Model Results and Findings.

David Kamau Irungu

September 13, 2024

Outline

Introduction

1. The project involved data preparation and credit scoring model development using machine learning techniques in the realm of insurance. A scorecard is crucial for assessing risk and, to some extent, determining premium rates for customers. The use of machine learning in the scorecard generation process enhances performance but may risk explainability in some cases.
2. Objectives
   - To develop a credit scoring model using machine learning techniques.
   - To format data and enhance model performance.
3. Assumptions
   - Current FX rate $= KES : 130.00$.
   - The data provided has been aggregated per customer level.
   - Sampling windows were well selected during data preparation.

ICEA LION GROUP

Data preprocessing

Data cleaning.

- *Missingness* Some features in the data had missing information:
    - Amount : 242 missing cases
    - Currency : 695 missing cases

  For the modeling purposes, Amount was imputes based on the other features.
- *Conversations* The data had different currencies(USD and KES), before proceeding with the modeling, the features like ‘Amount‘ were converted to similar currencies using the latest FX rate.

Data and Preprocessing

Feature engineering.

Some new features were engineered from the existing data. These features could be argued to have predictive power in the case of credit scoring. They include:

- *inactive days* This feature is gives the number of days since the customer last transacted. In most cases customers with higher number of transaction days tend to have higher chances of them defaulting.
- *income to balance ratio* This feature is gives the relation of the customer's income and the account balance balance, it is expected that if the ratio is significantly low the customer is less likely to default.

Model choice and fitting.

Logistic regression.

- Logistic regression is a machine learning technique used for binary classification, meaning it classifies a scenario into two main categories. In this case, it helps determine whether a customer is likely to default or not. It outputs the probability of someone becoming a defaulter, and from this, we determine the optimal threshold to classify individuals as defaulters or non-defaulters. The model is known for its simplicity, performance and explainability.
- Optimal Binning Process. In machine learning realm, generalization is key. Feature grouping (binning) helps with generalization, the features were binned optimally. Each been were used to generate the predictive power of the commonly known as weight of evidence (WoE). Each case were mapped to its corresponding WoE, this was used as the predictors. This approach has been know to yield a better performance compared to using raw values in modeling.

Model choice and fitting.

Model Fitting.

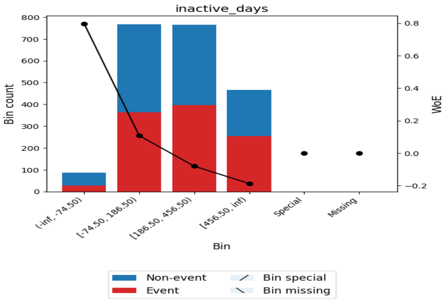
- *Oversampling* The data had class imbalance, the target class distribution was highly imbalance (more non-defaulter than defaulter over 86%). To prevent the model from being biased, there was need for oversampling the class with less occurrence (defaulters). This would ideally help the model sensitivity.
- *Parameter tuning* To enhance model performance, hyperparameter tuning was done for logistic regression. The best parameters were the used to fit the final model.
- *Data Split* The data was split into two sets, 80% for training set and 20% for testing set. The final model was trained using full dataset (for model that needs to go to prod).

ICEA LION GROUP

Model Performance

Variables

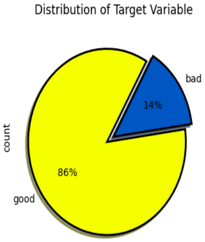
The newly engineered features displayed an impressive predictive power
The chart below shows descending monotinicity of inactive_days to default status:

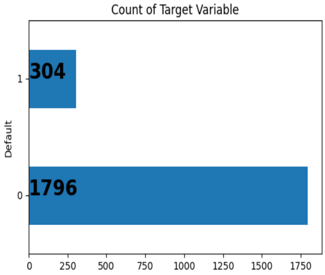


Model Performance

Target varaible

The target variable is highly imbalanced. The chart below shows the distribution of the target variable.



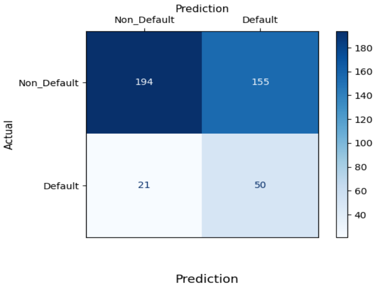


Model Performance

Performance metrics

The confusion matrix below derive this metrics.

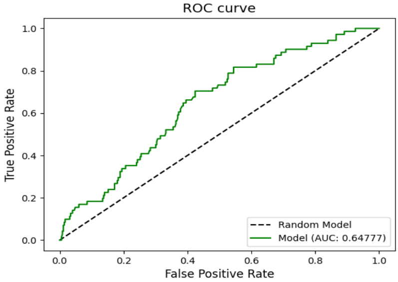
The f1_score::0.36 the recall_score::0.70 the precision_score::0.24



Model Performance

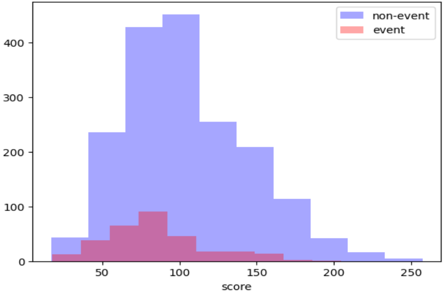
Performance metrics

The area under the ROC curve is 64% indicating a slightly better performing model. The AUC is greater than 50% indicating the model is better than random model.



Model output

After the model fitting, it is important to return a credit score for a customer, in this case a score range of $[0-300]$ was chosen. A customer having a higher score is less likely to default. The distribution of the scores clearly indicate some significant separability.



Model output: Suggested policy table

Policy table is generated once the model has been developed. It helps in optimal risk bands segmentation. In this case a 6-band risk category was assumed, it can be noted that for the first three bands i.e Low risk, medium low risk and medium risk are profitable and the cut-off score is between $[94 - 136]$. The decision here will be driven by the risk appetite.

Risk Category	Min Cut-off Score	Max Cut-off Score	Goods	Bads	Bad Rate (%)	Total Loss	Interest Income	Gross Margin
Low Risk	172	257	101	2	1.9	4544	29831	25287
Medium Low Risk	136	172	295	25	7.8	56800	87131	30331
Medium Risk	94	136	572	67	10.5	152224	168946	16722
Medium High Risk	72	94	426	101	19.2	229472	125823	-103649
High Risk	56	72	251	54	17.7	122688	74135	-48553
Very High Risk	16	56	150	55	26.8	124960	44304	-80656

Table: Snippet of the suggested policy table


THANK YOU