# Lending Club Loan Prediction

Capstone Project #2

# Introduction

The project goal is to train a classification model to predict bad loans on a major lending platform, Lending Club.

The typical lending process:

1. Applicants submit their loan applications to Lending Club
2. Individual lenders can directly browse and select loan applications that they want to fund.

Eventually, borrowers pay interests and principals back to lenders.

With this business model, Lending Club is considered P2P lending. There's still the risk of investors to run the risk of investing in a bad loan. This issue is to be addressed in this project by developing a predictive model to identify bad loans by using information available on loan applications. Then, investors can make more objective and data-driven assessment of loan applications to minimize risk.

We can download the dataset from the Lending Club website: https://www.lendingclub.com/info/statistics.action.

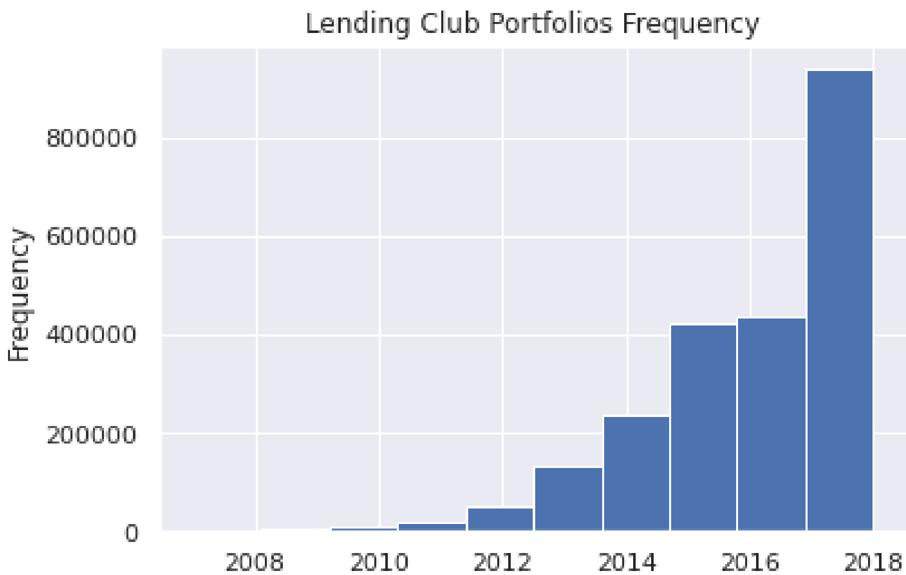
However, it requires signing up as a member to download the dataset. Thus, in this project, we will use a dataset available from Kaggle. The dataset was downloaded from : https://www.kaggle.com/wordsforthewise/lending-club?select=rejected_2007_to_2018Q4.csv.gz

Unfortunately, this was updated a year ago, so, it's not the most recent data.

# Exploratory Data Analysis

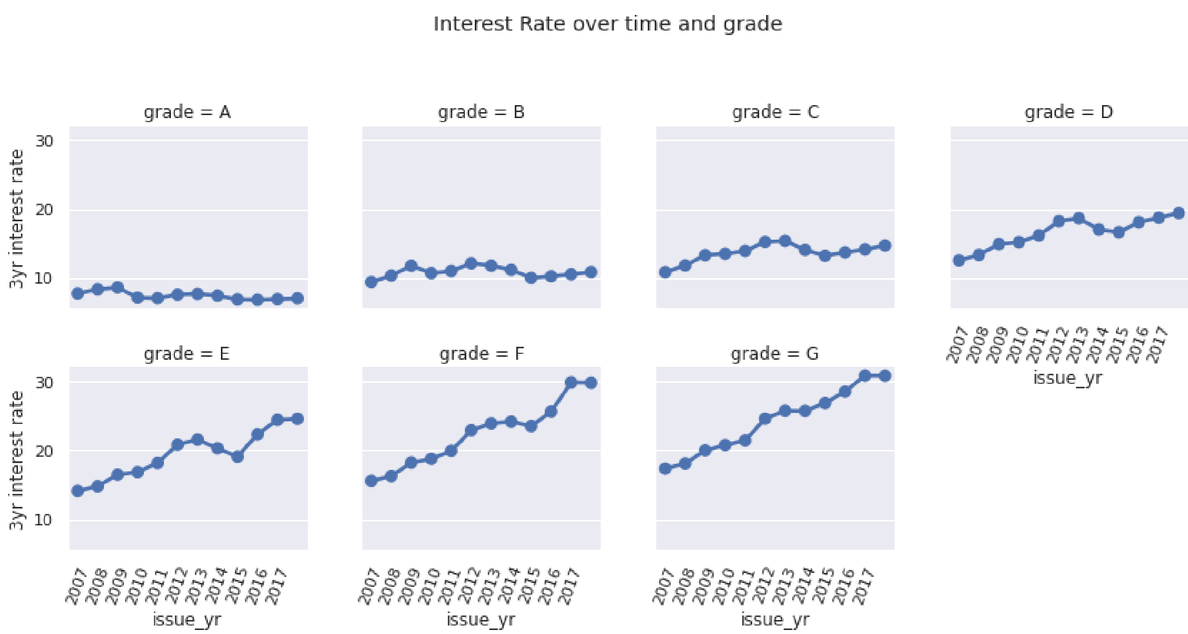
## Loan Application Frequency

Lending Club was launched back in 2012. Since then, the platform has gotten more exposure and popularity. Thus, we expected a significant increase in its loan portfolios over the years.




