CLASS ASSIGNMENT:
CREDIT RISK ANALYSIS AND MODELLING

HANOI, 2023

Table of Contents

1. Introduction

Predicting the repayment ability or probability of default of clients is important for financial institutions to minimize the risk of loan payment default. The purpose of credit risk analysis is to evaluate the creditworthiness of the borrower and predict the potential credit risk associated with the loan application. In this analysis, the author used the dataset "credit_risk_dataset" on Kaggle, which includes customers' information such as financial strength, repayment capacity and loan information. Then, the most identifiable features will be selected and machine learning models such as Logistic Regression, KNN, Random Forest will be applied to predict the probability of default of the borrower.

2. Theoretical Background

The financial credit risk indicates the risk associated with financing, in other words, a borrower cannot pay the lenders, or goes into default. This research is based on quantitative research approach with the application of machine learning to predict probability of default. Descriptive research was used to collect detailed information, while analytical research was used to analyze phenomenon or trend. Approaches applied to evaluate the creditworthiness of applicants could be the traditional statistical methods or advanced machine learning methods. Lately, machine learning techniques are studied extensively with more attentions. These techniques include artificial neural networks (ANNs), fuzzy set theory (FST), decision trees (DTs), case-based reasoning (CBR), support vector machines (SVMs), rough set theory (RST), genetic programming (GP), hybrid learning, and ensemble computing among others. In this research, the author focuses on the state-of-the-art approaches to credit risk assessment which are Logistic Regression, SVM, Random Forest, Decision Tree and KNN

3. Data

The dataset contains 32581 examples with 12 features breaking into 2 categories:

Demography:

Table 3.1. Demography column description

	Column	Data type	Description	Example
1	person_age	int	Customers' age	21, 22, 25
2	person_income	int	Customers' income	59000, 5600

3	person_emp_length	float	Years of employment	1.0, 4.0, 8.0
4	person_home_ownership	object	Does the customer own house?	Rent, own, mortgage

Credit loan

Table 3.2. Credit loan column description

	Column	Data type	Description	Example
1	loan_intent	object	Intention to make a loan	Personal, education
2	loan_grade	object	Grade of loan	B, C, D
3	loan_int_rate	float	Interest rate	16.2, 11.14
4	loan_amt	float	Loan amount	35000, 1000
5	cb_person_default_on_file	int	Historical default	Y, N
6	loan_percent_income	float	Percentage of loan over income	0.1, 0.59
7	cb_person_cred_hist_length	int	Credit length	2,3,4

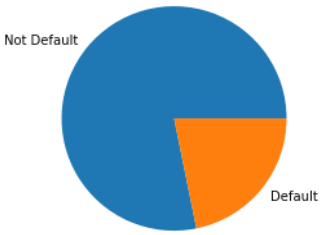
Target variable

Table 3.3. Target variable

	Column	Data type	Description	Example
1	loan_status	object	Loan status (1: default, 0: not default	1,0

