Assignment 01

Data Mining

Name: Kamil Ilyas

Roll No: 20i-2371

Section: CS-A

Submitted To: Mr. Farrukh Bashir

Date: March 24, 2024

**Task 1**

**Banking Dataset Classification**

**About Dataset:**

The dataset pertains to direct marketing campaigns of a Portuguese banking institution aimed at promoting long-term deposits among existing customers. The bank is facing a revenue decline and seeks to identify customers with a higher likelihood of subscribing to long-term deposits to focus marketing efforts effectively.

**Dataset Description:**

The dataset consists of two files: train.csv containing 32,950 examples and test.csv containing 8,238 observations. The train dataset includes 21 input features along with the target feature ('y'), ordered by date from May 2008 to November 2010, closely resembling the data analyzed in [Moro et al., 2014].

**Features:**

- age: Numerical (Discrete) - Represents whole numbers. No transformation needed.

- job, marital, education, default, housing, loan, month, day\_of\_week, poutcome: Nominal - Categories have no inherent order. One-Hot Encoding to convert categorical variables into numerical features.

- contact: Binary (Target Variable) - Binary categories. No transformation needed.

- duration, campaign, pdays, previous: Numerical (Discrete) - Represents whole numbers. No transformation needed.

- y (target variable): Binary (Target Variable) - Binary categories. No transformation needed.

**Preprocessing Steps:**

1. **Handling Missing Data:**

Identified missing values in numeric columns.

Employed mean and median imputation techniques to fill missing values.

Justification: Mean imputation is used when the data is normally distributed and doesn't have outliers, while median imputation is robust to outliers and skewed data distributions.

1. **Data Transformation and Scaling:**

Applied log transformation to the 'duration' column to handle skewness.

Utilized Z-score normalization, Decimal scaling normalization, and Min-Max scaling to scale numerical features.

Justification: Log transformation helps stabilize variance and make the data more Gaussian-like, while scaling ensures that features are on the same scale, which is crucial for certain machine learning algorithms like SVM and KNN.

1. **Encoding Categorical Variables:**

Employed one-hot encoding for nominal categorical variables like job, marital status, education, etc.

Justification: One-hot encoding is necessary to convert categorical variables into a numerical format that can be used for modeling without implying any ordinal relationship between categories.

**Evaluation Results:**

**Before Preprocessing:**

* Utilized various machine learning algorithms including KNN, SVM, Naive Bayes, Logistic Regression, and Decision Trees.
* Evaluated models on the original dataset without preprocessing.
* Performance metrics: Confusion Matrix, Precision, Recall, F1-score.

**After Preprocessing:**

* Repeated the same machine learning algorithms and evaluation process on the preprocessed dataset.
* Compared the performance metrics before and after preprocessing to assess the impact of preprocessing steps on model performance.

The evaluation involved training various machine learning algorithms on both the original and preprocessed datasets. Performance metrics such as confusion matrix, precision, recall, and F1-score were used to assess the effectiveness of preprocessing steps in improving model performance and generalization on unseen data. Comparisons were made between models trained on the original and preprocessed datasets to understand the impact of preprocessing on predictive performance.

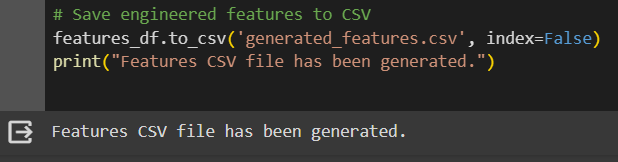
**Conclusion:**

Preprocessing steps such as handling missing data, transformation, scaling, and encoding categorical variables are crucial for preparing the dataset for machine learning models. The evaluation results before and after preprocessing demonstrate the effectiveness of these steps in improving model performance and generalization on unseen data.

**Task 2.1**

**1: Preprocessing and Feature Engineering:**

This section involves loading the retail customer transaction data, converting the visit dates from strings to datetime objects for easier manipulation, and then aggregating various customer-specific metrics. The key features extracted per customer include total revenue, maximum and minimum purchase in a day, total number of visit days, and the standard deviation of purchases. These features aim to capture both the breadth and variability of each customer's purchasing behavior. The aggregated features are then saved to a CSV file, generated\_features.csv, making them accessible for further analysis or model training.

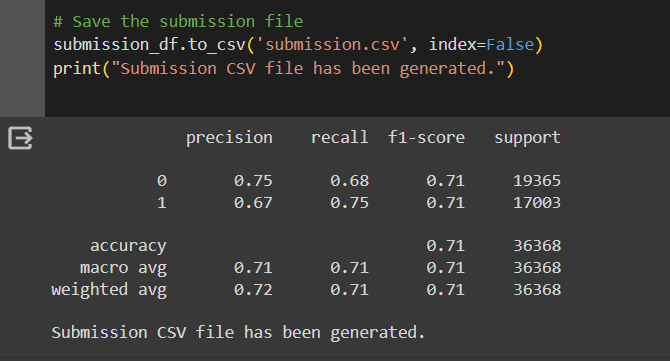


**2: Integrate Actual Churn Labels and Prepare for Model Training:**

After feature engineering, the next step merges the engineered customer features with actual churn labels extracted from a separate labels file. This process aligns each customer's transactional behavior with their churn status, creating a comprehensive dataset ready for predictive modeling. The dataset is subsequently split into training and testing sets, ensuring a balanced approach to model validation where the model's performance is tested on unseen data.

