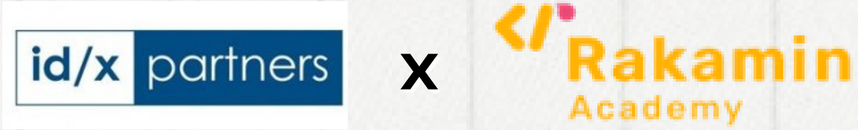
ID/X Partners Data Scientist Virtual Internship Program



Presented by Radhimas Januar Rachman

Process Pipeline

Business Understanding

Loan company needed to identify Credit risk prediction to evaluate the likelihood of borrowers repaying their loans which is a critical aspect of credit risk evaluation.

Exploratory Data Analysis

EDA to explore historical borrower data, including information on income, credit history, loan amounts, and other factors to gain insights of borrowers' ability to repay their loans.

Data Pre-Processing

Doing data pre-processing such as cleaning, transforming, and organizing the data. Handling missing values, outliers, and inconsistencies in the dataset to ensure data integrity.

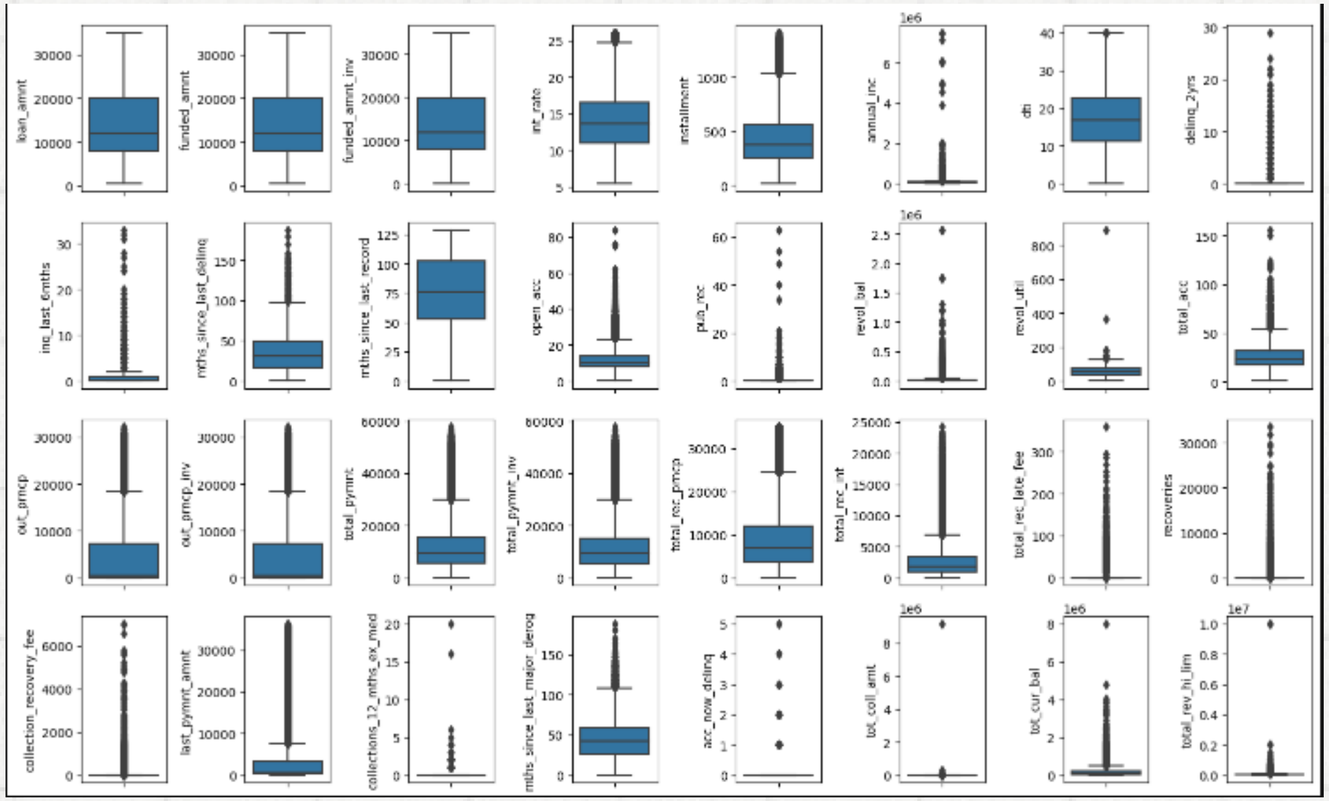
Modelling & Evaluation

Making a prediction model to accurately classify target feature. Model evaluation involves testing the models using different metrics and parameters.

