Convolve 3.0

Ayushmaan Pattnayak, Atharva Kulkarni, Purandhar V

Team CKMKB, IIT Madras

Please find the Solution CSV Here

Abstract

This document presents a comprehensive framework for developing a **Behaviour Score** predictive model to assess credit card default probabilities among customers of Bank A. The proposed solutions are categorized into two levels:

1. **Optimal Solution:** This approach employs a deep learning architecture, specifically Artificial Neural Networks (ANNs), to achieve superior classification performance in predicting defaults.
2. **Sub-optimal Solution:** This involves traditional machine learning classifiers applied to a PCA-transformed dataset. Among the evaluated classifiers, CatBoost demonstrated the best performance, with other models evaluated in increasing order of their effectiveness.

Preprocessing steps included handling missing values, identifying interdependent variables via Pearson correlation, imputing high-correlation features with autoencoders, and applying Principal Component Analysis (PCA) for dimensionality reduction. The final model evaluation focuses on critical performance metrics, validating the efficacy of the proposed Behaviour Score solutions.

Contents

1 Introduction

In the financial services sector, managing credit risk is crucial for ensuring profitability and maintaining a stable portfolio. To strengthen its risk management capabilities, Bank A has initiated the development of a **Behaviour Score** model to predict credit card default probabilities for active customers. This initiative aims to:

1. Enhance the bank's risk assessment framework.
2. Provide actionable insights for targeted customer interventions.
3. Improve overall decision-making for credit card management.

The project is divided into two primary approaches:

1. **Optimal Solution:**
   (a) Leverages Artificial Neural Networks (ANNs) for precise classification of default risks.
   (b) Focuses on maximizing performance through advanced deep learning architectures and optimized feature extraction techniques.
2. **Sub-optimal Solution:**
   (a) Implements traditional machine learning classifiers to evaluate the PCA-transformed dataset.
   (b) CatBoost, among the tested classifiers, showcased superior performance, followed by other models ranked in ascending order of effectiveness.

The raw dataset underwent rigorous preprocessing, including handling null values with targeted imputation strategies, analyzing feature interdependencies, and reducing dimensionality using PCA. The final Behaviour Score models are validated based on key evaluation metrics to ensure robustness and applicability in a real-world banking environment.

2 Dataset and Preprocessing

The dataset of 96,806 credit card records was rigorously preprocessed by handling missing values through targeted imputation (autoencoders for correlated features and 0 for others), addressing class imbalance with ADASYN, and applying PCA to reduce dimensionality to 176 features, ensuring a clean and balanced input for modeling.

2.1 Dataset Overview

The dataset provided by Bank A comprises detailed historical records of credit card usage by 96,806 customers, collectively referred to as the development dataset. Each record contains:

2.1.1 Features (Variables)

Attributes grouped into distinct categories:

- **On-us Attributes:** Details about credit limits, outstanding balances, and usage patterns.
- **Transaction-level Attributes:** Aggregated information on transaction counts and amounts across various merchant categories.
- **Bureau Tradeline Attributes:** Credit bureau data, including delin-

quencies and historical account information.

- **Bureau Enquiry Attributes:** Inquiries made by financial institutions to assess customer creditworthiness.

The dataset also includes a binary target variable, `bad_flag`, where:

- 1 represents a customer defaulting on their credit card payments.
- 0 indicates no default.

Additionally, a validation dataset of 41,792 records was provided, containing the same feature structure but without the target variable (`bad_flag`). This dataset is reserved for model evaluation.

2.2 Exploratory Data Analysis (EDA)

To prepare the raw dataset for predictive modeling, an in-depth Exploratory Data Analysis (EDA) was conducted to understand the data distribution, interdependencies, and potential anomalies:

2.2.1 Feature Distribution

- Variables exhibited diverse distributions, including skewed, normal, and multimodal patterns.
- Features with a high prevalence of missing values were identified for targeted imputation strategies.

