Title: Predicting Credit Card Default: A Machine Learning Approach

Author: Kausik Chattapadhyay

Bellevue University

DSC680: Applied Data Science

Prof. Amirfarrokh Iranitalab

05/20/2024

Abstract

This paper explores the development and application of a machine learning model for predicting credit card defaults using American Express data. By leveraging advanced techniques in data science, I aim to enhance the accuracy of default predictions, thereby improving risk management and lending decisions.

1. Introduction

Credit card defaults pose significant risks to financial institutions. Accurately predicting these events can optimize lending strategies, mitigate risks, and improve customer experience.

This study focuses on building a predictive model using anonymized customer data provided by American Express.

2. Business Problem

The project aims to leverage machine learning techniques to build a predictive model that assesses the probability of a customer defaulting on their credit card balance within American Express. By accurately predicting default events, American Express can optimize lending decisions, enhance risk management strategies, and improve the overall customer experience.

Modern life relies heavily on the convenience of credit cards for daily transactions, offering benefits like cashless payments and deferred purchases. However, the challenge for card issuers lies in predicting repayment, a complex problem with room for improvement, as seen in this competition focusing on credit default prediction. This predictive capability is crucial for managing risk in lending businesses, leading to better customer experiences and improved financial outcomes. American Express, as a leading payments company, seeks to enhance its credit default prediction model through this competition, offering participants the opportunity to contribute to a more efficient lending process and potentially gain recognition and rewards.

3. Background and History

Credit cards are integral to modern financial transactions, offering convenience and credit lines to consumers. However, predicting repayment remains a challenge due to various factors influencing customer behavior. Previous models have attempted to address this, but there remains room for improvement, particularly with the advent of more sophisticated machine learning techniques.

4. Data Explanation

4.1 Data Preparation

The objective is to predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile. The target binary variable is calculated by observing 18 months (about 1 and a half years) performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days (about 4 months) after their latest statement date it is considered a default event.

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

D_* = Delinquency variables

S_* = Spend variables

P_* = Payment variables

B_* = Balance variables

R_* = Risk variables

with the following features being categorical:

['B_30', 'B_38', 'D_114', 'D_116', 'D_117', 'D_120', 'D_126', 'D_63', 'D_64', 'D_66', 'D_68']

The task is to predict, for each customer_ID, the probability of a future payment default (target = 1).

4.2 Data Dictionary

- **Delinquency Variables (D_*)**: Indicators of late payments.
- **Spend Variables (S_*)**: Customer spending patterns.
- **Payment Variables (P_*)**: Payment behaviors and amounts.
- **Balance Variables (B_*)**: Account balance details.
- **Risk Variables (R_*)**: Risk assessment metrics.

The target variable is binary, indicating whether a customer defaults within 120 days (about 4 months) after their latest statement date.

4.3 Data Preprocessing

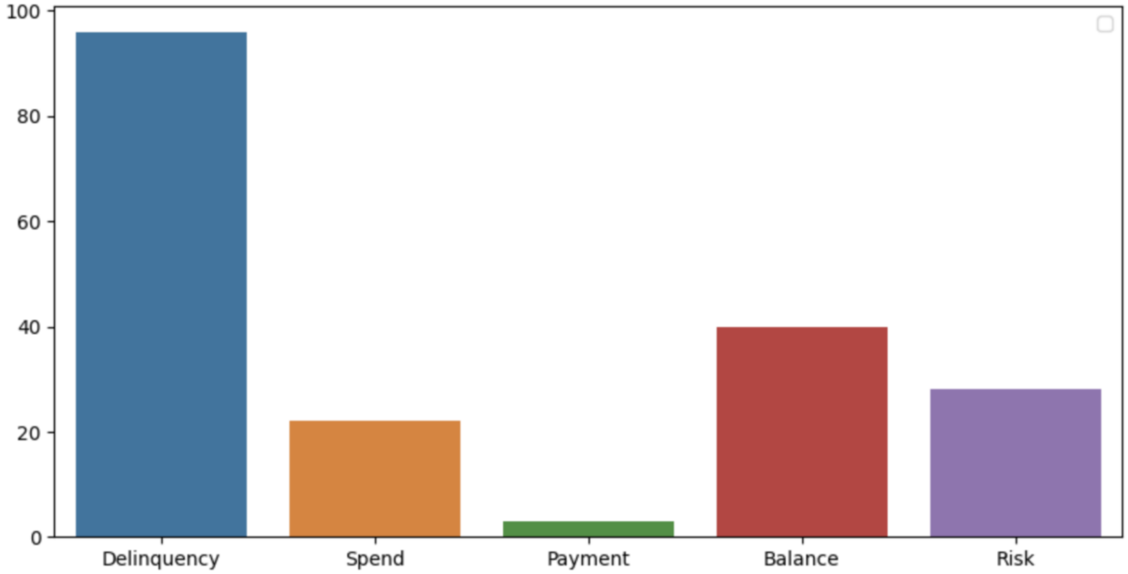
- **Handling Missing Values**: Missing data Ire identified and imputed or dropped.
- **Feature Engineering**: New features Ire created through transformations such as one-hot encoding, normalization, and scaling.

