1. Introduction

- Introduction
- Aims and objective of the work
- Brief Literature Review
- Problem Definition
- Plan Of Their Work

1.1 Introduction

The study focuses on a global retailer involved in e-commerce, offering a broad assortment of products across various markets. E-commerce provides opportunities like breaking geographical barriers and enabling faster customer exploration of products, but it also introduces challenges such as increased competition and difficulty in fostering personal customer relationships.

Customer retention is highlighted as more cost-effective and profitable than acquisition, with a need to build trust and loyalty. analyzing customer behavior is critical for improving retention, shifting from transactional metrics (e.g., purchase frequency, average spend) to non-transactional interactions (e.g., loyalty program engagement).

Understanding customer behavior across all stages of their journey, not just during purchases, helps retailers identify what drives customer loyalty and repeat interactions. This holistic approach can enhance the shopping experience and strengthen customer relationships.

1.1.1 The customer journey as defined by the global retailer

The customer journey, as outlined by the global retailer, is a comprehensive process encompassing all customer interactions with the retailer. It is not confined to a single session but comprises an ongoing flow of sessions until the customer chooses to disengage. This journey is dynamic, with sessions capable of starting or ending at any stage and moving in multiple directions.

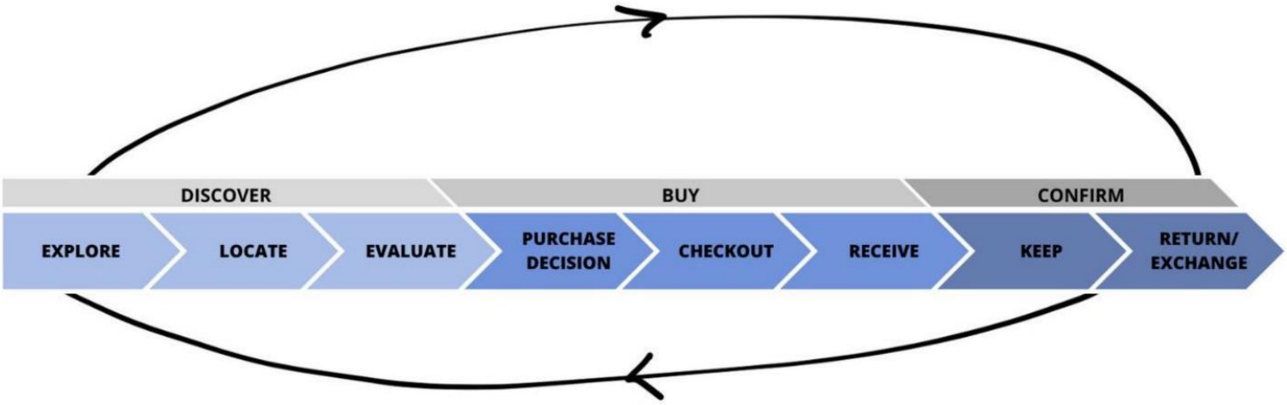


Figure 1.1: The customer journey and its eight stages.

The journey consists of eight key stages:

1. **Inspiration**: Customers are motivated to engage, sparking their interest.
2. **Exploration**: Customers browse the assortment, exploring brands and categories.
3. **Location**: Customers identify items of interest and locate the desired variant (e.g., size, color).
4. **Evaluation**: Customers assess the suitability of the item, considering style, quality, and reviews.
5. **Purchase Decision**: Customers decide whether to buy, assess urgency, and choose the timing of the purchase.
6. **Checkout**: Customers finalize the transaction, including applying discounts, selecting payment methods, and verifying memberships.
7. **Receiving Items**: Customers track delivery, receive notifications, and plan for parcel collection.
8. **Post-Purchase**: Customers decide whether to keep, return, or exchange the items, followed by potential refund processing.

1.2 Aims and objective of the work

1.2.1 Aim

The primary aim of this project is to develop and deploy a robust, data-driven predictive model for customer churn identification, leveraging advanced machine learning techniques.

The goal is to accurately detect customers who are at high risk of leaving the business by analyzing various factors such as customer demographics, usage patterns, transactional behavior, and engagement metrics. By gaining actionable insights into the drivers of churn, the model seeks to empower organizations to implement proactive and personalized retention strategies that address the specific needs and concerns of at-risk customers.

Additionally, the project aims to optimize resource allocation by prioritizing intervention efforts, reduce overall customer attrition rates, and enhance customer lifetime value (CLV). This initiative also seeks to contribute to long-term business growth by fostering stronger customer relationships, improving customer satisfaction, and maintaining a competitive edge in the market.

1.2.2 Objectives

- **Algorithm Analysis**: Compare and analyze multiple classification algorithms such as logistic regression, random forests, isolation forest, XGBoost, SVM(Support Vector Machine).
- **Dataset Evaluation**: Use a standardized dataset(E-Commerce) to ensure consistency in comparison and reliability of results.
- **Performance Metrics**: Evaluate algorithms using metrics like accuracy, precision, recall, F1 score, and ROC-AUC.
- **Insight Generation**: Identify which algorithm performs best and under what conditions.
- **Practical Implications**: Provide recommendations for businesses on selecting suitable models for churn prediction based on the findings.

1.2.3 Current scope

- **Algorithm Comparison**: Focuses on the evaluation of multiple classification algorithms, including traditional and advanced models, to determine their effectiveness in predicting churn.
- **Dataset Standardization**: Uses a consistent dataset to ensure reliable and unbiased comparison

- **Performance Insights**: Provides a detailed analysis of algorithm performance across multiple metrics, enabling a comprehensive understanding of their strengths and limitations.
- **Practical Recommendations**: Offers actionable insights for businesses to improve customer retention strategies based on research findings.

1.2.4 Future scope

- **Scalability Analysis**: Extend the study to evaluate how algorithms perform on larger, more complex datasets.
- **Real-Time Implementation**: Investigate the feasibility of deploying the best-performing algorithms in real-time customer relationship management systems.
- **Algorithm Optimization**: Explore techniques like hyperparameter tuning and feature selection to further improve algorithm performance.
- **Cross-Industry Application**: Adapt and test the models for other industries such as healthcare, finance, and telecommunications to assess their versatility.
- **Incorporation of Advanced Techniques**: Experiment with deep learning models and ensemble techniques for more nuanced predictions.
- **Customer Behavior Insights**: Integrate additional behavioral and psychographic data to refine churn prediction models and gain deeper insights into customer decision-making.

1.3 Brief literature review

Customer churn prediction has been extensively studied, with significant contributions highlighting the use of various machine learning algorithms to address this problem. Key findings from existing literature include:

- **Transactional and Behavioral Data**: Studies by Reinartz and Kumar (2003) emphasize the importance of combining transactional and behavioral data to enhance churn prediction accuracy.

- **Classification Algorithms**: Research has demonstrated the efficacy of logistic regression, decision trees, and random forests as robust baseline models. More recent works advocate the use of ensemble methods and deep learning for better performance on large datasets.

- **Feature Importance**: Variables like recency, frequency, and monetary value (RFM) have consistently been identified as critical predictors of churn. Additionally, customer engagement metrics and support interaction data have gained attention in modern studies.
- **Evaluation Metrics**: Precision, recall, and F1 score are commonly used to evaluate the performance of churn prediction models. ROC-AUC has been highlighted as an effective measure for comparing classification models.
- **Emerging Trends**: The application of advanced neural networks and explainable AI (XAI) is gaining traction, allowing for more nuanced predictions and better interpretability of results.

1.4 Problem definition

This study aims for a better understanding of how non-transactional behavior of customers affects customer churn, by finding and investigating behavioral features that potentially have an impact on customer churn, and later developing a model to predict customer churn in a binary classification problem. The study then proceeds with investigating the importance of each of the behavioral features in order to make conclusions about what pain points exist in the customer journey. A pain point is a critical moment or experience in the customer journey, where the customer encounters an event that may cause them to feel dissatisfied, and that may result in them falling off their journey. By identifying and addressing pain points in the customer journey, the company can improve customer satisfaction and increase customer retention.

The thesis finishes off with making suggestions, based on the findings of this study, regarding what actions the company could take in order to enhance customer retention.

The research questions that this thesis aims to answer are:

•What is non-transactional behavior and how can it be measured?

• What are the pain points of the customer journey?

• Where in the customer journey are the pain points located?

• How does non-transactional behavior of customers affect customer churn?

• How can the the global retailer use the discovered information to prevent customer churn?