2.2.2 Target Variable Distribution

- The `bad_flag` variable was highly imbalanced, with only 1.42% of records labeled as defaults (`bad_flag = 1`).

2.2.3 Correlation Analysis

Interdependencies among variables were examined using the Pearson Correlation Coefficient with a threshold of 0.5. This helped identify groups of highly correlated features for advanced imputation techniques.

2.3 Handling Missing Values

The dataset contained a significant number of null values with identifiable patterns. Two separate strategies were adopted based on feature interdependencies:

2.3.1 High-Correlation Features

- Features with significant interdependence (Pearson correlation ≥ 0.5) were processed using autoencoder-based imputation.
- This deep learning approach effectively preserved relationships between features by reconstructing missing values based on learned patterns.

2.3.2 Low-Correlation Features

Features with weak or no correlation to others were imputed with a default value of 0 to maintain simplicity without introducing bias.

2.4 Dimensionality Reduction

The dataset initially contained over 1,200 features, which posed challenges related to computation, noise, and overfitting. To address this, Principal Component Analysis (PCA) was applied:

2.4.1 PCA Implementation

- PCA transformed the feature space into principal components, capturing the maximum variance in the data.
- Several PCA-transformed datasets were generated, with varying levels of explained variance.

2.4.2 Final Selection

A dataset with 80% explained variance (reduced to 176 features) was chosen based on its superior performance in preliminary testing.

2.5 Challenges Addressed

2.5.1 Class Imbalance

The highly imbalanced target variable (`bad_flag`) required oversampling techniques for effective model training. ADASYN (Adaptive Synthetic Sampling) was employed to generate synthetic samples for the minority class (`bad_flag = 1`), ensuring a balanced dataset.

2.5.2 Feature Noise

By applying PCA, noise and redundancy in the raw features were significantly reduced, enhancing model generalizability.

2.5.3 Null Values

The targeted imputation strategy (autoencoders for high-correlation features and 0 for others) minimized the risk of information loss while preserving feature integrity.

3 Model Development

This section outlines the development and a short evaluation of two approaches for credit card default prediction: an Optimal Solution leveraging Artificial Neural Networks (ANNs) and a Sub-optimal Solution employing traditional machine learning classifiers

3.1 Optimal Solution

3.1.1 Model Development and Significance

1. **Model Architecture:**
   (a) The model is an Artificial Neural Network (ANN) built using the `Sequential` API from TensorFlow/Keras.
   (b) It comprises multiple `Dense` layers, each followed by a `Dropout` layer to prevent overfitting.
   (c) The `ReLU` activation function is used for the hidden layers to introduce non-linearity.
   (d) The output layer uses a `sigmoid` activation function, suitable for binary classification tasks, outputting probabilities for the 'bad_flag'.
2. **Model Compilation:**
   (a) The model is compiled using the `Adam` optimizer, which is efficient and well-suited for large datasets.
   (b) The loss function is `binary_crossentropy`, appropriate for binary classification.
   (c) The performance metric chosen is `accuracy`, along with additional evaluations like ROC AUC during testing.
3. **Model Significance:**
   (a) The ANN effectively captures complex patterns in the data, which traditional models may overlook.
   (b) Using `Dropout` layers addresses overfitting, making the model more robust to new, unseen data.
   (c) The high `ROC AUC` score of 0.9427 indicates the model's strong ability to differentiate between defaulters and non-defaulters.
   (d) Despite high overall accuracy, the

relatively lower recall for the minority class (defaulters) highlights areas for improvement, ensuring the model minimizes false negatives (missed defaulters).

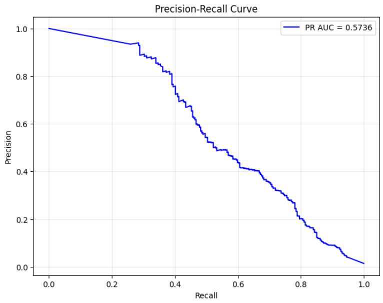
4. **Evaluation Metrics:**

   (a) The `Classification Report` indicates a high precision for non-defaulters and a moderate F1-score for defaulters.

