Credit Risk Loan Prediction by Machine Learning

Created by:
Rheza Paleva Uyanto
uyantorheza@gmail.com
https://www.linkedin.com/in/rheza-uyanto/

A data enthusiast who graduated from Data Science Bootcamp, Rakamin Academy. and also a registered pharmacist who graduated from Surabaya University since 2017, with 4+ years of experience in clinical pharmacy and quality data indicator in the hospital. Familiar with data preprocessing with Python, PostgreSQL and data visualization with Tableau. Interest with Data Analyst or Data Scientist for internship or fulltime position especially in the health-care setting. With this portofolio, I as a Data Scientist Intern in ID/X Partner, participate in project from lending company. We need to predict the credit risk from dataset that gave from the company.

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

As my final project from my probation as Data Scientist Intern in ID/X Partner, I participate in project from one lending company. I will collaborate with another department for provide the technology solution for them. I help to build he model to predict the credit risk form customer with company dataset such as : lending history from approved lending or rejected lending.

I also prepare the deck for visualize some insight. I make sure that our client understand with my data visualization and get some insights from my end-to-end recommendations based on my Programming Language and Framework/Methodology Data Science.

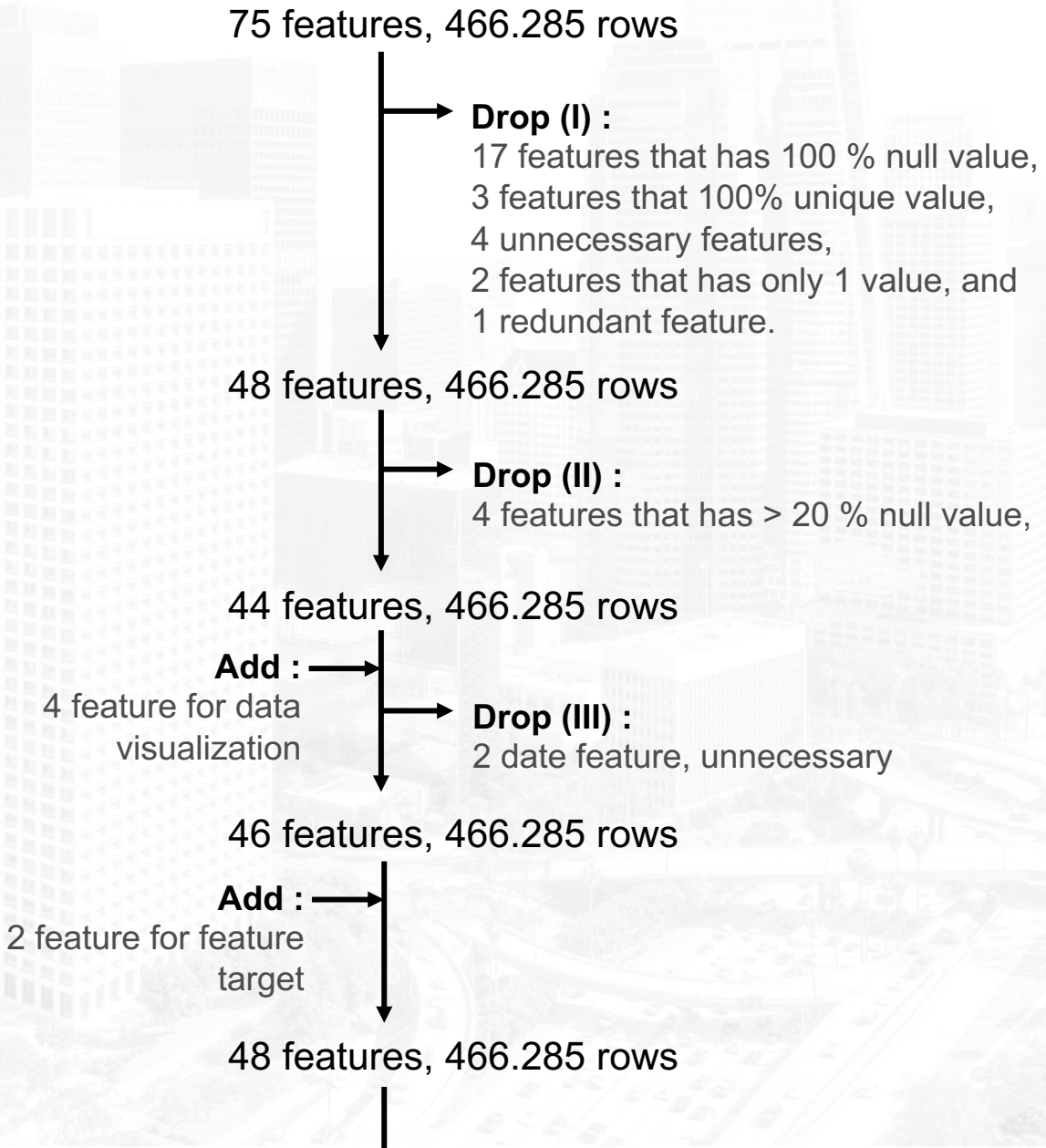
Source code : click in here

Data Preprocessing

Data set : loan_data_2007_2014.csv
Columns : 74 columns, Row : 466.285 rows

We do :

