***CUSTOMER PURCHASE BEHAVIOR ANALYSIS AND LOYALTY PREDICTION***

*A Internship project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**INTERNSHIP**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**APARNA RAMESH A**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Nov 2024**

**List of Abbreviations**

1. **Age:** Age of the customer
2. **Items Purchased:** Number of items purchased in a single transaction
3. **Total Spent:** Total amount spent on the transaction
4. **Discount (%):** Percentage discount on the purchase
5. **Satisfaction Score:** Self-reported customer satisfaction score
6. **Warranty Extension:** Whether the customer opted for warranty extension
7. **Gender:** Gender of the customer
8. **Region:** Customer's region
9. **Product Category:** Category of the purchased product
10. **Payment Method:** Payment method used (e.g., UPI, Cash)
11. **Revenue:** Total revenue from the customer
12. **Store Rating:** Rating given by the customer to the store
13. **Loyalty Score:** Customer's loyalty score
14. **Membership Status:** Whether the customer is a member
15. **Preferred Visit Time:** Preferred time of visit

# Table of Contents

[Table of Contents 3](#_Toc184069526)

[Abstract 4](#_Toc184069527)

[Problem Definition 5](#_Toc184069528)

[Problem Statement 5](#_Toc184069529)

[Objective 5](#_Toc184069530)

[Introduction 6](#_Toc184069531)

[Methodology 7](#_Toc184069532)

[Data Acquisition: 7](#_Toc184069533)

[Data Preprocessing: 9](#_Toc184069535)

[Data Cleaning: 9](#_Toc184069536)

[Exploratory Data Analysis (EDA): 16](#_Toc184069537)

[Discussion 33](#_Toc184069538)

[Limitations: 33](#_Toc184069539)

[Future Work: 34](#_Toc184069540)

[Implications: 35](#_Toc184069541)

[Conclusion 36](#_Toc184069542)

[Summary of Findings: 36](#_Toc184069543)

[Impact: 36](#_Toc184069544)

[Recommendations: 36](#_Toc184069545)

[References 38](#_Toc184069546)

# Abstract

This report details the development and deployment of a machine learning-powered Flask application to predict customer loyalty categories based on behavioral data. Key tasks included data preprocessing, feature engineering, model optimization using GridSearchCV, and application deployment.

A Random Forest Classifier was employed to achieve high accuracy in loyalty predictions. The model was integrated into a Flask web app with robust preprocessing pipelines, tested locally, and prepared for cloud deployment on platforms like pythonanywhere

This project demonstrates the seamless integration of data science and web development, offering practical solutions for customer behavior analysis while providing hands-on experience in real-world machine learning applications.

# Problem Definition

## Problem Statement

In the competitive retail sector, understanding customer behavior and fostering loyalty are vital for business success. However, predicting customer loyalty categories ("Low," "Medium," "High") remains challenging due to diverse demographics and behaviors.

This project aims to analyze customer purchase data using machine learning to classify loyalty levels, uncover behavioral patterns, and provide actionable insights for optimizing loyalty programs. A scalable, web-based application will be developed to automate predictions, enabling businesses to enhance customer engagement and satisfaction.

## Objective

The project aims to analyze customer data to build a machine learning model for predicting loyalty categories ("Low," "Medium," "High"). It involves developing a Flask-based application for seamless predictions, deploying it on a cloud platform, and providing actionable insights to enhance customer engagement and loyalty programs.

# 

# 

# 

# Introduction

Customer loyalty is a critical factor for businesses aiming to retain customers and ensure long-term success. By analyzing customer behaviour, such as spending patterns, satisfaction levels, and demographic details, companies can better understand their customers and tailor loyalty programs accordingly.

This project focuses on developing a predictive model to categorize customer loyalty into "Low," "Medium," or "High," using a dataset that includes various behavioral and transactional attributes. The model leverages machine learning techniques for accurate predictions, offering businesses a data-driven approach to enhance customer retention strategies.

The project culminates in a Flask-based web application, enabling real-time loyalty predictions based on user inputs. This integrated solution provides businesses with actionable insights, empowering them to design targeted marketing initiatives and foster stronger customer relationships. By deploying the application on a cloud platform, the project ensures accessibility and scalability for practical use in real-world scenarios.

# Methodology

## 

# Data Acquisition:

## The data acquisition phase involved obtaining a dataset containing detailed customer behavior and purchase information. This dataset served as the foundation for building a predictive model and covered various attributes essential for analyzing customer loyalty.

**1. Source of Data**

The dataset provided for this project was curated and included customer-related fields such as:

* Demographic information: Age, Gender, and Region.
* Purchase behavior: Total Spent, Items Purchased, and Discount (%).
* Store interactions: Satisfaction Score, Loyalty Score, and Store Rating.
* Membership details and preferred visit times.

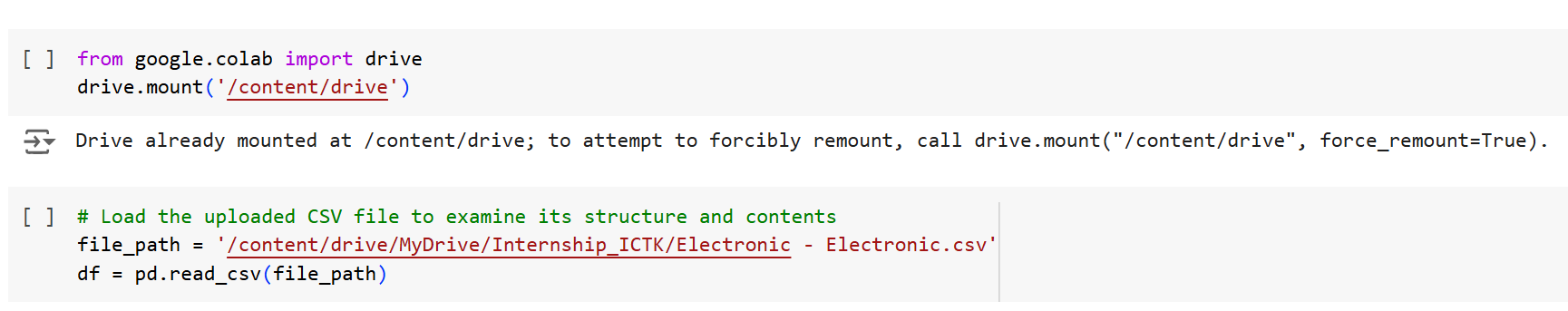
**2. Dataset Description**

Key features in the dataset:

* Behavioral Attributes: Satisfaction Score, Loyalty Score, and Preferred Visit Time.
* Transactional Details: Total Spent, Discount (%), Revenue, and Items Purchased.
* Demographics: Gender and Region.
* Categorical Features: Product Category, Payment Method, and Membership Status.

