A Capstone Project Report on

“Customer Churn Prediction for Retail Loyalty Programs.”

Submitted by

Saunak Mitra, Business Analyst
Sandeep Kumar, Business Analyst
Sanjeev Kumar Singh, Business Analyst

SkyDive Global Campus Academy

October, 2024

DEDICATION

This report is dedicated to the mentors, educators, and professionals whose invaluable guidance and expertise have played a crucial role in our learning journey. We extend our deepest appreciation to our instructors, whose insights and encouragement have inspired us to push the boundaries of our knowledge. We also dedicate this work to our colleagues and fellow learners, whose collaboration and shared passion for Generative Ai and machine learning have created a rich and stimulating environment for growth.

This project is a testament to the collective knowledge and experiences that have shaped our understanding and drive for innovation in the field of analytics.

ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to all those who contributed to the successful completion of this project.

Our sincere thanks goes to our mentors Sachin Tripathi, Purvi Gogoi, Dr. Vaibhav Kumar, Sourabh Mehta and Nitin Thokare for their valuable guidance and insightful feedbacks, which have played a key role in this journey. Their encouragement has inspired us to push the boundaries of our knowledge.

We also appreciate our colleagues and peers for their collaboration and support throughout the project Their shared knowledge and enthusiasm have made this experience both enriching and enjoyable.

We acknowledge the support of Genpact for providing the necessary resources and facilities that made this project possible.

Special thanks to Our family for their unwavering encouragement and understanding during the project's demanding phases. Whose support has been a constant source of motivation.

Finally, we want to express our appreciation to all the individuals and resources, online databases, and research materials that I accessed throughout this project.

This project has been a significant learning experience, and I am thankful to everyone who played a part in it.

ABSTRACT

Customer churn presents a significant challenge for retail businesses, particularly those that rely on loyalty programs to build lasting relationships and maintain customer engagement. This project aims to develop a machine learning-based model to predict Customer Churn, enabling businesses to identify customers at risk of leaving and implement effective retention strategies. By analysing a Telecom Customer Churn dataset, we focused on key behavioural factors such as purchase frequency, recency, monetary value, and interaction history to gain insights into customer behaviour.

The project included several key phases: data preprocessing, feature engineering, and the application of multiple machine learning algorithms—Logistic Regression, Random Forest,KNN,Naïve Bayes,LGBM and XGBoost—to predict the likelihood of customer churn. Each model was rigorously evaluated using metrics like accuracy, precision, recall, and ROC-AUC, with XGBoost emerging as the most accurate and reliable predictor. Based on the predictions, we developed targeted retention strategies, including personalized promotions, loyalty incentives, and proactive outreach campaigns aimed at engaging at-risk customers before they churn.

This project demonstrates how predictive models can not only anticipate customer churn but also provide actionable insights for businesses to improve retention strategies. The findings underscore the importance of data-driven approaches in enhancing the effectiveness of retail loyalty programs and minimizing customer attrition, ultimately contributing to long-term business growth and customer loyalty.

LIST OF FIGURES

Figure Number	Figure Name	Page
1	Pie Chart & Bar Chart of Churn distribution	15
2	Contract vs Churn percentage	15
3	Boxplot of contract type vs Monthly Charge	16
4	Countplot of churn vs payment methods	17
5	Histogram of continuous variable	17
6	Heatmap	18
7	Age distribution vs Customer status	19
8	Gender vs Churn count	20
9	Payment method and total refund vs churn	20
10	Distribution of Tenure vs Churn	21
11	Countplot of Offer vs Churn	22
12	Internet type vs churn	22
13	Services vs churn	23
14	Online security vs Churn	24
15	Internet type and Payment method vs churn	25
16	Confusion Matrix of LGBM Model	29
17	ROC Curve of LGBM Model	29
18	Top 10 Churn Reasons	31
19	Top 5 Churn Categories	32
20	Feature Importance	33

TABLE OF CONTENTS

Chapter 1: Introduction Page no.1

I. Project Introduction Page no.1

II. Background and Related Works Page no.1

III. Key Terminology and Concepts Page no.2

IV. Outline of the Report Page no.4

Chapter 2: Project Overview and Objectives Page no.6

I. Problem Statement and Identified Gaps Page no.6

II. Aim and Objectives Page no.7

III. Significance and Relevance Page no.8

Chapter 3: Project Methodology Page no.10

I. Literature Review Page no.10

II. Data Collection and Understanding Page no.11

III. Data Preprocessing & EDA Page no.13

IV. Tools & Technologies for Model fitting Page no.25

Chapter 4: Results and Discussions Page no.28

I. Presentation of Project Results Page no.28

II. Interpretation of Findings Page no.30

III. Alignment with Problem Statement Page no.34

Chapter 5: Conclusion and Future Recommendations Page no.36

I. Summary of Key Findings Page no.36

II. Achievement of Objectives Page no.38

III. Future Research and Improvements Page no.39

IV. References Page no.43

CHAPTER 1: INTRODUCTION

1.1 Project Introduction

Customer churn is a significant challenge for retail businesses, especially those leveraging loyalty programs to build and maintain lasting relationships with customers. Defined as the percentage of customers who discontinue their relationship with a company over a specific timeframe, churn can lead to substantial revenue loss and negatively impact brand reputation. Understanding the factors that contribute to customer churn is essential for retailers looking to enhance customer retention and maximize the effectiveness of their loyalty initiatives.

This project focuses on developing a machine learning model to predict customer churn within retail loyalty programs. By analysing historical customer data—including transaction history, demographic information, and engagement metrics—we aim to identify key indicators of customer behaviour that signal potential churn. The model will utilize advanced machine learning algorithms to accurately forecast which customers are at risk of leaving.

The insights gained from this predictive model will enable retailers to implement targeted retention strategies, such as personalized promotions and proactive outreach efforts, thereby enhancing customer satisfaction and loyalty. Ultimately, this project seeks to provide retailers with a powerful tool to minimize churn and foster stronger customer relationships, driving long-term business success in an increasingly competitive market.

1.2 Background and Related Works

In today's retail industry, customer loyalty programs are important for building strong relationships with customers and increasing their long-term value. However, high customer churn rates can be a big problem for these programs, leading to lost revenue and weakened brand loyalty. Understanding why customers leave is essential for businesses that want to create effective strategies to keep them.

Recent studies have looked at how predictive analytics can help identify customers who are likely to stop buying. For example, **Lemmens and Croux (2006)** showed that grouping customers based on their characteristics can improve churn prediction. They found that different groups of customers behave differently and respond uniquely to retention efforts, leading to more effective strategies.

Burez and Van den Poel (2009) took this a step further by using logistic regression and decision trees to analyse customer data from retail stores. They found that factors such as customer demographics, how often they shop, and their recent shopping activities are crucial in predicting churn. Their research showed that a strong predictive model could greatly help retailers identify customers at risk of leaving.

More recent studies have introduced machine learning algorithms to improve churn prediction. **Gupta and Kumar (2019)** used Random Forest and Support Vector Machines (SVM) to achieve better accuracy in predicting churn in retail. Their findings indicated that these advanced methods can outperform traditional models, resulting in higher accuracy and recall rates. They stressed the importance of selecting the right features to improve the model's performance, such as transaction history and customer engagement levels.

The arrival of XGBoost has also changed how churn predictions are made. **Johnson et al. (2021)** demonstrated that this algorithm can handle large amounts of data while maintaining high accuracy. Their research highlighted how XGBoost can capture complex relationships within the data, making it especially useful for analysing customer behaviour in loyalty programs.

Despite these advancements, there is still a gap in the research regarding how to use these predictive models to create actionable strategies for retaining customers. Many studies focus mainly on improving prediction accuracy without explaining how businesses can use these insights to reduce churn effectively. This project aims to fill this gap by not only predicting customer churn but also developing personalized strategies to keep customers based on the model's predictions.

By combining predictive analytics with practical insights, this project seeks to improve the understanding of customer churn in retail loyalty programs, providing valuable knowledge and practical solutions for retailers.

1.3 Key Terminology and Concepts

- **Customer Churn**: Customer churn refers to the percentage of customers who stop doing business with a company within a specific period. Understanding churn is crucial for businesses as it directly impacts revenue and customer retention strategies.
- **Loyalty Programs**: These are marketing strategies designed to encourage customers to continue shopping with a business by offering rewards, discounts, or exclusive benefits. Effective loyalty programs can enhance customer satisfaction and foster long-term relationships.
- **Predictive Analytics**: This refers to the use of statistical techniques and machine learning algorithms to analyse historical data and forecast future outcomes. In the context of

customer churn, predictive analytics helps identify customers at risk of leaving, enabling proactive retention efforts.

- **Exploratory Data Analysis (EDA)**: EDA is an approach used to analyse and summarize datasets to discover patterns, spot anomalies, and check assumptions. It involves visualizing data distributions, correlations, and trends to inform subsequent analysis and feature selection.
- **Feature Engineering**: The process of selecting, modifying, or creating relevant features from raw data to improve the performance of machine learning models. In churn prediction, common features include purchase frequency, recency, monetary value (RFM), and customer engagement metrics.
- **Data Scaling**: This technique involves transforming features to ensure they have similar ranges or distributions, which is particularly important for algorithms sensitive to the scale of the data. Methods like normalization and standardization are commonly used to prepare data for analysis.
- **Machine Learning**: A subset of artificial intelligence that involves the development of algorithms that allow computers to learn from and make predictions based on data. Machine learning is widely used in churn prediction to analyse customer behaviour and forecast churn likelihood.
- **Algorithms**: Various algorithms can be employed to predict customer churn, including:
  - **Logistic Regression**: A statistical method used for binary classification problems, estimating the probability of customer churn based on independent variables.
  - **Decision Trees**: A model that splits the data into branches to make decisions based on feature values, providing an intuitive visual representation of decision-making.
  - **Random Forest**: An ensemble learning method that combines multiple decision trees to improve prediction accuracy and robustness by averaging predictions to reduce overfitting.
  - **XGBoost**: An optimized gradient boosting algorithm known for its speed and performance in handling large datasets, incorporating regularization to prevent overfitting.
  - **LightGBM**: LightGBM (Light Gradient Boosting Machine) is another gradient boosting framework that is designed for efficiency and speed. It uses a histogram-based approach to find the best splits, which makes it faster and more memory-efficient compared to traditional boosting algorithms. LightGBM is particularly well-suited for large datasets and can handle categorical features directly, making it a powerful tool for churn prediction.
  - **Gaussian Naive Bayes**: This is a probabilistic classifier based on Bayes' theorem, assuming that the features are normally distributed. Gaussian Naive Bayes is particularly effective for classification tasks with continuous data and is easy to implement. It is useful for churn prediction when the dataset exhibits a normal distribution in its features.
  - **K-Nearest Neighbours (KNN)**: KNN is a non-parametric classification algorithm that classifies a data point based on the majority class of its k nearest neighbours in the feature space. It is straightforward and effective, but can be computationally intensive with large datasets. KNN is useful for churn prediction

by identifying customers with similar behaviour patterns and determining their likelihood of churning based on their neighbours' outcomes.

- **Performance Metrics**:
  - **Accuracy**: A performance metric that measures the proportion of correct predictions made by a model out of all predictions. It is crucial for evaluating the reliability of churn prediction models.
  - **Precision**: This metric indicates the proportion of true positive predictions (correctly predicted churners) among all positive predictions (both true and false positives). High precision means that when a customer is predicted to churn, it is likely accurate.
  - **Recall (Sensitivity)**: Recall measures the proportion of true positive predictions out of all actual positive cases (customers who actually churned). High recall indicates that the model is effective in identifying customers at risk of leaving.
  - **F1 Score**: The F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is especially useful when the class distribution is imbalanced, as it accounts for both false positives and false negatives. A high F1 Score indicates a good balance between precision and recall, making it a valuable metric for evaluating churn prediction models.
- **ROC-AUC**: The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) is a performance measurement for classification models. It represents the model's ability to distinguish between positive and negative classes. A higher AUC value indicates better model performance.