3.1. Data Exploring

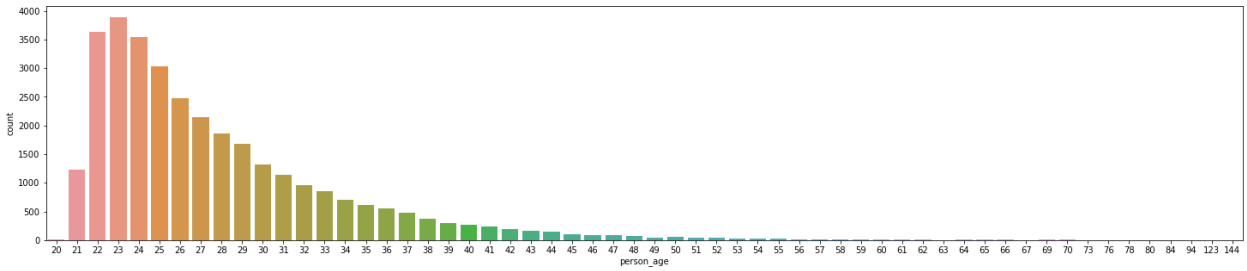
3.1.1. Check data imbalance



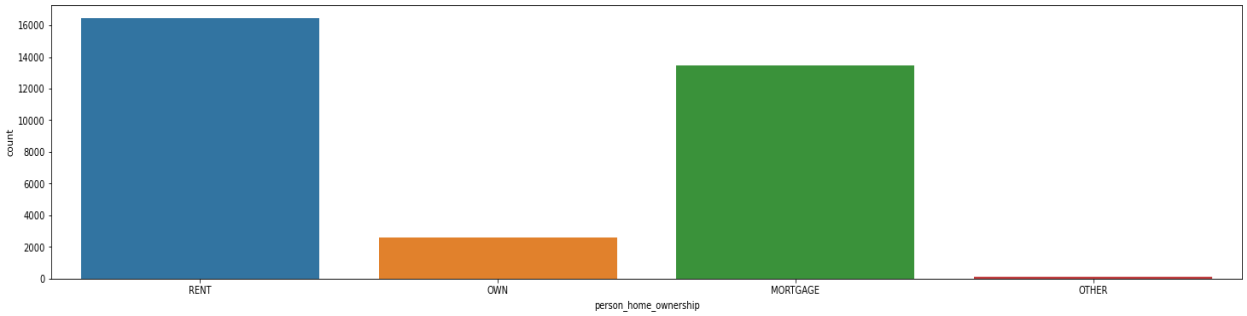
The dataset is imbalanced. It contains about 22% of values belong to the Default class and over 78% belong to Not Default. However, this imbalance percentage is acceptable.

3.1.2. Categorical and numeric variables

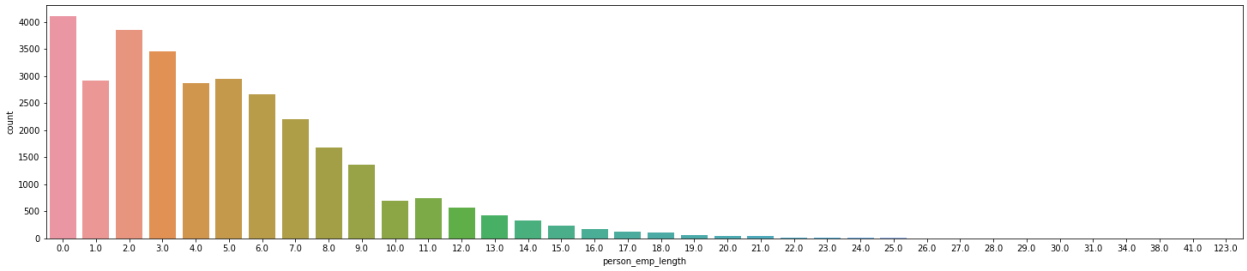
Age



Home ownership

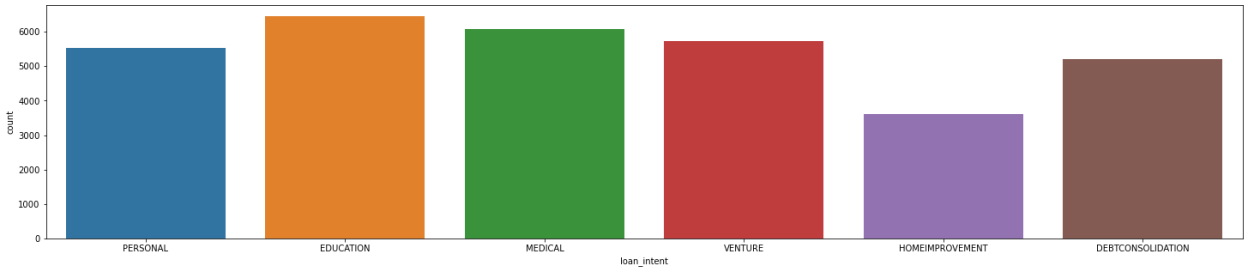


Years of employment



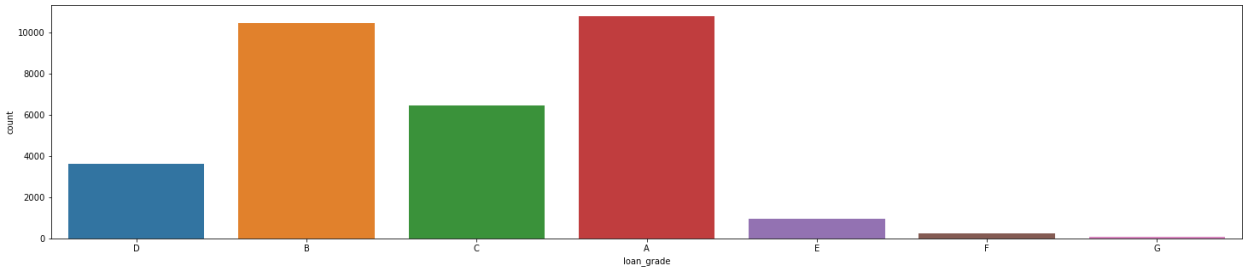
Most customers who borrow money are young, ranging from 20 to 30, and the number of borrowers gradually decreases as they are older. This is aligned with years of employment, which mostly hovers around 0 to 9 years. Many of them don't own a house but rent or mortgage.