## Interest Rate by Grade

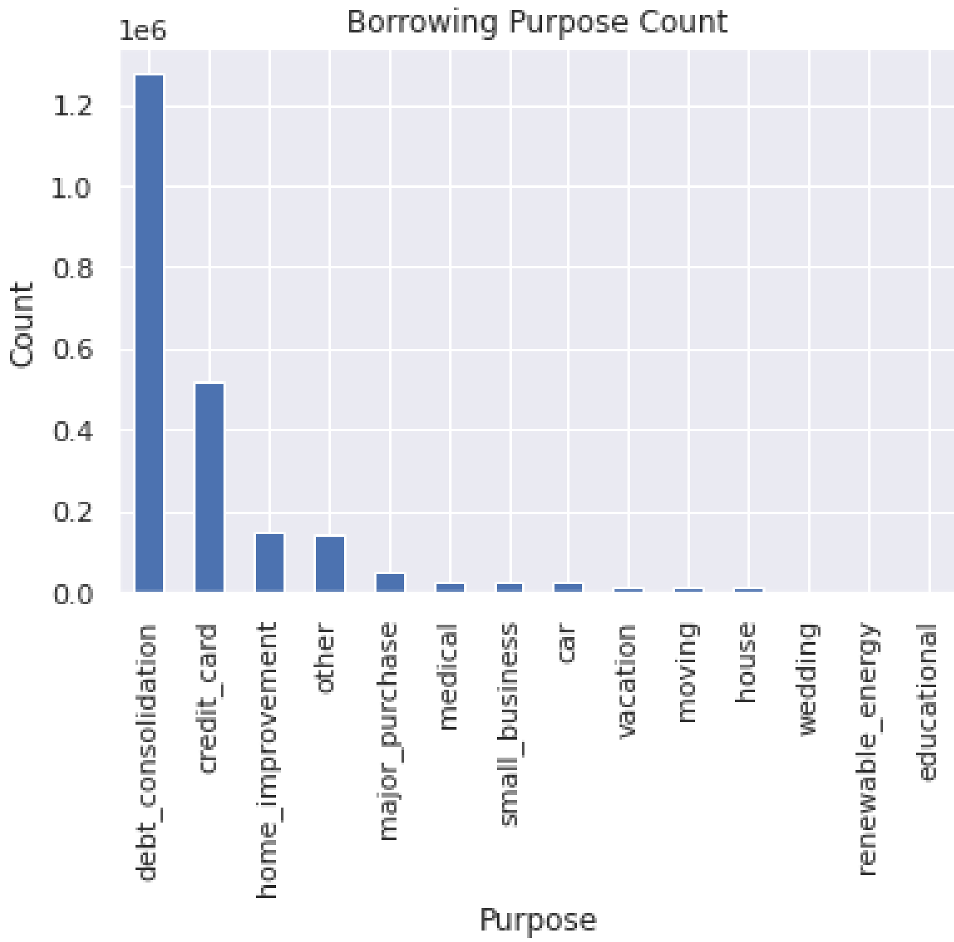
We can see that interest rates for grade D, E, F, G increase quickly from 2014.




We have very few data points from 2007 to 2014. Let's take a closer look to see if this increase in average rate is not due to the small number of observations.

| | int_rate | | | | | | |
|---|---|---|---|---|---|---|---|
| **grade** | **A** | **B** | **C** | **D** | **E** | **F** | **G** |
| **issue_yr** | | | | | | | |
| **2007.0** | 78 | 98 | 141 | 99 | 100 | 52 | 35 |
| **2008.0** | 318 | 594 | 580 | 419 | 285 | 111 | 86 |
| **2009.0** | 1203 | 1445 | 1348 | 817 | 308 | 105 | 55 |
| **2010.0** | 2567 | 2805 | 2070 | 1253 | 336 | 91 | 34 |
| **2011.0** | 5579 | 4722 | 2203 | 1261 | 272 | 54 | 10 |
| **2012.0** | 10753 | 16805 | 9902 | 5088 | 795 | 103 | 24 |
| **2013.0** | 17057 | 40313 | 24693 | 14505 | 3231 | 608 | 15 |
| **2014.0** | 35333 | 53460 | 44042 | 20510 | 7066 | 1980 | 179 |
| **2015.0** | 70132 | 91783 | 77457 | 32740 | 9450 | 1363 | 248 |
| **2016.0** | 66862 | 114783 | 92317 | 36707 | 9932 | 2364 | 530 |
| **2017.0** | 76300 | 108943 | 88425 | 34572 | 9837 | 1574 | 768 |
| **2018.0** | 122999 | 99155 | 76680 | 36228 | 9179 | 312 | 118 |