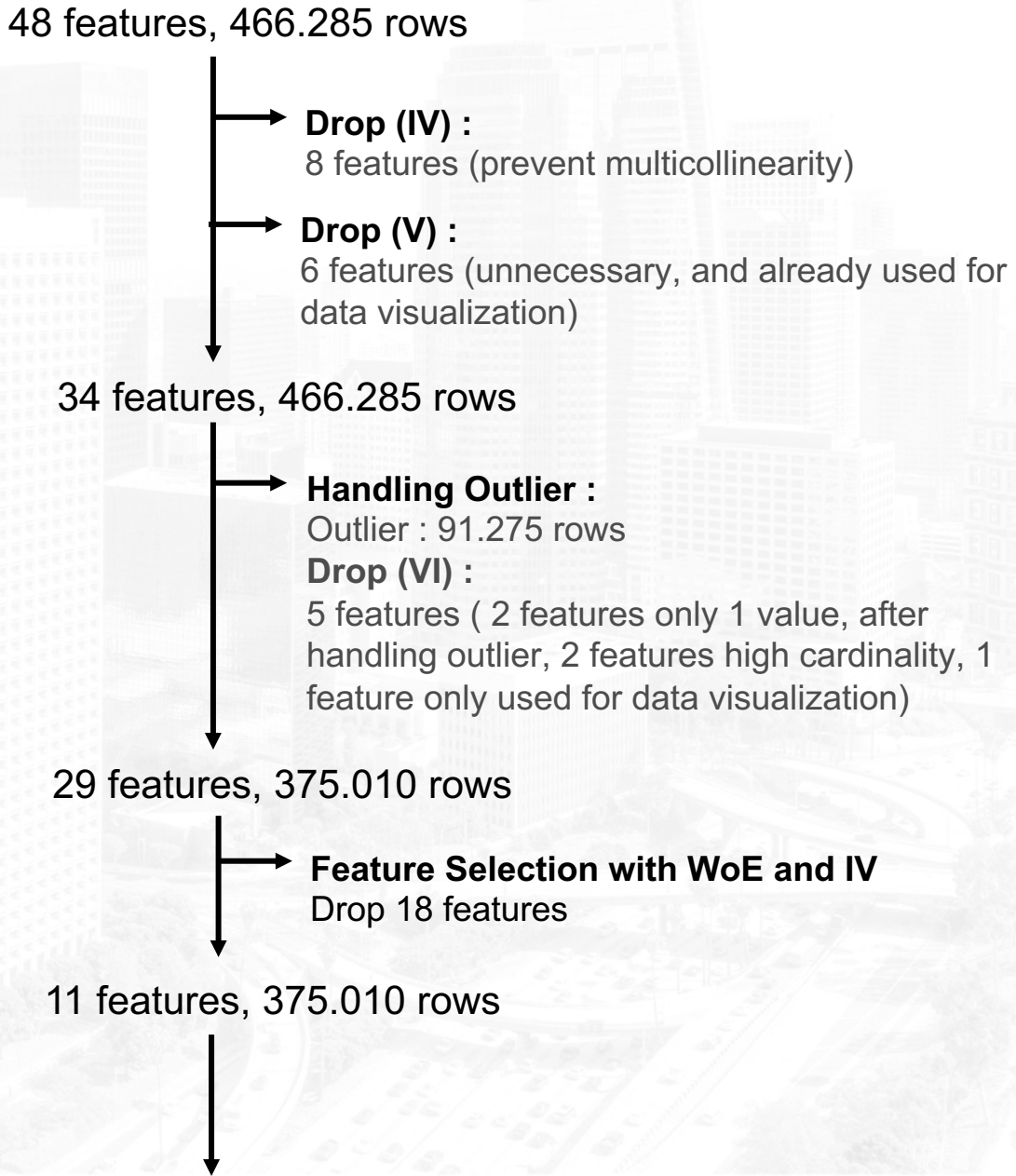
1. Drop some features (I)
2. Duplicate Checking : NO duplicate
3. Drop some features (II)
4. Make 4 feature engineering for date, with feature : issue_d and last_pymnt_d, drop 2 feature date that unnecessary (III), Date Feature engineering for data visualization.
5. Make Univariate Analysis
6. Feature Engineering for Loan_status (become a target feature (add 2 features), and home_ownership
7. Multivariate Analysis



Data Preprocessing

Data set : loan_data_2007_2014.csv
Columns : 74 columns, Row : 466.285 rows

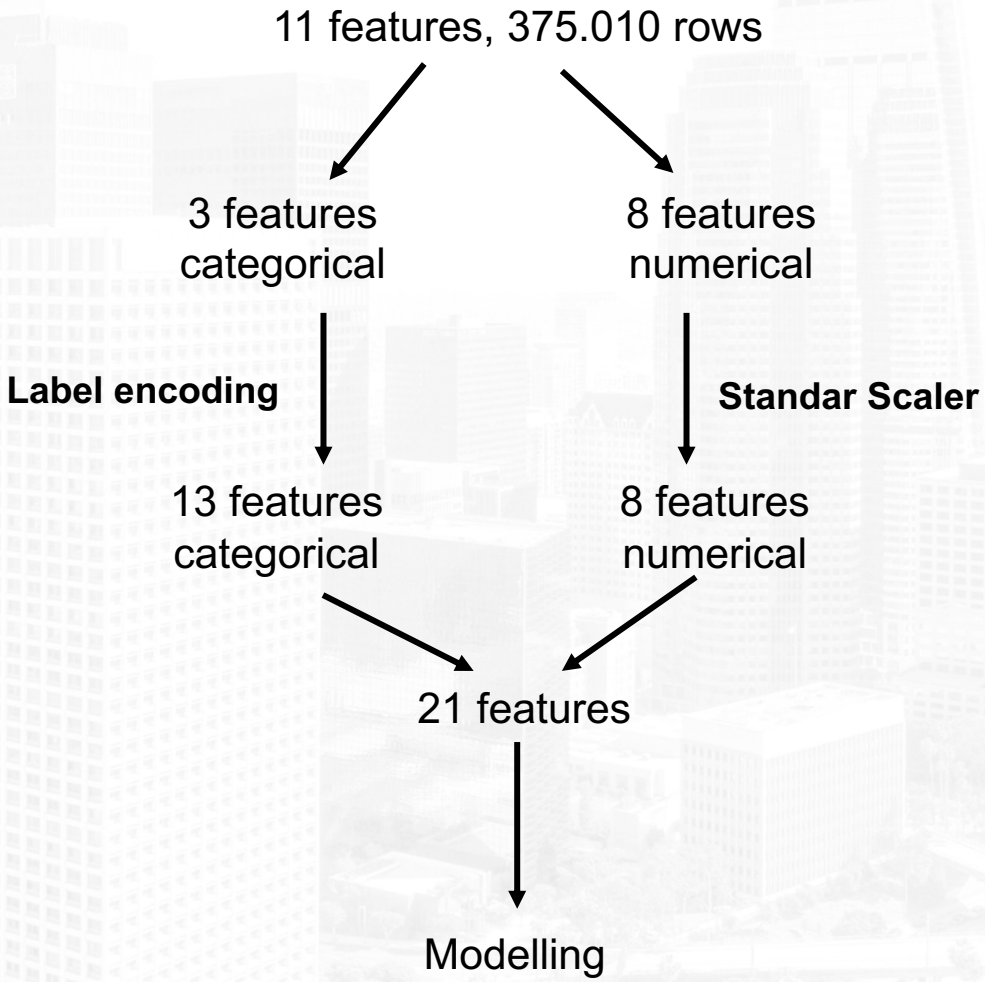
8. Drop some feature to prevent the multicollinearity (IV).
9. Drop Feature 6 features (V): 3 unnecessary, 3 feature already done for datavisualization
10. Handling Null Value : for data categoric (fill with mode), for data numeric (fill with median)
11. Feature Enginering : term (from string to numeric)
12. Handling Outlier (with Z score)
13. Dropping Feature (VI) after Handling outlier
14. Feature Selection with Weight of Evidence and Information value, only select feature that IV 0,2-0,5



