CREDIT LOAN STATUS PREDICTION

Wuri Nurma'rifah

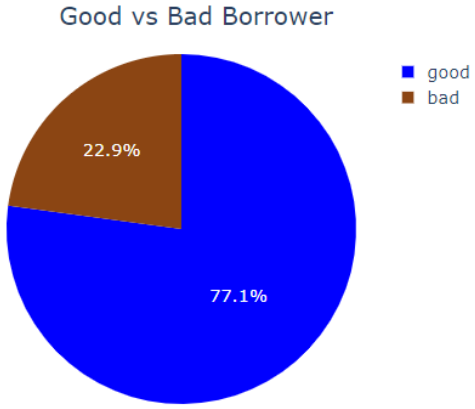
Pada perusahaan peminjam uang, penting untuk mengetahui jenis peminjam uang yang akan meminjam di perusahaan tersebut. Peminjam uang yang baik akan memberikan kelancaran dalam proses pinjam-meminjam uang. Berikut adalah model prediction yang akan memprediksi jenis peminjam uang, apakah peminjam uang tersebut baik dan memberikan kelancaran atau buruk

Data Preparation

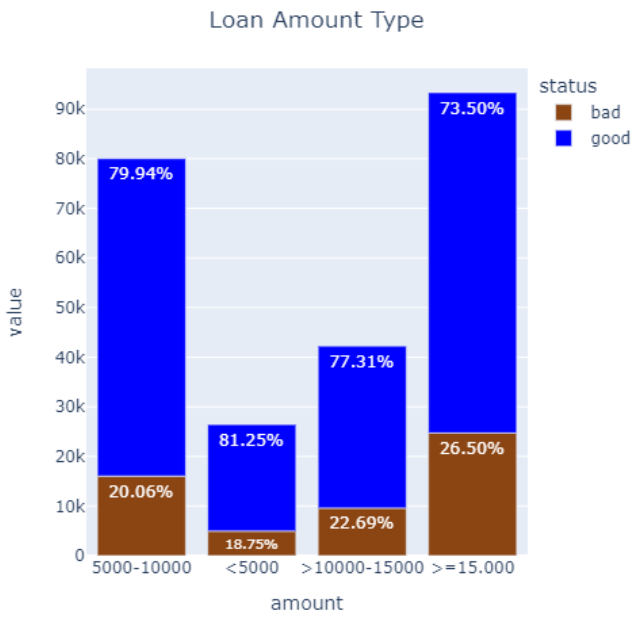
Awalnya dataset terdiri dari 466.285 baris dan 75 kolom. Namun tidak semua baris dan kolom kita gunakan dalam proses modeling. Berdasarkan kriteria jenis kolom, beberapa jenis kolom yang dipilih adalah loan_amount, term, int_rate, installment, grade, emp_length, home_ownership, annual_inc dan purpose pada variabel independent, dan loan_status pada variabel dependent. Untuk seleksi baris kita akan mengelompokkan tipe peminjam berdasarkan kolom loan_status dengan status good dan bad borrower saja. Untuk status peminjam current kita hapus.

Setelah proses seleksi baris dan kolom, kita juga melakukan proses imputasi missing value dengan nilai median untuk data numerik dan nilai modus pada data kategorik. Selain itu kita juga melakukan penanganan terhadap outlier dengan menggunakan nilai interquartile range dan proses encoding pada data bertipe object.