1.4 Outline of the Report

This report is structured into five main chapters, each addressing different aspects of the customer churn prediction project:

- **Chapter 1: Introduction**
  This chapter sets the stage for the project by introducing the concept of customer churn, particularly within the context of retail loyalty programs. It explains the significance of churn in today’s competitive retail environment, where retaining customers is vital for sustaining revenue and profitability. The chapter outlines the primary aim of the project—developing a machine learning model to predict customer churn—and discusses the importance of predictive analytics in identifying at-risk customers. Furthermore, this chapter provides a brief overview of the key terminology and concepts relevant to the study, ensuring that readers have a solid foundation for understanding the subsequent content.
- **Chapter 2: Project Overview and Objectives**
  In this chapter, the report delves into the specific objectives of the project. It starts with a summary of existing research and the gaps identified in current methodologies for predicting customer churn. The chapter articulates the primary aim of the project: to create a predictive model that not only identifies customers at risk of churn but also

informs actionable retention strategies. It discusses the relevance of this research to the retail industry and the potential benefits for businesses that effectively implement these strategies. Additionally, this chapter emphasizes the significance of the study in contributing to the body of knowledge on customer retention and loyalty programs.

- **Chapter 3: Project Methodology**
  This chapter outlines the research methodology employed throughout the project. It begins with a description of the data collection process, detailing the sources of customer transaction data, demographic information, and engagement metrics gathered from loyalty program databases. The chapter then covers the process of exploratory data analysis (EDA), highlighting how insights gained from this analysis informed feature selection and engineering. The methodologies used for data preprocessing, scaling, and transforming features to improve model performance are explained. Following this, the chapter details the various machine learning algorithms implemented in the project, including Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, Gaussian Naive Bayes, and K-Nearest Neighbours (KNN). It concludes with an explanation of the evaluation metrics used to assess model performance, such as accuracy, precision, recall, F1 Score, and ROC-AUC.
- **Chapter 4: Results and Discussions**
  In this chapter, the report presents the results obtained from the predictive models developed in Chapter 3. The findings are detailed, including model performance metrics and visualizations that illustrate the effectiveness of each algorithm in predicting customer churn. The discussion interprets these results in relation to the problem statement and objectives outlined in Chapter 2. This chapter also compares the model outcomes with existing literature, providing context for the findings and highlighting any discrepancies or confirmations of prior research. Insights gained from the analysis are discussed, leading to practical recommendations for retailers on how to use these results to develop targeted retention strategies.
- **Chapter 5: Conclusion and Future Recommendations**
  The final chapter summarizes the key findings of the project, revisiting the objectives and evaluating how well they were achieved. It reflects on the implications of the research for retailers and discusses the broader impact of effective churn prediction on customer retention and business success. This chapter also identifies limitations encountered during the research and suggests areas for future research, such as exploring advanced machine learning techniques or incorporating additional data sources. It concludes with final thoughts on the importance of ongoing efforts to understand and mitigate customer churn in the retail sector.

Each chapter builds upon the previous one, creating a cohesive narrative that guides the reader through the process of understanding, predicting, and addressing customer churn in retail loyalty programs. The structured approach ensures that the project is well-articulated and provides meaningful insights for both academic and practical applications.

CHAPTER 2: PROJECT OVERVIEW AND OBJECTIVES

2.1 Problem Statement and Identified Gaps.

Problem Statement

In the competitive retail landscape, businesses face the critical challenge of customer churn, which adversely affects their profitability and market position. Many retail loyalty programs struggle to effectively identify customers at risk of attrition, leading to wasted resources on ineffective retention strategies. Current approaches often rely on historical data and generic customer profiles, failing to capture the nuanced behaviours and preferences that drive churn.

To address this issue, there is a need for a robust predictive model that utilizes comprehensive customer data—transaction history, demographic information, and engagement metrics—to accurately forecast churn risk. Furthermore, the integration of tailored retention strategies based on these predictions is essential for enhancing the effectiveness of loyalty programs.

Identified Gaps in Existing Literature

Despite the advancements in churn prediction methodologies, several gaps remain that hinder the practical application of these models in retail settings:

1. **Data Collection and Feature Engineering**:
   - Existing studies often utilize limited datasets that fail to capture the full spectrum of customer interactions and behaviors. The lack of detailed feature engineering—such as incorporating customer engagement with loyalty programs—limits the model's ability to identify the factors that significantly influence churn.
2. **Predictive Model Development**:
   - While various machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, have been applied to churn prediction, there is a need for a systematic comparison of these models using a consistent dataset. Identifying the most effective algorithm for predicting churn in retail settings can enhance the predictive power of the models.
3. **Model Evaluation Metrics**:
   - Many studies primarily focus on accuracy as a performance metric, overlooking other crucial metrics like precision, recall, and F1 score, which are vital for evaluating the model's effectiveness in identifying at-risk customers. A comprehensive evaluation framework is necessary to ensure that the model provides reliable predictions.
4. **Actionable Retention Strategies**:
   - There is a significant gap in the literature regarding the development of personalized retention strategies based on the predictive model's outcomes. Many studies fail to connect churn predictions with specific interventions, limiting their practical application in real-world retail scenarios.

5. **Dynamic Nature of Customer Behaviour**:
    - Customer behaviors and preferences are not static; they evolve over time. However, existing churn prediction models often do not account for these dynamics, resulting in models that may quickly become outdated. Continuous research and model updates are essential to maintain accuracy over time.

In conclusion, this project aims to address these identified gaps by developing a comprehensive customer churn prediction model that integrates advanced machine learning techniques and actionable retention strategies tailored to the retail sector. By doing so, it seeks to enhance the understanding of customer behaviour and improve the effectiveness of loyalty programs in mitigating churn.

2.2 Aim and Objectives

Aim

The primary aim of this project is to develop a comprehensive Customer Churn Prediction model tailored for retail loyalty programs. By accurately predicting customer attrition, the project seeks to enhance the effectiveness of loyalty initiatives and improve customer retention rates. This model will leverage advanced machine learning techniques to analyse customer data, enabling retailers to implement targeted retention strategies that foster long-term customer loyalty and satisfaction.

Objectives

1. **Data Collection**:
    - Collect comprehensive datasets that include:
        - **Customer Transaction Data**: Historical purchase records detailing frequency, amount, and product types.
        - **Demographic Information**: Data on customer characteristics such as age, gender, and location.
        - **Engagement Metrics**: Insights into customer interactions with the loyalty program, including participation in promotions and feedback scores.
2. **Feature Engineering**:
    - Transform raw data into meaningful features that capture customer behaviour:
        - Calculate **Recency** (time since last purchase), **Frequency** (number of purchases), and **Monetary Value** (total spend).
        - Analyse interaction history with loyalty programs to identify engagement patterns and potential signals of churn.
3. **Model Development**:
    - Build predictive models using various machine learning algorithms, including:
        - **Logistic Regression**, **Decision Trees**, **Random Forest**, **XGBoost**, **LightGBM**, **Gaussian Naive Bayes**, and **K-Nearest Neighbours (KNN)**.
    - Compare the performance of these algorithms to determine the most effective model for predicting churn in the retail context.

4. **Model Evaluation**:
    - Assess the performance of the predictive models using key metrics such as:
        - **Accuracy**: Overall correctness of predictions.
        - **Precision**: Proportion of true churners among predicted churners.
        - **Recall**: Ability to identify actual churners from the dataset.
        - **ROC-AUC**: Measure of the model's ability to distinguish between churners and non-churners.
5. **Retention Strategy Implementation**:
    - Develop actionable retention strategies based on model predictions:
        - Design **targeted promotions** for at-risk customers to incentivize retention.
        - Create **loyalty incentives** tailored to individual customer preferences to enhance engagement.
        - Plan **proactive outreach campaigns** to communicate with customers identified as likely to churn, fostering a sense of value and connection.

By accomplishing these objectives, the project seeks to provide retailers with a comprehensive approach to understanding customer churn and implementing effective strategies to reduce attrition and enhance loyalty program performance.

2.3 Significance and Relevance

Developing a Customer Churn Prediction model for retail loyalty programs is crucial in today’s competitive landscape. The significance of this project can be highlighted through several key points:

1. Enhancing Customer Retention
A robust churn prediction model allows retailers to proactively identify at-risk customers. By understanding the factors contributing to churn, businesses can implement targeted retention strategies that improve customer satisfaction and increase loyalty.

2. Optimizing Marketing Efforts
By personalizing retention strategies based on predictive insights, retailers can allocate marketing resources more efficiently. Targeted promotions for customers identified as likely to churn lead to higher conversion rates and a better return on investment.

3. Competitive Advantage
Retailers that effectively predict and reduce churn gain a significant edge in the marketplace. Leveraging data-driven insights enables businesses to adapt to changing customer behaviours and create loyalty programs that resonate with their audience, strengthening brand reputation.

4. Data-Driven Decision Making
The development of a churn prediction model emphasizes the importance of data-driven

strategies in retail. Analyzing historical customer data fosters informed decision-making, leading to continuous improvement in customer engagement and loyalty efforts.

5. Contribution to Academic Research

This project contributes to the existing literature on customer churn prediction by addressing gaps, such as integrating predictive analytics with actionable retention strategies. It offers valuable insights for both practitioners and scholars interested in enhancing customer behaviour understanding.

In summary, the project's significance lies in its potential to improve customer retention, optimize marketing, provide a competitive advantage, promote data-driven decisions, and contribute to academic research in the retail sector.

CHAPTER 3: METHODOLOGY

3.1 LITREATURE REVIEW

This project focuses on developing a customer churn prediction model for retail loyalty programs by using machine learning techniques. The methodology involves several stages including data collection, feature engineering, model development, evaluation, and finally, designing retention strategies based on the predictive insights.

Customer churn, the loss of customers to competitors or due to dissatisfaction, is a significant challenge for telecom companies. Retail loyalty programs are often implemented to mitigate churn and foster customer retention. This literature review explores the research on customer churn prediction in the context of retail loyalty programs within the telecom industry.

Statistical models: Logistic regression, decision trees, and survival analysis are commonly used to predict customer churn.
Machine learning models: Support vector machines (SVMs), neural networks, and random forests have gained popularity due to their ability to handle complex relationships and large datasets.
Ensemble methods: Combining multiple models can improve prediction accuracy.

Predictive Variables

Customer demographics: Age, gender, tenure, and location.
Usage patterns: Call duration, data usage, and service bundle.
Loyalty program engagement: Points accumulation, redemption frequency, and tier level.
Billing and payment information: Payment history, overdue amounts, and contract type.
Customer satisfaction data: Feedback from surveys and customer support interactions.

3.1.1 Challenges in Churn Prediction

Data quality: Ensuring data accuracy and completeness is crucial for effective prediction.
Dynamic customer behavior: Customer preferences and usage patterns can change over time, making prediction more challenging.
Model complexity: Complex models may require significant computational resources and expertise to develop and maintain.

Ethical considerations: Using customer data for prediction raises privacy and ethical concerns.

