Credit Risk Loan Prediction

October 29th

Radif Ramadan

https://www.linkedin.com/in/radiframadan/

Objectives Analysis

Lendings

Automating lending decisions with credit risk model

Prediction

Building a credit risk prediction model using the Loan Dataset from Rakamin

Efficiency

Increased efficiency through automated lending decisions

Business Understanding

- **The company handles loan applications : The dataset supplied by ID/X partners from Rakamin includes both approved and declined loan data, indicating the company's involvement in processing loan applications.**

- **Managing credit risk is a priority for the company: The company's desire to develop a model for predicting credit risk indicates that managing credit risk is a top priority. This is because the company aims to prevent lending to individuals who are unlikely to repay, as this could lead to financial loses.**

- **The model will facilitate the company's lending decisions : The model under construction is expected to play a key role in shaping lending decisions. By predicting credit risk, the model will aid the company in determining whether to approve or deny loan applications, and may also impact the loan terms, such as the interest rate.**

Data Understanding

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 466285 entries, 0 to 466284
Data columns (total 74 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   id                           466285 non-null  int64
 1   member_id                    466285 non-null  int64
 2   loan_amnt                    466285 non-null  int64
 3   funded_amnt                  466285 non-null  int64
 4   funded_amnt_inv              466285 non-null  float64
 5   term                         466285 non-null  object
 6   int_rate                     466285 non-null  float64
 7   installment                  466285 non-null  float64
 8   grade                        466285 non-null  object
 9   sub_grade                    466285 non-null  object
 10  emp_title                    438697 non-null  object
 11  emp_length                   445277 non-null  object
 12  home_ownership               466285 non-null  object
