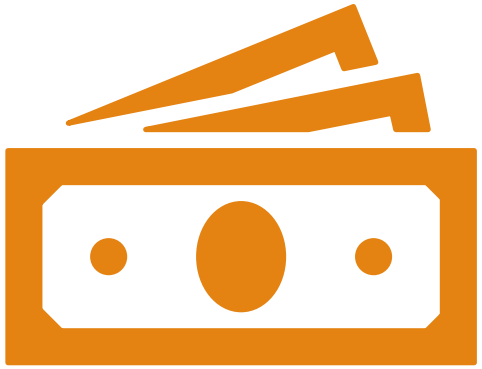
ERDÖS INSTITUTE SPRING 2024

APRIL 26, 2024



Predicting Loan Default

HOME CREDIT - CREDIT RISK MODEL STABILITY II

TEAM MEMBERS: BRADY ALI MEDINA, SZE HONG KONG

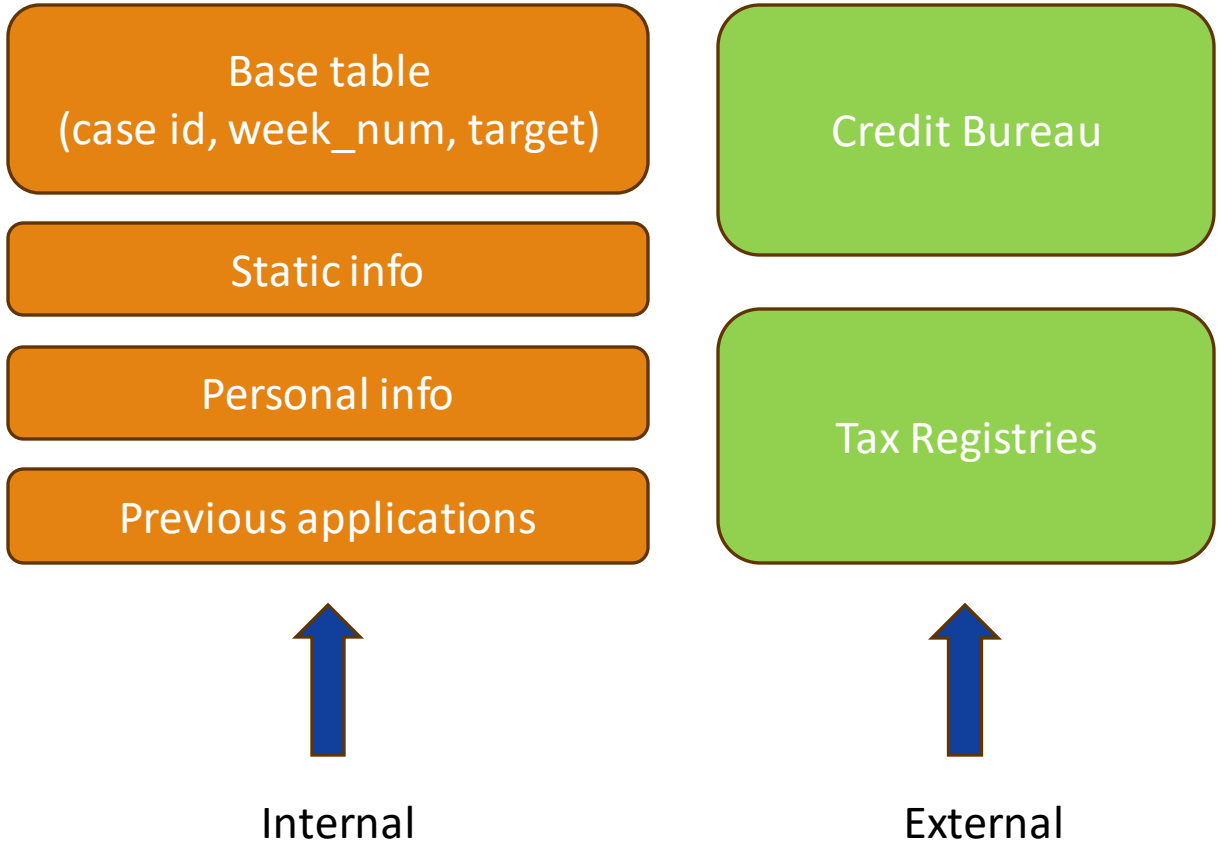
Introduction

- It is important for lenders to assess the risk of each borrower
- Particularly who with limited credit history

- This year, Home Credit hosts a Kaggle competition again
- Looking for a model predicting the risk of default that performs stably.