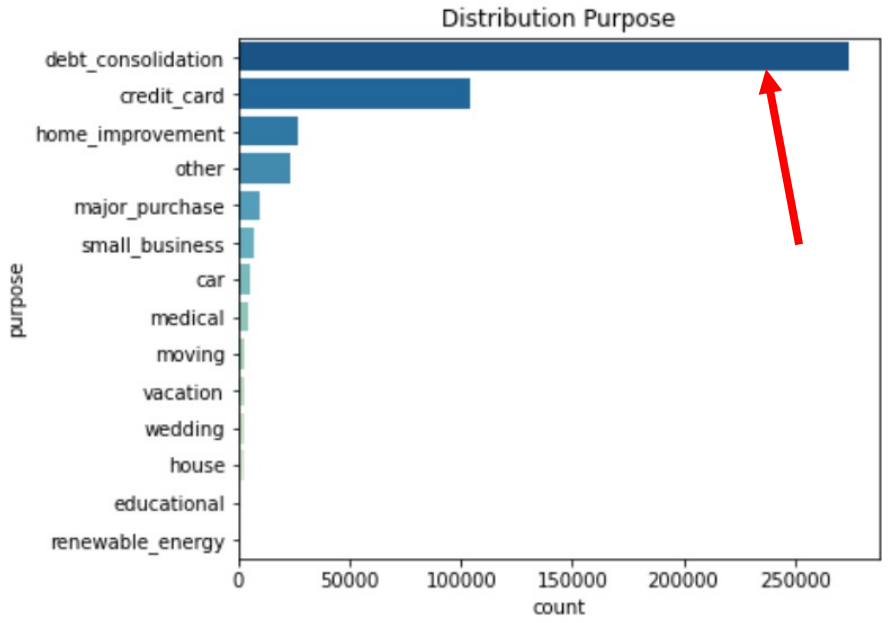
Data Preprocessing

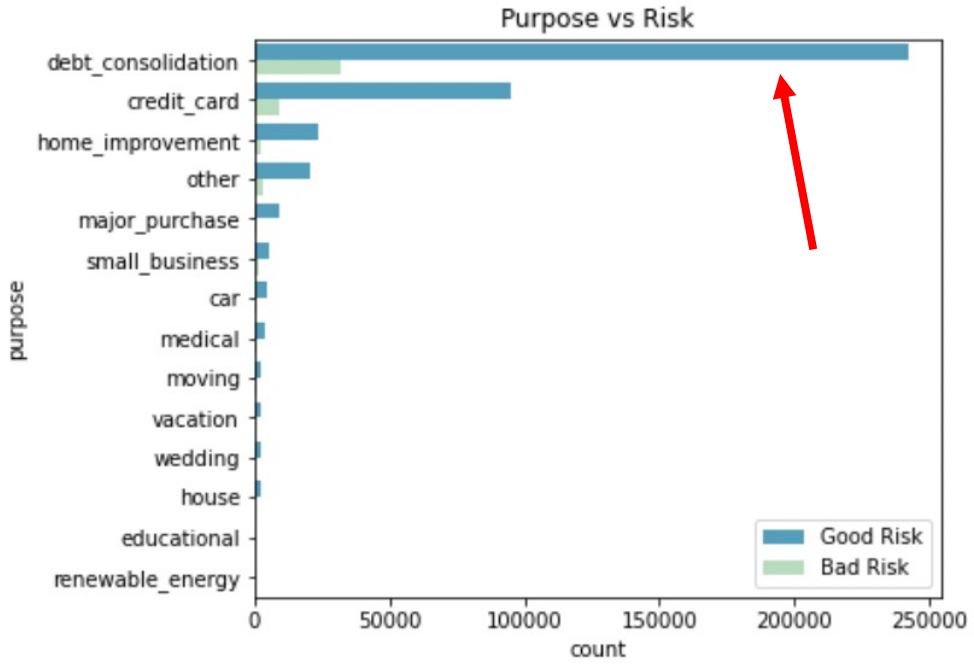
Data set : loan_data_2007_2014.csv
Columns : 74 columns, Row : 466.285 rows

15. 3 features categorical, handle for label encoding became 13 features.
16. 8 data numerical should be scaler
17. Modelling

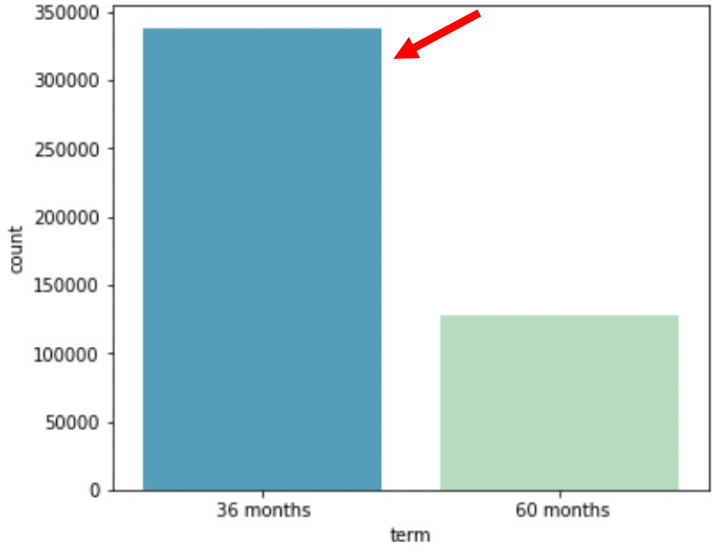


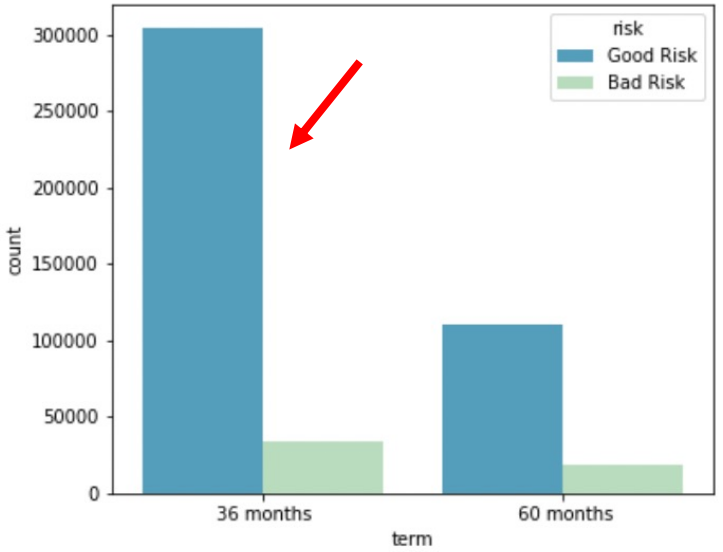
Analysis Univariate - Multivariate





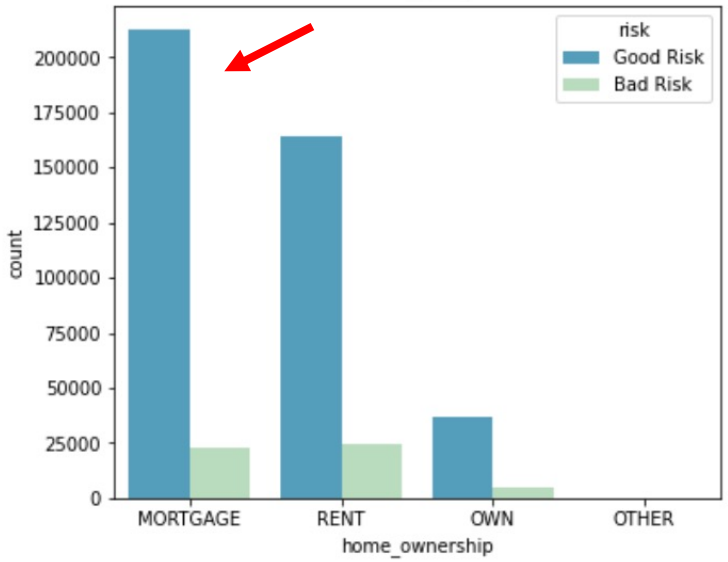
Purpose :
the high reason for credit loan is **debt consolidation**, and the good risk risk dominate the risk for debt consolidation.