 13  annual_inc                   466281 non-null  float64
 14  verification_status          466285 non-null  object
 15  issue_d                      466285 non-null  object
 16  loan_status                  466285 non-null  object
 17  pymnt_plan                   466285 non-null  object
 18  url                          466285 non-null  object
 19  desc                         125983 non-null  object
 20  purpose                      466285 non-null  object
 21  title                        466265 non-null  object
 22  zip_code                     466285 non-null  object
 23  addr_state                   466285 non-null  object
 24  dti                          466285 non-null  float64
 25  delinq_2yrs                  466256 non-null  float64
 26  earliest_cr_line             466256 non-null  object
 27  inq_last_6mths               466256 non-null  float64
 28  mths_since_last_delinq       215934 non-null  float64
 29  mths_since_last_record       62638 non-null   float64
 30  open_acc                     466256 non-null  float64
 31  pub_rec                      466256 non-null  float64
 32  revol_bal                    466285 non-null  int64
 33  revol_util                   465945 non-null  float64
 34  total_acc                    466256 non-null  float64
 35  initial_list_status          466285 non-null  object
 36  out_prncp                    466285 non-null  float64
 37  out_prncp_inv                466285 non-null  float64
 38  total_pymnt                  466285 non-null  float64
 39  total_pymnt_inv              466285 non-null  float64
 40  total_rec_prncp              466285 non-null  float64
 41  total_rec_int                466285 non-null  float64
 42  total_rec_late_fee           466285 non-null  float64
 43  recoveries                   466285 non-null  float64
 44  collection_recovery_fee      466285 non-null  float64
 45  last_pymnt_d                 465909 non-null  object
 46  last_pymnt_amnt              466285 non-null  float64
 47  next_pymnt_d                 239071 non-null  object
 48  last_credit_pull_d           466243 non-null  object
 49  collections_12_mths_ex_med   466140 non-null  float64
 50  mths_since_last_major_derog  98974 non-null   float64