Dataset description

- Case_id can be used to join other tables
- For tables with depth>1, there is a num_group columns for aggregation
- Default ratio is ~3.14%, a higly imbalanced dataset
- Lots of missing values



The evaluation metric

Gini = 2* AUC –1

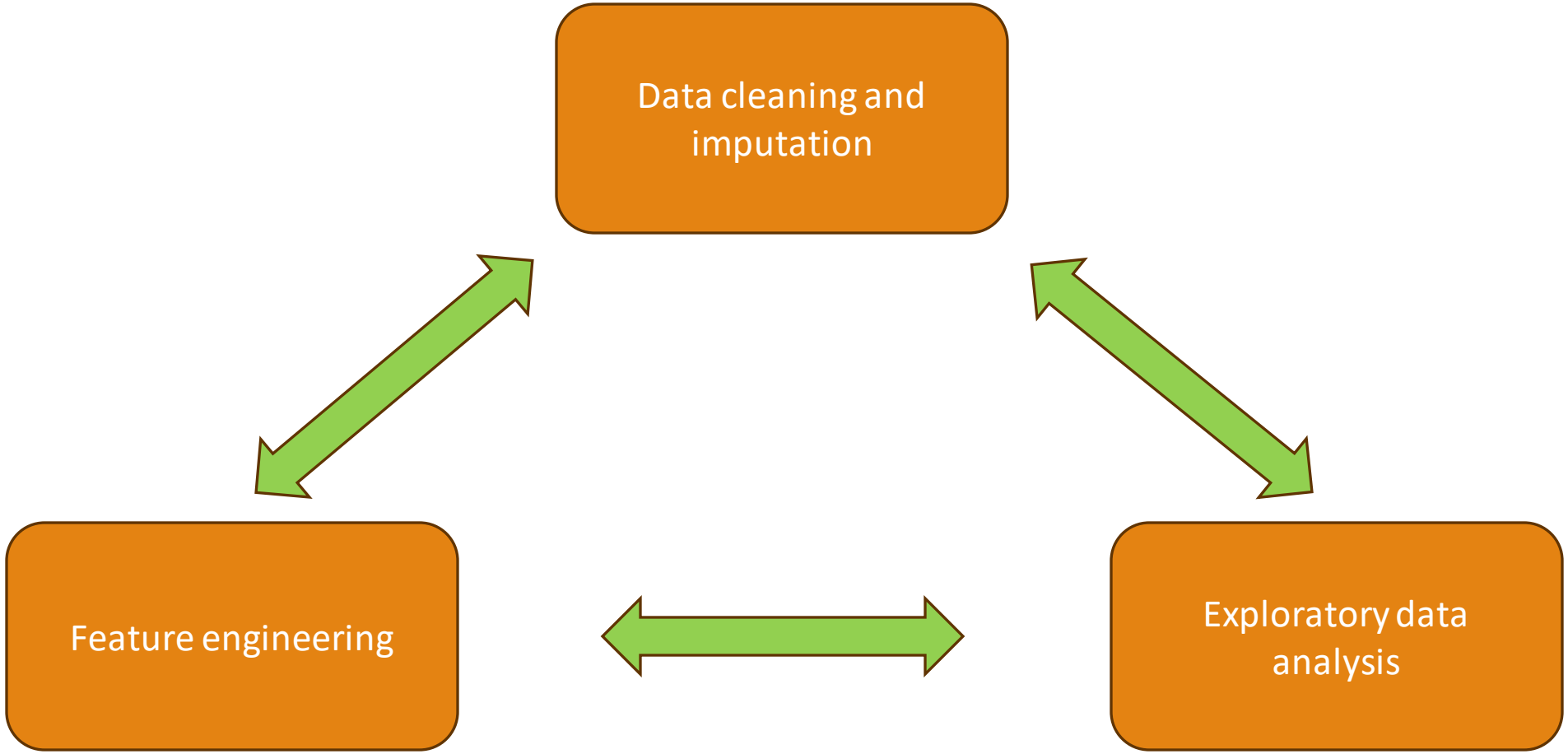
Fitted through weekly gini -> y = ax+b

Used to penalize drop in performance

Stability metric = mean(gini) + 88.0* min(0,a) - 0.5 * std(residuals)

