Data Project II

Preventing NPL Using a Probability of Default (PD) Model

Tanapon Tanavejwasin
6610424023
MADT 6004

Agenda

1. Business Requirement
2. Project Architecture
3. Probability of Default (PD Model)
4. Model Result Usage

Business Requirement

Business Requirement

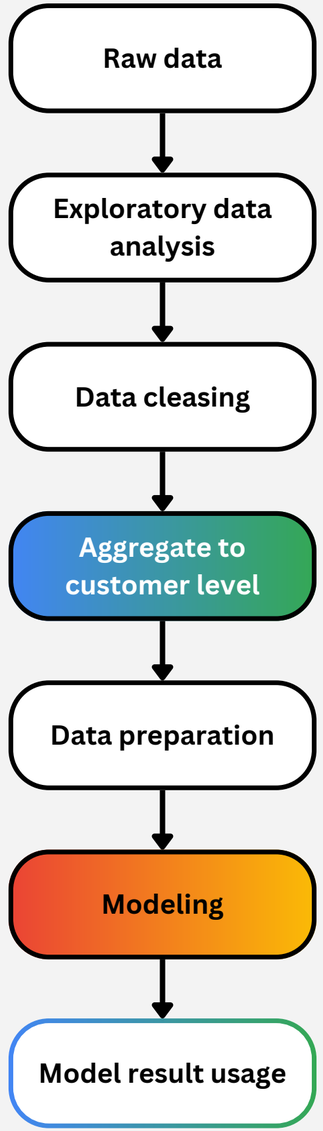
- The Business Unit would like to prevent non-profit loan (NPL) by applying machine learning model to assess borrower in underwriting process.
- The Business Unit need to cross-sell product with existing customer but still need to approve loan conservatively due to non-performing loan (NPL) is above threshold.

Project Architecture

Project Architecture

This project aims to conduct an experiment on a loan default prediction model using only behavioral data and income.

- First step: Retrieve the data source from Kaggle. This dataset contains four tables: application, bureau, bureau balance, and credit card balance. This project selects three out of four: application, bureau, and credit card balance.
- Second step: Perform exploratory data analysis.
- Third step: Clean the data.
- Fourth step: Aggregate the data at the customer level to predict customers who are likely to default.
- **Fifth step:** Perform data preparation.
- **Sixth step:** Train the model using XGBoost.
- **Finally:** Implement the model results.