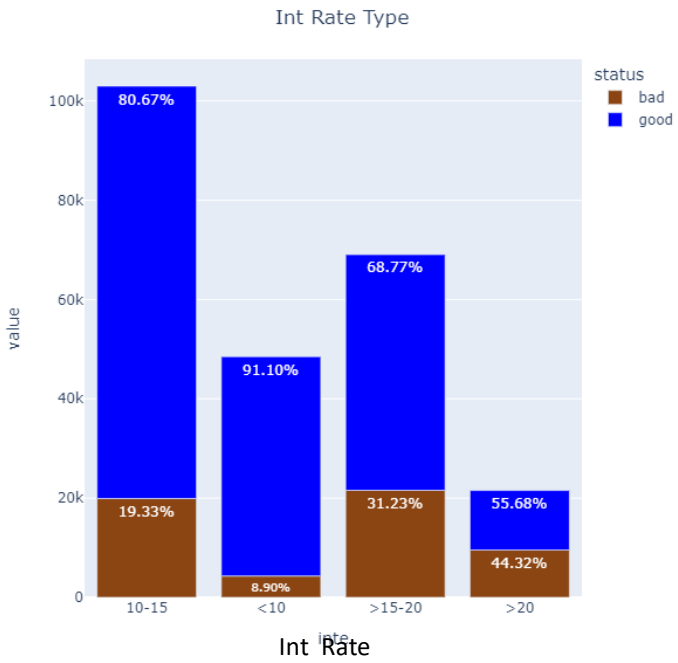
Visualization and Analysis



Berikut adalah jumlah prosentase peminjam dengan kategori good dan bad borrower. Terdapat 186.727 atau sekitar 77.1% good borrower dan 55.332 atau sekitar 22.9% bad borrower

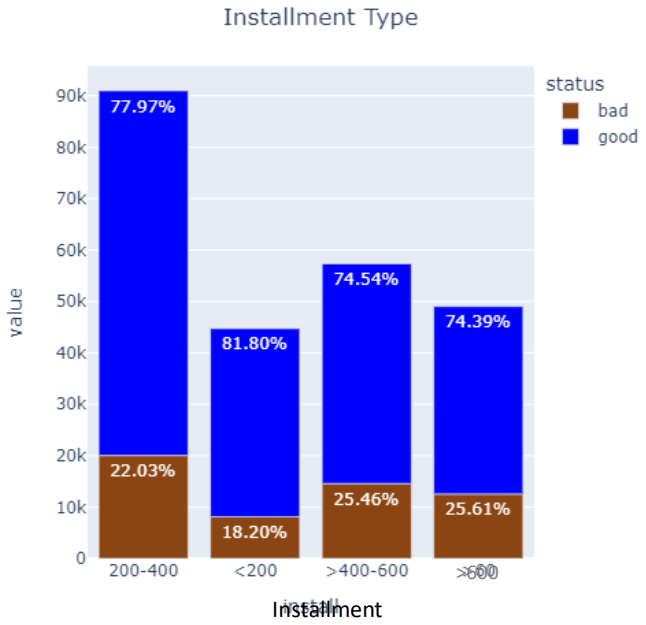


Untuk jumlah loan amount, rentang loan amount dengan proporsi bad borrower terbesar terdapat pada nilai >= 15.000. Sedangkan nilai loan amount dengan proporsi bad borrower terkecil terdapat pada rentang di bawah 5000



