HOME CREDIT Rakamin

SCORECARD MODELING

Data Science Project

PRESENTED BY:
Naufal Hadi Darmawan
GITHUB: https://github.com/NaufalHD12/home-credit-data-scientist-internship

Problem Statements

- How can we improve the accuracy of credit scoring models to minimize false rejections of creditworthy applicants?
- What machine learning techniques can be applied to enhance the prediction of loan repayment success?
- How can we ensure that loan terms are aligned with customers' financial capabilities?

Goals

- Develop a robust credit scoring model that accurately predicts a customer's likelihood of loan repayment.
- Implement machine learning models, including Logistic Regression and others, to compare performance and effectiveness.
- Ensure that the loan approval process is optimized to support both business sustainability and customer success.

Approaches:

- Data Preprocessing: Handle missing values, perform feature engineering, and normalize data for better model performance.
- Exploratory Data Analysis (EDA): Gain insights into key factors influencing credit risk using statistical analysis and visualizations.
- Machine Learning Models:

1. Logistic Regression: A baseline model for binary classification of loan repayment likelihood.
2. Advanced ML Techniques: Implement models such as XGBoost, LightGBM, and Random Forest to improve prediction accuracy.
3. Model Evaluation: Use metrics like Accuracy and F1-Score to assess model performance.

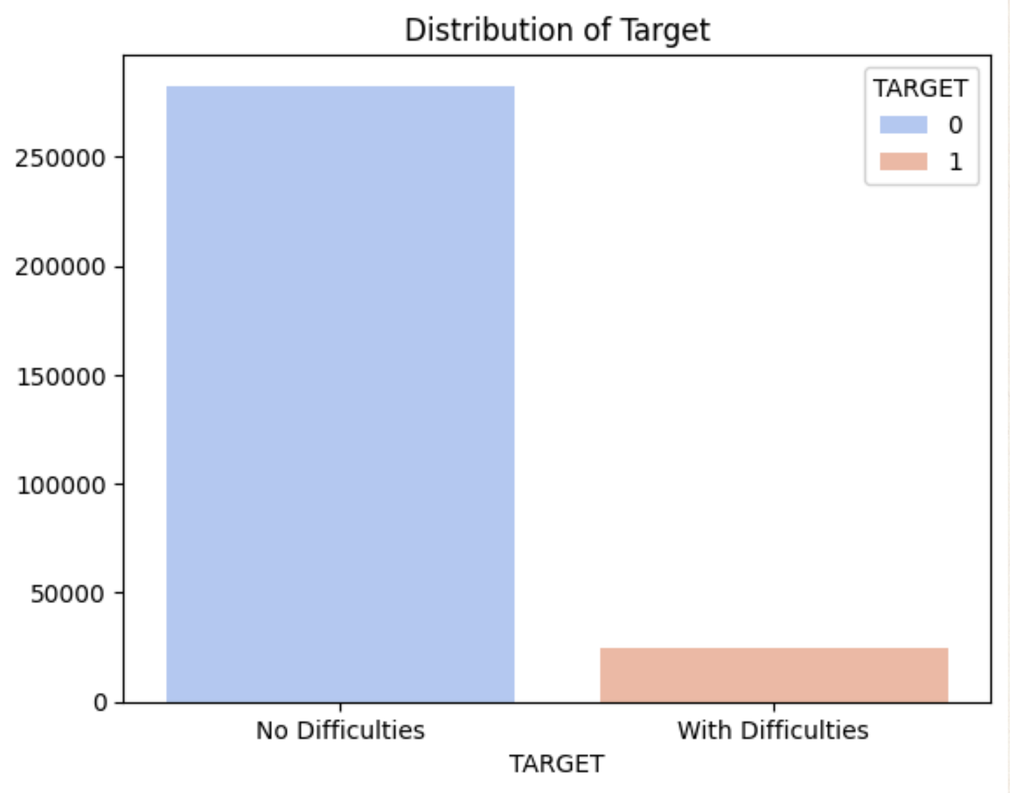
Project Domain

Background: Home Credit is currently utilizing various statistical methods and Machine Learning techniques to predict credit scores. Ensuring accurate credit assessment is crucial, as it prevents financially capable customers from being rejected while also mitigating the risk of loan defaults. By fully leveraging the potential of available data, we can optimize loan approvals with appropriate principal amounts, maturity periods, and repayment schedules that encourage successful repayments.**:**