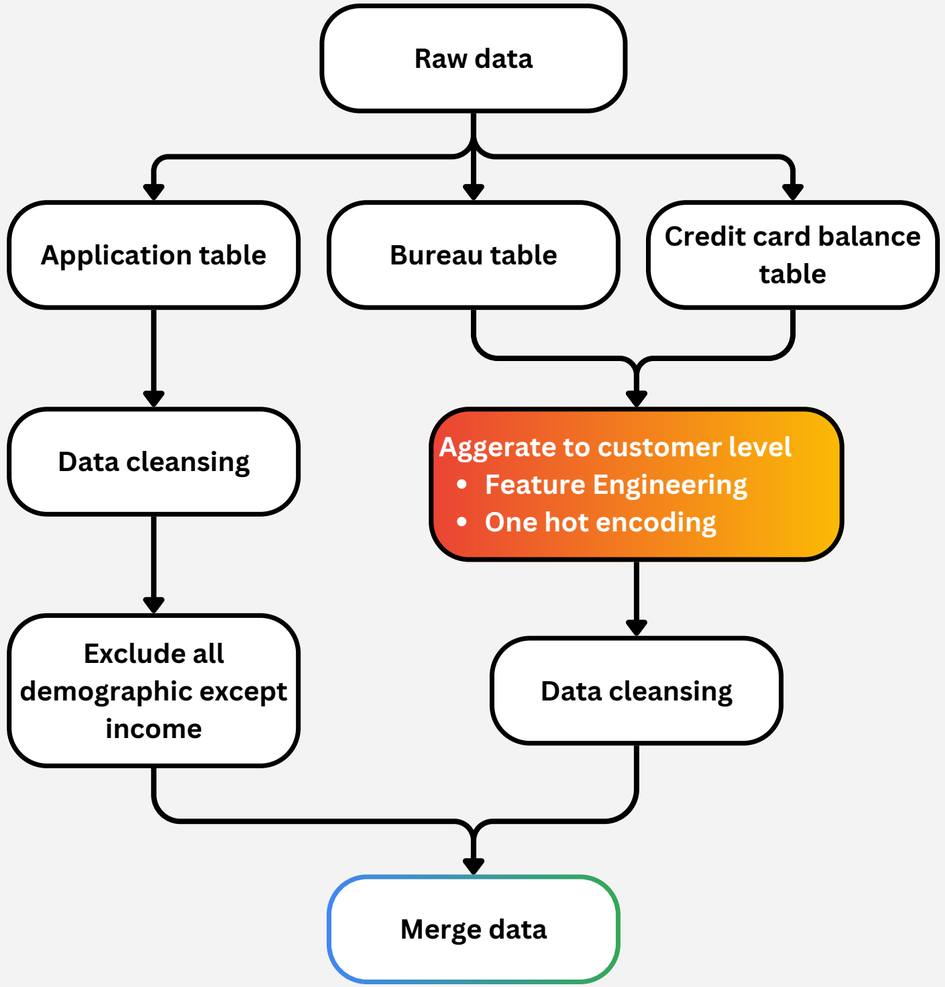


PD Model

PD Model

Feature Engineering

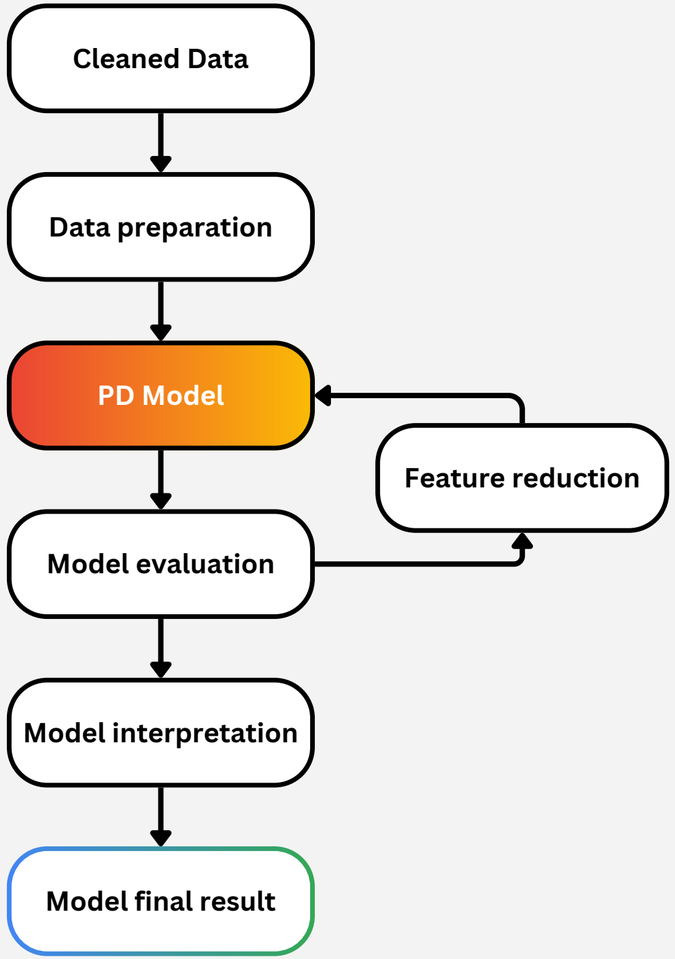
- **First step:** Perform exploratory data analysis (EDA) on all tables for understanding. Clean the application table data by assigning data types, removing outliers, filling null values, and performing label encoding. Exclude all demographic data except income.
- **Second step:** Aggregate the Bureau and Credit Card Balance tables from transaction data to the customer level, and perform feature engineering with statistics and one-hot encoding. Clean the data by filling null values and performing one-hot encoding.
- **Finally:** Merge all tables.



PD Model

Modeling

- **First step:** Perform data preparation for model training.
- **Second step:** Train the model using XGBoost.
- **Third step:** Evaluate the model using the AUC score.
- **Fourth step:** Perform feature reduction in four steps:
  - Remove features containing more than 90% missing values.
  - Remove features with zero contribution.
  - Keep one feature from each pair with a correlation greater than 80%.
  - Remove features with variance less than 1%.
- **Fifth step:** Evaluate the model using the AUC score.
- **Finally:** Implement the model results.