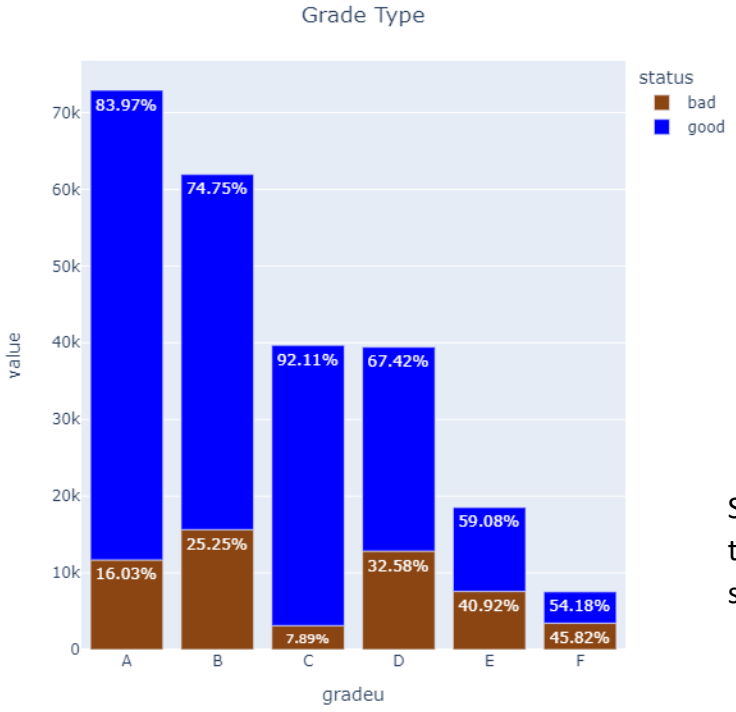
Int Rate

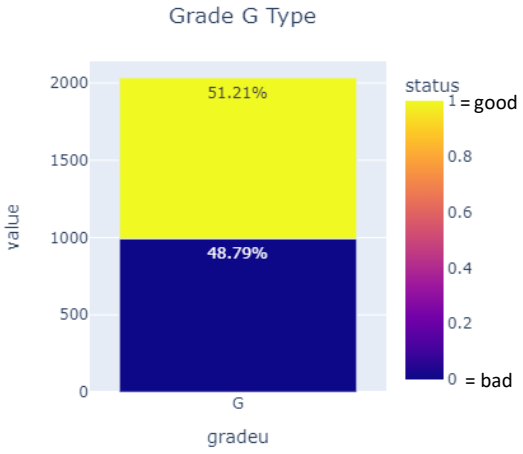
Pada kolom Int Rate, nilai INT Rate dengan jumlah proporsi bad borrower terbanyak terdapat pada nilai int rate >20. Sedangkan yang terendah terdapat pada int rate <10



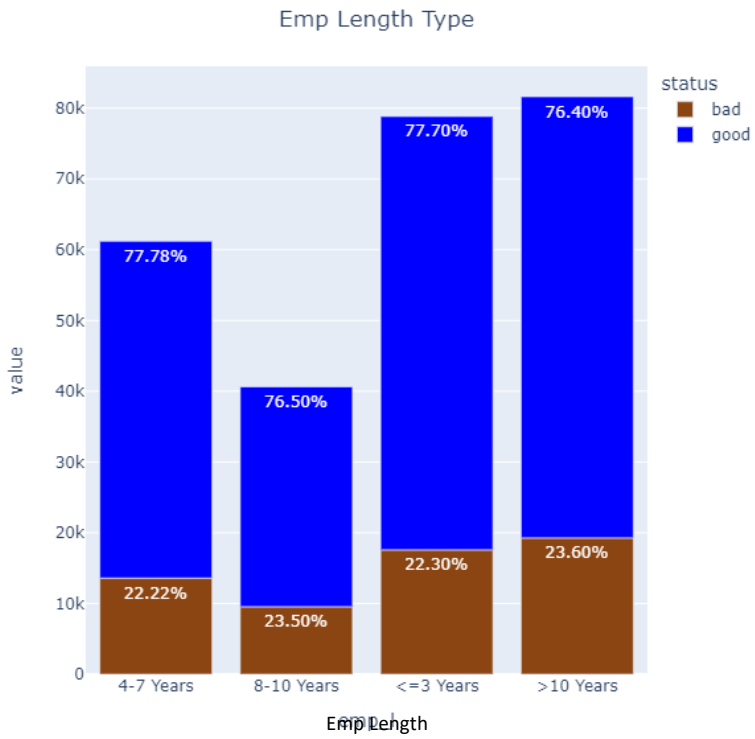
Installment

Untuk nilai installment, persentase installment dengan bad borower terbanyak terdapat pada nilai installment >600. Sedangkan yang terendah terdapat pada nilai <200 dengan jumlah antara keduanya tidak terlalu jauh berdasarkan grafik