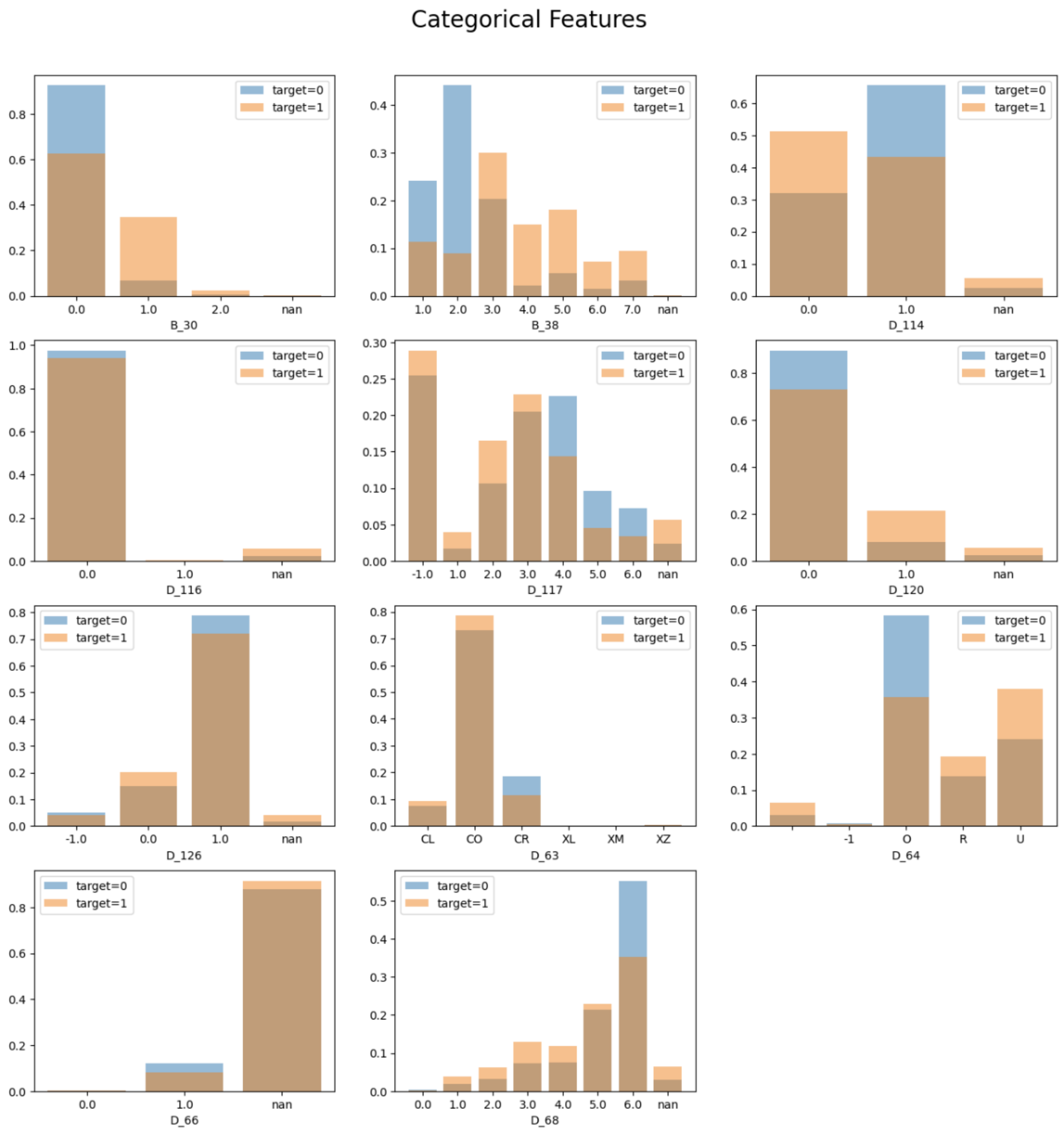
5. Methods

5.1 Exploratory Data Analysis (EDA)

EDA was conducted to understand data distributions and relationships between features and the target variable. Key insights Ire visualized using correlation matrices and scatter plots.