Source Code Link

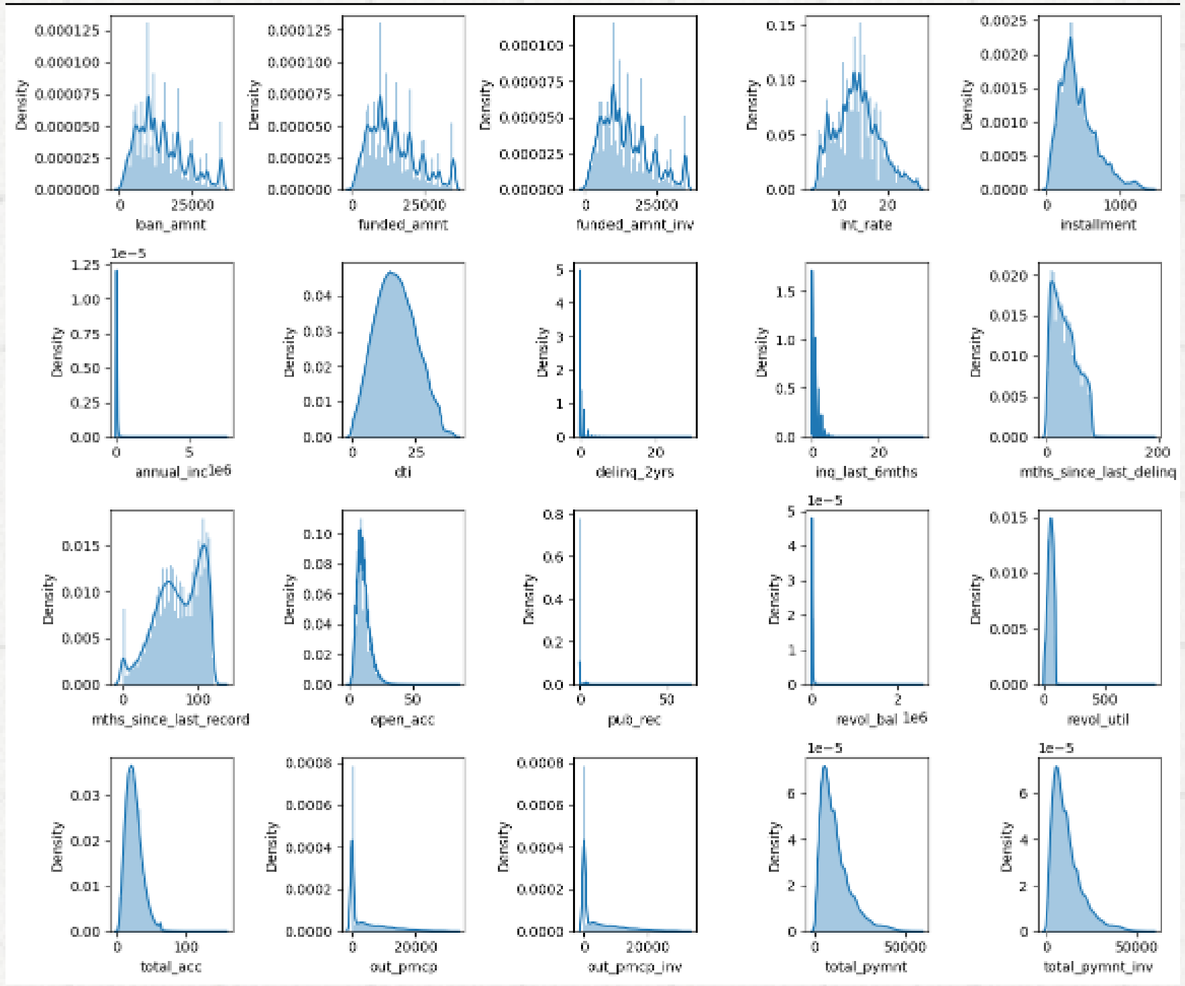
EDA (Exploratory Data Analysis)

- Most features generally have outlier values.
- Handling outliers is necessary to facilitate the subsequent processes.



