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 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, 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** |
|  |  |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Figure Name** | **Page** |
|  |  |  |

**TABLE OF CONTENTS**

Dedication

Acknowledgement

Abstract

List of Figures

List of Tables

**Chapter 1: Introduction** Page no.

1. Project Introduction Page no.
2. Background and Related Works Page no.
3. Key Terminology and Concepts Page no.
4. Outline of the Report Page no.

**Chapter 2: Project Overview and Objectives** Page no.

1. Problem Statement and Identified Gaps Page no.
2. Aim and Objectives Page no.
3. Significance and Relevance Page no.
4. Report Structure Page no.

**Chapter 3: Project Methodology** Page no.

1. Research Methodology Page no.
2. Data Collection and Processing Page no.
3. Tools and Technologies Page no.
4. Experimental Design or Data Sources Page no.

**Chapter 4: Results and Discussions** Page no.

1. Presentation of Project Results Page no.
2. Interpretation of Findings Page no.
3. Alignment with Problem Statement Page no.
4. Comparison with Literature Page no.

**Chapter 5: Conclusion and Future Recommendations** Page no.

1. Summary of Key Findings Page no.
2. Achievement of Objectives Page no.
3. Future Research and Improvements Page no.

**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 analyze 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 2: GUIDELINE ON REFERENCES**

1. In-text Citations:
   1. Use author-date citations within the main text (e.g., Smith, 2019).
   2. Place the author's last name and the publication year in parentheses.
2. Reference List:
   1. Include a reference list at the end of your document.
   2. List references alphabetically by the author's last name.
   3. If there are multiple works by the same author, list them chronologically.
3. Basic Format for Different Sources:
   1. Follow the general format: Author(s). (Year). Title. Source.
   2. Include additional information depending on the source type (e.g., page numbers for books, URLs for web pages).

**Examples**

1. Research Paper:
   1. Author(s). (Year). Title of the paper. Title of the Journal, Volume(Issue), Page numbers.
2. Example:
   1. Smith, J. (2019). Predictive Analytics in Healthcare. Journal of Health Informatics, 8(2), 112-125.
3. Book:
   1. Author(s). (Year). Title of the Book. Publisher.
4. Example:
   1. Johnson, R. (2020). Data Science Essentials. Wiley.
5. Online Web Page:
   1. Author(s) or Organization. (Year). Title of the Web Page. URL.
6. Example:
   1. OpenAI. (2021). GPT-3.5: Language Models. https://openai.com/models/gpt-3.5

**Additional Tips**

For multiple authors, list all authors in the same order as they appear in the source.

1. When citing a specific page or chapter in a book, include page numbers (e.g., pp. 45-67).
2. For online sources, provide the full URL and ensure it is accessible.
3. If there is no identifiable author, use the organization or website name as the author.
4. For in-text citations with multiple authors, use "et al." after the first author's name when there are three or more authors (e.g., Smith et al., 2018).