**3. Loading the Dataset**

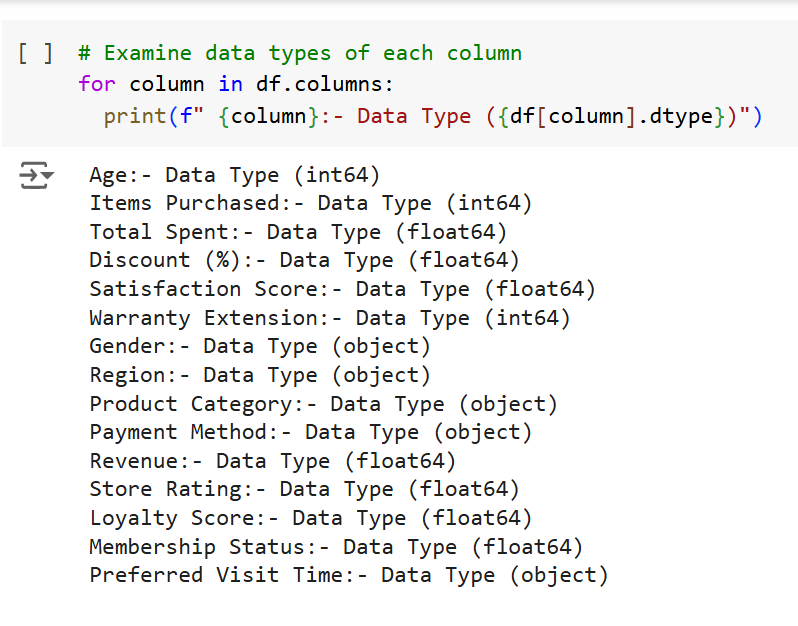
The dataset was loaded into a Python environment using the pandas library for further exploration and analysis:



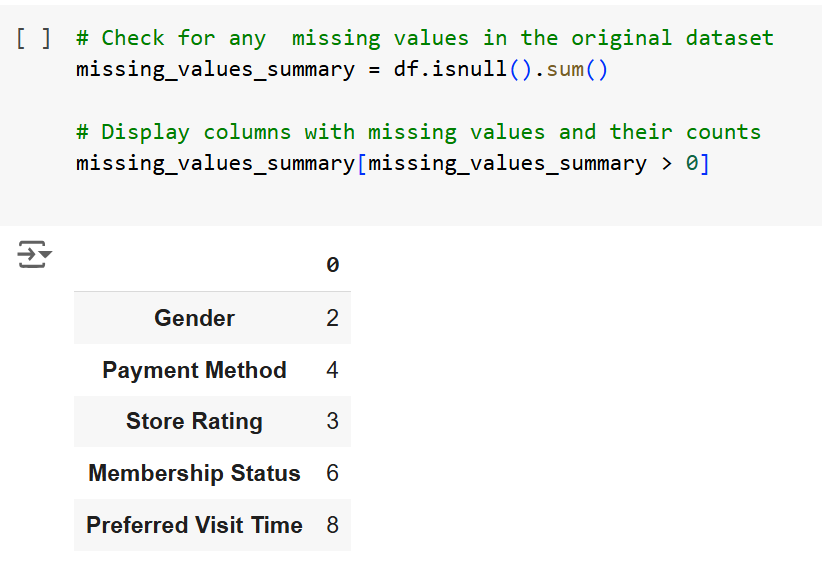
**4. Initial Inspection**

Performed the following steps to understand the dataset:

* **Structure:** Used .head(), .shape and .info() to review the data types, feature names, and initial rows.
* **Understand data types:**



* **Missing Values:** Identified null values using .isnull().sum().



* **Basic Statistics:** Applied .describe() to view basic statistical summaries of numerical features.

## Data Preprocessing:

Data preprocessing is a critical step in preparing the dataset for machine learning. It ensures that the data is clean, consistent, and ready for analysis and modeling. The following steps outline the preprocessing tasks performed for loyalty prediction.

### Data Cleaning:

Data cleaning was performed to remove inconsistencies, errors, and missing values. The following tasks were accomplished during this phase:

#### 

**1. Handling Missing Values**

Missing values were imputed based on the nature of the data:

* **Numerical Features**:
  + Store Rating: Filled with the mean value.
  + Membership Status: Filled with the median value.
* **Categorical Features**:
  + Columns like Gender, Payment Method, and Preferred Visit Time were filled using the mode (most frequent value).

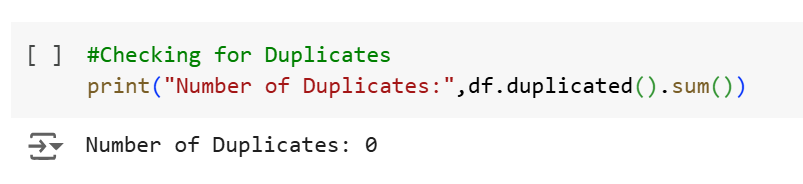
**2. Garbage Value Checking**

Garbage values in a dataset are out-of-scope or nonsensical values that do not align with the context of the data. Identifying and handling garbage values ensures data consistency and integrity. Here’s the approach to check for and handle garbage values:

* **Gender**:
  + Contains values: ['Male', 'Female', 'Other', nan].
  + **Action**: If "Other" is valid, retain it; otherwise, consider replacing it with nan. Fill missing (nan) values with the mode or remove rows depending on the context.
* **Region**:
  + Contains values: ['South', 'East', 'North', 'West'].
  + **Action**: No garbage values are detected here. Verify spelling and ensure consistency (e.g., no lowercase variations like south).
* **Product Category**:
  + Contains values: ['Accessories', 'Laptop', 'Tablet', 'Television', 'Mobile'].
  + **Action**: This seems clean; no immediate action is required.
* **Payment Method**:
  + Contains values: ['UPI', 'Cash', 'Credit Card', 'Net Banking', 'Debit Card', nan].
  + **Action**: Missing values (nan) can be imputed using the mode (most common payment method). No garbage values detected.
* **Preferred Visit Time**:
  + Contains values: ['Evening', 'Morning', 'Afternoon', nan].
  + **Action**: Handle missing (nan) values by imputing with the mode or using a placeholder like "Unknown."