 51  policy_code                  466285 non-null  int64
 52  application_type             466285 non-null  object
 53  annual_inc_joint             0 non-null       float64
 54  dti_joint                    0 non-null       float64
 55  verification_status_joint    0 non-null       float64
 56  acc_now_delinq               466256 non-null  float64
 57  tot_coll_amt                 396009 non-null  float64
 58  tot_cur_bal                  396009 non-null  float64
 59  open_acc_6m                  0 non-null       float64
 60  open_il_6m                   0 non-null       float64
 61  open_il_12m                  0 non-null       float64
 62  open_il_24m                  0 non-null       float64
 63  mths_since_rcnt_il           0 non-null       float64
 64  total_bal_il                 0 non-null       float64
 65  il_util                      0 non-null       float64
 66  open_rv_12m                  0 non-null       float64
 67  open_rv_24m                  0 non-null       float64
 68  max_bal_bc                   0 non-null       float64
 69  all_util                     0 non-null       float64
 70  total_rev_hi_lim             396009 non-null  float64
 71  inq_fi                       0 non-null       float64
 72  total_cu_tl                  0 non-null       float64
 73  inq_last_12m                 0 non-null       float64
dtypes: float64(46), int64(6), object(22)
memory usage: 266.8+ MB
```

- **Datasets comprises 74 columns with 466,285 rows, each column featuring various data types.**

- **Included 22 Categorical and 52 Numerical features.**
- **Data types include int64, float64 and object**

Statistical Descriptive

Numerical Columns

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	delinq_2yrs
count	4.662850e+05	4.662850e+05	466285.000000	466285.000000	466285.000000	466285.000000	466285.000000	4.662810e+05	466285.000000	466256.000000
mean	1.307973e+07	1.459766e+07	14317.277577	14291.801044	14222.329888	13.829236	432.061201	7.327738e+04	17.218758	0.284678
std	1.089371e+07	1.168237e+07	8286.509164	8274.371300	8297.637788	4.357587	243.485550	5.496357e+04	7.851121	0.797365