Data processing

- In some cases, we fill in null values with zeros (e.g. chilnum)
- Birth info: string -> int


Data cleaning and imputation
Feature engineering
Exploratory data analysis

Data imputation and feature engineering

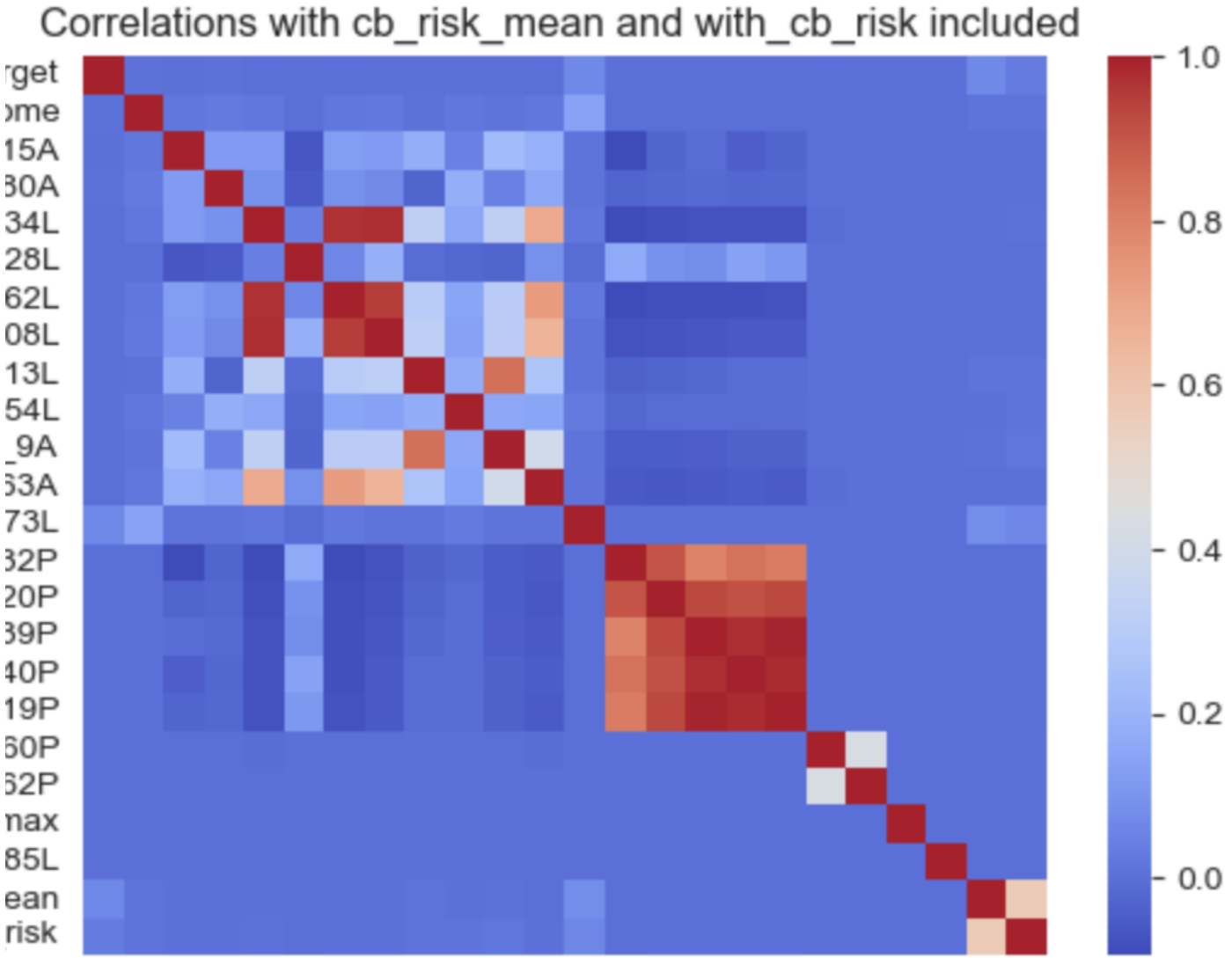
Imputation:

- Filled by medians
- Missing indicators

Feature engineering:

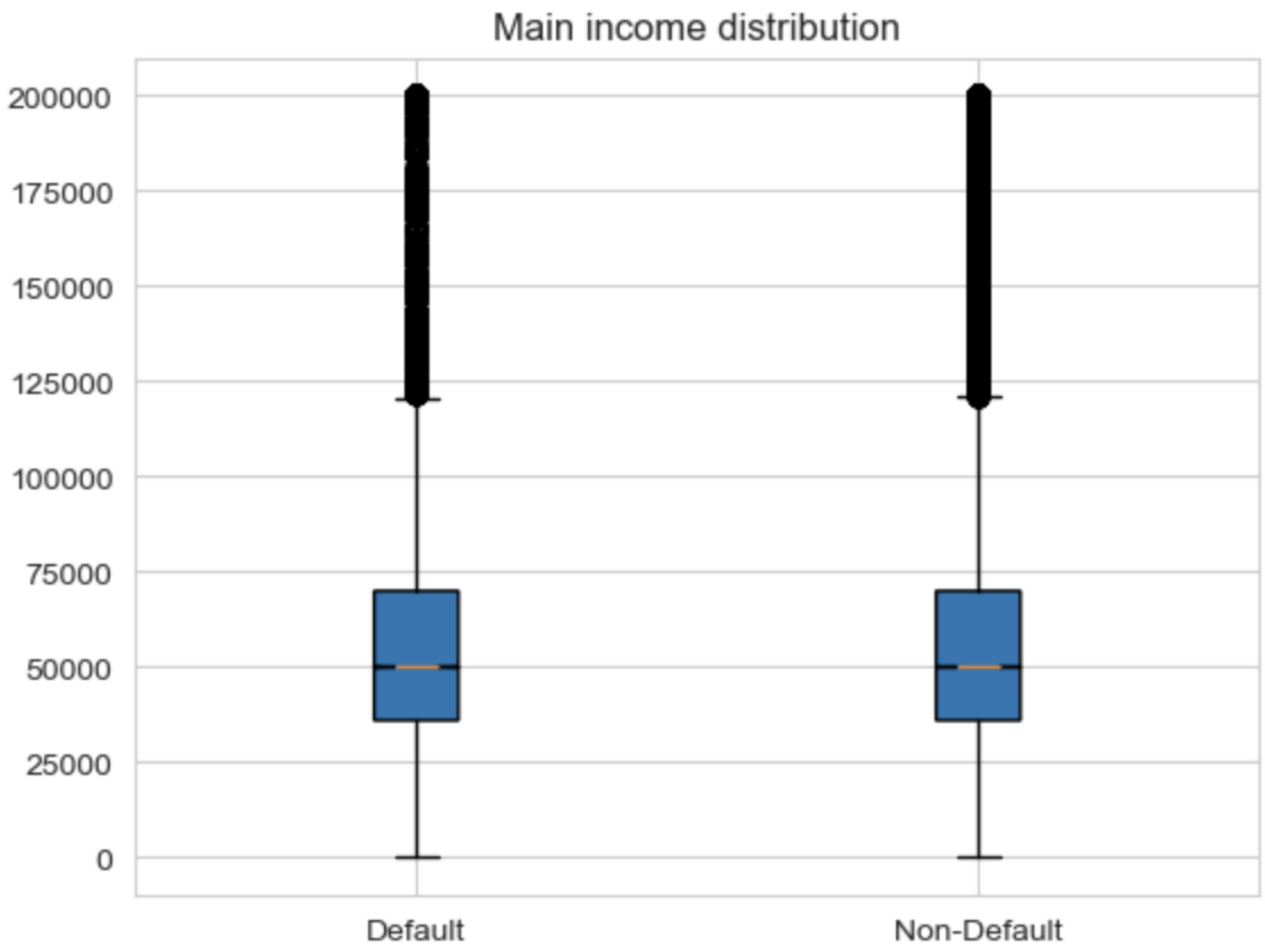
- Weighted dpd statistics
- Risk assesment of past credit history from credit bureau

(e.g. 5%-10% chance of default -> 7.5 (float)



Exploratory data analysis

- Fairily week correlations
- The weighted statistics raises the correlations with target slightly
- 'cb_risk_mean' has the highest correlation (~0.2), though it has >80% missing values



Exploratory data analysis

- The basic static information does not tell default and non-default apart.
- Things get better when stratified.