1.5 Plan of their work

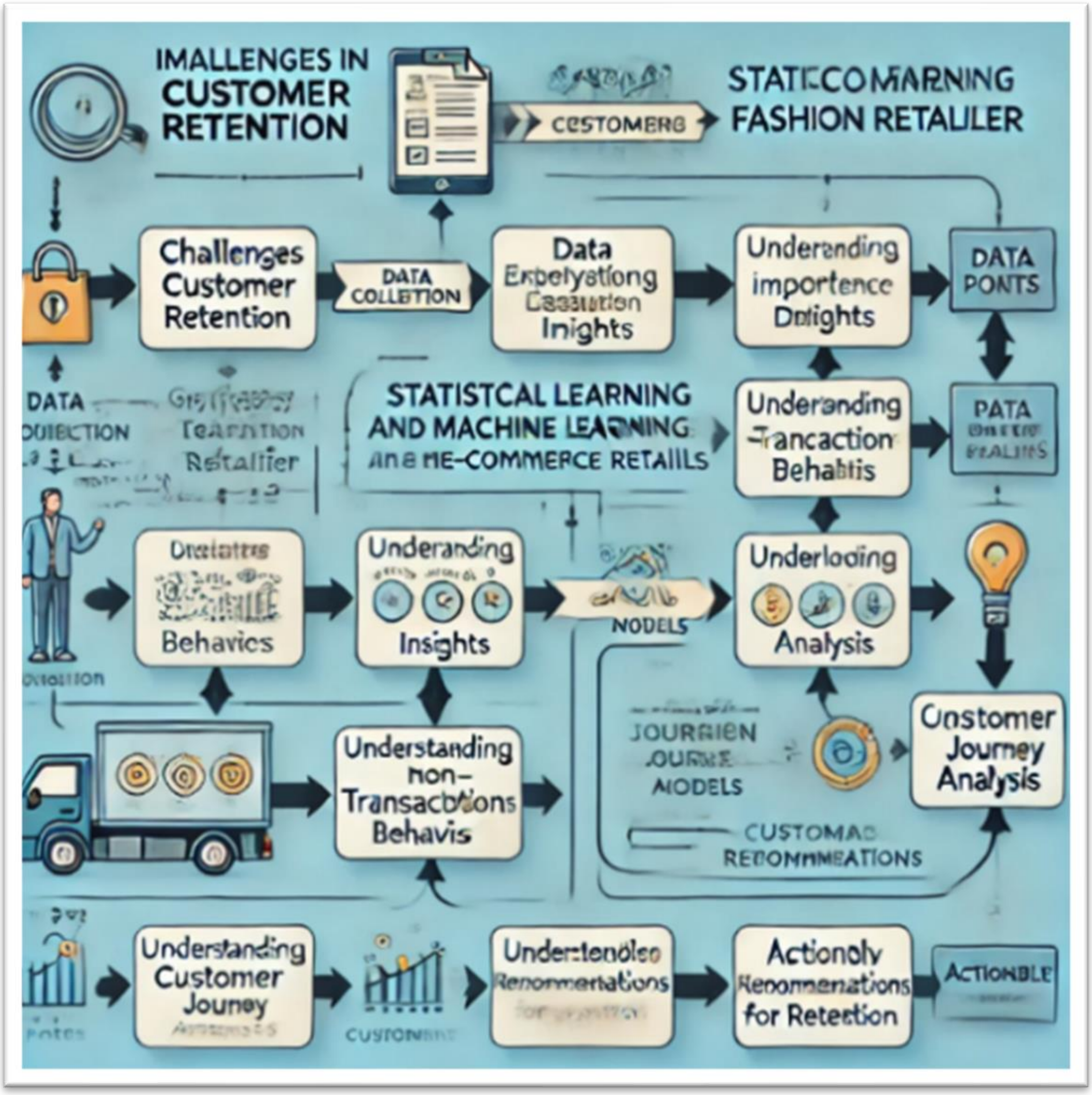

IMALLENGES IN CUSTOMER RETENTION
CESTOMERS
STATECOMARNING FASHION RETAULER
Challenges Customer Retention
DATA COLLETION
Data Erpelystiong Cassation Insights
Underanding importence Dnights
DATA PONTS
DATA
Tearntion
Retalifer
STATISTCAL LEARNING AND MACHINE LEARNING
ANE HE-COMMERCE RETAILS
Underanding Trancaction Behabis
PATA
Behavics
Underanding Insights
NODELS
Underloding Analysis
Understanding non-Transactbons Behavis
JOURRIEN OURKE MODELS
Onstomer Journey Analysis
CUSTOMAS REBONMMEATIONS
Understanding Customer Journey
CUSTONENT
Understendlee Renommentations
Actionly Renommenations for Reteetion
ACTIONBLE

Figure 1.2 – Plan of their work

Considering the challenges that an e-commerce fashion retailer faces, it is important to know not only how to make customers satisfied but how to make customers loyal and retain them. There have been cases where an increase in customer retention by 5% have led to profits being increased by over 100% (Buchanan & Gillies, 1990). Although such cases might be rare, Sterne (2002) has documented cases where the cost to acquire customers is 3, 5 or even 20 times larger than the cost to retain them. Therefore, it is likely that focusing on increasing customer retention will positively affect profitability.

Extensive amounts of customer-data are being gathered digitally everyday, however the data has little value in itself. Using analysis to turn data into insights and later, turning these insights into actions is what creates value (van den Driest et al., 2016). Furthermore, there is a great potential in using statistical learning as a tool for analysis, in order to draw useful insights from data.

The purpose of this study is to develop a statistical model using machine learning techniques and to analyze feature importance. The goal is to identify and understand the pain points of the customer journey and, thereby, to gain a better understanding of how non-transactional customer behavior affects customer churn. The developed statistical model should help to determine which parts of 3 the customer journey are critical for customer retention and to provide a better understanding of customers' non-transactional behavior.

2. Technology and literature review

- Machine learning techniques for churn

  Prediction
- Improved balanced random forests
- Churn prediction using random forest
- Clustering and classification for loyalty, return and churn prediction

2.1 Machine learning techniques for churn prediction

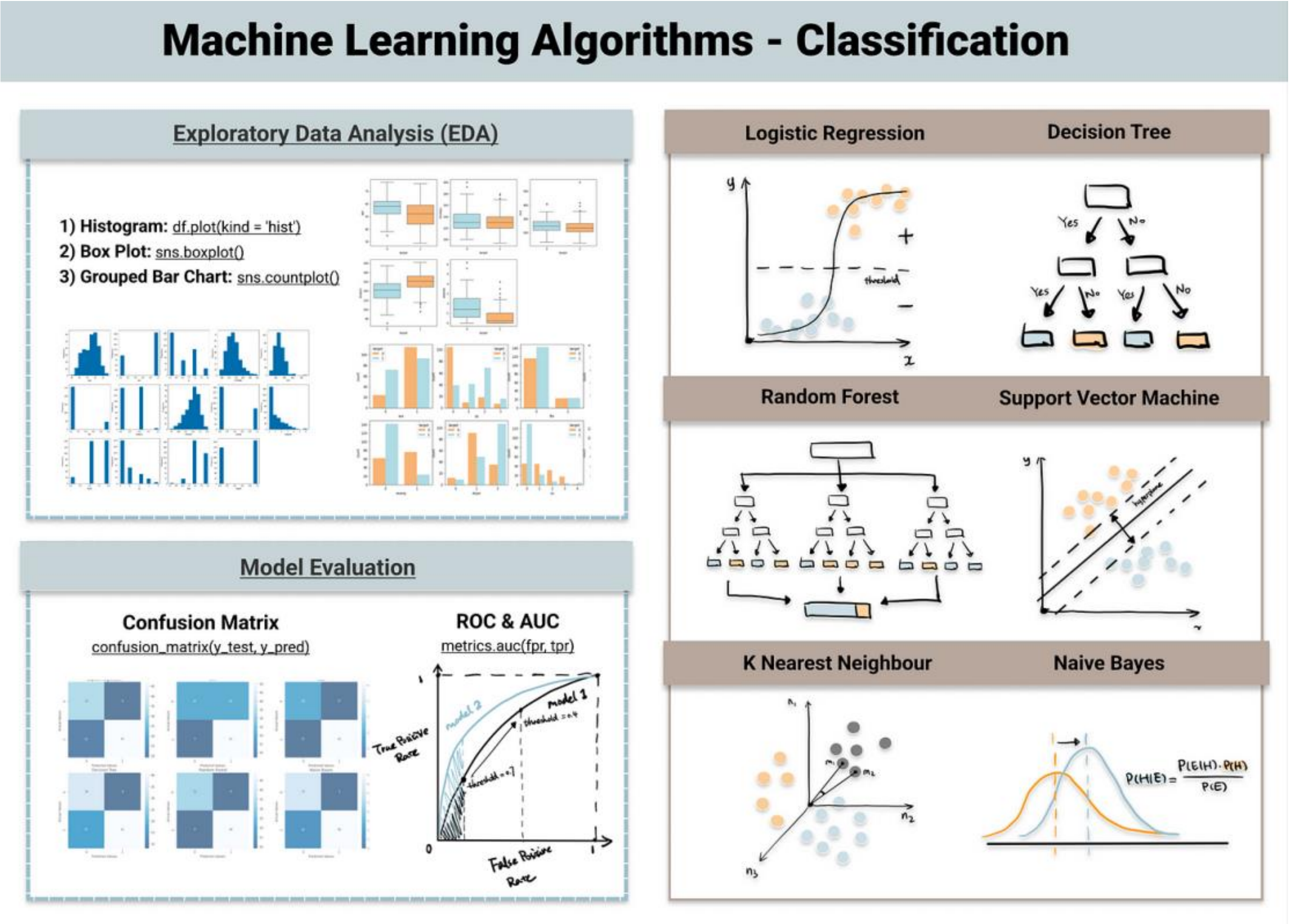


Figure 2.1 - ML Techniques for Churn Prediction

Various machine learning techniques are widely used for predicting customer churn, a critical aspect of Customer Relationship Management (CRM) strategies in fields like telecommunications and retail. Churn prediction models aim to identify early signals of customer churn, aiding organizations in retaining customers effectively.

Several machine learning techniques have been employed in churn prediction, including:

- **Decision Trees**: These generate classification rules by dividing datasets into smaller subsets. While effective in certain scenarios, decision trees may struggle with capturing complex, non-linear relationships.
- **Logistic Regression**: This method produces binary predictions based on multiple predictors. While commonly used, it performs well only after proper data transformations.

Advanced approaches have shown improved results over the years, particularly:

- **Random Forest**: An ensemble learning method based on decision trees, Random Forest has consistently delivered superior performance in churn prediction due to its ability to handle large datasets, reduce overfitting, and effectively capture complex relationships in data. Over time, it has been regarded as one of the most reliable techniques for churn prediction.
- **XGBoost**: A gradient-boosted decision tree algorithm known for its accuracy, efficiency, and scalability, especially in large-scale datasets. It often outperforms traditional methods by capturing intricate patterns in the data.
- **Isolation Forest**: A technique specializing in anomaly detection, which can identify atypical customer behaviors indicative of churn.

While traditional methods like decision trees and logistic regression have their place, **Random Forest** has emerged as a consistently strong performer over the years. When combined with modern boosting techniques like XGBoost or anomaly detection methods such as isolation forests, organizations can achieve even more robust churn prediction models. These advanced methods are particularly effective in capturing the complexities of customer behavior and delivering actionable insights.