min	5.473400e+04	7.047300e+04	500.000000	500.000000	0.000000	5.420000	15.670000	1.896000e+03	0.000000	0.000000
25%	3.639987e+06	4.379705e+06	8000.000000	8000.000000	8000.000000	10.990000	256.690000	4.500000e+04	11.360000	0.000000
50%	1.010790e+07	1.194108e+07	12000.000000	12000.000000	12000.000000	13.660000	379.890000	6.300000e+04	16.870000	0.000000
75%	2.073121e+07	2.300154e+07	20000.000000	20000.000000	19950.000000	16.490000	566.580000	8.896000e+04	22.780000	0.000000
max	3.809811e+07	4.086083e+07	35000.000000	35000.000000	35000.000000	26.060000	1409.990000	7.500000e+06	39.990000	29.000000

8 rows × 52 columns

...	total_bal_il	il_util	open_rv_12m	open_rv_24m	max_bal_bc	all_util	total_rev_hi_lim	inq_fi	total_cu_tl	inq_last_12m
...	0.0	0.0	0.0	0.0	0.0	0.0	3.960090e+05	0.0	0.0	0.0
...	NaN	NaN	NaN	NaN	NaN	NaN	3.037909e+04	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	3.724713e+04	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	0.000000e+00	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	1.350000e+04	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	2.280000e+04	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	3.790000e+04	NaN	NaN	NaN
...	NaN	NaN	NaN	NaN	NaN	NaN	9.999999e+06	NaN	NaN	NaN

Statistical Descriptive

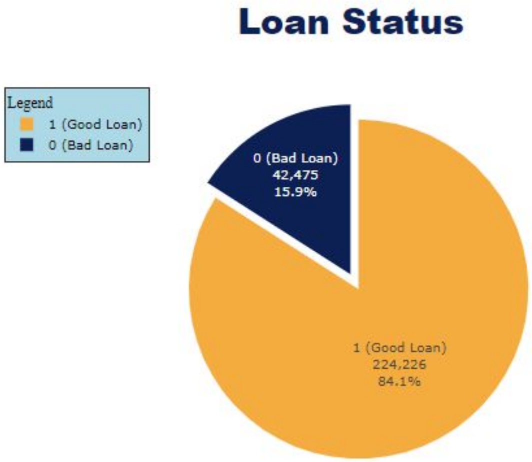
Categorical Columns

	count	unique	top	freq
term	466285	2	36 months	337953
grade	466285	7	B	136929
sub_grade	466285	35	B3	31686
emp_title	438697	205475	Teacher	5399
emp_length	445277	11	10+ years	150049
home_ownership	466285	6	MORTGAGE	235875
verification_status	466285	3	Verified	168055
issue_d	466285	91	Oct-14	38782
loan_status	466285	9	Current	224226
pymnt_plan	466285	2	n	466276
url	466285	466285	https://www.lendingclub.com/browse/loanDetail....	1
desc	125983	124436		234
purpose	466285	14	debt_consolidation	274195
title	466265	63099	Debt consolidation	164075
zip_code	466285	888	945xx	5304
addr_state	466285	50	CA	71450
earliest_cr_line	466256	664	Oct-00	3674
initial_list_status	466285	2	f	303005
last_pymnt_d	465909	98	Jan-16	179620
next_pymnt_d	239071	100	Feb-16	208393
last_credit_pull_d	466243	103	Jan-16	327699
application_type	466285	1	INDIVIDUAL	466285

Exploratory Data Analysis

Loan Status :

