Final Project

CREDIT RISK PREDICTION

Gilland Fausta Putra A

Latar Belakang

Credit risk adalah risiko yang harus ditanggung oleh sebuah bank atau lembaga pembiayaan lainnya ketika memberikan pinjaman kredit kepada seorang atau lembaga lain. Risiko berupa tidak bisa dibayarkannya pokok dan bunga pinjaman sehingga mengakibatkan kerugian.

Untuk memperkecil risiko kredit, biasanya dilakukan proses penilaian risiko sebelum diberikan pinjaman terhadap pihak peminjam. Manfaat atau keuntungan dari proses penilaian adalah memperkecil risiko bagi lembaga peminjam dalam memutuskan apakah aplikasi pengajuan pinjaman diterima atau ditolak oleh lembaga finansial.

Untuk menghitung *credit risk,* dapat dilakukan dengan perhitungan menggunakan *machine learning* berdasarkan data historis pinjaman.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Unnamed: 0                  466285 non-null  int64
 1   id                          466285 non-null  int64
 2   member_id                   466285 non-null  int64
 3   loan_amnt                   466285 non-null  int64
 4   funded_amnt                 466285 non-null  int64
 5   funded_amnt_inv             466285 non-null  float64
 6   term                        466285 non-null  object
 7   int_rate                    466285 non-null  float64
 8   installment                 466285 non-null  float64
 9   grade                       466285 non-null  object
 10  sub_grade                   466285 non-null  object
 11  emp_title                   438697 non-null  object
 12  emp_length                  445277 non-null  object
 13  home_ownership              466285 non-null  object
 14  annual_inc                  466281 non-null  float64
 15  verification_status         466285 non-null  object
 16  issue_d                     466285 non-null  object
 17  loan_status                 466285 non-null  object
 18  pymnt_plan                  466285 non-null  object
 19  url                         466285 non-null  object
 20  desc                        125983 non-null  object
 21  purpose                     466285 non-null  object
 22  title                       466265 non-null  object
 23  zip_code                    466285 non-null  object
```

Dataset Overview

- Dataframe memiliki total 466285 baris dan 75 kolom
- Dataframe masih memiliki null values di beberapa kolom
- Target klasifikasi adalah kolom loan_status dengan tipe data object
- Sisanya adalah *feature* (predictor)

Terlihat tidak ada data yang duplikat

```
[ ]  data_raw.id.nunique()

     466285

[ ]  data_raw.member_id.nunique()

     466285
```

Pembuangan fitur – fitur yang merupakan id unik, berupa free text dan nilai kosong semua

```
[ ]  cols_to_drop = ['id' , 'member_id' , 'url' , 'desc' , 'zip_code' ,
                     'annual_inc_joint' , 'dti_joint' , 'verification_status_joint' , 'open_acc_6m' ,
                     'open_il_6m' , 'open_il_12m' , 'open_il_24m' , 'mths_since_rcnt_il' , 'total_bal_il' ,
                     'il_util' , 'open_rv_12m' , 'open_rv_24m' ,
                     'max_bal_bc' , 'all_util' , 'inq_fi' , 'total_cu_tl' , 'inq_last_12m' , 'sub_grade']
```

Variabel Target

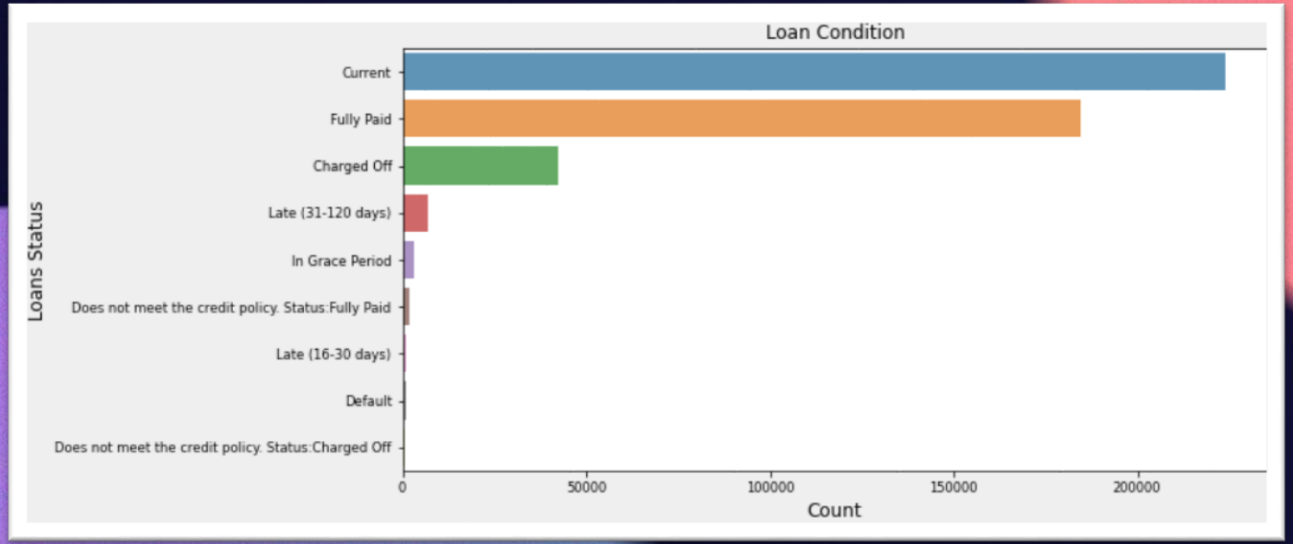
variabel `loan_status` dapat dijadikan variabel target karena mencerminkan performa tiapindividu dalam melakukan pembayaran terhadap pinjaman.