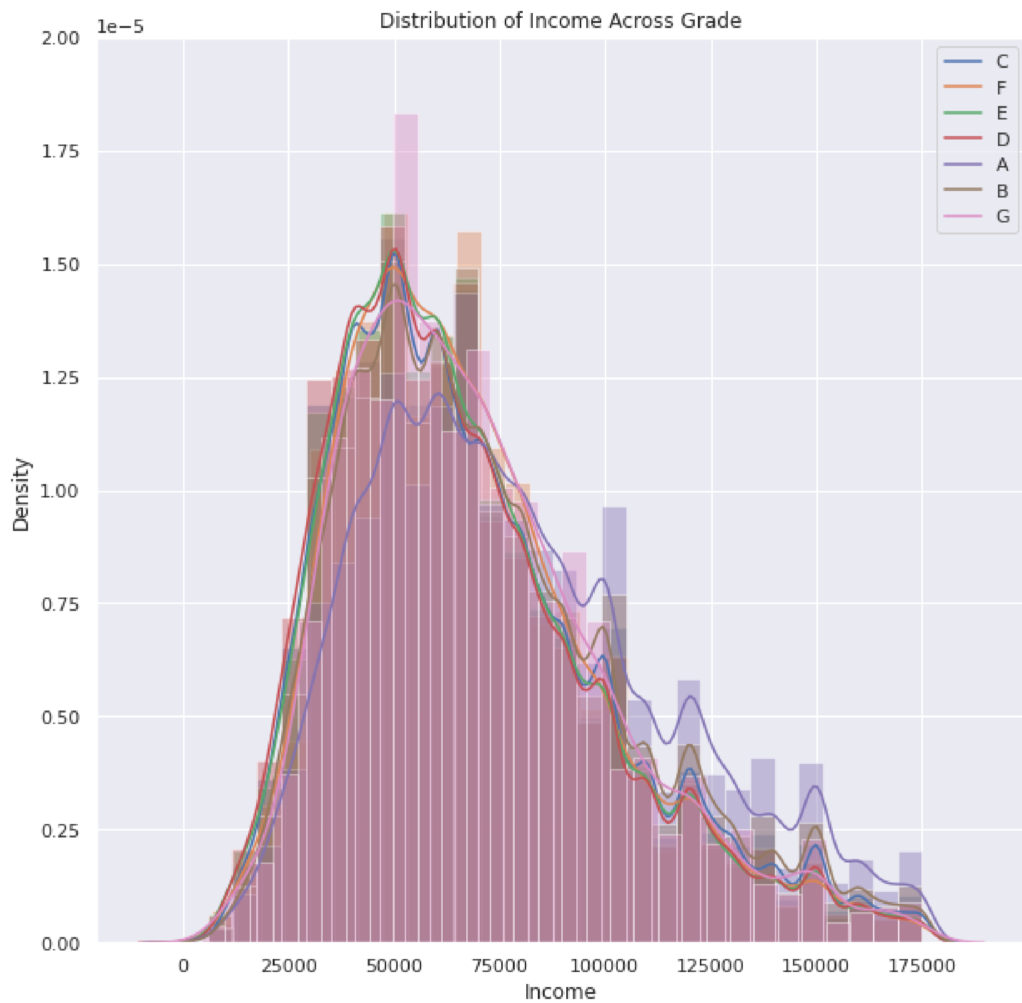
## Borrowing Purpose




Debt consolidation, credit card, and home improvement seems to be the top 3 borrowing purposes through Lending Club.

## Distribution of Income Across Grade

There seems to be a slight difference in income level across grade, with one grade has a high density of income above $100,000. Perhaps it may not be beneficial to dive too deep into it, except having a brief look at the median income across each grade. We do see that while other grades seem to have similar median incomes, grade A does stand out.




## Income by Professions

Because of the large number of observations that we have, we can construct a reliable distribution of annual incomes.

But first, further transformation:

1. Filtering for loans where the reported income is less than 1 million USD
2. Filtering for loans where Debt-to-Income ratio is less than 100 percent. If it is greater than or equal to 100, I wonder why we would have made such loans in the first place. I could have been more careful by capping dti value at 100.

Changing employment years into numeric.

1. Filling unknown values for home-ownership as Rent
2. Standardizing the values of employment title (emp_title)

Because people will most likely lie on their incomes when their income is low, we can filter out for data if:

1. Income is lower than $70,000 but has been verified by Lending Club2
2. Income is higher than $70,000 but lower than $120,000
3. Income is higher than $120,000 but has been verified by Lending Club

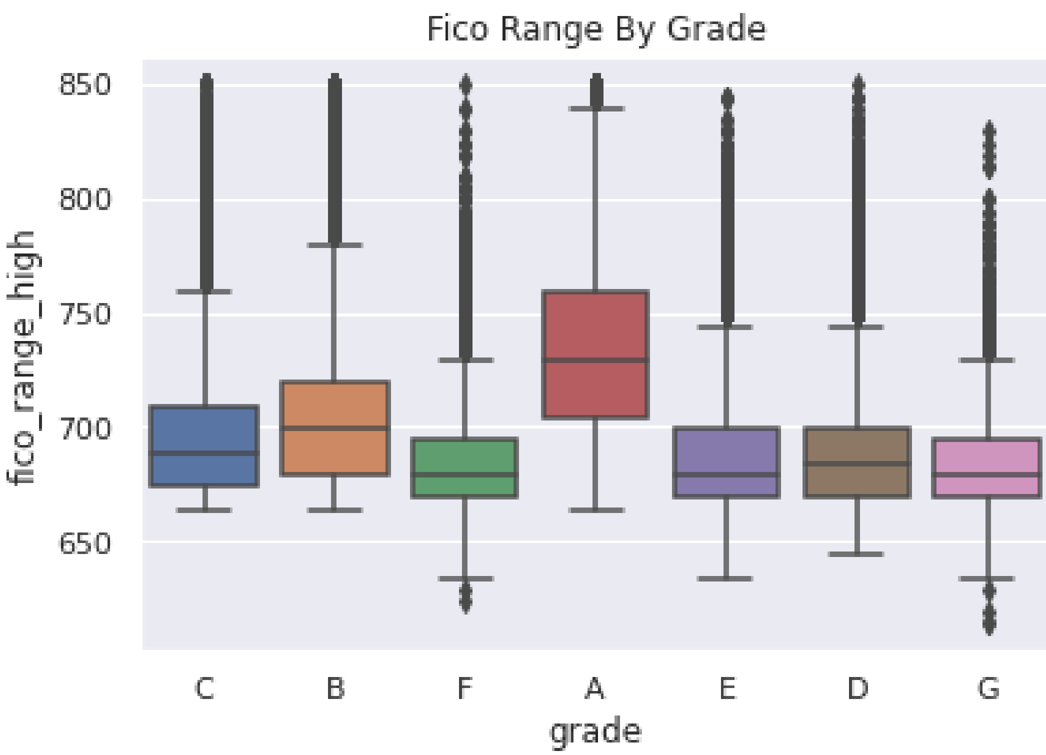
The choice of limit of $70,000 and $120,000 is arbitrary to filter out loans where income levels seem unrealistic.

