NYU

Application of SR 11-7 Guidelines to Machine Learning Models for Credit Risk

Eeshaan Asodekar

epa6822@nyu.edu

Outline

- Model Development
  - Statement of Purpose
  - Dataset assessment
  - Feature Engineering
  - Model search and final model selection
  - Model testing
- Model Validation
  - Data Quality
  - Conceptual Soundness
  - Quantitative Validation

PART 01

Model Development

Statement of Purpose

The purpose of this model is to leverage *explainable and interpretable machine learning techniques to predict credit card default*, adhering to the SR 11-7 guidelines

Dataset

Default of Credit Card Clients Dataset - UC Irvine Machine Learning Repository

Features:

- X1: Amount of the given credit (NT dollar)
- X2: Gender (1 = male; 2 = female)
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- X4: Marital status (1 = married; 2 = single; 3 = others)
- X5: Age (year)
- X6 - X11: History of past payment from April to September 2005
  The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above
- X12-X17: Amount of bill statement (NT dollar)
- X18-X23: Amount of previous payment (NT dollar)
- Target: default payment (Yes = 1, No = 0)

Dataset Assessment

Key steps performed:

- Checking for missing values and duplicates

  Although the dataset was free of NaNs, duplicates were detected and removed

- Identifying and handling outliers

  The IQR approach was used to identify outliers from the relevant features and they were removed

$$\text{Outliers}_i = \{x \in \text{Data}_i \mid x < Q1_i - 1.5 \cdot \text{IQR}_i \text{ or } x > Q3_i + 1.5 \cdot \text{IQR}_i\}$$

NYU

Data Augmentation

Based on analysis of features, these new features were augmented:

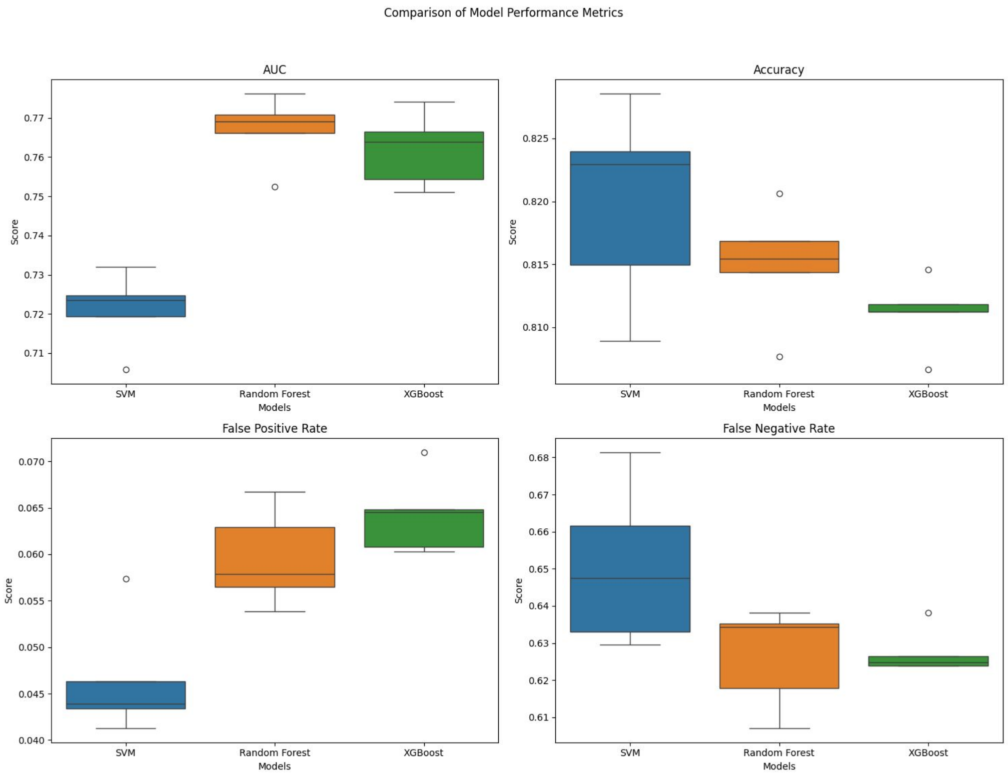
- Credit Limit Utilization
  - Represents the ratio of the bill amount to the credit limit
  - high ratio can indicate financial stress and potentially higher risk of default
- Average Delay in Payments
  - the average delay in payments over the last six months
  - would capture the general tendency of the customer to delay payments without focusing on a specific month
- Change in Bill Amount
  - Calculate the month-to-month percentage change in bill amount to capture trends in spending behavior