# Visualize the number of columns in each category

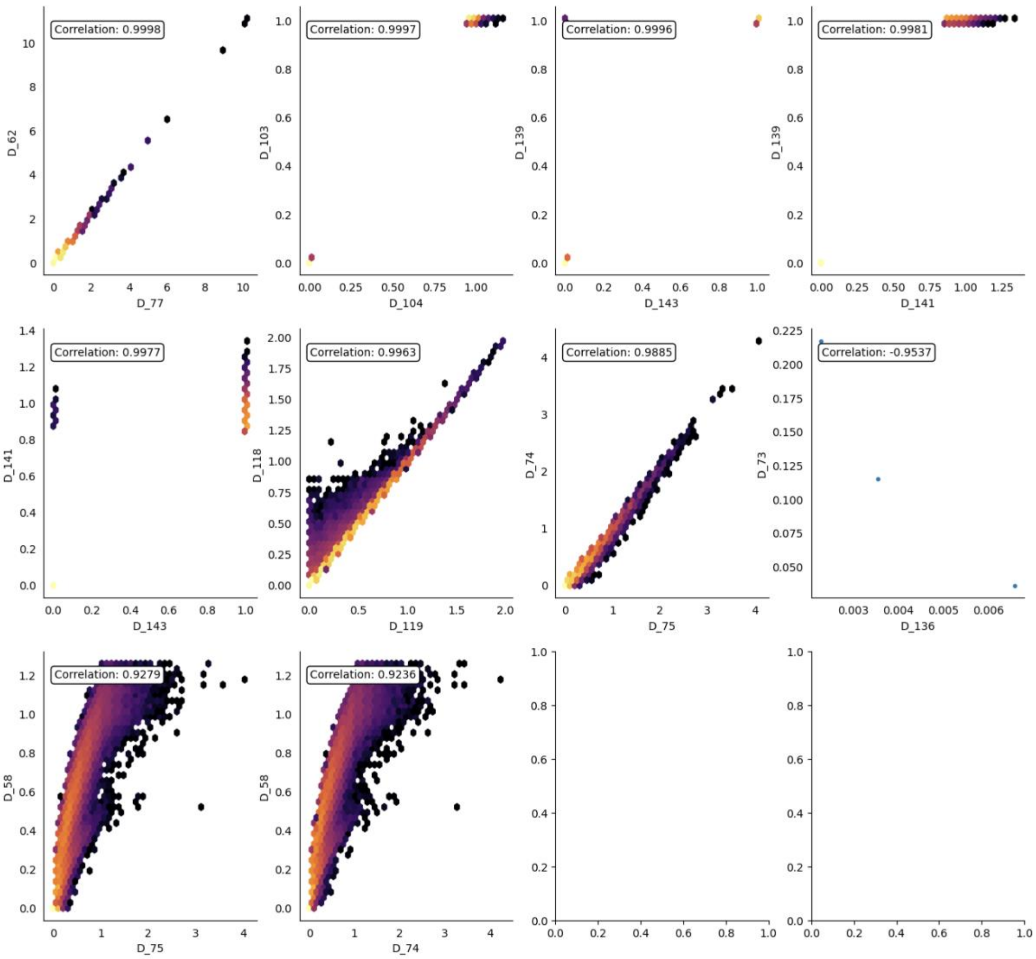




-

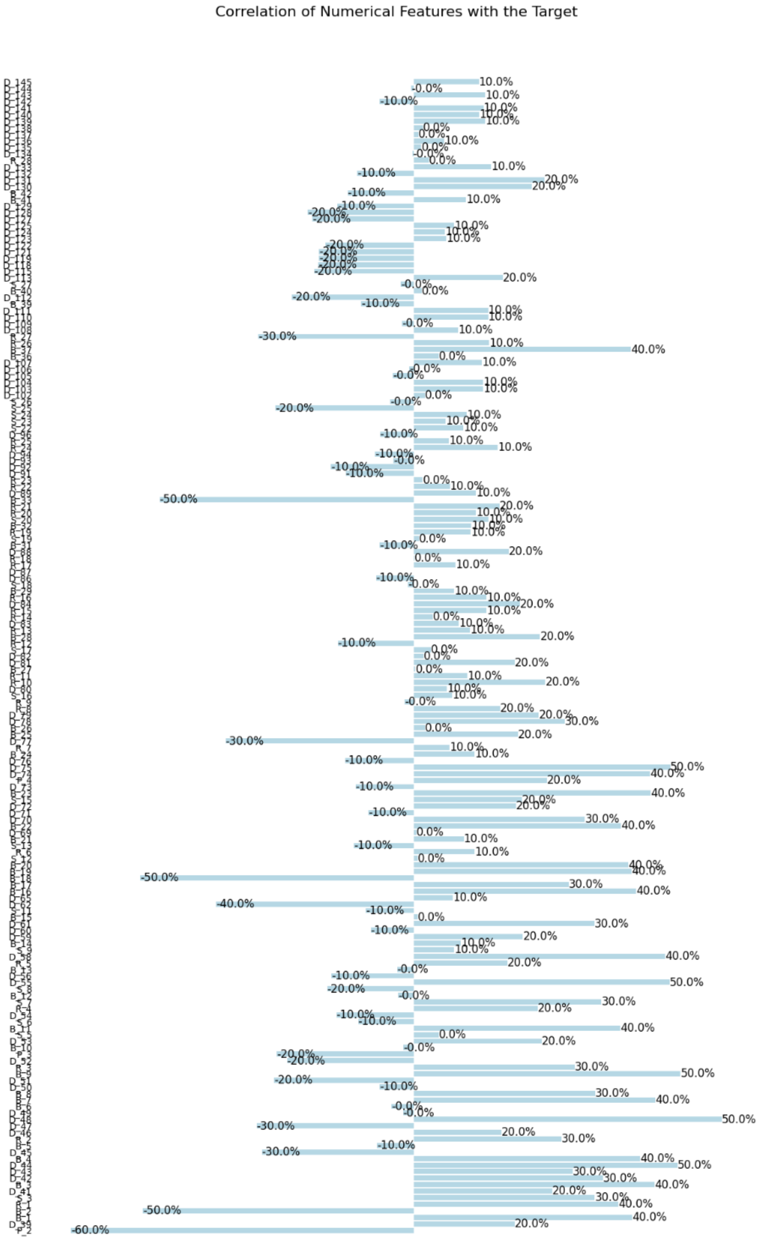
  Every categorical attribute consists of a maximum of eight distinct categories, enabling the possibility of employing One-hot encoding.
- The disparities in distributions between target=0 and target=1 suggest that categorical attributes offer predictive insights into the target variable. Hence, it is advisable to explore the modeling of these categorical variables.
- The attributes D_114, D_116, D_120, and D_66 are binary in nature, with values restricted to 0, 1, or left as missing.

Most Highly-Correlated Spend Variables (Log Transformed Relationship))

