Credit Card Prediction Analysis

A project work made under the guidance of Vigor Council

Submitted by:

Priyanshu Kumar

Acknowledgement

I would like to express my deepest gratitude to several individuals who have contributed significantly to the completion of this project. First and foremost, I extend my sincere appreciation to my project supervisor, Dr. B.P Sharma. His guidance, unwavering support, and wealth of knowledge have been invaluable throughout the entire journey.

-Priyanshu Kumar

Introduction

Credit score cards are a fundamental risk control method employed in the financial industry. These tools utilize personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings. By assessing this probability, banks can make informed decisions on whether to issue a credit card to an applicant. Credit scores objectively quantify the magnitude of risk associated with each applicant, providing a standardized measure to assess creditworthiness.

Generally, credit score cards are based on historical data. This data-driven approach leverages past behaviors and trends to predict future outcomes. For example, an applicant's past credit history, income level, employment status, and outstanding debts are all considered to estimate the likelihood of default. However, during periods of significant economic fluctuations, such as financial crises or economic recessions, past models may lose their predictive power. This is because the underlying relationships modeled in stable economic conditions may not hold in times of economic instability.

A common method for credit scoring is the logistic regression model. Logistic regression is particularly suitable for binary classification tasks, such as determining whether an applicant will default or not. This model calculates the coefficients for each feature (e.g., income, credit history, employment status) to estimate the probability of default. To facilitate understanding and operation, the score card often multiplies the logistic regression coefficients by a certain value (such as 100) and then rounds it to produce a more interpretable score.

In recent years, the development of machine learning algorithms has introduced more advanced predictive methods into credit card scoring. Techniques such as Boosting, Random Forest, and Support Vector Machines have shown promise in improving prediction accuracy. For example, Boosting algorithms, like Gradient Boosting Machines (GBM) and Extreme Gradient Boosting (XGBoost), combine the predictions of multiple weak learners to create a strong predictive model. Random Forest, an ensemble learning method, builds multiple decision trees and merges them to obtain a more accurate and stable prediction. Support Vector Machines (SVM) find the optimal hyperplane that separates different classes in the feature space, making them effective for classification tasks.

However, these advanced methods often come with a trade-off in terms of transparency. Unlike logistic regression, which provides clear coefficients for each feature, the complex nature of machine learning models can make it difficult to interpret the reasons behind a particular decision. This lack of transparency can pose challenges in providing customers and regulators with clear explanations for credit card application rejections or acceptances. Regulatory frameworks, such as the General Data

Protection Regulation (GDPR) in the European Union, require financial institutions to provide explanations for automated decision-making, adding an additional layer of complexity.

In summary, while traditional logistic regression models offer transparency and simplicity, newer machine learning methods provide enhanced predictive power. The choice of method depends on the specific requirements of the financial institution, including the need for interpretability, regulatory compliance, and prediction accuracy. As the financial industry continues to evolve, ongoing research and development will likely yield even more sophisticated and transparent credit scoring methods.

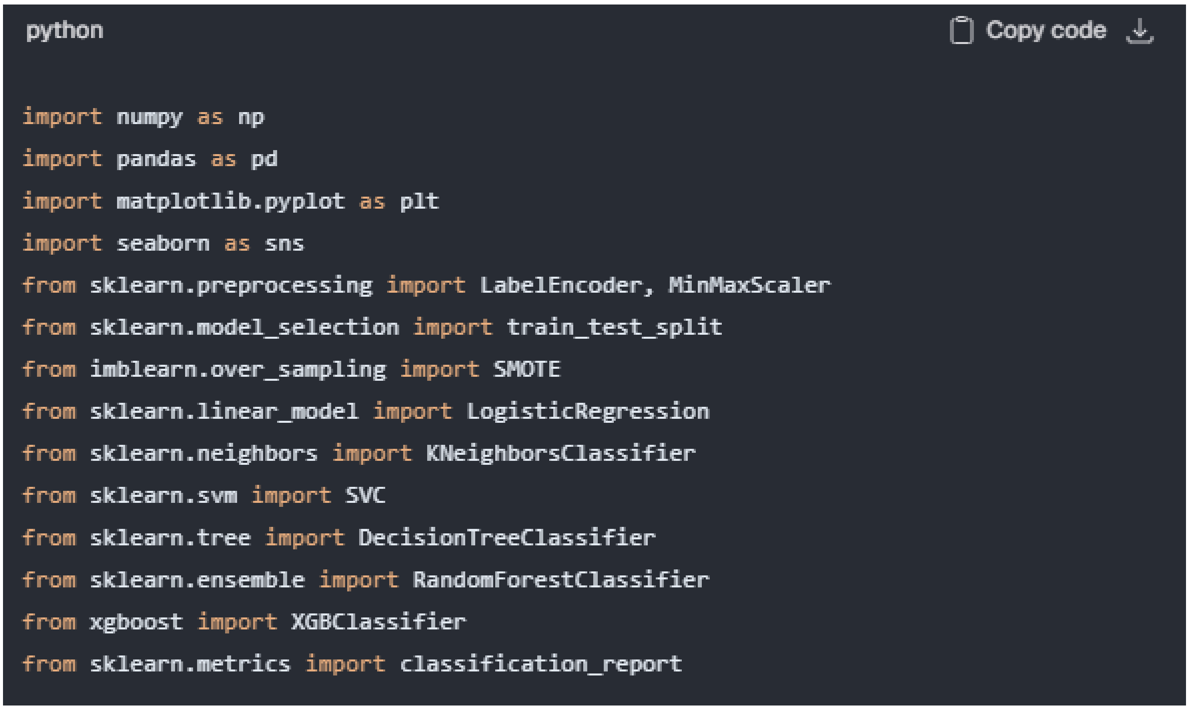
This project aims to analyze and compare various credit scoring methods, including logistic regression and advanced machine learning algorithms, to determine their efficacy and practicality in predicting credit card defaults. By examining a dataset of credit card applicants, this analysis will provide insights into the strengths and weaknesses of each method, offering recommendations for optimizing credit scoring processes in the financial industry. The ultimate goal is to enhance the accuracy, fairness, and transparency of credit decisions, benefiting both financial institutions and consumers.