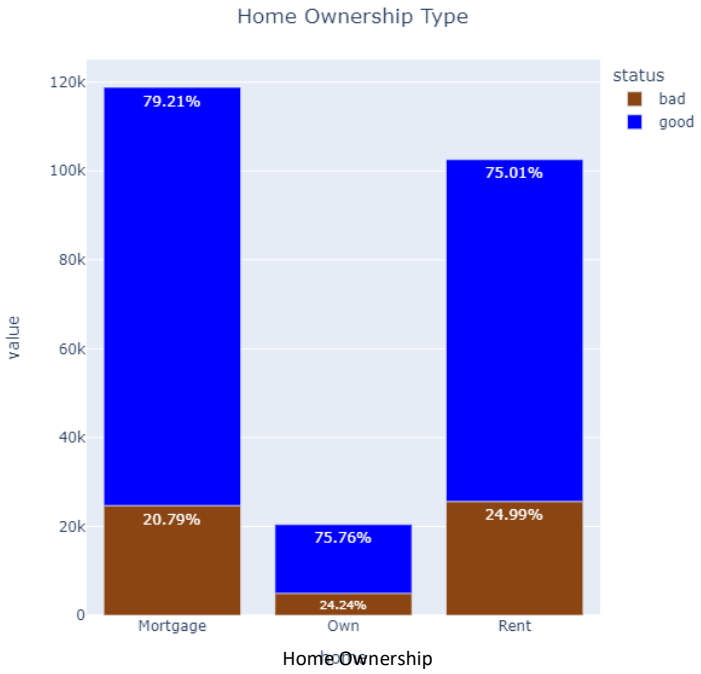




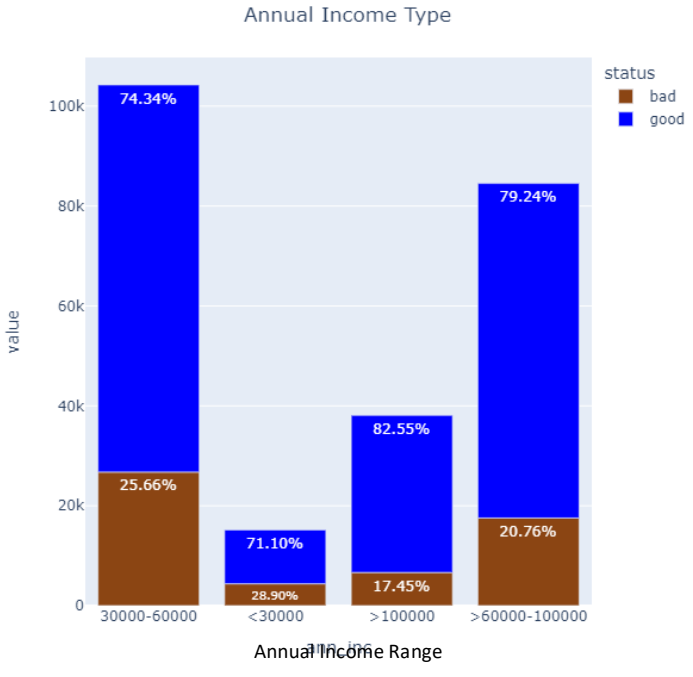
Sementara pada kolom grade, persentase bad borrower terbanyak terdapat pada grade G yaitu 48,79%, sedangkan yang terkecil terdapat pada grade C