From the loan_status category, separated into various categories, to be divided into 2 categories, namely good_loan and bad_loan, which will be used as data labels

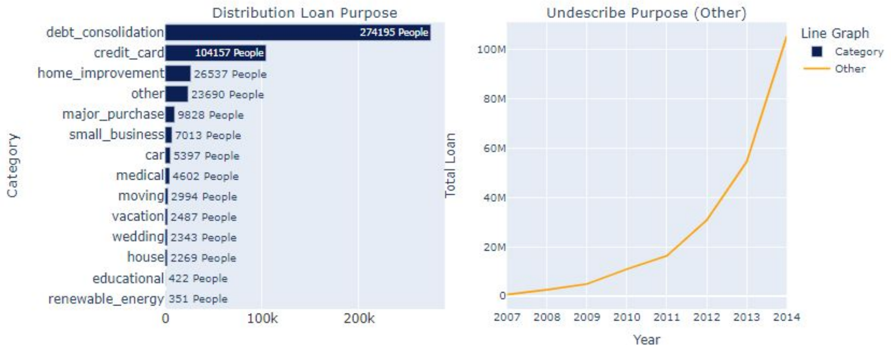


Exploratory Data Analysis

Loan Purpose :

The reason for borrowing credit is often for the purpose of debt consolidation and credit card usage. Additionally, there are some borrowing intentions not specified in the adjacent graph by year

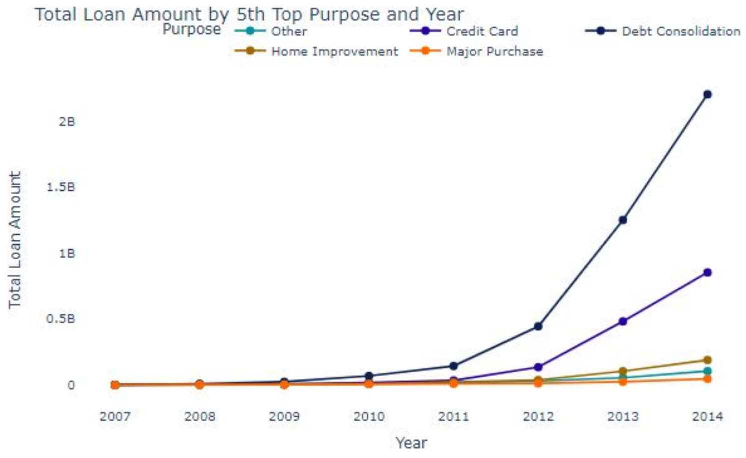
Loan Purpose



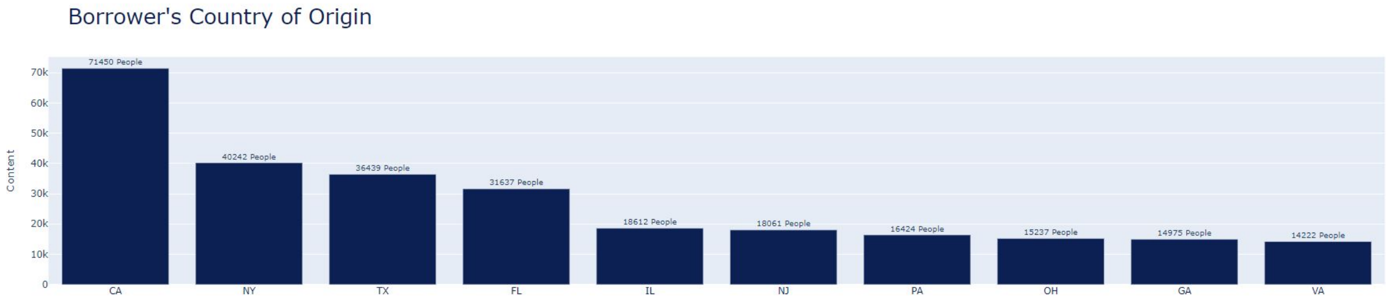
Exploratory Data Analysis

Total Loan Amount by Top 5th Purpose by Year :

Here are the top 5 loan totals based on the borrowing purposes per year, indicating that debt consolidation appears to have the highest value



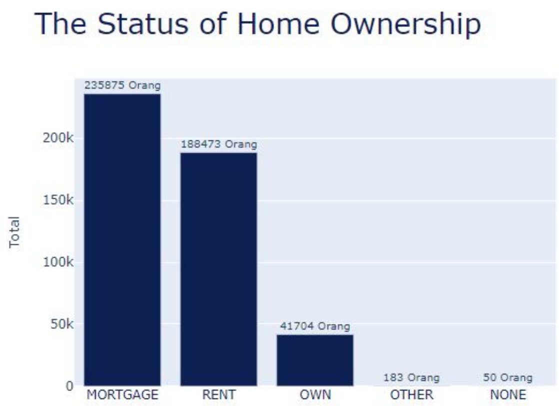
Exploratory Data Analysis



Borrower's Country of Origin :

The most active country utilizing the loan services is CA, according to the ISO 3166, which refers to Canada

Exploratory Data Analysis



The Status of Home Ownership :

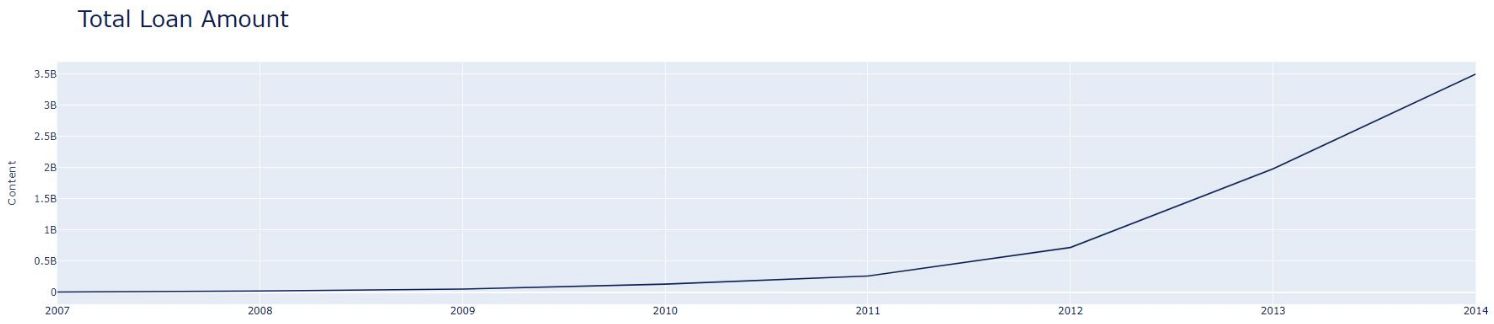
The majority of borrowers' home ownership status is Mortgage, which is used as collateral for the loan, while the remainder is solely Rent and Own.

Exploratory Data Analysis

Employment Title Borrower's :

There are several borrowers based on their job titles, and here it is evident that the job of a teacher has the highest number of loan applications

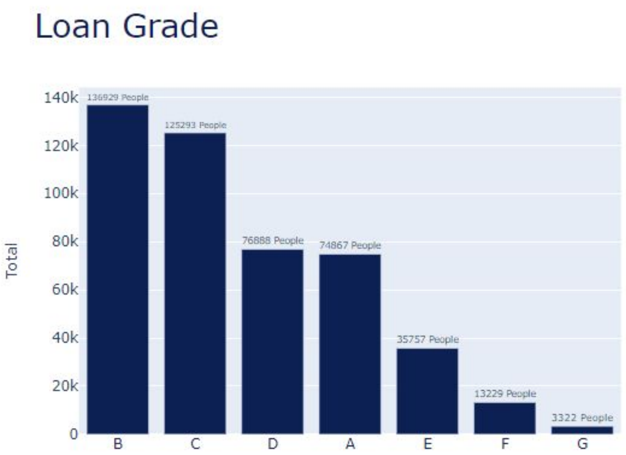
Exploratory Data Analysis



Total Loan Amount by Year :

Here are the total loan amounts per year from 2007 to 2014.

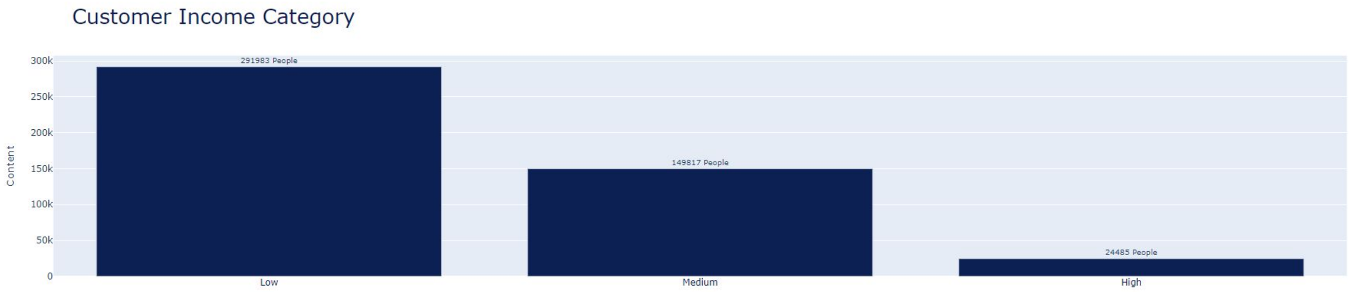
Exploratory Data Analysis



Loan Grade :

This visualization contains loan grades commonly taken by borrowers, and it appears that grade B loans are predominantly used.

Exploratory Data Analysis



Customer Income Category :

Several borrowers are categorized into income groups, and those with low income are the ones who utilize loans the most.

Data Preprocess & Cleaning

The Dataset is untidy.

- **Remove unnecessary data with numerous duplicates.**
- **Remove Data with NaN values in every row.**

Following the data cleansing process, 50 out of 74 columns remain

Clean Data

