CREDIT SCORING MODELLING

Ricky Hermanto

Credit scoring model adalah suatu model yang digunakan lembaga pembiayaan atau penyalur kredit untuk menentukan apakah seseorang layak atau tidak mendapatkan pinjaman. Hal ini dilakukan untuk meminimalkan kredit macet ‘*non-performing loop*’ atau pinjaman yang tidak *perform*.

Dalam projek ini saya akan membangun sebuah classification model credit scoring untuk memprediksi nasabah dari client mana yang termasuk dalam kategori ‘good loan’ dan ‘bad loan’ sehingga dapat digunakan di masa depan untuk memprediksi nasabah mana yang akan gagal bayar atau tidak.

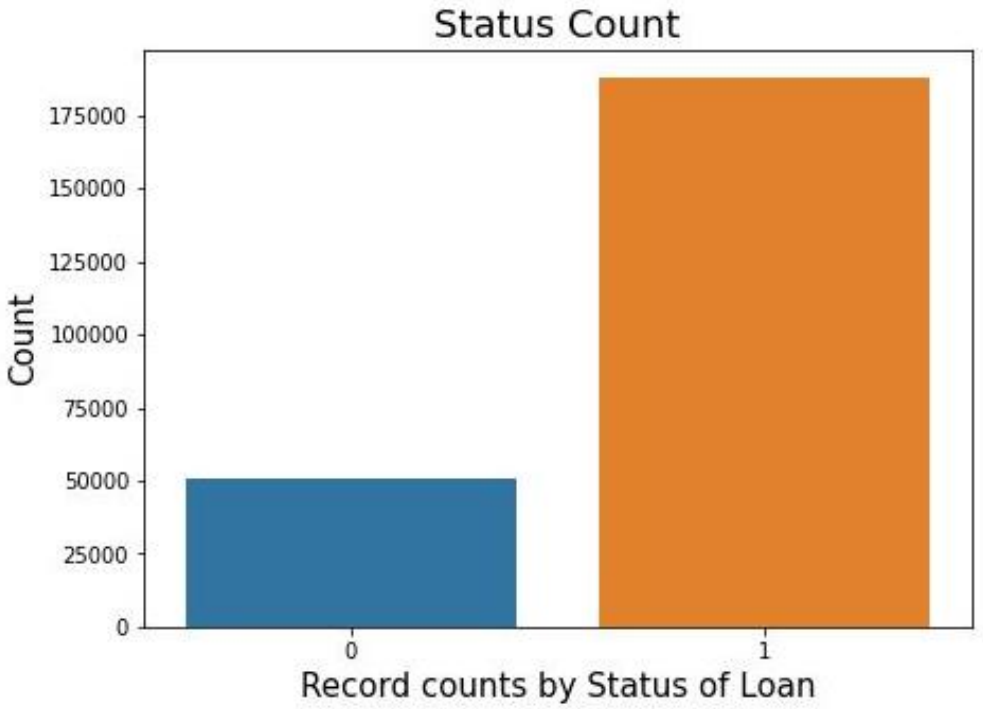
Dataset yang digunakan di sini memiliki historis data pinjaman selama 7 tahun yang diambil dari client PT. ID/X Partners. Dengan melihat status pinjaman, kita dapat memeriksa apakah pinjaman tersebut diberikan kepada nasabah atau tidak. Jika status pinjaman "charged off" atau "default" atau the issuance of loan got delayed by 31-120 days, itu adalah pinjaman macet(’bad loan’).

BUSINESS UNDERSTANDING

DATA UNDERSTANDING

Kumpulan datanya mencakup data - data nasabah yang pernah memakai produk kredit client PT.ID/X Partners. Dari total 466285 baris data yang ada terdapat 87.73% adalah label 1 atau klasifikasi baik kepada pelanggan yang taat bayar credit. Kumpulan data ini sangat tidak seimbang, dengan klasifikasi buruk menyumbang 12.27% dari total data yang ada.

Kumpulan datanya juga memiliki banyak nan value pada kolom - kolom featurenya. feature 'loan_status' merupakan label class yang akan digunakan untuk memprediksi nantinya, sebelum dipakai karena ada beberapa kategori dalam 'loan_status', maka harus terlebih dahulu dikelompokkan kedalam label 1 (good loan) dan 0 (bad loan).