Dataset Overview

The dataset contains **307,511 applicants with 122 features** providing extensive information about applicants' demographics, financial behavior, employment details, and loan history. With the majority **(282,686 or ~92%) classified as "No Difficulties" (TARGET = 0)**, meaning they had no issues repaying their loans. On the other hand, **24,825 applicants (~8%) fell into the "With Difficulties" category (TARGET = 1)**, indicating they struggled with repayment.

This dataset provides a comprehensive foundation for building a credit risk scorecard model, helping Home Credit Indonesia make data-driven lending decisions. 🚀



Exploratory Data Analysis

1

Performing an initial inspection using **.head()** to get a quick glimpse of the first few rows, followed by **.shape** to confirm the dataset's size.

#2

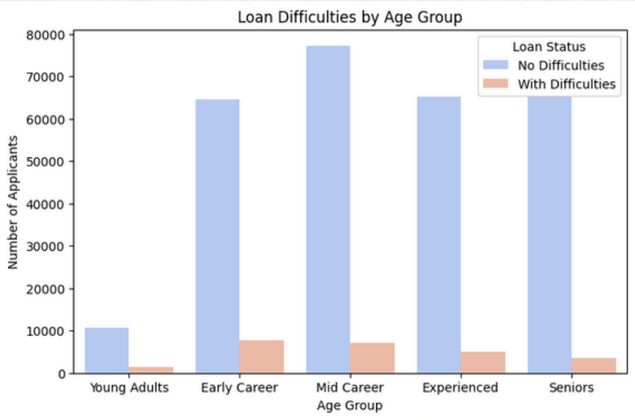
Next, **.info()** provided an overview of the dataset’s structure, showing the data types of each column and helping us identify potential missing values. A deeper look at missing using **.isnull()**, it revealed that some features had significant null values, requiring further handling.

#3

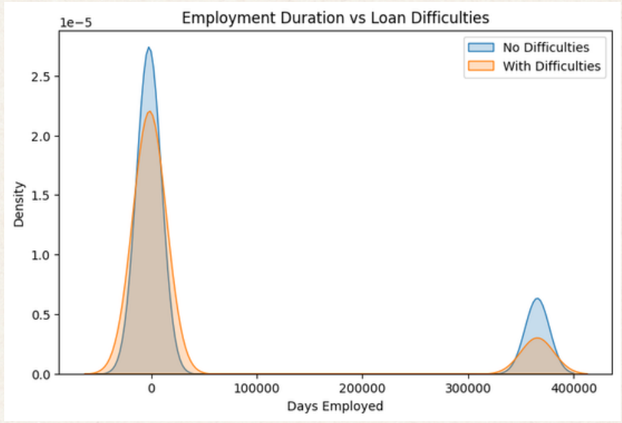
To ensure data integrity, we checked for duplicate entries but found none, indicating that each row represents a unique observation. This initial analysis helped us understand the dataset’s quality, completeness, and potential preprocessing needs before diving into deeper feature exploration and visualization.

Data Visualization

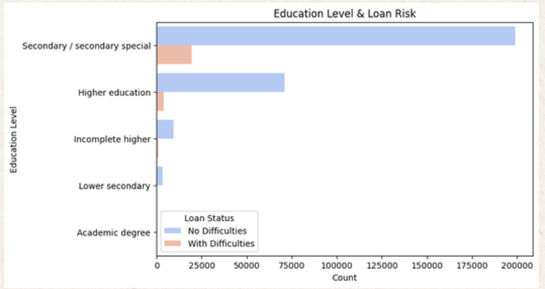
Leveraged chart for gaining some insights



- **Mid-Career (84,274 applicants) and Early Career (72,302 applicants)** represent the largest segments of loan applicants. This suggests that people in their 30s and 40s are the most active in seeking financial assistance, possibly due to homeownership, family expenses, or career investments.
- **Experienced professionals (70,077 applicants) and Seniors (68,699 applicants)** maintain a strong presence in the loan market, likely leveraging credit for retirement planning, business ventures, or real estate investments.
- **Young Adults (12,159 applicants)** make up the smallest group. This could be attributed to limited credit history, lower income levels, or a preference for alternative financial options.



