**Lending Club data analysis and improving Default Prediction using machine learning methods**

Credit Risk Project

Nguyễn Minh Đức

Student ID: 11204838

Lending Club data analysis and improving Default Prediction using machine learning methods

Abstract:

The aim of this paper is to examine how various demographic and financial variables are linked to the payment behavior of consumer credit clients, specifically in terms of credit default risk. The aim of this project is to develop a model that can accurately predict whether a customer is likely to default on a loan, based on a range of factors such as income, grade, and loan amount. To achieve this, the project will involve collecting and cleaning data from LendingClub's publicly available datasets, performing exploratory data analysis to identify trends, selecting and fine-tuning a machine learning algorithm. The model will be evaluated using performance metrics such as accuracy, precision, recall, AUC-ROC to determine its effectiveness in classifying customers. The result from this project could have a significant implication for lending institutions, helping them to make informed decisions about loan approvals and risk management.

1. **Introduction**

Over the past decade, the significance of risk management in credit has grown for both borrowers and lenders, particularly in developing countries. For this reason, financial institutions and banks have begun to review their lending policies. There are six main functional responsibilities associated with credit lending activities; including (1) assessing a customer’s credit risk, (2) making the credit granting decision with regard to credit terms and, where relevant, credit limits, (3) collecting receivables (debts) as the fall due and taking action against defaulters, (4) monitoring customer behavior and compiling management information, (5) assuming the risk of default or bad debt, (6) financing investments in receivables (debtor) (Summer and Wilson, 2000).

This project deal with the fourth step of credit lending activities. In this project, we will be using machine learning to develop a classification model that can predict whether a borrower is likely to default on a loan. These results allow financial institutions to adapt their lending policies and reduce their risks of credit defaults.

1. **Literature review**

Banking portfolios often have a value of billions of dollars in top banks, hundred of millions in smaller banks, so even small enhancements can be considered substantial. Therefore, striving to improve model performance in credit default is a valuable pursuit. But, first, we will review the models that will be used to classify default customers:

**2.1 Logistic Regression (Traditional method):**

A logistic regression model is a statistical model used to analyze the relationship between a binary dependent variable and one or more independent variables. It is commonly used for classification tasks where the dependent variable is binary (i.e., takes on only two possible values, such as 0 or 1).

The logistic regression model estimates the probability of the dependent variable being in one of the two possible categories as a function of the independent variables. This function is typically modeled using the logistic function, which maps any real-valued input to a value between 0 and 1. The logistic function is used because it ensures that the predicted probabilities always fall within the allowable range (i.e., between 0 and 1).

To build a logistic regression model, the modeler needs to first choose the independent variables that are expected to have a relationship with the dependent variable. The modeler then estimates the coefficients of the logistic function using a training dataset. The trained model can then be used to make predictions on new data by applying the logistic function to the new data's independent variables to generate the probability of the dependent variable being in one of the two categories. The predicted probability can then be used to make a binary classification decision based on a chosen threshold value.

The logistic function is of the form:

**2.2 Decision Tree:**

A decision tree model is a supervised learning algorithm used in machine learning for solving classification and regression problems. The model builds a tree-like structure by recursively splitting the data into subsets based on the most significant features, in order to achieve maximum homogeneity or purity within each subset.

The decision tree model starts with a single node, called the root node, which represents the entire dataset. The model then selects the most important feature to split the data and creates two child nodes, each representing a subset of the data. This process is repeated recursively for each child node, until a stopping criterion is met, such as achieving a certain level of homogeneity or reaching a maximum depth of the tree.

The decision tree model is an interpretable model, as it is easy to visualize and understand how the model arrives at its predictions. The model can be used for both classification and regression tasks, depending on the type of dependent variable. In classification problems, the model predicts the class label of a new data point by traversing the tree from the root node to a leaf node that corresponds to the predicted class. In regression problems, the model predicts the numerical value of a new data point by traversing the tree to a leaf node that corresponds to the predicted value.

The decision tree model has several advantages, such as being able to handle both numerical and categorical data, handling missing data, and being robust to outliers. However, it can suffer from overfitting if the tree is too deep or if the dataset is too small. Therefore, pruning techniques and ensemble methods such as random forests and boosting are commonly used to improve the model's performance.

**2.3 Random Forest:**

The random forest algorithm is a machine learning technique that utilizes ensemble learning to perform tasks such as classification and regression. This model consists of numerous decision trees, each constructed using a randomly selected subset of training data and features.

To build a random forest model, the algorithm starts by creating multiple decision trees. Each decision tree is trained on a random subset of features and a random subset of the training data. The final prediction is made by aggregating the predictions of all the trees in the forest. In classification problems, the prediction is made by majority voting, while in regression problems, the prediction is made by taking the average of the predictions of all the trees.

The random forest model has several advantages over a single decision tree, such as reducing the risk of overfitting, handling missing values, and providing a feature importance ranking. It is also a highly parallelizable algorithm, which makes it suitable for large datasets.

**2.4 XGBoost:**

XGBoost (Extreme Gradient Boosting) is a popular gradient boosting library used for regression, classification, and other machine learning tasks. It is known for its speed and performance, and is widely used in machine learning competitions and industry applications.

Gradient boosting is an ensemble learning method that combines multiple weak models to create a stronger model. The idea behind gradient boosting is to iteratively add weak models, each one correcting the errors of the previous model, until the final model is obtained. In XGBoost, the weak models used are decision trees, and the algorithm optimizes a loss function that measures the difference between the predicted and actual values.

XGBoost uses a number of techniques to improve performance and reduce overfitting, such as regularization, tree pruning, and parallelization. It also provides a built-in feature importance ranking that can be used to understand which features are most important for the model.

**2.5 CatBoost:**

CatBoost is a gradient boosting library that is designed to work with categorical data. It is an open-source library developed by Yandex and is known for its ability to handle categorical features without requiring any preprocessing.

Like other gradient boosting algorithms, CatBoost works by building an ensemble of decision trees. However, it differs from other gradient boosting libraries in its handling of categorical features. CatBoost uses an innovative technique called "ordered boosting," which is able to take into account the order of categorical variables and reduce the number of splits needed to separate the data into distinct categories.

CatBoost also includes a number of other features designed to improve performance, such as automatic handling of missing values, robust ranking methods, and efficient GPU support. Additionally, it provides built-in feature importance and visualization tools.

1. **Methodology**
   1. **Data Source:**

Data is obtained from the public data of Lending Club. Lending Club is a peer-to-peer lending platform that enables individuals to obtain loans from other individuals or institutional investors. This platform connects borrowers with investors, providing a more efficient and cost-effective alternative to traditional lending institutions. Lending Club’s data set is for the period from 2007 to 2018. There are more than 1.3 millons observations and and more than 150 variables. It is very hard to run analysis with all the variables and observations. So, we cleaned this data and keep several variables as our predictor variables for our models and then delete the left over.