Intention of borrowing



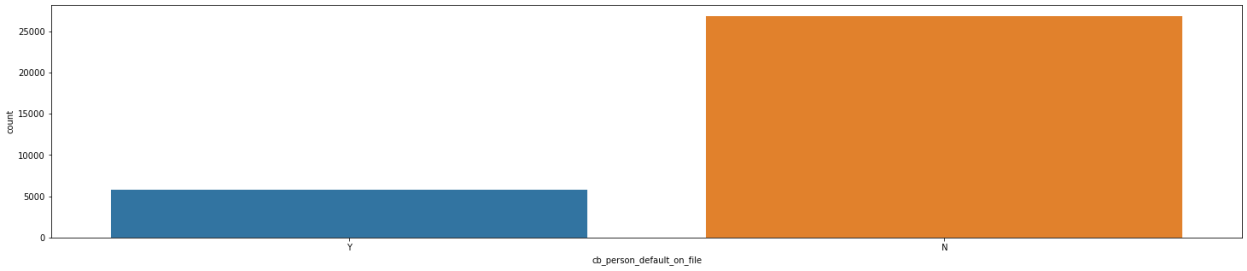
There are 6 types of loan intentions. Generally, difference between categories is small. Education is the most frequent reason one borrowing money. Other reasons: Personal, Medical, Debt Consolidation and Venture are also common intention while Home Improvement is the least to be seen.

Loan grade

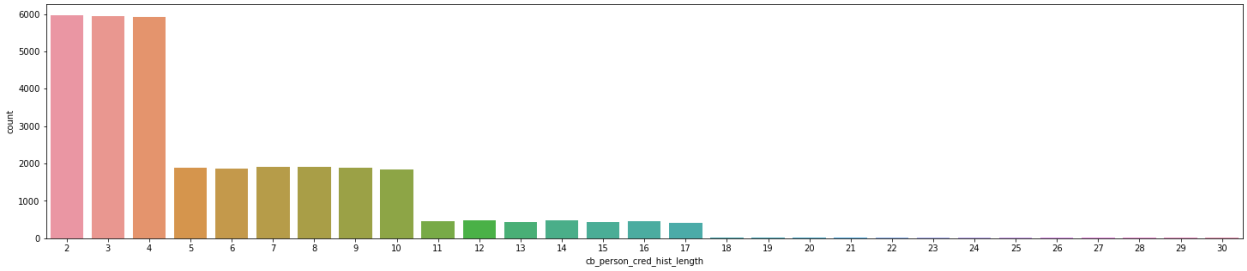


Loans are mostly graded A and B, followed by loans graded C and D. The remaining grades account for a little percentage of the total. This indicates a large proportion of customers in the dataset are highly graded.

