Predicting Credit Card Approvals

<p> Presented by: Ayesha Shabbir Mirza, September 9, 2024 </p>

TABLE OF CONTENTS. (Modules)

TABLE OF CONTENTS. (Modules)

1. Introduction

1. Introduction

- Objective: Build a predictive model for credit card approvals.
- Tools Used: Python, Scikit-learn, XGBoost, Python, Pandas, NumPy, Joblib.
- Importance: Helps in automating and streamlining the credit approval process, reducing manual errors and biases.

2. Data Overview

2.1 Dataset Description

- Source: application_record.csv
- Total Records: 438,557
- Columns: 17

2.2 Columns and Data Types

- ID: Integer data type with 438,557 non-null values.
- CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, OCCUPATION_TYPE: Categorical variables with some having missing values, e.g., OCCUPATION_TYPE has 134,203 missing values.
- CNT_CHILDREN, AMT_INCOME_TOTAL, DAYS_BIRTH, DAYS_EMPLOYED, FLAG_MOBIL, FLAG_WORK_PHONE, FLAG_PHONE, FLAG_EMAIL, CNT_FAM_MEMBERS: Numerical data types.

2.3 Missing Values

- OCCUPATION_TYPE is the only column with missing values (134,203 missing).

3. Data Analysis

3.1 Statistical Summary

- ID: Ranges from 5,008,804 to 7,999,952 with a mean of approximately 6,022,176.
- CNT_CHILDREN: Most records have 0 children, with a maximum of 19.
- AMT_INCOME_TOTAL: Ranges from 26,100 to 6,750,000 with a mean income of 187,524.
- DAYS_BIRTH: Represents ages with a range from -25,201 to -7,489 days (negative due to days before a reference date), with a mean age of approximately 15,998 days.

- DAYS_EMPLOYED: Shows employment duration with values ranging from -17,531 to 365,243 days (positive values represent years, with a mean of 60,564 days).
- FLAG_MOBIL: Almost all records have this flag set to 1.
- FLAG_WORK_PHONE, FLAG_PHONE, FLAG_EMAIL: Various binary flags with mean values indicating frequency of occurrence.
- CNT_FAM_MEMBERS: Family size ranges from 1 to 20 with a mean of about 2.19.

3.2 Categorical Data Distribution

- CODE_GENDER: Contains 2 values (F and M) with a predominance of female records.
- FLAG_OWN_CAR: 2 values indicating ownership, with more records showing no car ownership.
- FLAG_OWN_REALTY: Indicates property ownership with a higher number of records showing property ownership.
- NAME_INCOME_TYPE: Includes various income types, with "Working" being the most common.
- OCCUPATION_TYPE: Covers 18 types, with "Laborers" being the most frequent occupation.

4. Data Preprocessing

4.1 Preprocessing

- Dropped Columns: ID (irrelevant for modeling).
- Filled Missing Values: OCCUPATION_TYPE with 'Unknown'.
- Feature Engineering: Added AGE_YEARS and adjusted DAYS_EMPLOYED.
- Created Target Variable: APPROVED based on criteria such as income, age, employment, family size, and asset ownership.

4.2 Criteria for Approval/Target Variable

- Income: > 150,000
- Age: 21 to 65
- Employment Duration: 1 to 40 years
- Family Size: <= 4 members
- Assets: Owns car or realty

5. Model Building

5.1 Preprocessing Pipelines

- Categorical Features:
- Imputation: Replaced missing values with the most frequent value.
- Encoding: Applied One-Hot Encoding.
- Numerical Features:
- Scaling: Standardized numerical features.

5.2 Model Pipeline

- Classifier: XGBoost Classifier
- Hyperparameter Tuning: Randomized Search with the following parameters:
- n_estimators: [100, 200, 300]
- learning_rate: [0.01, 0.05, 0.1, 0.2]
- max_depth: [3, 4, 5, 6]
- subsample: [0.7, 0.8, 0.9, 1.0]
- colsample_bytree: [0.7, 0.8, 0.9, 1.0]