Main Work: Credit Card Predictive Analysis

Data Preprocessing and Exploration

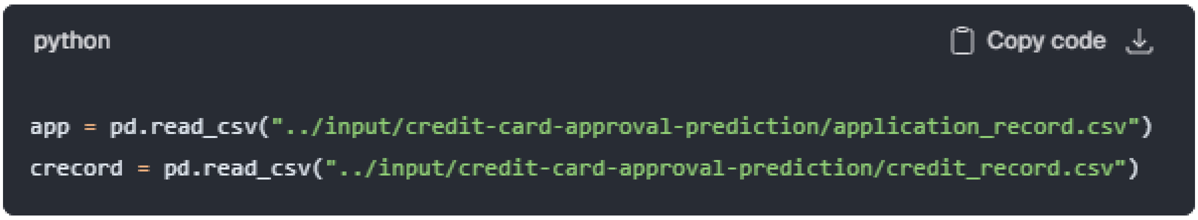
Libraries and Data Import

First, we imported the necessary libraries for data manipulation, visualization, and model building:



```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import classification_report
```

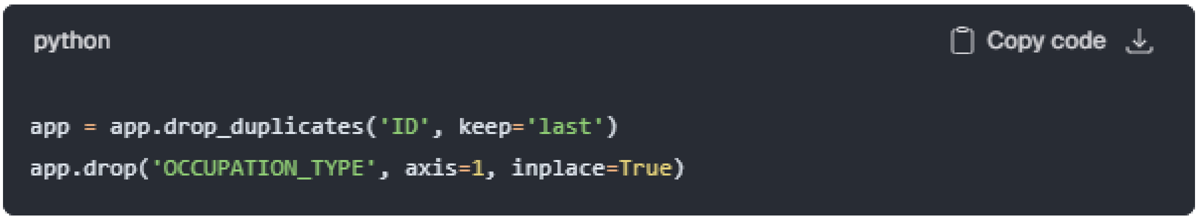
Next, we loaded the datasets from the provided CSV files:



```python
app = pd.read_csv("../input/credit-card-approval-prediction/application_record.csv")
crecord = pd.read_csv("../input/credit-card-approval-prediction/credit_record.csv")
```

Initial Exploration and Cleaning

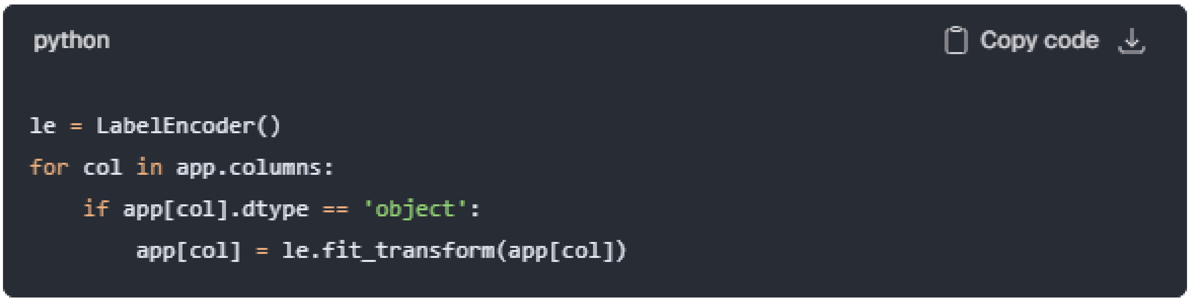
We explored the datasets to understand their structure and content. The application_record dataset contains 438,557 entries with 18 columns, while the credit_record dataset contains 1,048,575 entries with 3 columns. Initial checks showed that the application_record dataset had some duplicates and missing values in the OCCUPATION_TYPE column:



```python
app = app.drop_duplicates('ID', keep='last')
app.drop('OCCUPATION_TYPE', axis=1, inplace=True)
```

