FINAL TASK DATA SCIENTIST
ID/X PARTNER – VIX RAKAMIN

LOAN DATA 2007-2014

BY: NOOR KHARISMAWAN AKBAR

id/x partners

Rakamin Academy

PROBLEM UNDERSTANDING

BACKGROUND

Dalam kasus prediksi risiko kredit, melibatkan identifikasi tujuan bisnis, yang dalam hal ini adalah untuk mengurangi risiko kredit yang tidak terbayar, meningkatkan pengambilan keputusan kredit, dan meminimalkan kerugian perusahaan. Untuk memahami kebutuhan dan perspektif pemangku kepentingan, diperlukan komunikasi dengan tim manajemen perusahaan, tim risiko, dan tim keuangan. Kriteria keberhasilan dapat didefinisikan sebagai peningkatan akurasi prediksi risiko kredit dan pengurangan risiko kredit yang tidak terbayar.

PROBLEM UNDERSTANDING

PROBLEM STATEMENT

Bagaimana memprediksi risiko kredit dari pelanggan yang mengajukan pinjaman?

GOAL & OBJECTIVE

Membuat suatu model yang dapat menentukan suatu pengguna mampu atau tidak mampu membayar kredit.

ANALYTICS APPROACH

Melihat permasalahan yang ada, kami akan membangun model machine learning karena kami perlu membangun prediksi lebih dari sekedar menggunakan pendekatan analisis inferensial dan/atau deskriptif.

MODELLING

Kami akan mencoba 5 model unsupervised machine learning regression: Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, K-Nearest Neighbors