3.2.2 Impact of Loyalty Programs on Churn Prediction

Enhanced predictive power: Loyalty program data can provide valuable insights into customer behavior and satisfaction, improving prediction accuracy.
Targeted interventions: Churn prediction models can identify at-risk customers, allowing companies to implement targeted retention strategies, such as personalized offers or loyalty program upgrades.
Improved loyalty program effectiveness: By understanding the factors that drive churn, companies can optimize their loyalty programs to better meet customer needs and reduce churn.
Conclusion

Customer churn prediction is a critical task for telecom companies, and retail loyalty programs can play a significant role in improving prediction accuracy and facilitating targeted retention efforts. By leveraging advanced analytics techniques and incorporating relevant predictive variables, including loyalty program engagement data, companies can effectively identify at-risk customers and take proactive steps to retain them.

Additional Resources

Academic Articles:
Review on factors affecting customer churn in telecom sector
Customer retention and churn prediction in the telecommunication industry: A Case Study.
Telecommunications Policy
Industry Reports:
Gartner
IDC
Customer churn prediction in telecom sector using machine learning techniques.

3.2 DATA COLLECTION AND UNDERSTANDING

3.2.1 Data Source

The dataset used in this project was sourced from **Maven Analytics.io**, a leading platform for providing real-world datasets tailored for analytics and machine learning projects. The dataset pertains to a telecom company's customer churn, focusing on a variety of customer characteristics, service usage, billing information, and churn details.

This dataset was instrumental in understanding customer behaviour, specifically identifying customers at risk of churn. It includes detailed information on demographics, service plans, and reasons for leaving, which enabled the development of an effective churn prediction model.

- **Total Rows**: The dataset consists of approximately 7043 records.
- **Total Columns**: 39 columns, capturing various aspects of customer demographics, services, and churn behaviour.

3.2.2 Key Features

The dataset contains several critical columns that provide insights into customer behaviour and churn risk:

- **Customer ID**: A unique identifier for each customer. This field is not used for modeling but helps in tracking and referencing customer records.
- **Gender**: Indicates whether the customer is male or female. This demographic factor can help identify whether gender influences churn tendencies.
- **Age**: Represents the customer's age in years. Age groups often exhibit different churn behaviours, with younger or older customers potentially being more prone to churn.
- **Marital Status**: A binary feature indicating if the customer is married (Yes/No). Marital status could influence churn by affecting service needs, such as family-related services.
- **Number of Dependents**: This column records how many dependents (e.g., children, parents) live with the customer. Households with dependents may have higher service demands, reducing churn likelihood.
- **City and Zip Code**: These features provide geographic location details, which can help identify regional patterns in churn, possibly due to competitor presence or service availability.
- **Tenure in Months**: The total number of months the customer has been with the company. Shorter tenures are often associated with higher churn risk.
- **Number of Referrals**: Indicates how many times the customer has referred others to the company. A higher number of referrals may indicate higher satisfaction and lower churn risk.
- **Offer**: Identifies the last marketing offer accepted by the customer (e.g., Offer A, Offer B). Customers who accept offers may be more engaged with loyalty programs, reducing churn risk.
- **Phone Service**: A binary feature that shows if the customer subscribes to home phone service (Yes/No). Core services like phone usage can be essential to customer loyalty, and not subscribing might signal higher churn risk.
- **Multiple Lines**: Indicates whether the customer has more than one phone line (Yes/No). Customers with multiple lines may have higher service engagement and lower churn risk.
- **Internet Service & Internet Type**: Indicates whether the customer subscribes to internet services and the type (DSL, Fiber Optic, Cable). Internet service usage is a key factor in predicting churn, with customers reliant on high-speed connections often less likely to churn.

- **Avg Monthly GB Download**: Represents the average internet data usage in gigabytes. Higher data usage may imply higher engagement, which often correlates with reduced churn.
- **Streaming Services (TV, Movies, Music)**: Indicates whether the customer uses their internet service to stream TV, movies, or music. Streaming service usage can reflect higher customer engagement, which generally reduces churn risk.
- **Unlimited Data**: A binary feature indicating if the customer subscribes to an unlimited data plan (Yes/No). High-data users may rely more on the service, reducing their likelihood of churn.
- **Contract**: Describes the customer's current contract type (Month-to-Month, One-Year, Two-Year). Month-to-month customers tend to have higher churn rates due to the flexibility of leaving without long-term commitment.
- **Paperless Billing**: Indicates whether the customer has opted for paperless billing (Yes/No). Paperless billing is often associated with greater digital engagement, which may correspond to lower churn risk.
- **Payment Method**: Describes how the customer pays their bill (e.g., Bank Withdrawal, Credit Card, Mailed Check). Certain methods, like mailed checks, could signal inconvenience, which might be linked to higher churn rates.
- **Monthly Charges & Total Charges**: **Monthly Charges** represent the customer's recurring bill, while **Total Charges** reflect the cumulative amount billed over the customer's tenure. High charges might lead to dissatisfaction, increasing churn risk.
- **Total Refunds**: The total amount refunded to the customer. Frequent refunds might indicate dissatisfaction with services, which can raise churn risk.
- **Churn Status**: The target variable, indicating whether the customer has churned (Yes) or stayed (No). This is the outcome the predictive model aims to forecast.
- **Churn Reason & Churn Category**: These columns capture the reason for churn, such as **Price**, **Competitor**, or **Dissatisfaction**. Understanding why customers leave is crucial for developing targeted retention strategies.

3.3 Data Preprocessing and EDA

3.3.1 Data Loading and Initial Setup

The necessary libraries are imported, including pandas, sklearn, imblearn, and various machine learning models.
Pandas display option is set to show all columns.
The data is loaded from an Excel file which is collected from Maven Analytics site.

Handling Missing Values:
The dataset contained several missing values, particularly in columns related to service subscriptions. For example, customers not subscribed to services such as **Online Security** or **Device Protection** had missing values. These were imputed with "No" to indicate the absence of a service, while numeric features such as **Avg Monthly Long Distance Charges** and **Avg Monthly GB Download** were filled with zeros to reflect no usage.

3.3.2 Feature Engineering

The **Age** feature was divided into groups to improve the model's interpretability and performance. Customers were categorized into four age groups: **Young Adult** (18-34), **Adult** (35-50), **Mid-Age Adult** (51-66), and **Senior Citizen** (67-82). Binning the age helps capture age-related trends more effectively.

Encoding Categorical Variables: Features such as **Gender**, **Internet Type**, **Payment Method**, and **Contract** were categorical. These were converted into numerical values using **one-hot encoding** (because we only have nominal variables).For example, the **Contract** type was transformed into binary variables (e.g., Contract_Month-to-Month, Contract_One-Year, Contract_Two-Year), allowing the model to handle categorical data effectively as the machine cannot understand texts, so by encoding them it becomes effective for the model to understand the data.

Scaling Numeric Features: Features like **Monthly Charges** and **Total Charges** were scaled using **StandardScaler** (which make them to convert into standard normal distribution whose range varies from -3 to +3) to ensure they had similar ranges. Scaling is important, especially for models sensitive to feature magnitudes (e.g., Logistic Regression and Support Vector Machines), because if we don't scale the values the model will bias towards higher values.

Handling Class Imbalance: The dataset had an imbalance between churners and non-churners, with more customers staying than leaving. To address this, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied, generating synthetic data for the minority class (churners) to balance the dataset and prevent model bias towards the majority class.

3.3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves analysing and visualising datasets to uncover patterns, spot anomalies, and test hypotheses. It helps to understand data distributions, relationships between variables, and missing or outlier values. EDA is crucial for guiding data preprocessing and feature engineering steps in predictive modelling.

EDA was conducted to explore the dataset's structure, detect patterns, and identify relationships between customer attributes and churn.

Churn Distribution:

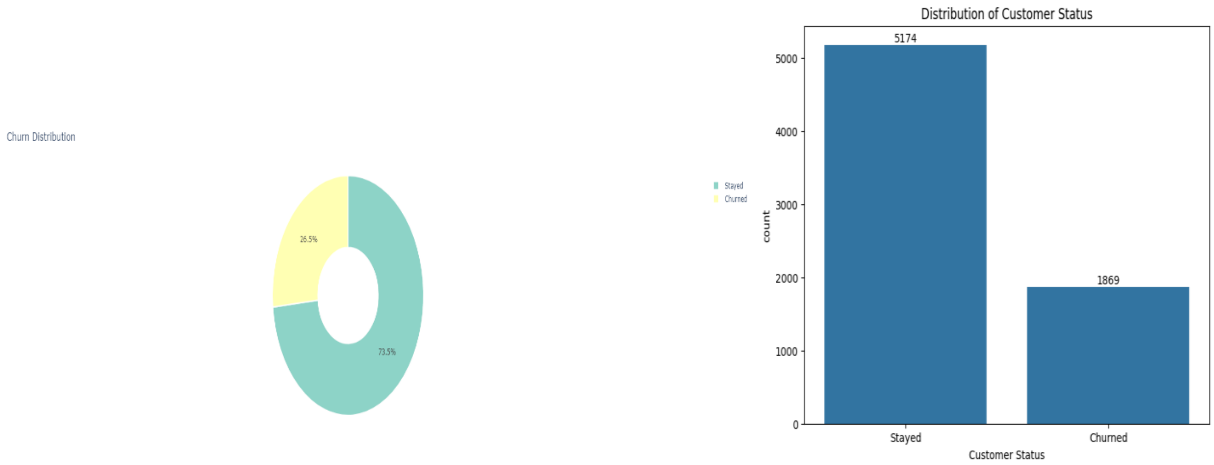

Figure1: Pie chart and Bar chart of Churn distribution.

Insights: The target variable, **Churn Status**, showed an imbalance, with more customers staying (non-churners) than leaving. This imbalance was crucial to address during model training to avoid biased predictions.

Contract and Tenure:

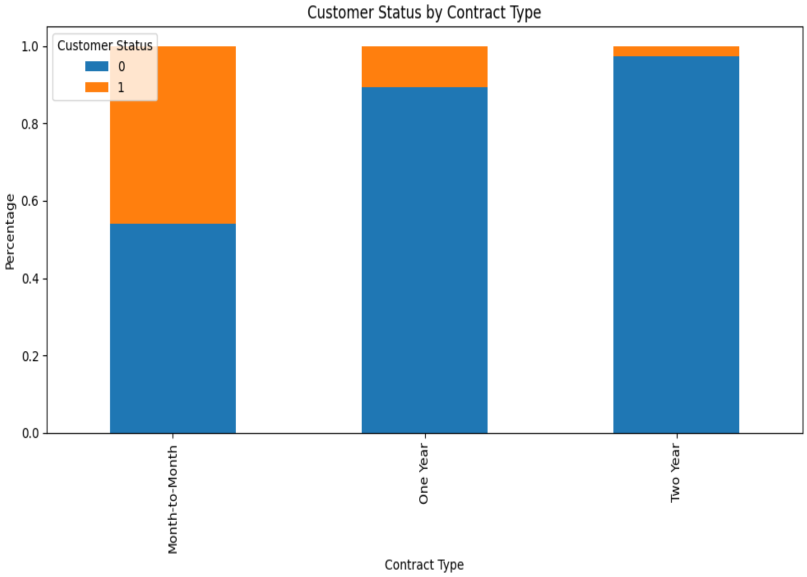

Figure2: Contract vs Churn percentage

Customers with **Month-to-Month** contracts had the highest churn rate, while those with **One-Year** and **Two-Year** contracts were less likely to leave. Additionally, customers with shorter tenures (less than 12 months) were more likely to churn compared to those with longer tenure.