Model Shortlisting

- SVM (Support Vector Machine)
    - Maximizes the margin between data classes, enhancing model generalization and robustness
    - Kernel trick to efficiently handle non-linear data separations
    - Effective in high-dimensional spaces
- Random Forest
    - Utilizes multiple decision trees to ensure stability and accuracy, reducing the risk of overfitting
    - Automatically ranks the importance of variables providing clear insights
    - Naturally adept at handling unbalanced datasets
- XGBoost
    - Ensemble approach capable of handling varied and complex data structures
    - Incorporates regularization to prevent overfitting
    - Extremely popular in Kaggle competitions for its performance

Model Shortlisting

K-fold cross validation results



Model Testing

Results of tuned XGBoost model:

	Accuracy	AUC	FPR	FNR
Train	80.69%	88.39%	14.49%	24.13%
Test	78.06%	79.53%	15.53%	33.35%

NYU

PART 02

Model Validation

Data Quality & Processing

Evaluating key points:

- Dataset
  - UCI dataset is well documented and widely used in research literature
  - Contains key, well defined attributes
- Dataset Cleaning treatment
  - Null and duplicates handled
  - Outliers handled using IQR approach
- Alignment with portfolio is essential

NYU

Conceptual Soundness

- **Does model capture key characteristics for the required portfolio?**
  - Captures the key demographic data (age, sex, education, marital status)
  - Financial behaviors (payment history, bill amounts, payment amounts) also captured
  - Critique: Model should also take in credit score as input feature

- **How is imbalanced dataset taken care of ?**
  - Imbalanced dataset with Y=1 for 22% of dataset, may lead to illusively high performance
  - SMOTE (Synthetic Minority Over-sampling Technique) used
  - SMOTE on the train and not the test, unaltered test dataset
  - Critique: SMOTE can lead to overfitting to synthetic data, bad test performance
    Can explore ADASYN for more realistic data generation

Conceptual Soundness

- **Model Section Process fair?**
  - K fold cross validation used to assess models' performance in robust manner
  - Ensures that the outperformance on a subset is subdued by underperformance on other subsets
  - AUC, ACC, FPR, and FNR capture overall correctness as well as type 1 & 2 errors
  - Model selection on the basis of the false negative rate is in line with significance of that error

- **Assumptions of XGBoost Satisfied?**
  - All input features preprocessed to a numerical format
  - Although robust, outlier handling is performed for more effective learning
  - Standardization performed, although not required, impacts interpretability
  - Model assumes that individual observations are independent of each other
  - Critique: More testing needs to be conducted to gauge the independence of the observations
  - Critique: Standardization can lead to reduction relative magnitude information

Conceptual Soundness

- **Augmented features logical?**
  - Credit Limit Utilization: incorporation is justified as it encapsulates risk through a single metric
  - Average Delay in Payments: smoothed indicator of an individual's payment habits, mitigating the impact of any one-off or atypical late payment
  - Change in Bill Amount: spikes or drops could indicate new financial undertakings or changes in fiscal behavior
  - Critique:
    - Average Delay may erode the information of the extreme values
    - Time weighted approaches giving maximum importance to latest payment cycle can be incorporated

Conceptual Soundness

- **Does the feature importance make business sense?**
  - Average Delay
    - High importance of the Average Delay feature is consistent with business expectations
    - Delays in past payments are a strong indicator of potential future defaults, reflecting the borrower's financial habits and stability
  - Most Recent Payment/Default
    - Importance of recent payment behavior or default status as a critical factor is also logically sound
    - Recent default or delay can indicate current financial distress, making this feature crucial for predicting short-term credit risk

