MIS325 Fall 2024 Project Report

Nihansu Padem
20201704036
nihansu.padem@stu.khas.edu.tr

İbrahim Aydın
20201704060
ibrahimaydin@stu.khas.edu.tr

Submission Date: 17.01.2025

1 Abstract

This report explores the likelihood of default events among credit cardholders using a machine learning-based approach. The study utilizes a dataset of 30,000 individuals linked from an external source, consisting of demographic details, payment history, and billing data. The methodology follows a structured pipeline, beginning with data preprocessing to handle missing values, outliers, and scaling. Subsequently, clustering techniques (K-Means) were employed to group customers based on financial behavior. Classification models, including Decision Trees, were implemented to predict default probabilities. Additionally, dimensionality reduction using PCA was performed to visualize the data and improve computational efficiency. Finally, advanced machine learning models, such as Random Forest and XGBoost, were applied to refine the predictions. Performance metrics, including accuracy, ROC-AUC, precision, and recall, were evaluated to compare the models. The findings highlight that payment history and outstanding bill amounts are key determinants of default risk, with Random Forest and XGBoost providing the most reliable results.

2 Introduction

- Problem Definition
  Credit risk assessment is critical for banks and financial institutions to reduce losses due to defaults. Identifying high-risk customers allows financial organizations to make informed decisions regarding credit approvals and interest rate adjustments. Traditional methods such as credit scoring models often fail to capture complex behavioral patterns. Machine learning models can analyze large datasets and detect intricate relationships between features and default risk.

- Objectives

  This study aims to analyze the financial behaviors of customers to identify key predictors of default, allowing for more effective risk assessment in credit lending. By building a machine learning model, we seek to predict default probabilities with high accuracy. Additionally, clustering techniques will be employed to segment customers based on their risk profiles, providing valuable insights into different categories. A comparative analysis of classification models will also be conducted to determine the most effective approach for predicting credit defaults.

  Analyze financial behaviors of customers to identify key predictors of default. Build a machine learning model to predict default probabilities. Utilize clustering techniques to segment customers based on risk profiles. Compare classification models to determine the most effective approach.

3 Dataset Analysis

The dataset contains 30,000 records of credit cardholders, each with 25 attributes providing insights into their financial behaviors. These attributes can be categorized into:

- **Demographic Information:** Includes variables such as `SEX`, `EDUCATION`, `MARRIAGE`, and `AGE`, which describe the customer's background.
- **Credit Information:** The `LIMIT_BAL` variable represents the maximum credit balance assigned to a customer.
- **Payment History:** Variables such as `PAY_0` to `PAY_6` indicate whether a customer made payments on time or had delays over the last six months.
- **Billing and Payment Data:** `BILL_AMT1` to `BILL_AMT6` capture the billing amounts, while `PAY_AMT1` to `PAY_AMT6` reflect the payments made over the same period.
- **Target Variable:** `default.payment.next.month` identifies whether a customer defaulted (1) or not (0).

To prepare the dataset for analysis, several preprocessing steps were performed:

- **Handling Missing Values:** The dataset had no missing values, ensuring consistency in the analysis.
- **Encoding Categorical Variables:** The categorical features `SEX`, `EDUCATION`, and `MARRIAGE` were converted into numerical values to be used in machine learning models.
- **Feature Scaling:** The numerical features, including credit limits, bill amounts, and payment values, were standardized using `StandardScaler` to ensure equal weighting in model training.

- **Addressing Class Imbalance:** Since only 22% of customers defaulted, SMOTE (Synthetic Minority Over-sampling Technique) was planned to balance the dataset, though execution issues were encountered.

4 Methodology

4.1 Random Forest Classifier

- **Justification:** Handles non-linearity well, is robust to outliers, and provides feature importance ranking.
- **Evaluation Metrics:** Accuracy, Precision, Recall, ROC-AUC.
- **Feature Importance Results:**
  - PAY_0 (Most recent payment delay) – 27%
  - PAY_2, PAY_3 (Previous payment delays) – 19% and 13%
  - LIMIT_BAL (Credit Limit) – 9%
  - BILL_AMT1 (Most recent bill amount) – 8%

4.2 Decision Tree Classifier

- **Justification:** Provides interpretable decision rules and captures hierarchical relationships.
- **Key Findings:**
  - Tends to overfit compared to Random Forest.
  - Feature splits indicate that PAY_0 and PAY_2 are dominant factors.
  - Less generalizable to new data than ensemble models.