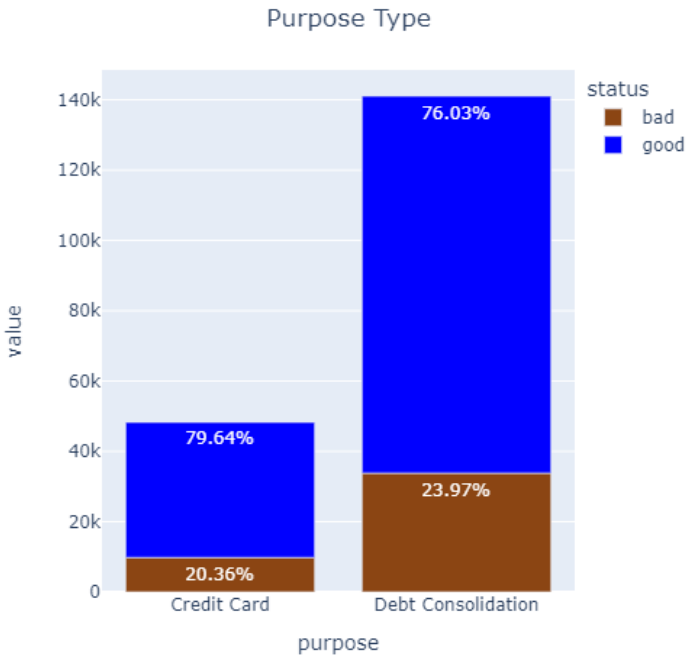
Untuk emp_length, persentase bad borrower tiap rentang tidak terlalu berbeda. Hanya yang tertinggi terdapat pada rentang emp_length >10 years