EDA (Exploratory Data Analysis)

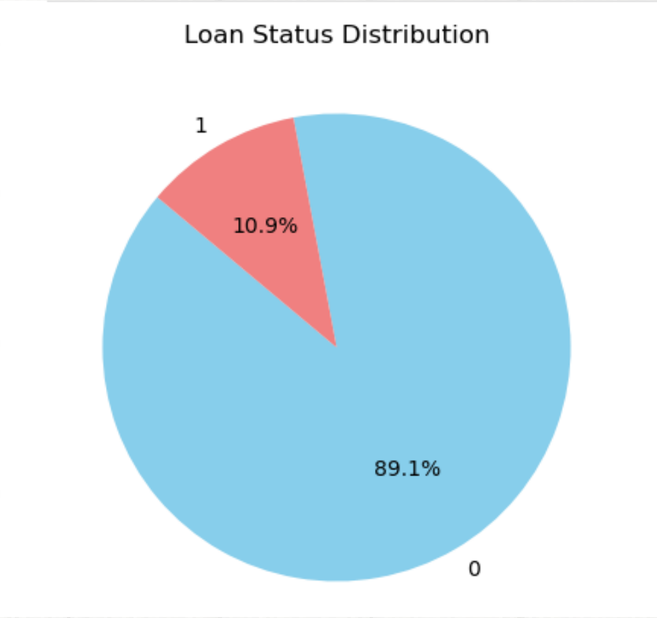
- The features are generally positively skewed, which may be caused by outliers and missing values.



EDA (Exploratory Data Analysis)

```
loan_status
Current                                                 224226
Fully Paid                                              184739
Charged Off                                              42475
Late (31-120 days)                                        6900
In Grace Period                                           3146
Does not meet the credit policy. Status:Fully Paid        1988
Late (16-30 days)                                         1218
Default                                                    832
Does not meet the credit policy. Status:Charged Off        761
Name: count, dtype: int64
```

- Data target which is ‘loan_status’ is transformed into ‘loan_status_label ’ with values good and bad loan status.
- Good loan status: Fully Paid, Current, In Grace Period, Does not meet the credit policy. Status:Fully Paid
- Bad loan status: Default, Charged off, Late (16-120 Days), Does not meet the credit policy. Status:Charged Off : status pinjaman Charged Off



Data Preprocessing

Missing Values

Features with missing value that have missing percentage above 75% will be dropped

Feature Engineering

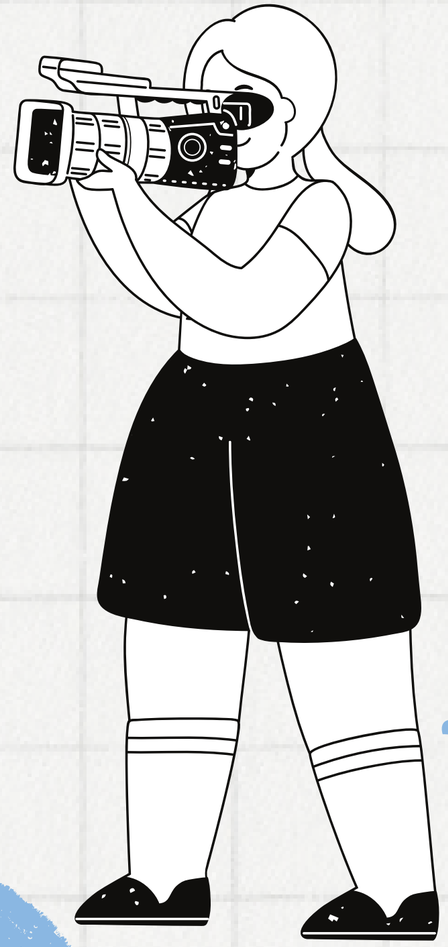
Changing features that have timestamp values into more suitable format for easier modelling purpose.

