# American Express - Default Prediction

Title: Predicting Credit Card Default: A Machine Learning Approach

Author: Kausik Chattapadhyay

Institution: Bellevue University

Course: DSC680: Applied Data Science

Date: **05/25/2024**

# Abstract

Objective: Develop and apply a machine learning model for predicting credit card defaults using American Express data

Goal: Enhance accuracy of default predictions to improve risk management and lending decisions

# Introduction

Problem: Credit card defaults pose significant risks to financial institutions

Solution: Accurate prediction can optimize lending strategies, mitigate risks, and improve customer experience

Focus: Build a predictive model using anonymized customer data from American Express

# Business Problem

Aim: Leverage machine learning to assess the probability of customer default on credit card balance

Benefits: Optimize lending decisions, enhance risk management strategies, and improve customer experience

Context: Credit cards offer convenience but predicting repayment is challenging

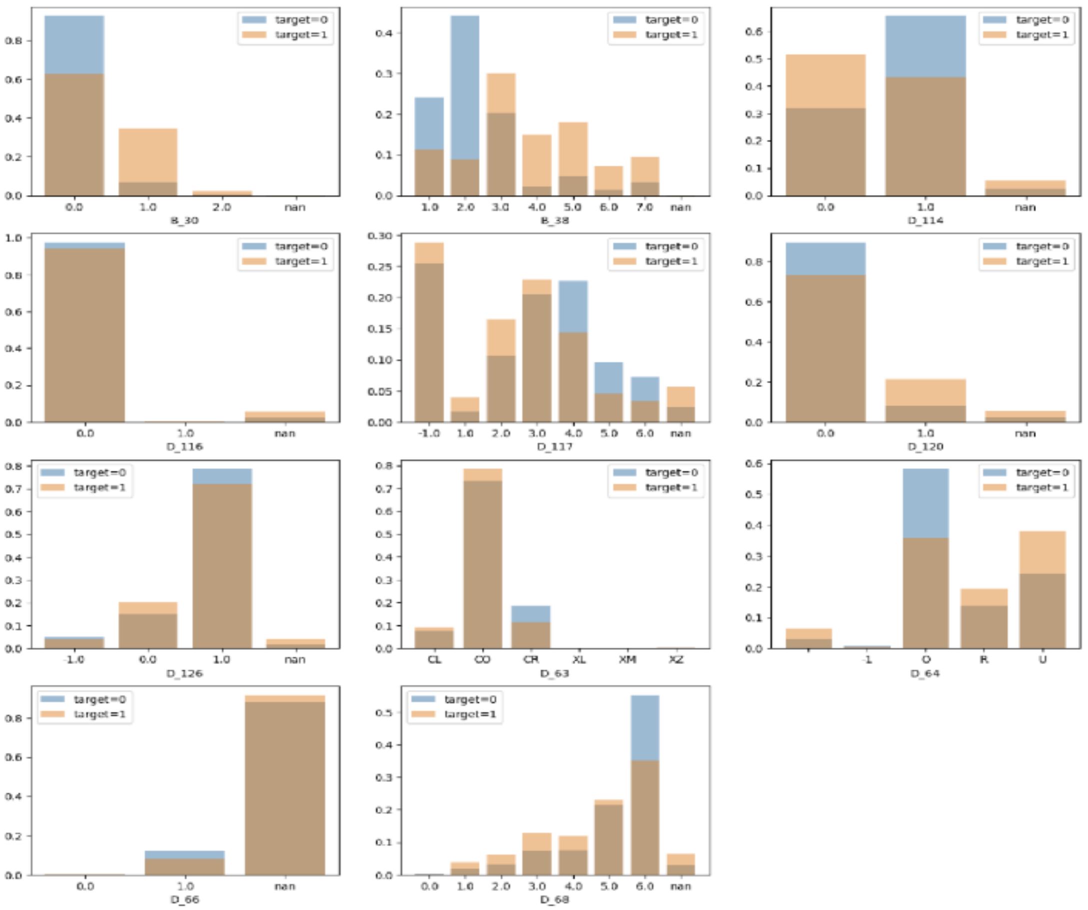
# Data Explanation

Objective: Predict the probability of future credit card payment defaults

Data: Anonymized and normalized customer profile features, categorized into delinquency, spend, payment, balance, and risk variables