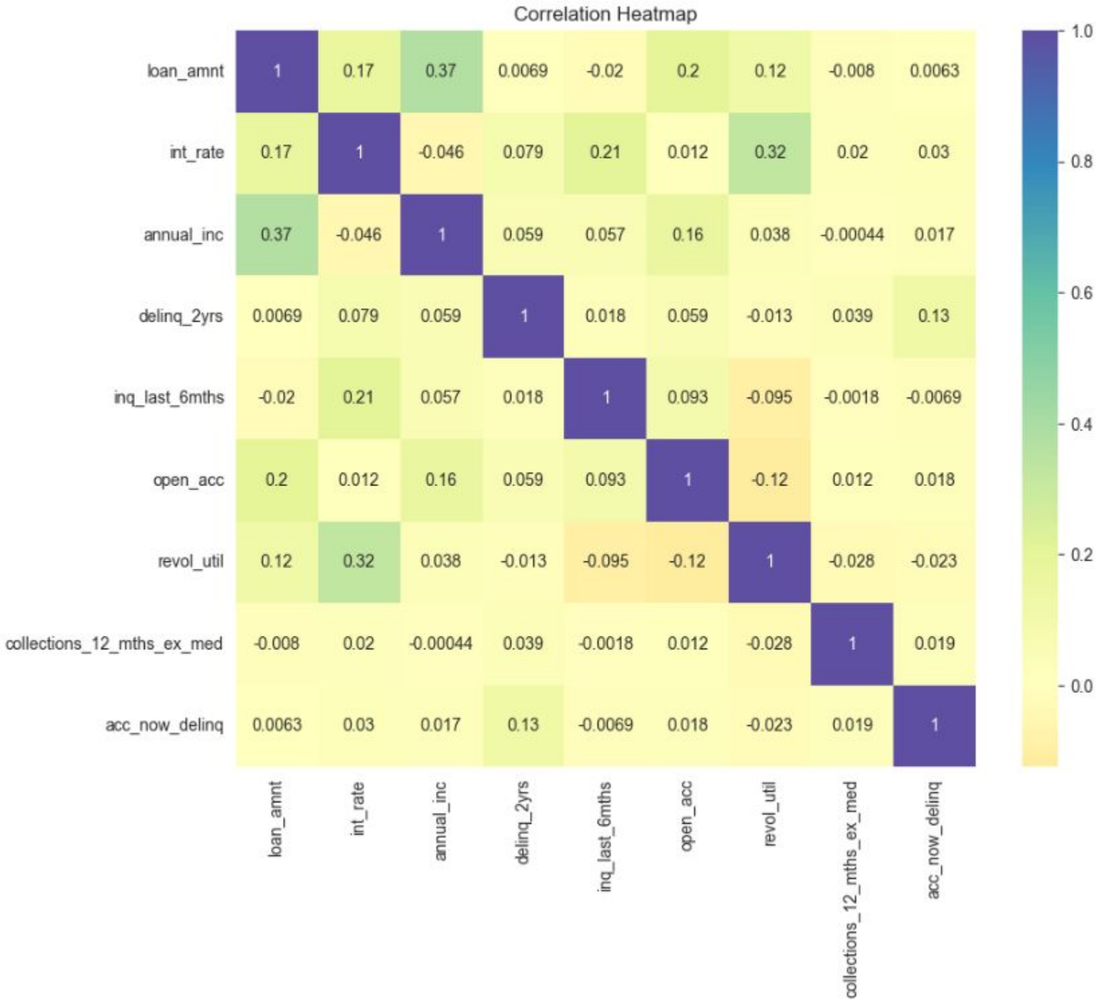
DATASET GENERAL INFO

Dataset Link:
https://www.kaggle.com/datasets/devanshi23/loan-data-2007-2014

Rows	Columns	Data Type
466.285	75	float64(46), int64(7), object(22)

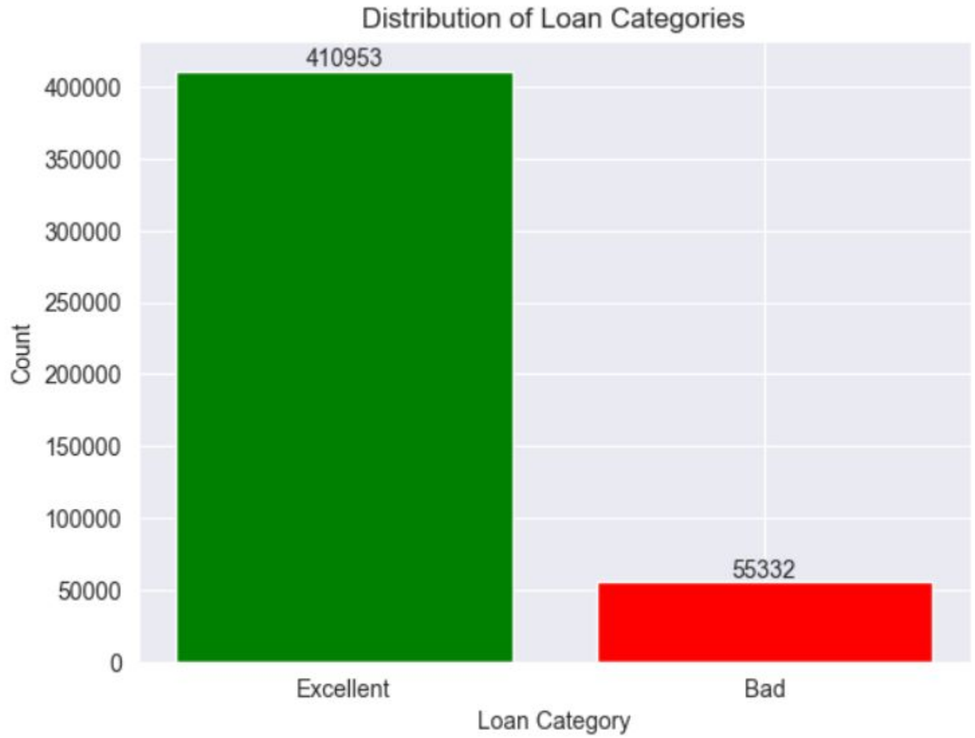
	Column	Dtype	null count	null perc.	unique count	unique sample
0	Unnamed: 0	int64	0	0.00	466285	[118572, 218560, 114701, 431706, 317301]
1	id	int64	0	0.00	466285	[8577135, 15411007, 2045036, 1172218, 16221044]
2	member_id	int64	0	0.00	466285	[7187822, 12439244, 10877614, 1265281, 17833906]
3	loan_amnt	int64	0	0.00	1352	[11975, 21550, 27200, 16625, 30300]
4	funded_amnt	int64	0	0.00	1354	[29375, 30775, 27150, 24425, 11575]
...	...	...	...	...	...	...
70	all_util	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
71	total_rev_hi_lim	float64	70276	15.07	14612	[262800.0, 12640.0, 32980.0, 22611.0, 26992.0]
72	inq_fi	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
73	total_cu_tl	float64	466285	100.00	0	[nan, nan, nan, nan, nan]
74	inq_last_12m	float64	466285	100.00	0	[nan, nan, nan, nan, nan]