2.2 Improved balanced random forests

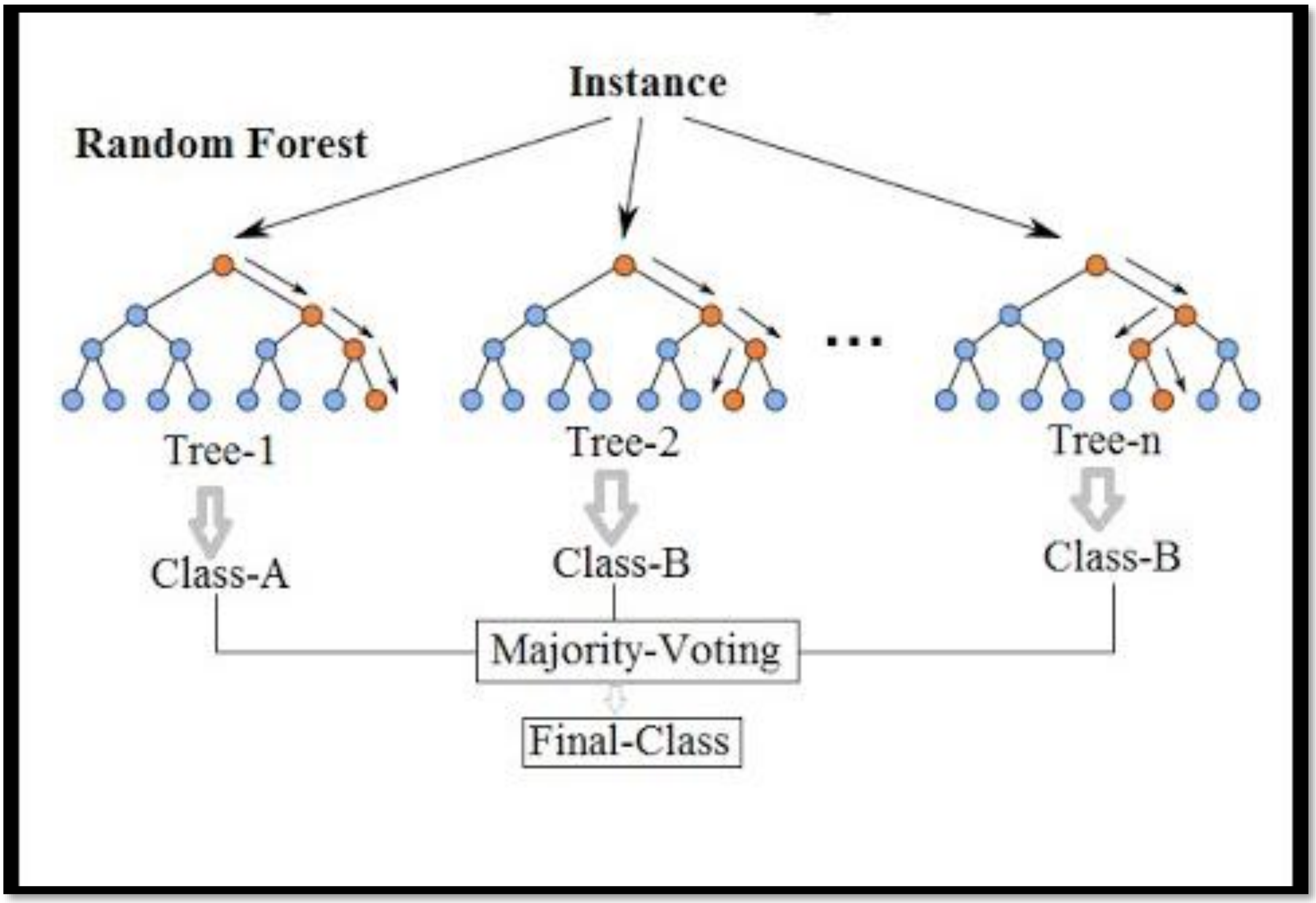


Figure 2.2 - Improved balanced random forests

The churn prediction problem has three significant characteristics:

1. **Data Imbalance**: The number of churners is often much smaller than non-churners.

2. **Data Noise**: Noise in the dataset can adversely affect the predictive performance.
3. **Customer Ranking**: Effective churn prediction requires ranking customers by their likelihood to churn.

Traditional approaches have been employed to address these challenges. For instance, decision-tree-based algorithms can rank customers, but their vulnerability to noisy data can reduce effectiveness. Neural networks, while often effective in churn prediction, lack interpretability, making it difficult to understand the rationale behind predictions. Additionally, genetic algorithms, although capable of producing accurate predictive models, struggle to quantify the likelihood associated with predictions.

To overcome these limitations, a method called **Improved Balanced Random Forests (IBRF)** has been proposed. IBRF integrates sampling techniques and cost-sensitive learning to enhance the classifier's performance, specifically for imbalanced datasets. IBRF builds upon earlier extensions of random forests, such as:

- **Weighted Random Forests**: Incorporates weights to handle imbalanced data.
- **Balanced Random Forests**: Ensures equal sampling from majority and minority classes.

IBRF combines these approaches by introducing "interval variables" to maintain a balanced distribution of classes while improving noise tolerance. Unlike standard balanced random forests, which can struggle with noise, IBRF heavily penalizes misclassification of the minority class and iteratively learns the most predictive features.

Evaluation of IBRF

IBRF was compared against standard methods, including decision trees, neural networks, and class-weighted support vector machines, using a dataset of approximately 1,500 customers, of whom 73 were churners. IBRF demonstrated superior performance, achieving an accuracy of 93.2% and excelling in metrics such as top-decile lift.

Further comparisons with other random forest variations, such as balanced and weighted random forests, revealed that IBRF outperformed these methods in distinguishing between churners and non-churners. Its scalability and reduced training time further highlight its potential for churn prediction tasks, making it a promising solution for classification challenges involving imbalanced datasets.

2.3 Churn prediction using random forest

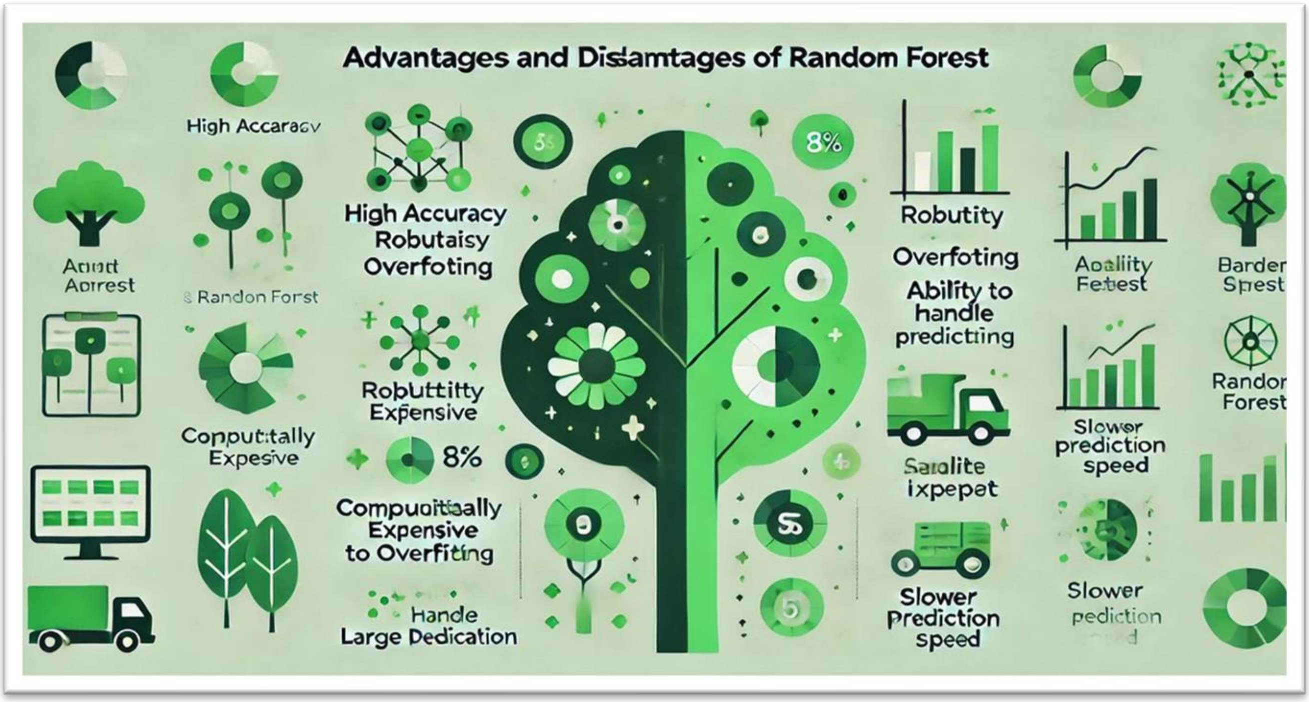


Figure 2.3 - Random forest

In the study conducted by Ullah et al. (2019), a churn prediction model is investigated, where classification and clustering techniques are applied to identify churn customers and determine the root causes of churn. The authors explain that finding churn factors (root causes) based on customer data and analyzing behavioral patterns allow Customer Relationship Management (CRM) to improve productivity as well as recommend and provide group-based retention offers.

The goal of the proposed churn prediction model is to identify churn customers and find the reasons behind churn to be able to improve customer retention, in the field of telecommunications. A problem that Ullah et al. (2019) found in existing churn models is that customers, or groups of customers, have different reasons for churn. Therefore, all churners should not be treated the same way. Once the data has been preprocessed, the proposed churn model classifies customers as churn or non-churn. The model then identifies factors behind churn. The last step of the proposed model is customer profiling through clustering. The outcomes of the proposed model are retention strategies.

There are 12 classification algorithms considered, namely (1) Random Forest, (2) Attribute Selected Classifier, (3) J48, (4) Random Tree, (5) Decision Stump, (6) AdaBoostM1 + Decision Stump, (7) Bagging + Random Tree, (8) Naïve Bayes, (9) Multilayer Perceptron, (10) Logistic Regression, (11) IBK, and (12) LWL. For clustering, k-means algorithms are considered.