6. Model Evaluation

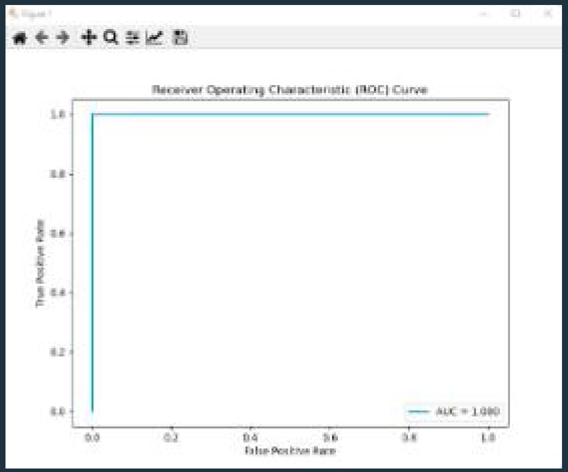
6.1 Performance Metrics

- Best Parameters:
  - subsample: 0.7
  - n_estimators: 200
  - max_depth: 6
  - learning_rate: 0.2
  - colsample_bytree: 0.9

- Confusion Matrix: Shows minimal misclassifications. Model made only 5 errors out of 87,712 predictions, which is exceptional.
- Classification Report: Precision, recall, and F1-score are all 1.00.
- AUC-ROC Score: Close to 1, indicating excellent performance.

6.2 ROC Curve & AUC Score

- Plot: Visualizes the trade-off between the True Positive Rate and False Positive Rate.
- The ROC curve shows a perfect AUC of 1.000, indicating that the model has excellent discriminatory power.



7. Model Testing Overview

7.1 Model Loading

- Loaded Model: trained_model.pkl
- Purpose: Predict credit card approvals based on user input.

7.2 User Input Collection

- Fields Collected:
  - Personal Information: Gender, Car Ownership, Real Estate Ownership
  - Income Information: Type, Amount
  - Employment Information: Duration
  - Family Information: Status, Number of Members, Number of Children

7.3 Example of Approved Credit Card

```
Enter your gender (M for Male, F for Female): F
Do you own a car? (Y for Yes, N for No): Y
Do you own any real estate (like a house or land)? (Y for Yes, N for No): Y
What type of income do you receive? (e.g., Salary from a job, Business income, etc.): BUSSINESS
What is your highest level of education? (e.g., Higher education, Secondary education, etc.): HIGHER
What is your current family status? (e.g., Married, Single, Divorced, etc.): MARRIED
What is your housing situation? (e.g., Do you live in a house, an apartment, or do you rent?): HOUSE
What is your occupation? (e.g., Laborer, Teacher, Sales staff, etc.): CEO
Enter your age (in years): 33
How many years have you been employed? 4
What is your total annual income? 1500000
How many family members live with you? 3
How many children do you have? 2
Credit card approved.
Probability of approval: 1.0000
PS E:\Qualcomm Internship\project_(ASM)>
* History restored
```

7.4 Example of disapproved Credit Card

```
PS E:\Qualcomm Internship\project_(ASM)> python predict_credit_card.py
Enter your gender (M for Male, F for Female): F
Do you own a car? (Y for Yes, N for No): N
Do you own any real estate (like a house or land)? (Y for Yes, N for No): N
What type of income do you receive? (e.g., Salary from a job, Business income, etc.): JOB
What is your highest level of education? (e.g., Higher education, Secondary education, etc.): SECONDARY
What is your current family status? (e.g., Married, Single, Divorced, etc.): SINGLE
What is your housing situation? (e.g., Do you live in a house, an apartment, or do you rent?): RENT
What is your occupation? (e.g., Laborer, Teacher, Sales staff, etc.): TEACHER
Enter your age (in years): 23
How many years have you been employed? 5
What is your total annual income? 200000
How many family members live with you? 2
Credit card disapproved.
Reasons for disapproval:
- No assets (car or real estate) owned.
Probability of approval: 0.0000
PS E:\Qualcomm Internship\project_(ASM)>
```

8. Model Evaluation after Testing

8.1 Performance Observations

- Accurate Predictions: Model correctly classified both cases.
- High Probability Score: Indicates high confidence in the approval prediction.

8.2 Disapproval Reasoning

- Useful Insights: Provides specific reasons for disapproval, which can be used for feedback and improvement.

Conclusion

- Summary: The model demonstrates strong performance and provides clear disapproval reasons.
- Future Work: Focus on refining the model and ensuring robust deployment.

References

- Taken help from various resources.

Thank you!