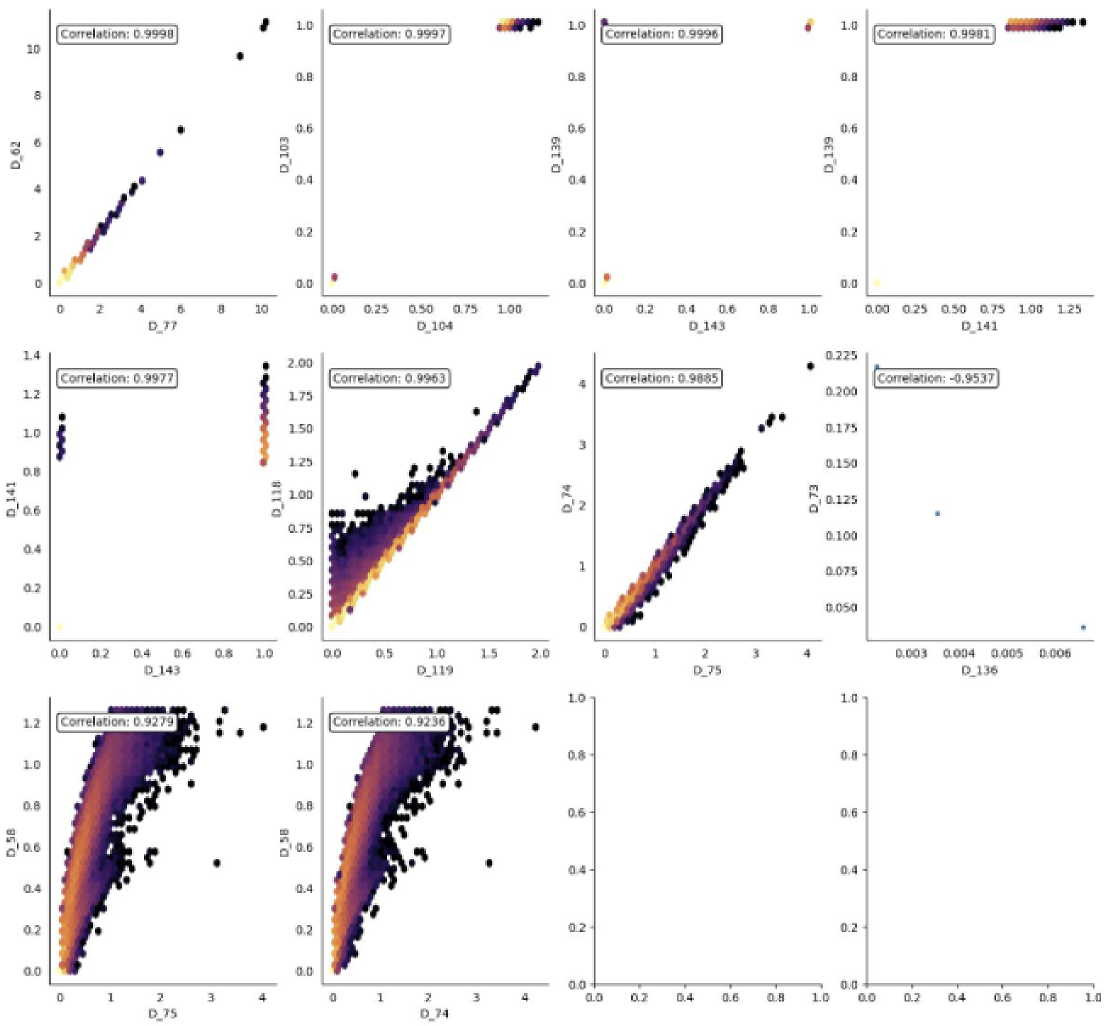
Target Variable: Binary indicator of default within 120 days after the latest statement date