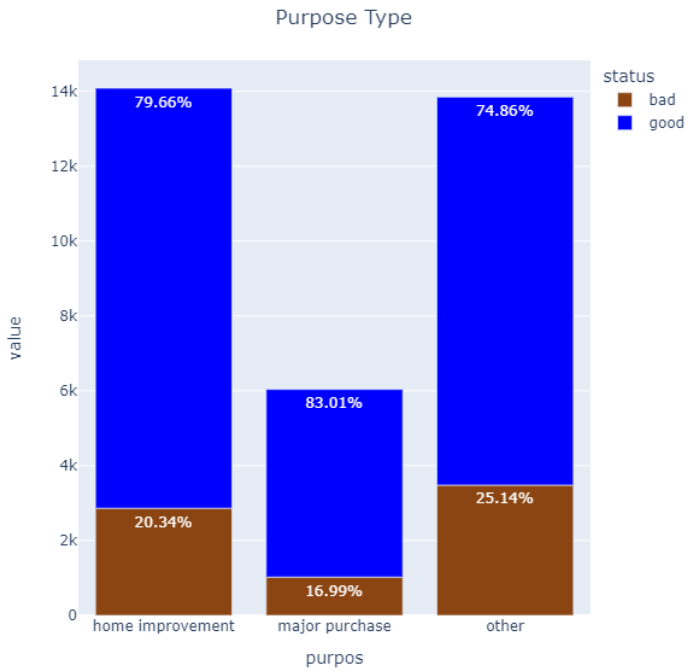


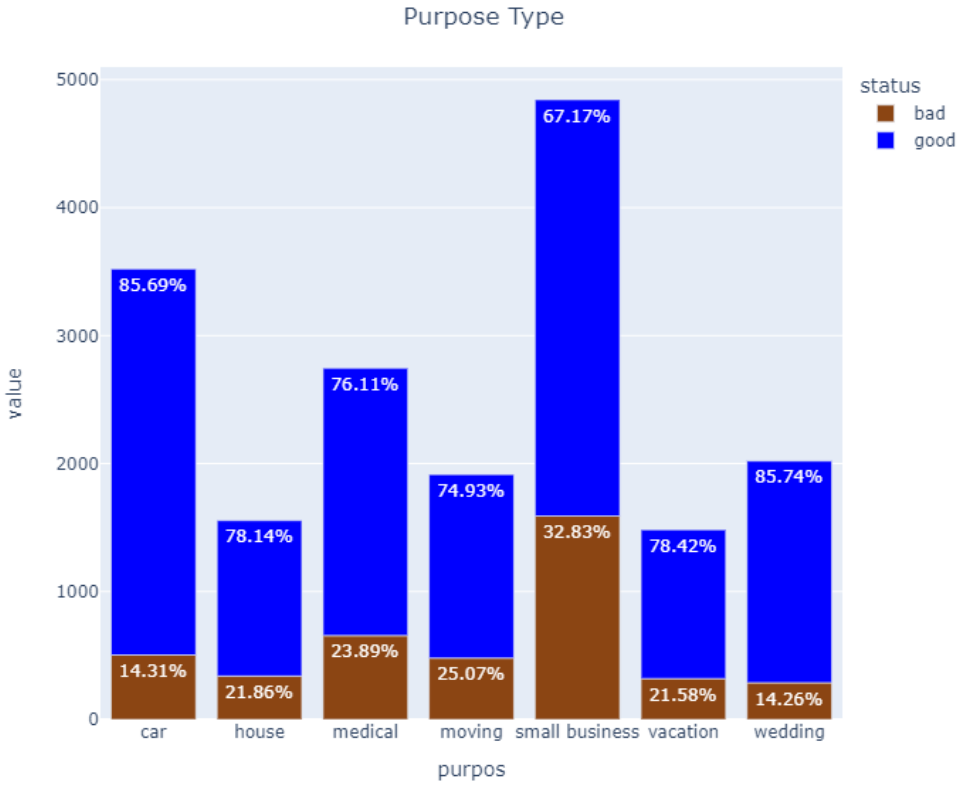
Tipe peminjam dengan status bad borrower terbanyak terdapat pada tipe peminjam dengan status kepemilikan rumah rent



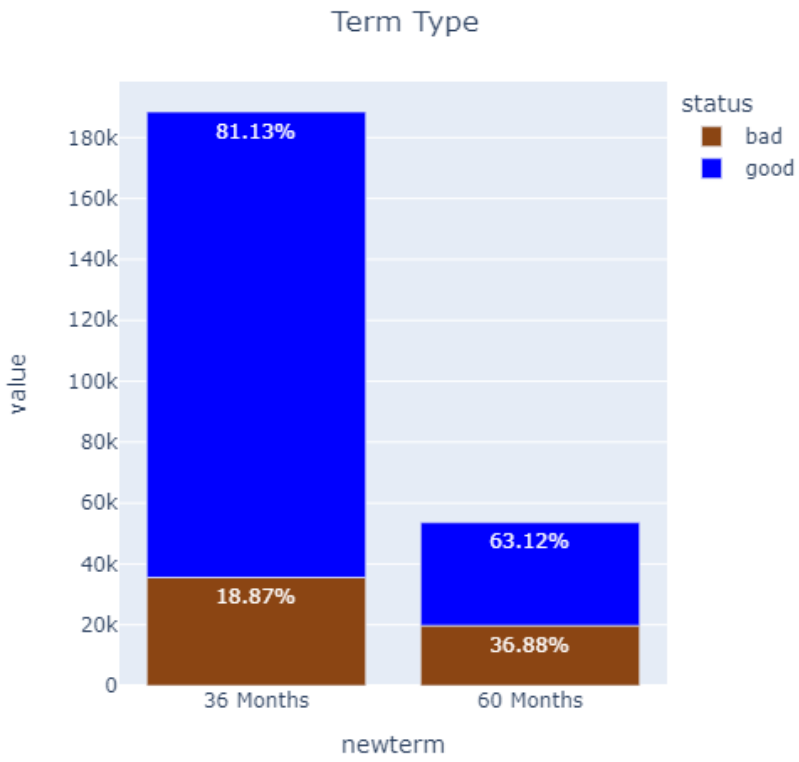
Tipe peminjam dengan persentase bad borrower terbanyak berdasarkan nilai annual income terdapat pada nilai annual income <30.000. sedangkan paling sedikit terdapat pada nilai annual income >100000







