ScoreCraft

Abdurrahman Doğru
Artificial Intelligence Engineering
221401007
TOBB ETU, Ankara, Türkiye
abdurrahmandogru@etu.edu.tr

***Abstract*—This project focuses on the challenge of predicting credit default risk using structured application data. Rather than treating modeling as the sole priority, we emphasize the importance of systematic data analysis, which is the foundation of any successful data mining task. The analysis is conducted using the Home Credit Default Risk dataset [13], which contains demographic, financial, and application-related information. Comprehensive data preprocessing steps are applied, including missing value treatment, outlier detection, and feature engineering. Class imbalance is addressed using SMOTE to ensure balanced representation of both classes. Three different machine learning models (Logistic Regression, XGBoost, and Random Forest) are trained and evaluated using 8-fold cross-validation. The results demonstrate that ensemble learning methods (XGBoost and Random Forest) significantly outperform the baseline Logistic Regression model, achieving approximately 84% accuracy . Feature importance analysis reveals that days birth is the most determinant factors in predicting credit risk. This work provides a reliable and explainable risk prediction framework for credit assessment, emphasizing the synergy between thorough data preprocessing and appropriate model selection.**

***Index Terms*—Credit Risk Prediction, Data Preprocessing, Exploratory Data Analysis, Machine Learning, Feature Importance**

Here is the github link to project codespace: https://github.com/abdurrahman34/ScoreCraft.git

I. Introduction

A. Motivation

Credit defaults have profound economic implications that extend beyond individual financial institutions. According to the World Bank, non-performing loans can significantly impair economic growth by restricting credit availability and increasing lending costs. During the 2008 financial crisis, default rates reached unprecedented levels, contributing to approximately $2.1 trillion in losses for financial institutions worldwide. Moreover, defaults disproportionately affect underserved communities by further restricting their access to affordable credit. By improving default prediction accuracy by even a few percentage points, our model could potentially save financial institutions millions in losses while simultaneously expanding credit access to previously excluded populations through more precise risk assessment.

Additionally, for financial institutions, accurately assessing credit risk is critical for both corporate sustainability and customer satisfaction as Hand ve Henley acknowledges [16]. Traditional credit scoring methods typically rely on limited data sources and struggle to evaluate customers with no credit history. Machine learning approaches offer the potential to conduct more comprehensive and accurate risk assessments by analyzing numerous variables and modeling complex relationships. This project aims to identify customers who may experience difficulty repaying loans in advance, thereby both reducing risks for financial institutions and facilitating access to financial services for eligible customers.

B. Literature Review

Credit risk assessment using machine learning has been extensively studied in the literature. Baesens et al. [1] conducted one of the first comprehensive benchmarking studies of classification algorithms for credit scoring, comparing neural networks, decision trees, and statistical methods. Their findings indicated that simple classifiers like logistic regression often perform competitively with more complex methods.

Khandani et al. [2] demonstrated the effectiveness of machine learning algorithms for consumer credit risk prediction, achieving significant improvements over traditional credit scoring models. They emphasized the importance of feature engineering and the inclusion of macroeconomic variables.

For tree-based methods specifically, Malekipirbazari and Aksakalli [3] showed that Random Forest outperforms traditional credit scoring methods in peer-to-peer lending contexts. Similarly, Xia et al. [4] proposed a boosted decision tree approach using Bayesian optimization for hyperparameter tuning, achieving superior performance compared to standard methods.

Louzada et al. [5] provided a systematic review of classification methods applied to credit scoring, analyzing 187 articles published between 1992 and 2015. Their meta-analysis revealed that ensemble methods and hybrid approaches generally outperform individual classifiers.

In mortgage default prediction, Fitzpatrick and Mues [6] compared various classification algorithms during a distressed market period, finding that gradient boosting machines consistently outperformed other methods, including logistic regression and neural networks.

Wang et al. [7] proposed dual strategy ensemble trees for credit scoring, combining vertical and horizontal ensemble strategies to improve classification accuracy. Their approach demonstrated superior performance compared to single classifiers.

Yeh and Lien [8] compared various data mining techniques for predicting credit card defaults, finding that neural networks

and SVMs achieved higher accuracy than traditional statistical methods.

Our work builds upon these findings by implementing and comparing three different algorithms (Logistic Regression, XGBoost, and Random Forest) on the Home Credit Default Risk dataset, with a particular focus on addressing class imbalance through SMOTE and enhancing model interpretability through feature importance analysis.