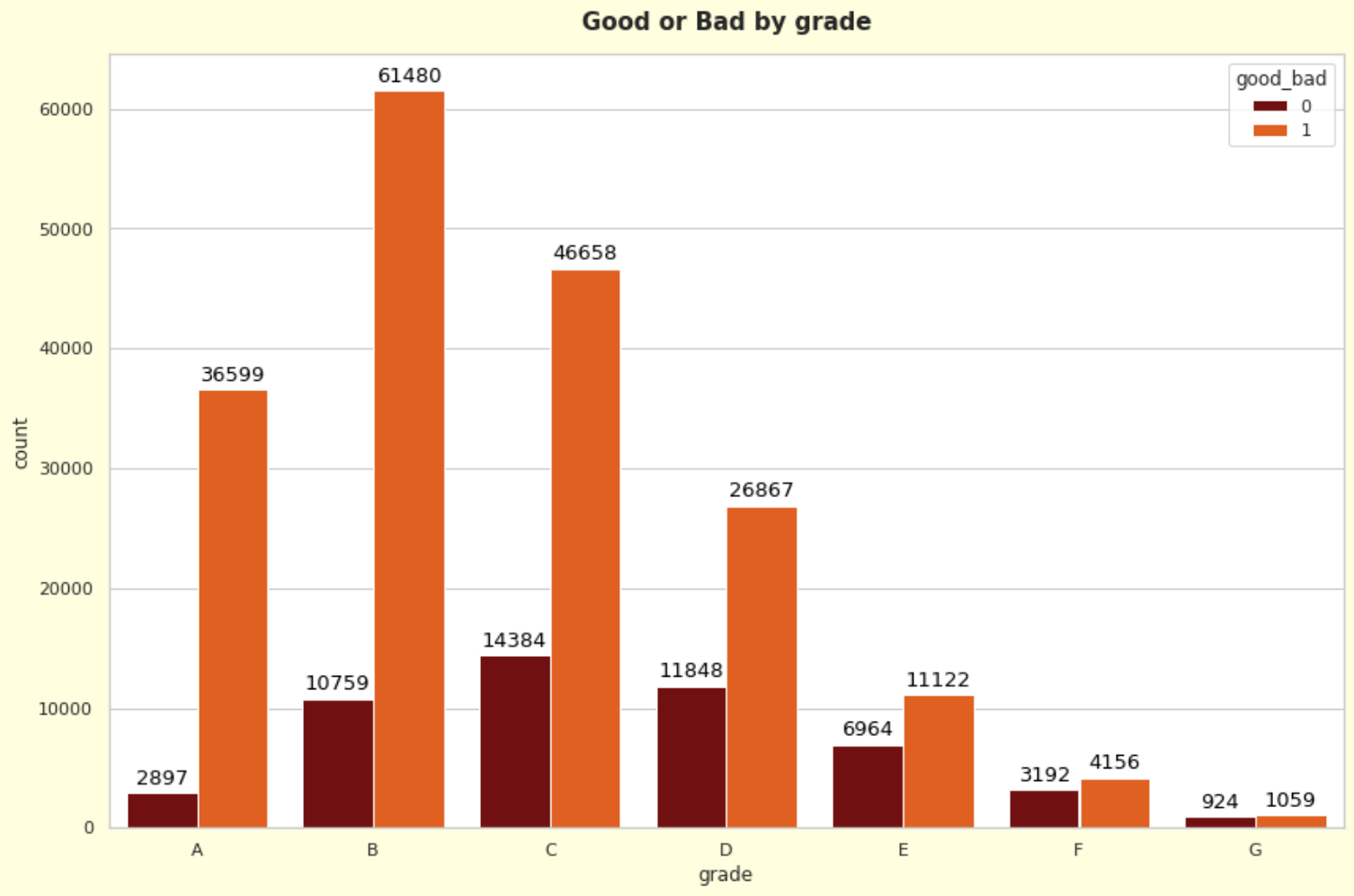


EXPLORATORY DATA ANALYSIS (MULTIVARIATE ANALYSIS)