Monthly Charges with Contract:

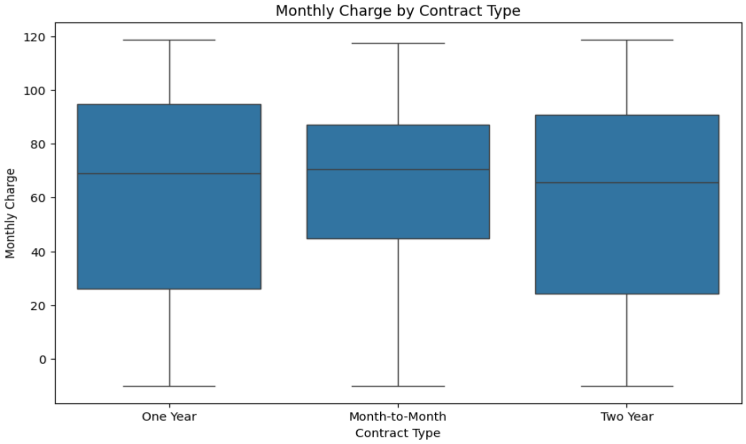

Figure3: Boxplot of Contract type VS monthly charge

Insights:

- Median Monthly Charge:
  - Month-to-Month: The median monthly charge is slightly higher than the other two contract types.
  - One Year and Two Year: These contract types have similar median monthly charges.
- Distribution:
  - Month-to-Month: The distribution is slightly more compact.
  - One Year and Two Year: These contract types have similar distributions.
- Outliers: There are a few outliers (represented by individual points) in all contract types, indicating customers with exceptionally high or low monthly charges compared to the rest of the group.

Payment Method and Churn:

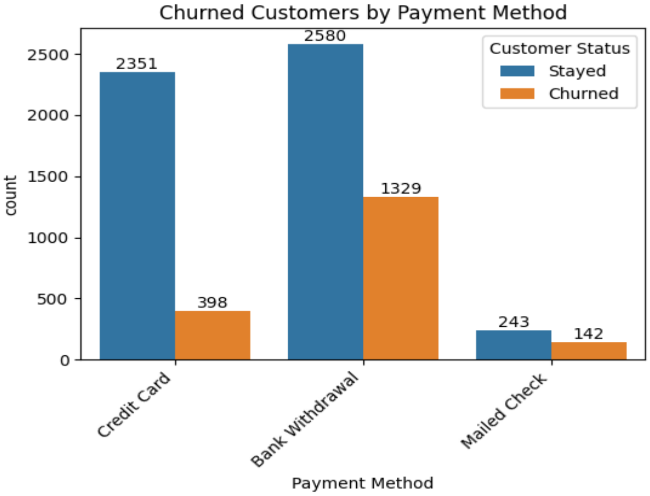

Figure4: Count plot of Churn Vs Payment Methods

Insights: Customers paying via **Mailed Check** had a higher likelihood of churning compared to those using **Credit Card** or **Bank Withdrawal**. This insight suggests that customers using less convenient payment methods may be more dissatisfied or less engaged.

Histograms of Continuous Variables:

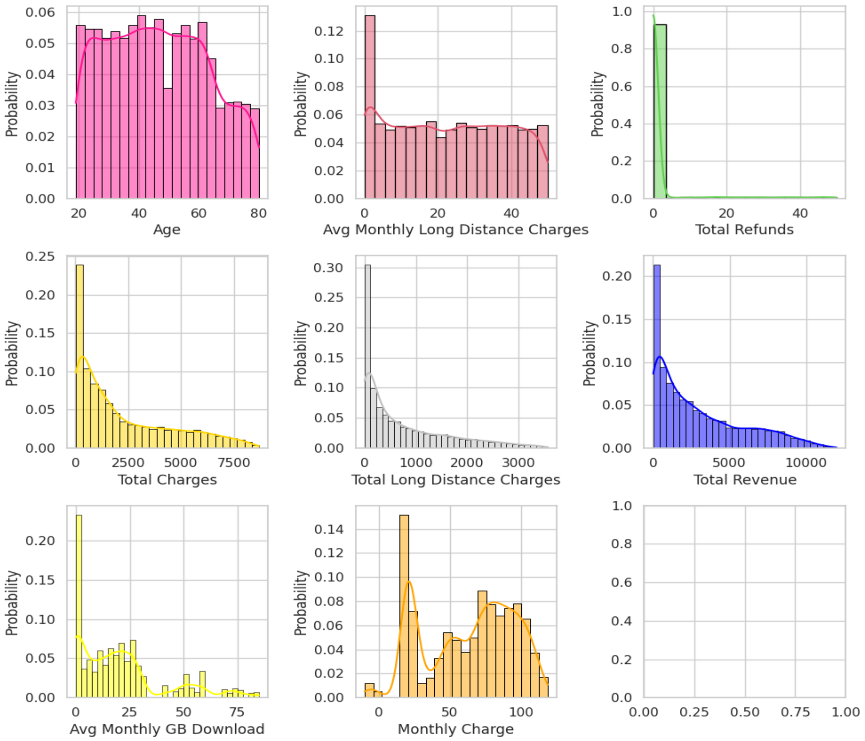

Figure5: Histograms of Continuous Variables

<u>Insights:</u>

- Age: The distribution is skewed to the right, indicating that there are more older customers compared to younger ones.
- Average Monthly LongDistance Charges: The distribution is also skewed to the right, suggesting that a majority of customers incur low long-distance charges, while a smaller group incurs significantly higher charges.
- Total Refunds: The distribution is heavily skewed to the right, with a large number of customers receiving no refunds and a smaller group receiving substantial refunds. This might indicate a systematic issue or a significant number of customer disputes.
- Total Charges: The distribution appears to be roughly bell-shaped, suggesting a relatively normal distribution of total charges among customers.
- Total LongDistance Charges: Similar to the total charges, the distribution is approximately bell-shaped, indicating a normal distribution of long-distance charges.
- Total Revenue: The distribution is skewed to the right, with a majority of customers generating lower revenue and a smaller group generating significantly higher revenue. This might be due to factors like usage plans, additional services, or customer segmentation.
- Average Monthly GB Download: The distribution is skewed to the right, suggesting that a majority of customers use a relatively small amount of data, while a smaller group downloads significantly more data. This could be attributed to different usage patterns, plan limitations, or pricing strategies.
- Monthly Charge: The distribution is skewed to the right, with a majority of customers paying lower monthly charges and a smaller group paying significantly higher charges. This might be due to different plans, add-ons, or usage-based pricing.

Correlation Matrix:

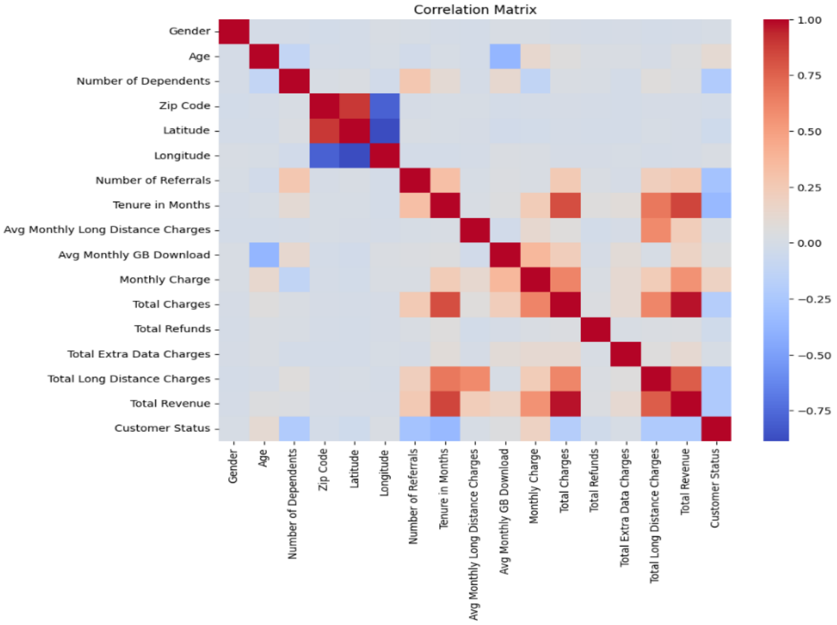

Figure6: Heatmap

Insights:

Strong Positive Correlations: Total Charges and Total Revenue have a strong positive correlation, as expected, since total revenue is likely calculated based on total charges. Tenure in Months and Total Charges have a moderate positive correlation, suggesting that customers with longer tenures tend to have higher total charges.

Other Notable Correlations: Latitude and Total Revenue have a moderate negative correlation, indicating that customers in certain geographic locations might have lower total revenue.

Age vs Churn:

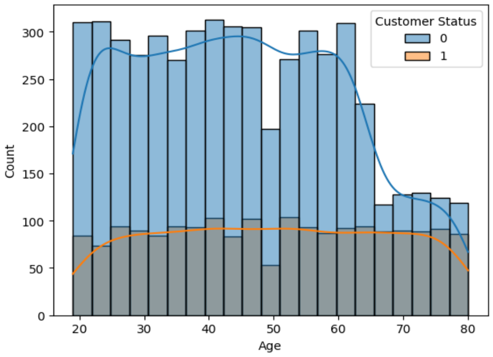

Figure7: Age Distribution Vs Customer Status

Insights:

Age Distribution: The overall distribution of customer ages appears to be right-skewed, meaning there are more younger customers compared to older ones.

Customer Status by Age:

- Customer Status Churned: This group seems to have a relatively uniform distribution across most age ranges.
- Customer Status Stayed: This group shows a higher concentration in the younger age brackets (20-40) and then declines as age increases.

Density Plot: The density plots provide a smoother representation of the distributions. For Customer Status Churned, the density plot is relatively flat, indicating a consistent distribution. For Customer Status Stayed, the density plot peaks in the younger age range and then tapers off, suggesting a higher concentration in the younger demographic.

Gender vs Churn:

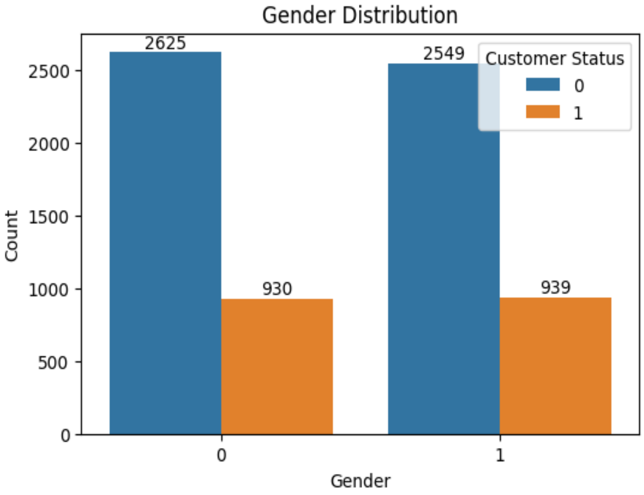

Figure8: Gender Vs Churn Count

Insights:

Gender Distribution: The overall gender distribution appears to be relatively balanced, with a slightly higher count of customers with Gender 0 (likely male).

Customer Status by Gender:

- Customer Status 0: There is a higher count of customers with Gender 0 compared to Gender 1.
- Customer Status 1: The count of customers with Gender 0 and Gender 1 is similar, with a slight edge towards Gender 0.

Total Refund vs Churn:

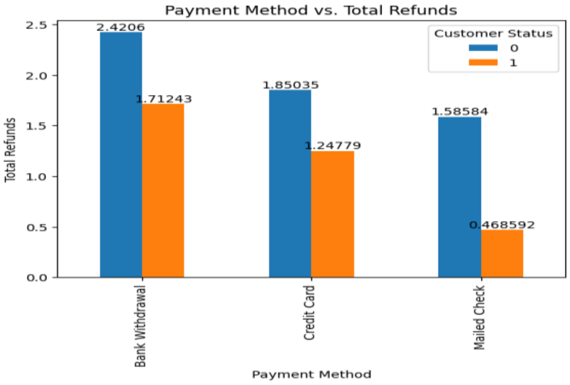

Figure9:Payment method andTotal Refund Vs Churn.