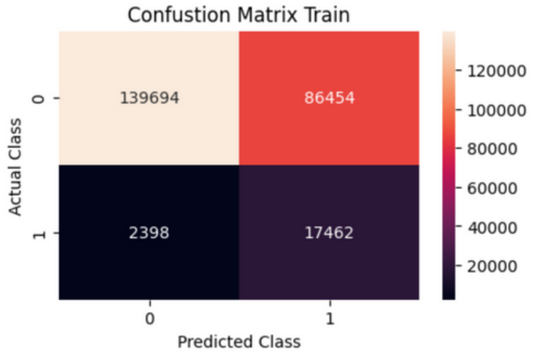


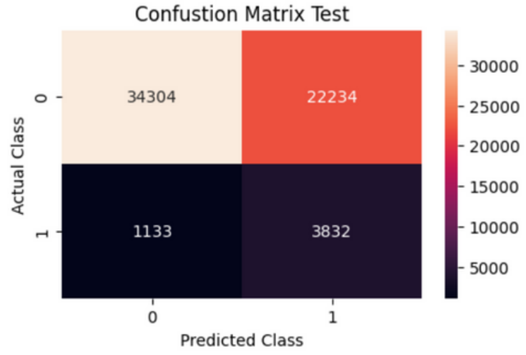
	Sk_id_curr	Predicted_Score	Predict	Actual
0	100002	0.894563	1	1
1	100003	0.254261	0	0
2	100004	0.252855	0	0
3	100006	0.505798	1	0
4	100007	0.551553	1	0

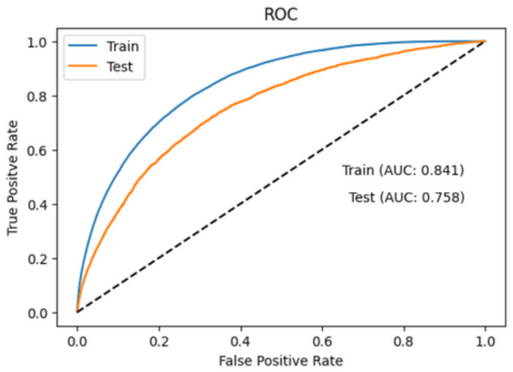
Example Result

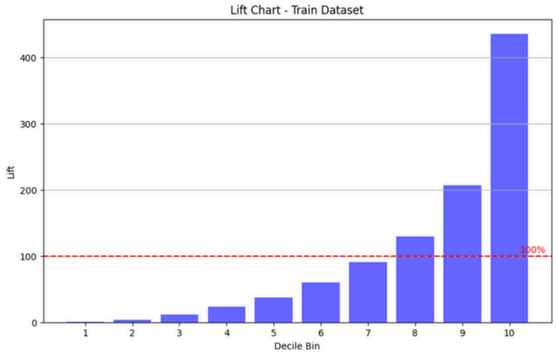
PD Model

Model Evaluation









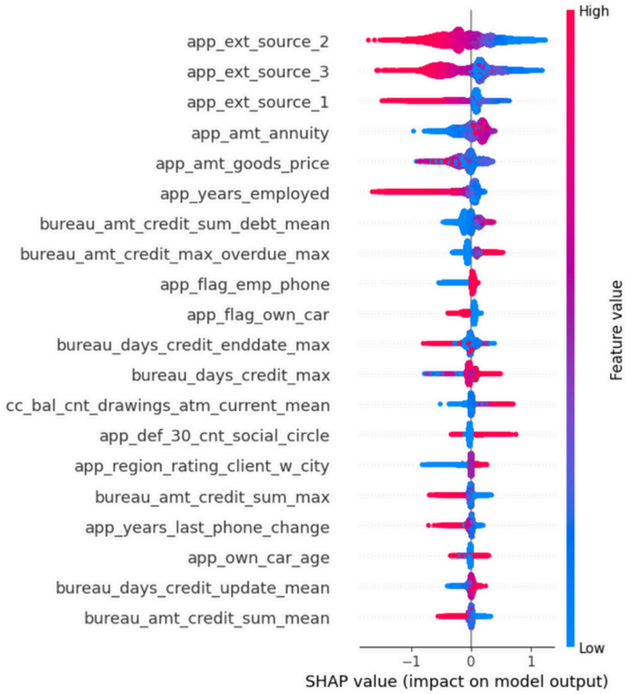
	model_name	parameters	rows	columns	%_target	accuracy	auc	precision	recall	f1-score
0	Train: XGBoost	<bound method XGBModel.get_params of XGBClassi...	246008	93	8	0.639	0.841	0.168	0.879	0.282
1	Test: XGBoost	<bound method XGBModel.get_params of XGBClassi...	61503	93	8	0.62	0.758	0.147	0.772	0.247

- This model performs well in terms of AUC on both the training and test sets. However, the slight drop from 0.8 to 0.75 suggests a bit of overfitting.
- High recall scores are beneficial as they reduce the risk of the bank missing potential defaulters, thereby preventing non-performing loans (NPL). In cases where there is conservative home loan approval, this approach can be effective. However, it could also lead to lost business opportunities and customer dissatisfaction.