* 1. **Data Exploration**

**3.2.1 Target Variable**:

Loan\_status: Current status of the loan. In the data set contain many status of Loan such as Fully Paid, Current, Charged Off, Late (31-120 days), In Grace Period,... We only keep 2 variable Fully Paid and Charger Of and then encode Fully Paid is 1, Charged Off is 0.

* + 1. **Independent Variables:**

Table 1: Independent Variables Description

|  |  |
| --- | --- |
| Variable | Description |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| term | The total number of monthly payments that must be made to repay the loan. The values can either be 36 or 60 months. |
| int\_rate | Interest rate on the loan. |
| installment | The amount of money that the borrower has to pay on a monthly basis when the loan is initiated. |
| sub\_grade | LC assigned loan subgrade. |
| purpose | A category provided by the borrower for the loan request. |
| emp\_title | The job title provided by the borrower while submitting the loan application. |
| emp\_length | Employment length in years. It can range from 0 to 10, where 0 indicates employment for less than one year and 10 indicates employment for ten years or more. |
| addr\_state | The state provided by the borrower in the loan application (USA). |
| home\_ownership | The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER |
| annual\_inc | The annual earnings declared by the borrower while enrolling for the loan. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified. |
| dti | Debt-to-income ratio. |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| pub\_rec | Number of derogatory public records. |
| revol\_bal | Total credit revolving balance. |
| revol\_util | The revolving credit utilization rate, which is the proportion of the borrower's used revolving credit amount to the total available revolving credit. |
| total\_acc | The total number of credit lines currently in the borrower's credit file. |
| mort\_acc | Number of mortgage accounts. |
| fico\_range\_high | The upper boundary the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary the borrower’s FICO at loan origination belongs to. |

1. **Exploratory Data Analysis**

|  |  |
| --- | --- |
| In our dataset, in cases where loans were Charged Off, the minimum loan amount was $900 and the maximum was $40,000, with an average of $15,565.04 and a standard deviation of $8,814.553. On the other hand, when loans were Full Paid, the maximum loan amount remained the same, but the minimum amount was slightly different ($500 compared to $900). Additionally, the mean and standard deviation of the loan amount were lower compared to the Charged Off cases. | Figure 1: Loan status by loan\_amount |

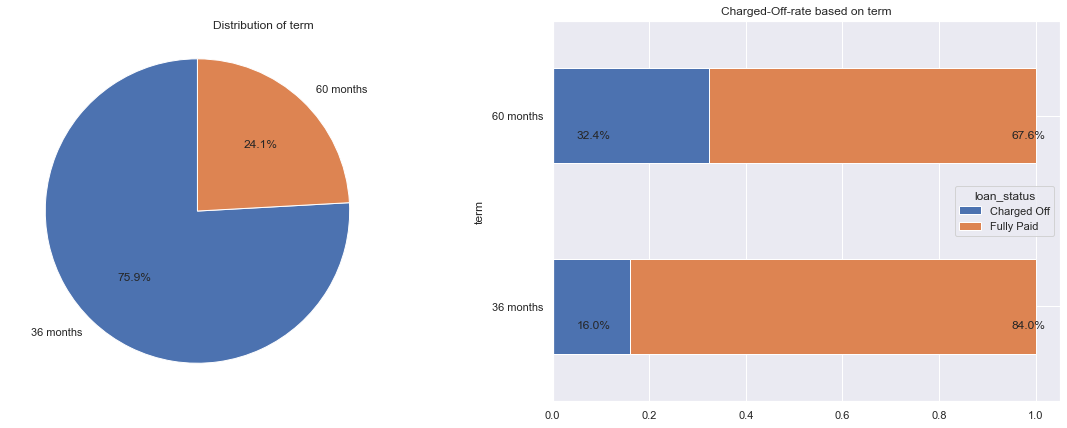
Figure 2 illustrates that approximately 75.9% of loans had a 36-month term, out of which 32.4% were charged off and 67.6% were fully paid. Additionally, around 24.1% of loans had a 60-month term, with 16% being charged off and 84% being fully paid.

Figure 2: Loan status by terms

Due to the presence of outliers in countinuous variable, we will need to eliminate them before conducting any further analysis on the remaining data. After remove outliners, we can see more detail about these variables:

|  |  |
| --- | --- |
| There was a very clearly trend is that Charged off status customers tended to have higher interrest rate on average, 15.71% versus 12.62%. The mean, standard deviation, minimum, and maximum values of the installment variable were greater for the Charged-off status compared to the Fully-paid status. | Figure 3: Loan status by int\_rate |
| In the group of Fully-paid status, the average annual income is $ 73988.66 with a standard deviation of 38990.31, while in the Charged-off status group, the average annual income is $68055.65 with a standard deviation of 36117.26. | Figure 4: Loan status by annual\_inc |
| It is apparent that the charged-off status has a higher average debt-to-income ratio (20.17) compared to the fully-paid status (17.81)    Figure 5: Loan status by dti | Figure 6: Loan status by revol\_bal  Surprisingly, fully-paid status has a higher average total credit revolving balance ($16471.25) compared to charged-off status ($15353.5). |
| Revol\_util refers to the revolving line utilization rate, which is the amount of credit that the borrower is utilizing in relation to all available revolving credit. On average, borrowers who have fully paid off their debts have a lower total credit revolving balance of 51.07 compared to those who have charged-off debts, whose average total credit revolving balance is 54.76. | Figure 7: Loan status by revol\_util |
| It has been observed that the average FICO score for fully-paid customers is higher than that of charged-off customers. Specifically, fully-paid customers have an average FICO score of 702.26, while charged-off customers have an average FICO score of 691.85. This suggests that individuals with a higher FICO score are more likely to pay off their debts on time, making them more attractive to lenders. | Figure 8: Loan status by fico\_range\_high |

|  |  |
| --- | --- |
| Figure 9: Loan status by addr\_state | From Figure 9, , it is evident that California has the highest number of borrowers, with approximately 200,000 loans. Following California is Texas and New York, with around 110,000 and 109,000 loans respectively. It is not difficult to comprehend why these three states have a large number of borrowers, |

as they are all known for their strong economic performance and large economies. For instance, California has the largest economy in the United States and the fifth largest in the world, while Texas has the second-largest economy in the United States and the tenth largest in the world. Similarly, New York has the third-largest economy in the United States and the fifteenth largest in the world.