Historical credit default

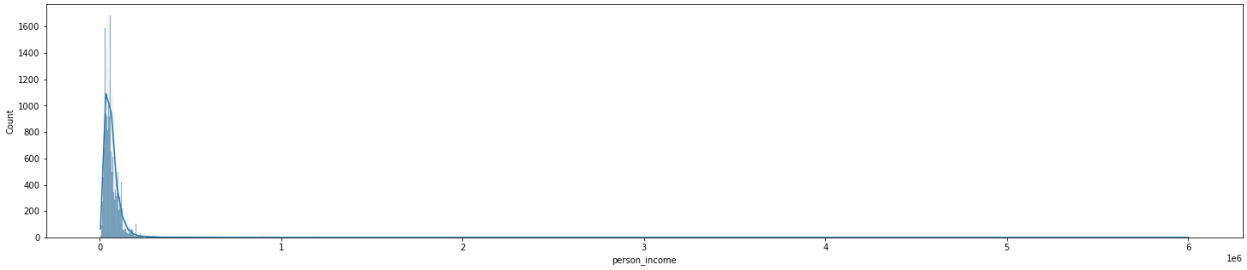


Historical credit length



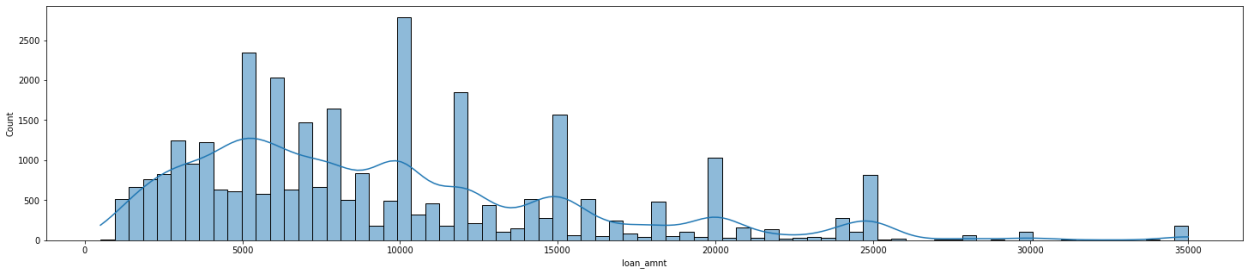
As can be seen, historical credit length could be divided into 3 groups. The first group of which credit terms last from 2 to 4 years has equally highest frequency with roughly 6000 counts. Meanwhile, the frequency of the second group with credit terms lasting from 5 to 10 years stands at just nearly a third of the first group. The last group takes up a very tiny proportion compared to the two formers.

Income



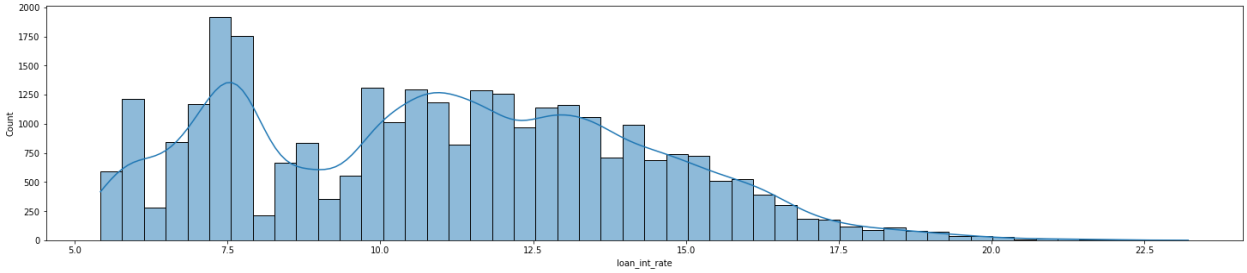
Income is the feature that shows obviously an enormous number of outliers. Most of the clients have income of 60000 and the average income in the data is approximately 66074. In the next preprocessing data step, this variable needs to be normalized.

Loan amount



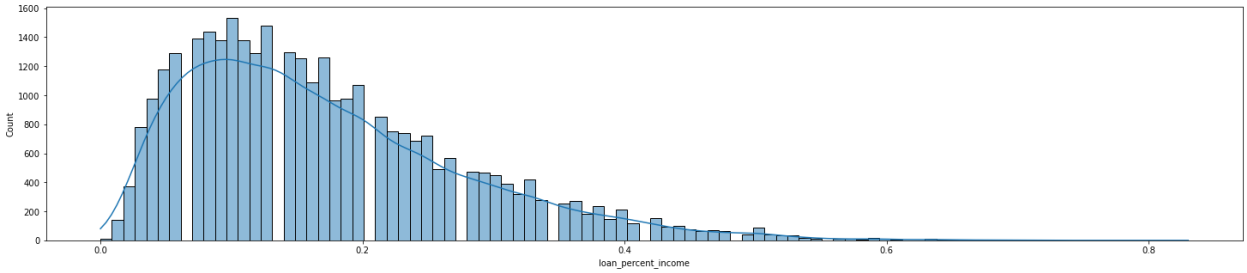
The average loan amount accounts for about 9589.37, which is over one-seventh of the average income. The most frequent loan amount is approximately 10000, followed by 5000. Outliers with a maximum of 35000 is observed.