```
[11] data['loan_status'].unique()
```

```
array(['Fully Paid', 'Charged Off', 'Current', 'Default',
       'Late (31-120 days)', 'In Grace Period', 'Late (16-30 days)',
       'Does not meet the credit policy. Status:Fully Paid',
       'Does not meet the credit policy. Status:Charged Off'],
      dtype=object)
```

```
[12] data['loan_status'].value_counts()
```

```
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: loan_status, dtype: int64
```



Terdapat 9 nilai unik pada kolom loan_status yang akan menjadi target model.
Dibagi menjadi dua kelompok, yaitu "good_loan" dan "bad_loan".
"good_loan" didefinisikan memiliki status *Current*, *Fully Paid*, dan *In Grace Period.*

Descriptive Statistics

	Unnamed: 0	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths
count	466285.000000	466285.000000	466285.000000	466285.000000	466285.000000	466285.000000	4.662810e+05	466285.000000	466256.000000	466256.000000
mean	233142.000000	14317.277577	14291.801044	14222.329888	13.829236	432.061201	7.327738e+04	17.218758	0.284678	0.804745
std	134605.029472	8286.509164	8274.371300	8297.637788	4.357587	243.485550	5.496357e+04	7.851121	0.797365	1.091598
min	0.000000	500.000000	500.000000	0.000000	5.420000	15.670000	1.896000e+03	0.000000	0.000000	0.000000
25%	116571.000000	8000.000000	8000.000000	8000.000000	10.990000	256.690000	4.500000e+04	11.360000	0.000000	0.000000
50%	233142.000000	12000.000000	12000.000000	12000.000000	13.660000	379.890000	6.300000e+04	16.870000	0.000000	0.000000
75%	349713.000000	20000.000000	20000.000000	19950.000000	16.490000	566.580000	8.896000e+04	22.780000	0.000000	1.000000
max	466284.000000	35000.000000	35000.000000	35000.000000	26.060000	1409.990000	7.500000e+06	39.990000	29.000000	33.000000

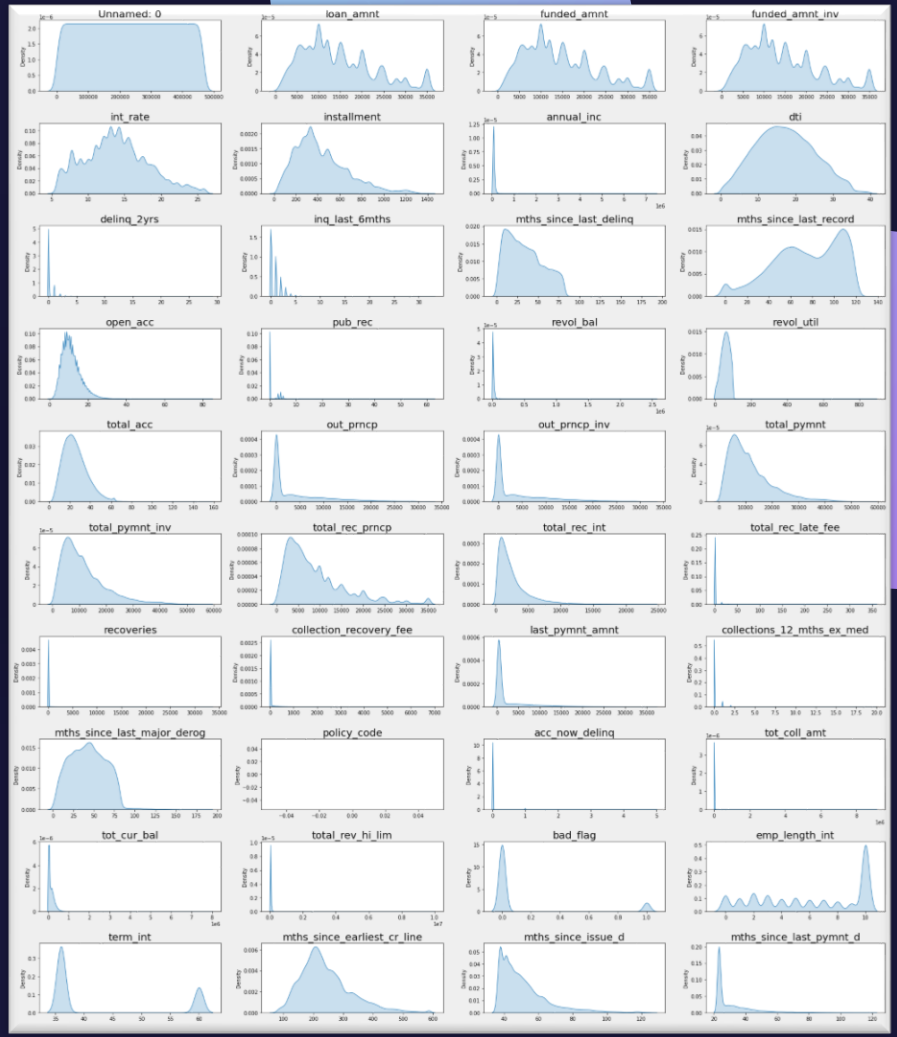
beberapa fitur *numerical* bernilai *null*

	grade	emp_title	home_ownership	verification_status	pymnt_plan	purpose	title	addr_state	initial_list_status	application_type
