Workplace Project: Profit scoring

December 6, 2024

Background Context

Profit scoring: Is essentially a system that estimates loan profitability. Similar to credit scoring,, however factors explaining profitability differ from variables explaining default probability.

The whole project was built on the basis that profit scoring is as equally important as credit scoring, however much emphasis is placed on credit scores by most credit institutions.

The are various way to develop a profit scoring system. Here is a high level overview of the approach we used:

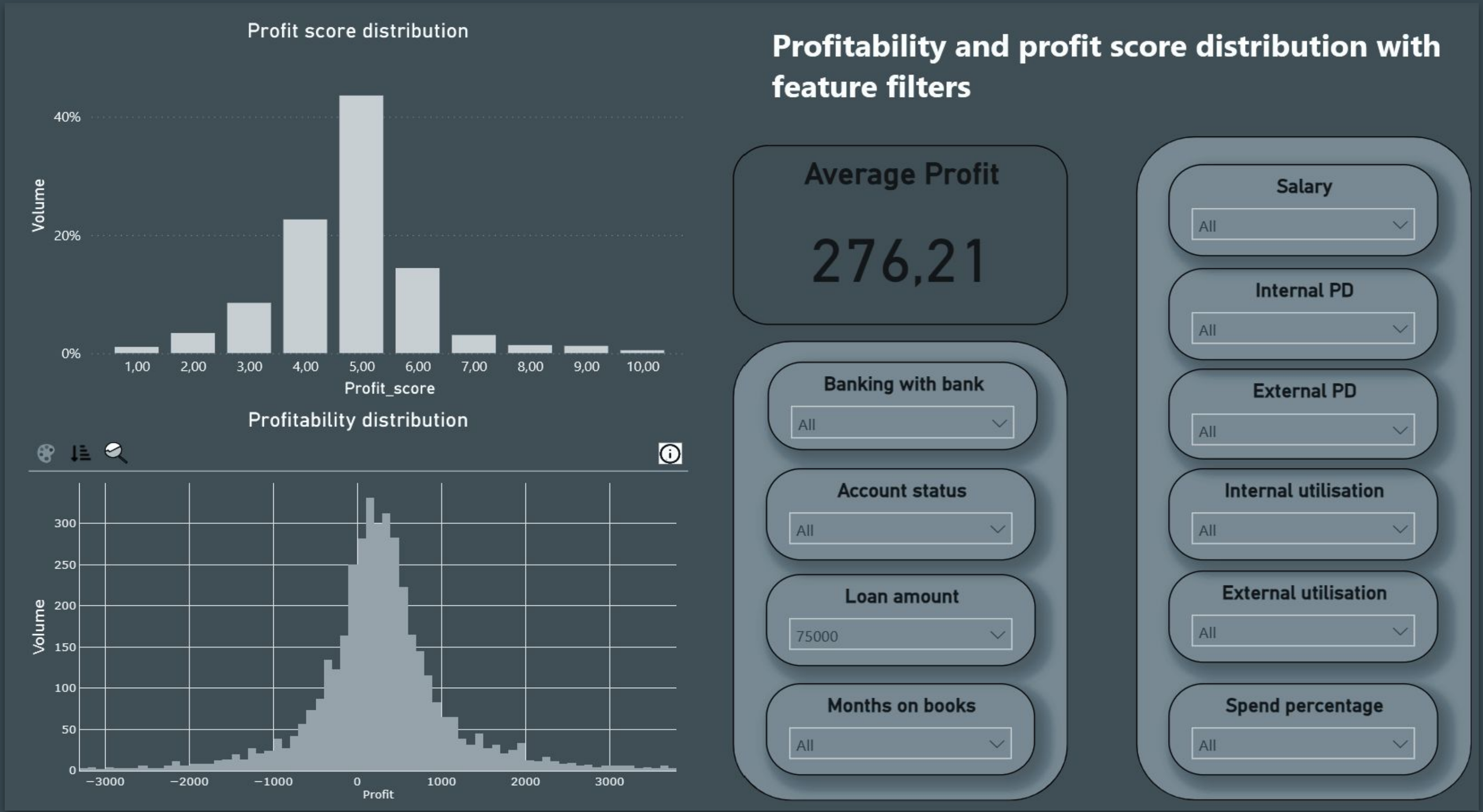
1) Profit score computation
- Comput relevant p**rofitability measures** (ROE etc) and normalize against industry measures.
- Determine the **consistency measure** of profitability measures in previous months.