| | **min** | **mean** | **median** | **max** | **count** |
|---|---|---|---|---|---|
| **emp_title** | | | | | |
| **other** | 0 | 55,211 | 50,000 | 3,330,432 | 98333 |
| **nurse** | 10,000 | 84,118 | 81,000 | 320,000 | 29515 |
| **teacher** | 0 | 74,662 | 75,000 | 367,500 | 22437 |
| **manager** | 2,500 | 84,846 | 82,000 | 4,800,000 | 22218 |
| **owner** | 0 | 93,432 | 85,000 | 1,000,000 | 13998 |
| **driver** | 0 | 74,401 | 75,000 | 1,000,000 | 10264 |
| **supervisor** | 500 | 76,947 | 75,340 | 780,000 | 9738 |
| **sales** | 0 | 87,748 | 80,000 | 2,000,000 | 8722 |
| **project manager** | 0 | 92,144 | 90,000 | 460,000 | 8616 |
| **general manager** | 20,000 | 93,502 | 84,500 | 700,000 | 6739 |
| **engineer** | 2,439 | 113,053 | 91,000 | 110,000,000 | 5973 |
| **truck driver** | 14,000 | 73,688 | 75,000 | 790,000 | 5680 |
| **office manager** | 0 | 66,101 | 70,000 | 300,000 | 5288 |
| **director** | 0 | 123,013 | 102,000 | 1,000,000 | 5090 |
| **president** | 7,956 | 129,495 | 100,000 | 3,000,000 | 4730 |
| **operations manager** | 15,300 | 86,669 | 84,000 | 450,000 | 4595 |
| **police officer** | 20,000 | 89,610 | 89,550 | 307,000 | 4502 |

Interesting Facts:

1. Teacher and Nurse - Their minimum salary is only around $12,000, which is lower than US' Poverty Level for individuals.
2. Relatively higher income people (>$400,000 and even $1,000,000) still use Lending Club to borrow money.
3. Some jobs traditionally associated with high income have very min income, such as attorney, director, engineer. And these salaries have been verified by Lending Club. Might be just a typo and missing a ‘0.

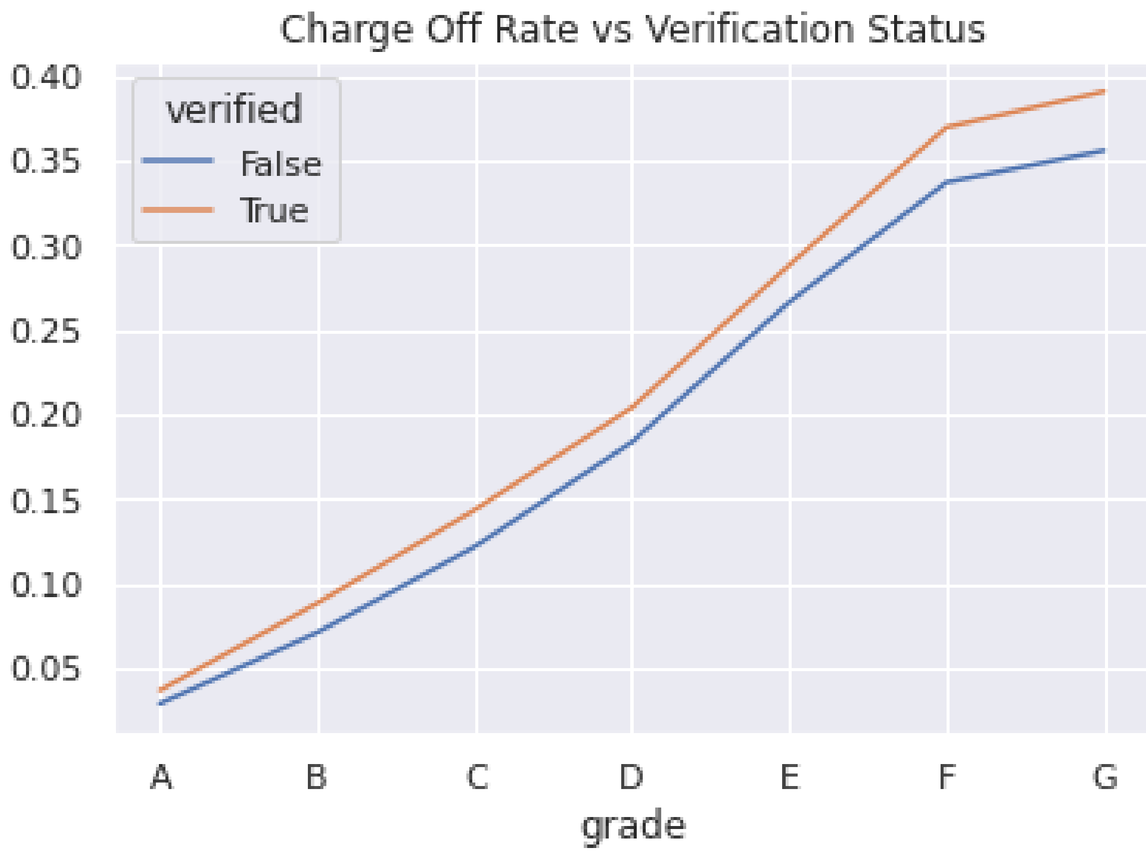
## Fico Range and Grade




No surprise here, Grade A has the highest Fico Score Range.