Modeling: a binary classification task

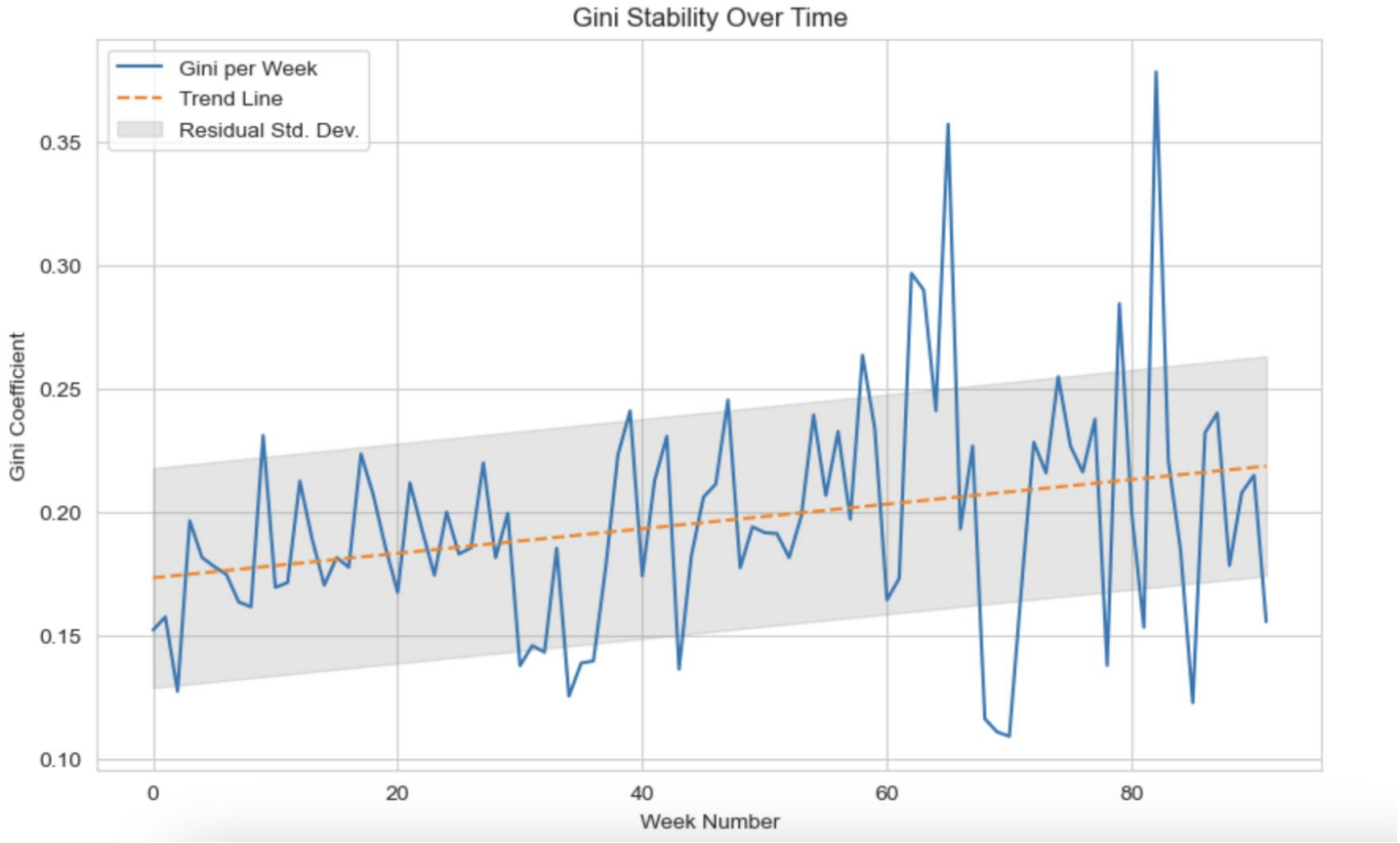
Baseline: Random forest

XGBoost

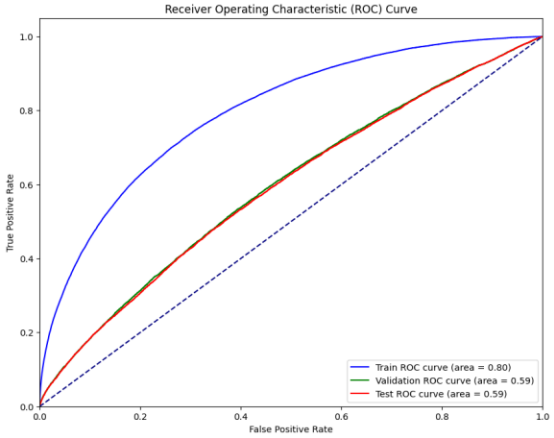