2) Profit score prediction
This is essential the bulk of our project where we aim to find the best ML model that can predict profit scores.

Project objective:

The main objective of this project is to build a profit scoring systems that helps to identify high value clients (high profitability with low to medium risk of defaulting).

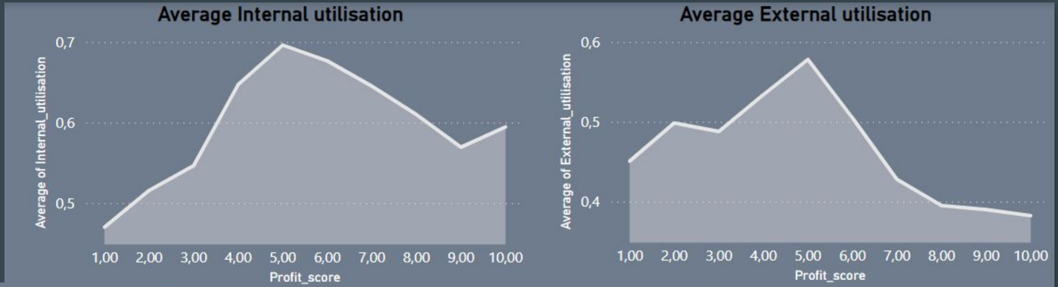
Exploratory data analysis

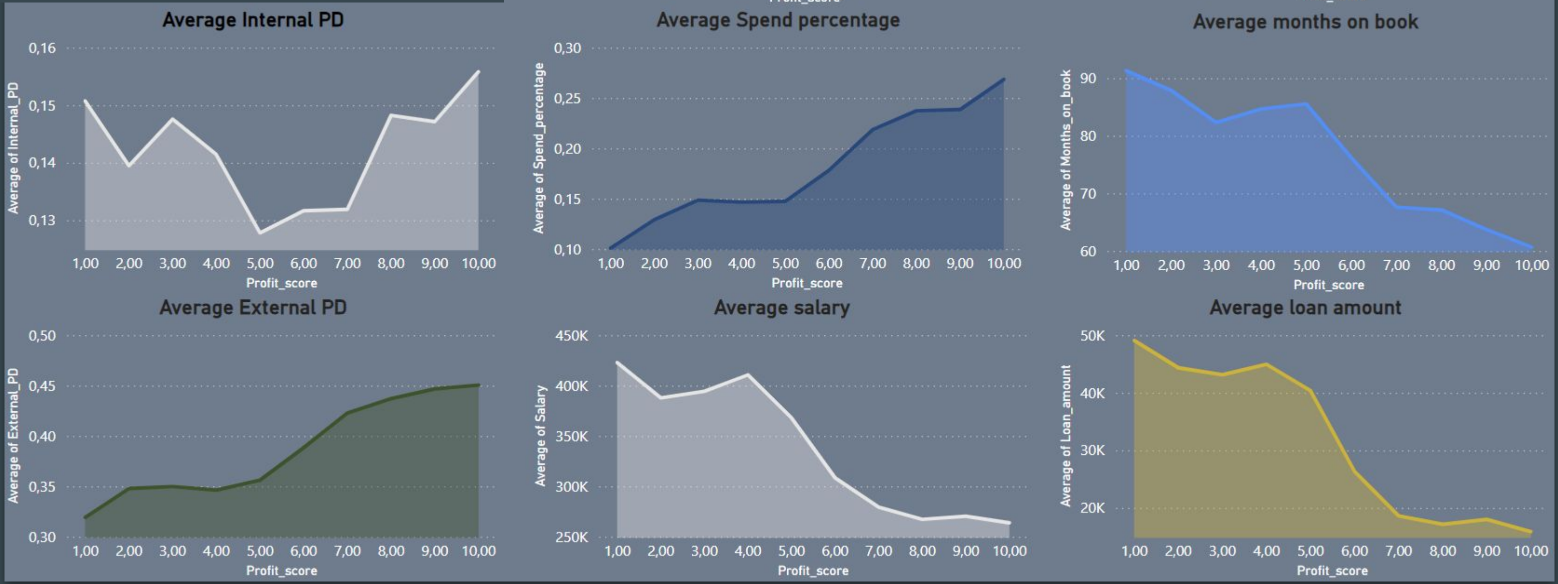


Extracted from : [dashboard - Power BI]

Exploratory data analysis

- Probability of default does not seem to have much influence on profitability
