Credit Card Application Scoring Analysis

Project Overview

The credit card department of a bank is increasingly adopting data science techniques to assess the creditworthiness of potential customers. This project focuses on analyzing credit card application data using data-driven methodologies to evaluate applicants' creditworthiness. The goal is to make informed decisions about approving or rejecting credit card applications based on a variety of factors.

Problem Statement

The primary objective of this project is to predict the creditworthiness of credit card applicants. The bank aims to mitigate risk and prevent defaults by accurately assessing applicants' ability to repay credit card debt. By analysing applicant data, credit scores, and other relevant information, the bank can make more informed and responsible lending decisions.

Data Description:

The project utilizes two main datasets:

1. Credit_Card.csv: This dataset contains information about credit card applicants, including attributes such as gender, car ownership, property ownership, income, education, and more.

2. Credit_card_label.csv: This dataset contains credit application labels, indicating whether an application was approved (0) or rejected (1).

Data Analysis Steps:

The project involves a comprehensive data analysis process, including data preparation, exploration, visualization, and model building which has been performed using Python. The following steps are covered:

1. Data Import and Libraries: Import necessary libraries for data analysis and visualization.

2. Data Loading and Merging: Load and merge the two CSV files to create a comprehensive dataset.

3.Data Preparation: Clean and prepare the dataset by handling missing values, duplicates, and renaming columns.

4.Data Exploration and Visualization: Explore the dataset through various visualization techniques to understand the distribution and relationships between variables.

5. Univariate Analysis: Analysed individual variables, including gender, car ownership, property ownership, income, education, marital status, and more.

6. Bivariate Analysis: Explore the relationship between variables and the credit application label to understand how different attributes affect application approval.

7. Multivariate Analysis: Create pair plots and box plot grids to visualize relationships across multiple variables.

8. Correlation Analysis: Calculate and visualize the correlation matrix to understand the relationships between numerical variables.

9. Missing Value Analysis: Identify and visualize missing values in the dataset.

10.Outlier Detection and Treatment: Detect and handle outliers in the data using various methods.

11.Data Transformation: Apply transformations like logarithmic and square root transformations to normalize skewed data.

12. Imputation: Handle missing values through basic imputation methods and advanced techniques like K-nearest neighbours (KNN) and Multivariate Imputation by Chained Equations (MICE).

13.Data Conversion and Label Encoding: Convert categorical values into numeric values using label encoding.

14. Model Training and Evaluation: Split the data into training and testing sets, standardize features, and train machine learning models such as Logistic Regression, XGBoost, Decision Tree, and Random Forest. Evaluate model performance using metrics like accuracy, precision, recall, F1-score, and more.

My Sql

The key analyses and their methods that are performed in MY Sql are summarized as follows:

1. Average Annual Income by Customer Group:

- Method: Used a Common Table Expression (CTE) to calculate the average annual income.

- Result: Average annual income value is retrieved from the CTE.

2. Grouping Customers by Income and Classifying:

- Method: Applied a CASE statement to classify customers based on their annual income compared to the average.

- Result: Customer groups ("Good," "Average," "Bad") along with relevant attributes are retrieved.

3. Female Owners of Cars and Property:

- Method: Used a simple SELECT query with conditions on gender, property ownership, and car ownership.

- Result: Details of 177 females owning both cars and property are retrieved.

4. Male Customers Staying with Families:

- Method: Another SELECT query with conditions on gender, family members, and marital status.

- Result: Details of 470 males who are staying with their families are retrieved.

5. Top Five Individuals with Highest Income:

- Method: Employed an aggregate function MAX along with grouping and ordering.

- Result: The top five individuals with the highest incomes, their genders, education, and income types are retrieved.

6. Creating a New Column for Credit Approval:

- Method: Used ALTER TABLE to add a new column named "credit approved."

- Result: The "credit approved" column is added to the table.

7. Updating Credit Approval Status:

- Method: Utilized an UPDATE statement with a CASE expression to populate the new column based on label values.

- Result: The "creditapproved" column is updated according to credit approval status.

8. Count of Married Men with Bad Credit:

- Method: SELECT query with conditions on gender, marital status, and credit approval.

- Result: The count of married men with bad credit is retrieved (51).

9. Highest Education Level and Total Count:

- Method: Aggregate function COUNT and DISTINCT used to find unique education levels and their counts.

- Result: Education levels and their corresponding total counts are retrieved.

10. Comparison of Bad Credit between Married Males and Females:

- Method: Another SELECT query with conditions on marital status, gender, and credit approval.

- Result: Comparison between married males and females in terms of bad credit counts.

In this script, various SQL queries were used to perform data analysis and generate insights from the "credit table" dataset. The script employed aggregate functions, CASE statements, filtering conditions, and table joins to extract meaningful information. Each analysis provides valuable insights into different aspects of credit approval and customer behaviour.

Importance and Impact

In the modern banking landscape, data-driven credit assessment is crucial for responsible lending and risk management. This project demonstrates the application of data science techniques to predict credit card applicants' creditworthiness, thereby helping banks make informed decisions. Accurate credit assessment minimizes the risk of defaults, reduces operational costs, enhances customer experience, and ensures compliance with regulatory standards.

Conclusion

The Credit Card Application Scoring Analysis project showcases the power of data-driven decision-making in the credit card department of a bank. By analysing applicant data and applying machine learning models, the bank can accurately predict creditworthiness and make informed lending decisions. This project has significant implications for risk mitigation, customer satisfaction, and financial stability in the banking sector.