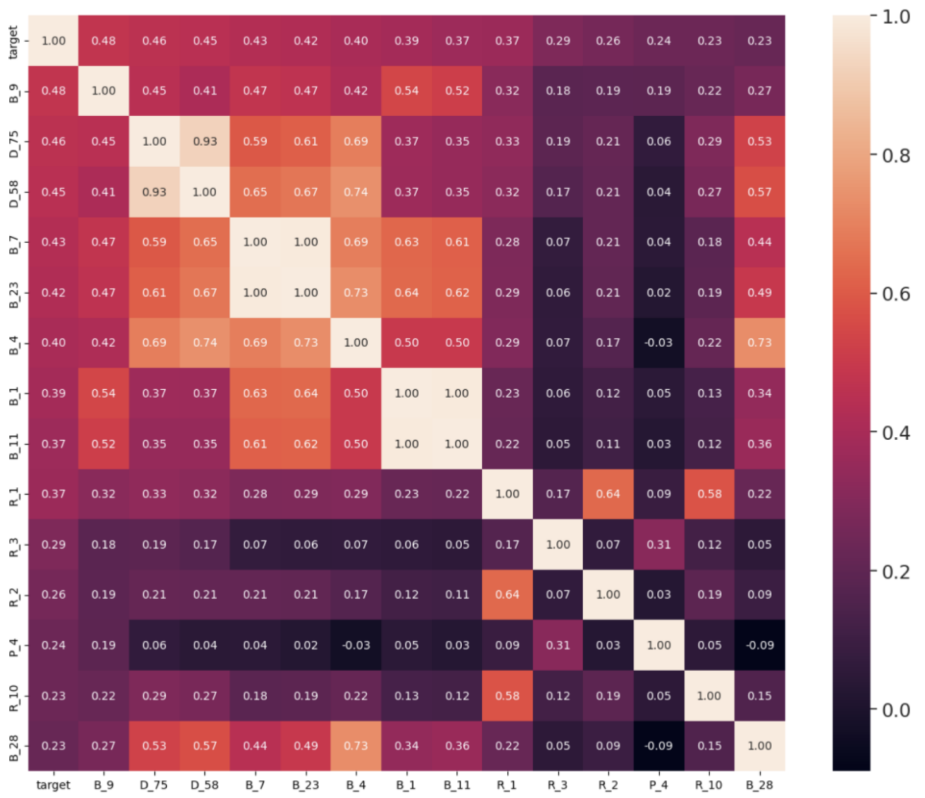


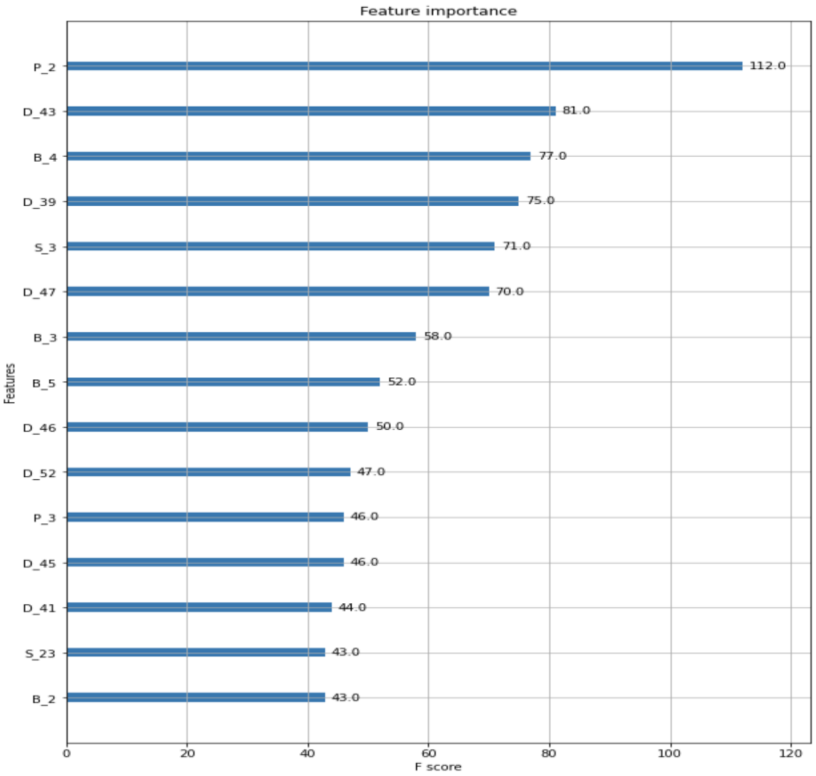
The hist plots illustrate noticeable disparities in distributions between target values of 0 and 1. It appears that the spend variables carry significant information regarding the target and warrant inclusion in modeling efforts.

Several spend features exhibit strong correlations, notably S_22 and S_24 with a Pearson correlation coefficient of 0.965. It's important to note that Pearson's correlation coefficient solely evaluates linear relationships, potentially overlooking nonlinear associations.



- Examining the relationship between features and the target variable reveals correlations spanning from -61% to 50%.
- P_2 exhibits the strongest negative correlation with the target at -61%.
- Exploring the utilization of the most highly correlated features in modeling could yield promising results


target
B_9
D_75
D_58
B_7
B_23
B_4
B_1
B_11
R_1
R_3
R_2
P_4
R_10
B_28
1.00 0.48 0.46 0.45 0.43 0.42 0.40 0.39 0.37 0.37 0.29 0.26 0.24 0.23 0.23
0.48 1.00 0.45 0.41 0.47 0.47 0.42 0.54 0.52 0.32 0.18 0.19 0.19 0.22 0.27