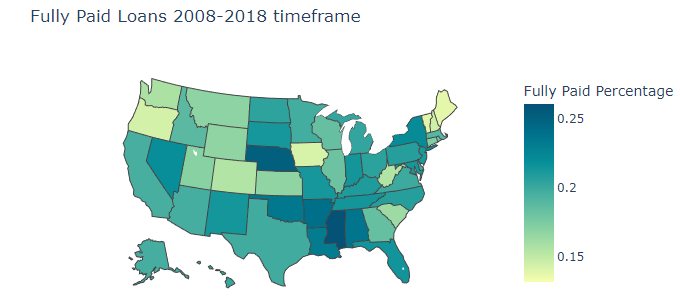
The charge-off rate ranges from 13.2% in Washington, DC to 26% in Mississippi. Despite having a default rate below the threshold of percent (20%), which is only exceeded by New York State with a 22% default rate, the three aforementioned states have relatively low default rates. Typically, states with high default rates are located in the central southern region of the United States. It is concerning to note that Mississippi and Nebraska, which have a relatively small number of loans, have the highest default rates among all states, with 26% and 25.2%, respectively. 

Figure 10: Loan status by addr\_state

Sub-grade has right-Skewed distribution, which means only few loans were labeled as F grade. There was also a clearly trend that the worse the sub-grade the higher percentage the loan was belong to a Charged-off customer.

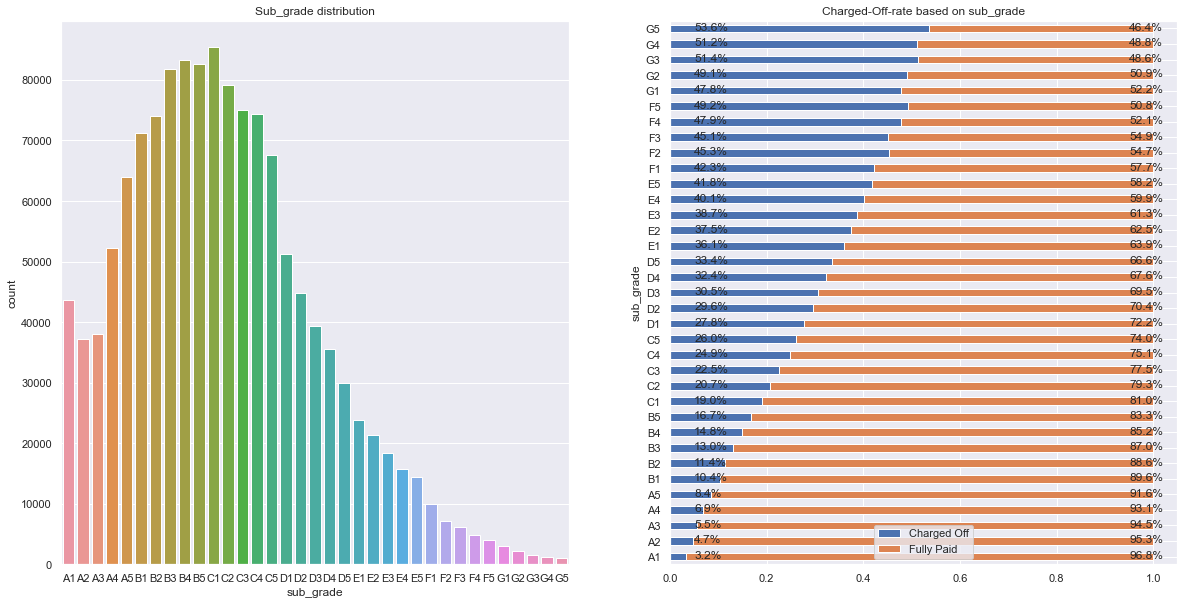
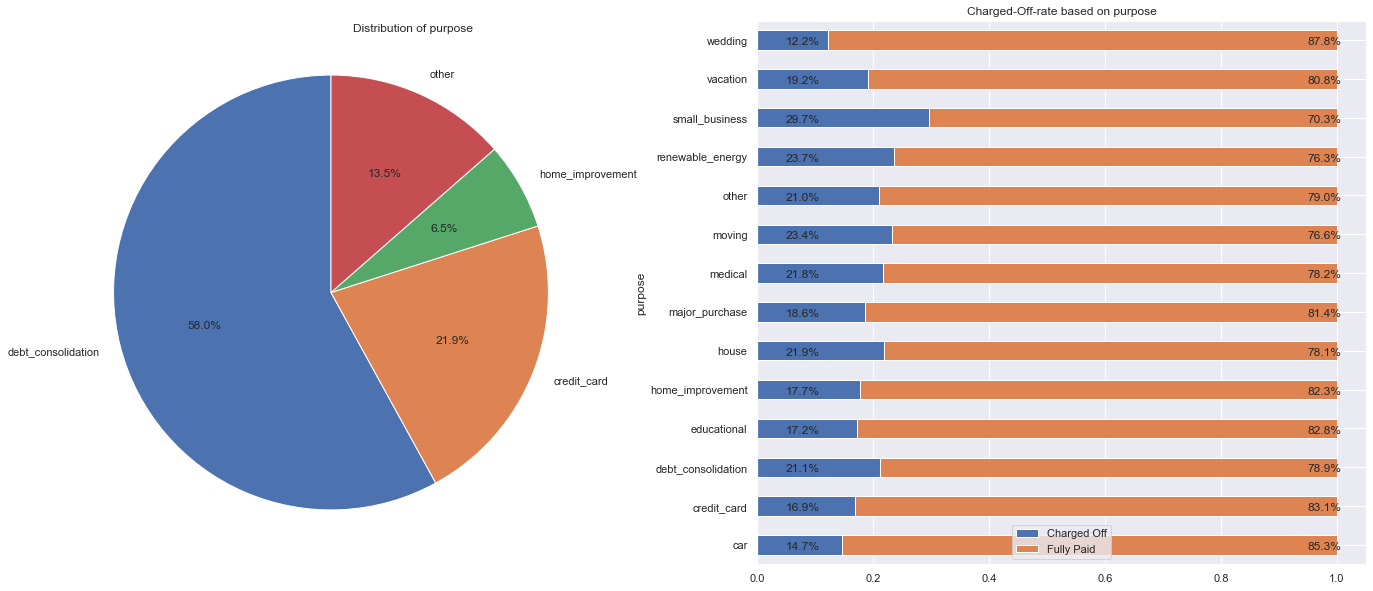


Figure 11: Sub-grade distribution and Loan status by Sub grade

As we can see from Figure 10, about half of the loans was about debt consolidation and 20% of the loans was credit card. If we dive deep into other purpose variable, we can see that small business purpose has the highest chance to become Charged-off customers (29,7%) followed up by renewable energy which have (23.7%). Although 58% of purpose was for debt consolidation but only one-fifth of these type of customers were Charged-off (21.1%).