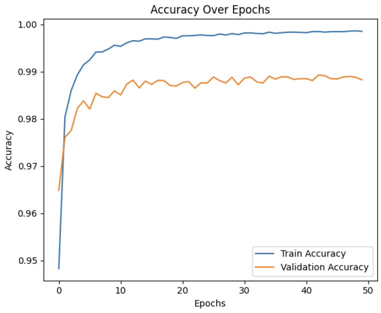
   (b) The `Confusion Matrix` shows the distribution of correct and incorrect predictions, providing insights into model performance on both classes.

   (c) The Precision-Recall Curve and its Area Under the Curve (AUC) offer a nuanced view of the trade-off between precision and recall, particularly important in imbalanced datasets.

   (d) **PR Curve**



   (e) **Accuracy over Epochs Curve**



3.2 Sub-optimal Solution

The sub-optimal solution evaluates traditional machine learning classifiers on the PCA-transformed dataset. Each classifier was trained on the resampled dataset (using ADASYN) and assessed on the test set, listed in order of performance

3.2.1 Decision Tree

The Decision Tree classifier was trained to predict defaults using a simple, interpretable tree-based model.

- **Confusion Matrix:**

$$\begin{bmatrix} 26335 & 2295 \\ 291 & 121 \end{bmatrix}$$

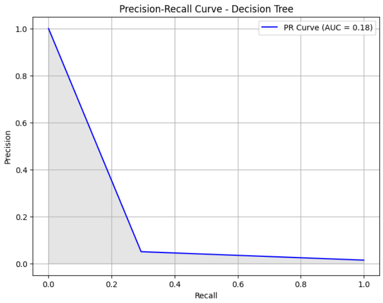
- **Classification Report:**

```
       precision  recall  f1-score
0.0      0.99      0.92      0.95
1.0      0.05      0.29      0.09
```

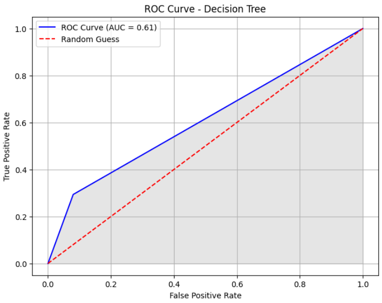
- **ROC-AUC Score:** 0.61

- **PR-AUC Score:** 0.09

- **PR Curve**



- **ROC Curve**



3.2.2 Random Forest

Random Forest, a powerful ensemble method, was evaluated next. It combines multiple decision trees to improve robustness and reduce overfitting.

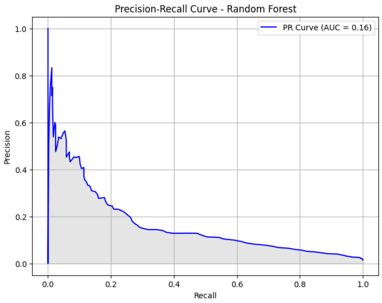
- **Confusion Matrix:**

$$\begin{bmatrix} 28370 & 260 \\ 328 & 84 \end{bmatrix}$$

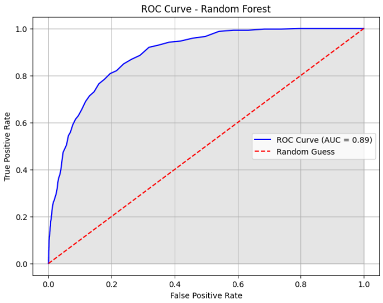
- **Classification Report:**

```
      precision   recall   f1-score
0.0     0.99       0.99      0.99
1.0     0.24       0.20      0.22
```

- **ROC-AUC Score:** 0.89
- **PR-AUC Score:** 0.22
- **PR Curve**



- **ROC Curve**



3.2.3 Naive Bayes

The Naive Bayes classifier, based on probability theory, was applied to predict defaults. Due to its assumptions, the results were suboptimal.

