Loan Approval Prediction

Henry

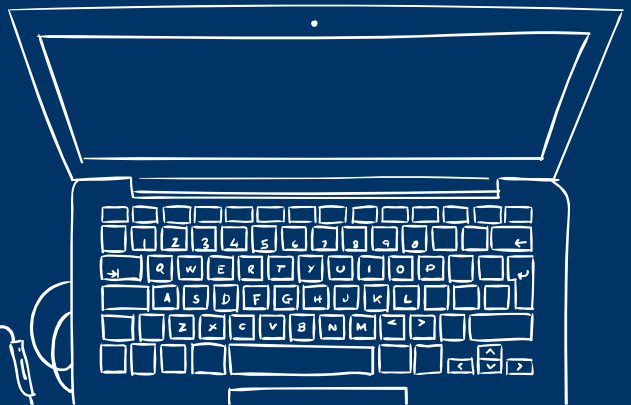
Purwadhika Startup and Coding School
Data Science and Machine Learning

Machine Learning

Metode analisis data yang mengotomatiskan pembuatan model analitik, sebuah cabang kecerdasan buatan yang didasarkan pada gagasan bahwa sistem dapat belajar dari data, mengidentifikasi pola, dan membuat keputusan dengan intervensi manusia yang minimal.

Bank Loan

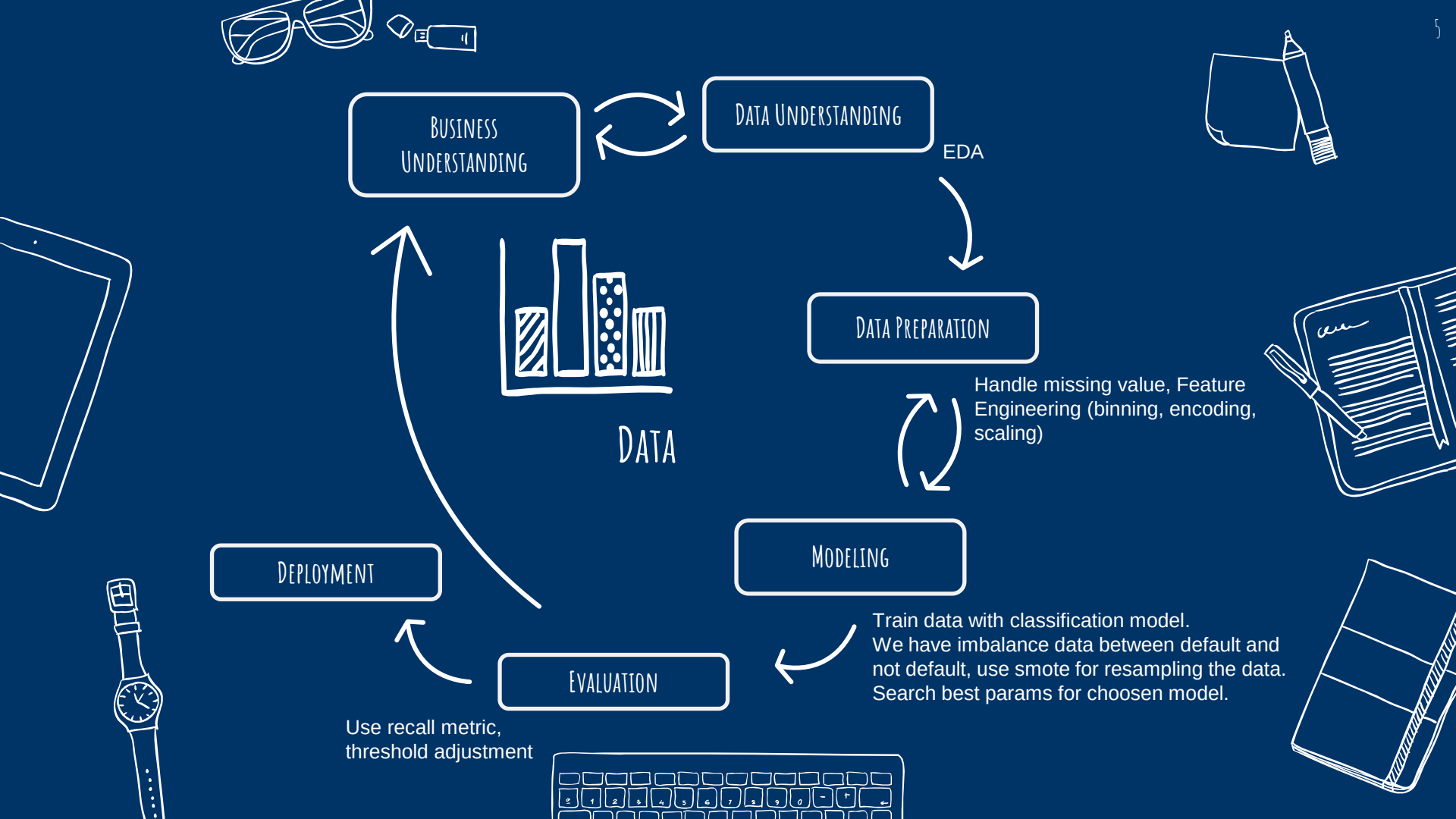
Penyediaan uang atau tagihan-tagihan berdasarkan perjanjian pinjam-meminjam antara bank dengan pihak lain. Pihak peminjam berkewajiban melunasi utangnya setelah jangka waktu yang telah ditetapkan dalam perjanjian.