**3: Model Training and Evaluation:**

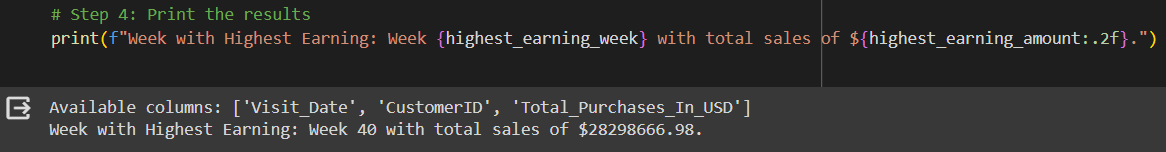
A RandomForestClassifier, a powerful ensemble learning method suitable for classification tasks, is trained on the prepared dataset. The model's performance is evaluated using standard metrics: accuracy, precision, and recall, providing insights into its predictive capabilities. Furthermore, predictions are generated for the test set and saved to a submission.csv file, which could be used for operational churn intervention strategies or further analysis.



**Task 2.2**

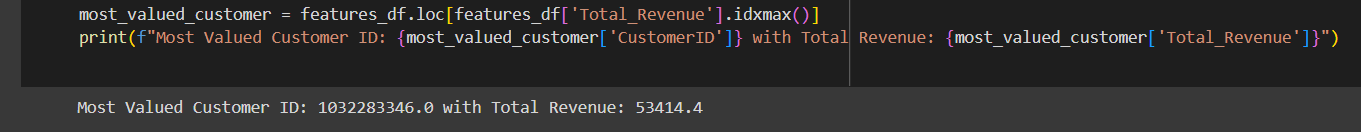
**1: Week with Highest Earning**

This part involves analyzing sales data aggregated on a weekly basis to identify the week with the highest total sales. This analysis helps understand customer purchasing patterns and identify peak sales periods, although it's simplified in the provided code and assumes weekly sales data is available.



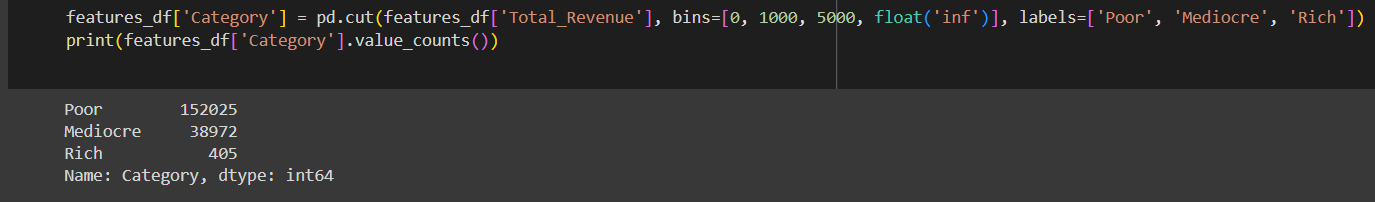
**2: Most Valued Customer**

Identifies the customer contributing the most to total revenue, highlighting their importance to the business. Understanding characteristics of such customers can aid in developing targeted retention or upselling strategies.



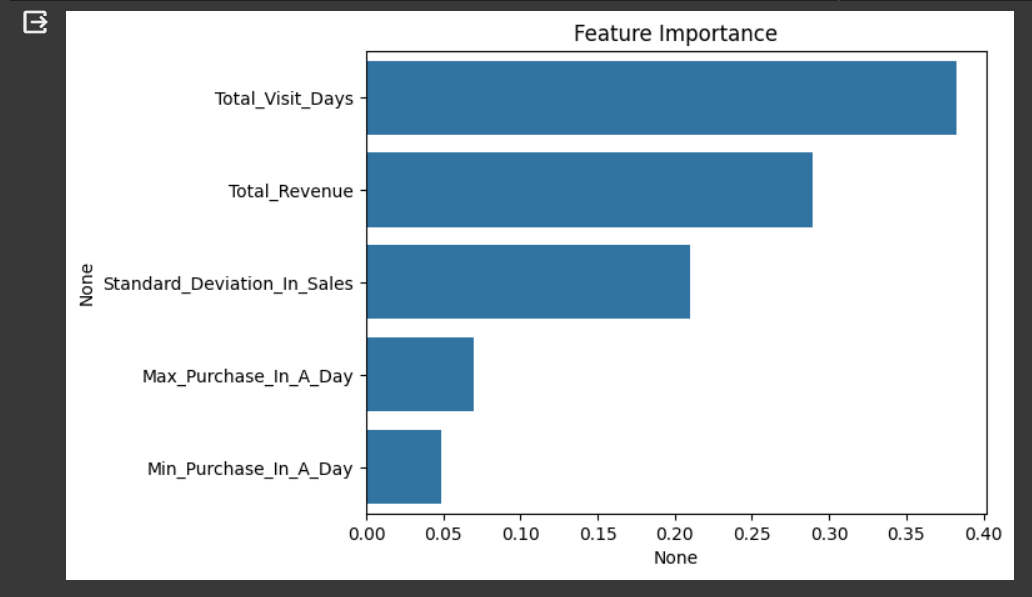
**3: Customer Categorization**

Customers are categorized into groups (Poor, Mediocre, Rich) based on their total revenue. This segmentation allows for more nuanced analysis and strategy development, enabling tailored marketing or service offerings to different customer segments.



**4: Churn Important Factors:**

By visualizing the importance of different features in predicting churn, this part provides insights into what factors most influence customer retention. Identifying these factors can guide efforts to mitigate churn, such as improving customer service or adjusting pricing strategies.



**5: Churn Rate and Historical Visits**:

Explores the relationship between the churn rate and the number of historical visits, aiming to understand how customer engagement correlates with retention. This analysis can uncover patterns that suggest the optimal engagement level to reduce churn risk.