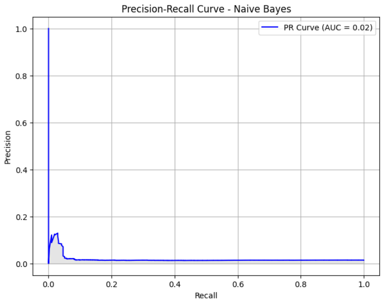
- **Confusion Matrix:**

$$\begin{bmatrix} 7754 & 20876 \\ 118 & 294 \end{bmatrix}$$

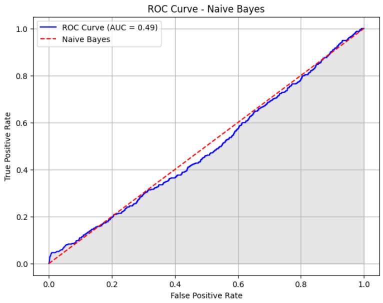
- **Classification Report:**

```
        precision     recall     f1-score
0.0       0.99         0.27        0.42
1.0       0.01         0.71        0.03
```

- **ROC-AUC Score:** 0.49
- **PR-AUC Score:** 0.03
- **PR Curve**



- **ROC Curve**



3.2.4 XGBoost

XGBoost, a gradient boosting framework, demonstrated strong results, particularly in terms of recall and overall AUC metrics.

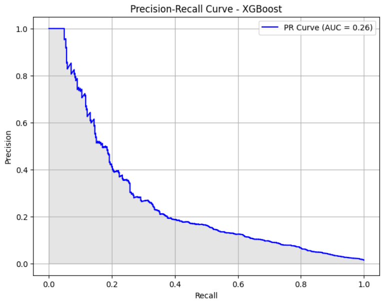
- **Confusion Matrix:**

$$\begin{bmatrix} 28192 & 438 \\ 274 & 138 \end{bmatrix}$$

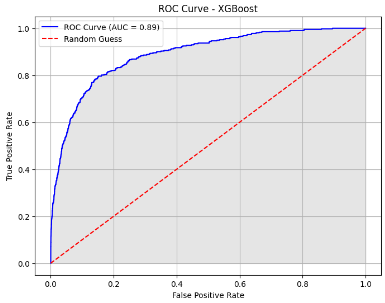
- **Classification Report:**

```
      precision  recall  f1-score
0.0     0.99      0.98     0.99
1.0     0.24      0.33     0.28
```

- **ROC-AUC Score:** 0.89
- **PR-AUC Score:** 0.28
- **PR Curve**



- **ROC Curve**



3.2.5 CatBoost

CatBoost emerged as the best-performing classifier among the traditional models. Its handling of categorical variables and efficient boosting mechanism contributed to its success.

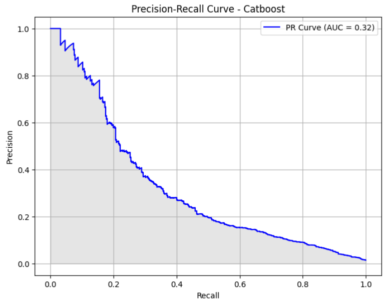
- **Confusion Matrix:**

$$\begin{bmatrix} 28284 & 346 \\ 264 & 148 \end{bmatrix}$$

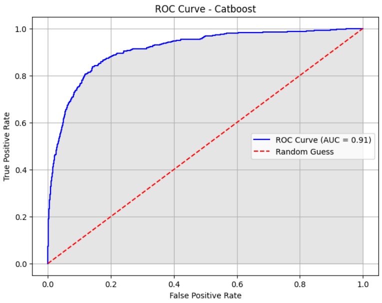
- **Classification Report:**

```
      precision  recall  f1-score
0.0     0.99      0.99     0.99
1.0     0.30      0.36     0.33
```

- **ROC-AUC Score:** 0.92
- **PR-AUC Score:** 0.33
- **PR Curve**



- **ROC Curve**



3.2.6 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) demonstrated competitive performance but was slightly outperformed by CatBoost.