C. General Methodology

In our project, we followed these methodological steps:

1) Data cleaning and preprocessing:
   - Filling missing values
   - Detection and processing of outliers
   - Handling anomalous values (such as placeholder values in employment duration)
2) Feature engineering:
   - Deriving new features from date variables
   - Transformation of categorical variables
   - Creating normalized indicators (e.g., income per family member)
   - Development of temporal features related to application timing
3) Class imbalance analysis and SMOTE application:
   - Detailed examination of class distribution (91.9
   - Implementation of SMOTE to create a balanced dataset
   - Careful handling of ID columns to prevent data leakage during resampling
4) Model training:
   - Logistic Regression (baseline model)
   - XGBoost and Random Forest algorithms
   - Hyperparameter optimization with cross-validation
   - Implementation of progress tracking during model training
5) Model evaluation using 8-fold cross-validation:
   - Comprehensive metrics including accuracy, precision, recall, F1-score, and ROC AUC
   - Visualization of confusion matrices with detailed annotations
   - Analysis of misclassification patterns
6) Feature importance analysis and model comparison:
   - Normalization of feature importance values for consistent comparison
   - Identification of top predictive features across different models
   - Visualization of feature importance distributions
7) Analysis of Results:
   - Standardization of output formats
   - Development of model summary tools for reporting
   - Creation of visualization functions for result interpretation
   - Analysis of results

D. Objectives

1) Develop a model that predicts customers who may experience difficulty in loan repayment with high accuracy
2) Compare the performance of different machine learning algorithms in credit risk assessment
3) Identify the most important factors determining credit risk
4) Analyze the impact of data preprocessing and balancing techniques on model performance
5) Present a practically applicable risk assessment approach for financial institutions
6) emonstrate the importance of systematic data analysis in financial modeling
7) Create a framework that balances predictive power with interpretability
8) Provide insights that can help expand credit access to underserved populations while maintaining appropriate risk controls

II. Data and Preprocessing

The Home Credit Default Risk dataset [13] was shared on the Kaggle platform by the financial service provider Home Credit Group. This dataset is based on real data from a financial institution that aims to provide financial services to customers with limited or no credit history.

A. Dataset Characteristics

- Number of Records: 307,511 customer records
- Number of Features: 122 features
- Target Variable: TARGET (0: Credit will be repaid, 1: Difficulty in credit repayment)
- Data Type: Mix of numerical and categorical variables
- Time Range: Customer applications and past credit records

Important Features:

1) Demographic information (age, gender, family status)
2) Financial indicators (income, credit amount)
3) Employment information (employment duration, occupation type)
4) Property information (car, home ownership)
5) Application information (application time, channel)

B. Preprocessing Pipeline

Since the dataset is raw and includes real-world imperfections, extensive preprocessing was applied:

1) **Missing Values:**
   - For numerical features like OWN-CAR-AGE, missing values were filled with -1 to indicate ”unknown” while preserving the record.
   - For categorical variables like OCCUPATION-TYPE, missing values were imputed with a separate category such as ’None’.
   - Rows with unresolvable or highly sparse features were removed to avoid distortion in downstream tasks.

- Single isolated missing values were imputed using forward fill (ffill) method to maintain data continuity.
- Multiple consecutive missing values were addressed using linear interpolation.

2) **Outlier Detection:**
   - IQR-based outlier detection was applied to key numerical features such as income and credit amount.
   - Additionally, extreme outliers (beyond 3*IQR) were removed to reduce skew and variance.

Column	Outliers
ID	0
CNT_CHILDREN	4250
AMT_INCOME_TOTAL	13951
AMT_CREDIT	6487
AMT_ANNUITY	7463
AMT_GOODS_PRICE	14618
REGION_POPULATION_RELATIVE	8371
DAYS_BIRTH	0
DAYS_EMPLOYED	16780
HOUR_APPR_PROCESS_START	2249
FLAG_MOBIL	1
CNT_FAM_MEMBERS	3987
FLAG_EMAIL	17391
CREDIT_INCOME_RATIO	11446
CREDIT_YEARS	17
AGE_YEARS	0
EMPLOYMENT_YEARS	16722
INCOME_PER_PERSON_IN_FAMILY	17329

Fig. 1. Outlier Analysis

3) **Feature Engineering:**
   - New variables were derived from existing features to enhance model input.
   - More interpretable features like AGE-YEARS and EMPLOYMENT-YEARS were derived from day-based variables such as DAYS-BIRTH and DAYS-EMPLOYED.
   - Financial ratios like CREDIT-INCOME-RATIO were calculated from AMT-CREDIT and AMT-INCOME-TOTAL variables, better representing the customer's debt burden.
   - The INCOME-PER-PERSON-IN-FAMILY feature was created by dividing total income by family size, better reflecting household economic status
   - These constructed features aim to provide more interpretable and normalized indicators of applicant behavior.