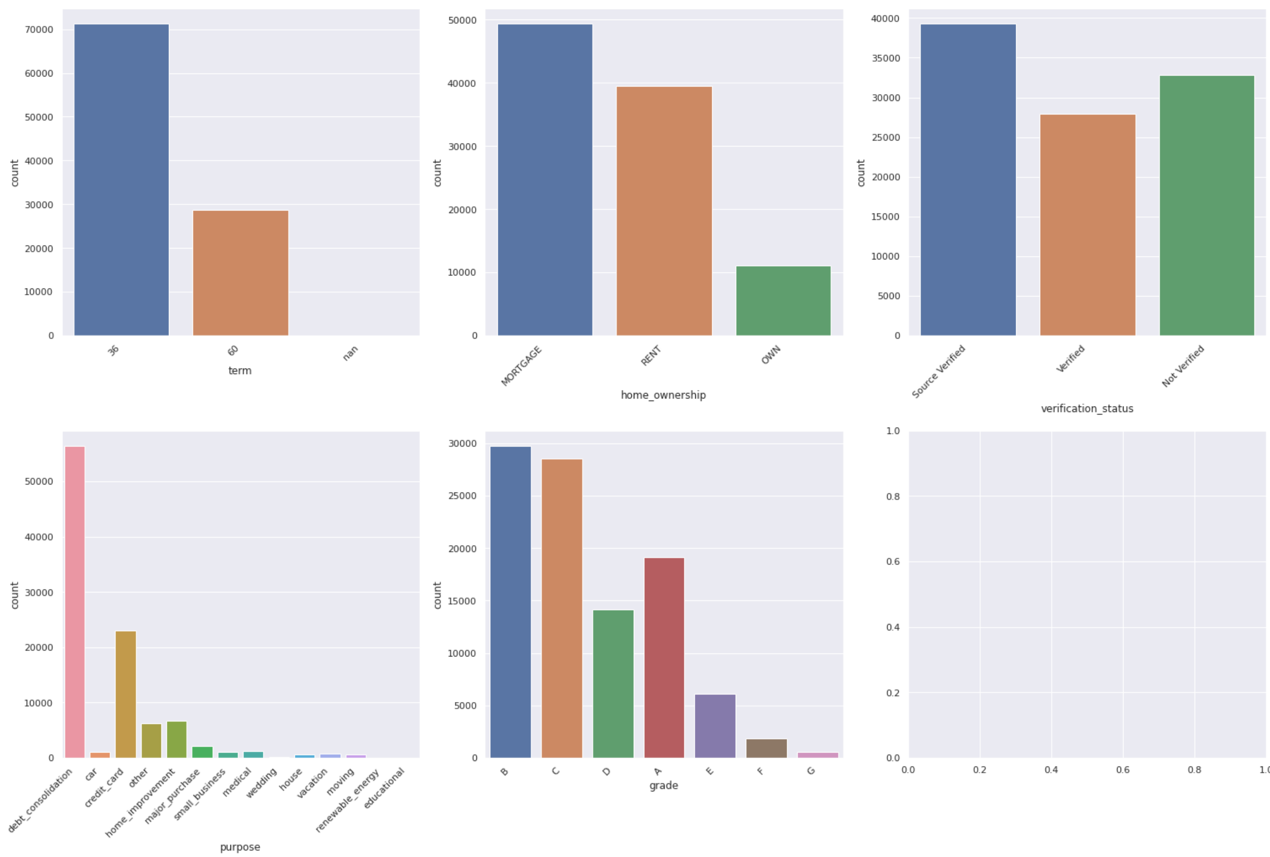
## Charge off rate vs Verification status

I define that a loan is considered charge-off when the value of loan_status is Charged Off or Default.

