Knowledge Discovery in Credit Card Applications: A Comprehensive Study

October 1, 2023

AVI AJMERA
SAN JOSE STATE UNIVERSITY

1 Introduction

In the age of digitalization, financial institutions are inundated with vast amounts of data. These datasets, if analyzed effectively, can provide valuable insights, particularly in the field of credit card applications. The approval or rejection of an application can be influenced by numerous factors, from personal income to property ownership. Understanding these factors can not only streamline the application process but also reduce the risk for institutions. This research aims to apply the Knowledge Discovery in Databases (KDD) methodology to a credit card application dataset, highlighting the importance of data preprocessing, exploratory analysis, and predictive modeling in deriving actionable insights.

2 Methodology

The KDD methodology, a comprehensive framework for data analysis, was employed for this study. The process involves several steps, including data selection, preprocessing, transformation, data mining, and interpretation. Each step is crucial, ensuring the data is clean, relevant, and ready for in-depth analysis. This research emphasizes the KDD process, demonstrating its application in real-world scenarios.

3 Data Description

The dataset under study comprises credit card application data with features such as gender, car ownership, property ownership, annual income, education level, marital status, and more. The target variable, labeled as 'label', indicates whether an application was approved (0) or rejected (1). Initial observations revealed the presence of missing values in some columns and potential anomalies in the 'Employed_days' column, which were addressed in the preprocessing steps.

4 Data Preprocessing

Data preprocessing is a pivotal step in the KDD process. For this dataset:
- Datasets were merged based on the 'Ind_ID' column.
- Missing values were identified and imputed. For categorical variables like 'GENDER' and 'Type_Occupation', the mode was used for imputation. For numerical variables, such as 'Annual_income' and 'Birthday_count', the median or mean was used, informed by their distribution.
- Categorical variables were encoded to convert them into numerical representations suitable for machine learning models.

5 Exploratory Data Analysis (EDA)

EDA was conducted to understand the relationships between different variables, especially concerning the target variable, 'label'.
- The distribution of the target variable indicated an imbalance, with more rejections than approvals.
- Bar plots revealed insights into categorical variables, showing, for instance, that males had a slightly higher rejection rate than females.
- Box plots for numerical variables, like 'Annual_income' and 'Birthday_count', were generated to understand their distribution and potential outliers.

6 Modeling

A Random Forest classifier, known for its robustness and ability to handle large datasets, was chosen for modeling. The model was trained and evaluated using standard metrics like accuracy, precision, recall, and the F1-score. The ROC and Precision-Recall curves provided insights into the model's performance across different thresholds.

7 Results

The Random Forest classifier showcased promising results, with performance metrics indicating its efficacy in predicting application outcomes. A feature importance analysis revealed the most influential variables in the model's decision-making process.

8 Discussion

The research underscores the importance of thorough data preprocessing and exploratory analysis before predictive modeling. The imbalances in the target variable, if not addressed, could have influenced the model's performance. The insights from the EDA were pivotal, providing a clear understanding of the relationships between various features and the application's outcome.

9 Conclusion

By leveraging the KDD process, this study offers a structured approach to analyzing credit card application data. The findings can guide financial institutions in refining their application processes, optimizing decision-making, and reducing potential risks.