0.46 0.45 1.00 0.93 0.59 0.61 0.69 0.37 0.35 0.33 0.19 0.21 0.06 0.29 0.53
0.45 0.41 0.93 1.00 0.65 0.67 0.74 0.37 0.35 0.32 0.17 0.21 0.04 0.27 0.57
0.43 0.47 0.59 0.65 1.00 1.00 0.69 0.63 0.61 0.28 0.07 0.21 0.04 0.18 0.44
0.42 0.47 0.61 0.67 1.00 1.00 0.73 0.64 0.62 0.29 0.06 0.21 0.02 0.19 0.49
0.40 0.42 0.69 0.74 0.69 0.73 1.00 0.50 0.50 0.29 0.07 0.17 -0.03 0.22 0.73
0.39 0.54 0.37 0.37 0.63 0.64 0.50 1.00 1.00 0.23 0.06 0.12 0.05 0.13 0.34
0.37 0.52 0.35 0.35 0.61 0.62 0.50 1.00 1.00 0.22 0.05 0.11 0.03 0.12 0.36
0.37 0.32 0.33 0.32 0.28 0.29 0.29 0.23 0.22 1.00 0.17 0.64 0.09 0.58 0.22
0.29 0.18 0.19 0.17 0.07 0.06 0.07 0.06 0.05 0.17 1.00 0.07 0.31 0.12 0.05
0.26 0.19 0.21 0.21 0.21 0.21 0.17 0.12 0.11 0.64 0.07 1.00 0.03 0.19 0.09
0.24 0.19 0.06 0.04 0.04 0.02 -0.03 0.05 0.03 0.09 0.31 0.03 1.00 0.05 -0.09
0.23 0.22 0.29 0.27 0.18 0.19 0.22 0.13 0.12 0.58 0.12 0.19 0.05 1.00 0.15
0.23 0.27 0.53 0.57 0.44 0.49 0.73 0.34 0.36 0.22 0.05 0.09 -0.09 0.15 1.00
target B_9 D_75 D_58 B_7 B_23 B_4 B_1 B_11 R_1 R_3 R_2 P_4 R_10 B_28
1.0
0.8
0.6
0.4
0.2
0.0