The experimental results showed that the Random Forest algorithm performed the best in terms of accuracy and recall. This means that Random Forest is the best at correctly identifying churned customers. However, the Random Forest algorithm is time consuming, taking 108.48 seconds to build the model. Another algorithm that revealed to be time consuming is the Multilayer Perceptron, taking 214.18 seconds to build the model. The Multilayer Perceptron does, however, have a low accuracy (82.04%).

One algorithm that both performed well and has low time consumption is J48, which is a decision tree based algorithm. It has much lower time consumption than Random Forest (4.08 seconds) but with lower accuracy compared to the Random Forest algorithm.

The authors also evaluated how each classification algorithms based on a secondary data set, available for the public. Based on the results of both data sets, the authors found Random Forest to perform better than the other algorithms, because it achieves high performance for both data sets. Also, when building the churn model on the secondary data set, Random Forest had much lower time consumption (1.39 seconds).

Ullah et al. (2019) concluded in their study that churn prediction is a significant issue of CRM in order to improve customer retention, and that Random Forest or J48 are appropriate machine learning techniques for a churn model.

2.4 Clustering and classification for loyalty, return and churn prediction

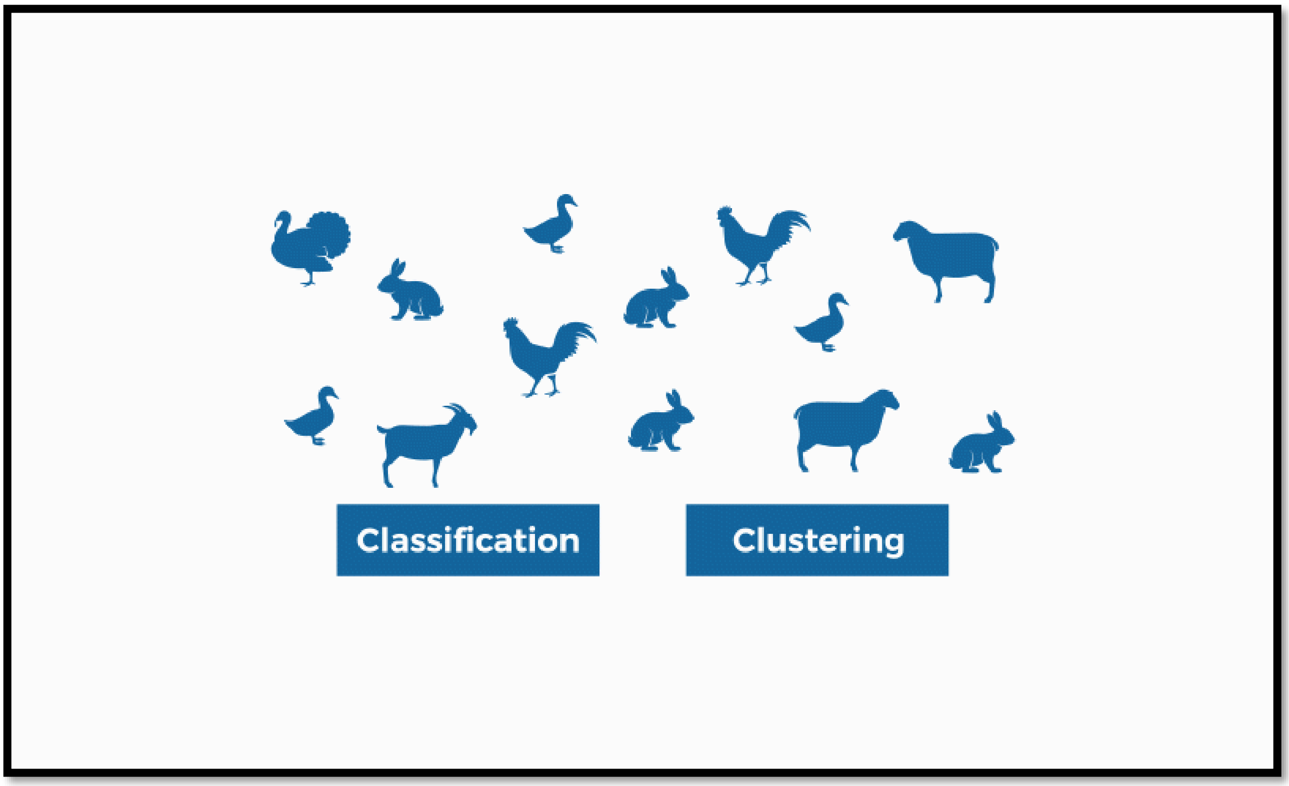


Figure 2.4 - Clustering and classification

Granov (2021) presented a method for loyalty, return and churn prediction, by using two different approaches. The aim of the study is to gain insights about customer behavior and how it affects purchases, returns and churn.

The first approach involves using the unsupervised machine learning method K-prototypes for clustering, to divide customers into different customer segments based on their behaviors. The second approach consists of three separate binary classification models, using the supervised classification technique Bias Reduced Logistic regression, which avoids the problem of overfitting by introducing a penalty in the log-likelihood function that is maximized when estimating the regression parameters. The binary classifiers classify the customers as Churners, Returners and Loyalists.

The K-prototypes algorithm successfully divided the customers into 6 segments defined as churned, potential, loyal, Brand Champions, indecisive shoppers and high-risky churners. The churn, return and loyalty classifiers accurately predicted 68%, 75% and 98% of the customers, respectively. Granov (2021) concluded that all models performed well based on their predictive accuracy, but also stated that data imbalance may cause this metric to not reflect the model’s true performance.

3. System Requirements Study

- **User Characteristics**
- **Hardware and Software Requirements**
- **Assumptions and Dependencies**

3.1 User Characteristics

- **Target Users**: This project targets:
    - **Data Scientists and Analysts**: To understand and implement customer churn prediction models.
    - **Business Decision Makers**: To leverage insights for customer retention strategies.
    - **E-commerce Platform Administrators**: To integrate the churn prediction system into customer relationship management (CRM) tools.
- **Skill Level**:
    - Users should have a basic understanding of data analysis and machine learning concepts.
    - Familiarity with tools like Python, Scikit-learn, and visualization libraries (e.g., Matplotlib, Seaborn) is beneficial.

3.2 Hardware and Software Requirements

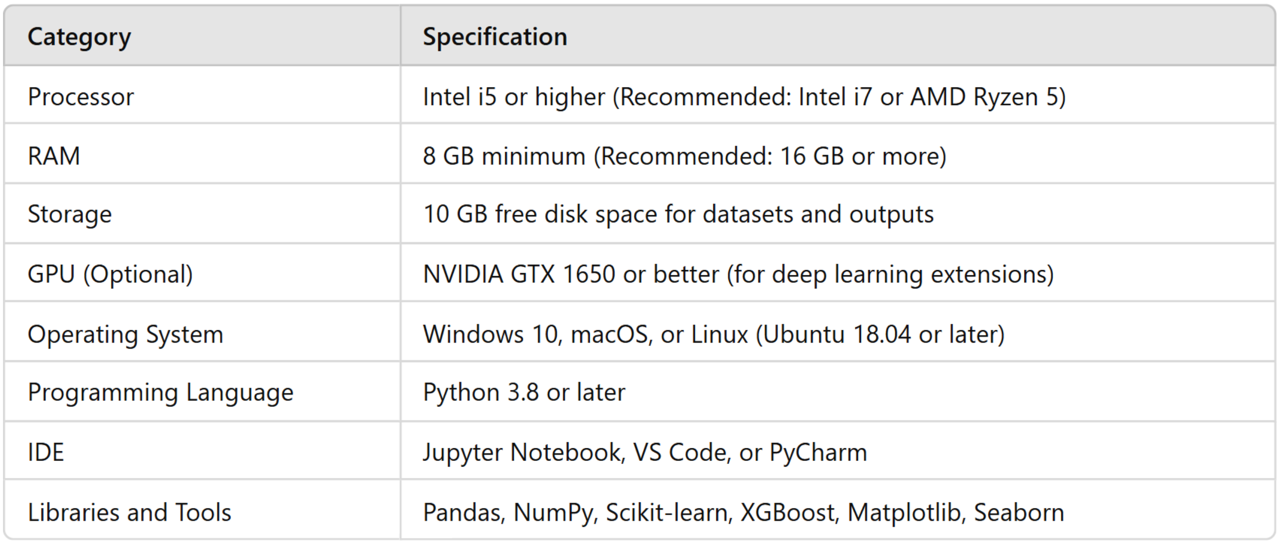
- **Hardware**:
    - Processor: Intel i5 or higher (recommended Intel i7 or AMD Ryzen 5)

  - RAM: 8 GB minimum (16 GB or more for large datasets)
  - Storage: 10 GB free disk space for dataset storage and model output.
  - GPU: (Optional) NVIDIA GTX 1650 or better for deep learning extensions.

- **Software**:

  - Operating System: Windows 10, macOS, or Linux (Ubuntu 18.04 or later).
  - Programming Language: Python 3.8 or later.
  - IDE: Jupyter Notebook, VS Code, or PyCharm.
  - Libraries and Tools:
    - Pandas, NumPy (Data manipulation)
    - Scikit-learn, XGBoost, Random Forest (Model development)
    - Matplotlib, Seaborn (Data visualization)
    - Isolation Forest (Anomaly detection)

Table 3.2 - System requirements



Category	Specification
Processor	Intel i5 or higher (Recommended: Intel i7 or AMD Ryzen 5)
RAM	8 GB minimum (Recommended: 16 GB or more)
Storage	10 GB free disk space for datasets and outputs
GPU (Optional)	NVIDIA GTX 1650 or better (for deep learning extensions)
Operating System	Windows 10, macOS, or Linux (Ubuntu 18.04 or later)
Programming Language	Python 3.8 or later
IDE	Jupyter Notebook, VS Code, or PyCharm
Libraries and Tools	Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn

3.3 Assumptions and Dependencies

- **Assumptions**:

  - The dataset provided is complete and sufficiently representative of the customer behavior for model training and validation.
  - Users have access to Python environments with all required libraries installed.
  - Data preprocessing steps, such as handling missing values and feature scaling, are completed prior to model training.