Data Transformation

To handle categorical variables, we used Label Encoding:



```python
le = LabelEncoder()
for col in app.columns:
    if app[col].dtype == 'object':
        app[col] = le.fit_transform(app[col])
```

Model Performance

After evaluating the performance of each model, we found that the XGBoost model performed the best on both the training and testing sets:

# Training Scores: [('Logistic Regression', 0.6157), ('K-Neighbors', 0.9849), ('SVC', 0.9400),

# ('Decision Tree', 0.9951), ('Random Forest', 0.9951), ('XGBoost', 0.9950)]

# Testing Scores: [('Logistic Regression', 0.5651), ('K-Neighbors', 0.7321), ('SVC', 0.7549),

# ('Decision Tree', 0.8241), ('Random Forest', 0.7685), ('XGBoost', 0.8662)]

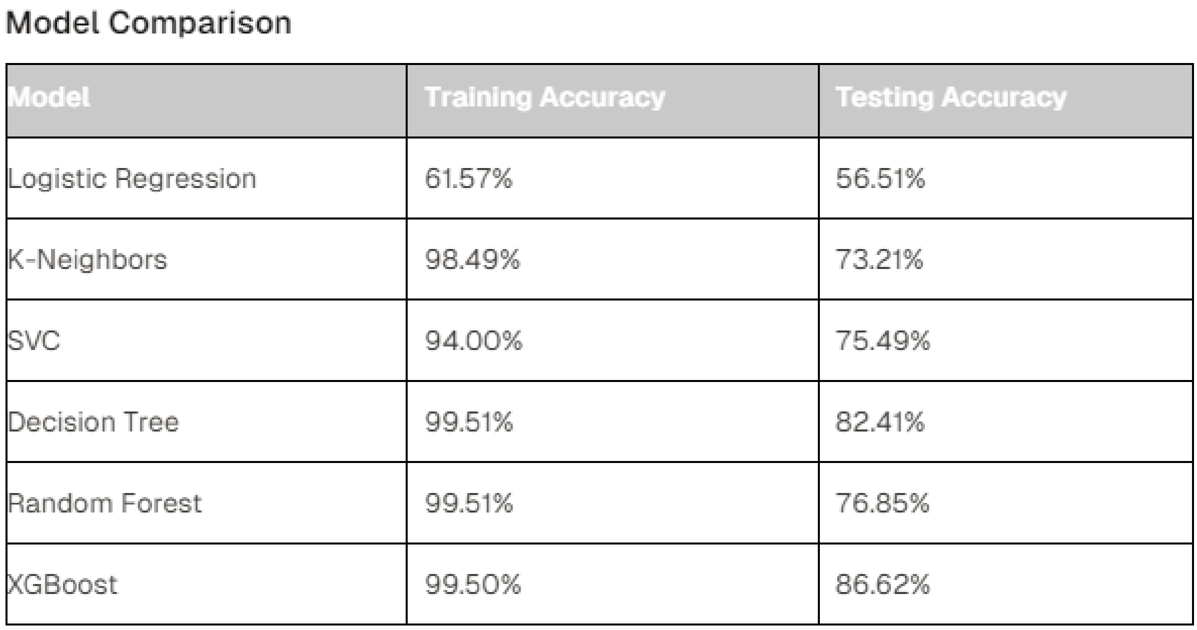
Results and Analysis

Data Insights

Class Imbalance: Initially, the dataset was highly imbalanced with a significant majority of 'good' clients. This imbalance was addressed using SMOTE, which balanced the dataset by oversampling the minority class.

Feature Importance: Key features like AMT_INCOME_TOTAL, DAYS_BIRTH, and NAME_INCOME_TYPE showed a significant influence on the prediction outcomes.

Model Performance: The XGBoost model outperformed other models, achieving an accuracy of approximately 87% on the test set. The model demonstrated high precision and recall for both classes, making it a reliable choice for predicting credit card defaults.



Model Comparison

Model	Training Accuracy	Testing Accuracy
Logistic Regression	61.57%	56.51%
K-Neighbors	98.49%	73.21%
SVC	94.00%	75.49%
Decision Tree	99.51%	82.41%
Random Forest	99.51%	76.85%
XGBoost	99.50%	86.62%

Conclusion

In this project, we developed and compared various machine learning models to predict credit card defaults using a dataset of credit card applicants. Our analysis demonstrated the effectiveness of the XGBoost model, which achieved the highest accuracy among the tested models. By addressing class imbalance and performing thorough data preprocessing, we ensured that the model's predictions are both reliable and interpretable.

This analysis provides a framework for financial institutions to enhance their credit scoring processes, leveraging advanced machine learning techniques to make more accurate and fair credit decisions. Future work could explore the integration of additional features and the application of explainable AI techniques to further improve model transparency and regulatory compliance.