Credit Card Default Prediction

Muhammad Fadly

Executive Summary

Background

1. The data record historical behaviors of credit card customers as well as the label if they were default or not.
2. Defaults potential in the future could cost the Bank (credit risk) and compliance problems

Objective

1. Machine learning model is made to predict probability of default on credit card customers in the future

Results Summary

1. Several supervised machine learning model has been build, Random forest model shows the highest performance
2. Random forest model generate accuracy 93%
3. Final model could be used to give early notification of potential defaults in advance for further business action

Background

Profits

- Credit card defaults cost the banks
- Predicting such events could help the Bank to get the most efficient cost of credit calculation
- Thus, maximize the profits
- The Bank itself would get the higher investment rating if default rate is lower

Comply

- There are also maximum threshold of defaults set by the regulators to be complied on
- Business team could act faster to manage potential default debtors

Data Introduction

	X	jumlah_kartu	outstanding	limit_kredit	tagihan	total_pemakaian_tunai	total_pemakaian_retail	sisa_tagihan_tidak_terbayar
0	1	2	36158	7,00E+06	23437	0	94	26323
1	2	2	268691	1,00E+07	254564	0	1012	0
2	3	3	6769149	2,80E+07	4159779	0	0	0
3	4	4	3496732	2,10E+07	111231	0	2536660	581334
4	5	2	9402085	1,00E+07	6099283	0	2666558	5951865
5	6	2	6227439	8,00E+07	2081248	0	3690250	4613435
6	7	2	3906290	4,00E+06	2043682	0	230400	3314046
7	8	4	9534837	2,00E+07	3692028	0	9327612	7881069
8	9	2	4145065	5,00E+06	4021399	0	335680	4122425
9	10	4	1818606	7,00E+06	1765911	0	0	1627786

Content

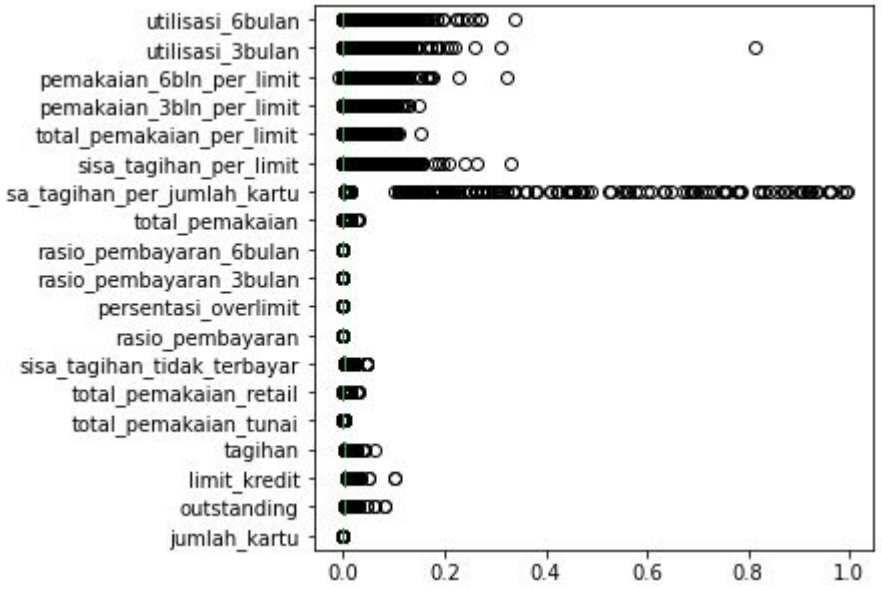
- 15K unique CC Holder
- 23 attributes
- Attributes related to credit card usage only

Condition

- 100 missing values in branch code column
- Many columns have mix format
- Imbalance target variables