- **A large peak near zero days**, indicating a high number of applicants with short employment durations.
- A **smaller peak around 350,000 days**, representing individuals with exceptionally long employment histories.
- While most individuals with long employment histories tend to encounter fewer loan issues, the initial spike near zero highlights a significant risk among those with very short employment durations.



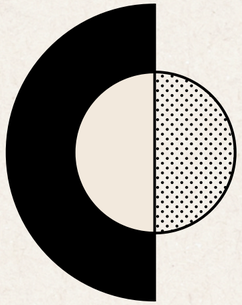
- **Most loan applicants have a basic or vocational education**, highlighting the need for financial products tailored to their income levels.
- **A significant portion holds college or university degrees**, likely seeking loans for homeownership, business investments, or further education. Their higher qualifications may contribute to better financial literacy and loan management.
- **Incomplete Higher Education** represents individuals who started but did not complete college, possibly young professionals or students with limited credit history and fluctuating income.
- **Lower-secondary group** may face financial instability due to early school dropout or unskilled employment.
- **Academic degree holders (PhDs/researchers) are a niche segment**, likely having alternative financial resources or lower reliance on loans.

Feature Selection

FEATURE SELECTION IS ESSENTIAL TO IMPROVE MODEL PERFORMANCE, REDUCE COMPLEXITY, AND ENHANCE INTERPRETABILITY.

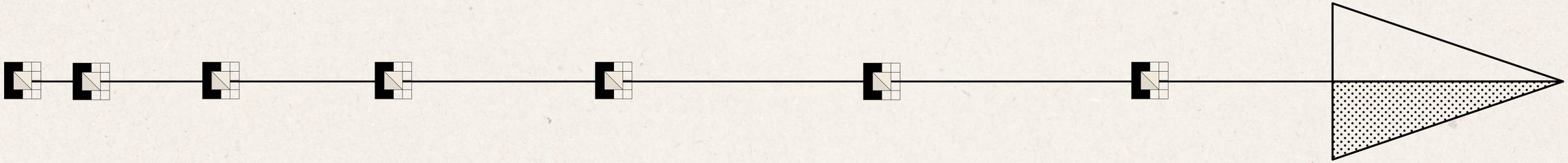
BY APPLYING THESE TECHNIQUES, WE RETAIN ONLY THE MOST INFORMATIVE 30 FEATURES, ENSURING BETTER PREDICTIVE PERFORMANCE AND REDUCING UNNECESSARY COMPLEXITY.

Drop columns with too many missing values (threshold: 50%)
Remove highly correlated numerical features (0.85)
Compute mutual information scores
Select top 30 important features based on mutual information scores



Data Preparation / Pre-Processing

prepare the dataset for machine learning modeling



Splitting Features and Target

Separates predictors (X) from the target (y) for model training.

Identifying Feature Types

Numerical features require scaling, while categorical features need encoding.

Handling Negative Time Values:

Converts negative values (e.g., DAYS_BIRTH) to positive for consistency.

Preprocessing Pipeline:

- Numerical Pipeline:
1. Converts negative values, imputes missing data (median), and scales features.
- Categorical Pipeline:
2. Imputes missing values (most frequent) and applies one-hot encoding.

ColumnTransformer Usage

Efficiently processes numerical and categorical features in a single step.