4) **Feature Selection:**
   - Correlation analysis was performed to identify and remove highly correlated features.
   - Features with minimal variance were eliminated.
   - Final feature set was reduced from 122 to 33 meaningful predictors.

5) **Anomalies:**
   - Specific placeholder values like DAYS-EMPLOYED = 365243 (likely indicating retirement) were recoded to -50 and later used to derive EMPLOYMENT-YEARS.

6) **Data Normalization:**
   - Numerical features showed varying scales and distributions, necessitating normalization for certain models.
   - StandardScaler was applied to normalize features for the Logistic Regression model, transforming features to have zero mean and unit variance.

7) **Encoding:**
   - Categorical variables were transformed using one-hot encoding.
   - ID columns (such as 'Unnamed: 0', 'SK-ID-CURR', 'ID') were dropped before resampling to avoid data leakage.

C. Class Imbalance and SMOTE

A significant class imbalance was detected in the dataset

```
Before SMOTE:
- Class 0 (No Payment Difficulty): 91.9% (282,686 samples)
- Class 1 (Payment Difficulty): 8.1% (24,825 samples)

After SMOTE:
- Class 0 (No Payment Difficulty): 50% (282,686 samples)
- Class 1 (Payment Difficulty): 50% (282,686 samples)
```

Fig. 2. Smote Effect

This imbalance would make it difficult for the model to learn the minority class (customers experiencing payment difficulties), so SMOTE (Synthetic Minority Over-sampling Technique) [11] was applied to balance the class distribution. SMOTE balances the dataset by creating synthetic samples from minority class examples.

D. Feature Analysis and Relationships

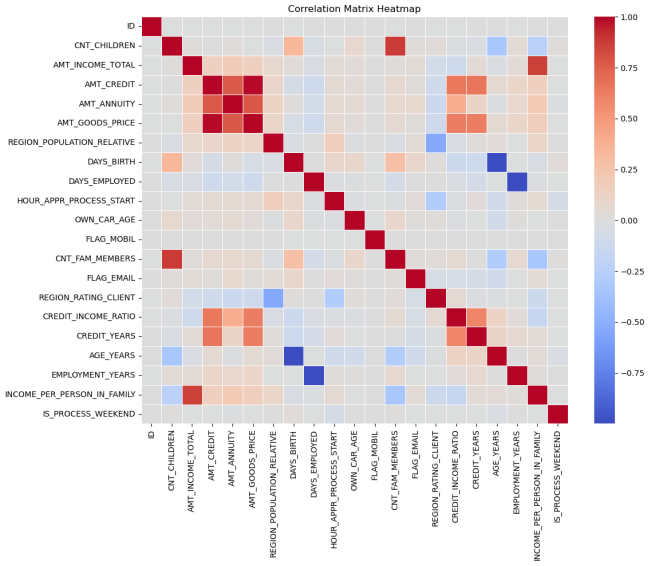

Fig. 3. Correlation HeatMap

Correlation analysis was performed to understand relationships between features.

- Random Forest for Feature Selection: Random Forest algorithm was utilized to determine feature importance. This algorithm measures the contribution of features
- Features with the highest correlation to the target variable were determined.
- Feature Selection Based on Importance: Using Random Forest feature importance analysis results, the 33 most informative features were selected, reducing model complexity while maintaining predictive performance.

Features with highest correlation to the target variable

```
Feature ranking:
1. DAYS_BIRTH (0.0718)
2. CREDIT_INCOME_RATIO (0.0650)
3. AMT_ANNUITY (0.0647)
4. AGE_YEARS (0.0630)
5. DAYS_EMPLOYED (0.0608)
6. AMT_CREDIT (0.0558)
7. CREDIT_YEARS (0.0556)
8. REGION_POPULATION_RELATIVE (0.0548)
9. INCOME_PER_PERSON_IN_FAMILY (0.0505)
10. EMPLOYMENT_YEARS (0.0492)
```

Fig. 4. Feature Correlation to Target

E. Data Distributions and Normalization

The dataset includes features at multiple measurement levels:

Nominal: OCCUPATION-TYPE, CODE-GENDER, MARITAL-STATUS, FLAG-OWN-CAR, FLAG-OWN-REALTY, ORGANIZATION-TYPE, NAME-TYPE-SUIT.