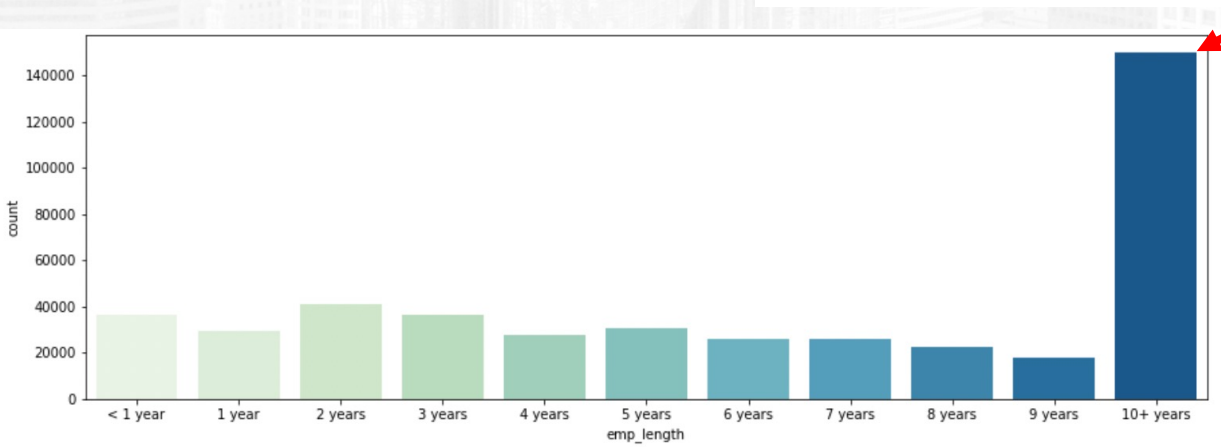


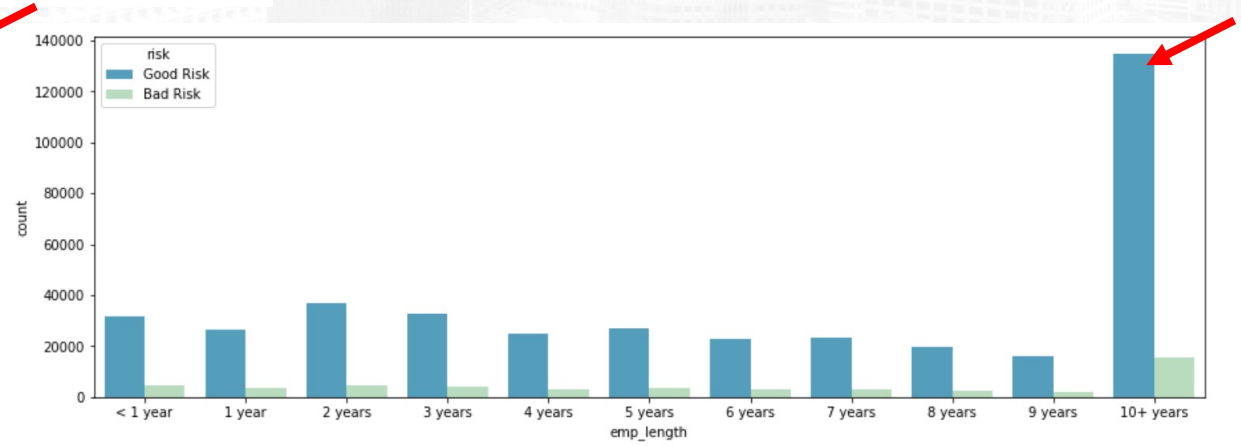
Term :
Most of borrowers used **36-months-term** for their lending.
And dominate with good risk category.

Analysis Univariate - Multivariate



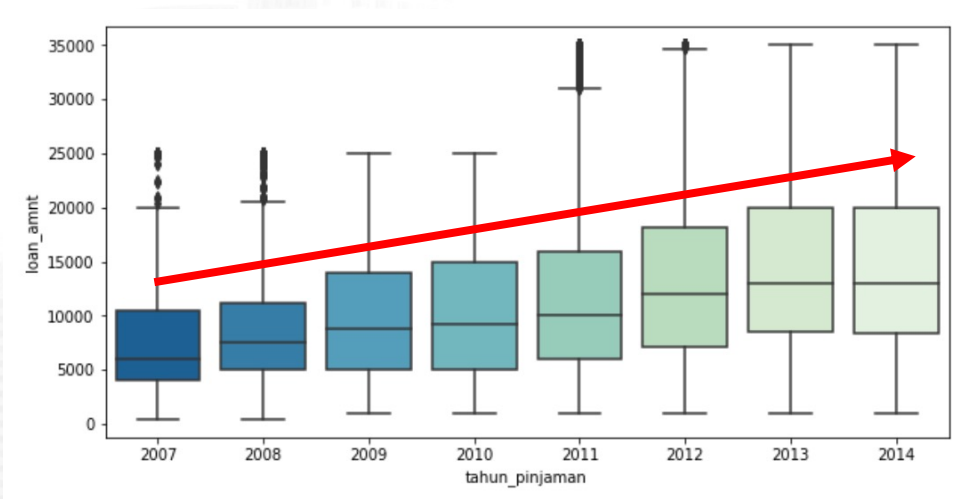
Home_ownership :
the highest home ownership is mortgage (jaminan), and many of them have a good risk for loan.