DATA CORRELATION



DATA CLEANING

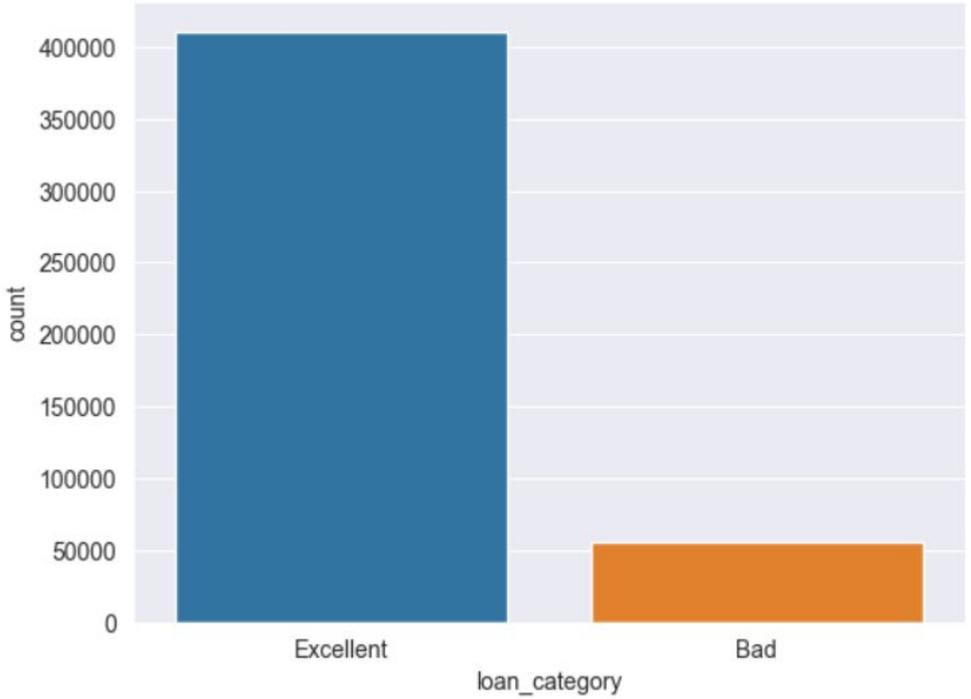
Making target column (Loan Categories) based on Loan Status.