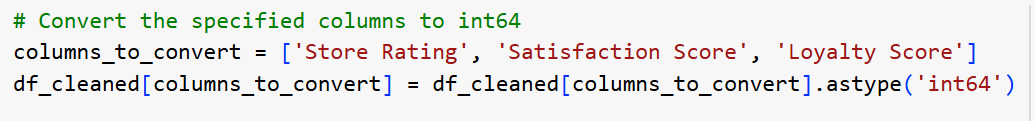
**3. Duplicate Entries**

Duplicate entries in a dataset can lead to biased results during analysis and modeling:

****

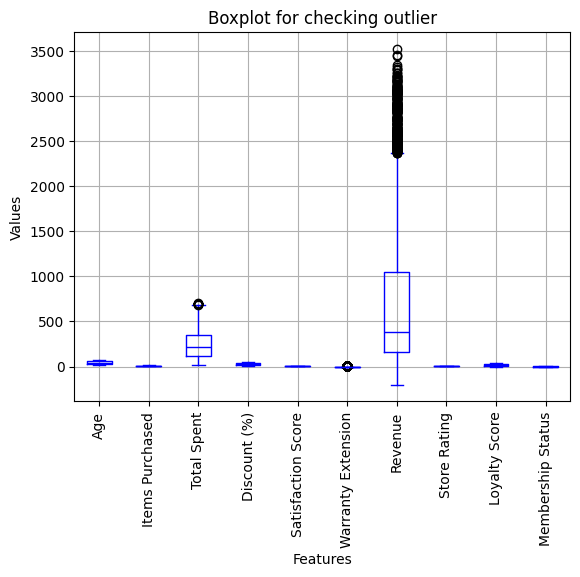
**4. Data Type Standardization**

Ensured numerical columns (Store Rating, Satisfaction Score, Loyalty Score) were converted to the int64 data type.



**5. Outlier Analysis and Interpretation**

Outliers are data points that differ significantly from other observations in the dataset. Identifying and interpreting outliers is essential for understanding the data and its impact on the predictive model. The boxplot provided highlights outliers in several features, and here is an explanation of the findings:



* **Total Spent**:
  + Outliers are observed, representing customers who have spent significantly more than the average.
  + **Interpretation**: These are genuine outliers that indicate high-value customers. These points are valuable for loyalty prediction, as they likely belong to highly loyal customers.
* **Revenue**:
  + Contains a high concentration of outliers in the upper range.
  + **Interpretation**: Similar to Total Spent, these outliers are genuine and reflect customers contributing significantly to revenue. These data points should be retained for modeling, as they are crucial for identifying high loyalty.
* **Warranty Extension**:
  + Exhibits higher values but is not considered an outlier based on the boxplot.
  + **Interpretation**: The higher values are legitimate and reflect customers opting for extended warranties, a normal pattern for some customer segments. This feature does not need special handling for outlier removal.
* **Other Features**:
  + Features like Age, Items Purchased, Satisfaction Score, Store Rating, and Loyalty Score show limited or no outliers.
  + **Interpretation**: These features are well-distributed, and no significant deviations are observed.

**Conclusion**

* **Retain Genuine Outliers**: Outliers in Total Spent and Revenue represent high-value customers and are critical for loyalty prediction. These values should not be removed but instead scaled or normalized to minimize their impact on the model.
* **No Action on Warranty Extension**: Despite its higher values, Warranty Extension is not an outlier and does not require additional processing.
* **Modelling Implications**:
  + Use techniques like Robust Scaler or Log Transformation to scale Total Spent and Revenue while preserving their importance.
  + Ensure that these outliers are included in training to improve the model's capability to predict high loyalty accurately.

This analysis confirms the importance of preserving key outliers for effective customer loyalty prediction while ensuring proper preprocessing to handle scale differences.

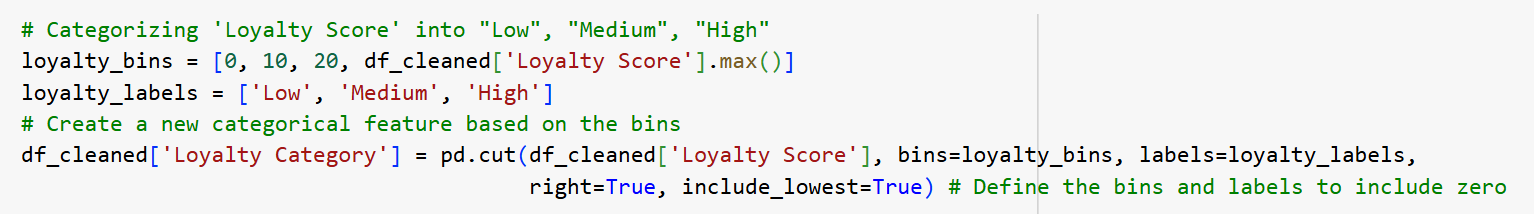
**Feature Engineering:**

**Target Variable Creation:**

The target variable in this project is the "Loyalty Category," derived from the original numerical Loyalty Score column. This step transforms a continuous variable into a categorical variable suitable for classification tasks. Here's how the target variable was created and prepared for modeling:

**1. Categorizing Loyalty Score**

* **Purpose**: Simplify the prediction task by categorizing the continuous loyalty scores into distinct categories based on predefined ranges.
  + The **Loyalty Score** was divided into three bins:
    - **Low Loyalty**: Customers with a score in the lowest range, indicating minimal loyalty.
    - **Medium Loyalty**: Customers with a moderate loyalty score.
    - **High Loyalty**: Customers with the highest loyalty scores, signifying strong loyalty.
  + **Ranges Used**:
    - Low: Loyalty Score in the range 0–10.
    - Medium: Loyalty Score in the range 11–20.
    - High: Loyalty Score above 20.
* **Implementation**:
  + Used Python's pandas.cut() method to categorize the loyalty scores.



**2. Encoding Loyalty Categories**

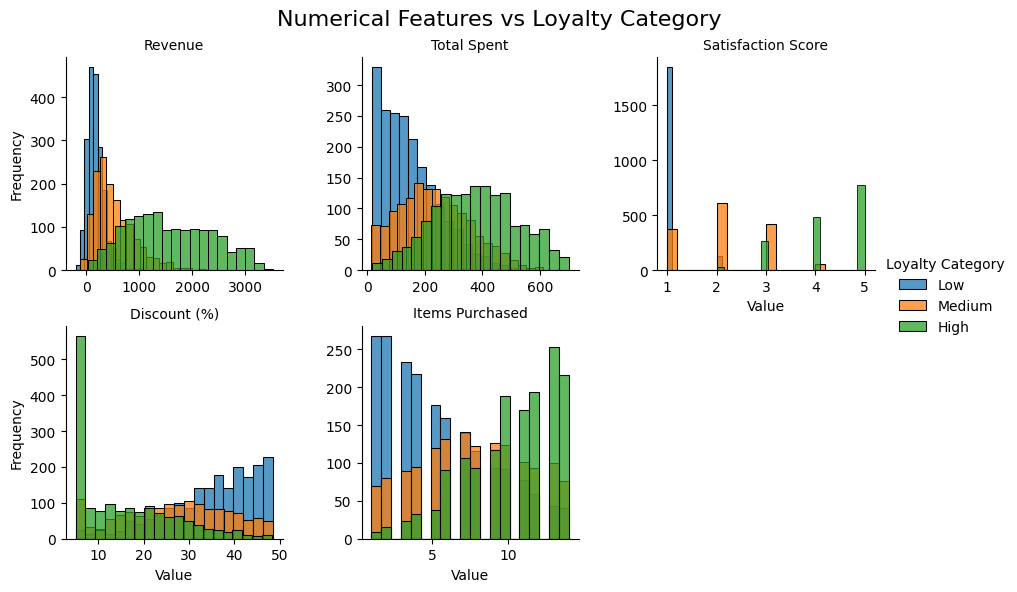
* **Purpose**: Convert the categorical "Loyalty Category" into numerical labels for compatibility with machine learning models.
  + Used **Label Encoding** to assign integer values to the categories:
    - **Low** → 0
    - **Medium** → 1
    - **High** → 2
  + Ensured consistency by saving the label encoder for reuse during deployment.
* **Implementation**:
  + Applied LabelEncoder from sklearn.