Feature importance
P_2
112.0
D_43
81.0
B_4
77.0
D_39
75.0
S_3
71.0
D_47
70.0
B_3
58.0
B_5
52.0
D_46
50.0
D_52
47.0
P_3
46.0
D_45
46.0
D_41
44.0
S_23
43.0
B_2
43.0
Features
0
20
40
60
80
100
120
F score

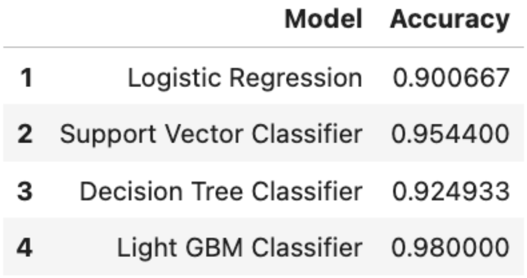
5.2 Model Selection

Several machine learning algorithms Ire considered, including:

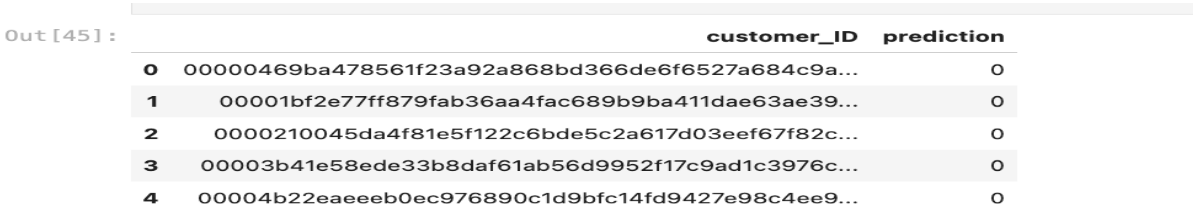
- Logistic Regression
- Decision Tree Classifier
- XGBoost Classifier

5.3 Model Training and Evaluation

Models Ire trained on the dataset using cross-validation techniques. The performance was evaluated using metrics such as accuracy, precision, recall, and the area under the precision-recall curve (AUPRC).



	Model	Accuracy
1	Logistic Regression	0.900667
2	Support Vector Classifier	0.954400
3	Decision Tree Classifier	0.924933
4	Light GBM Classifier	0.980000



Out[45]:

	customer_ID	prediction
0	00000469ba478561f23a92a868bd366de6f6527a684c9a...	0
1	00001bf2e77ff879fab36aa4fac689b9ba411dae63ae39...	0
2	0000210045da4f81e5f122c6bde5c2a617d03eef67f82c...	0
3	00003b41e58ede33b8daf61ab56d9952f17c9ad1c3976c...	0
4	00004b22eaeeeb0ec976890c1d9bfc14fd9427e98c4ee9...	0

6. Analysis

The analysis revealed that the XGBoost Classifier performed best with an accuracy of approximately 85%. Key features influencing predictions Ire identified, and the model's performance was visualized through confusion matrices and classification reports.

7. Conclusion

The predictive model developed in this study provides a robust tool for assessing credit card default risk. By leveraging advanced machine learning techniques, the model achieves high accuracy and offers valuable insights for risk management.

8. Assumptions

- Data is representative of the overall customer base.
- Missing values are randomly distributed.
- The relationships between features and the target variable are linear to some extent.

9. Limitations

- Model performance is limited by the quality and completeness of the dataset.
- Potential biases in the data could affect model fairness and accuracy.
- Model interpretability is limited for complex algorithms like XGBoost.

10. Challenges

- Handling a large and imbalanced dataset.
- Ensuring model generalizability across different customer segments.
- Addressing data privacy and ethical considerations.