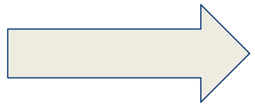


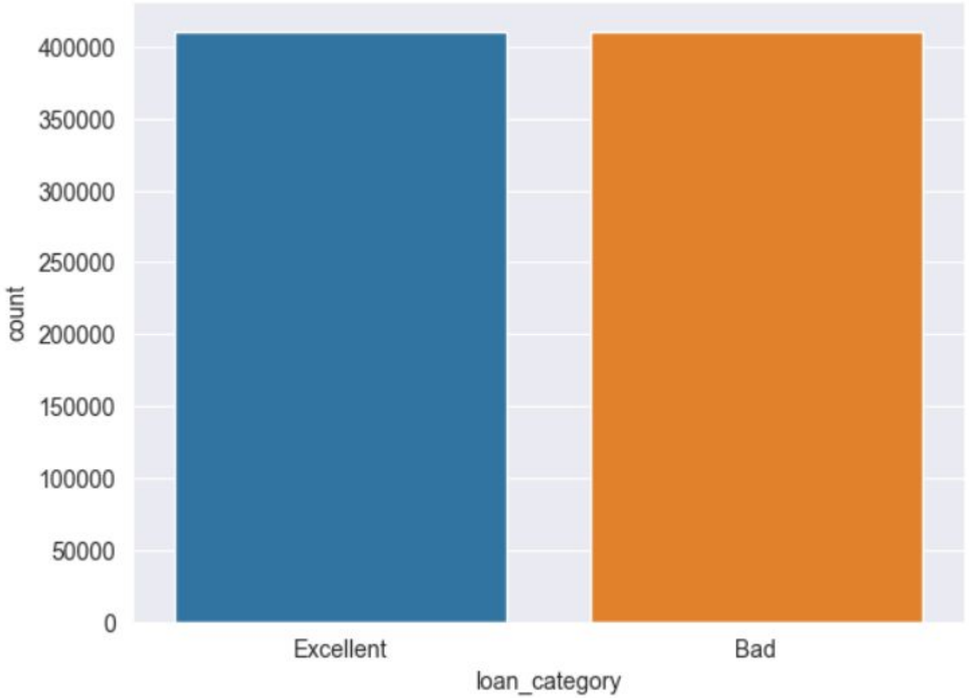
Loan Categories	Loan Status	Qty
Excellent	'Current', 'Fully Paid', 'Does not meet the credit policy. Status:Fully Paid'	410.953
Bad	'Charged Off', 'Late (31-120 days)', 'In Grace Period', 'Late (16-30 days)', 'Default', 'Does not meet the credit policy. Status:Charged Off'	55.332

DATA CLEANING

Oversampling for loan_category.



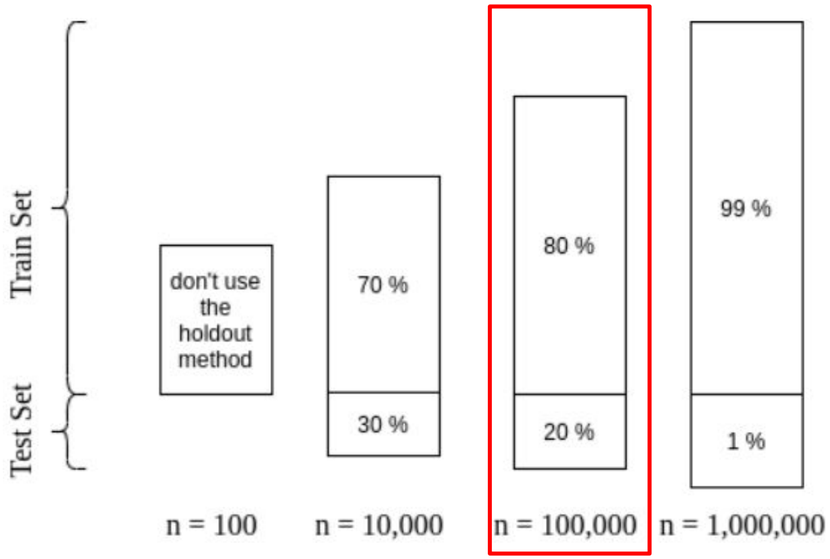




MACHINE LEARNING MODELLING

TRAIN-TEST SPLIT

We will use a ratio of **80:20**. This refers to the reference from **baeldung.com** for a dataset of size **n ~ 100,000+**.



