acquisition,” *Forbes*, Dec. 12, 2022. [Online]. Available: <https://www.forbes.com/councils/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition> [Accessed: 15-Jul-2025].

[6] B. Majhi, “Decision Tree Algorithm, Explained,” *KDnuggets*, Jan. 15, 2020. [Online]. Available: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html> [Accessed: 15-Jul-2025].

[7] “Machine Learning | Understanding Logistic Regression,” *GeeksforGeeks*, [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/> [Accessed: 15-Jul-2025].

[8] J. Li, “Introduction to Logistic Regression,” *Medium*, Apr. 9, 2020. [Online]. Available: <https://medium.com/data-science/introduction-to-logistic-regression-66248243c148> [Accessed: 15-Jul-2025].

[9] N. Donges, “Random Forest: A Complete Guide for Machine Learning,” *Built In*, Nov. 26, 2024. [Online]. Available: <https://builtin.com/data-science/random-forest-algorithm> [Accessed: 15-Jul-2025].

[10] Evidently AI, “Accuracy, Precision, and Recall,” Evidently AI, 2023. [Online]. Available: <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall> [Accessed: 18-Jul-2025].

[11] Evidently AI, “ROC Curve and AUC Explained,” Evidently AI, 2023. [Online]. Available: <https://www.evidentlyai.com/classification-metrics/explain-roc-curve> [Accessed: 18-Jul-2025].

[12] V7 Labs, “What is the F1 Score and how is it used in machine learning?,” V7, Jul. 2023. [Online]. Available: <https://www.v7labs.com/blog/f1-score-guide> [Accessed: 18-Jul-2025].

[13] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, “Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction,” Int. J. Inf. Manag. Data Insights, vol. 5, no. 1, p. 100331, Jun. 2025, doi: 10.1016/j.jjimei.2025.100331.

[14] H. Zhao, X. Zuo, and Y. Xie, “Customer Churn Prediction by Classification Models in Machine Learning,” in Proc. 2022 9th Int. Conf. Electr. Electron. Eng. (ICEEE), Mar. 2022, pp. 399–407, doi: 10.1109/ICEEE55327.2022.9772553.

[15] I. Huseyinov and O. Okocha, “A Machine Learning Approach To The Prediction Of Bank Customer Churn Problem,” in Proc. 3rd Int. Informatics and Software Eng. Conf. (IISEC), Ankara, Turkey, Dec. 15–16, 2022, doi: 10.1109/IISEC56263.2022.9998299.

[16] D. O. Orina, R. Rimiru, and W. Mwangi, “A Comparative Study of Predictive Data Mining Techniques for Customer Churn in the Banking Industry,” in Proc. 2023 Intelligent Methods, Systems, and Applications (IMSA), Jul. 2023, pp. 222–227, doi: 10.1109/IMSA58542.2023.10217514.

[17] C. Jothi Kumar, A. Anumala, S. E. P., and B. E. S., “Bank Customer Churn and Extra Benefits Prediction Using Machine Learning Model,” in Proc. 2024 Second International Conference on Advances in Information Technology (ICAIT), IEEE, 2024, doi: 10.1109/ICAIT61638.2024.10690731.

[18] A. Abhinav, “Telecom Customer Dataset,” Kaggle, 2018. [Online]. Available: <https://www.kaggle.com/datasets/abhinav89/telecom-customer/data>. [Accessed: 18-Jul-2025].

[19] H. Ribeiro, B. Barbosa, A. C. Moreira, and R. Rodrigues, “Customer Experience, Loyalty, and Churn in Bundled Telecommunications Services,” SAGE Open, vol. 14, no. 2, pp. 1–15, 2024, doi: 10.1177/21582440241245191.

[20] M. Thakur, “Data Dictionary – Telecom Churn Case Study,” Scribd, 2020. [Online]. Available: https://www.scribd.com/document/488111437/Data-dictionary. [Accessed: 18-Jul-2025].

[21] Evidently AI Team, “How to explain the ROC curve and ROC AUC score?,” *Evidently AI (Classification Metrics Guide)*, last updated Jan. 9, 2025. [Online]. Available: <https://www.evidentlyai.com/classification-metrics/explain-roc-curve?utm_source=chatgpt.com>