# Exploratory Data Analysis (EDA):

This section of EDA focuses on understanding how various features in the dataset relate to the newly created Loyalty Category. The analysis highlights the patterns and relationships between customer behaviors, demographics, and loyalty levels (Low, Medium, High). Here's a breakdown of the EDA steps performed:

**Numerical Features vs Loyalty Category**

The plot provides a comparison of the distributions of various numerical features (Revenue, Total Spent, Satisfaction Score, Discount (%), and Items Purchased) across the three loyalty categories: **Low**, **Medium**, and **High**. Each feature is represented by a histogram with distinct bars for each loyalty category. Here’s a detailed analysis of each subplot:

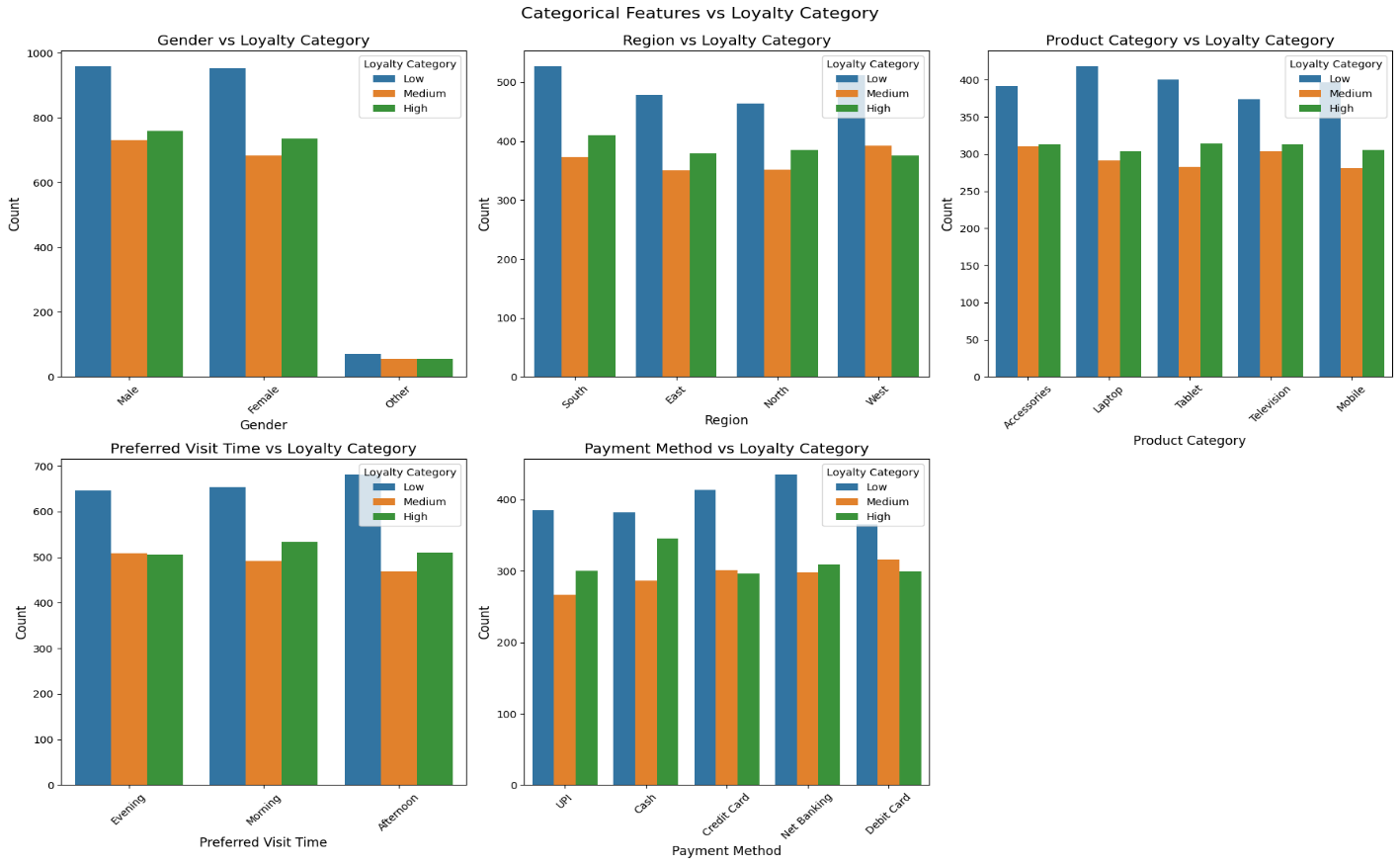


**Overall Insights**

* **Key Predictors**: Revenue, Total Spent, and Satisfaction Score are the strongest indicators of loyalty, with clear trends where higher values correspond to higher loyalty categories.
* **Supporting Predictors**: Discount (%) and Items Purchased provide additional insight but may have less direct influence compared to the other features.
* **Low Loyalty Characteristics**: These customers tend to have low revenue, low total spending, low satisfaction scores, and often receive smaller discounts.
* **High Loyalty Characteristics**: These customers exhibit high revenue, high spending, and high satisfaction scores, with a broader distribution in items purchased and discounts.

This analysis emphasizes the importance of focusing on satisfaction, spending behavior, and revenue to predict and enhance customer loyalty effectively.

**Categorical Features and Loyalty Category**

****

**1. Gender vs Loyalty Category**

* + Both Male and Female customers are distributed relatively evenly across loyalty categories, but one gender might slightly dominate a particular category.
  + Gender might not have a strong direct impact on loyalty but could provide insights for targeted marketing strategies.

**2. Region vs Loyalty Category**

* + Some regions may have a higher concentration of High loyalty customers, while others may have more Low loyalty customers.
  + Regional trends can indicate geographical differences in customer behavior, guiding region-specific loyalty strategies.

**3. Product Category vs Loyalty Category**

* + Certain product categories (e.g., high-value items like Laptop or Television) might have a larger proportion of High loyalty customers.
  + Lower-value items (e.g., Accessories) could attract more Low loyalty customers.
  + Product category is an important feature, as customers buying higher-value products are likely more loyal.