PURPOSE EDA

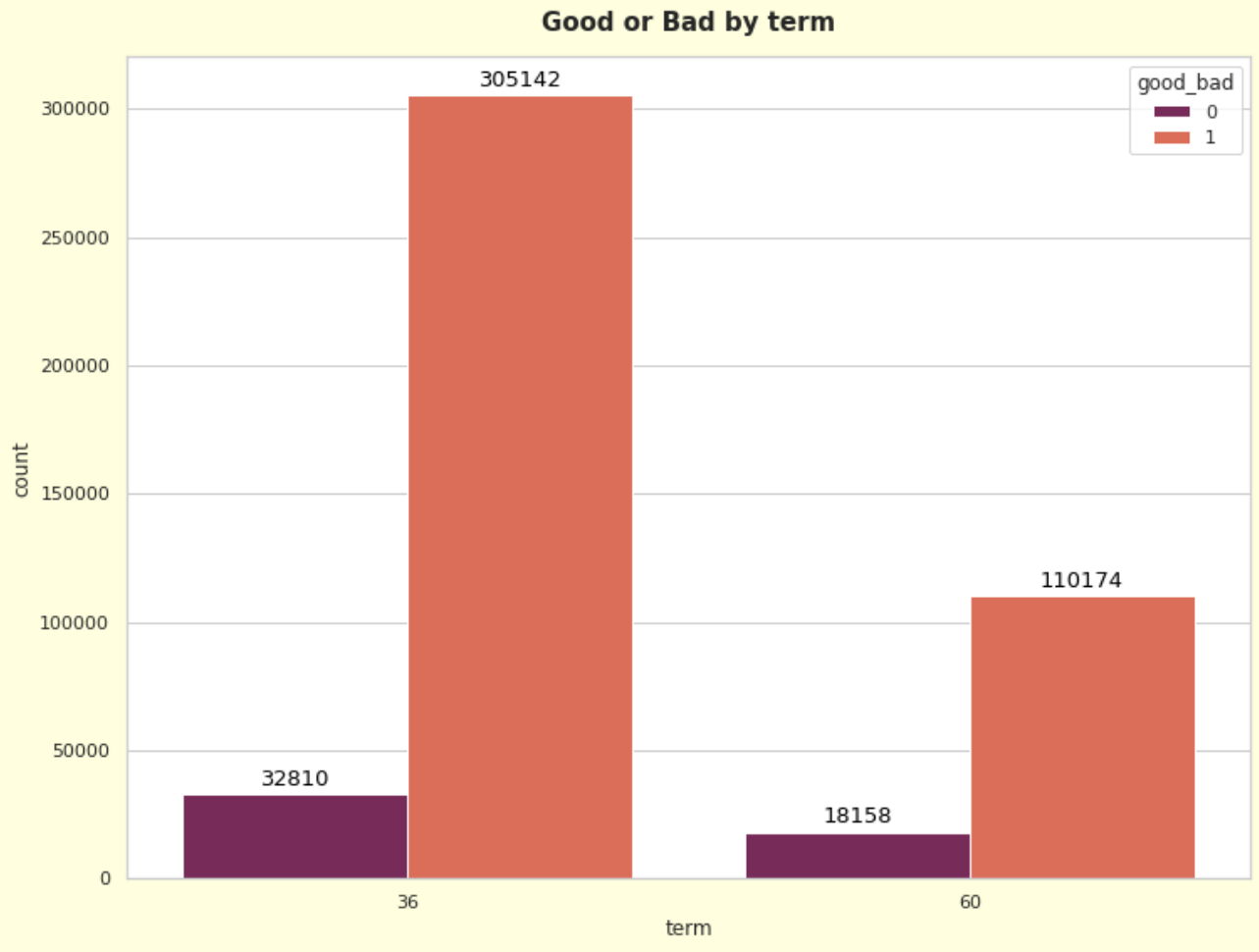
Tujuan dilakukannya EDA dengan Multivariate Analysis ini adalah untuk menganalisis karakteristik dari data historis nasabah sebelumnya. Yang hasil dari analisisnya akan dipakai untuk memprediksi nasabah baru dengan model machine learning. Selain untuk prediksi juga sebagai salah satu pertimbangan untuk pengambilan keputusan, langkah apa yang harus diambil selanjutnya.

Dari grafik terlihat bahwa nasabah paling banyak pada grade A, B,C dan paling sedikit pada D,E,F,G. ini akan berpengaruh pada resiko credit consolidation nantinya.