Categorical Features



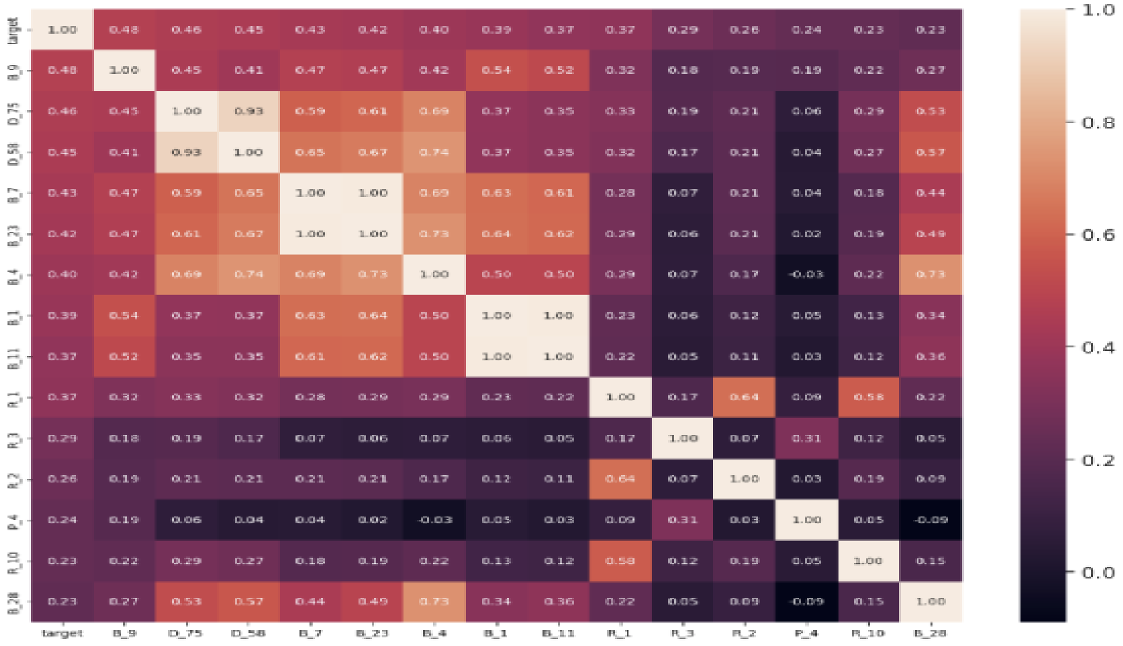

- Every categorical attribute consists of a maximum of eight distinct categories, enabling the possibility of employing One-hot encoding.
- The disparities in distributions between target=0 and target=1 suggest that categorical attributes offer predictive insights into the target variable. Hence, it 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.

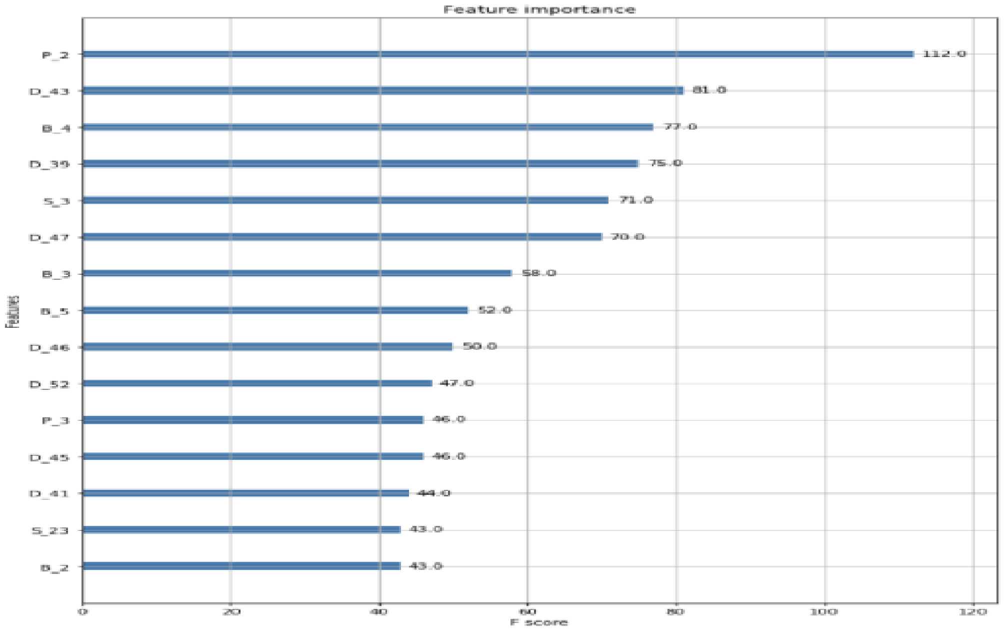
Most Highly-Correlated Spend Variables (Log Transformed Relationship))