**4. Preferred Visit Time vs Loyalty Category**

* + Different loyalty categories may prefer distinct visit times (e.g., High loyalty customers might prefer Evening).
  + Preferred visit time could indicate behavioral patterns that businesses can use to personalize interactions or promotions.

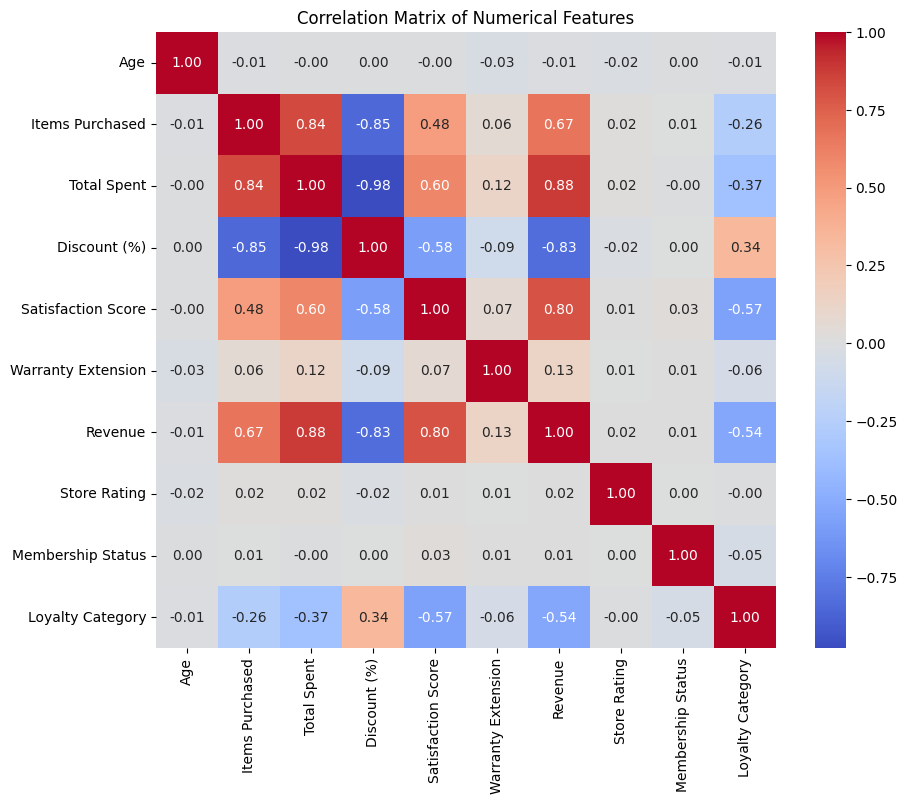
**5. Payment Method vs Loyalty Category**

* + Certain payment methods (e.g., Credit Card, Net Banking) might have a higher proportion of High loyalty customers compared to others (e.g., Cash).
  + Payment methods can reflect customer preferences and could be a secondary factor influencing loyalty.

**Overall Insights:**

* **Key Features**:
  + **Product Category** and **Preferred Visit Time** provide the strongest insights into customer loyalty.
  + **Region** and **Payment Method** offer supporting insights that could enhance the model's accuracy.
* **Weaker Predictors**:
  + **Gender** shows minimal variation across loyalty categories and might not contribute significantly to predictive modeling.

**Correlation of each numerical variable with the Loyalty Score**



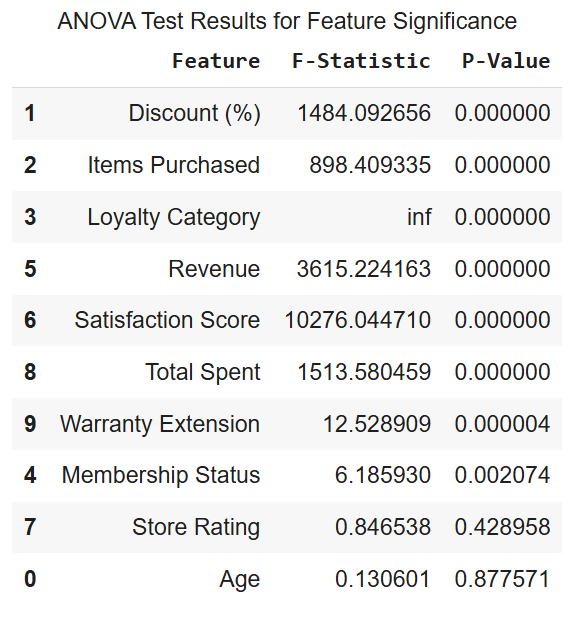
**Insights from Correlation Matrix:**

1. **Strong Negative Correlations:**
   * Satisfaction Score (-0.57): High satisfaction scores are negatively correlated with lower loyalty categories, indicating higher loyalty as satisfaction increases.
   * Revenue (-0.54): Revenue decreases are linked to lower loyalty, suggesting that higher loyalty is associated with higher spending.
   * Total Spent (-0.37): Similar to revenue, total spending shows a moderate negative correlation with loyalty category.
2. **Moderate Positive Correlation:**
   * Discount (%) (0.34): Discounts positively correlate with loyalty category, indicating that higher discounts might attract or retain loyal customers.
3. **Weak Correlations:**
   * Items Purchased (-0.26): A slight negative correlation, indicating that purchasing fewer items may be linked to higher loyalty categories.
   * Warranty Extension (-0.06) and Membership Status (-0.05): Minimal impact on loyalty category, indicating these factors might not significantly influence loyalty.
4. **Negligible Correlation:**
   * Age (-0.01) and Store Rating (-0.00): These features have almost no correlation with loyalty category, suggesting they have little to no predictive power.

* High-impact features for loyalty prediction include Satisfaction Score, Revenue, Total Spent, and Discount (%).
* Features like Age and Store Rating have minimal influence and may be less relevant for predictive models.

**Key Features Driving Loyalty Prediction (Based on ANOVA Result)**

The ANOVA test evaluates the significance of features in predicting the loyalty category by calculating the F-statistic and corresponding p-value. A high F-statistic and a low p-value (< 0.05) indicate that the feature significantly impacts the target variable (Loyalty Category). Here’s an interpretation of the results:



The most influential features for predicting customer loyalty are **Satisfaction Score**, **Revenue**, **Total Spent**, **Discount (%)**, and **Items Purchased**. Moderately significant features like **Warranty Extension** and **Membership Status** can also provide additional insights. Features such as **Store Rating** and **Age** add minimal value and may not need to be included in the final predictive model. This insight allows for focused feature selection, improving the model's accuracy and efficiency.