```
[ ] # For Categorical Data
    print('shape before drop = ',df.shape)

    df_clean = df.drop(columns=['member_id','id','emp_title','url','desc','title','zip_code','policy_code','application_type'], axis=1)
    df_clean.drop_duplicates(inplace=True)

    print('shape after drop = ',df_clean.shape)

    shape before drop =  (466285, 76)
    shape after drop =  (466285, 67)
```

Drop data yang tidak diperlukan di modeling dan memiliki duplikat

```
[ ] # For Numerical Data
    print('shape before drop = ',df_clean.shape)

    df_clean = df_clean.drop(columns=['annual_inc_joint','dti_joint','verification_status_joint','open_acc_6m','open_il_6m','open_il_12m','open_il_24m','mths_since_rcnt_il',
    'mths_since_rcnt_il','total_bal_il','il_util','open_rv_12m','open_rv_24m','max_bal_bc','all_util','inq_fi','inq_last_12m','total_cu_tl'], axis=1)

    print('shape after drop = ',df_clean.shape)

    shape before drop =  (466285, 67)
    shape after drop =  (466285, 50)
```

Drop data yang memiliki nilai NaN pada setiap barisnya

Data Preprocess & Cleaning

Labelling Data

We must transform the target column into a binary value based on the loan status conditions. Specifically, ['Charged Off', 'Default', 'Does not meet the credit policy. Status: Charged Off', 'Late (31-120 days)', 'Late (16-30 days)'] are considered bad statuses, while all other statuses are regarded as good.

Before

```
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
```

After

loan_status	good/bad
Fully Paid	1
Charged Off	0
Fully Paid	1
Fully Paid	1
Current	1
...	...
Current	1
Charged Off	0
Current	1
Fully Paid	1
Current	1

« »

Data Preprocess & Cleaning

Features Numerical Data

```
loan_status                    1.000000
recoveries                     0.435352
collection_recovery_fee        0.295281
total_rec_prncp                0.254255
total_pymnt_inv                0.194638
total_pymnt                    0.193977
int_rate                       0.174648
last_pymnt_amnt                0.170164
total_rec_late_fee             0.151624
out_prncp                      0.150442
out_prncp_inv                  0.150430
inq_last_6mths                 0.073109
term                           0.064644
revol_util                     0.051020
tot_cur_bal                    0.050865
dti                            0.049092
total_rev_hi_lim               0.037735
total_rec_int                  0.022833
mths_since_last_record         0.022542
total_acc                      0.022366
revol_bal                      0.018536
emp_length                     0.016499
installment                    0.015347
loan_amnt                      0.013181
funded_amnt                    0.012401
funded_amnt_inv                0.008686
pub_rec                        0.008279
open_acc                       0.005270
mths_since_last_delinq         0.004850
mths_since_last_major_derog    0.004253
collections_12_mths_ex_med     0.004126
delinq_2yrs                    0.002872
tot_coll_amt                   0.001178
acc_now_delinq                 0.000083
Name: loan_status, dtype: float64
```