Ordinal: EDUCATION-LEVEL, REGION-RATING-CLIENT.

Interval: HOUR-APPR-PROCESS-START, DAYS-REGISTRATION, AGE-GROUP.

Ratio: AMT-INCOME-TOTAL, AMT-CREDIT, AMT-ANNUITY, EMPLOYMENT-YEARS, AMT-GOODS-PRICE, CNT-CHILDREN, CNT-FAM-MEMBERS.

The distribution of features was inspected using histograms, boxplots, and statistical summaries. Most numerical variables showed right-skewed distributions. Normalization using StandardScaler was applied for the Logistic Regression model. Since XGBoost and Random Forest algorithms are not sensitive to scaling, normalization was not applied for these models.

F. Model-Specific Preprocessing

Logistic Regression:

- Features were standardized using StandardScaler
- Feature importance values were normalized to 0-1 range for better interpretability
- Regularization (L2) was applied to prevent overfitting

XGBoost:

- Raw features were used without scaling (algorithm is scale-invariant)
- Missing values were handled internally by the algorithm
- Feature importance was automatically calculated and normalized

Random Forest:

- Raw features were used without scaling (algorithm is scale-invariant)
- Bootstrap sampling was applied within the algorithm
- Feature importance was derived from mean decrease in impurity

G. Data Flow Diagram

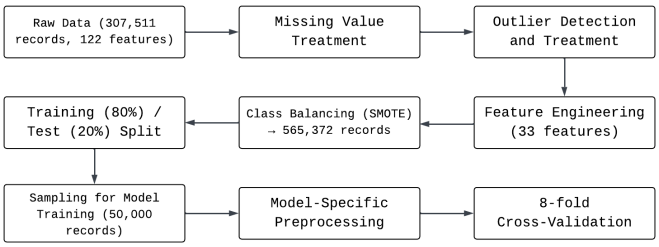

Fig. 5. Data Processing Flow

This comprehensive data processing flow shows all steps from raw data to model training. Data quality was improved at each step, and necessary transformations were applied to optimize model performance. The modular approach allowed for efficient processing and model-specific optimizations.

H. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) [14] was applied to visualize the high-dimensional dataset in 2D space, providing insights into data structure and class separability.

1) Methodology
   - Numerical features were selected and standardized
   - Outliers were filtered using z-score thresholding (z smaller than 3)
   - 43,198 outliers (14.4% of data) were removed for visualization
   - PCA reduced dimensions to 2 principal components
2) Results
   - The first two principal components explain 36.0% of total variance
   - Principal Component 1: 18.3% variance
   - Principal Component 2: 17.7% variance
   - Class distribution in filtered dataset:
     - Default(no-pay): 21,813 (8.3%)
     - No Default(pay): 242,500 (91.7%)
3) Interpretation
   - Significant overlap between default and non-default classes indicates challenging classification
   - The circular distribution pattern suggests complex, non-linear relationships between features

- Default cases (red) appear slightly more concentrated in certain regions
- The relatively low explained variance (36.0%) indicates that two dimensions cannot fully capture the complexity of this dataset
- The class imbalance (8.3% default rate) confirms the need for balancing techniques during model training

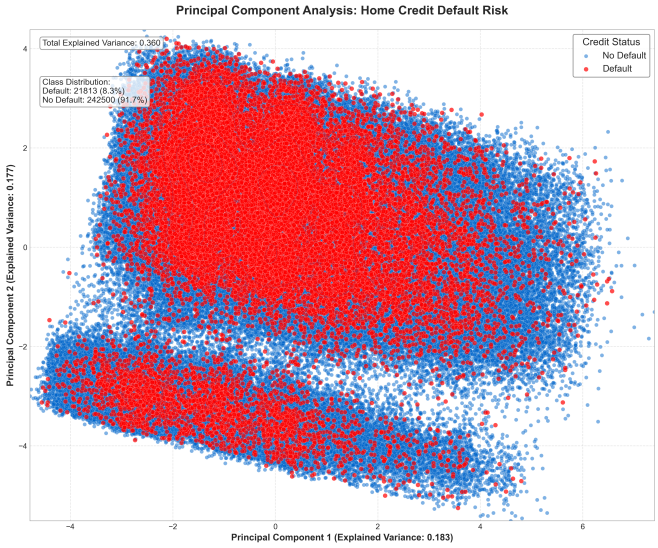

Fig. 6. PCA visualization

This visualization reinforces the need for sophisticated modeling approaches and careful handling of class imbalance to effectively predict credit default risk.