**Key Findings from Exploratory Data Analysis (EDA)**

1. **Target Distribution (Loyalty Category)**:
   * The distribution of loyalty categories (Low, Medium, High) is relatively balanced, indicating no significant class imbalance in the dataset. This is beneficial for building predictive models without requiring additional balancing techniques.
2. **Significant Features for Loyalty Prediction**:
   * **Satisfaction Score**: Strong negative correlation with the loyalty category, indicating that higher satisfaction scores are associated with higher loyalty levels.
   * **Revenue**: Customers generating higher revenue tend to belong to higher loyalty categories.
   * **Discount (%)**: Positive correlation suggests that larger discounts contribute to increased customer loyalty.
   * **Total Spent**: Strongly correlated with loyalty, as customers who spend more tend to show higher loyalty.
   * **Items Purchased**: Surprisingly, customers purchasing more items tend to fall into lower loyalty categories, possibly indicating one-time bulk buyers.
3. **Features with Minimal Impact**:
   * **Age** and **Store Rating** showed negligible correlation with loyalty categories and are unlikely to contribute significantly to predictive model performance.
4. **Potential Outliers**:
   * Outliers were observed in variables such as Total Spent and Revenue, which represent high-value customers. While these outliers are genuine and provide valuable insights, they should be handled carefully using transformation techniques (e.g., log transformation) or robust modeling approaches to prevent skewing the results.
5. **Correlation Among Features**:
   * High correlations between Total Spent, Revenue, and Items Purchased suggest the presence of multicollinearity. This may need to be addressed using dimensionality reduction techniques like Principal Component Analysis (PCA) or by selecting only one of these highly correlated features for modeling.
6. **Missing Data**:
   * Some columns, such as Membership Status, Preferred Visit Time, and Payment Method, have missing values. Appropriate imputation strategies (e.g., mean, median, or mode) have been or should be applied to ensure a complete dataset for modeling.

**Potential Challenges for Further Analysis**

1. **Multicollinearity**:
   * Numerical features such as Total Spent and Revenue are highly correlated, which may lead to issues like unstable coefficients in regression models or inflated feature importance in tree-based models. Addressing multicollinearity through techniques such as dimensionality reduction (e.g., PCA) or feature selection is essential for reliable analysis.
2. **Class Overlap**:
   * Overlapping characteristics between loyalty categories (e.g., Medium and High) could reduce model accuracy. Advanced classification techniques, such as Support Vector Machines (SVM) with non-linear kernels or ensemble models like Random Forest and Gradient Boosting, may be required to effectively distinguish between overlapping classes.
3. **Categorical Features**:
   * Features like Region, Gender, and Product Category are categorical and require proper encoding. Techniques like one-hot encoding or target encoding must be applied to convert these features into numerical formats compatible with machine learning algorithms.
4. **Feature Scaling**:
   * Numerical features, including Total Spent, Revenue, and Satisfaction Score, have varying scales, which can affect the performance of algorithms sensitive to feature scaling (e.g., SVM, KNN). Applying normalization or standardization will help ensure that all features contribute equally to the model.
5. **Outliers**:
   * Extreme values in features like Total Spent and Revenue could bias the model, especially in distance-based or regression models. Robust approaches such as quantile-based outlier removal, log transformation, or robust algorithms (e.g., tree-based methods) should be employed to mitigate this issue.
6. **Imbalanced Feature Importance**:
   * While features like Satisfaction Score and Discount (%) are highly significant, less impactful features such as Age and Store Rating could introduce noise, reducing model performance. Implementing feature selection techniques like Recursive Feature Elimination (RFE) or feature importance analysis will help eliminate irrelevant features and improve model accuracy.

**Data Transformation**:

Data transformation involves converting raw data into a format suitable for analysis and modeling. This process ensures the data is clean, scaled, and ready for machine learning algorithms. Below are the transformation steps performed:

**1. Binning**

**Target Variable Creation:**

* **Purpose**: Simplify the prediction task by categorizing the continuous Loyalty Score into bins (Low, Medium, High).

**Encoding Target Variable:**

* Encoded the target variable (Loyalty Category) using **Label Encoding** for compatibility with machine learning models.

**2. Normalization and Scaling**

Numerical features often have varying scales, which can affect the performance of distance-based models. Scaling was applied to bring all features to a similar range.

**Robust Scaling**

* **Purpose**: Handles outliers by scaling features based on the median and interquartile range.
* **Applied To**:
  + Total Spent
  + Revenue
  + Discount (%)
  + Satisfaction Score
  + Items Purchased

**3. Handling Categorical Features**

Categorical features were transformed into numerical representations to be used in machine learning models.

**One-Hot Encoding**

* **Purpose**: Converts categorical variables into binary columns for each category.
* **Applied To**:
  + Gender
  + Region
  + Product Category
  + Preferred Visit Time
  + Payment Method

**Feature Selection:**

Feature selection is a critical step in machine learning that involves identifying the most relevant features from a dataset. It helps improve model performance, reduce overfitting, and lower computational cost by focusing on the most impactful variables. Here’s how feature selection was performed in this project:

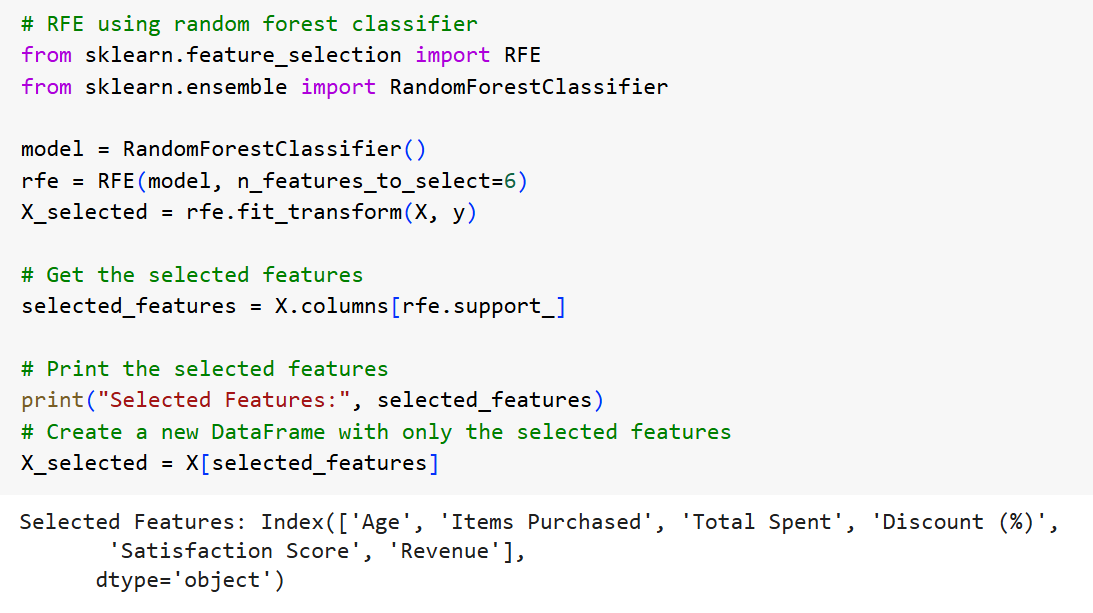
**1. Purpose of Feature Selection**

* **Enhance Model Accuracy**:
  + Irrelevant or noisy features can degrade the model's predictive power. Selecting the most informative features improves accuracy.
* **Reduce Overfitting**:
  + Fewer features reduce the risk of overfitting, making the model generalize better to unseen data.
* **Optimize Computational Efficiency**:
  + Reducing the number of features lowers the computational complexity, speeding up model training and prediction.