- **Dependencies**:
  - Internet connectivity for downloading additional libraries or updates.
  - Open-source libraries such as Scikit-learn and XG Boost for model building.

4. System Diagrams

- **Flow chart**
- **Use case diagram**
- **Sequence diagram**
- **Data flow diagram**
- **Architecture Diagram**

4.1 Flow chart

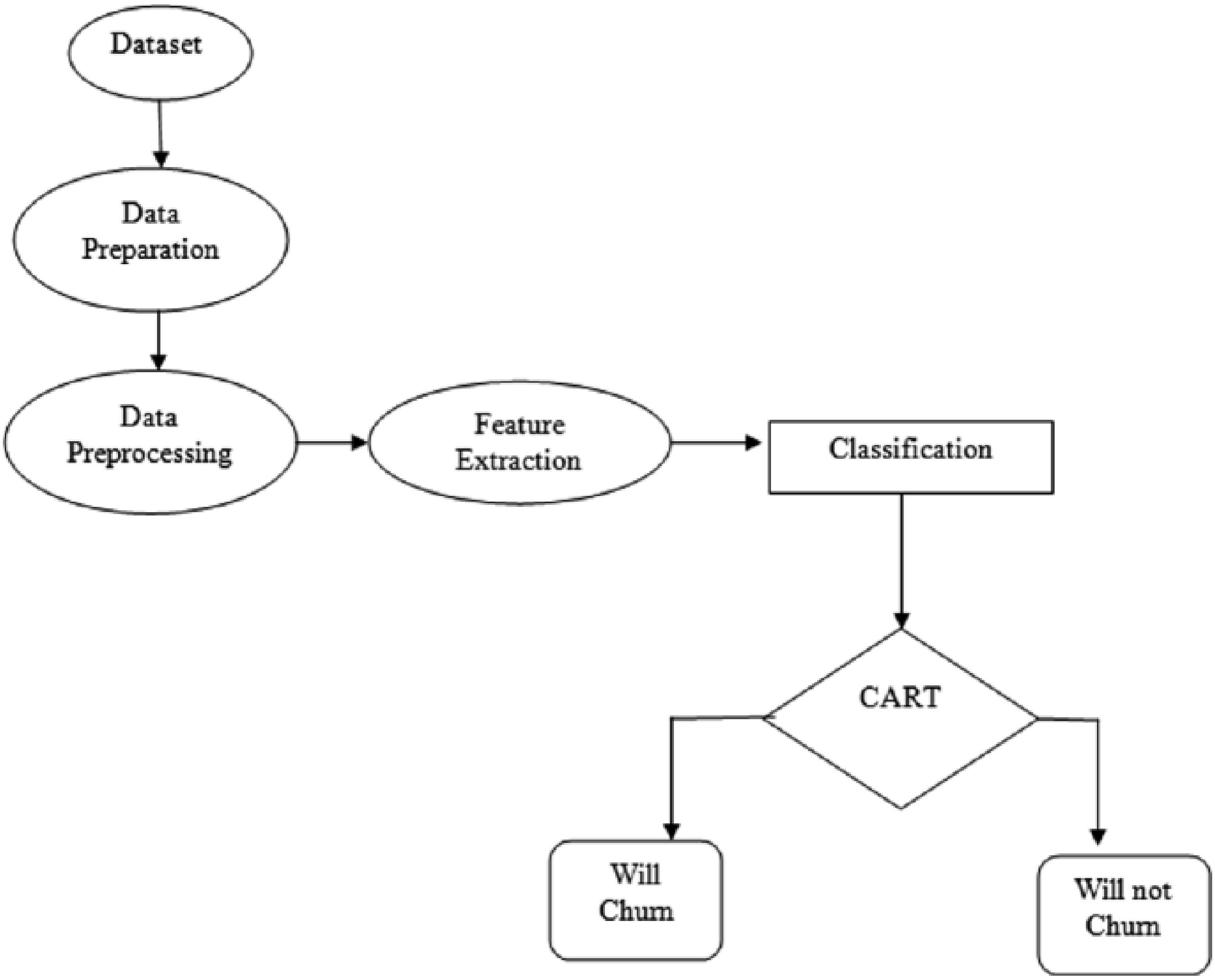


Figure 4.1 – Customer Churn or not

The Flow Chart outlines the step-by-step process of churn prediction. It begins with identifying the goal of predicting churn, followed by collecting customer data, such as transactional and behavioral metrics. The data is then preprocessed to handle missing values, normalize scales, and remove outliers. After preprocessing, relevant features are selected, such as usage patterns, satisfaction scores, and complaints. Machine learning models like Random Forest or Logistic Regression are trained on the data, and their performance is evaluated using metrics like accuracy, precision, recall, and AUC. Once the model is validated, it is deployed in production for real-time predictions, monitored regularly for performance, and retrained as needed.

4.2 Use case diagram

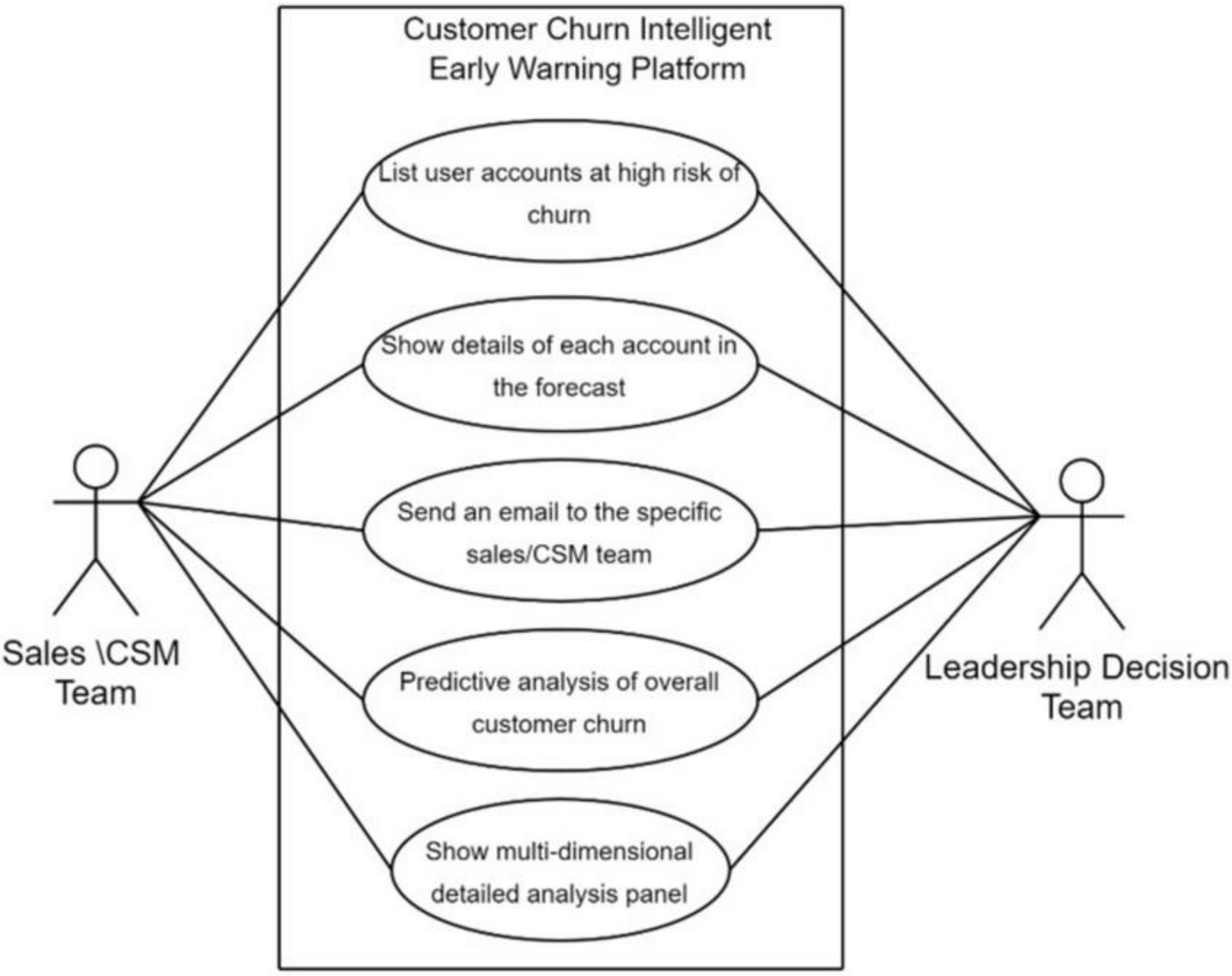

Customer Churn Intelligent Early Warning Platform
List user accounts at high risk of churn
Show details of each account in the forecast
Send an email to the specific sales/CSM team
Predictive analysis of overall customer churn
Show multi-dimensional detailed analysis panel
Sales \CSM Team
Leadership Decision Team

Figure 4.2 – Customer Churn

The Use Case Diagram highlights the interactions between the system and its users, including business analysts, data scientists, and the customer retention team. The business analyst uploads customer data and views churn risk reports, while the data scientist is responsible for building and refining the prediction model. The customer retention team uses the predictions to implement retention strategies, such as personalized offers or targeted communication, to prevent churn.