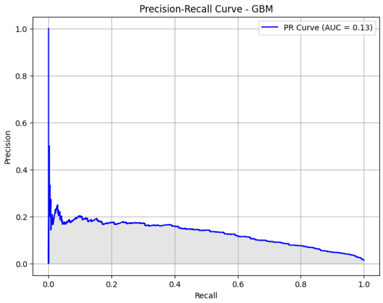
- **Confusion Matrix:**

$$\begin{bmatrix} 25316 & 3314 \\ 104 & 308 \end{bmatrix}$$

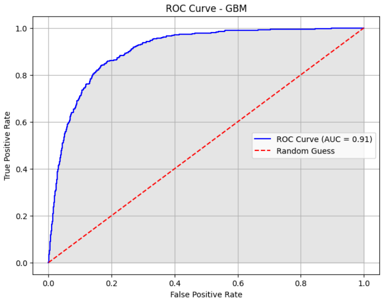
- **Classification Report:**

```
      precision    recall  f1-score
0.0     1.00        0.88      0.94
1.0     0.09        0.75      0.15
```

- **ROC-AUC Score:** 0.91
- **PR-AUC Score:** 0.15
- **PR Curve**



- **ROC Curve**



4 Conclusions and Future Scope

4.1 Conclusions

The Behaviour Score model effectively combines advanced data preprocessing techniques with machine learning approaches to predict credit card defaults. The conclusions are summarized as follows:

4.2 Data Preprocessing

1. **Handling Missing Values:** High-correlation features were imputed using autoencoder-based methods to preserve interdependencies, while low-correlation features were filled with default values for simplicity.
2. **Dimensionality Reduction:** Principal Component Analysis (PCA) reduced the feature space from over 1,200 to 176 components, retaining 80% of the variance while reducing noise and redundancy.
3. **Addressing Class Imbalance:** The highly imbalanced target variable was balanced using ADASYN, improving the model's ability to identify minority class instances (defaults).

4.3 Model Conclusions

1. **Tree-Based Models for Business Relevance:** Models like CatBoost offer high interpretability, allowing clear reasoning for classification decisions, which is crucial for practical implementation in financial risk management.
2. **CatBoost's Superior Performance:** CatBoost outperformed other traditional classifiers in terms of PR-AUC and ROC-AUC scores, demonstrating exceptional robustness and efficiency.
3. **Neural Networks' Predictive Power:** The neural network model achieved the best overall performance but lacked interpretability, which is critical for business applications.

4.4 Future Scope

The current project demonstrates a robust framework for predicting credit card defaults using advanced machine learning techniques. However, several areas of improvement and extensions can be explored to further enhance the model's performance and applicability:

1. **Incorporating Ensemble Methods and LSTMs:**
   - Future work could integrate advanced ensemble techniques combined with Long Short-Term Memory (LSTM) networks.
   - LSTMs, capable of capturing temporal dependencies in sequential data, can complement the ensemble methods' predictive power.
   - By taking a weighted average of the outputs from these models, a more accurate and generalized Behaviour Score can be achieved.
2. **Hyperparameter Tuning of CatBoost:**
   - CatBoost demonstrated exceptional performance in this project, particularly in terms of PR-AUC and ROC-AUC scores.
   - Future work can focus on hyperparameter tuning of CatBoost to further optimize its performance.
   - Although computationally expensive, such tuning can potentially yield higher PR-AUC values, enhancing the model's ability to identify minority class instances (defaults).

These future directions provide a roadmap for further improving the predictive capabilities and scalability of the Behaviour Score model, enabling its adoption in broader financial risk management applications.

5 References

- Comprehensive Overview of Big Data: Applications, Networking Technologies, and Challenges.