Loan interest rate



The average interest rate is over 11%. Most customers borrow with an interest rate of around 7.5%. Maximum interest rate can be seen at 23.22%

Debt to Income ratio



The average Debt to Income ratio is 0.17 and is mostly seen at 0.1. This indicates people tend to borrow an amount of money which equals to 10% of their income. The maximum DTI can be found at 83%.

3.1.3. Check missing values

There are two features that have null values: years of employment (‘person_emp_length’) and loan interest rate (‘loan_int_rate’):

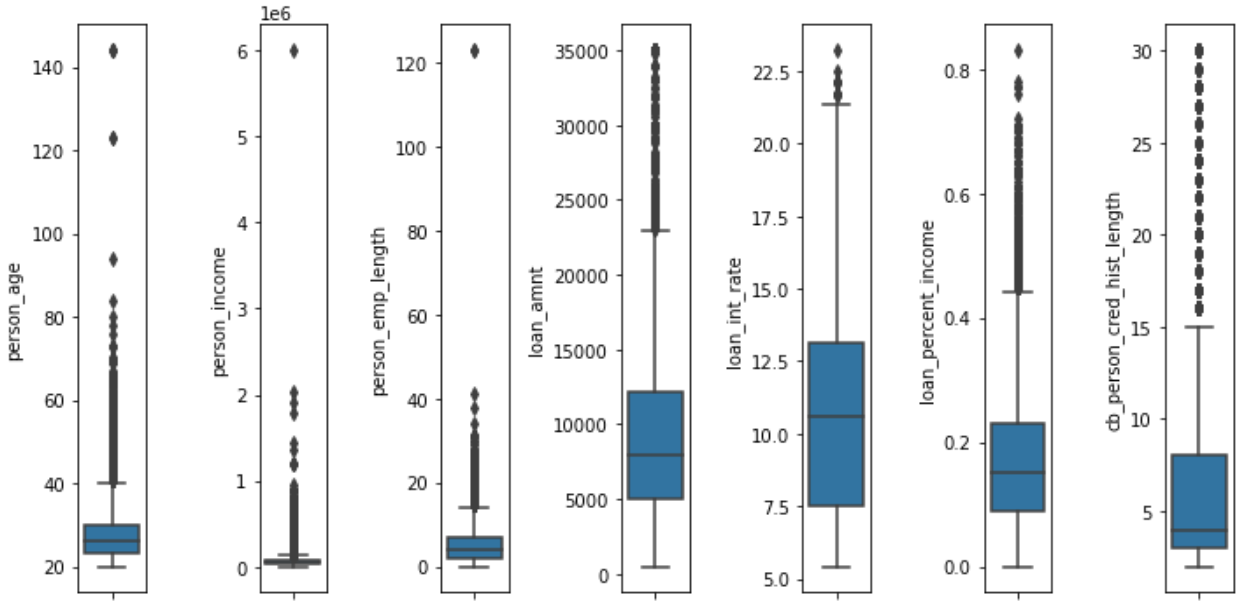
	Count	Percent
person_emp_length	895	2.747
loan_int_rate	3116	9.5639

Loan interest rate is one of the important features when analyzing credit risk so I will have more attention to this one.

3.1.4. Check duplicated rows

There are 165 duplicated rows in the data. The decision is to drop all these rows.

3.1.5. Check outliers



All features contain outlier values.

Age ('person_age') and years of employment ('person_emp_length') have outliers > 100 years, which is not reasonable. Hence, in the next data preprocessing, values > 100 in these two columns will be removed.

3.2. Data Preprocessing:

3.2.1. Drop duplicated rows

3.2.2. Fill missing values

- Loan interest rate: Since loan interest rate ('loan_int_rate') have outliers so filling null values by taking mean of interest rate may not reflect the true mean value. Hence, interest rate null values are filled by mode.
- Employment length: Employment length of null customers is filled by employment length of customers with the same age. That's because typically people of the same age tend to have the same year of employment.