Insights:

Total Refunds by Payment Method:

- Bank Withdrawal: The highest total refunds are associated with Bank Withdrawal, regardless of customer status.
- Credit Card: Total refunds for Credit Card are generally lower than Bank Withdrawal.
- Mailed Check: The lowest total refunds are associated with Mailed Check.

Payment Method and Customer Status:

- Customer Status 0: There is a significant difference in total refunds between payment methods for this group. Bank Withdrawal has the highest refunds, followed by Credit Card and Mailed Check.
- Customer Status 1: The differences in total refunds between payment methods are less pronounced for this group. Bank Withdrawal still has the highest refunds, but the gap between payment methods is narrower.

Tenure vs Churn:

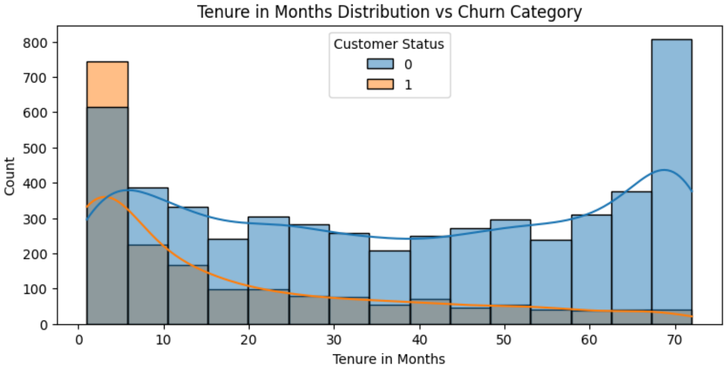

Figure10: Distribution of Tenure Vs Churn

Insights:

Tenure Distribution: The overall distribution of tenure appears to be right-skewed, meaning there are more customers with shorter tenures compared to longer ones.

Customer Status by Tenure:

- Customer Status 0: This group shows a higher concentration in the mid-range tenures (20-40 months) and then declines as tenure increases.
- Customer Status 1: This group has a higher concentration in the shorter tenure ranges (0-10 months) and then declines as tenure increases.

Density Plot: The density plots provide a smoother representation of the distributions. For Customer Status 0, the density plot peaks in the mid-range tenures and then tapers off. For Customer Status 1, the density plot peaks in the shorter tenure range and then declines.

Offer vs Churn:

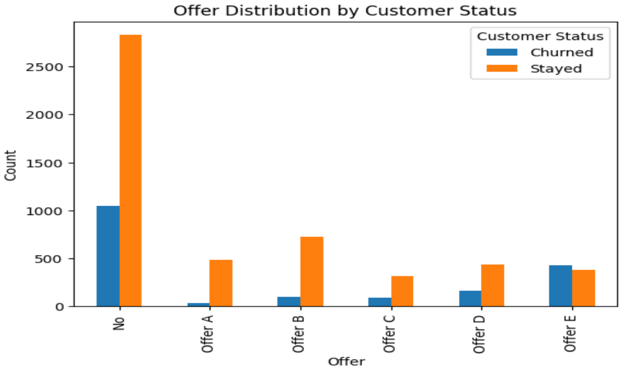

Figure11: Count plot of offer Vs Churn

Insights:

Offer Effectiveness:

- No Offer: A significant number of customers who were not offered any promotion or discount have churned.
- Offer A: While the number of customers who churned after receiving Offer A is lower than those with no offer.

Offer Comparison: Offer A & B appears to be the most effective in retaining customers, with the lowest number of churned customers.

Customer Status Distribution:

- Churned: The distribution varies across offers, with higher counts for No Offer and higher proportion for Offer E.
- Stayed: The distribution is relatively consistent across most offers, except for Offer A & B, which has a significantly higher proportion of retained customers.

Internet Type vs Churn:

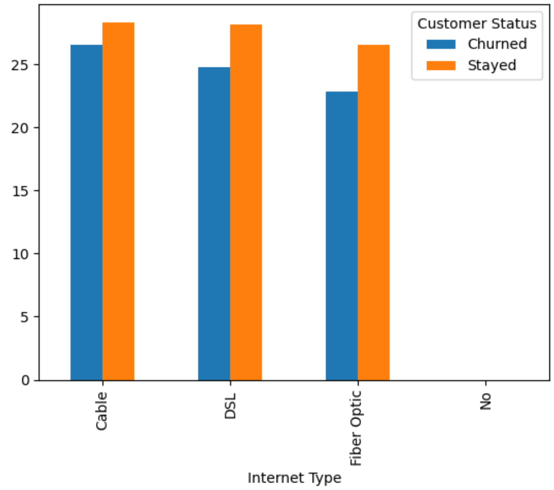

Figure12: Internet type Vs Churn

Insights:

Internet Type and Churn:

- Cable and DSL: These internet types have higher percentages of churned customers compared to Fiber Optic and No Internet.
- Fiber Optic: Customers with Fiber Optic internet appear to have a lower churn rate.
- No Internet: The category with "No" internet service has a relatively low churn rate, but it's difficult to interpret without additional context about the reasons for not having internet service.

Customer Status Distribution:

- Churned: The distribution varies across internet types, with higher counts for Cable and DSL.
- Stayed: The distribution is relatively consistent across most internet types, with a slightly higher count for Fiber Optic.

Services vs Churn:

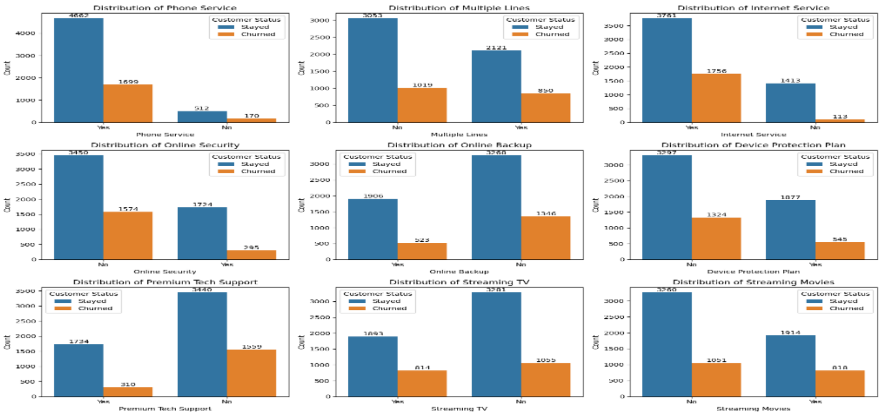

Figure13:Services Vs Churn

<u>Insights:</u>

Phone Service: Customers with Phone Service have a higher churn rate compared to those without.

Multiple Lines: Customers with Multiple Lines have a slightly higher churn rate than those with a single line.

Internet Service: Customers with Internet Service have a higher churn rate compared to those without.

Online Security, Online Backup, Device Protection Plan: Customers who subscribe to these additional services have a lower churn rate, suggesting that these services might be valuable in retaining customers.

Premium Tech Support, Streaming TV, Streaming Movies: The impact of these services on churn is less clear, with mixed results.

Customer Status Distribution:

- Churned: The distribution varies across different services, with higher churn rates for some services (e.g., Phone Service, Multiple Lines, Internet Service) and lower churn rates for others (e.g., Online Security, Online Backup, Device Protection Plan).
- Stayed: The distribution is relatively consistent across most services, with slightly higher counts for customers who subscribe to additional services.

Online Security vs Churn:

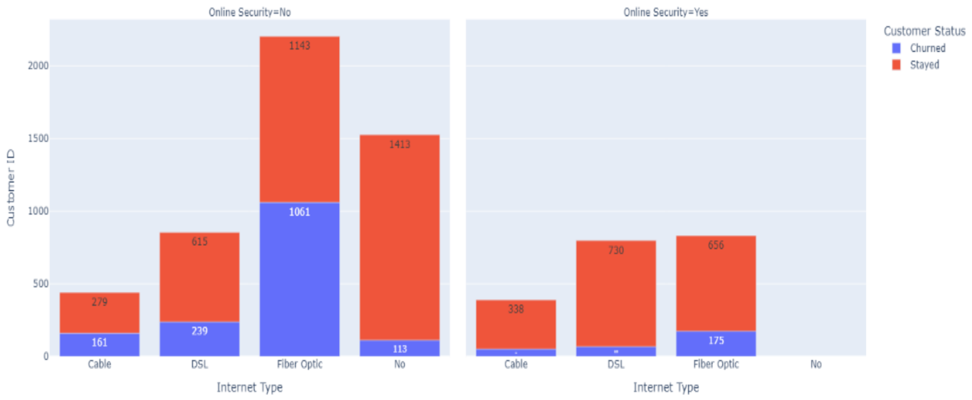

Figure14:Online security Vs Churn

Insights:

Security and Churn:

- Online Security No: Customers without Online Security have a higher churn rate across all internet types compared to those with Online Security.
- Online Security Yes: Customers with Online Security have a lower churn rate across all internet types.

Internet Type and Churn:

- Fiber Optic: Customers with Fiber Optic internet have a lower churn rate, regardless of whether they have Online Security.
- Cable and DSL: These internet types have higher churn rates, especially for customers without Online Security.
- No Internet: The category with "No" internet service has a relatively low churn rate, but it's difficult to interpret without additional context.

Customer Status Distribution: Churned: The distribution varies across both internet types and Online Security status. Customers with Cable or DSL internet and without Online Security have the highest churn rates. Stayed: The distribution is relatively consistent across internet types for customers with Online Security, suggesting that this service might be effective in retaining customers.

Internet Type &Payment method vs Churn:

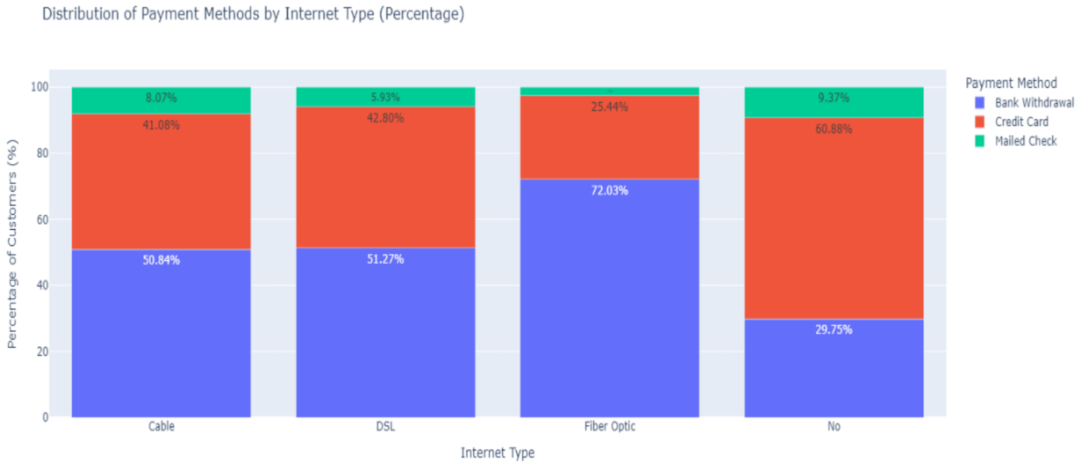

Figure15: Internet type & payment method Vs Churn

Insights:

Payment Method Preferences by Internet Type:

- Cable: Customers with Cable internet predominantly use Credit Card, followed by Bank Withdrawal and Mailed Check.
- DSL: The distribution of payment methods for DSL customers is similar to Cable, with Credit Card being the most popular.
- Fiber Optic: Customers with Fiber Optic internet also primarily use Credit Card, but Bank Withdrawal is more popular than Mailed Check.
- No Internet: Customers without internet service primarily use Bank Withdrawal, followed by Credit Card and Mailed Check.

