NAME: T.Sai Abhinav Chandra Znumber :23756830

INTRODUCTION TO DATA SCIENCE

CAP 5768

Problem: Credit Card Fraud Detection

Dataset link : <https://github.com/YBIFoundation/Dataset/blob/main/Credit%20Default.csv>

The dataset titled "Credit Default.csv" appears to be related to credit risk assessment, containing financial and demographic information about customers. It includes the following key features:

1. **Income**: Customer's income level, presumably in dollars. It helps to gauge financial capability.
2. **Age**: Age of the customer in years.
3. **Loan**: The amount of loan taken by the customer, likely in dollars.
4. **Loan to Income**: Ratio of loan amount to the customer's income, indicating how much of their income is allocated toward loan repayment.
5. **Default**: A binary feature indicating whether the customer has defaulted on the loan (0 for no default, 1 for default).

The dataset aims to help predict the likelihood of customers defaulting on their loans based on financial and demographic characteristics. It can be used for credit risk modeling, customer segmentation, and providing financial recommendations to mitigate risks. There are no missing values, and several features may require transformation, scaling, or feature engineering to train predictive models effectively.

Credit Default Analysis Report

1. Introduction

The purpose of this analysis is to predict loan default among customers based on their financial and demographic data, and to provide actionable insights for risk management and business decision-making. The dataset used is from a credit default study, and it includes features like income, age, loan amount, and loan-to-income ratio. We apply various machine learning techniques to explore and predict customer behavior and evaluate the risk of default.

1. Exploratory Data Analysis (EDA)

To understand the structure and characteristics of the dataset, we performed a thorough Exploratory Data Analysis (EDA):

* Data Overview: The dataset contains features such as 'Income', 'Age', 'Loan', 'Loan to Income', and 'Default'. After loading the data, we observed that it contains no missing values, and summary statistics were generated to understand the spread and characteristics of each feature.
* Missing Values: Using a heatmap, we confirmed that there were no missing values in the dataset, ensuring data completeness.
* Feature Distributions: Histograms were plotted for all numerical features to examine the distributions. It was noted that some features exhibited skewness, which was later addressed during the preprocessing stage.
* Correlation Analysis: A correlation heatmap was plotted to analyze the relationships between features. The 'Loan to Income' ratio showed moderate correlation with the default outcome, indicating its potential importance for prediction.
* Outlier Detection: Boxplots for all numerical features revealed the presence of outliers. These outliers were treated using the Interquartile Range (IQR) method to ensure the model's stability.

1. Data Preprocessing

* Outlier Handling: Outliers were removed using the IQR method. Values below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR were filtered out for each feature, reducing noise in the dataset.
* Feature Engineering: New features were created to enhance the model's predictive capabilities.
  + 'Income per Loan' was created to understand the ratio of income to loan amount.
    - 'Age Group' was created to categorize individuals into four age groups ('Young', 'Adult', 'Middle-aged', 'Senior').
* Feature Scaling and Normalization: The features were scaled using StandardScaler and normalized using MinMaxScaler to standardize the values for improved model performance.

1. Model Building and Evaluation

Two types of models were built: a classification model to predict loan default and a clustering model for customer segmentation.

* Decision Tree Classifier:
  + We used GridSearchCV to perform hyperparameter tuning with a parameter grid of 'max\_depth'

and 'min\_samples\_split'. The best model was selected based on cross-validation scores.

* + The model achieved an accuracy of over 70%, with an F1-score indicating a good balance between precision and recall.
  + A confusion matrix was plotted to visualize the classification performance, and an ROC curve showed the model's capability in distinguishing between default and non-default customers.
* K-Means Clustering:
  + PCA was used to reduce dimensionality before applying K-Means clustering, which divided customers into three distinct clusters.
  + The Silhouette Score was used to evaluate the quality of clustering, confirming well-defined groupings.

1. Business Insights and Recommendations

Based on the model results and analysis, several key insights were derived:

* Customer Default Prediction: The Decision Tree Classifier indicates that customers with lower 'Income per Loan' ratios and belonging to specific age groups (e.g., Young or Senior) are at higher risk of defaulting. These customers should be closely monitored or provided with additional financial management support.
* Customer Segmentation: Using K-Means clustering, customers were grouped into three segments. Each segment represents distinct characteristics:
  + Customers with a high 'Income per Loan' ratio might be offered premium financial services.
    - Customers with lower ratios should be targeted with educational programs to enhance their financial literacy.

1. Actionable Recommendations

* Risk Monitoring System: Implement a customer risk monitoring system based on features like 'Income per Loan' and 'Age Group' to identify high-risk customers.
* Targeted Financial Support: Provide financial literacy programs for younger and senior customers to help them manage their credit effectively.
* Tailored Credit Offers: Tailor financial services for different customer segments. Offer premium services to customers with a high 'Income per Loan' ratio and provide credit counseling for those with lower ratios.

1. Conclusion

This analysis provides a comprehensive approach to understanding the risk of loan default. By combining predictive modeling and customer segmentation, financial institutions can better allocate resources, enhance customer satisfaction, and mitigate risks. Further improvements can be made by incorporating more data, using advanced ensemble methods, or refining feature engineering to capture additional customer behaviors.

1. Future Scope

* Model Improvement: Consider using ensemble techniques like Random Forests or Boosting to enhance predictive accuracy.
* Additional Features: Incorporate more customer behavioral data (e.g., spending habits) to improve the model's robustness.
* Automation: Automate the risk monitoring and alert system for high-risk customers to streamline operations and provide proactive support.