Figure 12: Purpose distribution and Loan status by Purpose

|  |  |
| --- | --- |
| Figure 13: Top 5 jobs title | From the dataset, we observed 309948 unique job title, and teacher is the most popular job title among borrowers. We might consider to drop this variable |

|  |  |
| --- | --- |
| It seem like majority of borrowers were have 10+ years experiences (34.9%) and the more experiences customers have the less chance they were Charged-Off. This variables have too much categorical value (might not efficiency in machine learning models) so we gonna encode this variable into emp\_role (['< 1 year', '1 year', '2 years'] assigned as ‘Fresher’, ['3 years', '4 years', '5 years'] assigned as 'Middle Senior', ['6 years', '7 years', '8 years','9 years'] assigned as 'Senior' and ['10+ years'] assigned as 'Expert' ). Without loss of generality, we have: | Figure 14: Proportion of emp\_length |

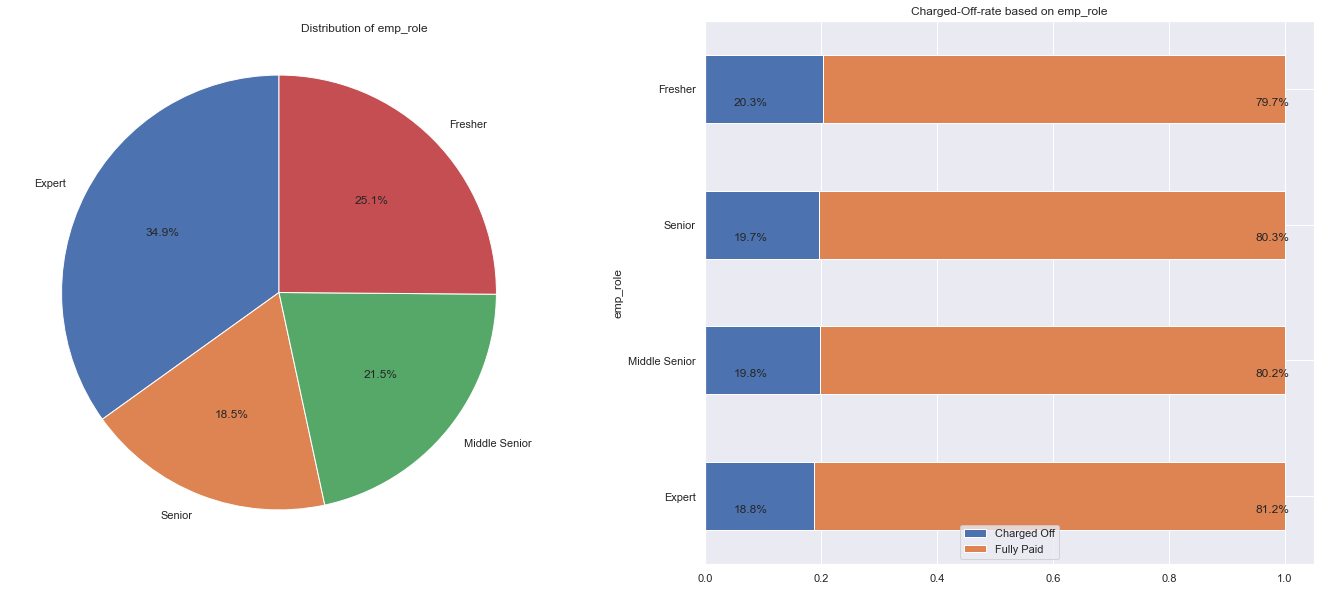


Figure 15: Emp\_role distribution and Loan status by Emp\_role

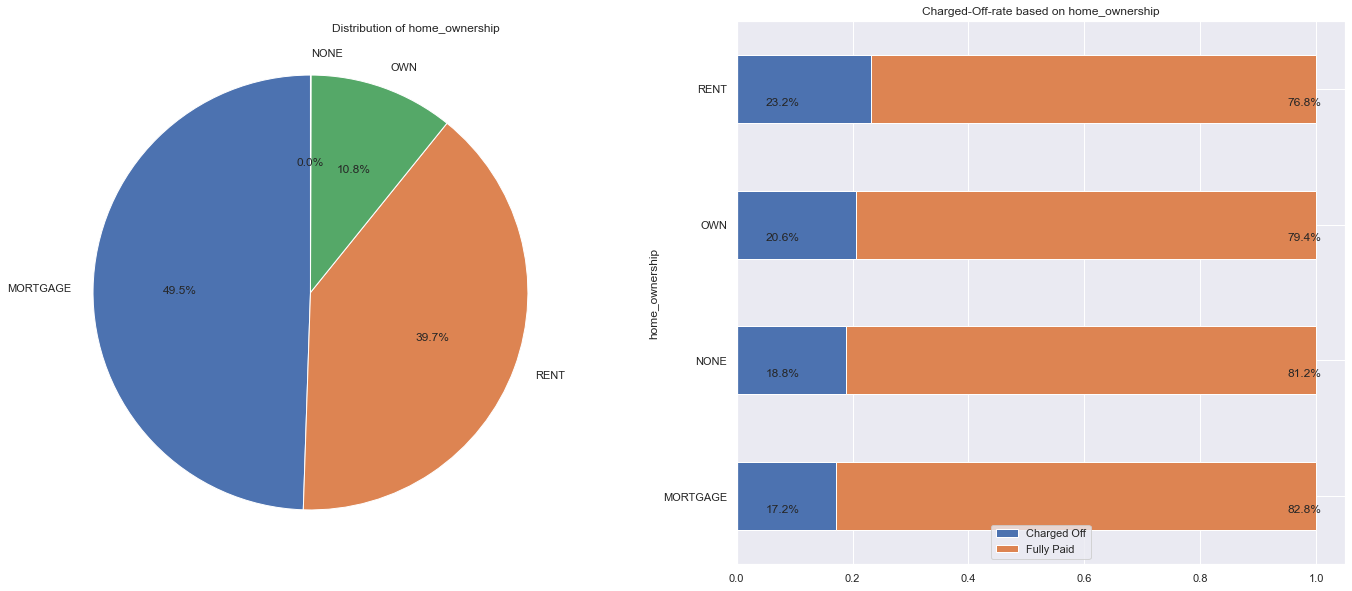
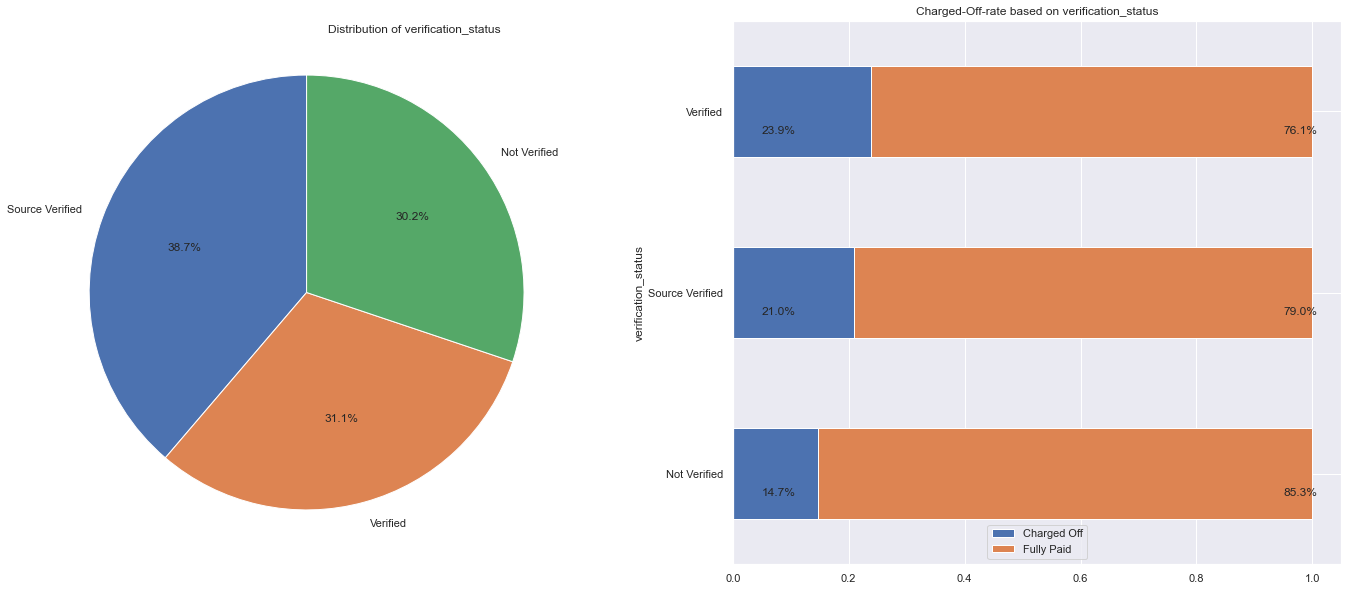