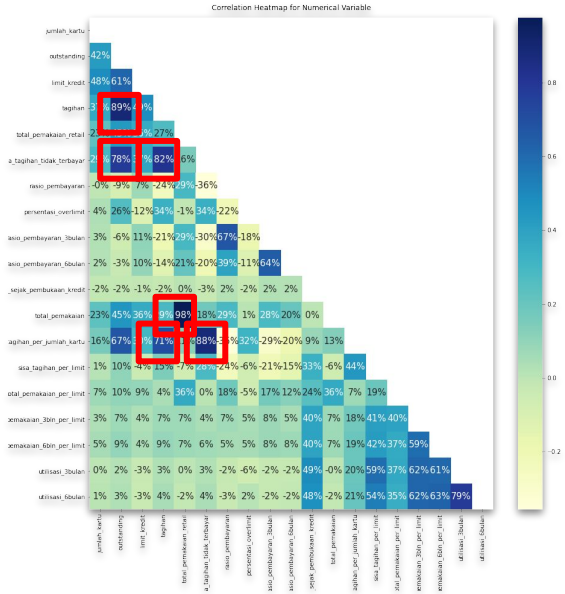
Exploratory Data Analysis

Outliers are imputed



1.5 IQR threshold is used as an imputation values

Highly correlated variables are dropped



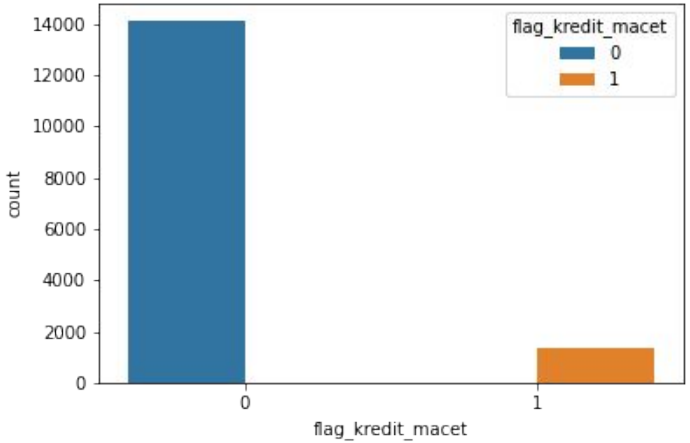
Dropped variables : 1) total_pemakaian_retail,
2) sisa_tagihan_tidak_terbayar, 3) outstanding = correlated with tagihan

Exploratory Data Analysis (II)

Insights

- Default customers tend to have higher limit
- Non Default customers have lower bill and higher Payment ratio
- Defaults customers are 2x more likely to overlimit

Parameters	Non Default	Default
jumlah_kartu	2.44	2.49
limit_kredit	Rp16,385,520	Rp17,542,310
tagihan	Rp5,122,565	Rp6,788,009
rasio_pembayaran (%)	52.9	15.5
persentasi_overlimit (%)	0.6	1.2



Target variable is imbalance

- Model risks ignoring insignificant default data
- Oversampling the default train data is done before modelling (after splitting)
- Synthetic Minority Oversampling Technique (SMOTE) is used

Modelling

1. **Split Data**
   - 80% trainset 20% test set
2. **Create Baseline model**
   - Logistic regression : accuracy 68%
3. **Compare with several model**
   - Comparison was made using KNN Classifier, SVC , Decision Tree, Random forest models
   - Random forest generate the best result : 89% Accuracy

	precision	recall	f1-score	support
0	0.94	0.94	0.94	2827
1	0.37	0.35	0.36	272
accuracy			0.89	3099
macro avg	0.65	0.65	0.65	3099
weighted avg	0.89	0.89	0.89	3099

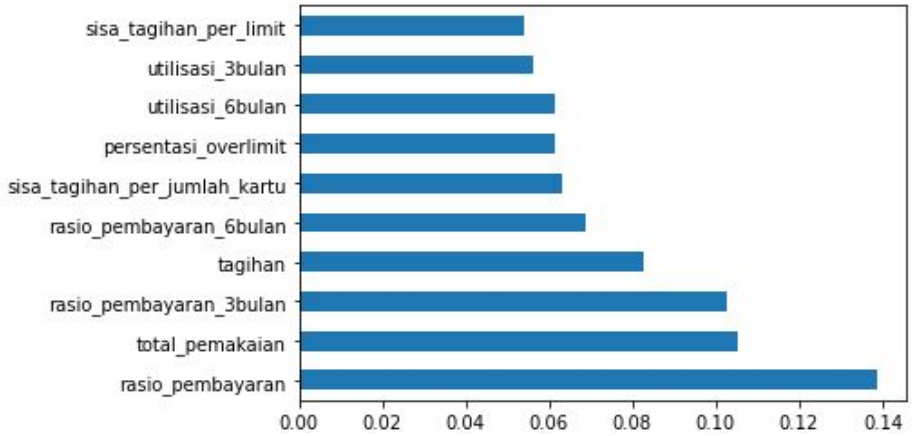
Cross Validation 3-Fold score result :
[0.90249403 0.94958875 0.94255771]

Classification Report

Modelling (II)

4. Improve Model

- Hyperparameter tuning : change max depth, min samples, and split according to the tuning result
- Feature Importances : All features have at least 4% significances, no feature dropped



5. Final Evaluation

- Accuracy score final is 93% and is considered high performance

Conclusion and Recommendation

- Payment Ratio is the most important parameter on predicting default
- Random forest model should be used to predict current debtors default probability
- Action items:
  - Create automated early warning notification using real-time data
  - Business team to prepare scenario on handling the potential default debtors
    - Restructuring
    - Communication with credit card holders
    - etc
  - By adding more personal attributes of credit card holders, another model could be generated : Credit card automatic approval
  - Improve ways on handling outliers and imbalance data to get the higher precision

THANK YOU

fadlymuham1@gmail.com |https://www.linkedin.com/in/muhammad-fadly-12ab10a2/ |https://github.com/fadlymuham1