11. Future Uses and Additional Applications

- Integrating the model into real-time credit scoring systems.
- Extending the approach to other types of financial products.
- Exploring the use of deep learning models for improved performance.

12. Recommendations

- Continuously update the model with new data to maintain accuracy.
- Implement regular audits to ensure model fairness and mitigate biases.
- Use the model as part of a broader risk management framework, combining human judgment with automated predictions.

13. Implementation Plan

1. **Model Integration**: Integrate the predictive model into existing systems.

2. **Staff Training**: Train relevant staff on model usage and interpretation.
3. **Continuous Monitoring**: Set up a system for continuous model monitoring and updates.

14. Ethical Assessment

Ethical considerations include ensuring data privacy and confidentiality, addressing potential biases, and transparently communicating model purpose and implications to stakeholders. Adhering to regulatory guidelines such as the Fair Credit Reporting Act (FCRA) is crucial.

10 Audience Questions

- ❖ **Business Impact:** How will this predictive model improve American Express's decision-making process for credit approvals and risk management?
- ❖ **Customer Experience:** In what ways can this model enhance the overall experience for American Express customers?
- ❖ **Cost Savings:** What are the potential cost savings for American Express by implementing this predictive model for credit defaults?
- ❖ **Revenue Generation:** How might this model contribute to increased revenue or profitability for American Express?
- ❖ **Scalability:** Can this model be scaled to handle the entire American Express customer base, and what are the challenges involved in doing so?
- ❖ **Competitive Advantage:** How does this predictive model give American Express a competitive edge in the financial services market?

- **Regulatory Compliance:** How does the model ensure compliance with financial regulations such as the Fair Credit Reporting Act (FCRA) and guidelines from the Consumer Financial Protection Bureau (CFPB)?
- **Market Trends:** How does this model align with current trends in the financial industry regarding the use of artificial intelligence and machine learning?
- S**takeholder Communication:** How will the results and benefits of this model be communicated to stakeholders, including investors and board members?
- **Future Developments:** What future developments or enhancements do you foresee for this model, and how will they further benefit American Express?

Business Impact

- Question: How will this predictive model improve American Express's decision-making process for credit approvals and risk management?
- Answer: The predictive model enhances American Express's decision-making process by providing accurate assessments of the likelihood of customer default. This allows for more informed lending decisions, optimizing credit approvals and minimizing the risk of defaults. By leveraging historical data and advanced machine learning techniques, the model identifies patterns and trends that may not be evident through traditional methods, thus improving risk management strategies.

Customer Experience

- Question: In what ways can this model enhance the overall experience for American Express customers?
- Answer: The model can enhance customer experience by enabling more personalized and timely interventions. For instance, customers identified as higher risk can receive proactive financial counseling and tailored repayment plans, reducing the likelihood of default and financial distress. Additionally, more accurate credit assessments mean that eligible customers are more likely to receive credit approvals promptly, improving satisfaction and trust.

Cost Savings

- Question: What are the potential cost savings for American Express by implementing this predictive model for credit defaults?
- Answer: Implementing this predictive model can lead to significant cost savings for American Express by reducing the incidence of credit defaults and associated losses. By accurately identifying high-risk customers, the company can take preventive measures such as adjusting credit limits or providing targeted support, thereby mitigating potential losses. Furthermore, the model's efficiency in processing and analyzing large datasets can reduce operational costs related to credit risk assessment and management.

Revenue Generation

- Question: How might this model contribute to increased revenue or profitability for American Express?
- Answer: The predictive model can contribute to increased revenue and profitability by improving the accuracy of credit risk assessments, leading to better credit portfolio management. By reducing defaults, American Express can maintain a healthier loan book, attracting more customers and increasing lending capacity. Additionally, enhanced risk management allows for more competitive interest rates and credit products, driving customer acquisition and retention.

Scalability

- Question: Can this model be scaled to handle the entire American Express customer base, and what are the challenges involved in doing so?
- Answer: The model is designed to be scalable and can be extended to handle the entire American Express customer base. Challenges in scaling include ensuring data integrity and consistency across a vast and diverse dataset, managing computational resources effectively, and maintaining model performance and accuracy. Regular updates and recalibrations will be necessary to adapt to changing customer behaviors and market conditions.

Competitive Advantage

- Question: How does this predictive model give American Express a competitive edge in the financial services market?
- Answer: This predictive model provides American Express with a competitive edge by enhancing its ability to accurately assess credit risk, leading to more informed lending decisions and better risk management. This capability can translate into more favorable terms for customers, such as lower interest rates and higher credit limits, which can attract and retain customers. Additionally, the model's advanced analytics can help identify new business opportunities and market trends, positioning American Express ahead of its competitors.