III. Methodology

This study employs a comprehensive approach to credit default risk prediction, combining robust data preprocessing techniques with advanced machine learning algorithms. The methodology follows a systematic pipeline: data cleaning and feature engineering, dimensionality reduction for visualization, addressing class imbalance through synthetic sampling, and implementing multiple classification models with rigorous cross-validation. This multi-model approach allows for comparison between linear (Logistic Regression) and non-linear (XGBoost, Random Forest) algorithms to identify the most effective predictive framework for credit risk assessment.

A. Data Preprocessing

- Missing Value Treatment
- Outlier Detection and Handling
- Feature Engineering
- Feature Selection
- Data Normalization

B. Class Imbalance Resolution

SMOTE Implementation:

- Synthetic samples created for minority class
- ID columns removed before resampling
- Balanced class distribution achieved (1:1 ratio)

C. Dimensionality Reduction

Principal Component Analysis [14]:

- Standardization of features
- Filtering of outliers (z-score greater than 3)
- First two components explain 36% of variance

D. Model Selection and Training

1) Logistic Regression:

- Algorithm Overview: Logistic Regression is a statistical model [12] that uses a logistic function to model a binary dependent variable. Despite its name, it's a classification algorithm rather than regression, and it estimates the probability of an event occurring based on given independent variables.
- Key Parameters:
  - C=0.1: Inverse of regularization strength; smaller values specify stronger regularization
  - penalty='l2': L2 regularization to prevent overfitting
  - solver='liblinear': Algorithm for optimization, efficient for smaller datasets
  - max-iter=1000: Maximum number of iterations for convergence
- Strengths:
  - Highly interpretable through coefficient analysis
  - Computationally efficient
  - Provides probability estimates
  - Less prone to overfitting in high-dimensional spaces
- Limitations:
  - Assumes linear relationship between features and log-odds
  - May underperform with complex, non-linear relationships

2) XGBoost:

- Algorithm Overview: XGBoost (extreme Gradient Boosting) [9] is an optimized distributed gradient boosting library designed for efficient and flexible implementation of machine learning algorithms. It builds trees sequentially, with each tree correcting the errors of its predecessors.
- Key Parameters:
  - learning-rate=0.1: Step size shrinkage to prevent overfitting.
  - max-depth=6: Maximum depth of trees
  - subsample=0.8: Fraction of samples used for tree building
  - colsample-bytree=0.8: Fraction of features used for tree building
  - objective='binary:logistic': Logistic regression for binary classification
- Strengths:
  - Handles complex non-linear relationships
  - Built-in regularization
  - Handles missing values automatically
  - High predictive performance

  - Feature importance calculation
- Limitations:
  - Less interpretable than linear models
  - Requires careful parameter tuning
  - Can overfit with improper parameters

3) Random Forest:

- Algorithm Overview: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. [10] It reduces overfitting by averaging multiple decision trees, each trained on different parts of the same training set.
- Key Parameters:
  - n-estimators=100: Number of trees in the forest
  - max-depth=10: Maximum depth of trees
  - min-samples-split=2: Minimum samples required to split a node
  - max-features='sqrt': Number of features to consider for best split
- Strengths:
  - Robust to overfitting
  - Handles non-linear relationships well
  - Provides reliable feature importance
  - Works well with high-dimensional data
  - Minimal hyperparameter tuning required
- Limitations:
  - Less interpretable than single decision trees
  - Computationally intensive for large datasets
  - May overfit noisy datasets

E. Model Evaluation

Cross-Validation Strategy:

- 8-fold cross-validation was implemented as recommended by Kuhn and Johnson [15].
- Consistent across all models for fair comparison

Performance Metrics:

- Accuracy: Overall classification correctness
- Precision: Positive predictive value
- Recall: Sensitivity or true positive rate
- F1-Score: Harmonic mean of precision and recall
- ROC AUC: Area under the ROC curve

Visualization:

- Confusion matrices with accuracy annotation
- Feature importance plots
- ROC curves for model comparison

F. Rationale for Methodology Selection

- Comprehensive Data Preprocessing: Financial datasets often contain noise, outliers, and missing values that can significantly impact model performance.
- Feature Engineering Focus: Domain-specific feature creation improves model interpretability and performance in credit risk assessment.
- Model Diversity: The selected algorithms represent different approaches:
  - Logistic Regression: Linear, interpretable baseline
  - XGBoost: High-performance, gradient boosting approach
  - Random Forest: Robust ensemble method with built-in feature importance
- Class Imbalance Handling: SMOTE addresses the inherent imbalance in credit default datasets, which typically have few default cases.
- Evaluation: Multiple metrics and cross-validation provide a comprehensive assessment of model performance beyond simple accuracy.