Business Problems

Masih banyak Bank yang salah dalam memberikan pinjaman kepada nasabah. Banyak nasabah yang pada akhirnya gagal dalam melakukan pembayaran pinjaman. Hal tersebut diakibatkan oleh banyak faktor.

Tujuan dari machine learning ini adalah untuk meminimalisir resiko kesalahan pemberian pinjaman kepada nasabah dan dapat melakukkan prediksi secara real time.


Business Understanding
Data Understanding
EDA
Data Preparation
Handle missing value, Feature Engineering (binning, encoding, scaling)
Modeling
Train data with classification model.
We have imbalance data between default and not default, use smote for resampling the data.
Search best params for choosen model.
Evaluation
Use recall metric, threshold adjustment
Deployment
Data

DATASET

Source : Kaggle (Bank Loan Status Dataset | Kaggle)

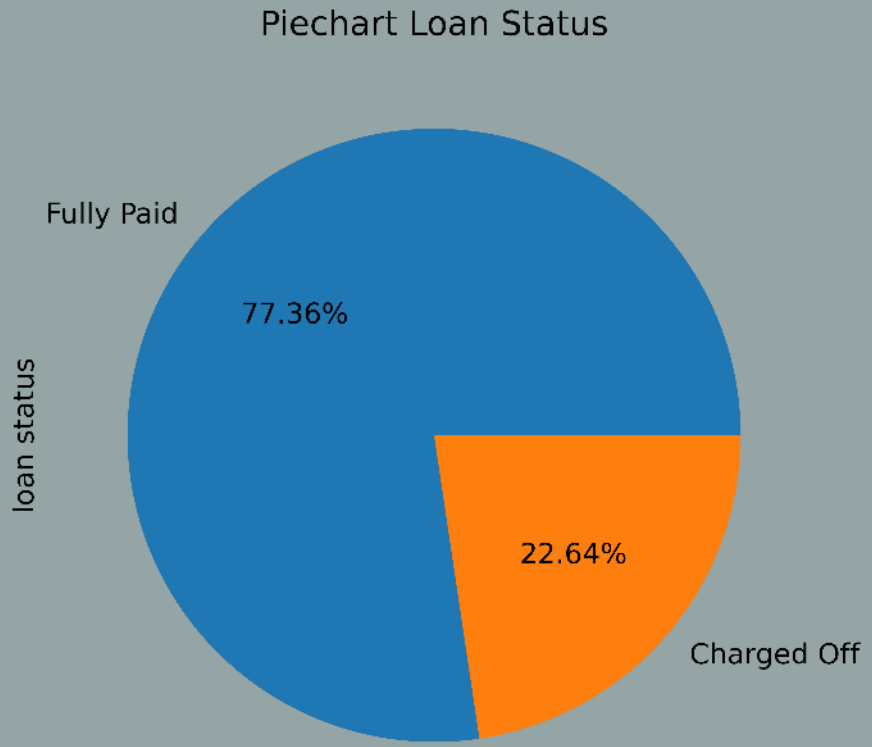
Shape of Dataset : (10514, 19)

	loan id	customer id	loan status	current loan amount	term	credit score	annual income	years in current job	home ownership	purpose	monthly debt	years of credit history	months since last delinquent	number of open accounts	number of credit problems	current credit balance	maximum open credit	bankruptcies	tax liens
0	14dd8831-6af5-400b-83ec-68e61888a048	981165ec-3274-42f5-a3b4-d104041a9ca9	Fully Paid	445412.0	Short Term	709.0	1167493.0	8 years	Home Mortgage	Home Improvements	5214.74	17.2	NaN	6.0	1.0	228190.0	416746.0	1.0	0.0
1	4771cc26-131a-45db-b5aa-537ea4ba5342	2de017a3-2e01-49cb-a581-08169e83be29	Fully Paid	262328.0	Short Term	NaN	NaN	10+ years	Home Mortgage	Debt Consolidation	33295.98	21.1	8.0	35.0	0.0	229976.0	850784.0	0.0	0.0
2	4eed4e6a-aa2f-4c91-8651-ce984ee8fb26	5efb2b2b-bf11-4dfd-a572-3761a2694725	Fully Paid	99999999.0	Short Term	741.0	2231892.0	8 years	Own Home	Debt Consolidation	29200.53	14.9	29.0	18.0	1.0	297996.0	750090.0	0.0	0.0
3	77598f7b-32e7-4e3b-a6e5-06ba0d98fe8a	e777faab-98ae-45af-9a86-7ce5b33b1011	Fully Paid	347666.0	Long Term	721.0	806949.0	3 years	Own Home	Debt Consolidation	8741.90	12.0	NaN	9.0	0.0	256329.0	386958.0	0.0	0.0
4	d4062e70-befa-4995-8643-a0de73938182	81536ad9-5ccf-4eb8-befb-47a4d608658e	Fully Paid	176220.0	Short Term	NaN	NaN	5 years	Rent	Debt Consolidation	20639.70	6.1	NaN	15.0	0.0	253460.0	427174.0	0.0	0.0