Figure 16: Home\_ownership distribution and Loan status by Home\_ownership

Approximately 10.8% of borrowers were homeowners, whereas half of borrowers used their house as collateral for a mortgage, and 39.7% rented their house while their loan was active. Out of all mortgage borrowers, only 17.2% were classified as Charged-off borrowers, while the group of homeowners had 20.6% of Charged-off borrowers. It is expected that renters had the highest proportion of Charged-off borrowers (23.2%).

 Figure 17: verification\_status distribution and Loan status by verification\_status

The "verification\_status" variable is evenly split into three equally categories, and borrowers who have been verified by LC have the highest Charged off rate of 23.9%. Similarly, Source Verified customers have a Charged off rate of 21.9%. Interestingly, unverified borrowers have the lowest Charged off rate of just 14.7%.

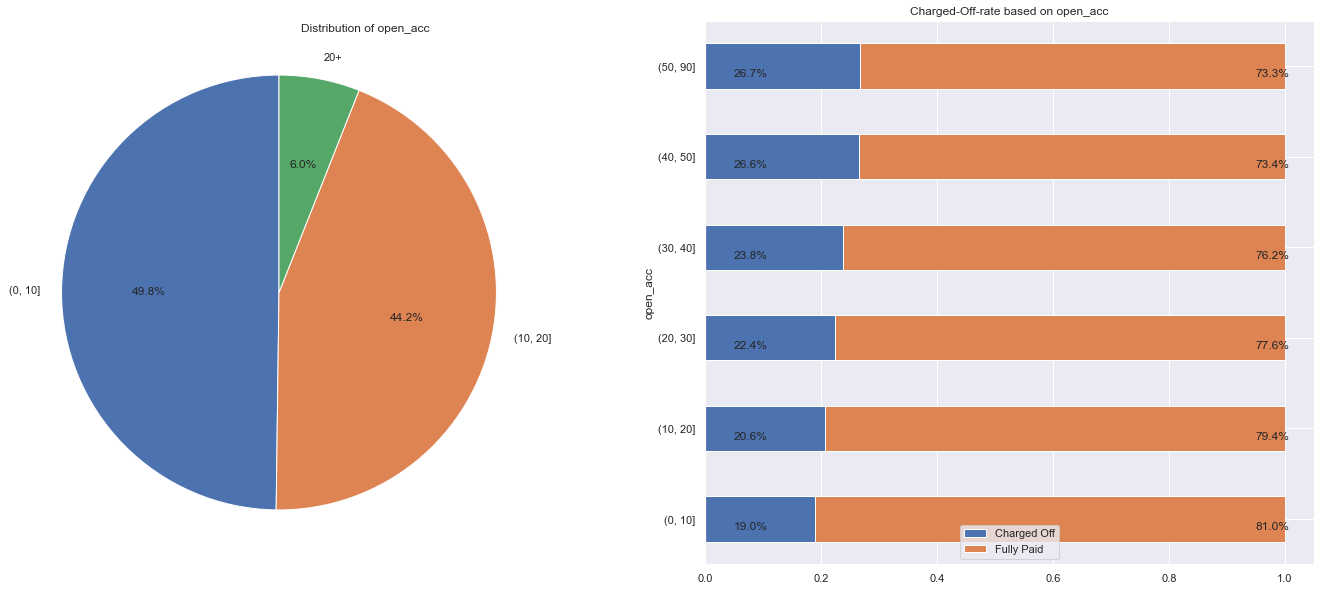


Figure 18: open\_acc distribution and Loan status by open\_acc

Just 6% of borrowers possess over 20 active credit lines on their credit file, while half of borrowers have less than 10 active credit lines. 44.2% of borrowers hold between 10 to 20 credit lines. It's noteworthy that borrowers with a higher number of credit lines are more prone to becoming a charged-off borrower. For instance, borrowers with more than 50 credit lines have a 26.7% likelihood of becoming charged-off, while those with less than 10 credit lines have only a 19% chance.