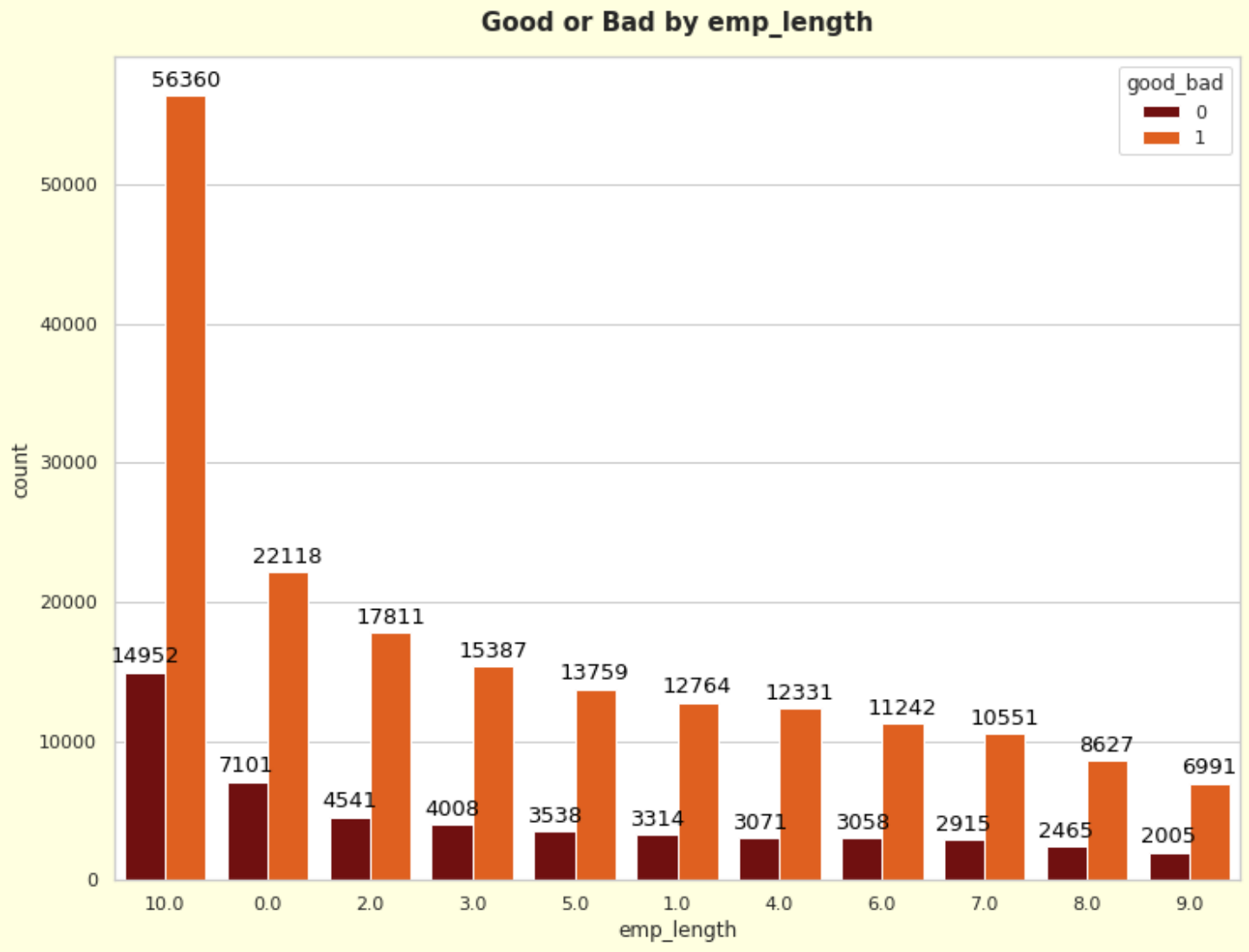
GRADE

Dari grafik terlihat bahwa para nasabah lebih banyak memakai term 36 months dibanding term 60 months.