Feature Transformation

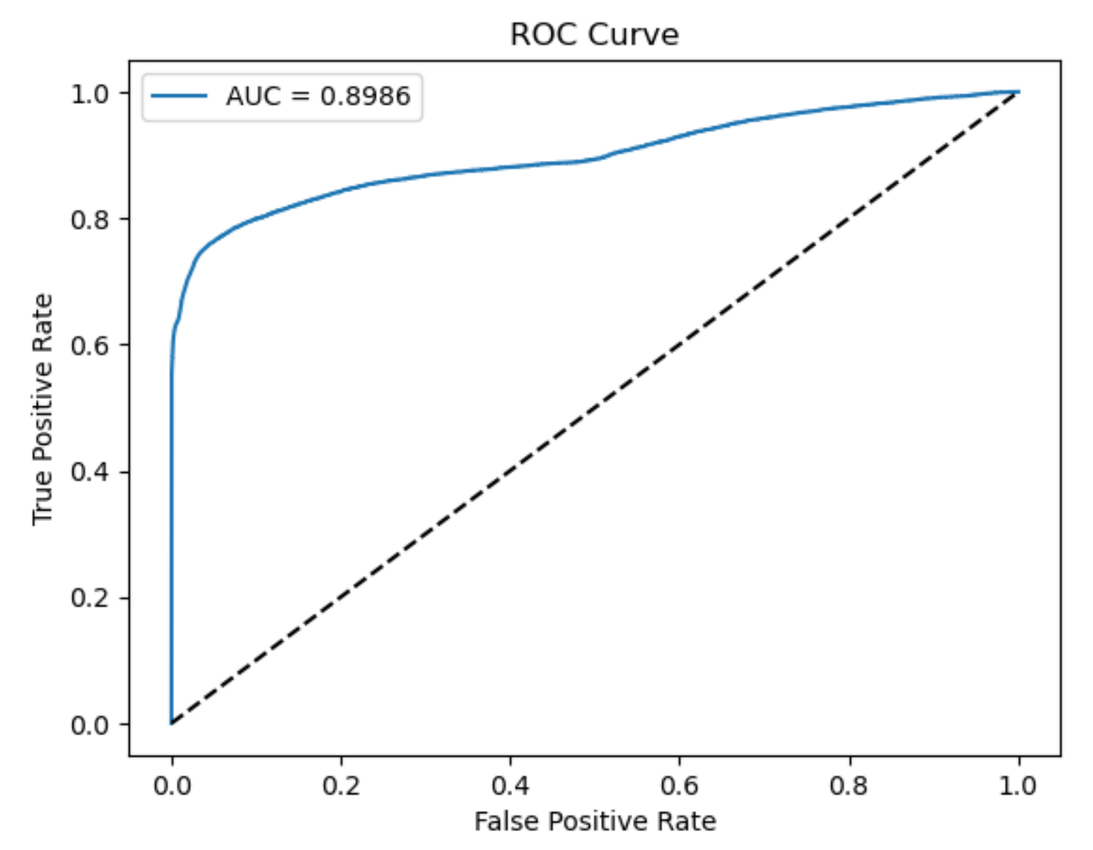
- Categorical features are encoded with one-hot encoding.
- Numerical features are standardized with standardization

Feature Selection

- Selecting one of two features that have 0.7 correlation score.
- Total selected features for data split train % test are 53 features and 466285 rows.

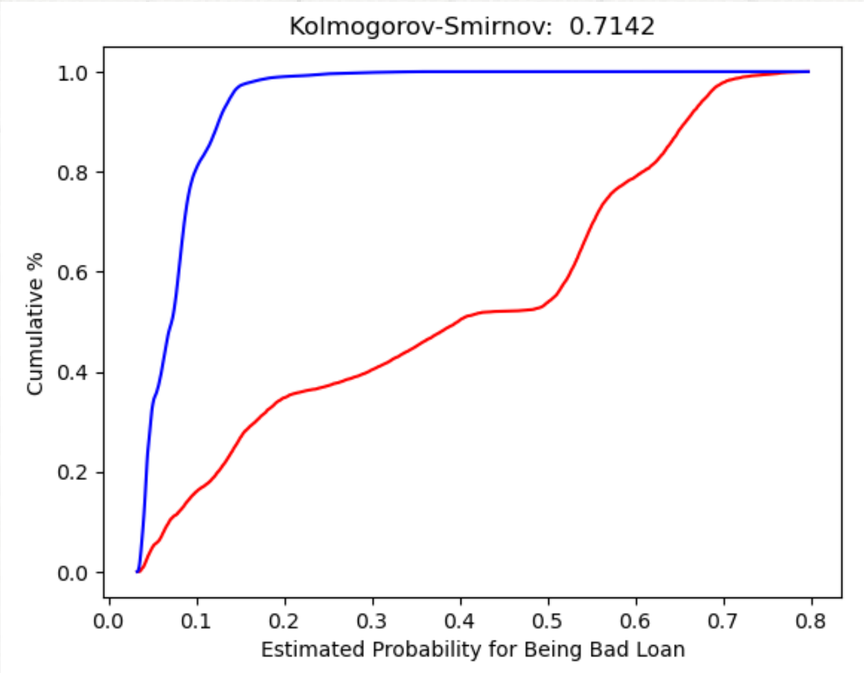


Modelling & Evaluation



- The model chosen is Random Forest for this modeling task, assessing model performance using standard credit risk metrics, including AUC (Area Under the ROC Curve) and KS (Kolmogorov-Smirnov). The obtained AUC score is 0.857.