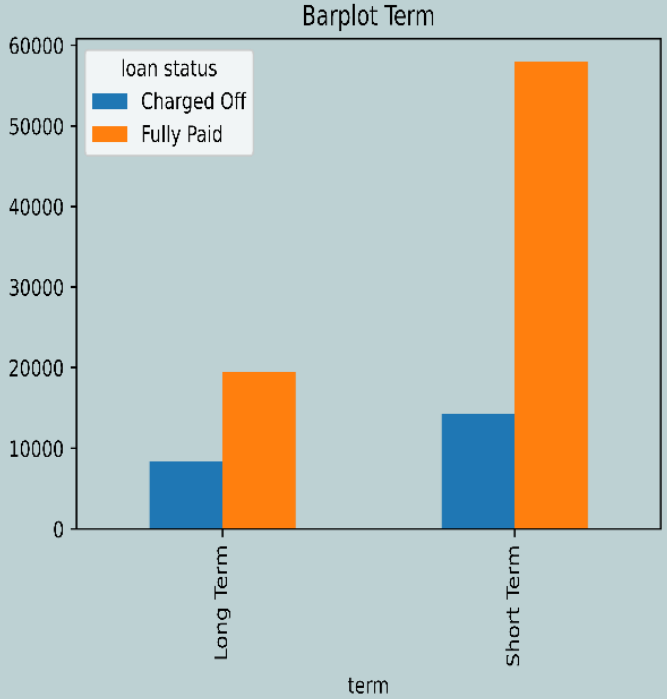
EDA

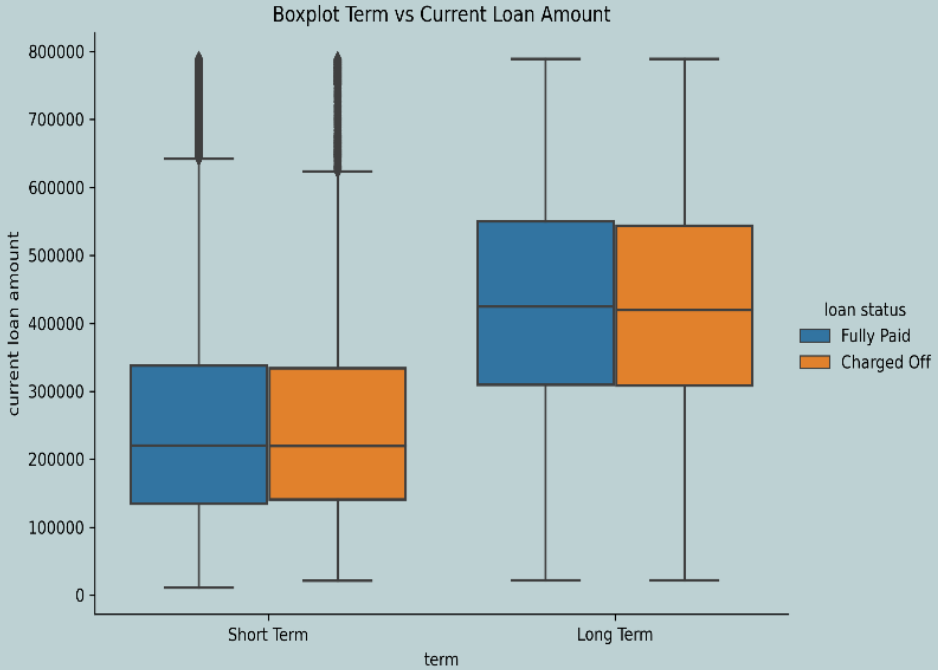
Pieplot Loan Status

Persentase nasabah yang berhasil bayar sebanyak 77.36% dibandingkan nasabah yang gagal bayar