PD Model

Model Interpretation

	feature_name	feature_importance	pct	cum_pct
1	app_ext_source_3	0.06153	6.153	6.153000
2	app_ext_source_2	0.05741	5.741	11.894000
3	app_flag_emp_phone	0.03483	3.483	15.377000
4	bureau_credit_type_microloan	0.02625	2.625	18.002001
5	cc_bal_cnt_drawings_atm_current_mean	0.02227	2.227	20.229000

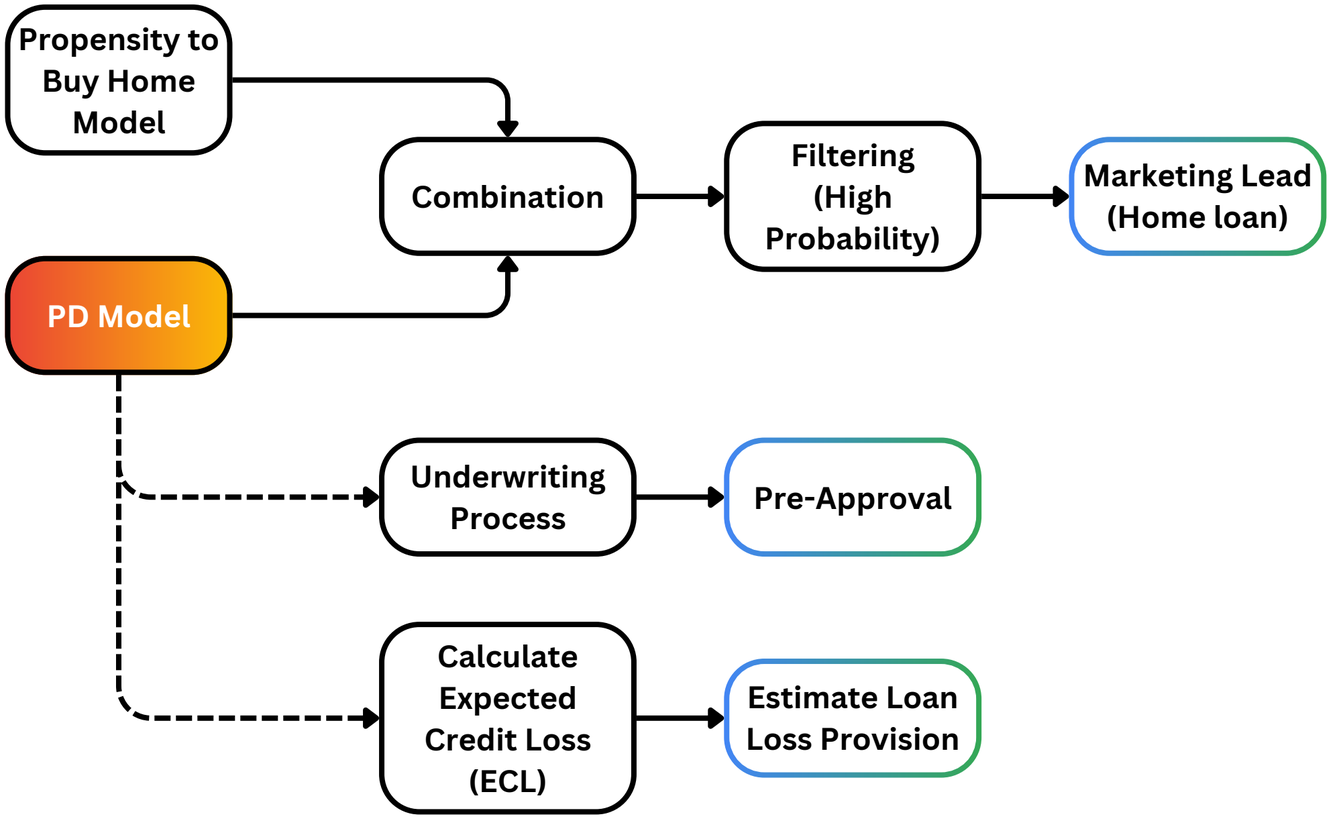


Example of model interpretation

- Top 3 feature importances come from external sources which indicate that the low value in the feature likely to be default
- Loan annuities indicate that the large size of the loan annuity likely to be default.
- Years employed indicate that having the number of years is lower likely to be default.

Model Result Usage

Model Result Usage



Summary

Summary

The PD model is used for assessing loan approval to reduce credit risk in terms of the underwriting process.

The PD model can be combined with other models, such as the propensity model, for lead selection.

The PD model is one of the metrics used to estimate expected credit loss.

This project experiment is based on behavioral data and income, and the results are quite good. In the future, hyperparameter tuning and feature reduction with RFECV can be performed to optimize the model performance while still using a manageable number of features.