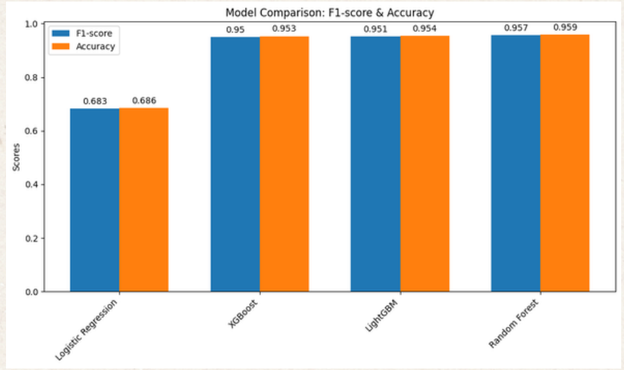
Handling Class Imbalance with SMOTE

- The dataset is imbalanced (fewer loan defaulters than non-defaulters).
- SMOTE generates synthetic samples to ensure both classes are equally represented, preventing bias in the model.

Data Splitting

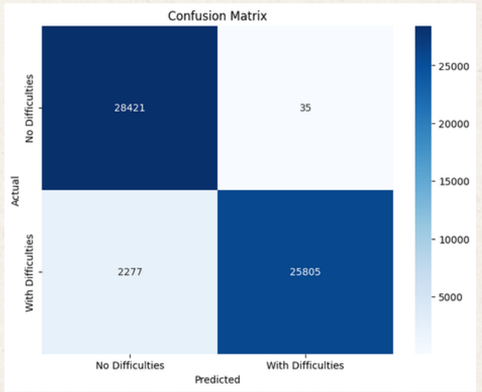
It sets aside 10% of the data for testing and 90% for training. 90% ensures the model has enough examples to learn patterns. 10% test set is enough to evaluate performance without sacrificing training data. Since we used SMOTE to handle class imbalance, we still have enough minority class examples in both training and test sets.

Modeling



I chose **Random Forest** because it achieved the highest performance among all tested models, **with an accuracy of 0.9571 and an F1-score of 0.9591**, outperforming Logistic Regression, XGBoost, and LightGBM. Its ensemble approach provides strong generalization, reducing overfitting while maintaining high predictive power, making it the best choice for this classification task.

	precision	recall	f1-score
0.0	0.93	1.00	0.96
1.0	1.00	0.92	0.96
accuracy			0.96
macro avg	0.96	0.96	0.96
weighted avg	0.96	0.96	0.96



Random Forest
Type: Ensemble of Decision Trees
How It Works:

- Random Forest is an ensemble learning method that creates multiple decision trees and combines their predictions.
- Each tree is trained on a random subset of the data (bootstrap sampling).
- The final prediction is made by majority voting (classification) or averaging (regression).

Key Features

✅ Robust to overfitting due to averaging multiple trees.
✅ Handles missing data and categorical variables well.
✅ Good for feature importance analysis.

Limitations

❌ Slower training & prediction time than single models.
❌ Not as interpretable as Logistic Regression.

Conclusion

Recommendation Based on EDA

- Consider age groups when structuring repayment plans, offering flexible options for young borrowers.
- Incorporate employment duration as a key risk factor to identify high-risk applicants.
- Offer credit-building products for young professionals and short-term employees.
- Provide budgeting and loan management workshops for secondary-educated customers.
- Design short-term, lower-interest loans for individuals with limited job history.
- Expand awareness of revolving loans to increase their adoption among qualified customers.

Impact of Scorecard Model on Home Credit Indonesia

Implementing a scorecard model can significantly enhance Home Credit Indonesia's risk assessment and loan approval process. By leveraging key features such as employment duration, credit history, financial behavior, and demographic factors, the model helps identify customers more likely to face loan difficulties.

Key Benefits

- Improved Risk Prediction
- Better Credit Decisions
- Optimized Customer Segmentation
- Fair and Transparent Lending
- Enhanced Profitability

By integrating a robust scorecard model, Home Credit Indonesia can enhance loan portfolio quality, improve customer experience, and drive sustainable growth while maintaining a competitive edge in the financial market. 🚀

Thank you

Reference:

- https://youtu.be/YaKMeAlHgqQ?si=NknPMUiUrWsybZzL
- https://towardsdatascience.com/select-features-for-machine-learning-model-with-mutual-information-534fe387d5c8/#:~:text=The%20MI%20score%20will%20fall,this%20feature%20and%20the%20target.
- https://youtu.be/hCwTDTdYirg?si=g_U00aLgg4hdoS11