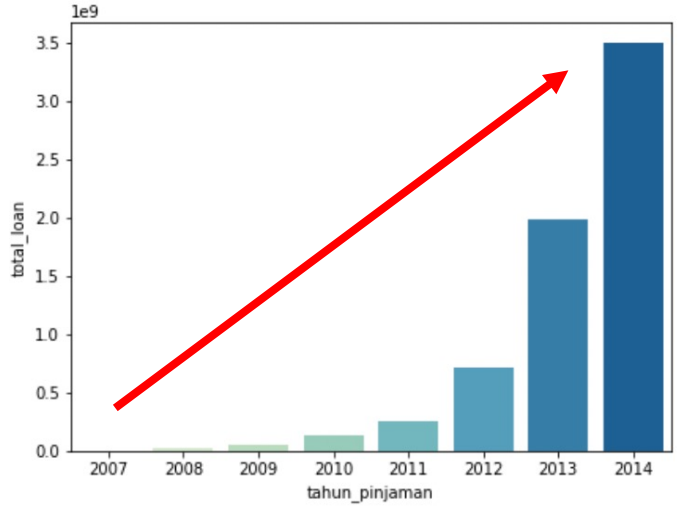


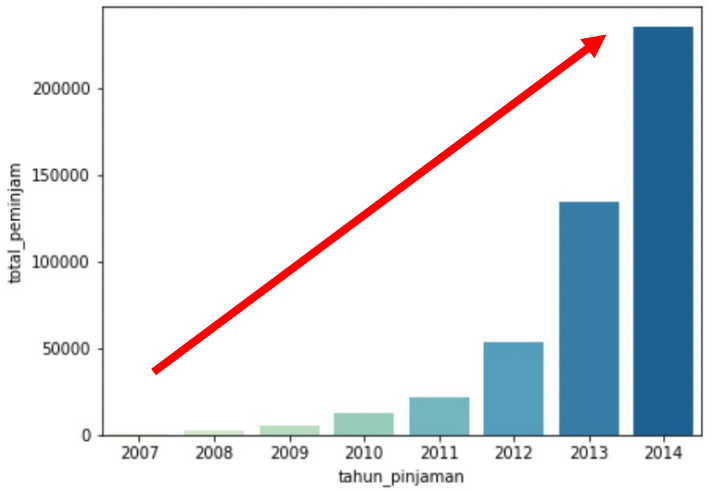


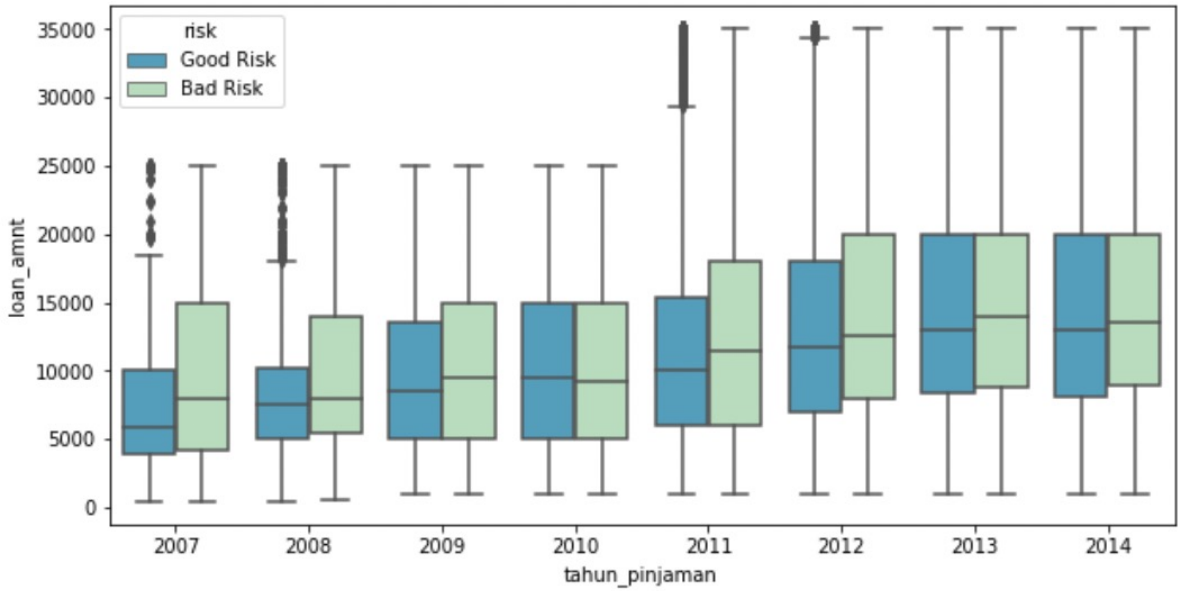
Term :
Most of borrowers have worked for more than **10 years** and most of them have a good-risk loan

Analysis Univariate - Multivariate





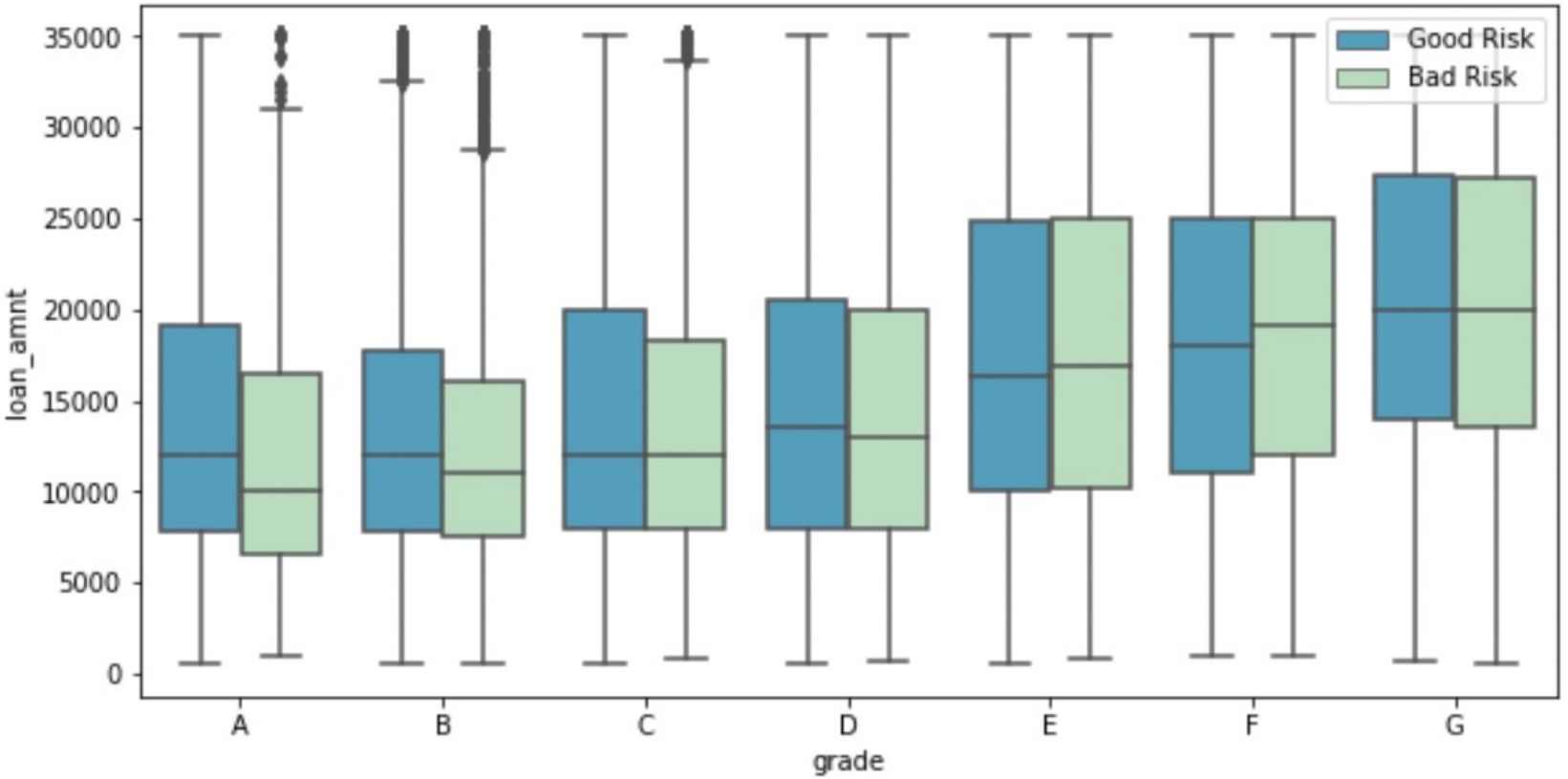




The total of loan per year increased gradually from 2007 to 2014, the number of person who loan also increased significantly.