3.2.3. Drop outlier values

- Age: drop age > 100
- Employment length: drop year of employment > 100

3.2.4. Change datatype

- person_emp_length: from “float” to “int”

3.2.5. Encode

- Label encoder: historical loan default (‘cb_person_default_on_file’)
- Ordinal encoder: Loan grade (“loan_grade”)
- Dummy encoder for KNN model: categorical variables
- WOE encoder for Logistic model

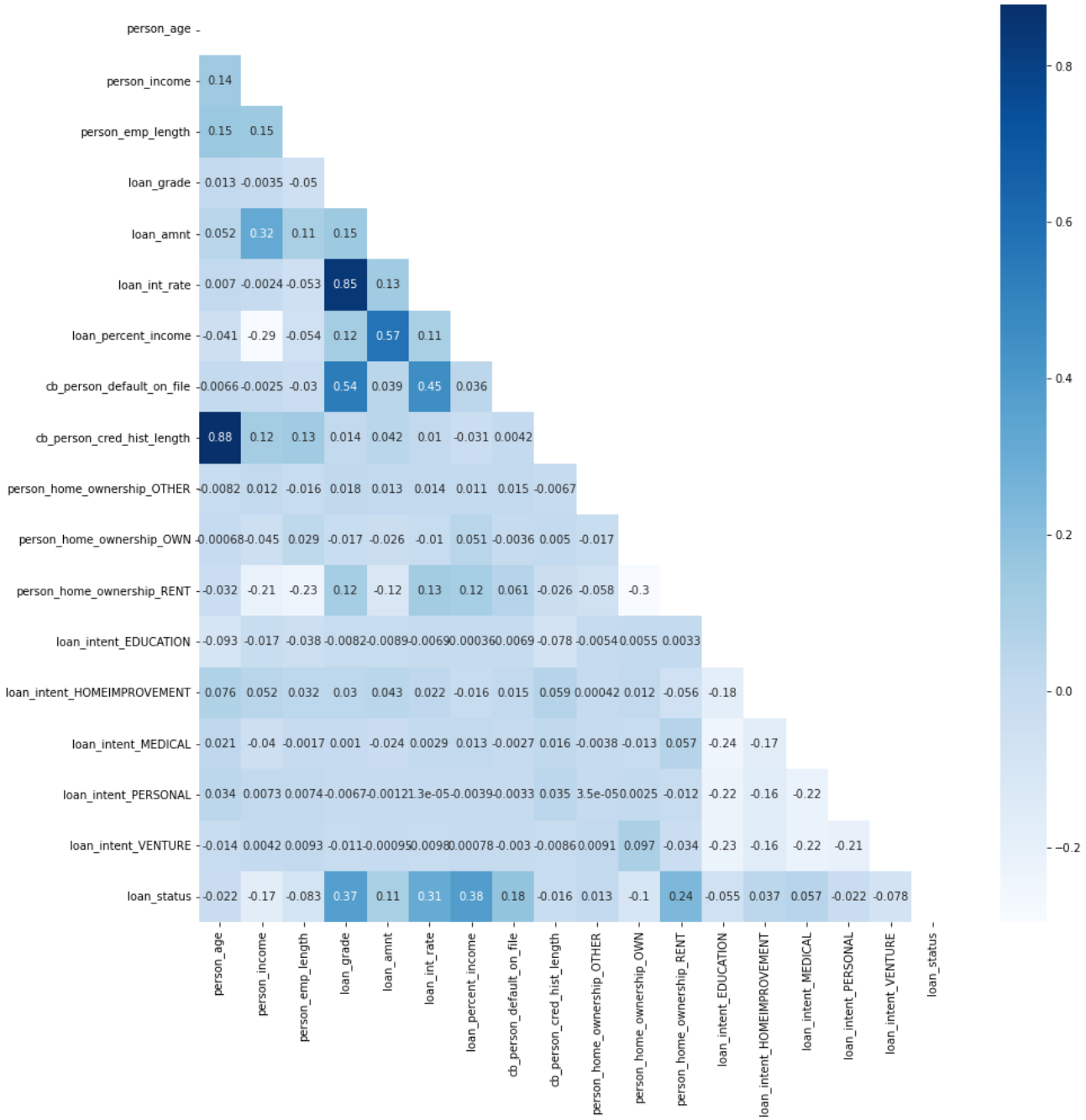
4. Feature Selection

4.1. VIF score

	Feature	VIF
1	person_age	42.241885
2	loan_int_rate	31.302362
3	loan_percent_income	9.639560
4	cb_person_cred_hist_length	8.765606
5	loan_amnt	8.651179
6	loan_grade	7.685429
7	person_income	4.541630
8	person_emp_length	2.645242
9	person_home_ownership_RENT	2.585721
10	loan_intent_EDUCATION	2.152026
11	loan_intent_MEDICAL	2.103197
12	loan_intent_VENTURE	2.068288
13	loan_intent_PERSONAL	2.008266
14	loan_status	1.876796
15	target	1.711369
16	loan_intent_HOMEIMPROVEMENT	1.678201
17	person_home_ownership_OWN	1.260909
18	person_home_ownership_OTHER	1.010294

From the table above, age (“person_age”), loan interest rate (“loan_int_rate”) have very high values of VIF, indicating that these two variables are highly correlated with others.