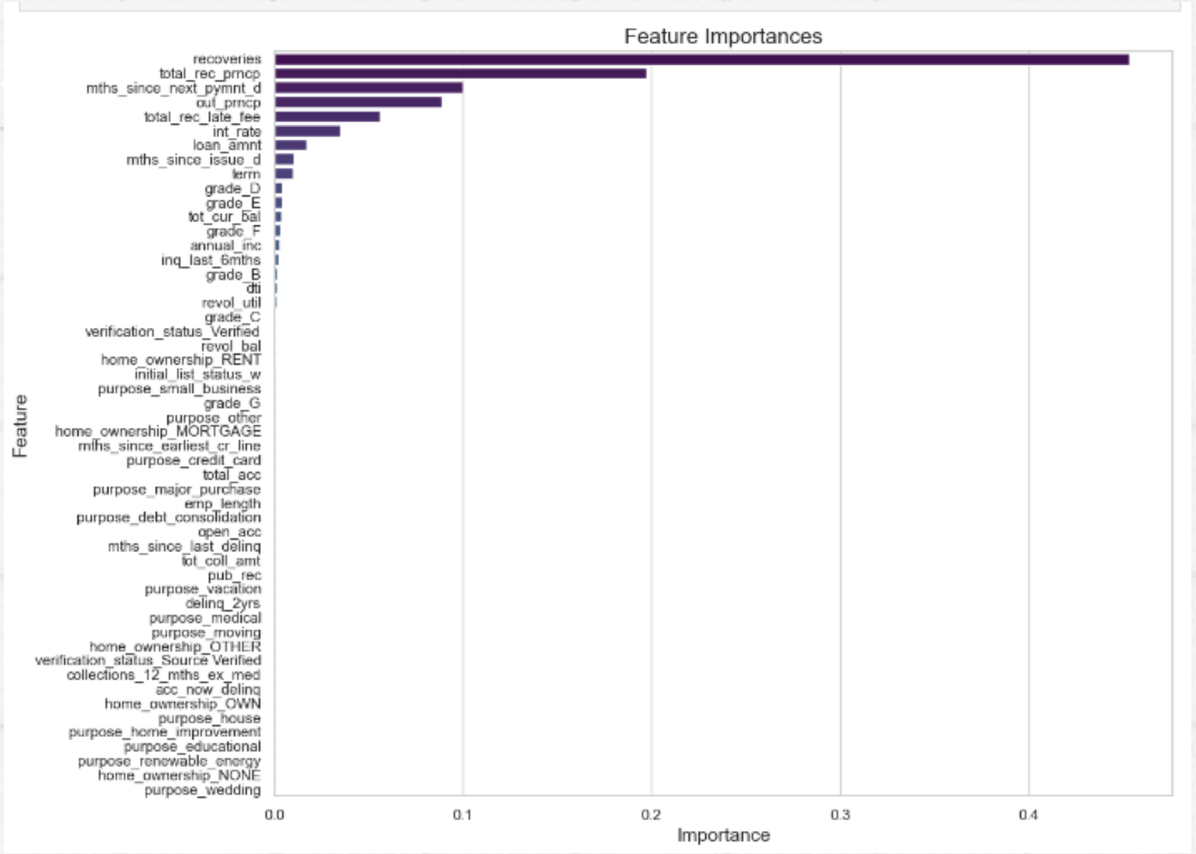
Modelling & Evaluation



	index	y_actual	y_pred_proba	Cumulative N Population	Cumulative N Bad	Cumulative N Good	Cumulative Perc Population	Cumulative Perc Bad	Cumulative Perc Good
0	322307	0	0.032922	1	0	1	0.000011	0.0	0.000012
1	245818	0	0.033074	2	0	2	0.000021	0.0	0.000024
2	299920	0	0.033120	3	0	3	0.000032	0.0	0.000036
3	362747	0	0.033175	4	0	4	0.000043	0.0	0.000048
4	378542	0	0.033201	5	0	5	0.000054	0.0	0.000060

As for the result of the Kolmogorov-Smirnov metric obtained is 0.56. The result indicates the model's ability to classify predicted credit applicants with high accuracy.

Feature Importance



- Based on the analysis of feature importance, we can observe that several features make significant contributions in determining the model predictions.
- The most important feature is "recoveries," indicating the amount of funds recovered from borrowers who failed to pay.
- The next significant features are "total_rec_prncp," which represents the total principal received by the lender, and "mths_since_next_pymnt_d," indicating the number of months until the next payment.

Conclusion

- Based on the evaluation metrics, the developed model demonstrates quite good performance for credit risk modeling. With an AUC value reaching 0.857 and a KS of 0.56 in the credit risk modeling world, AUC values above 0.7 and KS above 0.3 are often considered signs of good performance, indicating the model's ability to classify credit applicants with high accuracy. The model shows strong capability in distinguishing between credit applicants potentially having good loan and bad loan statuses.

Thank you