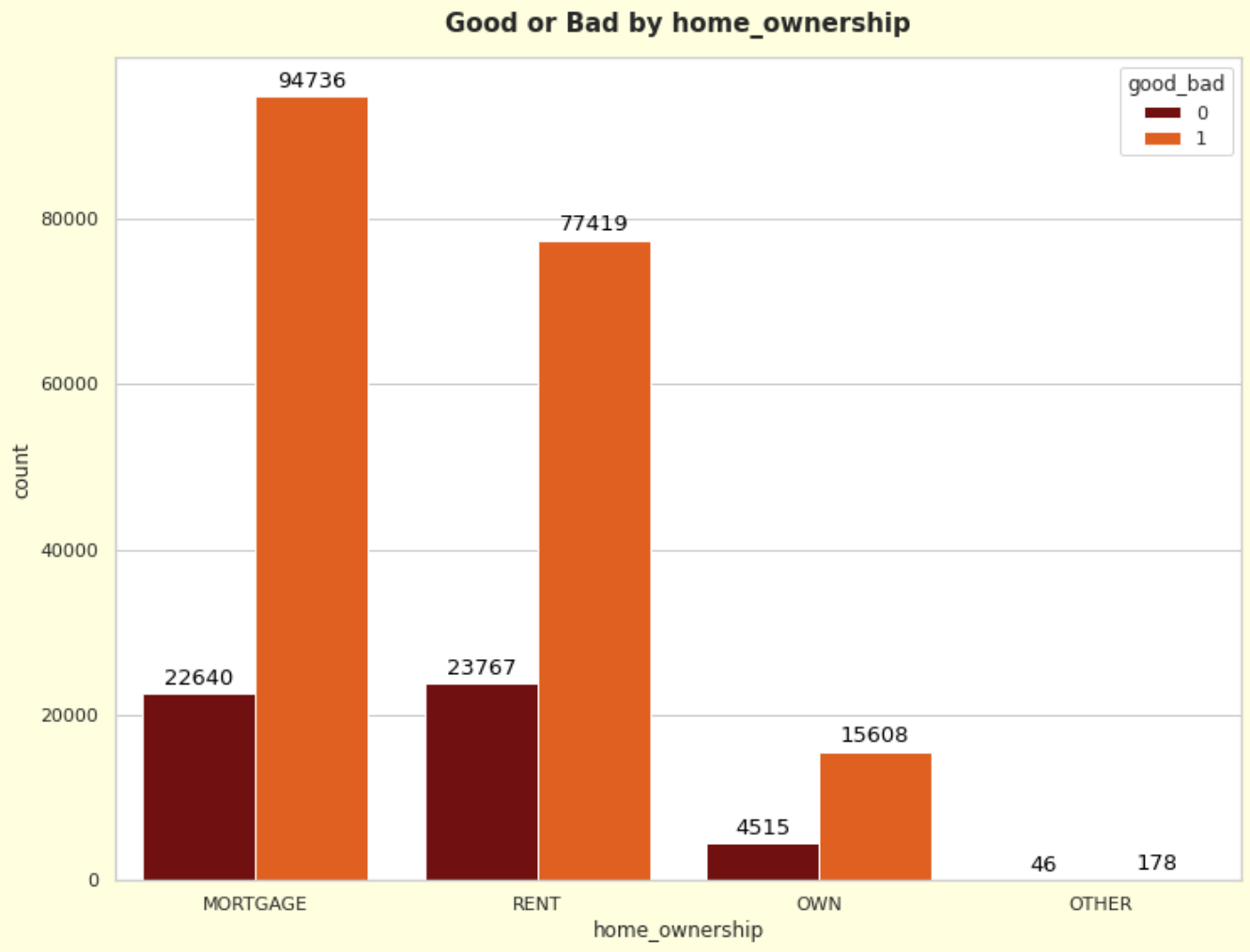
TERM

emp_length adalah lamanya para nasabah sudah bekerja, terlihat di grafik bahwa rata – rata lamanya para nasabah bekerja paling banyak berkisar dari 5 – 10 tahun.