- Account with high profit scores seem to be those with low loan amounts, high spend percentage, utilisation about 60% and little utilisation with other external products
- Interesting behaviour with high profit scores being assigned high PDs by external vendor.





Extracted from : [dashboard - Power BI]

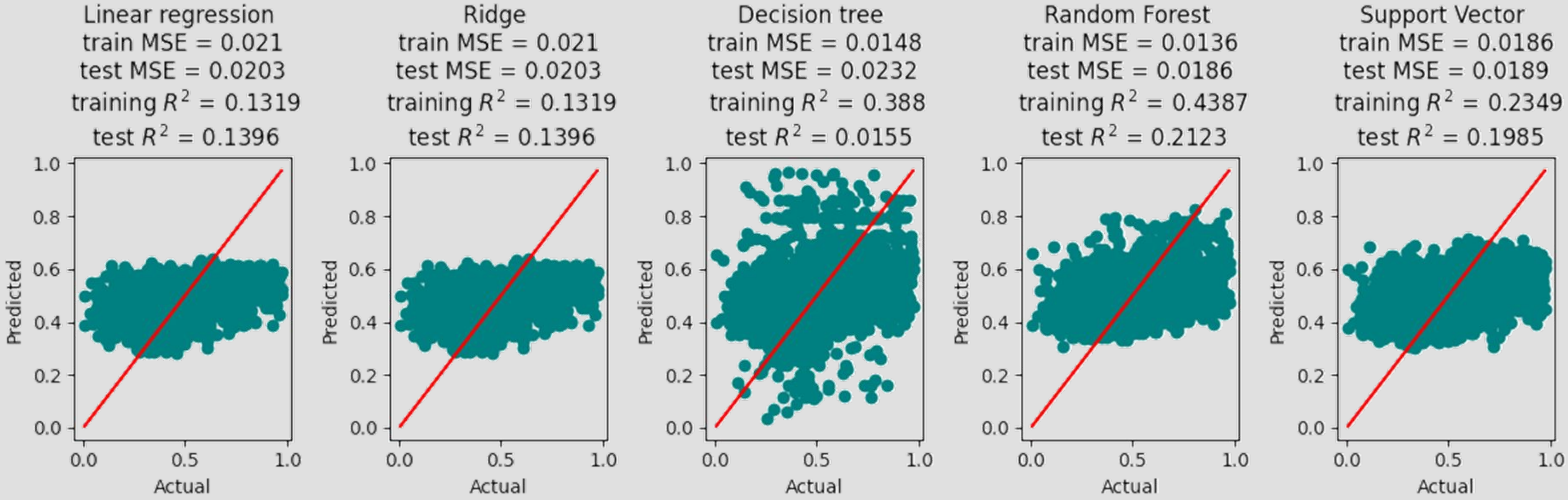
Modelling & Evaluation

Our dataset has both profit score classes and probabilities of profitability, therefore we can use both regression and classification models.