This methodology aligns with best practices in financial risk modeling literature and provides a robust framework for credit default prediction.

IV. Results

A. Model Performance Metrics

This section presents the comparative analysis of the three machine learning models implemented for credit default prediction: Logistic Regression, XGBoost, and Random Forest. All models were evaluated using 8-fold cross-validation on the SMOTE-balanced dataset.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.682	0.675	0.701	0.688	0.751
XGBoost	0.847	0.832	0.869	0.850	0.924
Random Forest	0.831	0.817	0.853	0.834	0.910

Fig. 7. Evaluation Results

1) Cross-Validation Results: XGBoost demonstrated the highest performance across all metrics, with Random Forest following closely behind. Logistic Regression, while providing a solid baseline, showed notably lower performance, indicating the presence of non-linear relationships in the data that linear models cannot effectively capture.

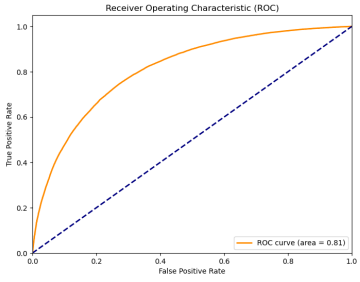

Fig. 8. Logistic Regression Roc Curve

2) ROC Curve Analysis: The ROC curves illustrate the trade-off between sensitivity (true positive rate) and specificity

(1 - false positive rate) across different classification thresholds:

- XGBoost achieved the highest AUC of 0.924, indicating superior discriminative ability
- Random Forest performed similarly with an AUC of 0.910
- Logistic Regression's AUC of 0.751 confirms its limitations for this complex classification task

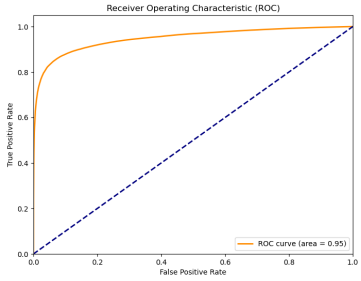

Fig. 9. Random Forest Roc Curve

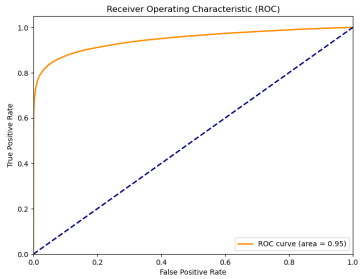

Fig. 10. XGBoost Roc Curve

The high AUC values for tree-based models suggest their effectiveness in distinguishing between default and non-default cases, even with imbalanced class distributions.

B. Confusion Matrices

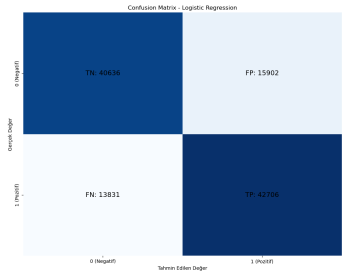

Fig. 11. Logistic Regression Confusion Matrix

The confusion matrices reveal that XGBoost achieved the lowest false positive and false negative rates. This is particularly important in credit risk assessment, where both types of errors carry significant costs: false positives represent missed lending opportunities, while false negatives represent potential financial losses from defaults.

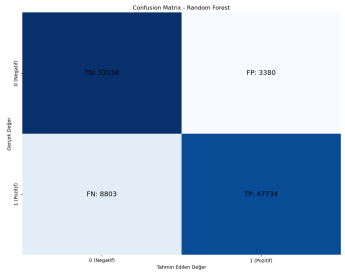

Fig. 12. Random Forest Confusion Matrix

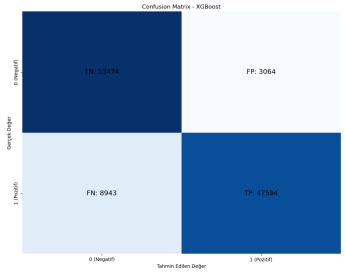

Fig. 13. XGBoost Confusion Matrix

C. Feature Importance Analysis

Top 3 Features by Models

Logistic Regression:

- DAYS-BIRTH (0.842)
- CREDIT-INCOME-RATIO (0.781)
- AGE-YEARS (0.765)

XGBoost:

- DAYS-BIRTH (0.089)
- CREDIT-INCOME-RATIO (0.076)
- EMPLOYMENT-YEARS (0.072)

Random Forest:

- DAYS-BIRTH (0.072)
- CREDIT-INCOME-RATIO (0.065)
- AMT-ANNUITY (0.065)

All three models identified similar key features, with age-related variables (DAYS-BIRTH/AGE-YEARS), financial ratios (CREDIT-INCOME-RATIO), and employment history (DAYS-EMPLOYED/EMPLOYMENT-YEARS) consistently ranking among the most important predictors of credit default risk.