4.3 Sequence diagram

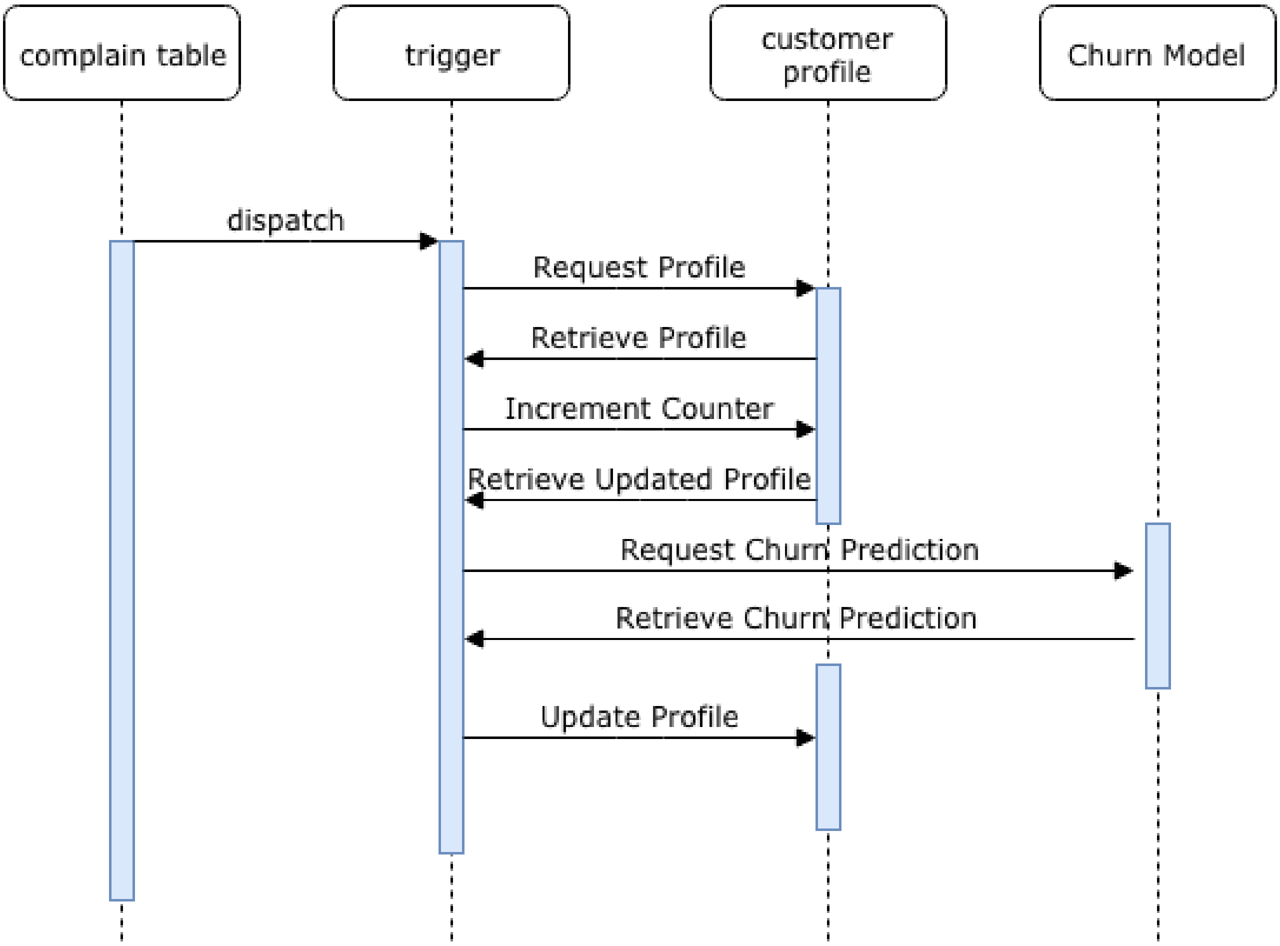


Figure 4.3 – Sequence diagram

The Sequence Diagram shows the flow of interactions within the system components. The user uploads customer data through a web interface, which the backend system processes. The backend sends the data to a model server for churn prediction. Once predictions are made, the backend retrieves the results, stores them in a database, and generates a churn risk report displayed to the user.

4.4 Data flow diagram

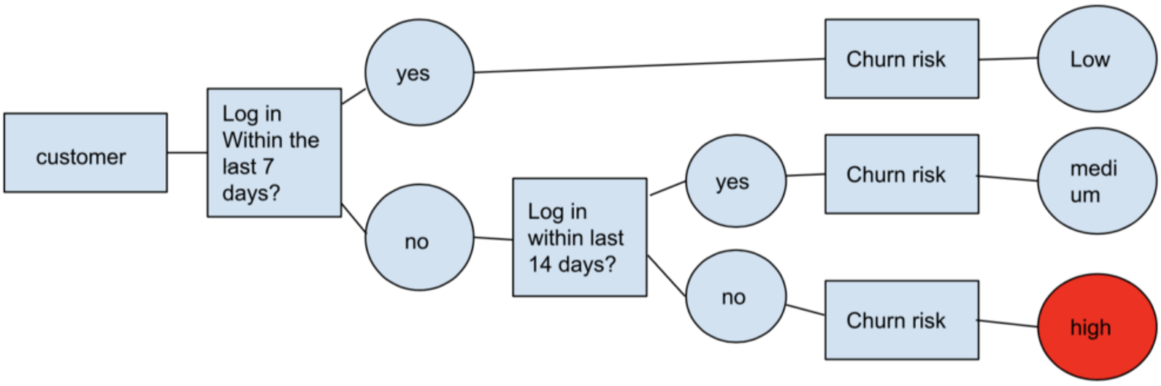


Figure 4.4 – DFD Level 0

The Data Flow Diagram (DFD) represents the flow of data through the system. In the Level 0 context diagram, external entities such as customer databases provide input data, and the system outputs churn risk predictions and actionable insights. In the Level 1 DFD, the processes include collecting and preprocessing data, applying machine learning algorithms, generating churn risk predictions, and continuously updating the model with new data to improve its performance.

4.4 Architecture Diagram

The Architecture Diagram describes the system's structure, consisting of multiple layers. The data layer handles the collection and storage of customer data from various sources like CRM systems and logs. The processing layer cleans and transforms the data into usable formats. The model layer contains machine learning models that predict churn. The application layer provides dashboards and APIs for accessing predictions and reports. A monitoring layer tracks the model's performance to ensure continuous improvements, and a storage layer securely stores data, features, and prediction results.

Together, these diagrams provide a comprehensive visualization of the customer churn prediction system, enabling efficient implementation and effective decision-making.

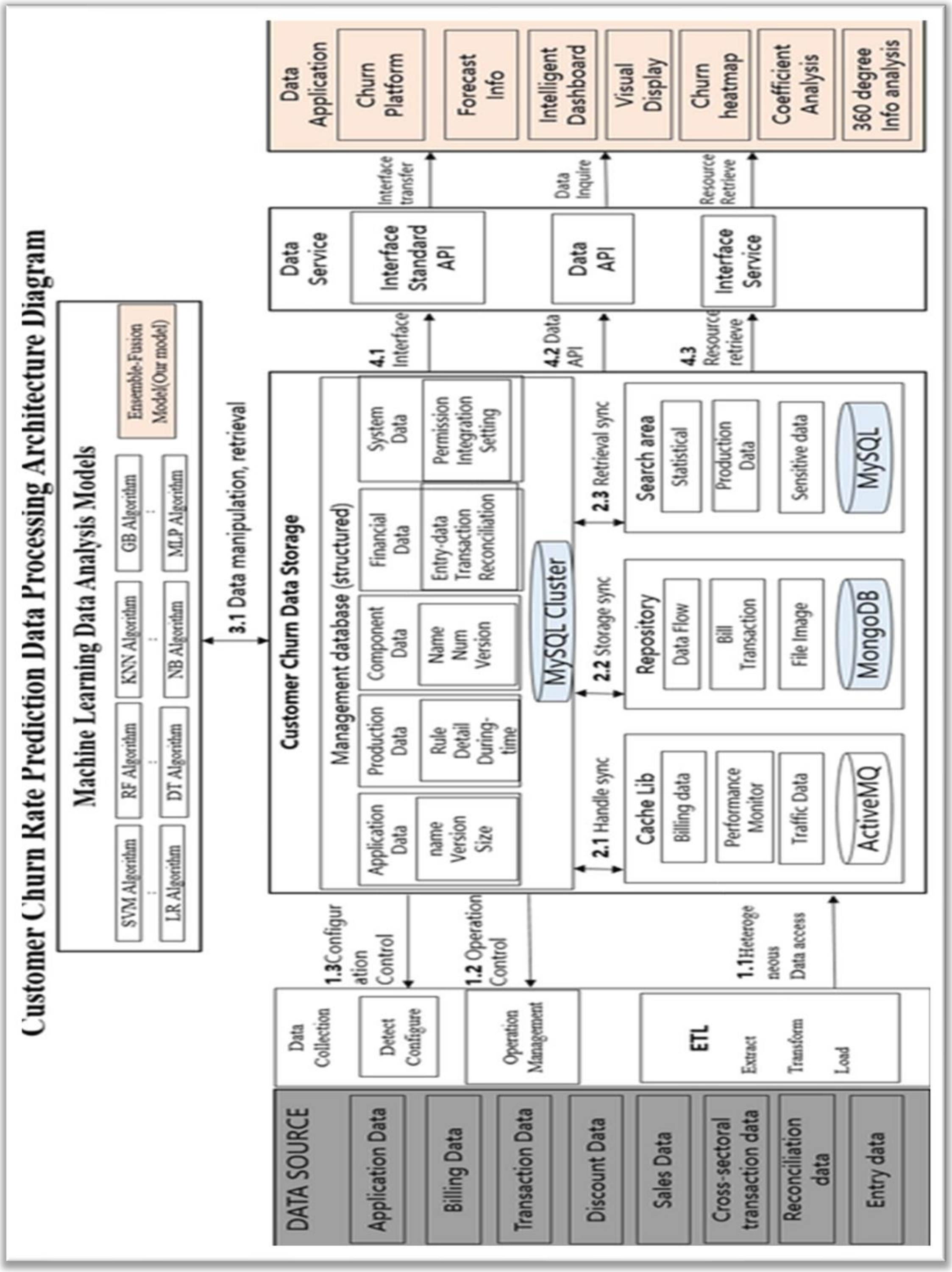
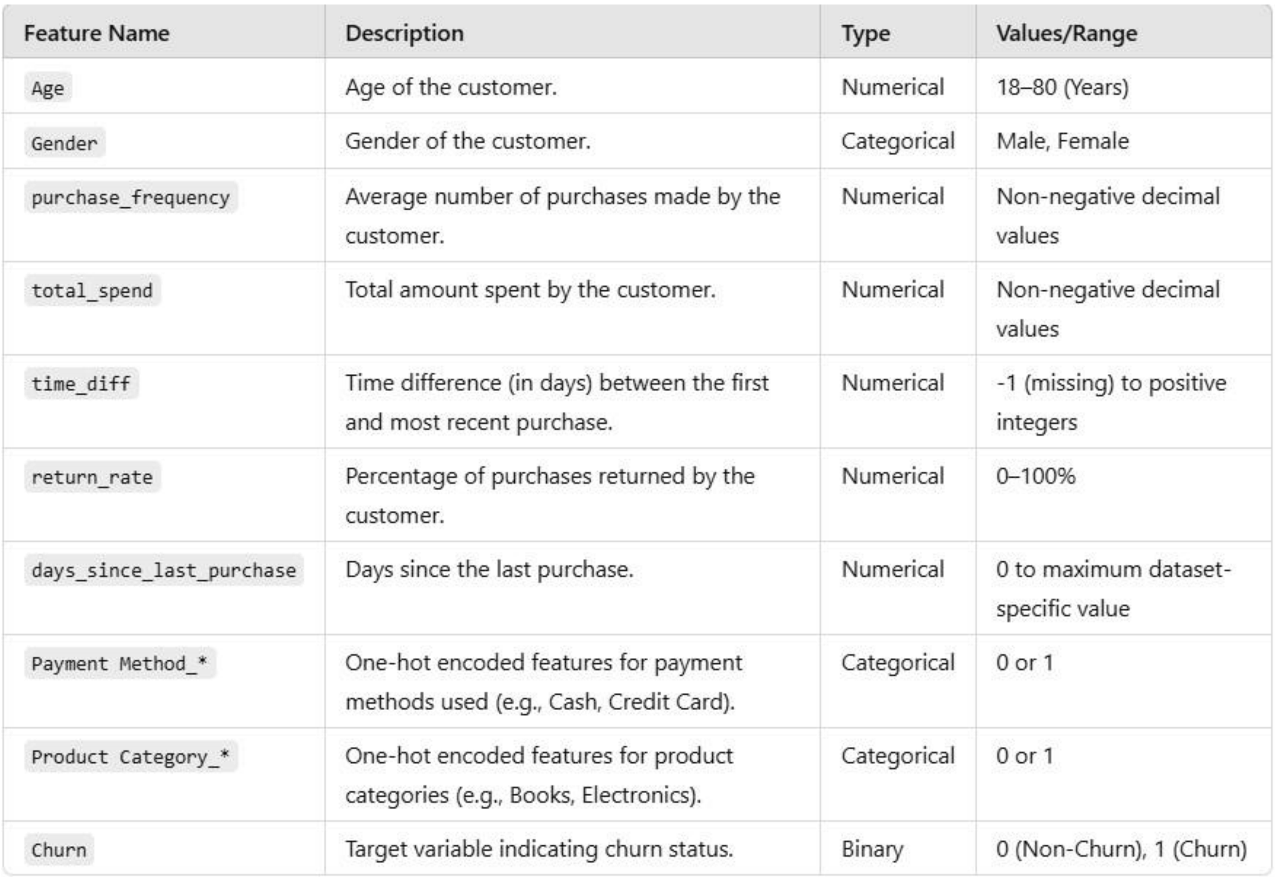