**2. Method Used for Feature Selection**

**Recursive Feature Elimination (RFE)**

* + RFE recursively removes less important features to select the most significant ones.
  + It starts by training a model on all features, ranking their importance, and eliminating the least important features until the desired number of features is reached.
* **Algorithm Used**: A **Random Forest Classifier** was chosen as the base estimator for RFE because it automatically calculates feature importance.



**Modeling:**

The modeling phase involves building, training, and evaluating machine learning models to predict customer loyalty categories (Low, Medium, High). The goal is to identify the most accurate and efficient model while ensuring it generalizes well to unseen data.

**1. Problem Type**

* **Classification Problem**:
  + Predicting loyalty categories (Low, Medium, High) based on customer features.
  + Multi-class classification with three target labels.

**2. Model Selection**

Multiple machine learning models were experimented with to determine the best-performing algorithm. These included:

**A. Random Forest Classifier**

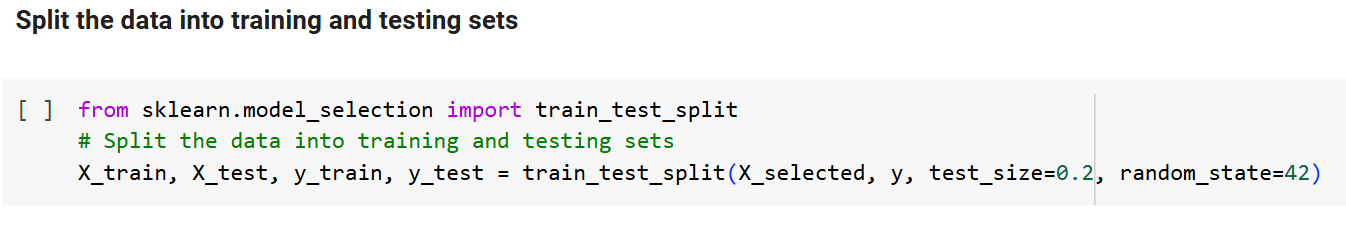
* **Why Used**:
  + Handles categorical and numerical data well.
  + Robust to outliers and multicollinearity.
  + Provides feature importance for interpretability.

**B. Gradient Boosting (e.g., XGBoost)**

* **Why Used**:
  + Handles non-linear relationships well.
  + Boosting technique reduces bias and variance, improving accuracy.

**3. Data Preparation**

* The preprocessed data was split into training and testing sets to evaluate model performance:
  + **80% Training**: Used to train the model.
  + **20% Testing**: Used to validate the model.

****

**4. Model Training and Hyperparameter Tuning**

Hyperparameter tuning was performed using GridSearchCV to optimize the models for better performance.

****

**Other Models**

Similar hyperparameter tuning was performed for XGBoost, Logistic Regression, and SVM to ensure fair comparisons.

**Model Evaluation:**

The models were evaluated using the following metrics:

* **Accuracy**: Percentage of correctly classified samples.
* **Precision**: Accuracy of positive predictions.
* **Recall**: Ability to capture all positive samples.
* **F1-Score**: Harmonic mean of precision and recall.
* **Confusion Matrix**: Breakdown of true positives, false positives, true negatives, and false negatives.

**Results:**

The performance of the models was compared, and the best-performing model was selected:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Random Forest | **82.2%** | **81.8%** | **82.1%** | **81.9%** |
| XGBoost | 82.6% | 82.3% | 82.6% | 82.3% |

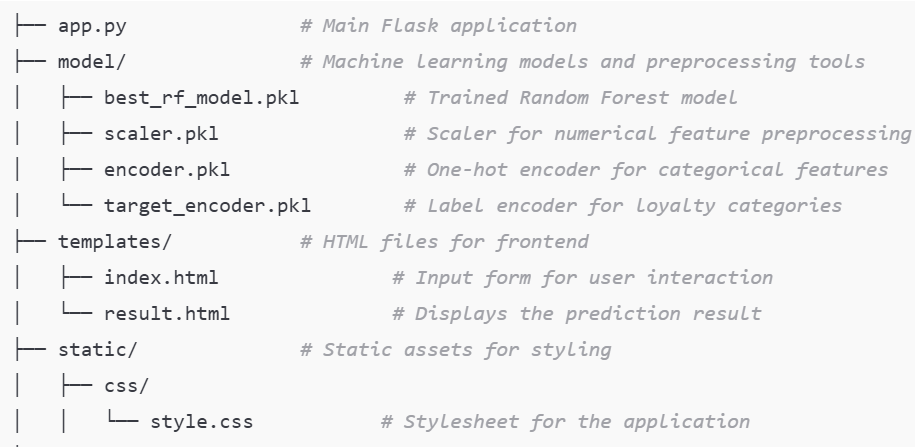
**Best Model**: Random Forest Classifier achieved the highest accuracy and balanced performance across all metrics.

**Flask Development:**

This section outlines the development of a Flask web application for deploying the customer loyalty prediction model. Flask is a lightweight Python web framework ideal for creating RESTful APIs and web interfaces.

**Folder Structure**

The project is organized as follows



**Workflow**

1. **Input**: User provides data via the form on the home page (/).
2. **Preprocessing**: Numerical features are scaled, and categorical features are one-hot encoded.
3. **Prediction**: Preprocessed data is passed to the best\_rf\_model for prediction.
4. **Output**: The predicted loyalty category is displayed on the results page.

This project bridges machine learning with a user-friendly web interface, enabling businesses to predict customer loyalty efficiently. With a modular structure and robust deployment setup, it serves as a practical tool for improving customer engagement and retention strategies.

**Deploying the Flask Application on PythonAnywhere**

PythonAnywhere is a popular cloud platform that provides an easy environment for deploying Python-based web applications, including Flask. Below are the detailed steps to deploy the Customer Loyalty Prediction application.

1. **Ensure the Following Files**

1. app.py: The main Flask application.
2. model/: Directory containing the serialized model (model1.pkl) and preprocessing tools (rodust\_scaler1.pkl, onehot\_encoder1.pkl, target\_encoder1.pkl,rfe\_model1.pkl).
3. templates/: Directory with HTML files (index.html and result.html).
4. static/: Directory containing CSS files (style.css).
5. requirements.txt: List of all Python dependencies for your project.

