Comprehensive Data Analysis Report

Table of Contents

5. Modelling
   - Model Training and Hyperparameter Tuning
   - Model Evaluation
     - Performance Metrics
     - Confusion Matrix

6. Conclusion
   - Summary of Gradient Boosting Results and Findings
   - Summary of Findings

7. Recommendations and Further Research

   - Ideas for Further Research
   - Recommendations

1. Introduction

Abstract

This report explores the development and application of a predictive model to determine credit card eligibility using a dataset from Kaggle. By leveraging various data analysis and machine learning techniques, the goal is to accurately predict the eligibility of applicants for a credit card. The resulting model achieved high precision and offers concrete recommendations for financial institutions to improve their credit card issuance process.

Problem Statement

Determining credit card eligibility is a crucial task for financial institutions, as it directly impacts both their risk management and customer satisfaction. The challenge lies in accurately predicting the eligibility of applicants based on a variety of factors, including demographics, financial history, and behavioural patterns.

Objective

The objective of this project is to develop a predictive model that can accurately assess the eligibility of applicants for a credit card. This model aims to assist financial institutions in making more informed and efficient decisions, thereby reducing the risk of defaults and improving customer satisfaction.

Background

Using the Credit Card Eligibility dataset from Kaggle, this project applies a combination of data wrangling, exploratory data analysis, and machine learning techniques to develop a robust predictive model. The dataset contains various features that are potential determinants of credit card eligibility, including age, income, employment status, and credit history. By analyzing these features and their relationships with eligibility outcomes, the project seeks to provide valuable insights and recommendations for the credit card issuance process.

This introduction provides a clear problem statement, objective, and background, setting the stage for the rest of the report.

2. Data Wrangling

The Credit Card Eligibility dataset from Kaggle contains various factors that determine or influence an individual's eligibility for a credit card. It includes demographic variables such as gender, employment status, family size, total income, education, and occupation. These elements collectively capture the complex nature of credit card assessments and are essential for evaluating creditworthiness and eligibility.

Initial Dataset Overview

The dataset includes the following columns:

- ID: An identifier for each individual (customer).
- Gender: The gender of the individual.
- Own_car: A binary feature indicating whether the individual owns a car.
- Own_property: A binary feature indicating whether the individual owns a property.
- Work_phone: A binary feature indicating whether the individual has a work phone.

- Phone: A binary feature indicating whether the individual has a phone.
- Email: A binary feature indicating whether the individual has provided an email address.
- Unemployed: A binary feature indicating whether the individual is unemployed.
- Num_children: The number of children the individual has.
- Total_income: The total income of the individual.
- Education: The education level of the individual.
- Occupation: The occupation of the individual.

Steps Taken:

1. Data Cleaning:
   - Missing Values: Missing values were handled by imputing them with the median for numerical columns and the mode for categorical columns.
   - Outliers: Outlier data points were examined individually. Incorrect entries were corrected or made null, while legitimate outliers were retained.

2. Feature Reduction:
   - Dimensionality Reduction: Features with high collinearity or little variance were removed to reduce dimensionality and improve model performance.
   - Correlation Analysis: Redundant features were identified and removed based on correlation analysis.

3. Normalisation:
   - Scaling: Numerical features were scaled to a standard range to ensure all features contributed equally to the model.

The final dataset, after cleaning and processing, was significantly reduced in size, making it more manageable and suitable for analysis. This prepared dataset served as the foundation for subsequent exploratory data analysis and model training, leading to more accurate and efficient predictions.