MLP

LightGBM

Model: XGBoost



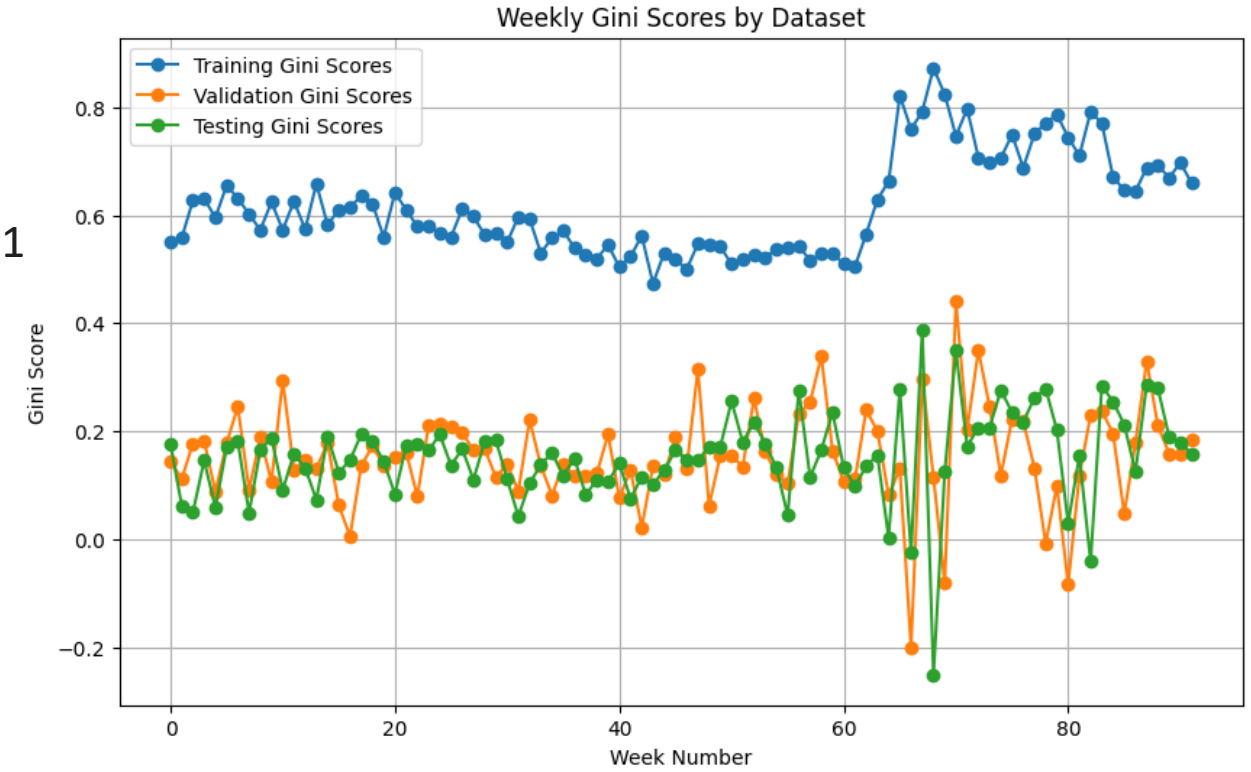
- No imputation required
- Runs pretty fast
- Performance not satisfactory