NORMALIZATION

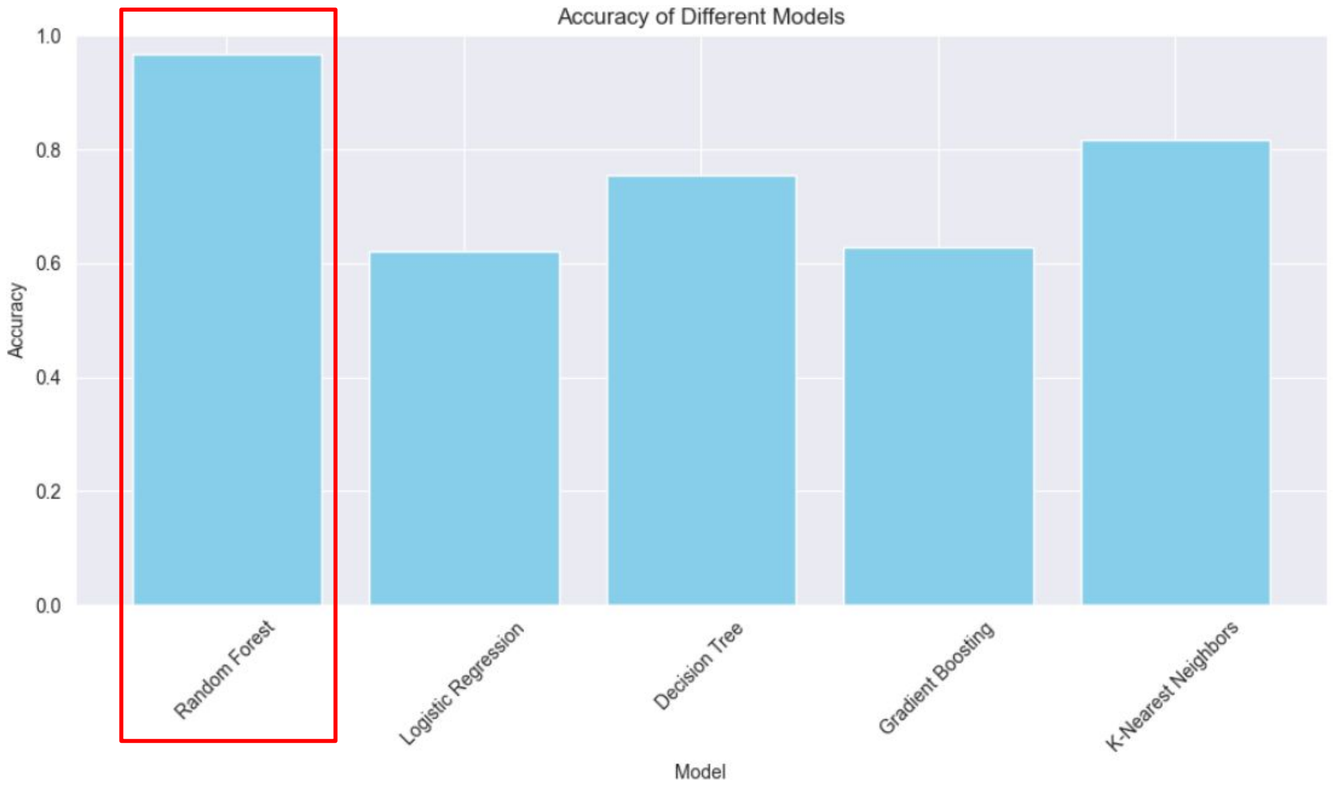
We will make all data **standardized** using **StandardScaler** (μ=0, σ=1) to avoid the ML model become bias.

EVALUATION METRICS

For comparing, we mainly use **accuracy.**

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

ML MODELING RESULT



Model	Accuracy
'Random Forest'	96,80%
'Logistic Regression',	62,29%
'Decision Tree',	75,50%
'Gradient Boosting',	62,86%
'K-Nearest Neighbors'	81,79%

The best model is **Random Forest** with accuracy **96,80%.**

THANK YOU

Contact me!

+6281227223150

akbar.noorkharismawan@gmail.com

http://www.linkedin.com/in/n-k-akbar

https://github.com/baramizzo58

https://public.tableau.com/app/profile/akbar2070