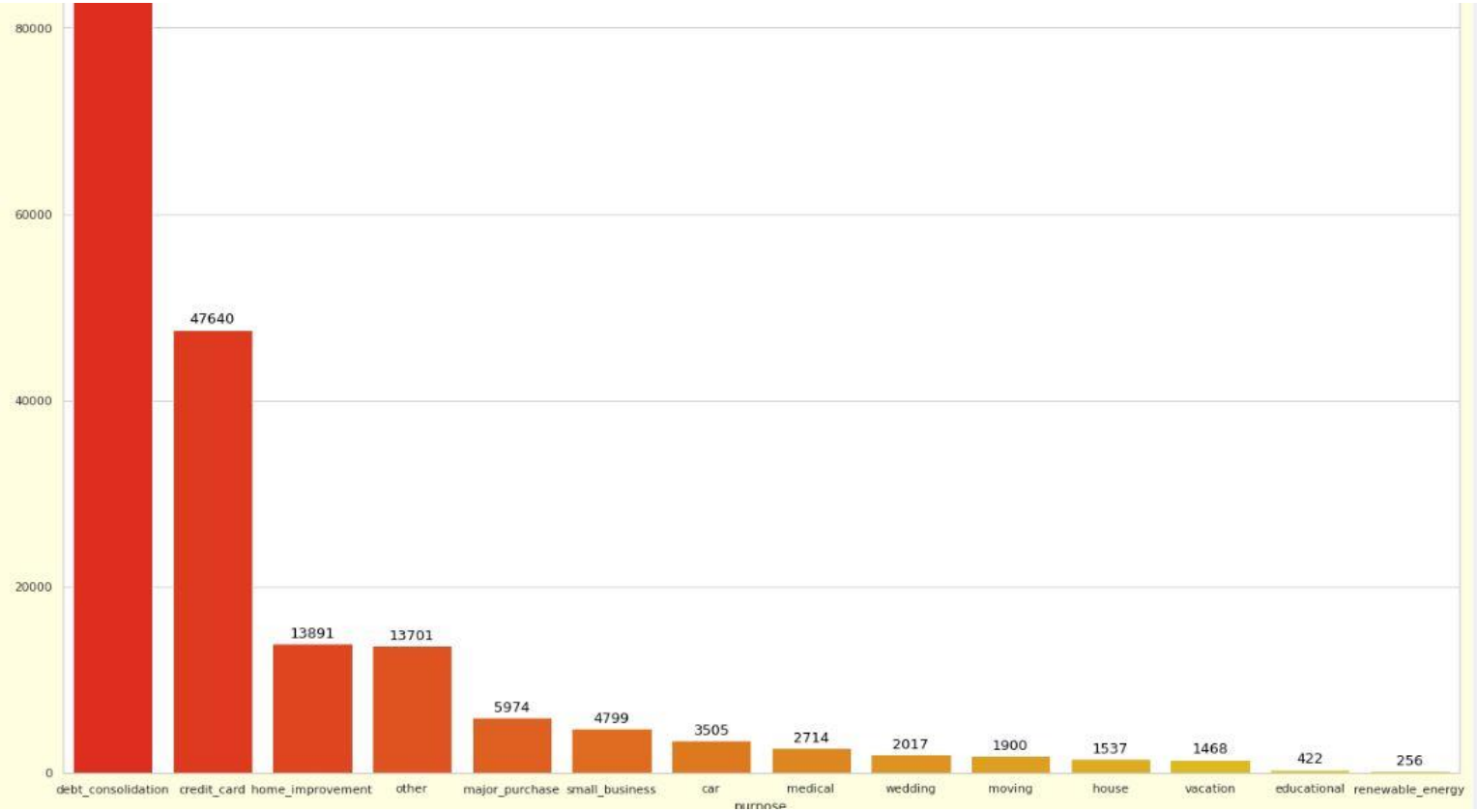
EMP_LENGTH

Home_ownership adalah status kepemilikan tempat tinggal para nasabah, terlihat bahwa para nasabah kebanyakan berasal dari status home_ownership = mortgage.



HOME_OWNERSHIP

Purpose, ini merupakan alasan kenapa mengambil pinjaman. terlihat dari graph, bahwa posisi teratas adalah untuk debt consolidation atau untuk penyelesaian utang. lalu di posisi setelahnya adalah credit card, ini mungkin untuk pembayaran kartu kredit, lalu ada home improvements ini asumsi saya dipakai untuk renovasi rumah , dll.



PURPOSE

CONCLUSION BASED ON EDA

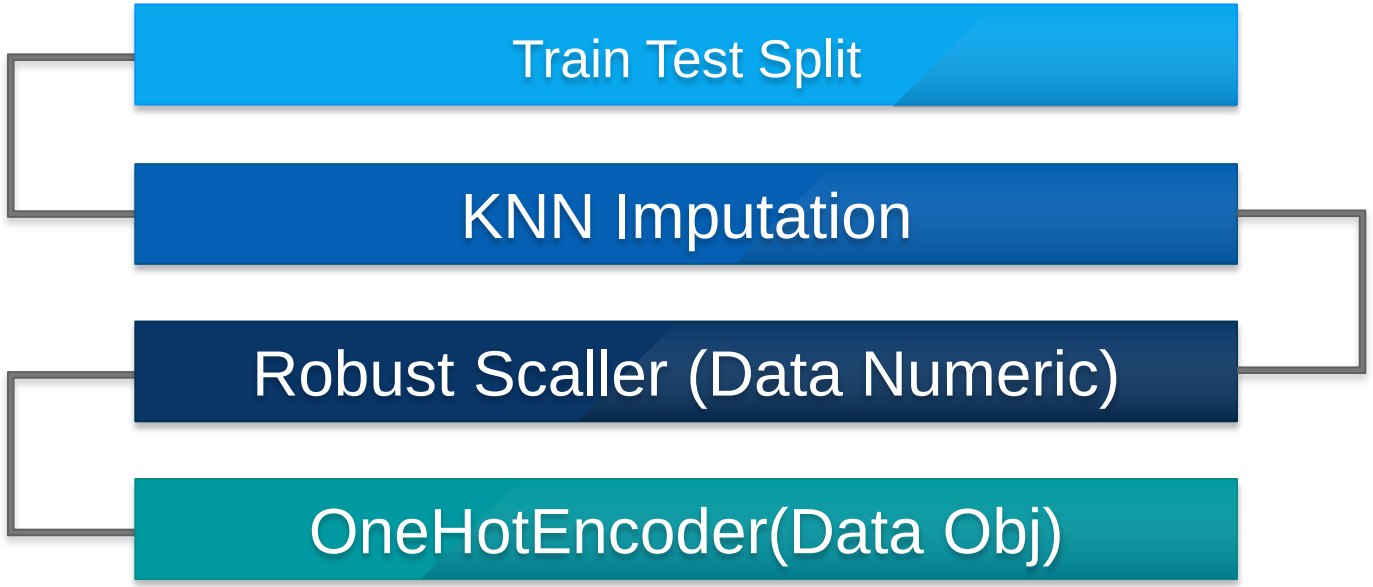
menurut saya dari hasil pengamatan graph diatas. bila di summary semuanya, para nasabah rata - rata, memiliki masa kerja diatas **5 - 10 tahun**, purpose pemakaian kredit untuk **debt consolidation/penyelesaian utang**, term paling disukai **36 months** dibanding **60 months**, lalu home_ownershipnya paling banyak adalah **Mortgage atau rumah cicilan**, **grade rata - rata paling banyak A dan B**.