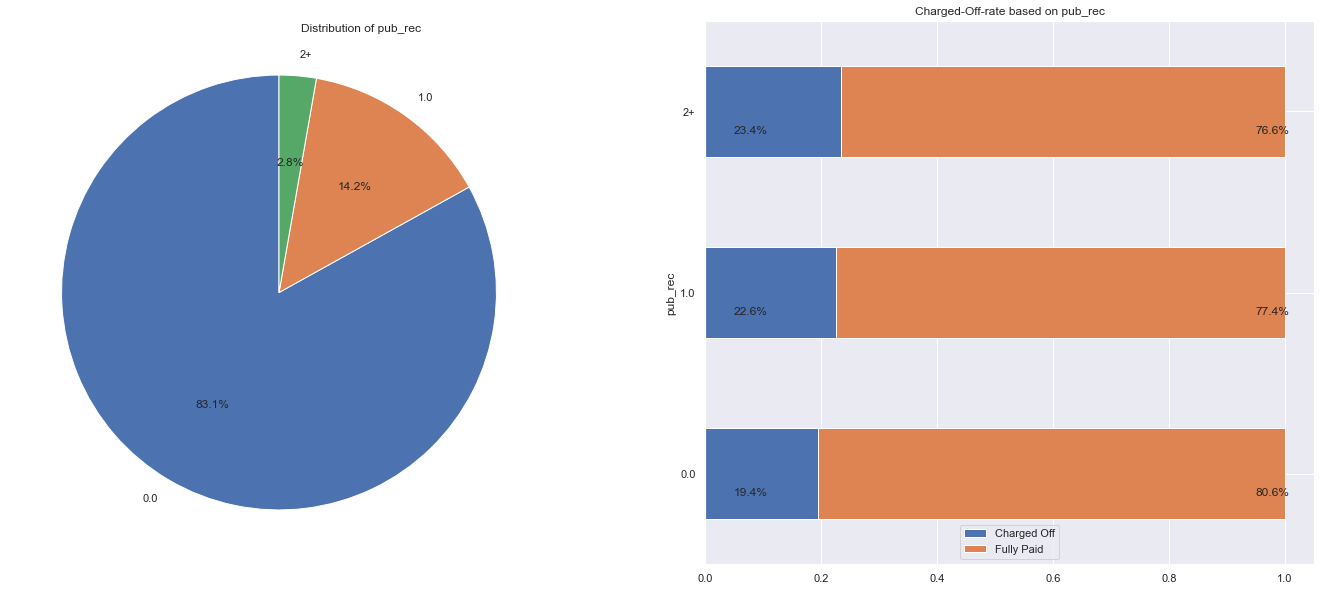


Figure 19: pub\_acc distribution and Loan status by pub\_acc

The vast majority of borrowers, nearly all, have no negative public records associated with their credit file. Only 14.2% of borrowers have a single derogatory public record, and merely 2.8% have more than two derogatory public records. There is a definite pattern indicating that borrowers with more derogatory public records are more likely to become charged-off. For example, non-derogatory public record borrowers have a charge-off rate of only 19.4%, while those with more than two records have a charge-off rate of 23.4%.

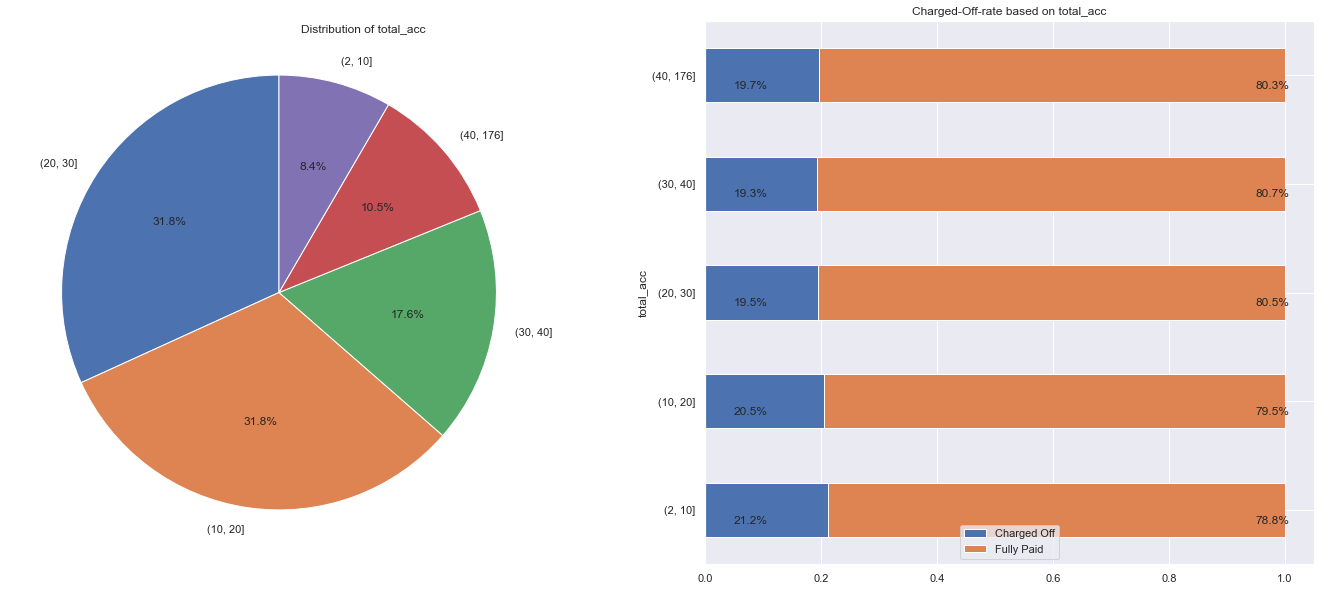


Figure 20: total\_acc distribution and Loan status by total \_acc

The majority of borrowers have a total number of accounts (total\_acc) ranging from 10 to 40. Additionally, there is a clear trend indicating that borrowers with a higher number of total\_acc are more likely to be fully paid compared to those with less than 10 total\_acc.