3. Exploratory Data Analysis (EDA)

```
%matplotlib inline
import os
from autoviz.AutoViz_Class import AutoViz_Class
# Define the AutoViz class
AV = AutoViz_Class()

# Path to the dataset
dataset_path = 'dataset.csv'

# Generate visualizations
dft = AV.AutoViz(dataset_path)

# Save visualizations
def save_current_figure(fig, filename):
    if not os.path.exists('plots'):
        os.makedirs('plots')
    fig.savefig(os.path.join('plots', filename))

# Example: Save current figure (you might need to iterate over all figures if multiple are generated)
fig = plt.gcf()  # Get the current figure
save_current_figure(fig, 'autoviz_output.png')

# Display saved images
from IPython.display import Image, display

def display_saved_images(image_folder='plots'):
    images = os.listdir(image_folder)
    for image in images:
        display(Image(filename=os.path.join(image_folder, image)))

# Display the saved images
display_saved_images()
```

This code snippet uses the `AutoViz` library to generate visualisations for a dataset. It defines functions to save these visualisations into a 'plots' directory and then display them. The process includes:

1. Initialising `AutoViz_Class`.
2. Generating visualisations from `dataset.csv`.
3. Saving visualisations to 'plots' directory.
4. Displaying saved visualisations.

Variable Classification and Dataset Overview

Using the AutoViz library, the dataset was analysed to classify and summarise its variables. The dataset contained 9,709 rows and 20 columns. The classification results are as follows:

- Numeric Columns: 3
- Integer-Categorical Columns: 3
- String-Categorical Columns: 5
- String-Boolean Columns: 0
- Numeric-Boolean Columns: 8
- Discrete String Columns: 0
- NLP String Columns: 0
- Date Time Columns: 0
- ID Columns: 1 (removed due to low information)

In total, 20 predictors were classified, with one variable (ID) removed for being low-information. This classification ensured all variables were correctly typed for subsequent analysis.

Demographic and Eligibility Analysis

Gender Distribution
- The dataset comprises 6,323 males and 3,386 females.

Car Ownership
- Among the individuals in the dataset, 3,570 own a car while 6,139 do not.

Eligibility Criteria
- Eligibility was determined based on age (>= 18), car ownership, and property ownership.
- According to these criteria, 1,060 males and 1,315 females are eligible for a credit card.

Income Type Distribution
- Males:

- Working: 2,955
- Commercial Associate: 1,428
- Pensioner: 1,410
- State Servant: 528
- Student: 2

- Females:

- Working: 2,005
- Commercial Associate: 884
- Pensioner: 302
- State Servant: 194
- Student: 1

Observations

- The dataset indicates a higher number of males compared to females.
- A significant proportion of individuals do not own a car.
- The majority of both males and females are categorised as working, with fewer individuals in other income types.
- Eligibility based on the given criteria shows a higher number of eligible females than males.

These insights provide a clear understanding of the demographic characteristics and eligibility distribution within the dataset, helping to inform the development of predictive models for credit card eligibility.

4. Data Preprocessing and Model Training

Encoding Categorical Variables

Categorical variables were encoded using label encoding. This transformation converts categorical text data into numerical data, which is essential for machine learning algorithms to process the information effectively.

Splitting Data into Features and Target

The dataset was split into features (X) and the target variable (y). The target variable is the one we aim to predict, while the features are the input variables used for making predictions.

Polynomial Feature Transformation

A polynomial feature transformation was applied to the feature set (X) to include interaction terms and enhance the model's ability to capture nonlinear relationships.

Feature Scaling

The features were scaled using the `StandardScaler` to standardize the data. This process ensures that each feature contributes equally to the model performance by bringing all variables to the same scale.

Feature Selection

Using `SelectKBest` with mutual information as the scoring function, the top k features were selected. This step reduces dimensionality and retains the most informative features for model training.

Train-Test Split

The dataset was divided into training and testing sets, with 80% of the data used for training and 20% for testing. This split allows for evaluating the model's performance on unseen data.

Model Training and Hyperparameter Tuning

A Gradient Boosting Classifier was used for training. GridSearchCV was employed to perform hyperparameter tuning, identifying the best combination of parameters for the model. The best parameters included :

- Learning Rate: 0.01
- Max Depth: 3
- Minimum Samples per Leaf: 1
- Number of Estimators: 100.