count	466285	438697	466285	466285	466285	466285	466265	466285	466285	466285
unique	7	205475	6	3	2	14	63099	50	2	1
top	B	Teacher	MORTGAGE	Verified	n	debt_consolidation	Debt consolidation	CA	f	INDIVIDUAL
freq	136929	5399	235875	168055	466276	274195	164075	71450	303005	466285

berapa fitur *categorical* memiliki terlalu banyak nilai unik dan hanya memiliki satu nilai unik

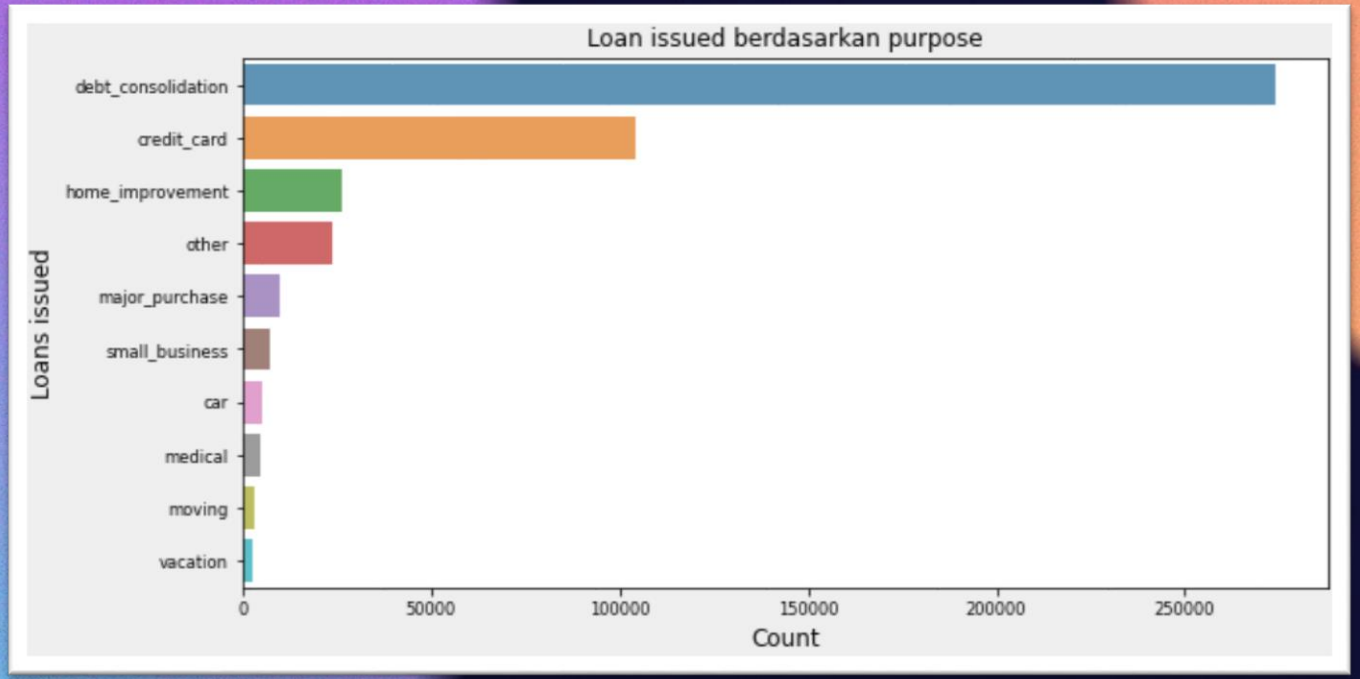
Univariate Analysis

Mayoritas fitur *numerical* tidak terdistribusi secara normal



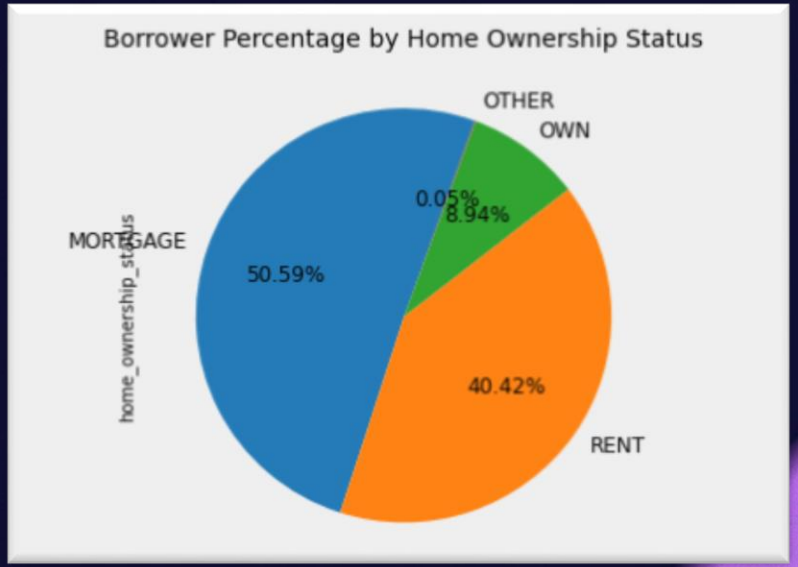
Multivariate Analysis

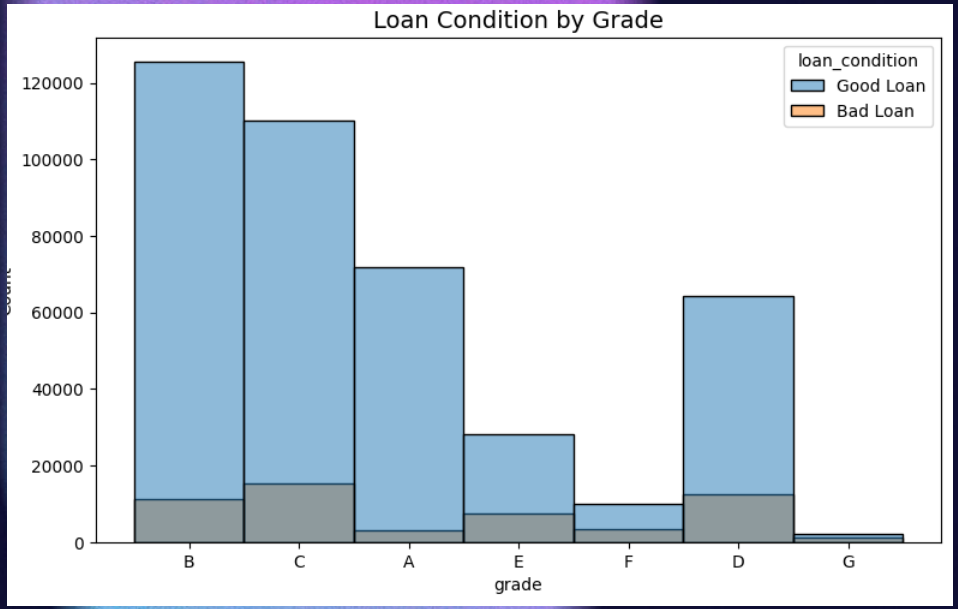
Tampaknya terdapat beberapa fitur independen yang sangat berkorelasi positif kuat satu sama lain, kemungkinan redundan



Terlihat bahwa tujuan peminjaman terbanyak terdapat dalam kategori debt_consolidation.

Peminjam dengan *status MORTAGE* memiliki persentase tertinggi





- Peminjam dengan *grade* B memiliki persentase tertinggi
- Semua subkategori didominasi oleh Good Loan