Figure 4.5 – Architecture Diagram

5. Data Dictionary

Table 5.1 – Data Dictionary



Feature Name	Description	Type	Values/Range
`Age`	Age of the customer.	Numerical	18–80 (Years)
`Gender`	Gender of the customer.	Categorical	Male, Female
`purchase_frequency`	Average number of purchases made by the customer.	Numerical	Non-negative decimal values
`total_spend`	Total amount spent by the customer.	Numerical	Non-negative decimal values
`time_diff`	Time difference (in days) between the first and most recent purchase.	Numerical	-1 (missing) to positive integers
`return_rate`	Percentage of purchases returned by the customer.	Numerical	0–100%
`days_since_last_purchase`	Days since the last purchase.	Numerical	0 to maximum dataset-specific value
`Payment Method_*`	One-hot encoded features for payment methods used (e.g., Cash, Credit Card).	Categorical	0 or 1
`Product Category_*`	One-hot encoded features for product categories (e.g., Books, Electronics).	Categorical	0 or 1
`Churn`	Target variable indicating churn status.	Binary	0 (Non-Churn), 1 (Churn)

6. Result, Discussion, and Conclusion

6.1 Results

Model Performance

- **Random Forest**:
  - Precision: 0.99 (Non-churners), 0.95 (Churners)
  - Recall: 0.99 (Non-churners), 0.96 (Churners)
  - F1-Score: 0.99 (Non-churners), 0.96 (Churners)
  - Accuracy: 0.98
  - AUC: 0.99
- **XGBoost**:
  - Precision: 0.92 (Non-churners), 0.44 (Churners)
  - Recall: 0.77 (Non-churners), 0.73 (Churners)
  - F1-Score: 0.83 (Non-churners), 0.55 (Churners)
  - Accuracy: 0.76
  - AUC: 0.83
- **Isolation Forest**:
  - Precision: 0.80 (Non-churners), 0.20 (Churners)
  - Recall: 0.88 (Non-churners), 0.12 (Churners)
  - F1-Score: 0.84 (Non-churners), 0.15 (Churners)
  - Accuracy: 0.73
  - AUC: 0.50
- **Ensemble Learning**:
  - Precision: 0.91 (Non-churners), 0.83 (Churners)
  - Recall: 0.97 (Non-churners), 0.61 (Churners)

  - F1-Score: 0.94 (Non-churners), 0.70 (Churners)
  - Accuracy: 0.90
  - AUC: 0.92

- **Logistic Regression**:

  - Precision: 0.79 (Non-churners), 0.72 (Churners)
  - Recall: 0.82 (Non-churners), 0.76 (Churners)
  - F1-Score: 0.80 (Non-churners), 0.73 (Churners)
  - Accuracy: 0.80
  - AUC: 0.76

Feature Importance

- Key features influencing churn prediction:

  - purchase_frequency
  - total_spend
  - time_diff
  - days_since_last_purchase

6.2 Discussion

Model Performance Analysis

The performance of different machine learning models for predicting customer churn is analyzed using various metrics. The models include Random Forest, XG Boost, Isolation Forest, Ensemble Learning, and Logistic Regression. Below is a detailed discussion of the results:

1. Random Forest

- **Metrics Overview:**

  - **Precision:** 0.99 (Non-churners), 0.95 (Churners)

    - **Recall:** 0.99 (Non-churners), 0.96 (Churners)
    - **F1-Score:** 0.99 (Non-churners), 0.96 (Churners)
    - **Accuracy:** 0.98
    - **AUC:** 0.99
- **Discussion:**
    - **Strengths:**
        - Random Forest outperforms other models in almost every metric, especially with high precision, recall, and F1-scores for both churners and non-churners.
        - The high AUC score (0.99) indicates that the model is excellent at distinguishing between churned and non-churned customers.
        - Its ability to handle non-linear relationships and feature importance analysis makes it the most robust model.
    - **Weaknesses:**
        - Despite its high accuracy, it may require higher computational power for training and prediction.

2. XG Boost

- **Metrics Overview:**
    - **Precision:** 0.92 (Non-churners), 0.44 (Churners)
    - **Recall:** 0.77 (Non-churners), 0.73 (Churners)
    - **F1-Score:** 0.83 (Non-churners), 0.55 (Churners)
    - **Accuracy:** 0.76
    - **AUC:** 0.83

- **Discussion:**
  - **Strengths:**
    - XG Boost performs reasonably well in recall for the churn class (0.73), making it suitable for identifying churners.
    - The model is scalable and efficient, particularly for large datasets.
  - **Weaknesses:**
    - Precision for churners (0.44) is relatively low, indicating a high rate of false positives.
    - AUC (0.83) and overall accuracy (0.76) are significantly lower compared to Random Forest, highlighting limited performance.

3. Isolation Forest

- **Metrics Overview:**
  - **Precision:** 0.80 (Non-churners), 0.20 (Churners)
  - **Recall:** 0.88 (Non-churners), 0.12 (Churners)
  - **F1-Score:** 0.84 (Non-churners), 0.15 (Churners)
  - **Accuracy:** 0.73
  - **AUC:** 0.50
- **Discussion:**
  - **Strengths:**
    - Isolation Forest is designed for anomaly detection and works well in detecting non-churners with a high recall (0.88).

- **Weaknesses:**
    - The model struggles significantly with churn detection, evidenced by low precision (0.20) and recall (0.12) for churners.
    - AUC of 0.50 indicates that the model performs no better than random guessing for distinguishing churners.

4. Ensemble Learning

- **Metrics Overview:**
    - **Precision: 0.91 (Non-churners), 0.83 (Churners)**
    - **Recall: 0.97 (Non-churners), 0.61 (Churners)**
    - **F1-Score: 0.94 (Non-churners), 0.70 (Churners)**
    - **Accuracy: 0.90**
    - **AUC: 0.92**
- **Discussion:**
    - **Strengths:**
        - Ensemble Learning combines the strengths of multiple models, achieving a good balance across all metrics.
        - Precision (0.83) and F1-score (0.70) for churners are higher compared to most models except Random Forest.
        - AUC (0.92) highlights strong overall classification performance.
    - **Weaknesses:**
        - The recall for churners (0.61) is lower compared to other models, which might result in some churned customers being missed.