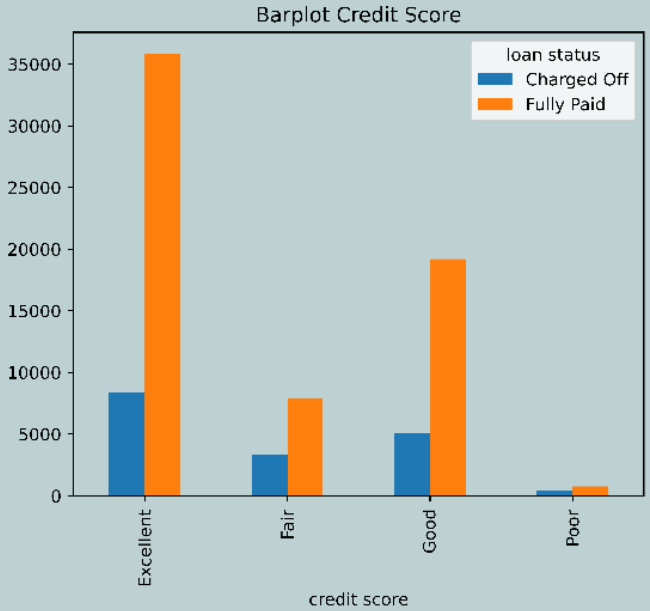
Grafik Loan Status berdasarkan term

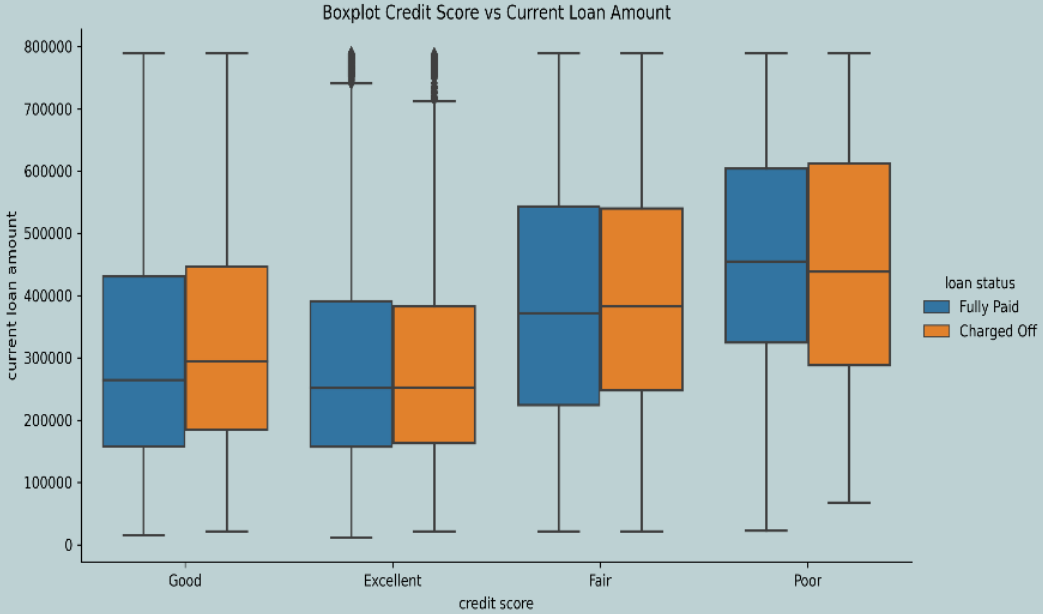




- Persentase nasabah yang gagal bayar pada long term lebih tinggi dibandingkan pada short term.
- Nasabah yang melakukan pinjaman dengan jangka waktu yang panjang memiliki pinjaman yang lebih besar dibanding dengan nasabah yang meminjam dengan jangka waktu yang pendek. Pihak bank harus lebih berhati – hati dalam meminjamkan nasabah dengan jangka waktu yang panjang karena memiliki potensi untuk gagal bayar yang tinggi dengan loan amount yang tinggi pula.

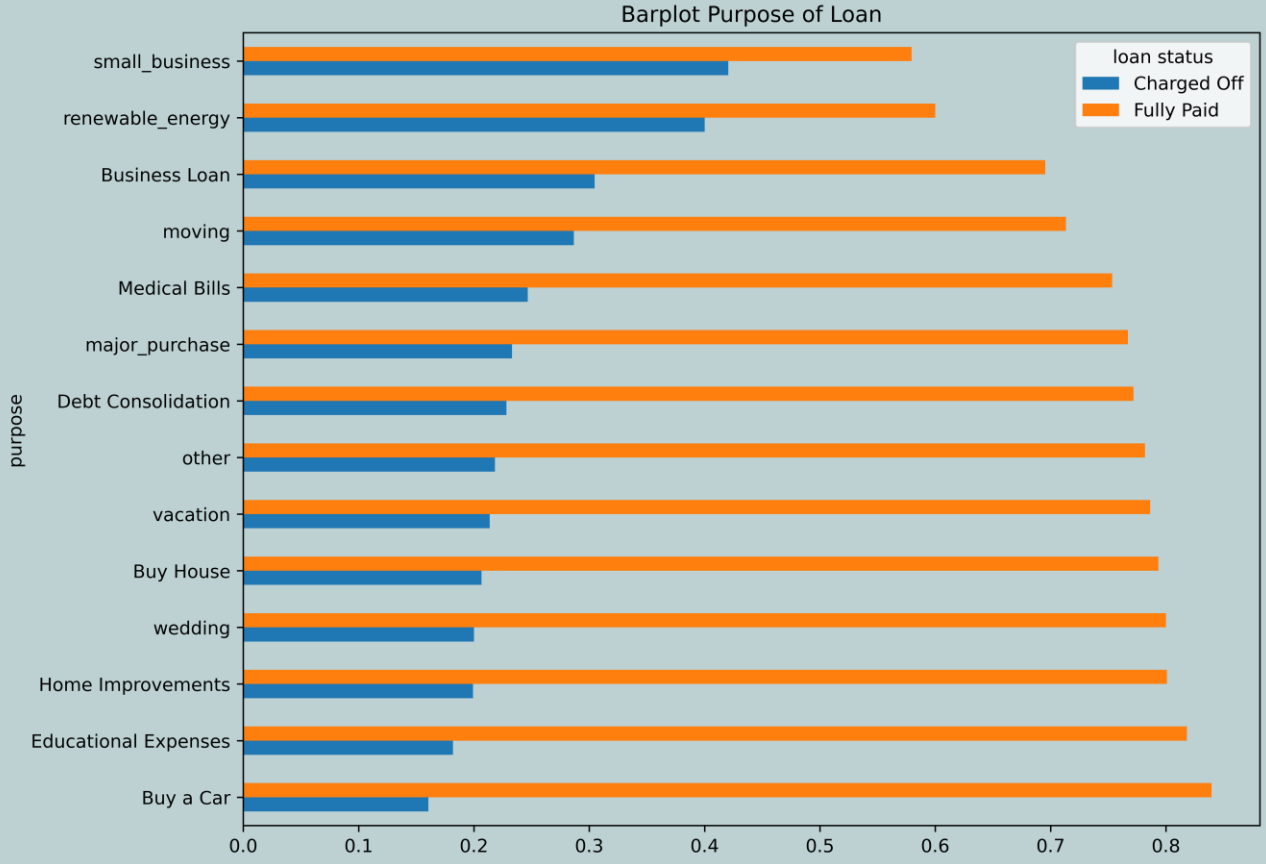
Grafik Loan Status berdasarkan credit score