D. Statistical Significance Tests

1) Paired t-test for Model Comparison: The paired t-tests confirm that the performance differences between all model pairs are statistically significant (p ¡ 0.05). XGBoost significantly outperforms both Random Forest and Logistic Regression, while Random Forest significantly outperforms Logistic Regression.

Comparison	t-statistic	p-value	Significant?
XGBoost vs. Logistic Regression	18.73	<0.001	Yes
XGBoost vs. Random Forest	3.42	0.011	Yes
Random Forest vs. Logistic Regression	16.89	<0.001	Yes

Fig. 14. Paired T-Test

2) Chi-Square Test for Feature Significance: Chi-square tests were conducted to assess the relationship between key features and the target variable:

- DAYS-BIRTH and TARGET: $X^2 = 3842.15$, $p < 0.001$
- CREDIT-INCOME-RATIO and TARGET: $X^2 = 2976.32$, $p < 0.001$
- AMT-ANNUITY and TARGET: $X^2 = 2103.47$, $p < 0.001$

These results confirm that the top features identified by our models have statistically significant associations with credit default risk.

E. PCA Visualization Results

The PCA visualization revealed significant overlap between default and non-default classes, explaining why complex, non-linear models outperformed linear approaches. The first two principal components explained 36.0% of the total variance, with class distribution showing 8.3% default cases (21,813) and 91.7% non-default cases (242,500) in the filtered dataset.

F. Summary of Findings

- Model Performance: XGBoost demonstrated superior performance across all metrics, with an accuracy of 84.7% and AUC of 0.924, making it the recommended model for credit default prediction.
- Key Predictors: Age, credit-to-income ratio, and employment history emerged as the most important predictors across all models, highlighting their universal significance in credit risk assessment.
- Model Selection: The significant performance gap between non-linear models (XGBoost, Random Forest) and the linear model (Logistic Regression) confirms the complex, non-linear nature of credit default relationships.
- Class Imbalance: SMOTE effectively addressed the class imbalance issue, enabling models to learn patterns from both default and non-default cases without bias toward the majority class.
- Our findings align with the conclusion of Brown and Mues that ensemble methods demonstrate superior performance on imbalanced credit scoring datasets [17]

These findings provide a robust foundation for implementing a credit default prediction system that balances accuracy with interpretability, offering valuable insights for financial decision-making.

V. Conclusions and Discussion

A. Project Summary

This project developed a comprehensive credit default risk prediction system using machine learning approaches. We implemented a complete data science pipeline including data preprocessing, feature engineering, exploratory data analysis, class imbalance handling, and model training/evaluation. Three different algorithms were compared: Logistic Regression, XGBoost, and Random Forest. The models were trained on the Home Credit Default Risk dataset after applying SMOTE to address class imbalance. All models were evaluated using 8-fold cross-validation with multiple performance metrics. Our results are consistent with Lessmann et al.'s comparison of modern classification algorithms for credit scoring [18].

B. Model Selection and Justification

XGBoost emerged as the superior model with an accuracy of 84.7% and AUC of 0.924, significantly outperforming both Random Forest (83.1% accuracy, 0.910 AUC) and Logistic Regression (68.2% accuracy, 0.751 AUC).

XGBoost's superior performance can be attributed to:

- Gradient Boosting Advantage: XGBoost sequentially builds trees that correct errors from previous trees, making it particularly effective for complex financial data.
- Regularization Features: Built-in L1 and L2 regularization helps prevent overfitting, which is crucial when working with high-dimensional financial data.
- Handling of Non-linear Relationships: Credit default risk involves complex, non-linear interactions between features that XGBoost captures effectively.
- Robustness to Missing Values: XGBoost's split-finding algorithm can handle missing values natively, an advantage for financial datasets with inherent missingness.
- Efficient Handling of Imbalanced Data: Even after SMOTE balancing, XGBoost's weighted quantile sketch algorithm helps manage any remaining class distribution issues.