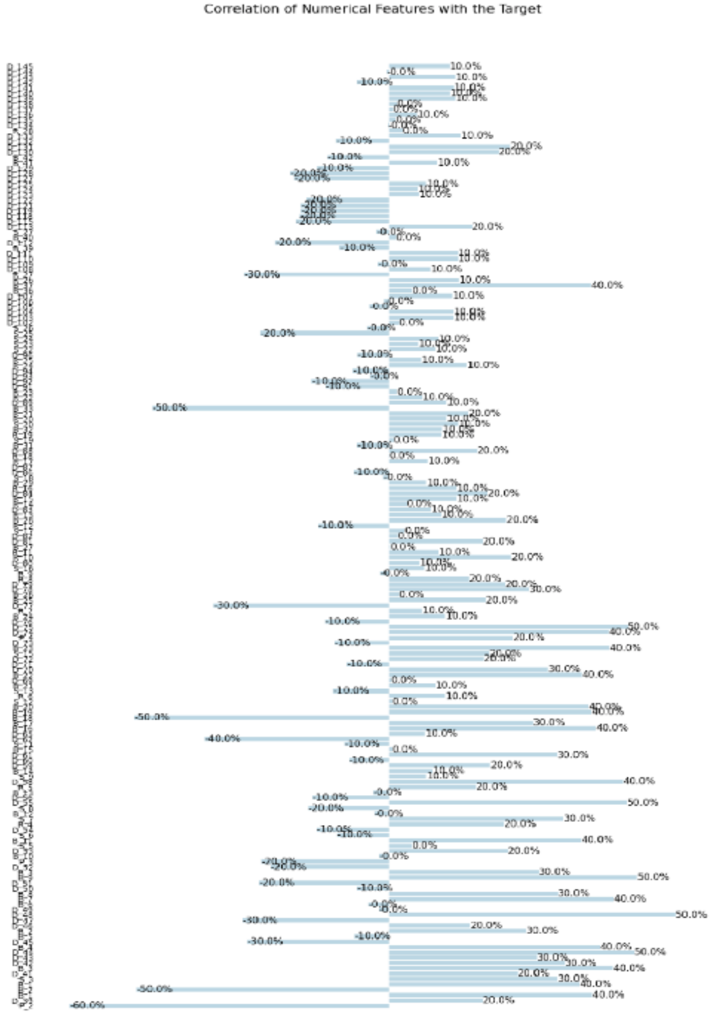



The hist plots illustrate noticeable disparities in distributions between target values of 0 and 1. It appears that the spend variables carry significant information regarding the target and warrant inclusion in modeling efforts.

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

# Data Preprocessing

Steps

- Handle missing values by imputation or removal
- Feature engineering through one-hot encoding, normalization, and scaling

Categories: Delinquency , Spend , Payment , Balance , Risk

# Methods

Exploratory Data Analysis

| Understand data distributions and relationships | Use visualizations like correlation matrices and scatter plots |
| --- | --- |

Model Selection: Considered Logistic Regression, Decision Tree Classifier, and XGBoost Classifier

Model Training and Evaluation: Used cross-validation and evaluated with metrics like accuracy, precision, recall, and AUPRC

# Analysis

| | 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 |

Best Model: XGBoost Classifier with approximately 85% accuracy

Insights: Identified key features influencing predictions

Visualization: Performance visualized through confusion matrices and classification reports

# Conclusion and Recommendations

Conclusion: The model provides a robust tool for assessing credit card default risk with high accuracy

Recommendations

- Continuously update the model with new data
- Implement regular audits to ensure fairness
- Integrate the model into a broader risk management framework

# Future Uses and Ethical Considerations

Future Uses

- Integrate into real-time credit scoring systems
- Extend to other financial products
- Explore deep learning models for improved performance

Ethical Considerations

- Ensure data privacy and confidentiality
- Address potential biases
- Adhere to regulatory guidelines like the Fair Credit Reporting Act

# Key Benefits of Predictive Model for American Express

**Enhanced Decision-Making**: Accurate default risk assessments optimize credit approvals and risk management.

**Improved Customer Experience**: Personalized interventions and timely credit assessments boost satisfaction.

**Cost Savings**: Reduces credit defaults and operational costs in risk assessment.

**Revenue Generation**: Maintains a healthier loan book and offers competitive credit products to drive customer acquisition.

**Scalability**: Scales across the customer base, addressing data integrity and computational challenges.

# 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