Modelling:

- Regression
- Classification
- Final model selection and improvements

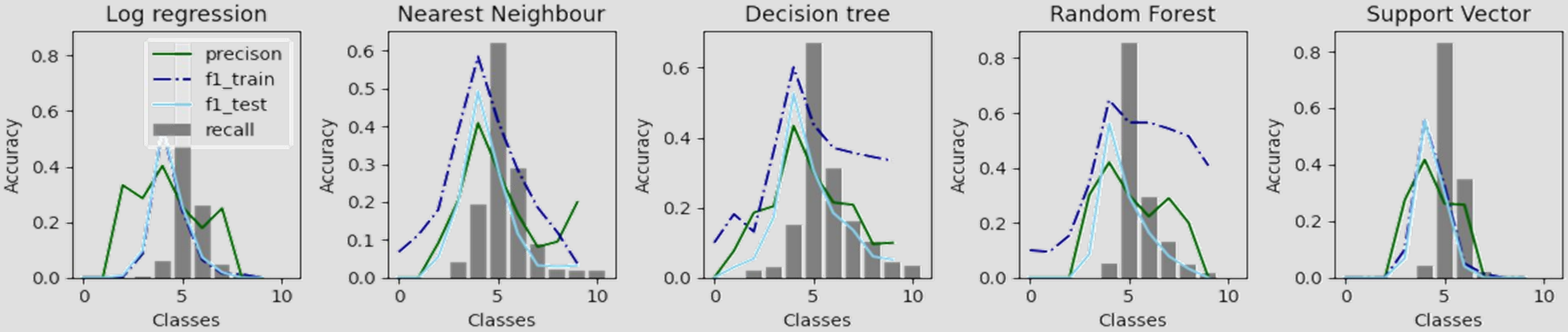
Regression models



- Ran various parameters for the 5 models, and used the optimal parameters for each model to evaluate the performance of each..
- The predictions have poor accuracy across all models, however random forest seems more promising.
- This is due to class imbalance as seen most of the predictions are concentration between 0.3 -0.8.

Classification models

- Tested a few parameter combinations with grid search and cross validation and compared the best combination of each model



- High F1 train score in comparison to the low F1 test scores indicates overfitting across all models.
- This is due to the class imbalance in the data.
- For the classification models random forest too has the best accuracy, precision and recall.

Classifier	Accuracy	Precision	Recall	F1 Train	F1 Test
Log regression	0.361531	0.289964	0.361531	0.257727	0.267603
Nearest Neighbour	0.320575	0.266454	0.320575	0.406993	0.281591
Decision tree	0.346416	0.295716	0.346416	0.438712	0.306127
Random Forest	0.380302	0.299180	0.380302	0.509458	0.291734
Support Vector	0.369332	0.269558	0.369332	0.280485	0.271025

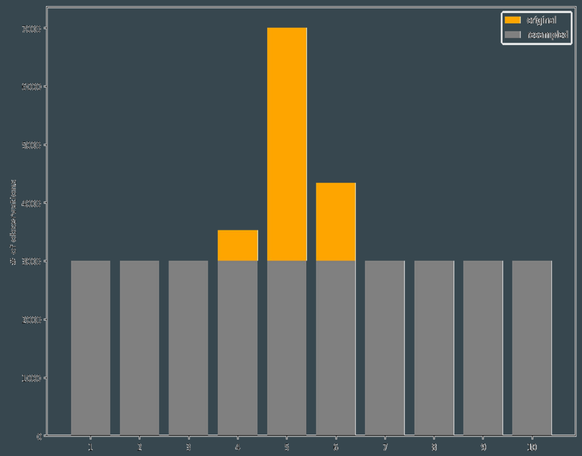
Final model

Classification reports indicate a very low accuracy. We know that our classes are heavily imbalanced. We rebalance the data by upscaling and refit the random forest classification model.