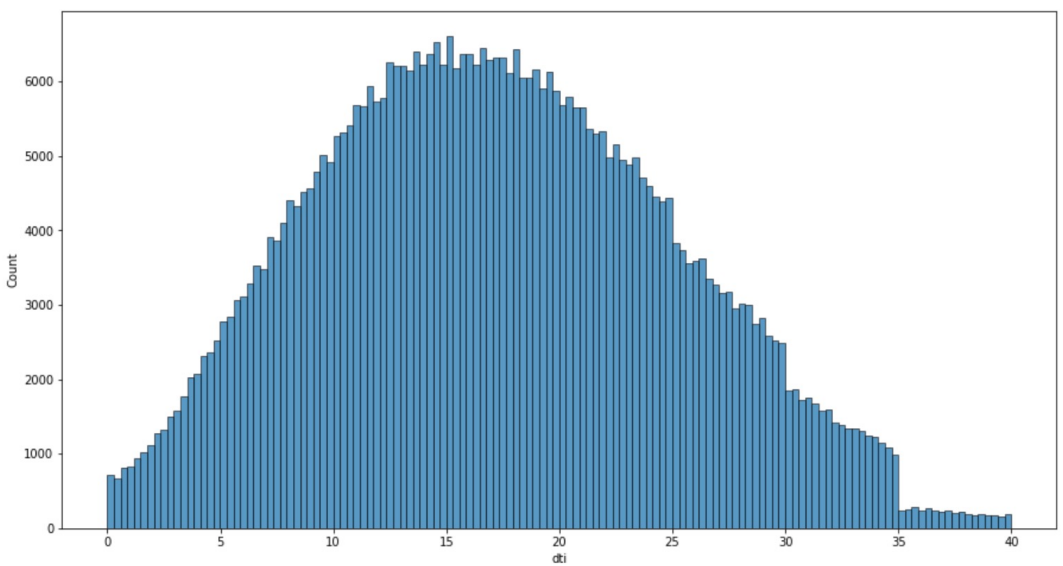
If we compare the total loan between good risk and bad risk over the year, the maximum loan amount from 2007-2010 is 25.000, and the median of bad risk is high than good risk. But from 2011-2014, the maximum loan amount is increased to 35.000, and we don't see any difference between bad and good risk for median loan amount.