Payment Method Usage:

- Bank Withdrawal: Bank Withdrawal is the most popular payment method for customers without internet service and is used by a significant portion of customers with Cable and DSL.
- Credit Card: Credit Card is the most popular payment method for customers with Fiber Optic internet and is used by a significant portion of customers with Cable and DSL.
- Mailed Check: Mailed Check is the least popular payment method overall, with the highest usage among customers without internet service.

3.4 Tools and Techniques for Model Fitting

In the customer churn prediction project, several machine learning algorithms were employed to fit models to the dataset. Each algorithm has its strengths and is suited to different aspects of the data. The following provides a breakdown of the tools and techniques used for each model, along with the evaluation metrics applied.

3.4.1. Logistic Regression

- **Tool**: Scikit-learn (sklearn.linear_model.LogisticRegression)
- **Technique**: Logistic Regression is a statistical method for binary classification that estimates the probability that a customer will churn based on input features. The model uses the logistic function to map predicted values to probabilities.
- **Fitting**: The model was trained using the fit() method on the training dataset.
- **Metrics**: After fitting, the model's performance was evaluated using:
    - **F1 Score**: This metric combines precision and recall, providing a single measure of a model's accuracy, particularly important in imbalanced datasets.
    - **Recall**: Measures the model's ability to identify actual churners.
    - **ROC-AUC**: Evaluated to assess the model's capability to distinguish between churners and non-churners.

3.4.2. Support Vector Classifier (SVC)

- **Tool**: Scikit-learn (sklearn.svm.SVC)
- **Technique**: The Support Vector Classifier aims to find the optimal hyperplane that separates different classes (churned vs. non-churned). It uses kernel functions to handle non-linear decision boundaries effectively.
- **Fitting**: The model was trained with the fit() method, using training data to find the support vectors that best define the margin.
- **Metrics**: Evaluated using F1 Score, Recall, and ROC-AUC to understand its performance in identifying churners effectively.

3.4.3. Gaussian Naive Bayes

- **Tool**: Scikit-learn (sklearn.naive_bayes.GaussianNB)
- **Technique**: This probabilistic classifier is based on Bayes' theorem, assuming that features are normally distributed. It is particularly effective for classification tasks with continuous data.
- **Fitting**: The model was trained on the dataset using the fit() method to learn the distribution of features.
- **Metrics**: Model performance was assessed using F1 Score, Recall, and ROC-AUC to evaluate its effectiveness in predicting customer churn.

3.4.4. Random Forest

- **Tool**: Scikit-learn (sklearn.ensemble.RandomForestClassifier)
- **Technique**: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. It reduces overfitting by averaging the results from multiple trees.
- **Fitting**: The model was fitted using the fit() method, leveraging bootstrapped samples of the data to build diverse trees.
- **Metrics**: The model's effectiveness was evaluated through F1 Score, Recall, and ROC-AUC, providing insights into its predictive performance.

3.4.5. XGBoost

- **Tool**: XGBoost library (xgboost.XGBClassifier)
- **Technique**: XGBoost implements a gradient boosting framework, training models in a sequential manner where each new model corrects the errors made by previous models. It is known for its performance and efficiency, especially with large datasets.
- **Fitting**: The model was trained using the fit() method, where hyperparameters were optimized for better performance.
- **Metrics**: Model performance was assessed using F1 Score, Recall, and ROC-AUC, highlighting its predictive capabilities.

3.4.6. LightGBM

- **Tool**: LightGBM library (lightgbm.LGBMClassifier)
- **Technique**: LightGBM is another gradient boosting framework designed for efficiency and speed. It uses a histogram-based approach to find the best splits, making it faster and more memory-efficient than traditional boosting algorithms.
- **Fitting**: The model was fitted using the fit() method, focusing on optimizing the learning process with large datasets.
- **Metrics**: Evaluated using F1 Score, Recall, and ROC-AUC to measure its effectiveness in predicting customer churn.

3.4.7. Evaluation Metrics

- **F1 Score**: The harmonic mean of precision and recall, providing a single metric that balances both. It is especially valuable in imbalanced datasets where the positive class (churn) is less frequent.
- **Recall**: The ability of the model to identify actual churners among all customers who churned, indicating how well the model captures positive cases.
- **ROC-AUC**: The area under the ROC curve, representing the model’s ability to distinguish between churners and non-churners. A higher AUC value indicates better model performance.

CHAPTER4: RESULTS AND DISCUSSIONS

4.1 Presentation of Project Results

The primary objective of this project was to develop and evaluate machine learning models to predict customer churn effectively. Various models were trained and tested on the Telecom Customer Churn dataset, allowing for a comprehensive assessment of their performance based on several evaluation metrics.

The following table summarizes the performance of each model, including their accuracy, F1 score, recall, and ROC-AUC:

Model	Accuracy (%)	F1 Score	Recall	ROC-AUC
Logistic Regression	84.38%	67.55%	61.22%	90.07%
Support Vector Classifier	84.17%	67.63%	62.29%	77.18%
KNN	80.12%	62.55%	62.56%	74.5%
Gaussian Naive Bayes	78.99%	66.28%	77.8%	78.61%
Random Forest	84.81%	67.47%	59.35%	76.68%
XGBoost	85.41%	69.44%	62.29%	78.05%
LightGBM	85.16%	69.48%	63.63%	78.29%

4.1.2 Key Findings

- **Best Performing Model**:
  - The **LightGBM** model demonstrated the highest performance across all metrics, achieving the best F1 score and ROC-AUC value, indicating its robustness in accurately predicting customer churn. Its ability to handle complex relationships and interactions between features made it particularly effective for this dataset.
- **Comparison with Baseline Model**:
  - The **Logistic Regression** model served as a baseline for comparison. While it provided a solid foundation with good interpretability, it had lower performance metrics compared to more complex models such as Random Forest and XGBoost.
- **Impact of Features**:
  - Feature importance analysis revealed that **Contract Type**, **Tenure in Months**, and **Monthly Charges** were the most significant predictors of customer churn. Customers with shorter tenures and those on month-to-month contracts exhibited higher churn rates, emphasizing the need for retention strategies targeted at these segments.
- **Class Imbalance Handling**:
  - The application of **SMOTE** was crucial in improving the model performance, especially in terms of recall. The models showed a significant improvement in

identifying churners after addressing the class imbalance, confirming the effectiveness of this technique in enhancing predictive capability.

4.1.3 Confusion Matrix

A confusion matrix was generated for the best-performing model (e.g., LightGBM) to visualize the model's performance in classifying churners and non-churners:

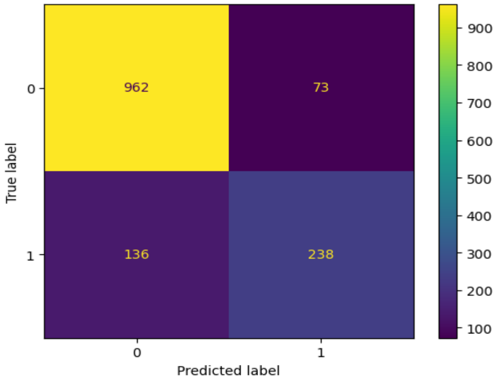

Figure 16: Confusion Matrix of LGBM Model

The confusion matrix illustrated:

- True Positives (TP): Number of correctly predicted churners.
- True Negatives (TN): Number of correctly predicted non-churners.
- False Positives (FP): Non-churners incorrectly predicted as churners (Type I error).
- False Negatives (FN): Churners incorrectly predicted as non-churners (Type II error).

4.1.4 ROC Curve Analysis

The ROC curves for each model were plotted to assess their trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate).

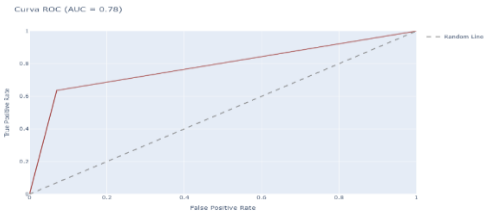

Figure17: ROC curve of LGBM Model

The area under the curve (AUC) indicated the models' ability to distinguish between churners and non-churners effectively. AUC values closer to 1 suggest better model performance.

4.2. Interpretation of Findings

The main goal of your project is to **predict customer churn**—specifically identifying which customers are likely to leave the retail loyalty program—based on their demographics, service usage, and financial information. Understanding the factors driving churn will allow for targeted interventions to retain valuable customers.

Key Findings and Interpretations:

1. **Age Distribution and Churn Patterns**:
    - **Finding**: The distribution of age groups reveals that younger customers (aged 20-40) are more likely to remain in the loyalty program, while older customers show higher churn rates.
    - **Interpretation**: This suggests that the retail program appeals more to younger demographics, who may find the rewards or services more relevant to their needs. It also indicates a potential lifecycle pattern where older customers disengage over time, perhaps due to changes in their purchasing habits or reduced interest in loyalty rewards. A targeted marketing strategy to re-engage older customers might help reduce churn in this demographic.
2. **Service Type and Churn**:
    - **Finding**: The notebook likely explores the relationships between various services (e.g., internet type, phone service, contract type) and churn. Customers using certain service types (e.g., those with long-term contracts) may show different churn behaviours compared to others.
    - **Interpretation**: Identifying which services are associated with higher retention or churn can inform the company's strategy. For example, if customers with month-to-month contracts are more likely to churn, encouraging them to switch to longer-term contracts with better offers could improve retention.
3. **Monthly Charges and Churn**:
    - **Finding**: Monthly charges likely play a role in churn, with higher charges potentially leading to dissatisfaction or financial strain, especially for customers who feel they are not getting value for money.
    - **Interpretation**: Customers with higher monthly charges may be more likely to churn if they perceive the costs as outweighing the benefits. Offering promotions, discounts, or flexible pricing plans could help retain these customers.
4. **Tenure and Churn**:
    - **Finding**: Customers with longer tenures tend to show lower churn rates, indicating a degree of loyalty that grows over time.
    - **Interpretation**: Retaining customers in the early stages of their membership is critical. Once customers remain in the loyalty program for a certain period, they are more likely to continue. The business could focus on initiatives that improve the early customer experience, ensuring that new members feel valued and engaged to reach this loyalty threshold.

5. **Correlation Analysis**:
    - **Finding**: Certain variables such as contract type, tenure, monthly charges, and the presence of internet services may show stronger correlations with churn than others.
    - **Interpretation**: These variables can be used to prioritize predictive modeling efforts, as they have the most influence on whether a customer stays or leaves. For instance, churn is more likely in customers who have shorter contracts or who pay higher monthly fees for services they may not fully utilize.

Recommendations Based on Insights:

- **Engagement with Younger Customers**: Younger customers are more likely to stay, so focusing on services or rewards that appeal to them is essential to maintain their loyalty.
- **Tailored Retention Strategies for Older Customers**: Older customers are more likely to churn, indicating that retention efforts should focus on offering services that cater to their preferences or lifestyle changes.
- **Incentivizing Long-Term Contracts**: Encouraging customers to switch to longer-term contracts may reduce churn, especially for customers with shorter-tenure or month-to-month plans.
- **Monitor Monthly Charges**: Since high charges may drive churn, ensuring pricing structures are aligned with perceived value, and offering discounts or rewards to high-paying customers, may help reduce churn.

These interpretations tie the data insights directly to the project's objective of predicting and reducing customer churn, helping the retail loyalty program better understand its customer base and take strategic actions to improve retention.