4.2. Correlation



The average correlation coefficient between target variable (“loan_status”) and others is 0.138

From the correlation heatmap above, we realize that age (“person_age”) has high correlation with historical credit length (“cb_person_cred_hist_length”) with correlation coefficient of 0.88. Loan interest rate (“loan_int_rate”) and loan grade (‘”loan_grade”) also have high correlation (0.85), indicating that these two variables are highly correlated. This is expected as the credit length increases as people become older and the lower grade of a loan (more risk), the higher interest rate is.

Based on VIF score and correlation coefficient, age (“person_age”) and loan interest rate (“loan_int_rate”) should be dropped in order to avoid multicollinearity in the model.

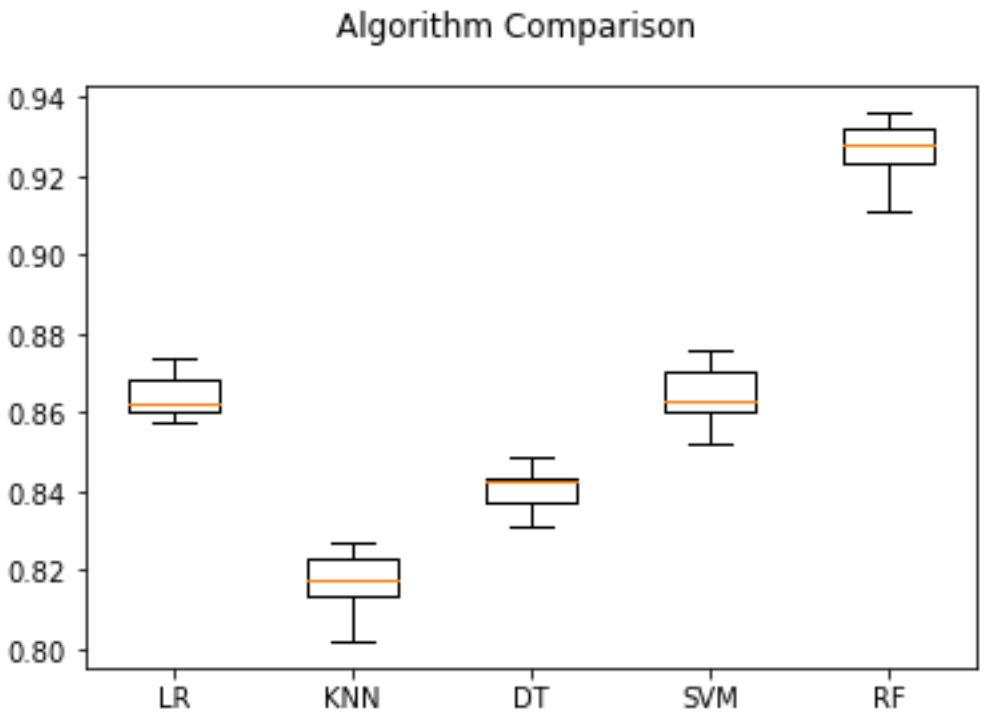
5. Results

Machine learning models which are used to classify default or not default are: Logistic Regression, SVM, Random Forest, Decision Tree, KNN.

Before training, numeric data: 'person_income', 'loan_amnt', 'cb_person_cred_hist_length' are normalized to the range of [0,1].

Train set and test set is splited by 70/30.

Using k-folds cross validation (k=10) on the train set, Random Forest gives the best mean accuracy of 0.927 (of 10 folds). Next comes Logistic Regression and SVM algorithm with mean accuracy of 0.864



From the results above, I chose to tune hyperparameter for 3 algorithms: Logistic Regression, SVM and Random Forest

5.1. Logistic Regression

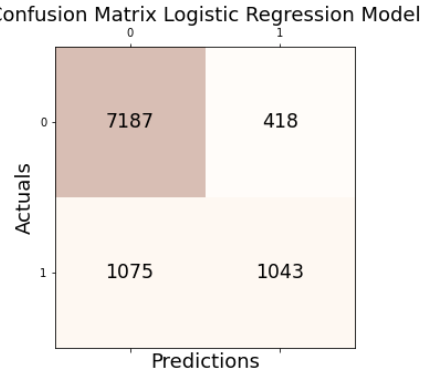
5.1.1. Coefficients

Top 5 coefficient of independent variables that have large weight on target variable