Untuk purpose peminjaman, tipe peminjam dengan purpose small business memiliki persentase bad borrower terbanyak yaitu sekitar 32,83% dari keseluruhan peminjam dengan purpose small business, diikuti oleh other dimana peminjam dengan purpose other ini adalah peminjam yang maksud atau purposenya selain yang disebutkan di atas atau yang terdapat pada grafik



Terakhir untuk term peminjaman, peminjam untuk term 60 Months memiliki persentase bad borrower lebih banyak dibanding dengan term 36 Months

Modeling

```
Train features: (193647, 9)
Train target: (193647,)
Test features: (48412, 9)
Test target: (48412,)
```

```
y_train.value_counts()
```

```
1    149475
0     44172
```

Modeling dilakukan dengan membagi jumlah train set dan test set menjadi 80% dan 20% dengan hasil seperti pada gambar. Disini terdapat imbalance data dimana jumlah y features dengan status good borrower atau (1) 3 kali lebih banyak daripada status bad borrower atau (0). Maka dari itu dilakukan proses oversampling

Cross Validation

Setelah melakukan oversampling, dilakukan proses cross validation pada data training dengan beberapa model yaitu Decision tree, Random forest, K-Neighbors, Logistic regression dan Naïve Bayes. Setelah proses training, maka didapat nilai score dan standard deviasi seperti pada gambar. Model Logistic regression memiliki nilai score tertinggi yaitu sekitar 76,52% dan standard deviasi 0,00897.

	score_mean	score_std
LogisticRegression	0.765181	0.008971
GaussianNB	0.741505	0.023230
RandomForestClassifier	0.737130	0.014790
KNeighborsClassifier	0.734903	0.002783
DecisionTreeClassifier	0.642851	0.021431

Evaluation

	train score	test score	difference
DecisionTreeClassifier	0.997612	0.663079	0.334532
RandomForestClassifier	0.997598	0.712592	0.285006
KNeighborsClassifier	0.815524	0.608196	0.207328
LogisticRegression	0.636658	0.616376	0.020282
GaussianNB	0.625750	0.651719	0.025968

Selain dengan menggunakan cross validation, dilakukan juga uji nilai akurasi pada training dan testing yang bertujuan untuk melihat akurasi saat data training dan testing dan perbedaannya. Terlihat bahwa model Logistic regression memiliki nilai perbedaan akurasi terkecil antara data training dan data test yaitu sekitar 2,03%, diikuti oleh model Gaussian NB dengan nilai sekitar 2,59%. Decision tree memiliki nilai akurasi yang sangat tinggi pada data training yaitu sekitar 99.76% sedangkan pada data testing nilai akurasinya berkurang menjadi sekitar 66,30%

Cross Validation for Some Metrics

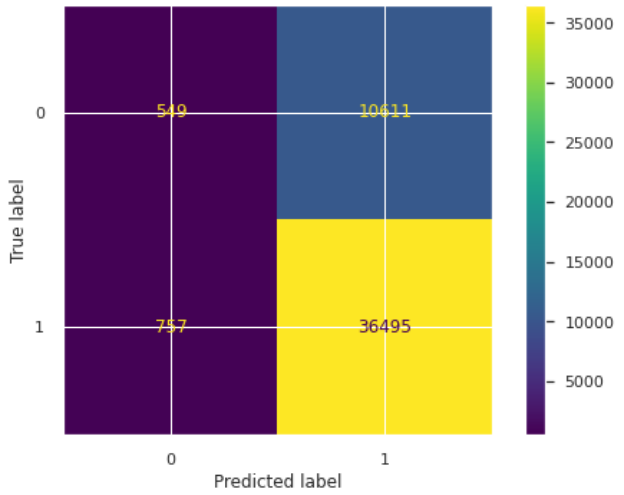
	Model	fit_time	score_time	test_accuracy	test_precision	test_recall	test_f1
0	LogisticRegression	3.081161	0.073173	0.765181	0.776250	0.977700	0.865148
1	GaussianNB	0.109800	0.090489	0.741505	0.811672	0.866645	0.837338
2	RandomForestClassifier	54.300572	2.392725	0.736333	0.787008	0.902639	0.840635
3	KNeighborsClassifier	0.794664	4.138112	0.734903	0.783488	0.907057	0.840718
4	DecisionTreeClassifier	1.966749	0.101278	0.642149	0.786760	0.735550	0.759749

Selanjutnya dilakukan cross validation untuk mengukur performa model dengan beberapa metrik yaitu accuracy, precision, recaal dan f1-score.