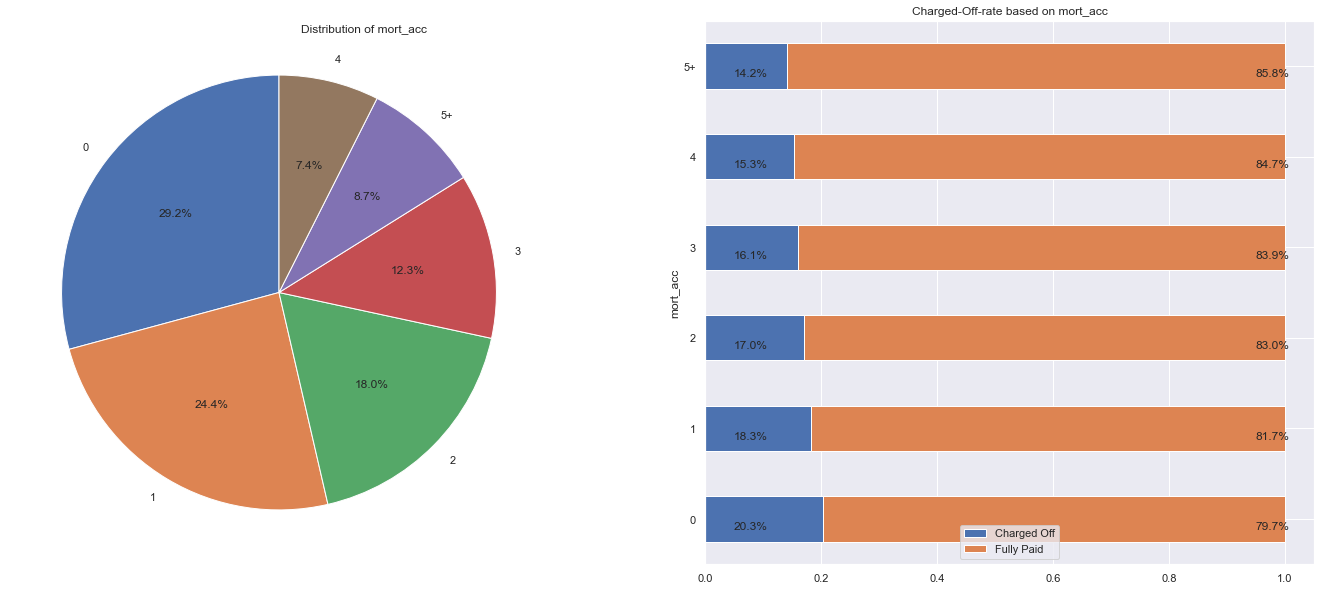


Figure 21: total\_acc distribution and Loan status by total \_acc

Approximately half of all borrowers have either no mortgage accounts or just one. Interestingly, borrowers with more than five mortgage accounts were found to have the lowest percentage of charge-offs at 14.2%, while those who have never had a mortgage account were found to have the highest rate of charge-offs at 20.3%.

1. **Data preprocessing**

Now that we have got some more understranding of the data, it’s time to go furthermore and do the next step (data preparation step) of this project. In this data preparation phase, a number of preprocessing activities have been applied to each variable in our dataset, a few activities in this phase are:

* Data cleaning
* Features selection
* Data Transform
* Data splitting
  1. **Data cleaning**

|  |  |
| --- | --- |
| Fortunately, there are only a few variables in our data that contain missing values, and the percentage of missing values in these variables is within an acceptable range (with the highest percentage being only 6.5% in the emp\_title variable). To address this issue, we plan to impute categorical variables with their respective mode values, and numerical variables with 0. | Figure 22: Feature null Percentage |

Also, the variables annual\_inc and dti contain numerous outliers. Therefore, we will remove these outliers using the IQR rule.

* 1. **Features selection**

Data is typically composed of numerous variables, but not all of them have a significant impact on the target variable. Since it can be challenging to identify irrelevant features from the outset, most variables are initially included in the model, and the model itself learns to choose the most effective features. However, it's important to keep in mind that irrelevant features can adversely impact the model's performance in terms of both speed and accuracy. In our project, we manually selected features due to the need to handle certain categorical variables (such as the emp\_title variable, which had 309,948 unique values). Dealing with variables with such a large number of unique labels posed a significant challenge. Additionally, in choosing numeric variables, we simply using correlation marix to check if there was correlation between features.

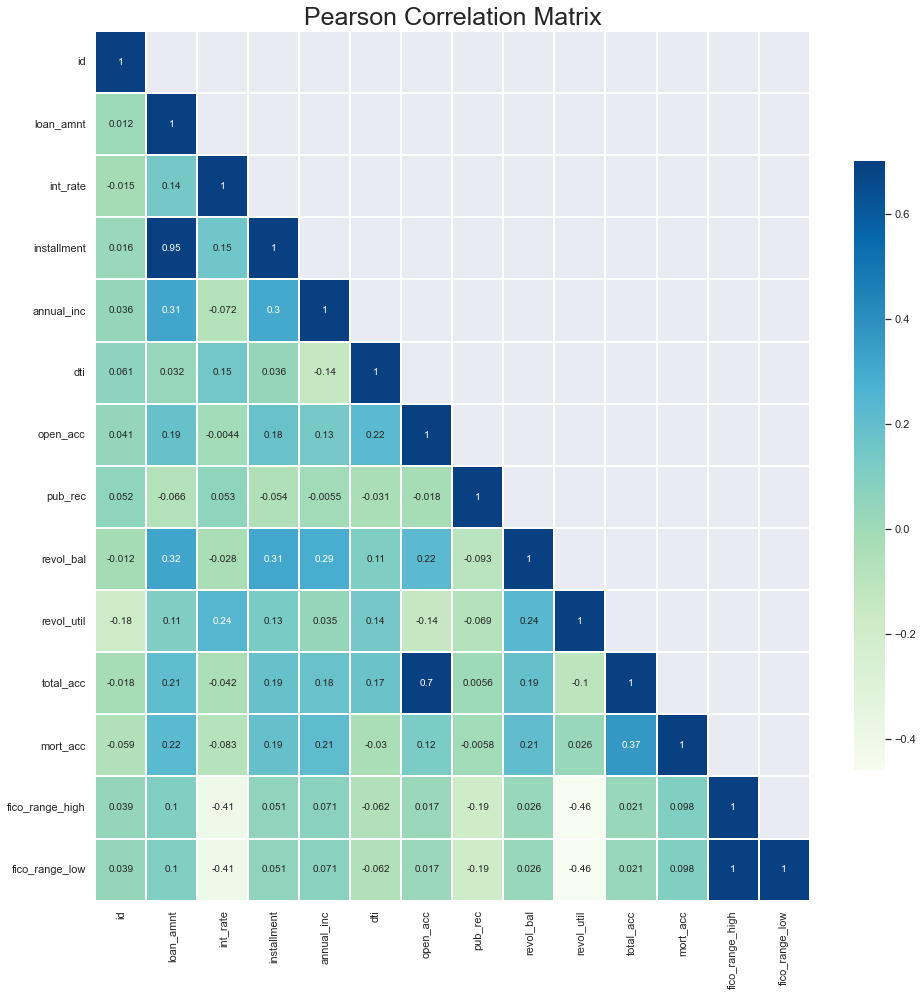


Figure 23: Correlation Matrix

The correlation matrix shows that there is a significant correlation between loan\_amnt and installment, as well as a moderately strong correlation between total\_acc and open\_acc. Furthermore, there is a perfect correlation (identity correlation) between fico\_range\_high and fico\_range\_low. To address this issue, we will select one variable from each pair of strongly correlated variables.

* 1. **Data Transform**
     1. **Dealing with categorical variables:**

**5.3.1.1 Ordinal encoding:**

There are some features have a natural ordered relationship between each other, and machine learning algorithm may be able to understand this relationship (e.g in ‘term’ variable: 36 months has been assigned as 0 while 60 months assigned as 1, ‘emp\_role’: assigned 'Intern' as 0, 'Junior' as 1,'Middle Senior' as 2,'Senior' as 3,'Expert' as 4)

**5.3.1.2 One-Hot encoding:**

If there are categorical features without an inherent order, ordinal encoding is not suitable. This method can cause the model to incorrectly assume a natural order between categories, resulting in suboptimal performance or unpredictable outcomes. To circumvent this issue, one-hot encoding can be utilized. In our dataset, we are using one-hot encoding to encode the remaining categorical variables

e.g: encoding ‘home\_ownership’

|  |  |
| --- | --- |
| Original data: | Data after encoding |

* + 1. **Dealing with numerical variables**

**5.3.2.1 Normalization (Min-Max Scaler)**

The Min-Max scaler retains the original distribution's form, and it does not substantially modify the information contained in the initial data. We use this technic to scaler numerical variables in our datasets.