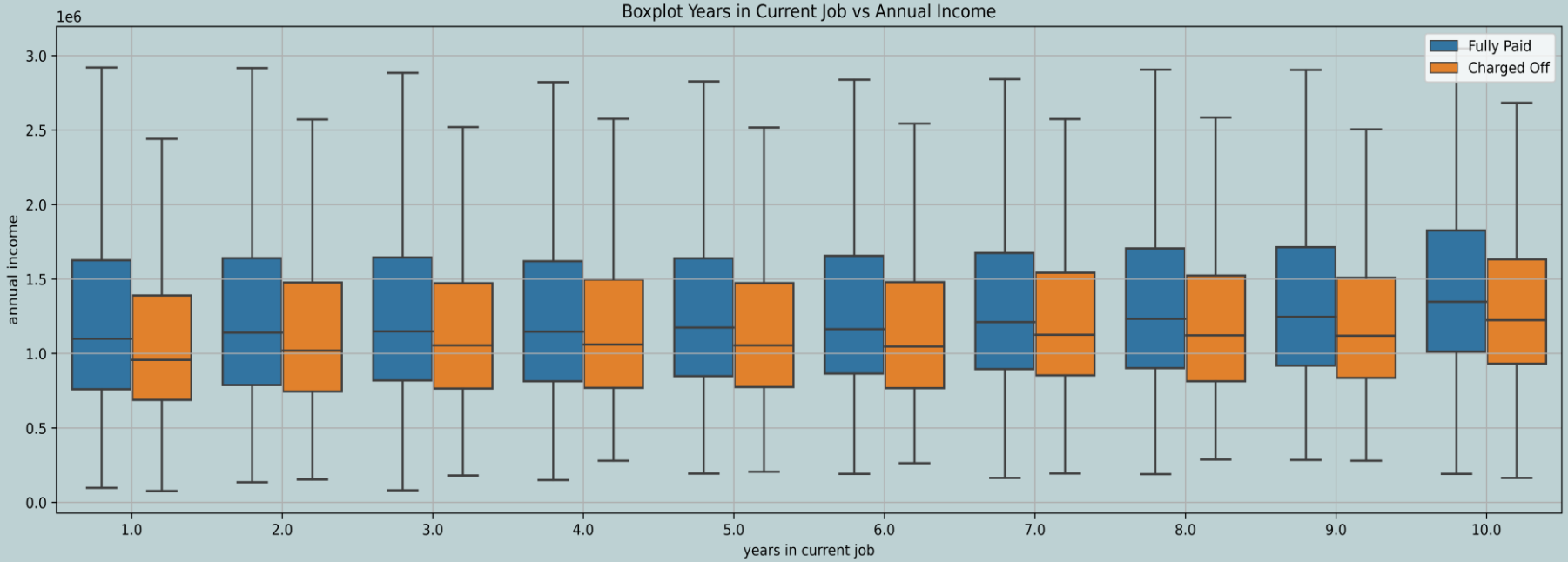
- Semakin buruk credit score nasabah, semakin tinggi persentase nasabah tersebut gagal bayar.
- Nasabah dengan credit score fair dan poor memiliki distribusi pinjaman yang lebih besar dibanding nasabah dengan credit score good dan excellent. Oleh sebab itu, pihak bank harus lebih teliti lagi dalam memberikan pinjaman kepada nasabah dengan credit score fair dan poor.

Grafik Loan Status berdasarkan credit score



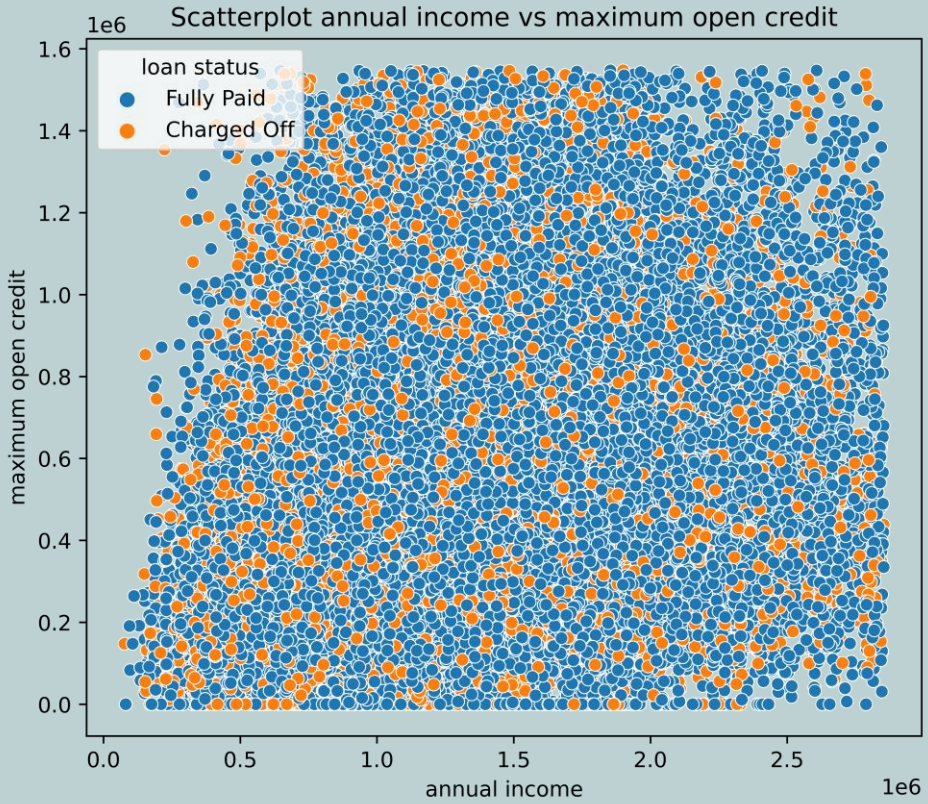
- Persentase tertinggi nasabah gagal bayar memiliki tujuan untuk bisnis, bukan untuk personal. Peminjam dengan tujuan personal seperti membeli mobil, melanjutkan sekolah, perbaikan rumah, dan sebagainya cenderung lebih aman untuk diberikan pinjaman.
- Small business menjadi tujuan nasabah meminjam yang memiliki persentase gagal bayar tertinggi. Bank seharusnya lebih teliti dalam memberikan pinjaman kepada small business karena ada kemungkinan bisnis tersebut mengalami kerugian sehingga nasabah menjadi gagal bayar pinjaman.