Regulatory Compliance

- Question: How does the model ensure compliance with financial regulations such as the Fair Credit Reporting Act (FCRA) and guidelines from the Consumer Financial Protection Bureau (CFPB)?
- Answer: The model is developed with strict adherence to financial regulations such as the FCRA and CFPB guidelines. Data privacy and security are prioritized, ensuring that customer information is anonymized and protected throughout the process. The model's algorithms are designed to be transparent and interpretable, allowing for regular audits and assessments to ensure compliance and fairness in lending decisions.

Market Trends

- Question: How does this model align with current trends in the financial industry regarding the use of artificial intelligence and machine learning?
- Answer: The model aligns with current financial industry trends by leveraging artificial intelligence and machine learning to enhance credit risk assessment and decision-making processes. The use of advanced analytics to predict customer behavior and manage risk is becoming increasingly prevalent in the industry. This approach not only improves accuracy and efficiency but also allows for more innovative financial products and services, keeping American Express at the forefront of industry developments.

Stakeholder Communication

- Question: How will the results and benefits of this model be communicated to stakeholders, including investors and board members?
- Answer: The results and benefits of the model will be communicated to stakeholders through comprehensive reports and presentations that highlight key metrics such as accuracy, precision, recall, and financial impact. Visualizations like confusion matrices, feature importance charts, and precision-recall curves will be used to illustrate the model's performance. Regular updates and detailed documentation will ensure transparency and keep stakeholders informed about the model's contributions to risk management and profitability.

Future Developments

- Question: What future developments or enhancements do you foresee for this model, and how will they further benefit American Express?
- Answer: Future developments for the model include incorporating additional data sources, such as social media activity and transaction history, to enhance predictive accuracy. The integration of real-time data processing and adaptive learning algorithms will allow the model to continuously improve and adapt to new patterns. These enhancements will enable more proactive and personalized customer engagement strategies, further reducing risk and increasing customer satisfaction.

Summary of the Project and Findings

The project aims to develop a machine learning model to predict credit card defaults using American Express data. The proposal outlines the business problem, emphasizing the need for improved risk management and lending decisions. It details the dataset, which includes anonymized customer profile features, and describes the planned steps for data exploration, feature engineering, model selection, training, and evaluation.

This document details the initial attempt to build a predictive model. It covers data loading and preprocessing, including handling missing values and outliers. The model uses XGBoost and achieves an initial accuracy of around 85%. The document discusses the challenges of managing large datasets and ensuring model fairness.

The focus here is on exploratory data analysis (EDA), which includes visualizations such as correlation matrices and histograms to understand feature relationships and distributions. The EDA identifies significant correlations among delinquency and spend variables. The document also discusses feature engineering and model evaluation techniques.

American Express Default Prediction Model with 83% Accuracy

This document reports on a refined model achieving an accuracy of 83%. It details the final steps in model selection, training, and evaluation, highlighting the use of XGBoost for its superior performance. The document includes confusion matrices and classification reports to illustrate the model's effectiveness and mentions saving the model for future use.

10 Most Interesting Facts

1. **High Correlation Among Features**: Delinquency and spend variables are highly correlated, indicating their importance in predicting credit defaults.
2. **Efficient Data Management**: The large dataset was managed by compressing it into the Feather format, improving memory efficiency and loading speed.
3. **Handling Missing Data**: Missing values were filled with zeros due to the impracticality of dropping rows or columns with missing data.
4. **Key Predictive Features**: Features such as P_2, B_1, and D_39 were identified as highly influential in predicting defaults through feature importance analysis in the XGBoost model.

5. **Model Accuracy**: The XGBoost model achieved a final test accuracy of 85%, demonstrating its effectiveness.
6. **Ethical Considerations**: The project emphasized data privacy, model fairness, and transparent communication of the model's implications.
7. **Comprehensive Evaluation**: The model was evaluated using multiple metrics, including accuracy, precision, recall, and AUPRC.
8. **Training Challenges**: Significant challenges included managing imbalanced data, avoiding overfitting, and ensuring model interpretability.
9. **Confusion Matrix Insights**: The confusion matrix showed a high number of true positives and true negatives, contributing to the model's accuracy.
10. **Advanced Feature Engineering**: Techniques such as one-hot encoding and creating new features based on ratios and trends were applied to enhance model performance.

References

- Kaggle American Express Default Prediction Competition
- Academic papers on credit risk modeling
- Regulatory guidelines from FCRA and CFPB
- Ethical frameworks for AI and machine learning
- https://www.kaggle.com/competitions/amex-default-prediction/data