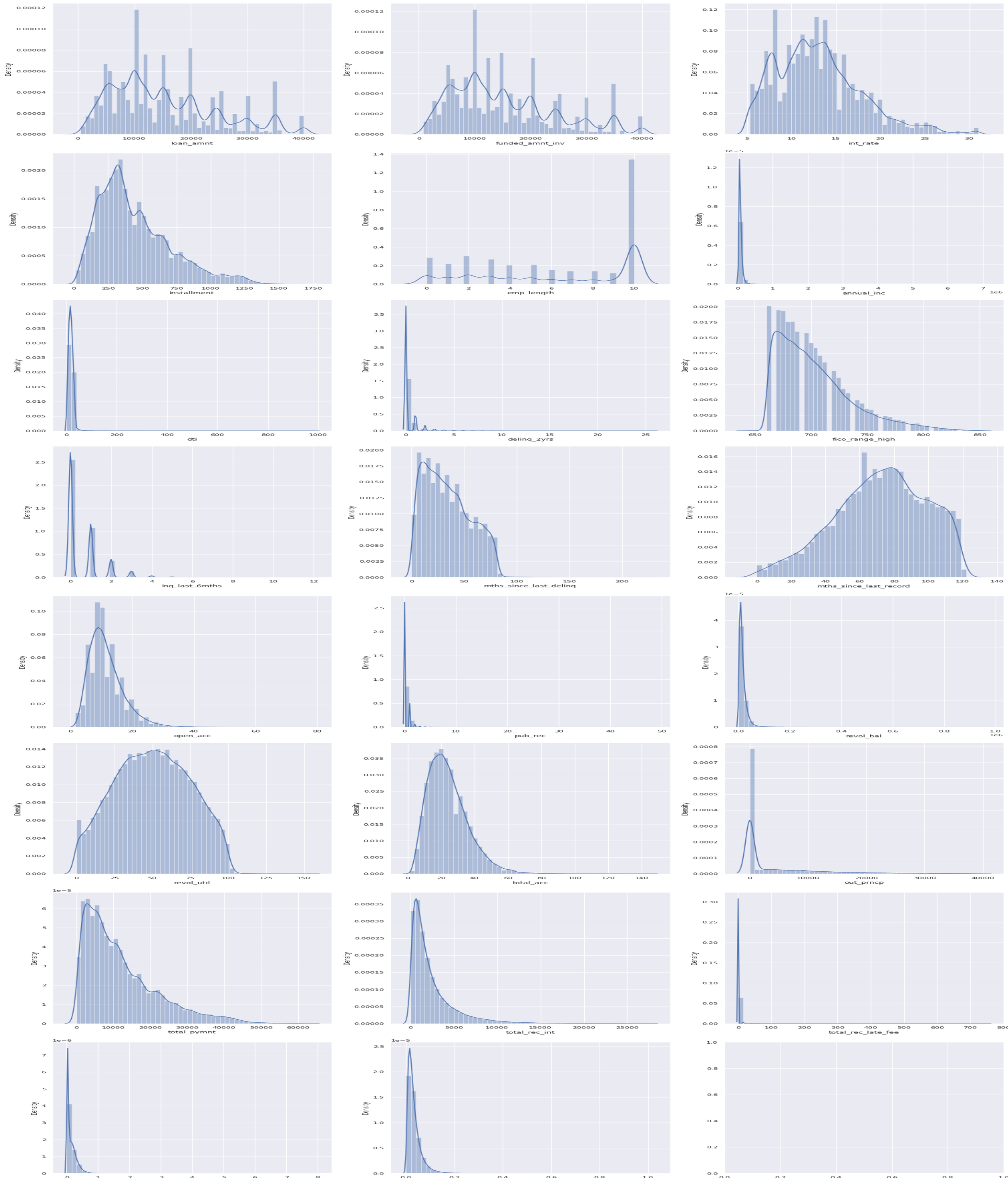

Charge Off Rate vs Verification Status
verified
False
True
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
A
B
C
D
E
F
G
grade