Grafik years in current job, annual income



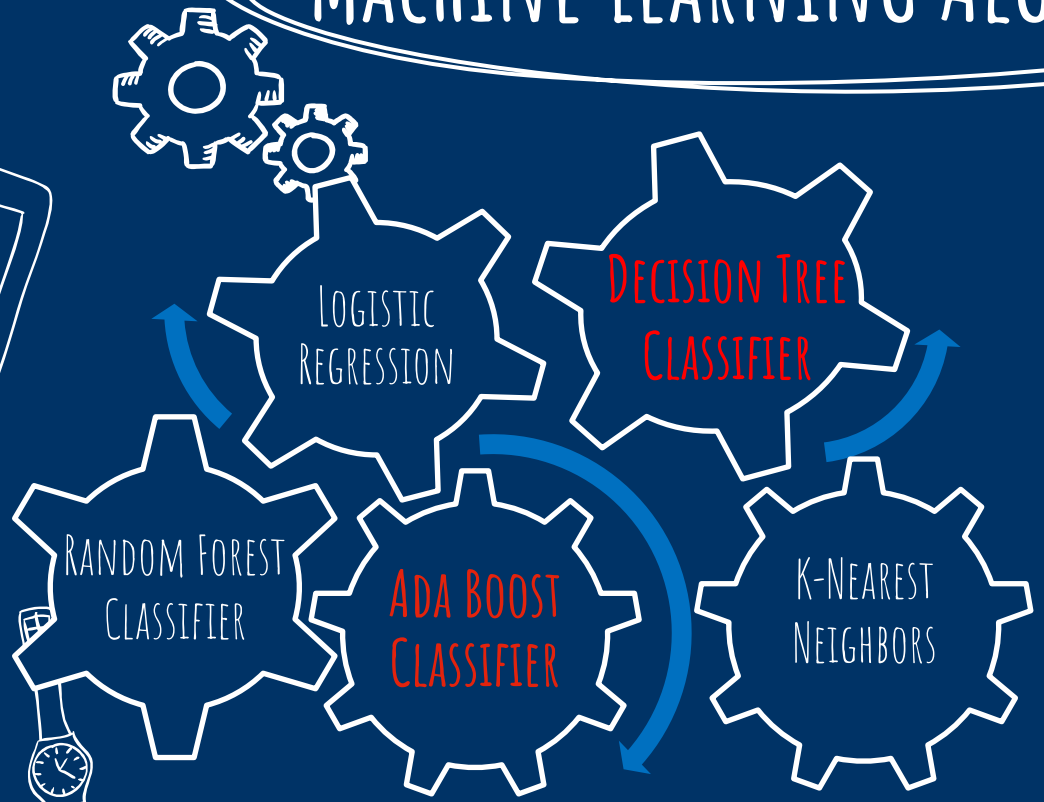
- Pengalaman nasabah dalam bekerja tidak berpengaruh signifikan dalam status peminjaman nasabah tersebut. Namun, semakin lama pengalaman nasabah dalam bekerja membuat annual income semakin besar. Hal tersebut berdampak pada semakin besarnya pinjaman yang dapat diajukan oleh nasabah itu.

Grafik annual income vs Max open credit



- Semakin besar pendapatan nasabah, plafon yang diberikan dalam pinjaman relatif makin besar. Hal tersebut tentu berpengaruh terhadapt current loan amount dan monthly debt dari nasabah tersebut.

Machine Learning Algorithm



Recall Score tiap model :

Logistic Regression Score : 0.5762
Random Forest Classifier Score : 0.6448
Decision Tree Classifier Score : 0.6680
K-Nearest Neighbors Score : 0.4754
AdaBoost Classifier Score : 0.6944

RECALL & PRECISION

RECALL

Recall dapat didefinisikan sebagai rasio dari *jumlah total contoh positif yang diklasifikasikan bernilai benar* dibagi dengan *jumlah total contoh positif*. **High Recall** menunjukkan kelas dikenali dengan baik (FN rendah).

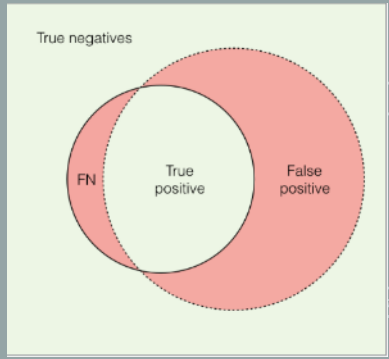
$$Recall = \frac{TP}{TP + FN}$$

PRECISION

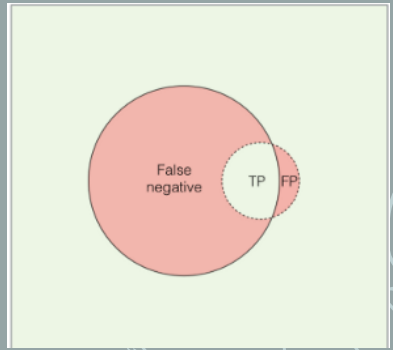
Precision merupakan pembagian dari *jumlah total contoh positif yang diklasifikasikan bernilai benar* dengan *jumlah total contoh positif yang diprediksi*. **High Precision** menunjukkan contoh berlabel positif memang positif (FP rendah).