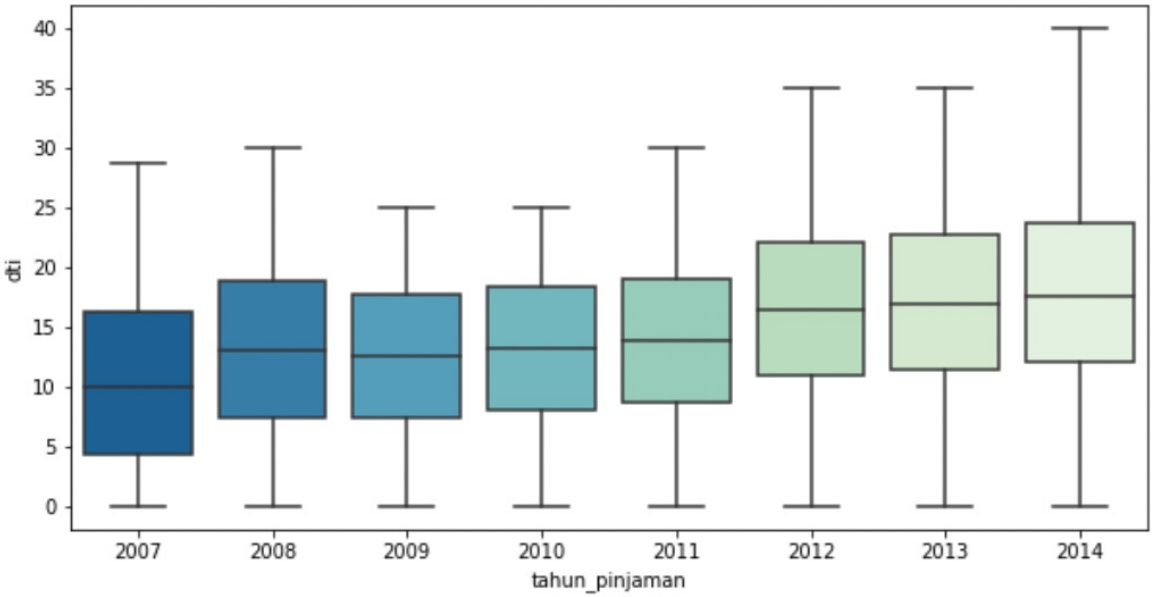
Analysis Univariate - Multivariate

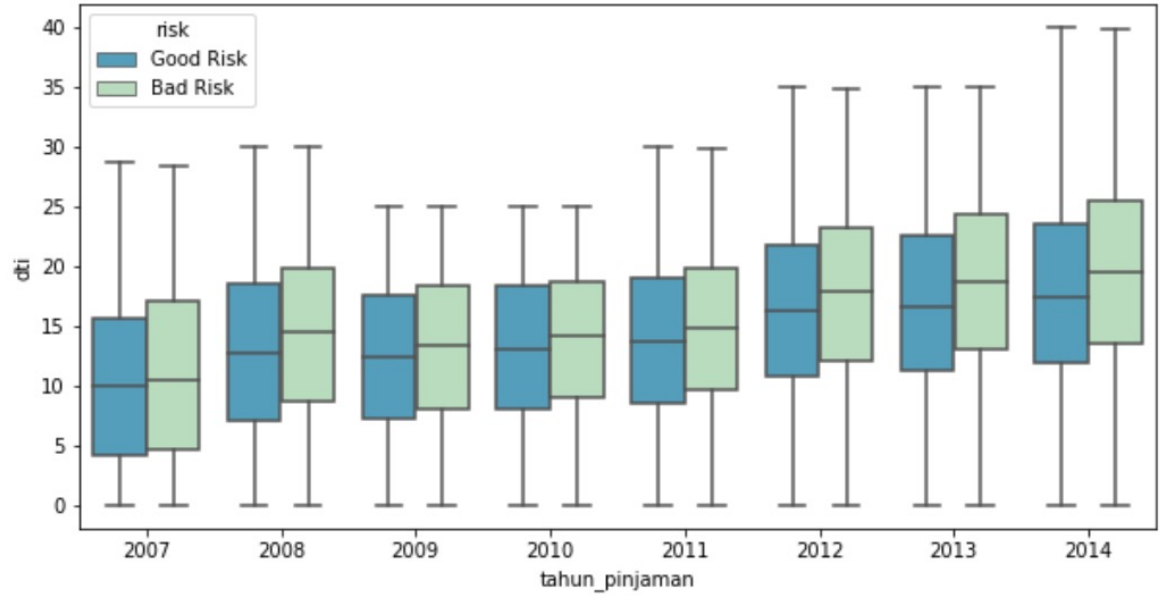


The relation of Loan Amount, Grade, and the Risk
From Grade A to Grade G, the median of total amount increase gradually.
The Good Risk have a median loan bigger than the bad risk in grade A to D, but grade E until G, they don't have any differences.

Analysis Univariate - Multivariate





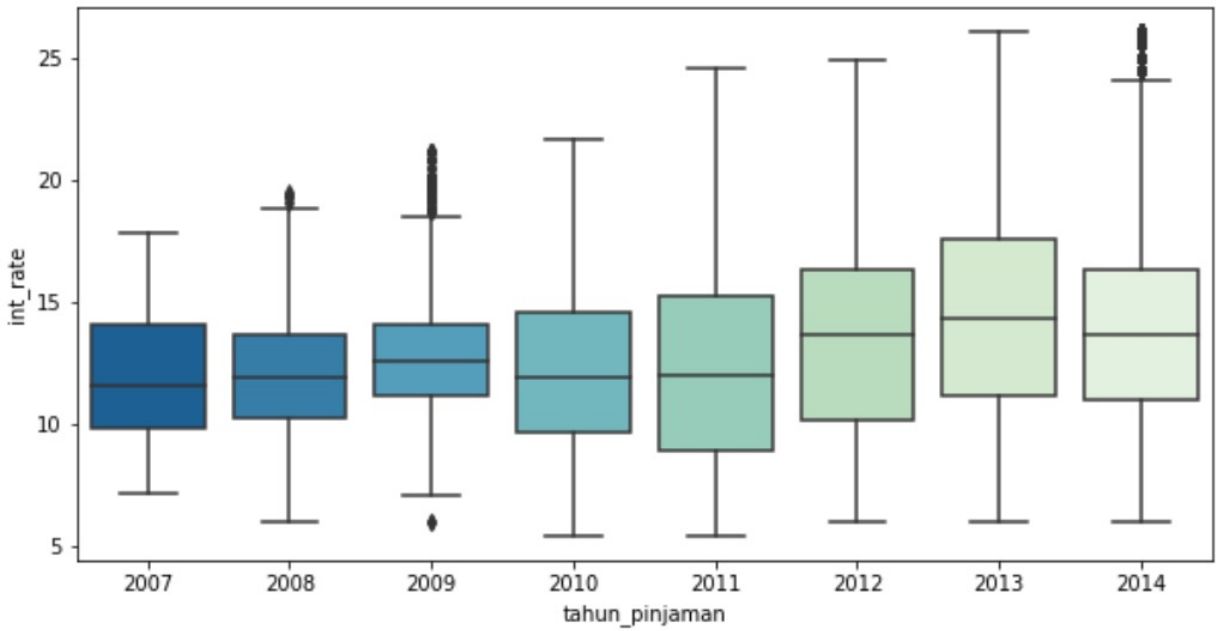


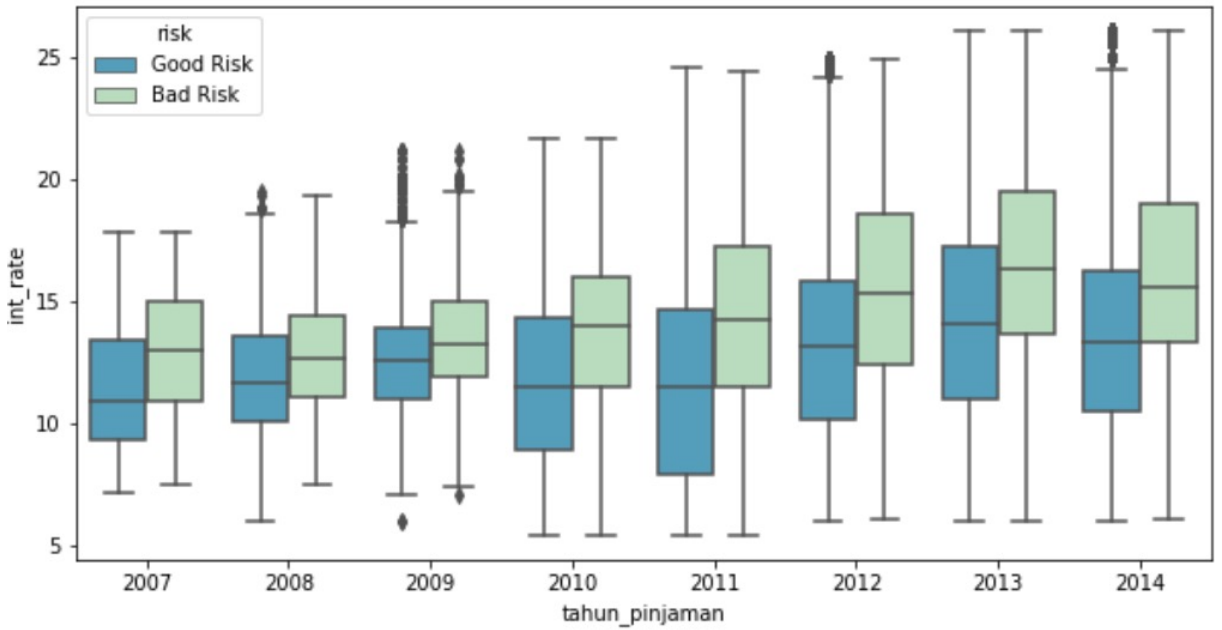
The debt-to-income (DTI) ratio is the percentage of your gross monthly income that goes to paying your monthly debt payments and is used by lenders to determine your borrowing risk.

In this case DTI value is 0-40, with median 16,87.