```
mths_since_last_record          86.566585
mths_since_last_delinq          53.690554
tot_coll_amt                    15.071469
tot_cur_bal                     15.071469
emp_length_int                   4.505399
revol_util                       0.072917
collections_12_mths_ex_med       0.031097
delinq_2yrs                      0.006219
inq_last_6mths                   0.006219
open_acc                         0.006219
pub_rec                          0.006219
total_acc                        0.006219
acc_now_delinq                   0.006219
mths_since_earliest_cr_line      0.006219
annual_inc                       0.000858
dtype: float64
```

Mengatasi Nilai yang Hilang

Mengelompokkan Data

Mengkonversi Tipe Data

MODELING

```
[ ]  from sklearn.model_selection import train_test_split
```

```
[ ]  X = data_model.drop('bad_status', axis=1)
     y = data_model['bad_status']
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

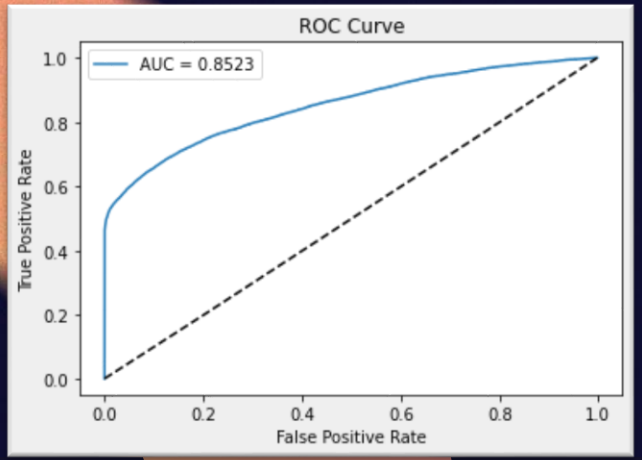
```
[ ]  from sklearn.ensemble import RandomForestClassifier
```

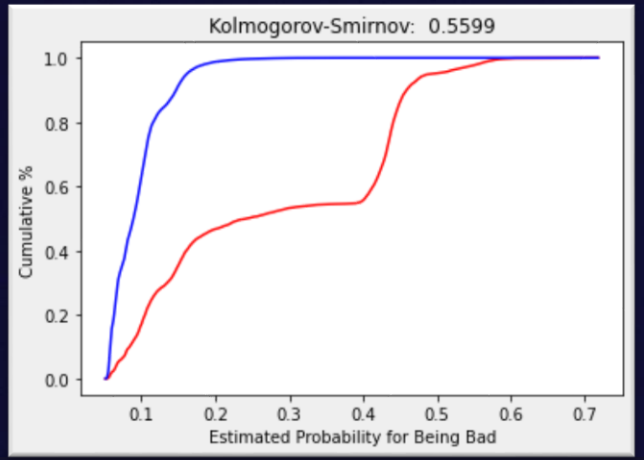
```
[ ]  rfc = RandomForestClassifier(max_depth=4)
     rfc.fit(X_train, y_train)

     RandomForestClassifier(max_depth=4)
```

	feature	importance
91	recoveries	4.437028e-01
89	out_prncp	1.954760e-01
90	total_rec_late_fee	1.032745e-01
78	int_rate	6.562778e-02

VALIDATION

	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	321938	0	0.051625	1	0	1	0.000011	0.0	0.000012
1	282547	0	0.051780	2	0	2	0.000021	0.0	0.000024
2	458381	0	0.051789	3	0	3	0.000032	0.0	0.000036
3	308835	0	0.051905	4	0	4	0.000043	0.0	0.000049
4	268474	0	0.051992	5	0	5	0.000054	0.0	0.000061






THANKS!