# Visualization

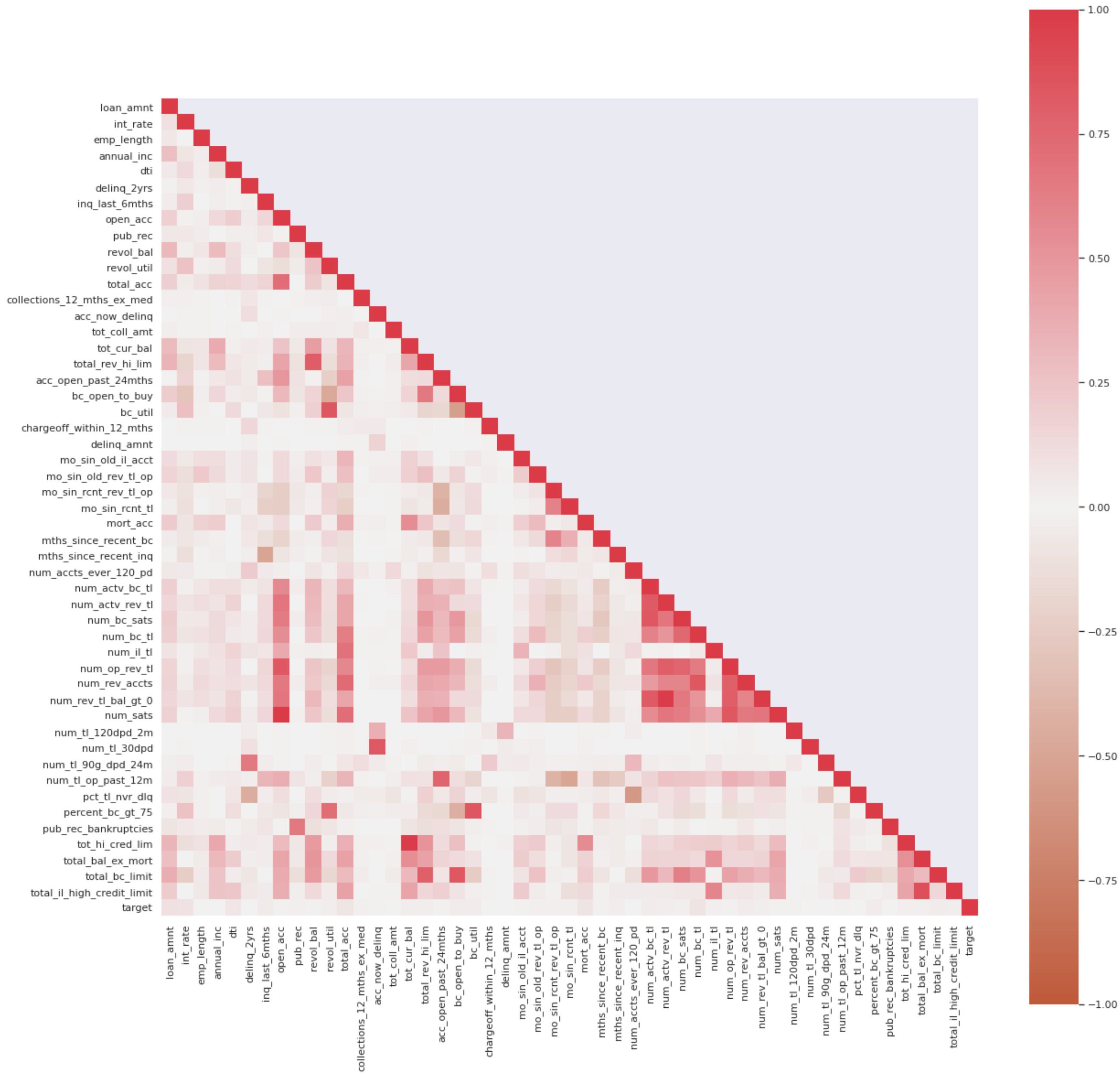
## Categorical Features



# Numerical Features



# Correlation Heatmap




We can see from the heatmap that there are several highly correlated features. Highly correlated features can affect the accuracy of the regression model. We will find and drop highly correlated features.

Once we've done that, we're ready for modeling.

# Classification Model

In this projects, several widely used algorithms are compared using Pycaret module, including:

1. CatBoost Classifier
2. Light Gradient Boosting Machine
3. Random Forest Classifier
4. K Neighbors Classifier

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| 0 | CatBoost Classifier | 0.7568 | 0.8360 | 0.7805 | 0.7612 | 0.7707 | 0.5120 | 0.5121 | 63.8516 |
| 1 | Light Gradient Boosting Machine | 0.7328 | 0.8031 | 0.7754 | 0.7308 | 0.7524 | 0.4627 | 0.4637 | 3.6287 |
| 2 | Gradient Boosting Classifier | 0.6678 | 0.7274 | 0.7102 | 0.6734 | 0.6913 | 0.3324 | 0.3329 | 1032.0488 |
| 3 | Random Forest Classifier | 0.6219 | 0.6707 | 0.5766 | 0.6589 | 0.6150 | 0.2469 | 0.2490 | 11.3751 |
| 4 | K Neighbors Classifier | 0.5327 | 0.5439 | 0.5638 | 0.5528 | 0.5583 | 0.0623 | 0.0623 | 744.4443 |

CatBoost Classifier was found to be the best performing model for our purpose.

Then, we set up CatBoost Classifier and train it on our data.

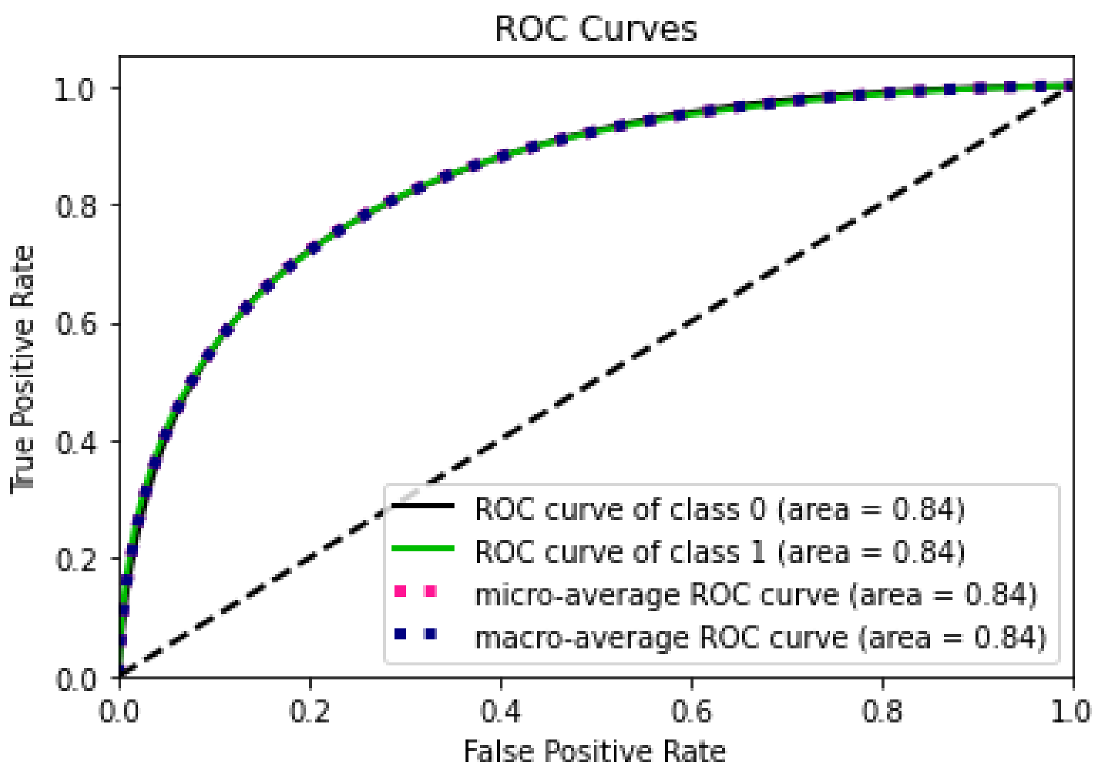
| | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| 0 | 0.7580 | 0.8372 | 0.7820 | 0.7621 | 0.7719 | 0.5143 | 0.5145 |
| 1 | 0.7582 | 0.8373 | 0.7818 | 0.7625 | 0.7720 | 0.5147 | 0.5148 |
| 2 | 0.7557 | 0.8351 | 0.7786 | 0.7606 | 0.7695 | 0.5098 | 0.5099 |
| 3 | 0.7566 | 0.8360 | 0.7801 | 0.7611 | 0.7705 | 0.5115 | 0.5117 |
| 4 | 0.7565 | 0.8367 | 0.7794 | 0.7613 | 0.7702 | 0.5113 | 0.5114 |
| 5 | 0.7582 | 0.8364 | 0.7800 | 0.7635 | 0.7717 | 0.5149 | 0.5150 |
| 6 | 0.7580 | 0.8370 | 0.7798 | 0.7633 | 0.7715 | 0.5145 | 0.5146 |
| 7 | 0.7586 | 0.8380 | 0.7812 | 0.7634 | 0.7722 | 0.5156 | 0.5157 |
| 8 | 0.7576 | 0.8363 | 0.7810 | 0.7620 | 0.7714 | 0.5135 | 0.5137 |
| 9 | 0.7561 | 0.8351 | 0.7784 | 0.7613 | 0.7698 | 0.5106 | 0.5107 |
| **Mean** | 0.7574 | 0.8365 | 0.7802 | 0.7621 | 0.7711 | 0.5131 | 0.5132 |
| **SD** | 0.0010 | 0.0009 | 0.0012 | 0.0010 | 0.0009 | 0.0020 | 0.0020 |