* 1. **Data splitting**

Data splitting is about partitioning available data into two portions; usually, one for training and the other is for validation purposes. The first portion of the data is used to develop a predictive model, and the other to evaluate the model's performance. We divide our data set into train set with account for 70% proportion of original datasets while validation set contain 30%.

1. **Model Building**

This project is about default classification problem, so we gonna try both traditional approach (Logistic) and Tree-Based Ensemble Models (like RandomForest, XGBoost, CatBoost)

* 1. **Evaluation Model Performance Metrics:**

In this project, model performance is defined as the model’s ability to correctly predict potential defaults. To measures the performance of the model the following performance measures were derived:

* Accuracy: The ratio of correctly classified instances to the total number of instances in the evaluation data.
* Precision: The proportion of true positives to the total number of positive predictions made by the model.
* Recall: The ratio of true positive predictions to the total number of positive instances in the data.
* Area Under receiver operating Curve (AUC) - The Receiver Operating Curve (ROC): A measure of the model's ability to classify instances correctly over a range of decision boundary thresholds.
  1. **Models Performance**

Table 2: Comparison of Model Performance between Algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Accuracy | Precision | Recall | AUC |
| Train | Logistics | 0.803 | 0.548 | 0.055 | 0.708 |
| Decision Tree | 1.000 | 1.000 | 1.000 | 1.000 |
| Random Forest | 1.000 | 1.000 | 1.000 | 1.000 |
| XGBoost | 0.808 | 0.616 | 0.096 | 0.741 |
| CatBoost | 0.813 | 0.668 | 0.117 | 0.751 |
| Test | Logistics | 0.803 | 0.553 | 0.056 | 0.708 |
| Decision Tree | 0.706 | 0.276 | 0.291 | 0.550 |
| Random Forest | 0.803 | 0.551 | 0.070 | 0.706 |
| XGBoost | 0.805 | 0.565 | 0.088 | 0.722 |
| CatBoost | 0.805 | 0.563 | 0.097 | 0.726 |

In the train set, Decision Tree and Random Forest algorithm seem like to be overfit. Both gradient boosting methods outperform the Logistics model in every metric. When comparing the performance of each model, the CatBoost model have the highest score in overall.

* 1. **Features important**

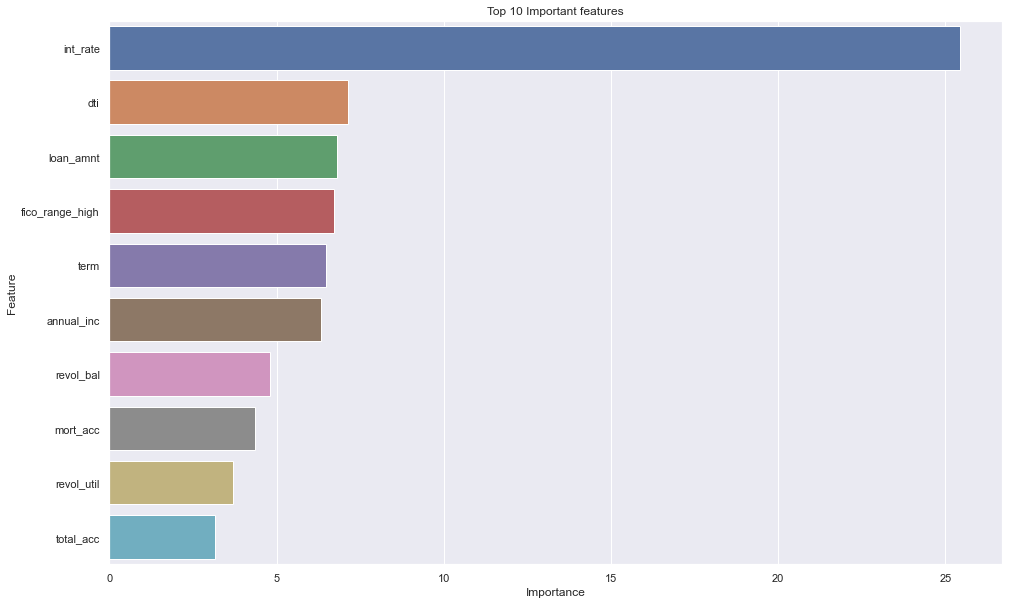


Figure 24: Features important score from CatBoost model

Among all features we selected 10 with the largest importance values (see above). The top 3 features are int\_rate, dti and loan\_amnt. The importances of dti and loan\_anmt, however, are quite close to each other and the rest of the features so this ranking might slightly change for a different train-test split. We will take another closer look at these important features.

|  |  |
| --- | --- |
| Figure 25: Distribution of int\_rate | Figure 26: Distribution of dti |
| From the Distribution plot for the feature int\_rate (top1) the loans is more likely to be returned if the loan int\_rate is lower (specificly lower than 15) . This is logical because lower interest typically led to lower monthly payments, which are easier for borrowers to repay. Others features also share a similar trend, loan of which borrower whom has lower dti and loan\_amnt usually be good loans. This also make sense, because lower debt-to-income ratio it is means borrower’s debt at a manageable level and lower loan amount led to less money and time borrowers have to spend. | Figure 27: Distribution of loan\_amnt |

1. **Conclusion**

This project, we utilize a very large sample of loans (1.3 million observations) to inspect the factors influencing the risk of loan default and create model to predict Default Customer. Overall result table shows that model performance can be improved using tree-based ensemble models rather than Logistics Regression (traditional method). This implies that banks and lenders can utilize these models that may assist them in cutting down costs resulting from loan defaults, as well as detecting and steering clear of risky lending prospects. Result from Cat Boost model shows that high interest rate, large loans and high debt-to-income borrowers are associated with a higher risk of default. Our result has important practical implications. First, Commercial banks and investors are suggested to pay more attention to large loans and high interest rate loan. A stricter credit approving procedure may need to be developed for these types of loans in order to better evaluate their risk of default. Second, customers are typically concentrated in the western and southern regions of the country. Nevertheless, the eastern and central states, where customer numbers are lower, have the highest rates of default. It is advisable for banks to establish loan policies to manage lending in these areas.

Our paper has some limitations. We were unable to take into account certain customer traits such as educational level and marital status due to our restricted access to the financial institution. Furthermore, our data solely originated from one lending institution, and therefore, may not be representative of the credit evaluation procedures at other banks or lenders. We hope to expand this project in the future with more extensive data as they become accessible.

1. **References**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in neural information processing systems (pp. 6638-6648).

Özdemir, Ö. and Boran, L. (2004), An empirical investigation on consumer credit default risk, Turkish Economic Association Discussion Paper, No. 2004/20.

Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained (https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725)