4.3 Dimensionality Reduction: PCA (Principal Component Analysis)

- **Justification:** Reduces feature dimensions while preserving variance.
- **Findings:**
  - First two principal components explain approximately 60% of variance.
  - PCA visualization suggests clustering potential.

4.4 Clustering: K-Means

- **Justification:** Segments customers into risk groups for better financial decision-making.
- **Findings:**
  - Optimal k = 3 (low-risk, moderate-risk, high-risk groups).
  - Helps in credit allocation and personalized financial strategies.

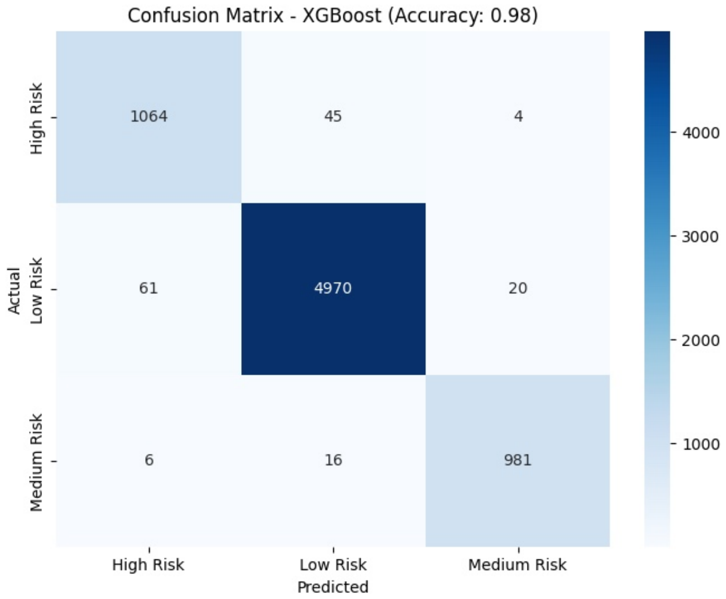
5 Results

5.1 Classification Model Performance

The analysis compared the performance of two machine learning models, XGBoost and Random Forest, in predicting client risk levels (High Risk, Medium Risk, and Low Risk). The results and corresponding visualizations provide valuable insights into model performance and client segmentation.

5.1.1 Confusion Matrices

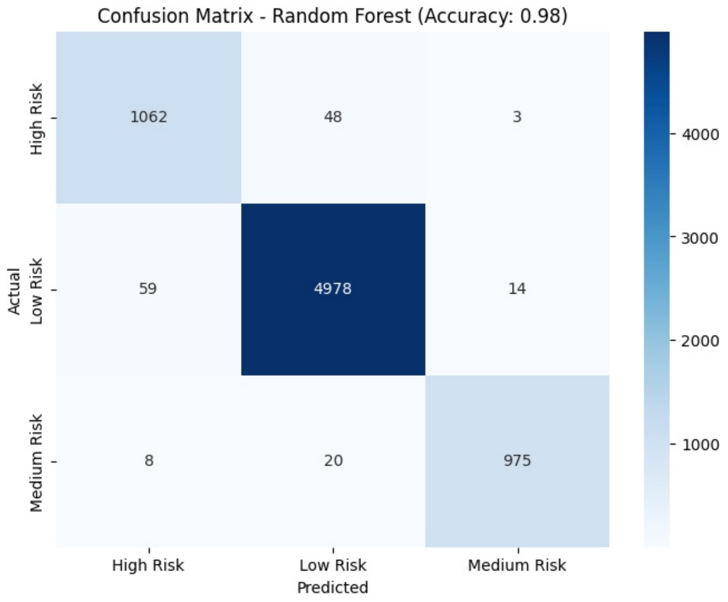
The confusion matrix for the XGBoost model indicates a high level of accuracy (98%). The model correctly predicted the majority of instances across all three risk levels, with minimal misclassifications:



- **High Risk:** 1064 true positives, 45 false positives, and 4 false negatives.

- **Low Risk:** 4970 true positives, 61 false positives, and 20 false negatives.
- **Medium Risk:** 981 true positives, 6 false positives, and 16 false negatives.