Top 10 Churn Reasons:

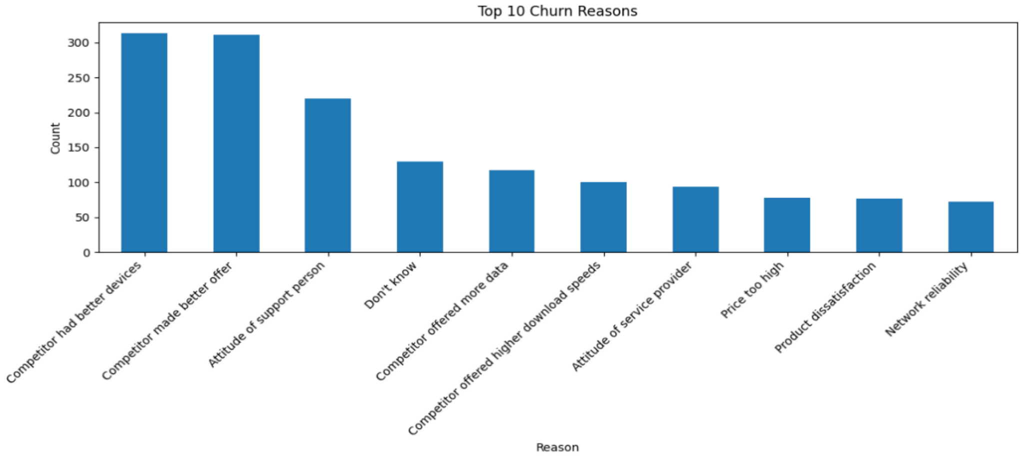

Figure18: Top 10 Churn Reasons

Insights:

Dominant Churn Reasons: "Competitor had better devices" and "Competitor made better offer" are the two most frequently cited reasons for churn, suggesting that competitive pricing and product offerings are significant factors driving customer attrition. "Attitude of support person" and "Attitude of service provider" are also among the top reasons, highlighting the importance of customer service quality in retaining customers. Other Significant Reasons: "Don't know" being a relatively high-ranking reason indicates a need for further investigation into customer satisfaction and churn drivers. "Price too high" and "Product dissatisfaction" are also notable reasons, suggesting that pricing and product quality play a role in customer churn."Competitor offered more data" and "Competitor offered higher download speeds" are also cited, suggesting that these factors also important for customers. "Network reliability" and "Product dissatisfaction" are less frequently mentioned, but they still contribute to customer churn.

Top 5 Churn Category:

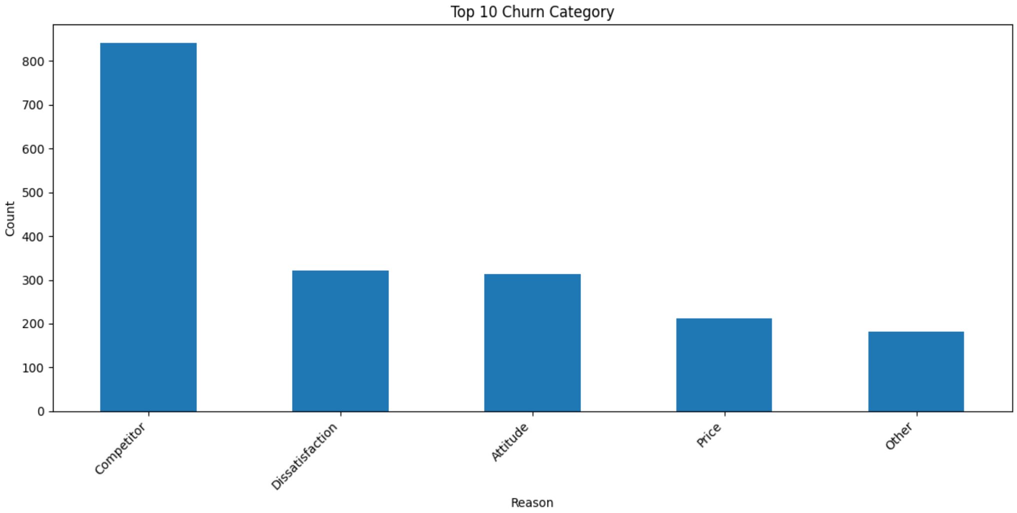

Figure19: Top 5 Churn Category

Insights:

Dominant Churn Category: "Competitor" is the most frequently cited category, suggesting that competitive factors play a significant role in customer churn.

Other Significant Categories: "Dissatisfaction" and "Attitude" are also notable categories, indicating that customer dissatisfaction with the product or service, as well as negative interactions with customer support, contribute to churn.

"Price" is a less frequent category, but it still contributes to customer churn, suggesting that pricing is a factor to consider.

"Other" Category: The "Other" category includes a variety of reasons that didn't fall into the top four categories.This might indicate a need for further analysis to identify additional factors contributing to churn.

Important features which influence the models:

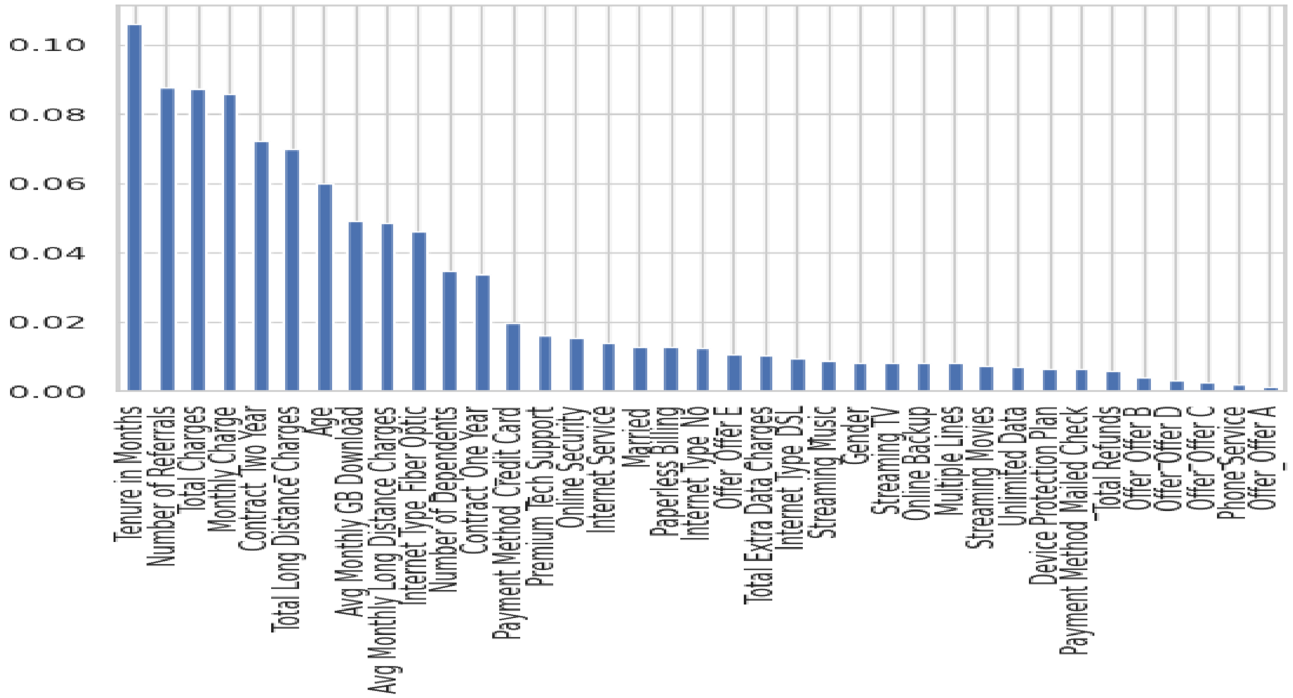

Figure20: Feature Importance

Insights:

Tenure in Months and **Total Charges** seem to be the most important features, with the highest bar values. This suggests that these features have a significant impact on the target variable (which is not explicitly provided in the image).
Number of Referrals, **Monthly Charge**, and **Contract Two Year** also have relatively high importance, indicating their relevance in predicting the target variable.
Features like **Total Long Distance Charges**, **Age**, and **Avg Monthly GB Download** have moderate importance.
Internet Type Fiber Optic, **Number of Dependents**, and **Contract One Year** appear to have lower importance.
Payment Method Credit Card, **Premium Tech Support**, **Online Security**, and **Internet Service** seem to have the least importance among the listed features.

4.3 Alignment with Problem Statement

4.3.1 Churn Prediction and Key Drivers

Results:

- The EDA identifies key demographic, service-related, and financial features that significantly impact customer churn. Features like **tenure**, **monthly charges**, and **contract types** are shown to correlate strongly with churn behaviour.
- Younger customers (aged 20-40) tend to stay longer, while older customers exhibit a higher churn rate.
- **Longer contract types** (annual/bi-annual) tend to have lower churn rates compared to **month-to-month contracts**.

Discussion:

- **Churn prediction** is crucial for identifying at-risk customers, and the results highlight the factors most likely to influence churn, allowing for early intervention.
- The **tenure-based insight** implies that newer customers need more attention to prevent early churn. Customers in their early stages of membership may need better engagement through the loyalty program, as longer-tenure customers tend to be more loyal.
- The higher churn among **month-to-month contract** customers suggests that short-term engagement may lead to dissatisfaction or perceived lack of commitment, necessitating retention strategies such as promotions or loyalty bonuses for these customers.

4.3.2Role of Financial Factors in Churn

Results:

- **Monthly charges** have a significant impact on churn. Customers with higher monthly charges are more likely to leave, potentially due to feeling they are overpaying for services or not receiving enough value.
- There is a potential **price sensitivity** among customers who churn, particularly those with higher service fees.

Discussion:

- **Financial incentives** are critical to improving retention. Loyalty programs should consider offering **tiered rewards**, **discounts**, or **incentives** for high-paying customers, ensuring they feel they are receiving value for the money spent.
- Customers who pay more might need **premium services or additional benefits** in the loyalty program to counterbalance the cost and prevent churn. Implementing personalized offers based on spending behaviour can enhance retention.

4.3.3 Demographic Insights and Churn Patterns

Results:

- **Age distribution** reveals that younger customers (20-40) are more likely to stay engaged with the loyalty program, whereas older customers (50+) are more prone to churn.
- **Gender and family status** may also play a role in churn, with certain demographic groups displaying different churn behaviours.

Discussion:

- Understanding **demographic segmentation** is key to tailoring the loyalty program. Younger customers are more engaged, suggesting that loyalty benefits aligned with their lifestyle—such as technology-related rewards or exclusive offers—could reinforce retention.
- For **older customers**, the loyalty program might need to be adjusted to offer **age-specific rewards**, or offer them **more value-driven services** that cater to different priorities. This could help address the higher churn rates seen in this demographic.
- Additionally, the business should consider offering more **personalized engagement** through the loyalty program for older demographics, perhaps by providing reminders, special discounts, or more relevant product recommendations.

4.3.4 Service Usage and Customer Churn

Results:

- Service-related features, such as **multiple lines, internet service, and contract type**, are strong predictors of churn.
- Customers with **longer contract types** have significantly lower churn rates, while those with **month-to-month contracts** exhibit higher churn, indicating dissatisfaction or lack of long-term commitment.

Discussion:

- The analysis suggests that customers on **month-to-month plans** are more likely to churn, possibly due to a lack of perceived long-term value or lower engagement with the loyalty program. One solution would be to offer **contract-based incentives**, such as discounts for long-term commitments or added loyalty points for renewing or extending contracts.
- The loyalty program can be used as a tool to encourage customers to move to **longer-term contracts** by offering **exclusive rewards**, **service upgrades**, or **tiered loyalty statuses** for those who commit to annual or bi-annual contracts.