Based on 10-fold CV CatBoost:

- Accuracy = 75%
- AUC = .84

## ROC Curves

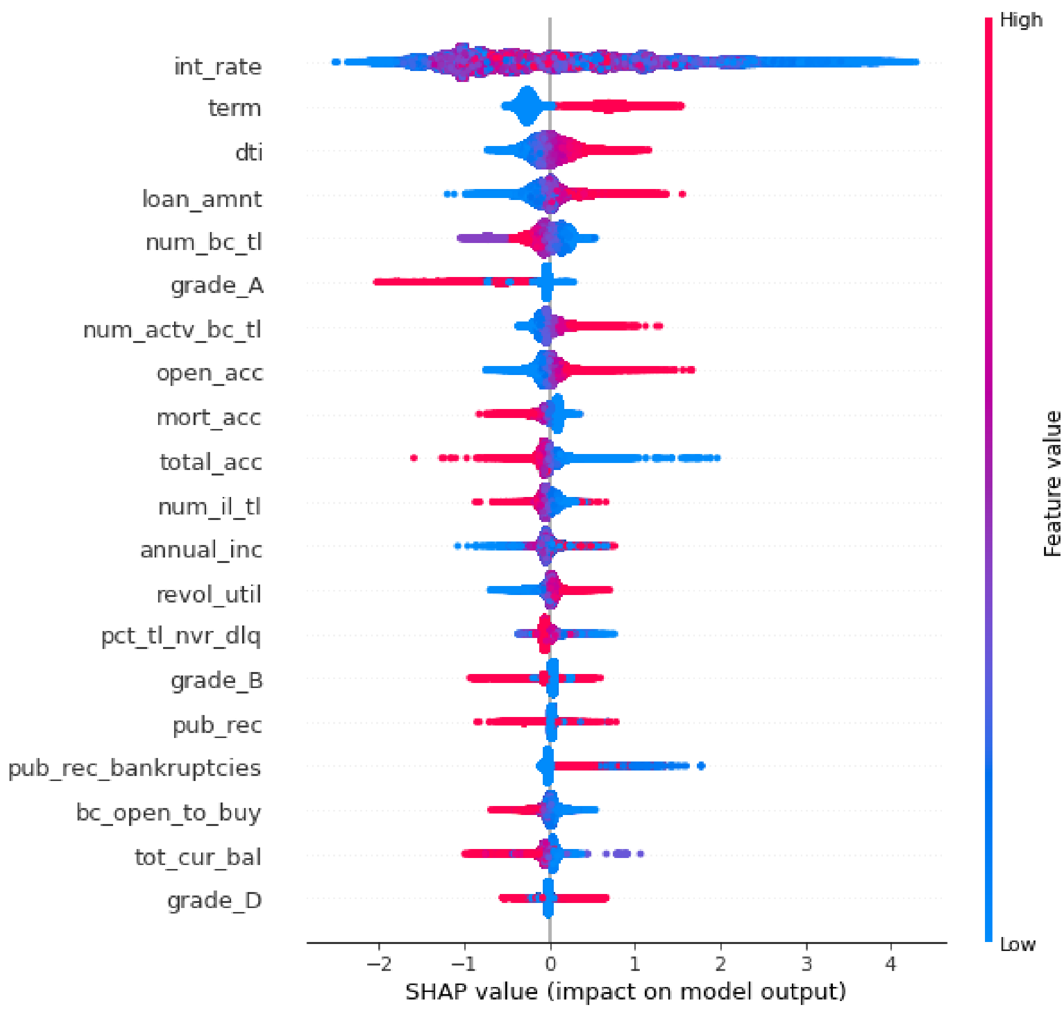
The performance of the model can also be seen in the ROC Curves. Using scikitplot package - ROC curves were plotted.




## Feature Importance

SHAP value can be generated with the Pycaret model. In this project, we used it to determine the feature importance.

Interest rate was found to have the highest feature value.




## Prediction

Making predictions with our model on the test data resulting in similar performance of the training data.

- Accuracy = 75%
- AUC = .84

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|---|
| **0** | CatBoost Classifier | 0.7571 | 0.8364 | 0.7809 | 0.7614 | 0.771 | 0.5125 | 0.5127 |

# To Improve

There are things I would like to do in the future to improve accuracy:

1. Ensemble machine learning model
2. PCA