Dari hasil diatas, LogisticRegression memiliki nilai akurasi dan recall yang paling tinggi dari keseluruhan model. Pada kasus ini selain nilai akurasi yang kita perhitungkan, recall juga akan kita perhitungkan karena kita lebih ingin model kita dapat mengklasifikasi lebih banyak False Positive(FP) daripada False Negative (FN). FP pada kasus ini yaitu model memprediksi peminjam good borrower, tetapi sebenarnya tidak. Maka FP lebih baik daripada FN. FN yaitu model memprediksi peminjam merupakan bad borrower tetapi sebenarnya good, dan hal ini dapat menyebabkan semakin banyak peminjam yang memiliki predikat good borrower.

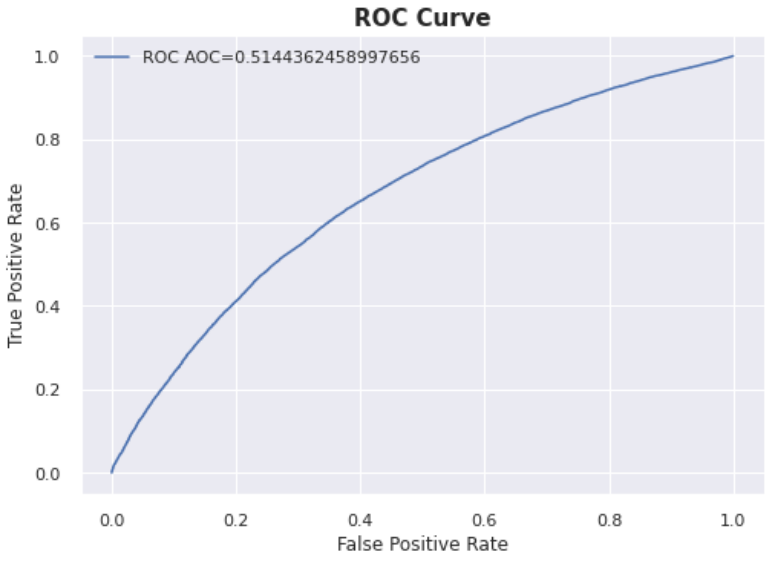
Dari keseluruhan proses diatas, maka dipilih model Logistic Regression karena memiliki nilai accuracy dan recall yang cukup tinggi

Training Test Set



Selanjutnya dilakukan proses training data test menggunakan model logistic regression dengan parameter default dan didapatkan hasil seperti gambar

```
Training Accuracy : 76.74%
Test Accuracy : 76.52%
Precision Score : 77.47%
Recall Score : 97.97%
F1 Score : 86.52%
```



	precision	recall	f1-score	support
0	0.42	0.05	0.09	11160
1	0.77	0.98	0.87	37252
accuracy			0.77	48412
macro avg	0.60	0.51	0.48	48412
weighted avg	0.69	0.77	0.69	48412

Dari hasil thresholds adjusment dapat dilihat bahwa hasilnya terdapat 37.252 berhasil diklasifikasi sebagai good borrower (1) dan sebanyak 11.160 diklasifikasi sebagai bad borrower (0). Selain itu model di evaluasi ROC AUC dan didapatkan ROC AUC score yaitu sekitar 0,514

Sumber gambar : https://www.badcredit.org/how-to/bad-credit-loans-with-preapproval/