Model Evaluation

The best Gradient Boosting model was evaluated using accuracy scores and a confusion matrix. The model achieved:

- Train Accuracy: 87.20%
- Test Accuracy: 85.32%

The confusion matrix provided insights into the model's performance in correctly classifying eligible and non-eligible credit card applicants.

5. Summary of Gradient Boosting Results and Findings

The Gradient Boosting model, optimised with the best parameters (learning rate of 0.01, max depth of 3, min samples per leaf of 1, and 100 estimators), achieved impressive performance metrics with a best cross-validation score of 87.15%, train accuracy of 87.20%, and test accuracy of 85.32%. The confusion matrix further illustrated the model's performance, highlighting a high number of true negatives (1657) and a minimal number of false positives (2), although it did note a significant number of false negatives (283) and no true positives. This analysis underscored that income, education, and occupation are key determinants of credit card eligibility. Higher income levels showed a strong correlation with approval likelihood, indicating financial stability, while educated individuals exhibited better eligibility, emphasising the importance of financial literacy. Additionally, the type of occupation significantly impacted approval rates, reflecting income stability. Overall, the model's high test accuracy demonstrates its reliability in predicting creditworthiness, offering valuable insights for banks to refine their credit assessment processes, reduce risks, and improve financial inclusion.

6. Recommendations and Further Research

Ideas for Further Research

1. **Incorporate Additional Demographic Features:**
    - Future research could explore incorporating more detailed demographic information such as marital status, number of dependents, and geographic location. These additional features may provide deeper insights into the factors influencing credit card eligibility.
2. **Longitudinal Data Analysis:**
    - Analyzing data over a longer time period could help identify trends and patterns in credit card eligibility. This could include examining changes in eligibility criteria over time and understanding how economic conditions impact credit card approval rates.
3. **Integration of Behavioral Data:**
    - Incorporating behavioral data, such as spending patterns and repayment history, could enhance the predictive power of the model. This would require accessing additional data sources and integrating them with the current dataset.
4. **Advanced Feature Engineering:**
    - Implement advanced feature engineering techniques such as interaction terms, polynomial features, and domain-specific transformations. These techniques can capture complex relationships between variables, potentially improving model performance.
5. **Use of Ensemble Learning Techniques:**
    - Explore the use of ensemble learning techniques like stacking, bagging, and boosting. Combining multiple models can often result in better predictive performance and robustness compared to individual models.
6. **Hyperparameter Optimization:**
    - Conduct extensive hyperparameter optimization using techniques like Bayesian optimization or genetic algorithms. These methods can find the

optimal set of hyperparameters more efficiently than traditional grid search or random search.

-

7. **Incorporate Real-Time Data:**
    - Integrate real-time data to create a dynamic model that updates and improves as new data comes in. This can make the model more responsive to recent trends and changes in applicant behavior.
8. **Model Interpretability and Explainability:**
    - Focus on model interpretability by using techniques such as SHAP values or LIME. This will help in understanding which features are most influential in predicting credit card eligibility, making the model more transparent and trustworthy.

Recommendations

1. Implement Enhanced Eligibility Criteria:

- Based on the analysis, financial institutions should consider using income, education, and occupation as key determinants in their credit card eligibility criteria. Higher income levels, better educational backgrounds, and stable occupations were shown to be strong indicators of creditworthiness.

2. Focus on Financial Literacy Programs:

- Educated individuals have shown better eligibility, highlighting the importance of financial literacy. Financial institutions should invest in educational programs that promote financial literacy among potential applicants. This could improve their eligibility and reduce default rates.

3. Refine Marketing Strategies:

- Financial institutions can use the findings to refine their marketing strategies, targeting segments with higher eligibility rates such as individuals with stable jobs and higher education levels. Tailored marketing campaigns can increase approval rates and customer satisfaction.