Saya berkesimpulan, untuk memberi saran kepada client untuk membuat sebuah **Program KPR(Kredit Cicilan Rumah) untuk menarik nasabah baru berdasarkan hasil EDA tadi**, skemanya bisa diatur oleh tim marketing. bila program KPR sudah ada, maka perlu dimaksimalkan kembali program tersebut. karna dari data terlihat jelas **home_ownershipnya rata - rata mortgage**, **purposenya untuk debt consolidation** salah satunya asumsi saya adalah untuk **pembayaran cicilan rumah** tadi. jadi dengan adanya program ini (bila belum ada) nasabah akan lebih tertarik juga untuk menggunakan **term 60 months** atau mungkin ada term yang lebih dari 60 months misal 120 months contohnya. karna untuk KPR yang saya pernah dapet info jarang sekali yang memakai short term seperti 36 months. untuk penawarannya mungkin bisa ke nasabah dengan **grade A atau grade B**, karna terlihat para nasabah dengan good loan rata - rata paling banyak antara grade A dan B. bila sudah berhasil baru mungkin bisa ditawarkan ke nasabah lain dengan grade dibawah B seperti C, D, atau E.

MODELLING

FEATURE ENGINEERING

Ada beberapa step yang saya terapkan untuk memproses datanya.



ExtraTreesClassifier
ExtraTreesClassifier

FEATURE SELECTION

Setelah Feature Engineering, saya menerapakan Feature Selection untuk seleksi fitur – fitur mana yang mempunyai korelasi yang kuat. Untuk feature selection saya menggunakan beberapa selection seperti:

- WOE-IV
- Random Forest (feature Importances)
- ExtraTreesClassifier(feature Importances)
- LinearSVC
- Recursive feature elimination (RFE)
- SelectKBest + ChiSquare

Setelah semua selection berjalan lalu dilakukan sistem voting.

ExtraTreesClassifier
ExtraTreesClassifier

HASIL FEATURE SELECTION

	index	IV	RF	Extratrees	Chi_Square	RFE	L1	final_score
23	last_pymnt_amnt	0	1	1	1	1	1	5
3	total_rec_int	1	1	1	0	1	1	5
41	mths_since_last_pymnt_d	0	1	1	0	1	1	4
21	total_pymnt	0	1	1	0	1	1	4
4	term	1	0	0	1	0	1	3
5	mths_since_issue_d	0	1	0	0	1	1	3
42	out_prncp	0	0	1	1	0	1	3
11	grade_A	0	0	0	1	1	1	3
32	verification_status_Not Verified	0	0	0	0	1	1	2
26	home_ownership_RENT	0	0	0	0	1	1	2
1	total_rev_hi_lim	1	0	0	0	0	1	2
28	grade_F	0	0	0	0	1	1	2
29	purpose_debt_consolidation	0	0	0	0	1	1	2
0	purpose_house	1	0	0	0	1	0	2
38	purpose_major_purchase	0	0	0	0	1	1	2
45	purpose_vacation	0	0	0	0	1	1	2
50	grade_C	0	0	0	0	1	1	2
51	purpose_small_business	0	0	0	0	1	1	2
53	pymnt_plan_n	0	0	0	0	1	1	2
22	purpose_credit_card	0	0	0	0	1	1	2
27	purpose_car	0	0	0	0	1	1	2
12	verification_status_Source Verified	0	0	0	0	1	1	2
2	mths_since_earliest_cr_line	1	0	0	0	0	1	2
17	initial_list_status_w	0	0	0	0	1	1	2
16	initial_list_status_f	0	0	0	0	1	1	2
15	grade_B	0	0	0	0	1	1	2
14	grade_E	0	0	0	0	1	1	2
8	grade_G	0	0	0	0	1	1	2
47	collections_12_mths_ex_med	0	0	0	0	1	0	1
43	inq_last_6mths	0	0	0	0	0	1	1

ExtraTreesClassifier
ExtraTreesClassifier

MODEL

Saya mencoba membuat beberapa model sekaligus:

- Decision Tree
- Random Forest
- Logistic Regression dengan penalty L1 dan L2
- XGBoost
- MLP (Multi Layer Perceptrom)