5. Logistic Regression

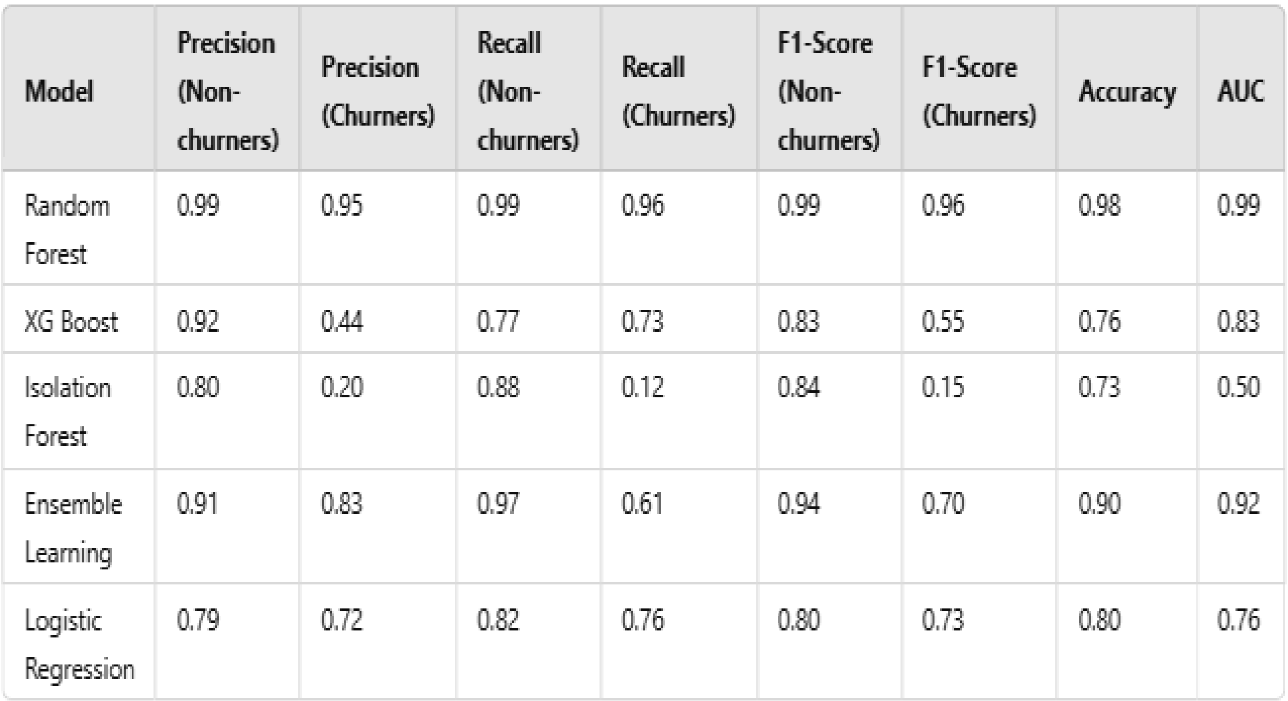
- **Metrics Overview:**
  - **Precision:** 0.79 (Non-churners), 0.72 (Churners)
  - **Recall:** 0.82 (Non-churners), 0.76 (Churners)
  - **F1-Score:** 0.80 (Non-churners), 0.73 (Churners)
  - **Accuracy:** 0.80
  - **AUC:** 0.76

- **Discussion:**
  - **Strengths:**
    - Logistic Regression provides a baseline for model comparison with balanced performance across metrics.
    - It is computationally efficient and easy to interpret.
  - **Weaknesses:**
    - The accuracy (0.80) and AUC (0.76) are lower compared to ensemble models and Random Forest, suggesting limited performance in distinguishing churners.
    - The model may not capture complex non-linear relationships in the data.

Performance Visualization

Below is a table summarizing the key metrics for all models:

Table 6.1 - Performance Visualization



Model	Precision (Non-churners)	Precision (Churners)	Recall (Non-churners)	Recall (Churners)	F1-Score (Non-churners)	F1-Score (Churners)	Accuracy	AUC
Random Forest	0.99	0.95	0.99	0.96	0.99	0.96	0.98	0.99
XG Boost	0.92	0.44	0.77	0.73	0.83	0.55	0.76	0.83
Isolation Forest	0.80	0.20	0.88	0.12	0.84	0.15	0.73	0.50
Ensemble Learning	0.91	0.83	0.97	0.61	0.94	0.70	0.90	0.92
Logistic Regression	0.79	0.72	0.82	0.76	0.80	0.73	0.80	0.76

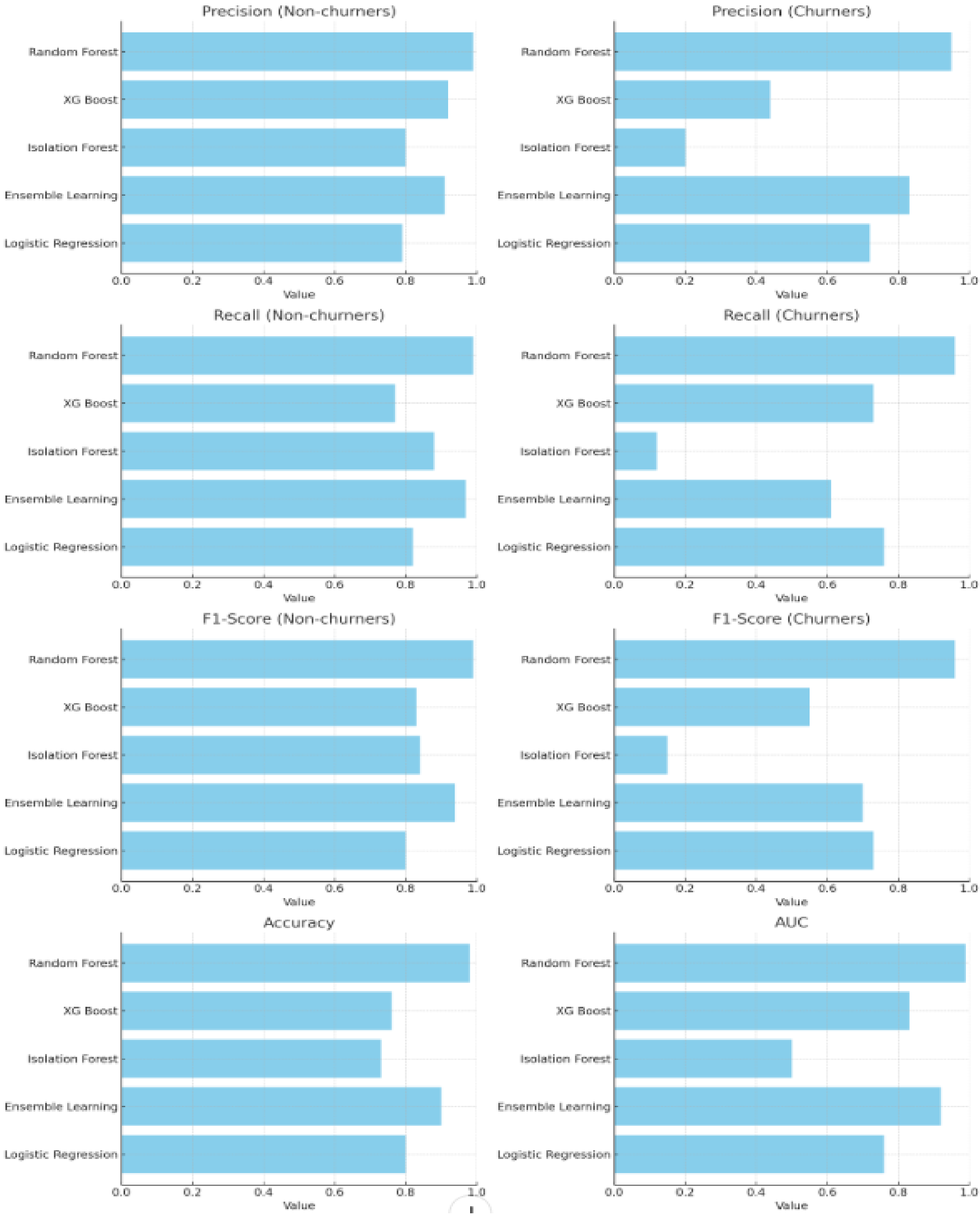


Figure 6.2 – Performance Visualization

6.3 Conclusion

- **Best Model**: Random Forest delivers the best performance across all metrics, making it the ideal choice for churn prediction.
- **Alternative Model**: Ensemble Learning provides a balanced performance and could be used as a fallback option.
- **Improvements Needed**: Models like XG Boost and Isolation Forest require further optimization and tuning to improve their ability to identify churners.
- **Practical Recommendation**: While Random Forest is highly accurate, Ensemble Learning may be used when computational efficiency is a concern, as it provides near-optimal performance with reduced complexity.

7. References

[1] Saran Kumar A,Chandrakala D "A Survey on Customer Churn Prediction using Machine Learning Techniques" November 2016,International Journal of ComputerApplications,154(10):13-16,DOI:10.5120/ijca2016912237 volume Article number: 28 (2019)

[2] Abdelrahim Kasem Ahmad, Assef Jafar & Kadan Aljoumaa "Customer churn prediction in telecom using machine learning in big data platform" Journal of Big Data, Article number: 28 (2019)

[3] Damandeep Singh , Vansh ,Dr. M. Kanchana, Associate Professor, "Survey Paper on Churn Prediction on Telecom", SRM, India

[4] Nasebah Almufadi , Ali Mustafa Qamar, Rehan Ullah Khan, Mohamed Tahar Ben Othman, "Deep Learning-based Churn Prediction of Telecom Subscribers", International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 2743-2748

[5] N Lakshmi Kalyani and Kolla Bhanu Prakash, "Soil Color as a Measurement for Estimation of Fertility using Deep Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 13(5), 2022.

[6] Kriti, "Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning", Iowa State University Ames, Iowa 2019

[7] Adnan Amin , Feras Al-Obeidat , Babar Awais Adnan , Jonathan Loo , Sajid Anwar, "Customer churn prediction in telecommunication industry under uncertain situation", Center for Excellence in Information Technology, Institute of Management Sciences, Peshawar, 25000 Pakistan. College of Technological Innovation, Zayed University,144534 Abu Dhabi, United Arab Emirates. Computing and Communication Engineering, University West London.

[8] Pronay Ghosh, "Project report on customer churn prediction using supervised machine learning".