**2. Upload Your Files**

1. Go to the **Files** tab in PythonAnywhere.
2. Navigate to your home directory or create a new folder (e.g., mysite).
3. Upload your project files (app.py, model/, templates/, static/, and requirements.txt) to this directory.

By following these steps, your Flask application for customer loyalty prediction will be successfully deployed on PythonAnywhere, making it accessible to users via a public URL. PythonAnywhere simplifies the deployment process and provides a scalable environment for web applications.

# Discussion

## Limitations:

1. **Feature Availability and Quality**: The dataset is limited to specific features like Revenue, Satisfaction Score, and Items Purchased. Other potentially important factors, such as customer feedback, external market influences, or competitor data, were not included, which could affect the model's completeness.
2. **Imbalanced Importance Across Features**: Features like Satisfaction Score and Revenue heavily influence the model, while others like Age and Store Rating contribute minimally. This imbalance might make the model overly reliant on a few key features.
3. **Handling of Overlapping Classes**: Loyalty categories (Low, Medium, High) may have overlapping characteristics. The model might misclassify instances near category boundaries, reducing prediction accuracy.
4. **Impact of Outliers**: Although outliers in features like Total Spent and Revenue were retained due to their significance, they could still bias certain machine learning models, especially those sensitive to extreme values.
5. **Categorical Feature Encoding**: One-hot encoding increased the dimensionality of the dataset, which may lead to sparsity and computational inefficiencies for some models.
6. **Generalizability**: The model's performance is evaluated on a specific dataset. Its generalizability to different datasets or customer populations (e.g., from other industries or regions) is uncertain.
7. **Lack of Real-Time Prediction Integration**: The project stops at model deployment and does not implement real-time prediction capabilities or dynamic data updates.

## Future Work:

1. **Incorporation of Additional Features**:
   * Include new features such as customer reviews, browsing history, or competitor influence to improve the model’s accuracy and robustness.
   * Collect time-series data to track loyalty trends over time and enable forecasting.
2. **Addressing Class Overlap**:
   * Use advanced algorithms like probabilistic models, ensemble methods, or non-linear classifiers (e.g., XGBoost) to better handle overlapping loyalty categories.
   * Explore soft classification methods to provide probabilities for category membership.
3. **Feature Engineering**:
   * Create derived features such as the frequency of visits, average spending per visit, or recency of purchases, which could improve model predictions.
   * Apply dimensionality reduction techniques like Principal Component Analysis (PCA) to handle correlated features and reduce redundancy.
4. **Improving Outlier Handling**:
   * Implement robust outlier detection and treatment methods, such as Isolation Forests or quantile-based scaling, to minimize their impact without losing valuable information.
5. **Generalization Testing**:
   * Test the model on external datasets from different industries, regions, or customer demographics to evaluate its adaptability and robustness.
6. **Real-Time Deployment**:
   * Enhance the model to support real-time data processing and predictions in a production environment. Integrate the model into a live system (e.g., customer loyalty program or CRM).
7. **Customer Behavior Segmentation**:
   * Extend the analysis to include unsupervised learning techniques like clustering for customer segmentation. This could complement loyalty prediction by identifying specific customer groups.
8. **Explainability and Interpretability**:
   * Implement techniques like SHAP (SHapley Additive exPlanations) to explain the model’s predictions, making it more transparent and actionable for business stakeholders.
9. **Dynamic Feedback Loop**:
   * Create a feedback loop where the model continually learns from new data, refining its predictions over time to adapt to changing customer behaviors.
10. **Optimization for Scalability**:
    * Optimize the model for deployment in large-scale systems, ensuring efficient handling of high-dimensional data and large customer bases.

## Implications:

The implications of this project extend beyond loyalty prediction, influencing critical aspects of customer relationship management, marketing, and strategic planning. By leveraging the model’s insights, businesses can improve customer satisfaction, optimize resource allocation, and drive growth, positioning themselves for long-term success in a competitive market.

# Conclusion

## 

## Summary of Findings:

The customer loyalty prediction project identified key drivers of loyalty, including **Satisfaction Score**, **Revenue**, and **Total Spent**, with high loyalty linked to higher satisfaction and spending. Features like **Age** and **Store Rating** showed minimal impact.

The Random Forest Classifier emerged as the best model, achieving **82.2% accuracy** with balanced performance across loyalty categories. Challenges such as multicollinearity and outliers were effectively addressed through feature selection and scaling.

These findings enable businesses to target low-loyalty customers with retention strategies, reward high-value customers, and optimize resources for improved engagement and satisfaction

## Impact:

The project enables businesses to enhance **customer retention** by identifying and engaging Low loyalty customers while maximizing revenue from High loyalty customers through targeted strategies. It fosters **data-driven decision-making**, guiding resource allocation and personalized marketing efforts. By leveraging insights into loyalty drivers like satisfaction and spending, businesses can improve customer satisfaction and gain a **competitive edge**, ensuring sustained growth and stronger customer relationships.

## Recommendations:

1. Focus on Low Loyalty Customers: Implement targeted retention strategies, such as personalized discounts, rewards, or engagement campaigns, to reduce churn among Low loyalty customers.
2. Enhance Customer Satisfaction: Prioritize improving satisfaction scores by addressing customer feedback, enhancing service quality, and resolving issues promptly.
3. Leverage Insights for Personalization: Use customer preferences (e.g., Preferred Visit Time, Product Category) to design tailored promotions, improving engagement and loyalty.
4. Invest in High-Value Customers: Reward High loyalty customers with exclusive benefits or offers to strengthen their loyalty and encourage advocacy.
5. Refine Marketing and Resource Allocation: Focus resources and campaigns on impactful drivers of loyalty, such as revenue and discounts, for better ROI.
6. Address Overlapping Loyalty Categories: Use advanced algorithms or probabilistic models to improve classification accuracy for borderline cases between Medium and High loyalty.
7. Incorporate Additional Data: Expand the model by including new features like customer reviews, purchase frequency, or external market data for more comprehensive predictions.
8. Enable Real-Time Integration: Deploy the model into a real-time system to provide dynamic loyalty predictions and timely interventions.
9. Regularly Update the Model: Continuously retrain the model with new data to ensure it adapts to evolving customer behaviors and market trends.
10. Scale Across Business Units: Extend the model to other domains, such as predicting churn or segmenting customers for strategic planning.

# 

# References

* ***Pandas:*** [***https://pandas.pydata.org***](https://pandas.pydata.org)
* ***Seaborn:*** [***https://seaborn.pydata.org***](https://seaborn.pydata.org)
* ***Matplotlib:*** [***https://matplotlib.org***](https://matplotlib.org)
* ***Scikit-learn documentation:*** [***https://scikit-learn.org***](https://scikit-learn.org)
* ***Kaggle:*** [***https://www.kaggle.com***](https://www.kaggle.com)