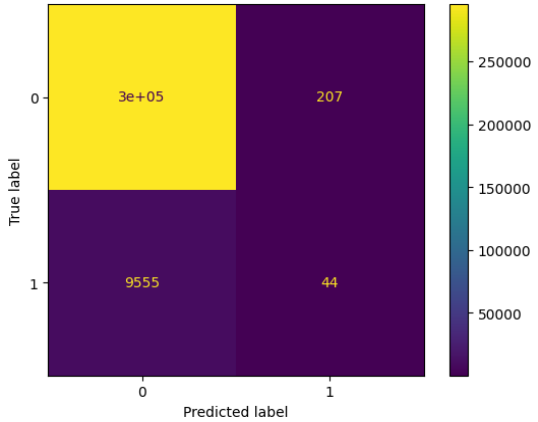
Training ROC AUC Score: 0.7963112370103307
Validation ROC AUC Score: 0.5947242876872081
Testing ROC AUC Score: 0.5914501713516093

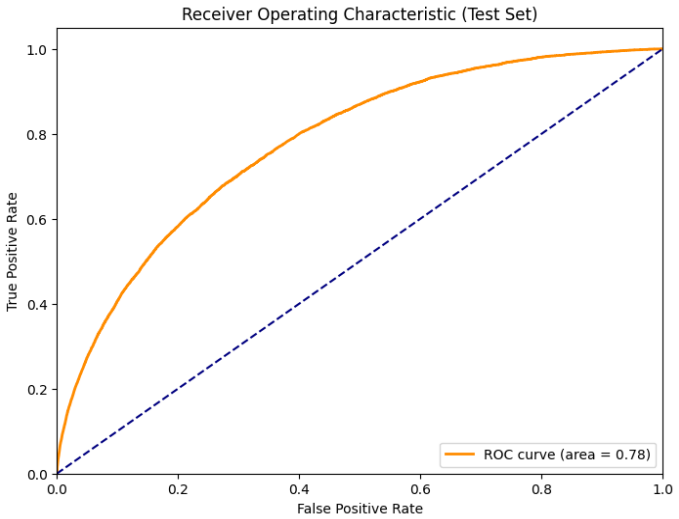
Model: MLP

- Imputation required
- Running Time: 53 min 55 sec
- Performance not satisfactory



- **Stability score train set:** 0.5771
- **Stability score valid set:** 0.1086
- **Stability score test set:** 0.1113