```
affect_num_cols
```

```
['recoveries',
 'collection_recovery_fee',
 'total_rec_prncp',
 'int_rate',
 'last_pymnt_amnt',
 'total_rec_late_fee',
 'out_prncp',
 'out_prncp_inv']
```

ita akan menggunakan numerical yang memiliki korelasi di atas > 0,1 dengan loan_status sebagai feature numeric. Namun ada beberapa olom yang berkolasi diatas > 0,1 memiliki isi data yang ambigu

We have several numerical features related to Loan_Status,

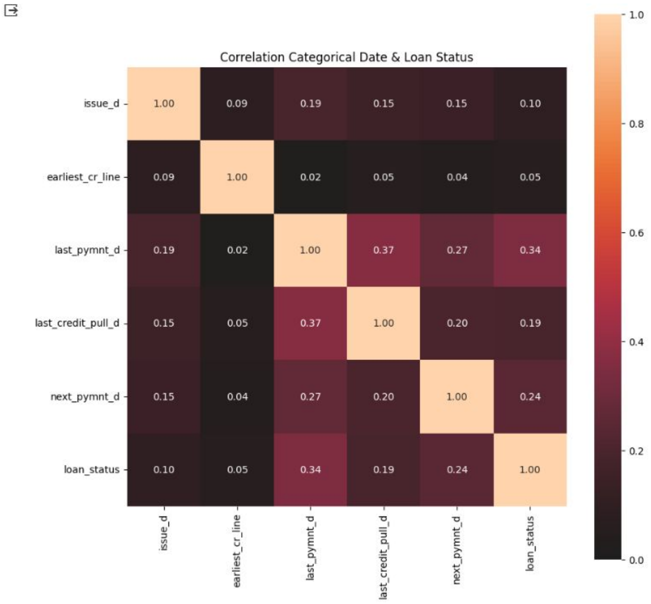
and we will select some columns to be used as features with a correlation value > 0.1.

It is worth noting that some columns with correlation values > 0.1 have ambiguous data values.

After consideration, these columns are the ones that become the numerical features.

Data Preprocess & Cleaning

Features Categorical Data (date)



Kita akan menggunakan kategori date yang memiliki korelasi di atas > 0,1 dengan loan_status sebagai feature kategori (date)

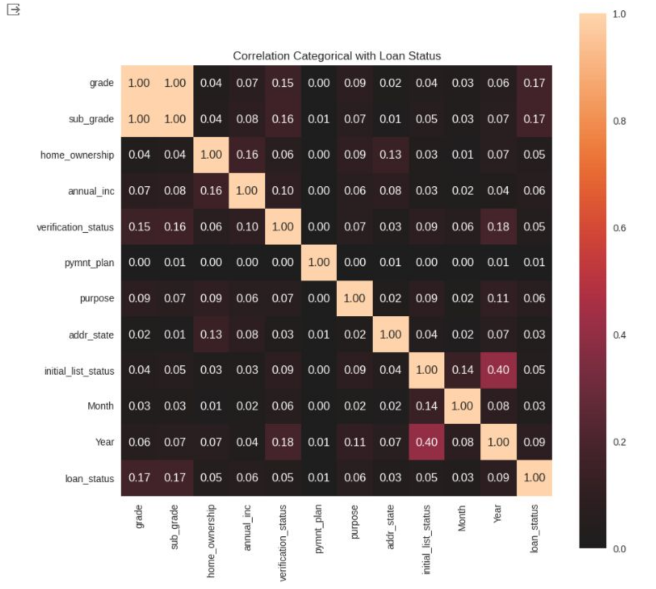
After that, we have categorical features that contain date information, and we will similarly apply it by using columns that have a correlation > 0.1.

```
affect_date_cols
```

```
['issue_d', 'last_pymnt_d', 'last_credit_pull_d', 'next_pymnt_d']
```

Data Preprocess & Cleaning

Features Categorical Data (date)



Kita akan menggunakan kategori yang memiliki korelasi di atas > 0,1 dengan loan_status sebagai feature kategori

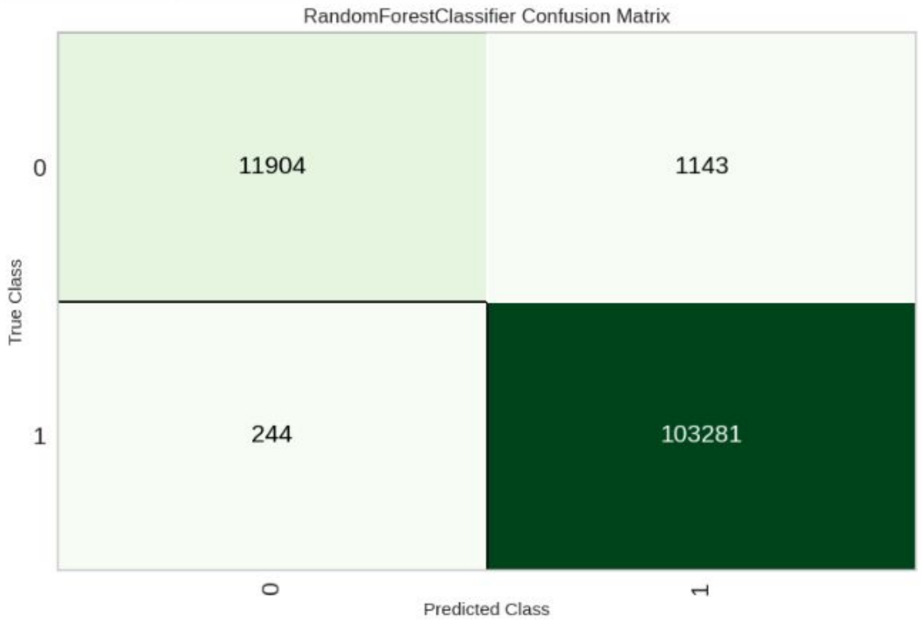
Similar to the previous categorical feature, we will use columns as features that have a correlation with loan_status > 0.1.

However, we encounter an issue with the sub_grade column, which evidently holds a better position for use as a feature grade