$$Precission = \frac{TP}{TP + FP}$$

High Recall, Low Precision

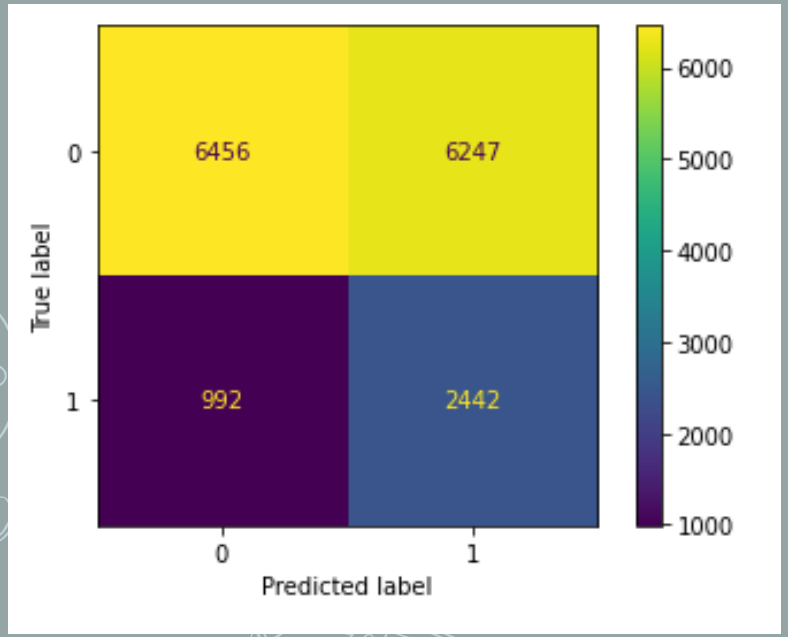


Low Recall, High Precision:

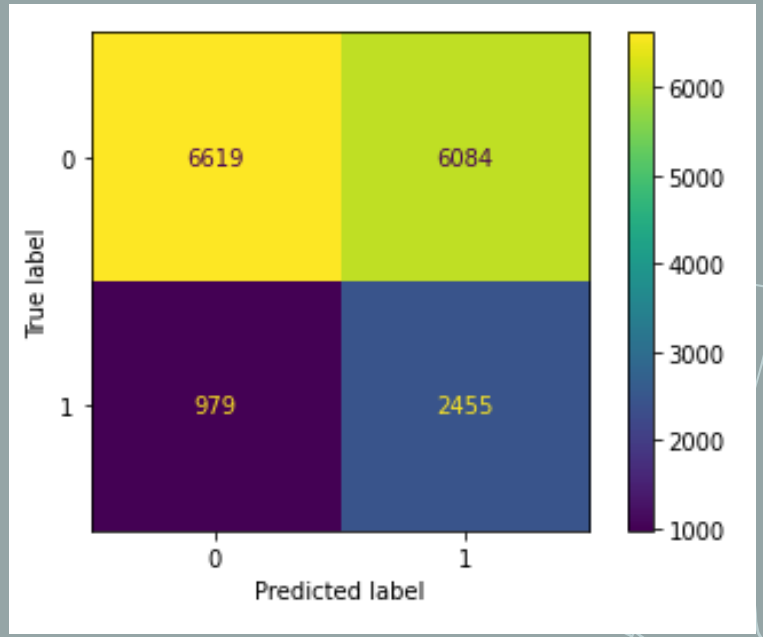


Confusion Matrix Between DTC dan ABC

Decision Tree Classifier

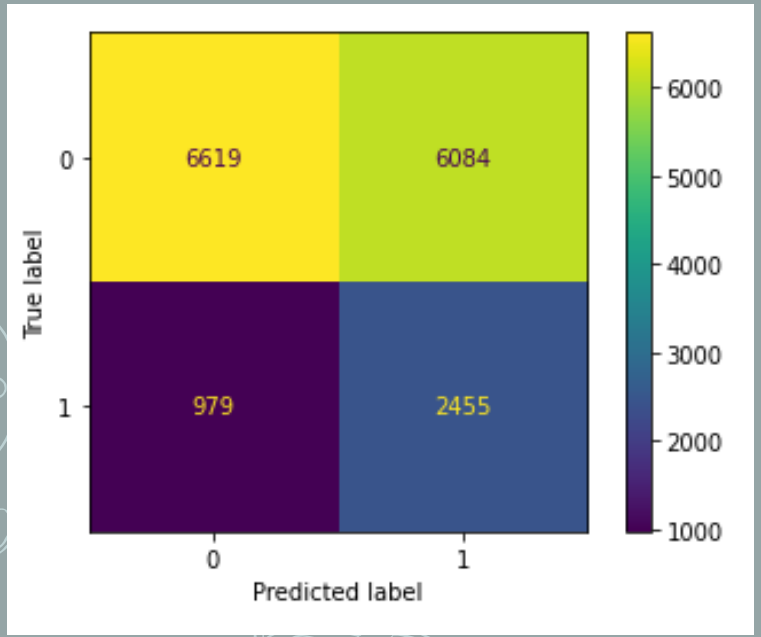


Ada Boost Classifier

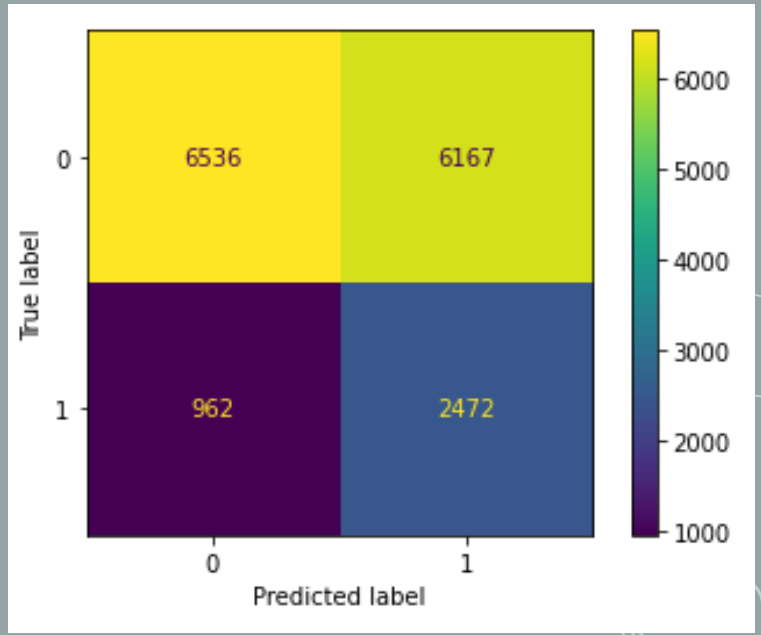


Confusion Matrix ABC after hyperparameter tuning

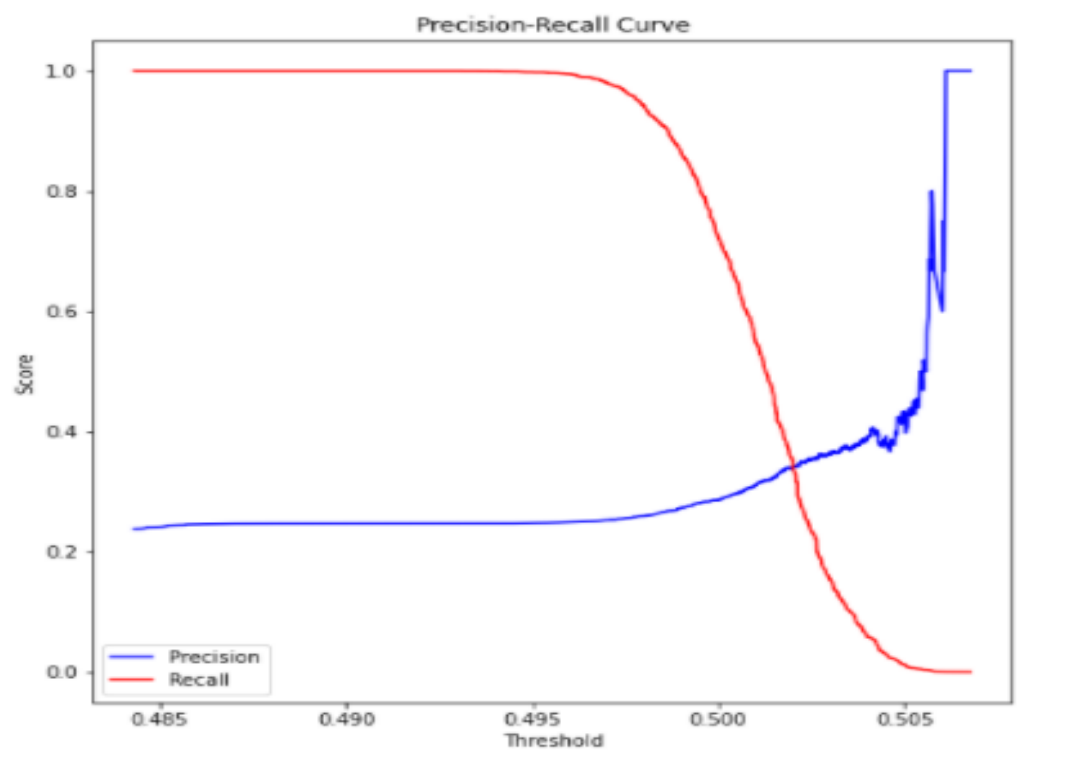
Before Hyperparameter Tuning



After Hyperparameter Tuning



THRESHOLD ADJUSTMENT BERDASARKAN PRECISION-RECALL CURVE



Threshold yang digunakan bernilai 0.499394525800627, hal ini berarti probability di bawah nilai tersebut akan diklasifikasikan sebagai 0 dan probability di atas nilai tersebut akan diklasifikasikan sebagai 1.

Confusion Matrix before and after threshold adjustment

Confusion Matrix

	prediction 0	prediction 1	total actual
0	6619	6084	12703
1	979	2455	3434
total prediction	7598	8539	16137

Confusion Matrix

	prediction 0	prediction 1	total actual
0	4766	7937	12703
1	459	2975	3434
total prediction	5225	10912	16137

Choosen Algorithm

ADA Boost Classifier

Confusion Matrix

	prediction 0	prediction 1	total actual
0	4766	7937	12703
1	459	2975	3434
total prediction	5225	10912	16137

Classification Report

	precision	recall	f1-score
0	0.91	0.38	0.53
1	0.27	0.87	0.41
accuracy			0.48
macro avg	0.59	0.62	0.47
weighted avg	0.78	0.48	0.51

Conclusion

Modeling Report

Ada Boost Classifier merupakan model yang akhirnya dipakai. Model tersebut memiliki nilai recall 0.87 dan precision 0.27

Dengan mengatur threshold, kita dapat mengurangi potensi untuk memberikan pinjaman kepada nasabah yang salah sebesar 4.05%