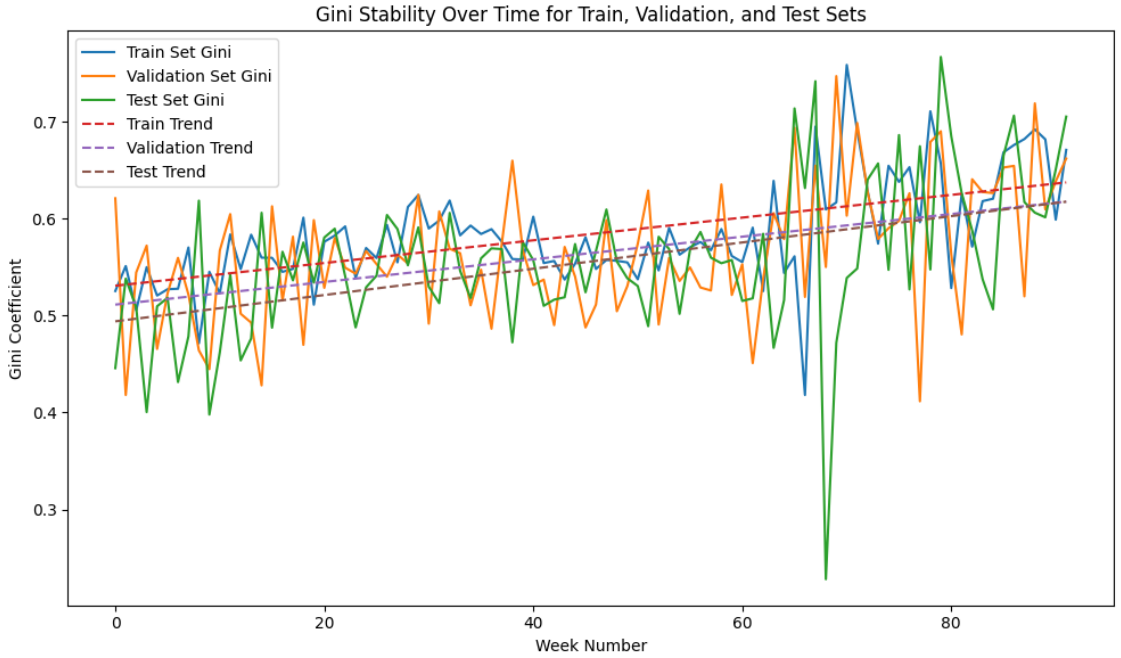


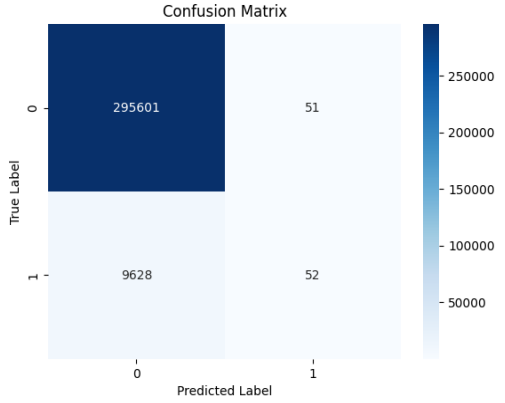
The AUC score on the train set is: 0.7922
The AUC score on the valid set is: 0.7823
The AUC score on the test set is: 0.7784

Model: Light GBM

- No Imputation required
- Runs pretty fast
- Performance satisfactory

- **Stability score train set:** 0.5621
- **Stability score valid set:** 0.5343
- **Stability score test set:** 0.52108



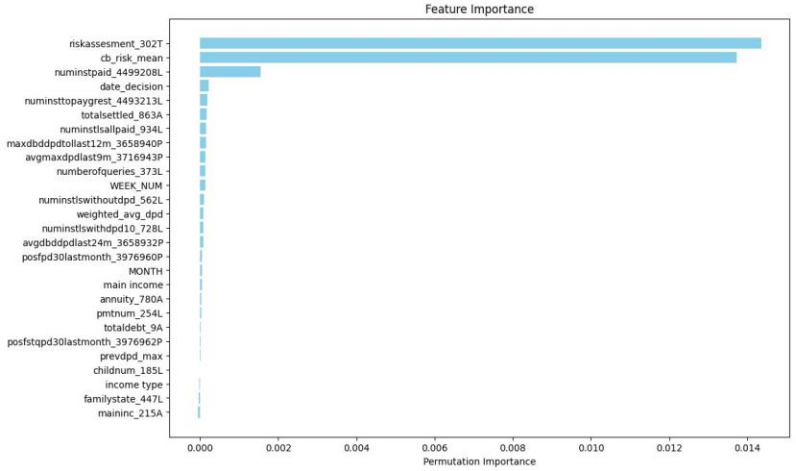


Summary of Performance (in stability metric)

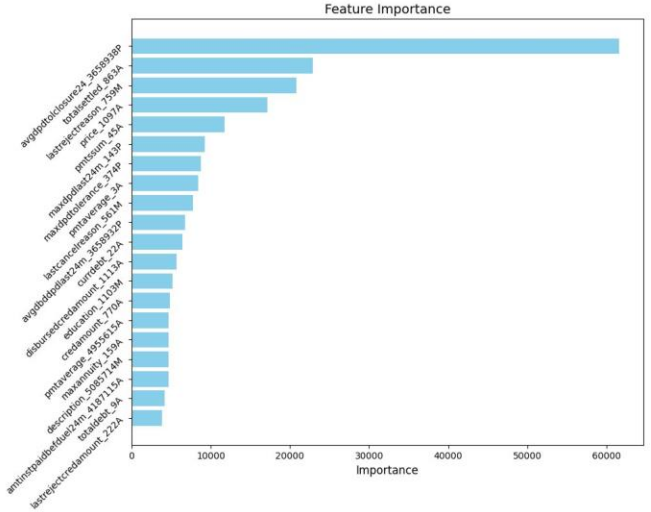
	XGBRFClassifier (baseline)	**XGBoost**	**MLP Classifier**	**LightGBM**
Training	-0.000	0.236	0.5771	0.562
Validating	-0.000	0.209	0.1086	0.534
Testing	-0.000	0.165	0.1113	0.521

Future work

- We are going to add the best features of our other models to the Light GBM.
- We are going to add external data such as the M2 Money Stock(measure of the total money supply in circulation).



Feature importance of MLP



Feature importance of LightGBM

Acknowledgments

Thank you to Roman Holowinsky, Matt Osborne, Alec Clott and the Erdös Institute for their support throughout the Summer 2024 boot camp.

Thank you Soumen Deb for his mentorship throughout the project.