Initial model

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	28
2.0	0.00	0.00	0.00	107
3.0	0.00	0.00	0.00	252
4.0	0.24	0.04	0.07	682
5.0	0.43	0.83	0.56	1416
6.0	0.28	0.30	0.29	878
7.0	0.22	0.13	0.16	370
8.0	0.28	0.07	0.11	198
9.0	0.15	0.04	0.06	113
10.0	0.00	0.00	0.00	58
accuracy			0.37	4102
macro avg	0.16	0.14	0.13	4102
weighted avg	0.29	0.37	0.29	4102

Improved model



	precision	recall	f1-score	support
1.0	0.71	1.00	0.83	567
2.0	0.56	0.83	0.67	612
3.0	0.48	0.61	0.54	609
4.0	0.44	0.09	0.16	635
5.0	0.34	0.24	0.28	625
6.0	0.28	0.21	0.24	591
7.0	0.51	0.55	0.53	576
8.0	0.69	0.70	0.69	595
9.0	0.83	0.83	0.83	618
10.0	0.84	0.93	0.88	572
accuracy			0.59	6000
macro avg	0.57	0.60	0.56	6000
weighted avg	0.56	0.59	0.56	6000

- Balancing the class improves the accuracy from 37% to 59% accuracy, a 22% improvement.
- More improvements can be done further improve the accuracy, especially feature related improvements.