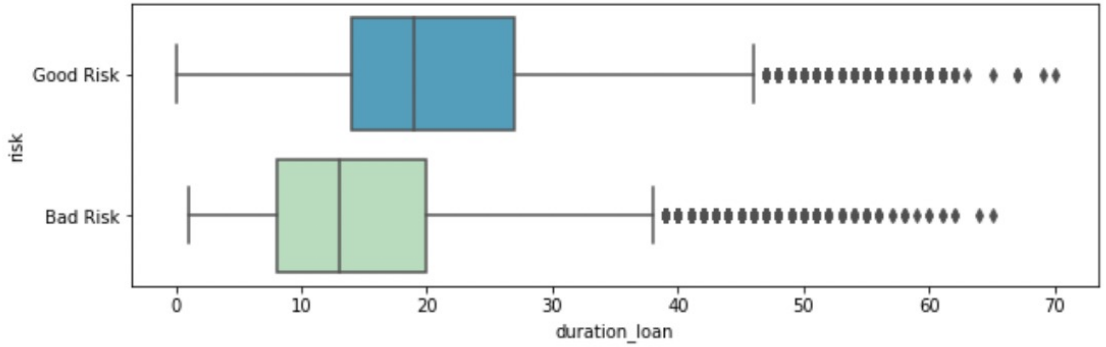
From 2007-2014, DTI increase significantly, and the dti from bad risk is higher than good risk by the year.

Analysis Univariate - Multivariate





The interest rate is similar from 2007-2014, from 10 – 15 % , but the bad risk get higher interest rate than good risk borrower.



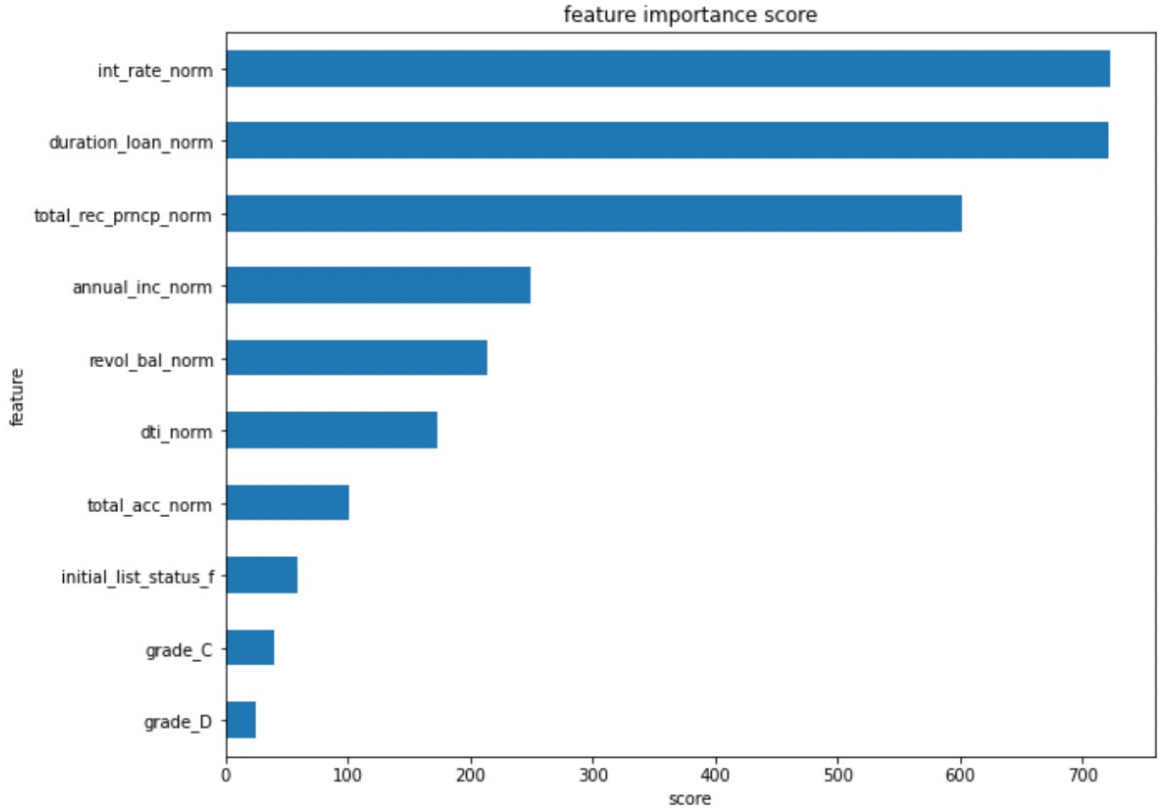
If we compare the risk and duration loan, The good risk people trend loan longer than bad risk people.

Modelling

Modelling	Accuracy (Train)	Accuracy (Test)	Precision	Recall	F1 Score	AUC
Kneighbors Classifier	0,93	0,92	0,71	0,32	0,44	0,76
XGB Classifer	0,93	0,94	0,89	0,40	0,55	0,87
Random Forest Classifier	0,99	0,94	0,90	0,44	0,59	0,89
Gradient Boosting Classifier	0,93	0,93	0,89	0,40	0,55	0,87
LGBM Classifier	**0,94**	**0,94**	**0,89**	**0,48**	**0,62**	**0,92**

Modelling (Under sampling)	Accuracy (Train)	Accuracy (Test)	Precision	Recall	F1 Score	AUC
Kneighbors Classifier	0,82	0,73	0,22	0,73	0,34	0,81
XGB Classifer	0,78	0,81	0,30	0,74	0,42	0,88
Random Forest Classifier	1,00	0,84	0,35	0,75	0,47	0,89
Gradient Boosting Classifier	0,78	0,80	0,29	0,75	0,42	0,88
LGBM Classifier	0,83	0,86	0,39	0,78	0,52	0,92

Recommendation




```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 375010 entries, 0 to 466284
Data columns (total 21 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   int_rate                375010 non-null  float64
 1   annual_inc              375010 non-null  float64
 2   dti                     375010 non-null  float64
 3   revol_bal               375010 non-null  int64
 4   total_acc               375010 non-null  float64
 5   total_rec_prncp         375010 non-null  float64
 6   duration_loan           375010 non-null  float64
 7   risk1                   375010 non-null  int64
 8   grade_A                 375010 non-null  uint8
 9   grade_B                 375010 non-null  uint8
 10  grade_C                 375010 non-null  uint8
 11  grade_D                 375010 non-null  uint8
 12  grade_E                 375010 non-null  uint8
 13  grade_F                 375010 non-null  uint8
 14  grade_G                 375010 non-null  uint8
 15  home_ownership_MORTGAGE 375010 non-null  uint8
 16  home_ownership_OTHER    375010 non-null  uint8
 17  home_ownership_OWN      375010 non-null  uint8
 18  home_ownership_RENT     375010 non-null  uint8
 19  initial_list_status_f   375010 non-null  uint8
 20  initial_list_status_w   375010 non-null  uint8
dtypes: float64(6), int64(2), uint8(13)
memory usage: 38.5 MB
```

The best model is LGBM Classifier, and based on feature importance, some feature are importance such as : interest rate, duration loan and total_rec_prncp. Some recommendation should be considered :

- The interest rate indicated the most importance feature, based on analysis, the interest rate for bad rate is higher than the good rate. The client should consider for calculate the lower interest rate but reliable.
- For duration loan, the good risk people trend loan longer than bad risk people.