```
# Fitur kategorikal yang akan kita gunakan
affect_cat_cols = ["grade"]
```

Model Evaluation

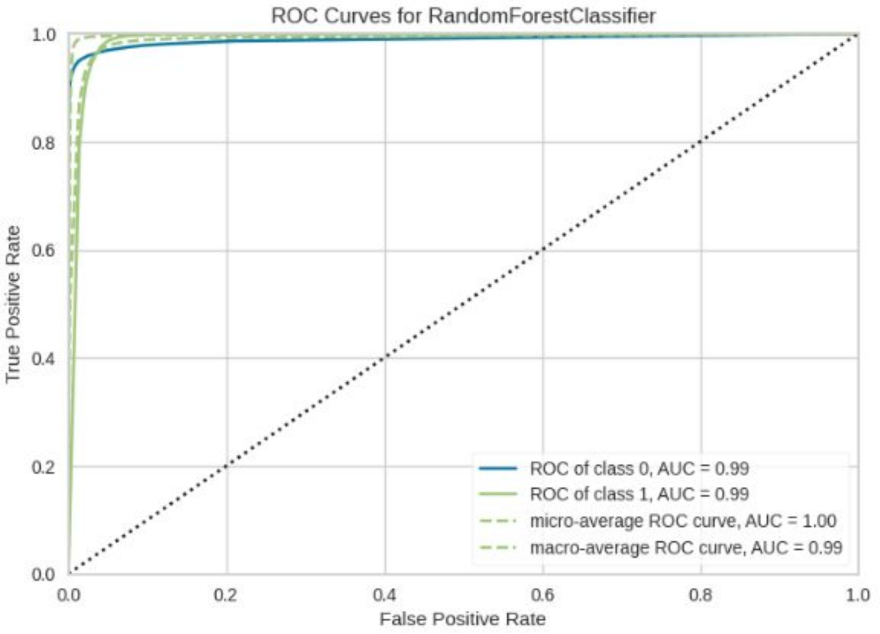
Random Forest Classifier Evaluation

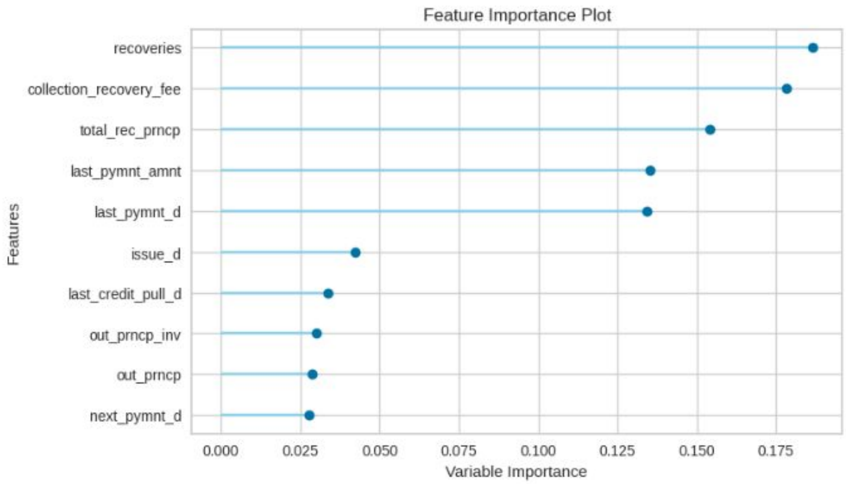


This model provides predictions with 103281 True Positives and 11904 True Negatives.

Model Evaluation

ROC AUC Evaluation & Feature Importance





Thanks!

I apologize for any shortcomings in this project.

I welcome any critique and suggestions to improve it in the future.