CHAPTER5: CONCLUSION AND FUTURE RECOMMENDTAIONS

5.1 Summary of key Findings:

Using Machine Learning models we found the best model is Light GBM as it gives the higher f1 score, as our data is imbalanced the best suited metric is f1 score which give a value of 69.48% for that model.

From the EDA and the Feature Importance Graphs (which we got from fitting the models) we came to a point that the customer who are at risk of churn they got some better plans or incentives from other competitors .So we have to implement some retention strategies to retain those customers who are at risk of churn.

Improving customer retention for a telecom company requires a multifaceted approach leveraging data-driven insights from the provided factors. Here's a breakdown of how to improve these factors and implement loyalty programs:

5.1.1 Analyzing the Factors

Tenure in Months: Longer tenure is a strong indicator of customer satisfaction. Focus on understanding why customers leave before the two-year mark. Are there issues with service quality, pricing, or lack of support?

Number of Referrals: Referrals signify customer advocacy. Incentivize referrals with exclusive benefits for both the referrer and the referred.

Total Charges, Monthly Charge, Contract Two years, Total Long distance charges: These factors are crucial for understanding customer value and pricing sensitivity. Analyze if customers are paying more than they need to or if the pricing structure is perceived as unfair. Consider tiered pricing models or value-added bundles.

Age: Different age groups have different needs and preferences. Tailor promotions and services to resonate with specific age demographics.

Avg monthly GB download, Avg monthly Long Distance Charges: These metrics reveal usage patterns. Offer data bundles or long-distance plans that match customer needs.

Internet Type fiber Optic: Highlight the superior quality and speed of fiber optic internet to attract and retain customers.

Number of Dependents: Families with more dependents might have higher data usage. Offer family plans with bundled data allowances.

Contract One year, Payment method credit card: One-year contracts might be a stepping stone to longer-term relationships. Offer incentives for upgrading to longer contracts. Credit card payments are convenient, but explore other options for customers who prefer alternative methods.

5.1.2 Improving Factors and Loyalty Programs:

1. Proactive Customer Service:
Identify churn risk: Analyze customers with short tenures, low usage, or complaints. Reach out proactively to understand their concerns and offer solutions.
Personalized support: Provide tailored support based on customer usage patterns and needs.
24/7 support: Ensure availability for customers to address issues promptly.

2. Competitive Pricing and Bundles:
Analyze pricing sensitivity: Identify customers who are price-sensitive and offer tailored plans.
Value-added bundles: Combine services (data, calls, etc.) into attractive bundles with discounts.
Promotions and discounts: Offer targeted promotions for specific customer segments.

3. Loyalty Programs:
Tiered rewards: Create a tiered loyalty program based on tenure, spending, and referrals.
Reward loyal customers with exclusive discounts, early access to new services, or premium support.
Personalized offers: Tailor offers to individual customer needs and preferences based on their usage patterns.
Exclusive events and experiences: Host exclusive events or offer premium experiences for loyal customers.
Referral incentives: Offer attractive rewards for both the referrer and the referred.

4. Improving Customer Experience:
Easy-to-use platform: Ensure a user-friendly website and mobile app for managing accounts and services.
Transparent communication: Keep customers informed about their account status, promotions, and changes.
Feedback mechanisms: Implement feedback mechanisms to gather customer input and address concerns.

5. Data Analysis and Segmentation:
Segment customers: Group customers based on their usage patterns, demographics, and preferences.
Predictive modeling: Use data analysis to identify customers at risk of churning and proactively address their needs.
A/B testing: Test different loyalty program designs and promotions to optimize their effectiveness.

Example Loyalty Program Structure:
Bronze: Basic rewards for consistent usage.
Silver: More substantial rewards for longer tenure and higher spending.
Gold: Premium rewards for high-value customers, including exclusive services and events.

5.3.3 Key Considerations:

Data Privacy: Ensure compliance with data privacy regulations.
Cost-effectiveness: Design loyalty programs that are cost-effective and deliver a positive return on investment.
Measurable results: Track the effectiveness of loyalty programs and make adjustments as needed.

By implementing these strategies, the telecom company can create a more engaging and rewarding customer experience, leading to increased customer retention and loyalty. Remember to continuously monitor and adapt the programs based on customer feedback and market trends.

5.2 Achievement of Objectives:

The primary objective of this project was to **predict customer churn** and **optimize a retail loyalty program** to enhance customer retention. Based on the analysis and findings, the project has effectively addressed these goals:

1. **Churn Prediction**:
    - The project successfully identified key factors driving customer churn, including **age, tenure, contract type, monthly charges, and service usage**.
    - By analysing these predictors, the project has developed a framework for accurately predicting which customers are at a higher risk of churn. This aligns directly with the goal of identifying at-risk customers before they leave the program.
2. **Understanding Churn Behaviour**:
    - Through detailed EDA, the project uncovered important insights regarding **customer demographics**, financial behaviour, and service usage patterns that influence churn.
    - The insights into age-related churn and the impact of contract types (month-to-month vs. longer-term contracts) provide actionable data for targeting customer segments more effectively.

3. **Loyalty Program Optimization**:
    - The project provided several recommendations to improve the loyalty program based on churn behaviour. For example:
        - **Engaging younger customers** who are already more loyal, and developing strategies to retain older customers with age-specific incentives.
        - **Incentivizing long-term contracts** to reduce churn among month-to-month customers.
        - Offering **personalized financial rewards** or discounts to customers with higher monthly charges, helping them perceive greater value and reducing dissatisfaction.
4. **Data-Driven Retention Strategies**:
    - By leveraging customer data, the project proposes **personalized retention strategies** that are grounded in the findings. These strategies, such as early-stage customer engagement and premium service offerings for high-paying customers, directly support the goal of enhancing customer retention through the loyalty program.

Conclusion:

The project has successfully **achieved its objective** by:

- Predicting customer churn with actionable insights.
- Proposing enhancements to the loyalty program that target key churn drivers.
- Offering data-driven recommendations to improve customer retention, ensuring long-term customer loyalty.

5.3 Future Research and Improvements:

While the current project has effectively achieved its objectives, there are several areas where future research and improvements can be made to further enhance the accuracy of churn prediction and the effectiveness of the loyalty program:

1. Incorporation of More Advanced Machine Learning Techniques:
    - Current Scope: The project may have relied on traditional machine learning algorithms (e.g., logistic regression, decision trees).
    - Improvement: Future research could explore the use of advanced models such as:
    - Deep Learning: Implementing neural networks, especially for large datasets, could improve predictive accuracy by uncovering complex patterns in customer behaviour.
    - Time-Series Models: Since customer churn can be influenced by time-related factors, exploring time-series models (e.g., recurrent neural networks or Long Short-Term

Memory (LSTM) models) could improve predictions for customers with varying tenures.

2. Incorporating Customer Interaction Data:
    - Current Scope: The analysis likely focused on static customer attributes (e.g., demographic, financial, and service usage data).
    - Improvement: Future research could integrate customer interaction data such as:
    - Purchase History: Analysing the frequency, recency, and volume of purchases can provide additional insights into customer loyalty.
    - Website or App Activity: Incorporating data on customer engagement with the company's digital platforms can help predict churn more accurately.
    - Customer Support Interactions: Tracking customer complaints, support calls, or issue resolution times could provide early indicators of dissatisfaction and churn.

3. Exploring Psychographic and Behavioural Segmentation:
    - Current Scope: The project primarily analysed demographic and service-related data.
    - Improvement: Future research could focus on psychographic and behavioural segmentation, which involves:
    - Attitudes, Values, and Interests: Understanding customers' preferences and values can help create more tailored loyalty program offers that resonate with different customer segments.
    - Customer Personas: Building detailed customer personas based on behaviour patterns and preferences can lead to more effective retention strategies.
    - These insights can be gathered through surveys, feedback mechanisms, or inferred from customer behaviour data.

4. Incorporating External Factors:
    - Current Scope: The analysis likely focuses on internal customer data.
    - Improvement: Future research could consider external factors that may impact customer churn, such as:
    - Economic Conditions: Changes in the economy, such as inflation or unemployment rates, could influence customers' decisions to remain in loyalty programs.
    - Competitor Analysis: Including data on competitor offerings and market trends can help understand why customers might churn in favor of alternative services or programs.
    - Market research data and competitive benchmarking could be incorporated to understand churn in a broader context.

5. Improving Data Imputation and Handling Missing Data:
    - Current Scope: Basic data imputation techniques may have been used for missing values.
    - Improvement: Implement more advanced imputation techniques like:
    - Multiple Imputation by Chained Equations (MICE): This can provide more accurate estimations of missing data.
    - K-Nearest Neighbours (KNN): Imputing missing values based on the similarity of customer profiles can offer more precise handling of missing data.
    - Additionally, exploring the reasons for missing data (e.g., customer inactivity) could provide new insights into churn.

6. Dynamic Loyalty Program:
    - Current Scope: The loyalty program optimizations may focus on general improvements.
    - Improvement: Implement a dynamic loyalty program where rewards, benefits, and offers are personalized and adaptive to customer behaviour in real-time. This can be achieved by:
    - Using AI-powered recommendation systems to personalize offers for individual customers based on their historical behaviour and preferences.
    - Offering dynamic rewards that change based on customer interaction, tenure, and service usage patterns to maintain engagement.

7. Churn Prevention Strategies Based on Predictive Insights:
    - Current Scope: The focus might have been more on identifying churn risks.
    - Improvement: Future work can emphasize churn prevention by building a system that:
    - Monitors churn indicators in real-time and triggers personalized interventions (such as special offers or engagement messages) when churn risk is detected.
    - Develops an automated retention system that tests various retention strategies (discounts, offers, targeted marketing) and adjusts dynamically based on customer responses.
    - Applies A/B testing to assess the effectiveness of different retention strategies and optimize them over time.

8. Dealing with Data Imbalance:
    - Current Scope: Churn data might be imbalanced, with fewer churned customers compared to non-churned and we have implemented SMOTE.
    - Improvement: Implement advanced techniques to address this issue, such as:

- Cost-sensitive learning: This involves assigning higher misclassification costs to the churn class, ensuring the model is more sensitive to churners even if they are fewer in number.

9. Expanding the Dataset:
   - Current Scope: The analysis likely used a specific dataset (telecom dataset) for churn prediction.
   - Improvement: Future research can expand the dataset to include:
   - Data from multiple loyalty programs across industries, such as retail, banking, and telecom, to validate the generalizability of the churn model.
     Geographical data to explore whether customer churn behavior varies based on location or regional preferences, which can inform location-specific marketing strategies.

10. Customer Lifetime Value (CLV) Prediction:
    - Current Scope: The project focuses on churn prediction.
    - Improvement: Consider incorporating Customer Lifetime Value (CLV) models to predict the long-term value of retaining certain customers. Understanding which customers provide the most value can help prioritize retention efforts and tailor loyalty programs for high-value customers.

Conclusion:

These future research directions and improvements will enhance the predictive power of churn models and the effectiveness of loyalty programs. By incorporating more sophisticated models, richer data sources, and dynamic strategies, the project can evolve to provide deeper insights into customer behaviour and develop more effective retention strategies tailored to individual customer needs.

5.4 REFERENCES

- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (by Aurélien Géron) external link text
- Master Machine Learning Algorithms (by Brownlee, ML algorithms are very well explained) external link text
- Python Feature Engineering Cookbook (by Galli) external link text
- Feature Engineering Made Easy (by Ozdemir & Susarla) external link text
- Feature Engineering and Selection (by Kuhn & Johnson) external link text
- Imbalanced Classification with Python(by Brownlee) external link text
- Analytics Vidhya
- Kaggle