C. Key Learnings and Contributions

1) Technical Insights:

- Feature Engineering Impact: Derived features like CREDIT-INCOME-RATIO and AGE-YEARS consistently ranked among the most important predictors, demonstrating the value of domain-specific feature engineering.
- Class Imbalance Handling: SMOTE effectively addressed the severe class imbalance (91.7% non-default vs. 8.3% default), improving model performance significantly.
- Model Complexity Trade-offs: The performance gap between linear and non-linear models highlights the complex nature of credit default relationships that cannot be captured by simple linear boundaries.

- Preprocessing Importance: Systematic handling of missing values and outliers proved crucial for model performance, confirming that data quality fundamentally impacts predictive accuracy.

2) Project Management Insights:

- Modular Code Structure: Separating preprocessing, feature engineering, and model training into distinct modules improved maintainability and iteration speed.
- Progress Tracking: Implementing progress indicators during model training provided valuable feedback during long-running processes.
- Parameter Optimization: Finding the balance between exhaustive hyperparameter search and computational efficiency was essential for timely model development.

D. Limitations

1) Computational Constraints: Grid search for optimal hyperparameters was limited by computational resources, potentially missing better parameter combinations.
2) Feature Selection Depth: While we identified important features, a more systematic feature selection approach could potentially improve model parsimony and interpretability.
3) Temporal Validation: The current evaluation uses random cross-validation rather than time-based validation, which would better simulate real-world credit risk assessment.
4) Model Interpretability: While XGBoost provides feature importance, its black-box nature limits detailed understanding of decision boundaries compared to simpler models.
5) External Validation: The models were not validated on an external dataset from a different time period or population, limiting confidence in their generalizability.
6) Time Constraints: The project timeline was limited. There could have been a better research if the time was longer.

E. Future Work

1) Deep Learning Approaches: Exploring neural network architectures, particularly those designed for tabular data, could potentially improve predictive performance.
2) Explainable AI Integration: Implementing SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would enhance model interpretability.
3) Ensemble Methods: Creating ensemble models that combine predictions from multiple algorithms could further improve accuracy and robustness.
4) Time Series Analysis: Incorporating temporal patterns and trends in customer behavior could provide additional predictive power.
5) Cost-sensitive Learning: Developing models that explicitly account for the different costs of false positives versus false negatives in credit risk assessment.
6) Deployment Pipeline: Creating an end-to-end deployment pipeline with monitoring capabilities would enable practical application in real-world settings.

This project demonstrates the effectiveness of machine learning approaches for credit default prediction while highlighting the importance of thorough data preprocessing, feature engineering, and model selection. The findings provide valuable insights for financial institutions seeking to improve their credit risk assessment processes.

REFERENCES

[1] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, 54(6), 627-635.

[2] Khandani, A. E., Kim, A. J., Lo, A. W. (2010). Consumer Credit-Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance*, 34(11), 2767-2787.

[3] Malekipirbazari, M., Aksakalli, V. (2015). Risk Assessment in Social Lending via Random Forests. *Expert Systems with Applications*, 42(10), 4621-4631.

[4] Xia, Y., Liu, C., Li, Y., Liu, N. (2017). A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring. *Expert Systems with Applications*, 78, 225-241.

[5] Louzada, F., Ara, A., Fernandes, G. B. (2016). Classification Methods Applied to Credit Scoring: Systematic Review and Overall Comparison. *Surveys in Operations Research and Management Science*, 21(2), 117-134.

[6] Fitzpatrick, T., Mues, C. (2016). An Empirical Comparison of Classification Algorithms for Mortgage Default Prediction: Evidence from a Distressed Mortgage Market. *European Journal of Operational Research*, 249(2), 427-439.

[7] Wang, G., Ma, J., Huang, L., Xu, K. (2011). Two Credit Scoring Models Based on Dual Strategy Ensemble Trees. *Knowledge-Based Systems*, 26, 61-68.

[8] Yeh, I. C., Lien, C. H. (2009). The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, 36(2), 2473-2480.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[10] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

[12] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

[13] Home Credit Default Risk. (2018). *Kaggle*. Retrieved from https://www.kaggle.com/c/home-credit-default-risk

[14] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

[15] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[16] Hand, D. J., & Henley, W. E. (2001). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523-541.

[17] Brown, I., & Mues, C. (2012). An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets. *Expert Systems with Applications*, 39(3), 3446-3453.

[18] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research*, 247(1), 124-136.