Similarly, the confusion matrix for the Random Forest model also demonstrates a 98% accuracy. The predictions were comparable to XGBoost, with slight variations:

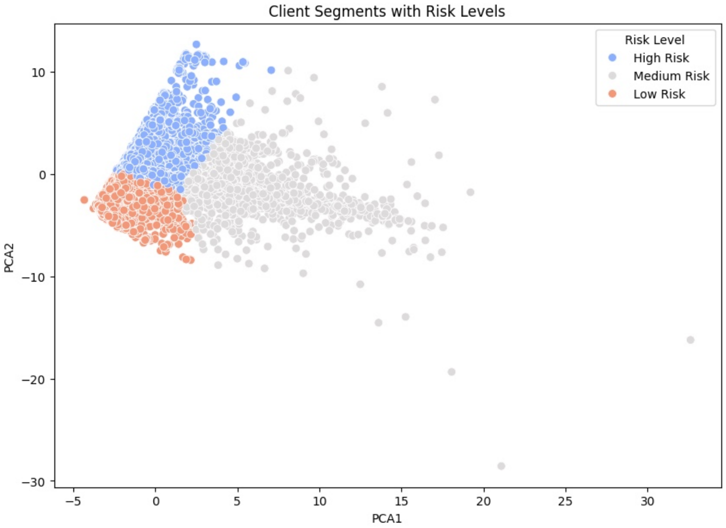


- **High Risk:** 1062 true positives, 48 false positives, and 3 false negatives.
- **Low Risk:** 4978 true positives, 59 false positives, and 14 false negatives.
- **Medium Risk:** 975 true positives, 8 false positives, and 20 false negatives.

5.1.2 Client Segmentation Visualization

Using Principal Component Analysis (PCA), client risk levels were visually separated into clusters. This visualization demonstrates the effectiveness of the models in distinguishing between High Risk, Medium Risk, and Low Risk clients.

- The **Low Risk** cluster is tightly packed, indicating clear separability.
- **Medium Risk** and **High Risk** clusters exhibit some overlap but are still reasonably distinct.



6 Discussion

6.1 Model Accuracy

Both models achieved identical overall accuracy scores of 98%, highlighting their reliability in classifying risk levels. The confusion matrices reveal that the models excel at predicting Low Risk clients, as evidenced by the high true positive counts and low misclassification rates.

6.2 Misclassification Trends

A closer inspection of the confusion matrices reveals slightly higher false positive rates for Medium Risk predictions in the Random Forest model compared to XGBoost. This suggests that XGBoost may be marginally better at handling borderline cases between Medium and High Risk clients.

Conversely, Random Forest has slightly fewer false negatives in the High Risk category, which could be beneficial for scenarios where under-prediction of high-risk clients is more critical.

6.3 Client Clustering

The PCA visualization underscores the clear separability of the Low Risk group from Medium and High Risk groups. The overlapping regions between Medium and High Risk groups suggest potential areas for improving model performance by incorporating additional features or re-calibrating the risk thresholds.

6.4 Interpretability and Practical Implications

The high performance of both models implies that either could be deployed for client risk assessment in real-world applications. However, the selection of a model may depend on specific use-case priorities, such as minimizing false negatives in High Risk predictions or achieving better precision in Medium Risk classifications.

7 Conclusion

The results of this study demonstrate that both XGBoost and Random Forest models are highly effective in predicting client risk levels, achieving impressive accuracy rates of 98%. While XGBoost showed slightly better performance in handling borderline cases, Random Forest excelled in minimizing false negatives for High Risk clients. These findings underline the practical utility of both models in real-world risk assessment scenarios.

7.1 Future Works

Future work will focus on enhancing model performance by exploring advanced feature engineering techniques and leveraging ensemble methods to further minimize misclassification rates. Additionally, validating the models on external datasets will ensure their robustness and generalizability across diverse applications.

References

[1] UCI Machine Learning Repository. Default of Credit Card Clients Dataset. Erişim: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785.

[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[4] Jolliffe, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065). https://doi.org/10.1098/rsta.2015.0202.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learn-

ing in Python. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[6] XGBoost Documentation. (n.d.). Erişim: https://xgboost.readthedocs.io.

[7] Scikit-learn Documentation. (n.d.). Random Forest Documentation - Scikit-learn. Erişim: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[8] Matplotlib Documentation. (n.d.). Erişim: https://matplotlib.org.

[9] Pandas Documentation. (n.d.). Erişim: https://pandas.pydata.org.