Final model

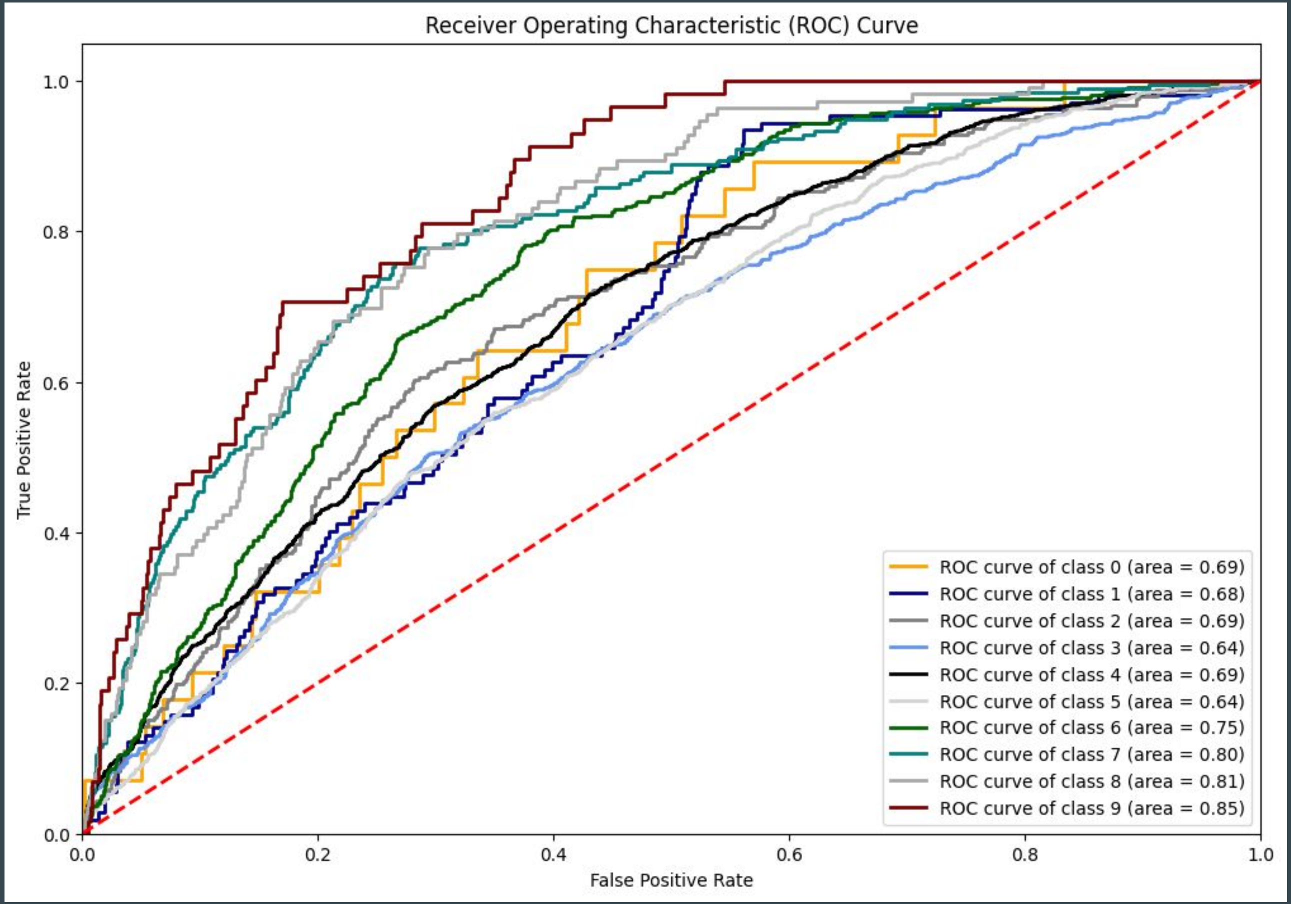
Model: Random Forest Classifier

Optimal parameters:

- N estimators = 100
- Max features = log2
- Max depth = 10

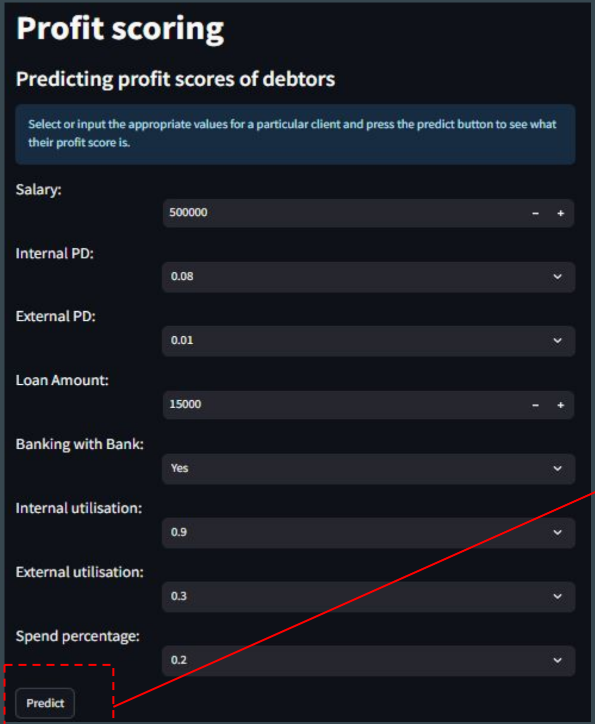
Weighted performance metrics:

- Accuracy: 59%
- Precision: 56%
- Recall: 59%
- F1_score: 56%



Streamlit app

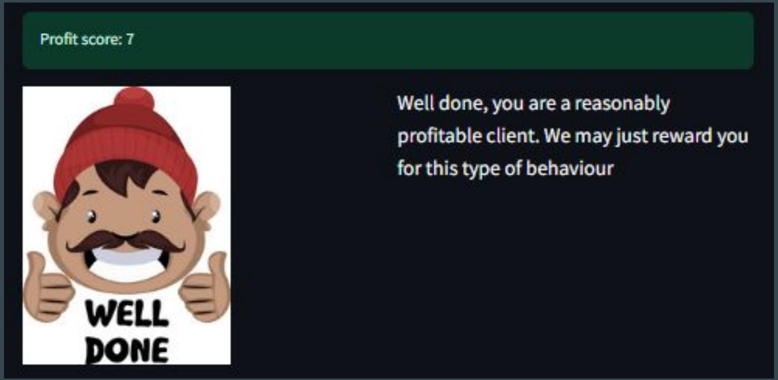
Input:



Press predict to get profit score prediction.

Output:

This is the output the app gives out. Your profit score, a sticker and text personalized for each profit score.



Link to [Streamlit app]

Compiled by

Vuyiswa Kubalasa

...