Variable	Coefficient
Debt to Income Ratio (“loan percent income”)	6.544
Loan amount (loan_amt)	-1.78
Income (“person_income”)	1.03
Own house (“person_home_ownership_OWN”)	-0.82
Loan_grade	0.46

The coefficients of independent variables are quite reasonable. The variable that influences the risk of default the most is Debt to Income ratio (loan_percent_income), which indicates the more customer borrows within a given income, the higher probability of default that customer has.

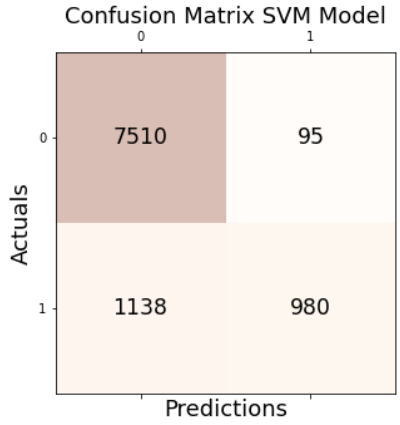
5.1.2. Confusion matrix and Classification Report




```
Classification report:
              precision    recall  f1-score   support

           0       0.87      0.95      0.91      7605
           1       0.71      0.49      0.58      2118

    accuracy                           0.85      9723
   macro avg       0.79      0.72      0.74      9723
weighted avg       0.84      0.85      0.84      9723


ROC_AUC score:
 0.8617697839314751
```

The number of predictions is 9723. According to roc-auc score, Linear Regression model is capable of distinguishing 86% between default or non-default. For class 1, recall is lower than precision (0.49 < 0.71), which means there are more FN1 (type error II, predict non-default for customers who actually default) than FP1 (type error I, predict default for customers who actually non-default).

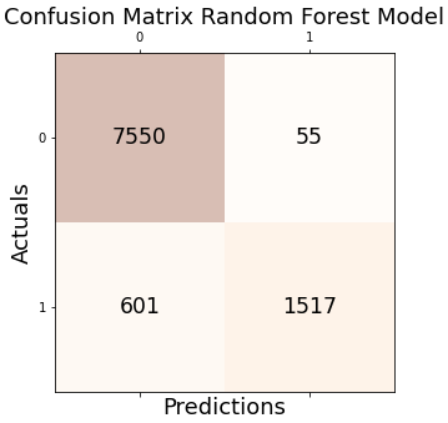
5.2. SVM




```
Classification report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.92      7605
           1       0.91      0.46      0.61      2118

    accuracy                           0.87      9723
   macro avg       0.89      0.73      0.77      9723
weighted avg       0.88      0.87      0.86      9723


ROC_AUC score:
 0.8710937650358002
```

From the result above, SVM model is capable of distinguishing 87.1% a bit higher than Linear Regression model. For class 1, recall is much lower than precision (0.46 < 0.91), which means the model produces more type error II (predict non-default for customers actually default) than type error I (predict default for customers actually non-default).

5.3. Random Forest




```
Classification report:
              precision    recall  f1-score   support

           0       0.93      0.99      0.96      7605
           1       0.97      0.72      0.82      2118

    accuracy                           0.93      9723
   macro avg       0.95      0.85      0.89      9723
weighted avg       0.93      0.93      0.93      9723


ROC_AUC score:
 0.927802145474841
```

According to roc-auc score, Random Forest model is capable of distinguishing 92.7%, higher than both LR and SVM. For class 1, recall is still lower than precision, but it is as high as 72% predicting default for customers who are actually default. This is by far the most accurate model in terms of accuracy score and recall score.

6. Conclusion

Random Forest model with the tuned hyperparameter: {'n_estimators': 2000, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 20, 'bootstrap': False} brings the highest accuracy.

From the Linear regression model result, Debt to Income ratio is the most influencing feature to customer's default, which also include Income and Loan amount variable. This is quite reasonable as the larger amount of money people borrow given an income, the more risk associated with them. Other significant factors are house owning and loan grade, which is appropriate since owning a house guarantees a higher quality of collateral. Hence, when deciding to lend to clients, financial institutions should take these factors above into consideration.