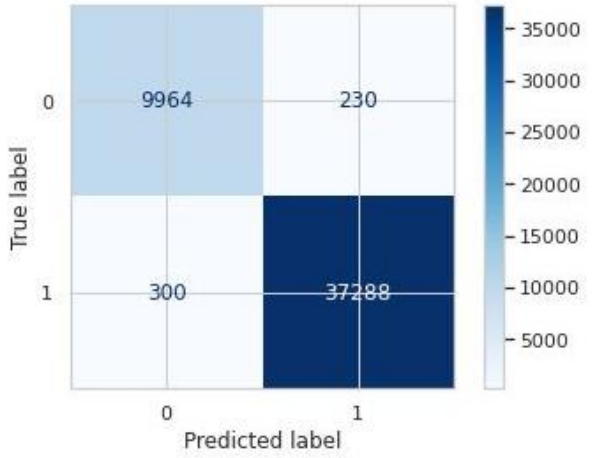
Saya putuskan untuk mengambil MLP model sebagai model saya, karna hasilnya yang paling baik menurut saya.

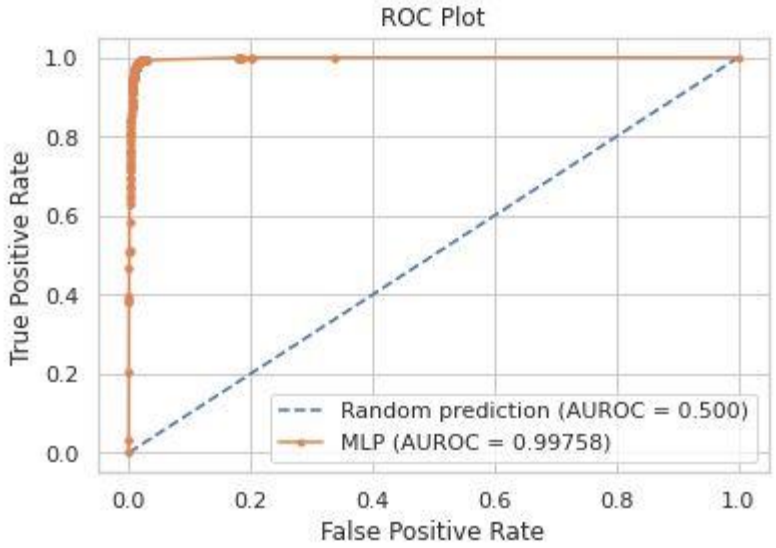
ExtraTreesClassifier
ExtraTreesClassifier

MODEL

```
Training Accuracy: 0.9907757668984497
Testing Accuracy: 0.9889079569712445
              precision    recall  f1-score   support

           0       0.97      0.98      0.98     40774
           1       1.00      0.99      0.99    150353

    accuracy                           0.99    191127
   macro avg       0.98      0.99      0.99    191127
weighted avg       0.99      0.99      0.99    191127

              precision    recall  f1-score   support

           0       0.97      0.98      0.97     10194
           1       0.99      0.99      0.99     37588

    accuracy                           0.99     47782
   macro avg       0.98      0.98      0.98     47782
weighted avg       0.99      0.99      0.99     47782
```

```
Akurasi dari Multilayer Perceptron adalah: 98.89
Recall dari Multilayer Perceptron adalah: 99.20187293817176
Precision dari Multilayer Perceptron adalah: 99.38696092542246
```





ExtraTreesClassifier
ExtraTreesClassifier

CROSS VALIDATION

Cross Validation

```
kfold = KFold(10, shuffle=True)

scores = cross_validate(MLPClassifier(max_iter=120, random

print('Accuracy : ',scores['test_accuracy'].mean())
print('Precision : ',scores['test_precision'].mean())
print('Recall/ sensitivity : ',scores['test_recall'].mean(
print('AUC ROC : ',scores['test_roc_auc'].mean())
```

```
Accuracy :  0.9886358319756636
Precision :  0.9936846737618339
Recall/ sensitivity :  0.9918595483035771
AUC ROC :  0.9972453225793908
```

ExtraTreesClassifier
ExtraTreesClassifier

BACKTEST

```
MLP_model = joblib.load('/content/drive/MyDrive/Dataset/MLP_model_V.2_ricky.sav')
print(MLP_model.score(X_test_final, y_test))
```

```
0.9889079569712445
```

```
kfold = KFold(10, shuffle=True)

scores = cross_validate(MLPClassifier(max_iter=120, random_state=5, hidden_layer_

print('Accuracy : ',scores['test_accuracy'].mean())
print('Precision : ',scores['test_precision'].mean())
print('Recall/ sensitivity : ',scores['test_recall'].mean())
print('AUC ROC : ',scores['test_roc_auc'].mean())
```

```
Accuracy :  0.986584848547753
Precision :  0.9926983098575294
Recall/ sensitivity :  0.9902301830980514
AUC ROC :  0.9963222452330195
```

THANK YOU