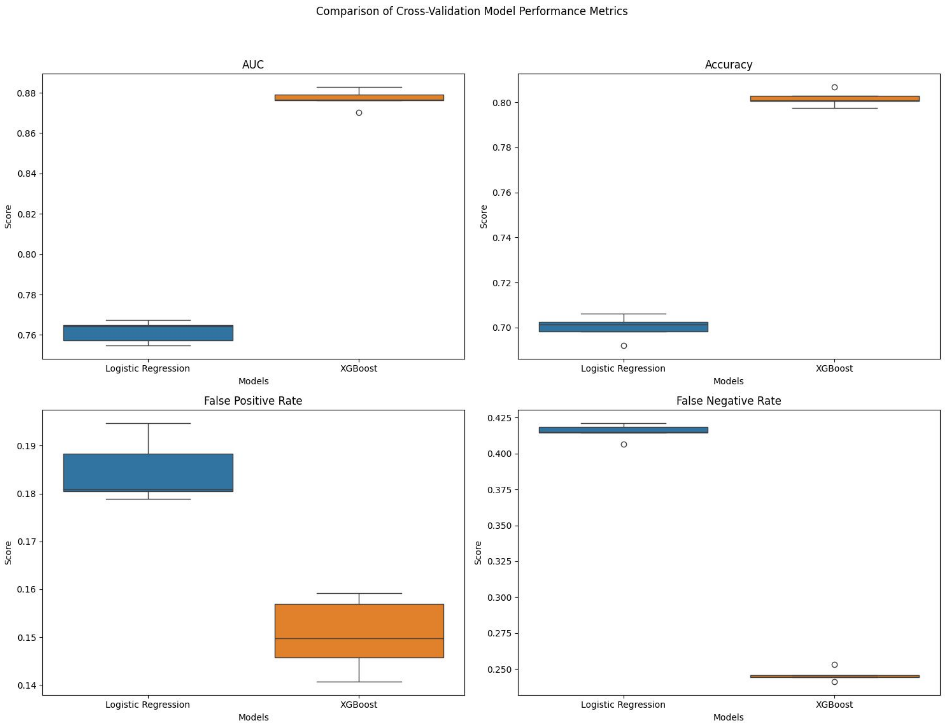
Quantitative Validation

Benchmarking against Logistic Regression

- Baseline Performance:
    - Logistic Regression is a well-established and well-understood model
    - If XGBoost performs significantly better than Logistic Regression, it provides evidence that the more complex model is capturing more complex patterns
- Feature Importance comparison
    - We can compare and contrast the feature importance
    - A radical shift would indicate insidious logical error in the XGBoost approach

NYU

Quantitative Validation

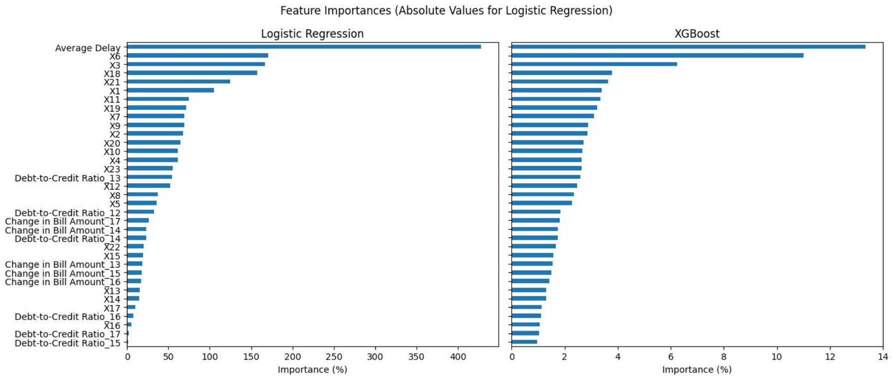


	Accuracy	AUC	FPR	FNR
Logistic Regg	70.02%	72.32%	18.41%	41.56%
XGBoost	76.72%	77.00%	16.73%	31.65%

Quantitative Validation

Feature Importance

- Average Delay is given highest importance by both the models underscoring the information gained by an summarized feature of delayed payments
- Most recent payment/default is second most important factor highlighting a possible a domino effect
- The performance enhancement of XGBoost can be explained by even importance distribution across recent payment history
- Feature importance assigned by the XGBoost is not a stark departure from that assigned by Logistic Regression



NYU

Quantitative Validation

Sensitivity Analysis

Average Delay

	Accuracy	AUC
+10 %	76.97%	77.15%
	79.36%	78.49%
-10%	75.504%	78.238%

X6

	Accuracy	AUC
+10 %	74.95%	77.249%
	79.36%	78.49%
-10%	75.771%	76.09%

On varying the two key feature inputs by +/- 10% we see that there is only a significant yet small difference in the accuracies and the AUC

NYU

Key Takeaways

- Machine Learning models add additional checkpoints in the model development and model validation process
  - Feature Engineering
  - Data Leakage
  - Hyperparameter tuning

- Model Validation for ML models is absolutely critical: high accuracy% (high FNR%) gives illusive confidence

NYU

Thank you