**Final Project Research Paper work**

Northeastern University College of Professional Studies, Analytics Program

ALY6110: Big Data and Data Management

Project name: Loan Prediction

CRN: 20434

**Group Epsilon**

Yiwei Wei

Email: [wei.yiw@northeastern.edu](mailto:wei.yiw@northeastern.edu)

Jiaming Ma

Email: [ma.jianm@northeastern.edu](mailto:ma.jianm@northeastern.edu)

Sahil Gawande

Email: gawande.s@northeastern.edu

Feb 14, 2023

**Dataset background**



The datasource is from [LendingClub site](https://www.lendingclub.com/info/download-data.action), which also can be seen in Kaggle. LendingClub is a US peer-to-peer lending company and the world's largest peer-to-peer lending platform. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee. Above is the introduction of LendingClub works [1].

**Summary**

The whole project is built on **Databricks** platform, and the programming language used is **Python**. In terms of dataset, there are 1.6 million rows and 150 variables. The size of the dataset is 1.8 GB. The purpose of this project is to learn to clean dirty loan data, perform exploratory analysis and later predictive modeling on loan data. **The purpose of the modeling is to predict the probability that a loan will be paid or charged off**. In general, before embarking on a big data project, we need to have a framework that includes the choice of big data environment, the choice of programming language, the business purpose of the project, data cleaning, data processing, data analysis and visualization, predictive modeling, etc. And all our visual graphing is based on Python internal libraries like matplotlib and seaborn. We will build on these components to achieve our goals. Here are our project’s goals:

* Demonstrate how to use databricks to do the whole project, including importing files, opening cluster, starting, etc.
* Cleaning of data from Loan 2007-2017
* Data pre-processing and Exploratory Analysis
* Feature engineering and building three machine learning models to predict the probability that a loan will be paid or charged off and compare each model’s performance

**Research**

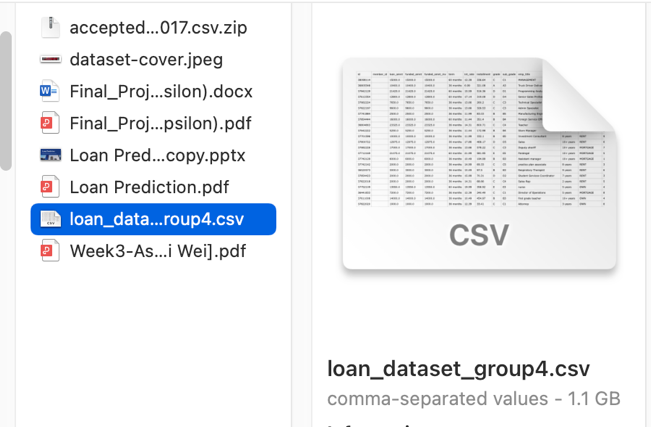
This is our first time doing big data analysis and prediction about loans, so we don't know many proper nouns in it, so we summarized a descriptive table of some key variables after LendingClub official website and google query.

|  |  |  |
| --- | --- | --- |
| Data fields | Data type | Data description |
| id | int | To identify the unique borrower. |
| loan\_status | chr | "Fully Paid" or "Charged Off." |
| loan\_amnt | num | The listed amount of the loan applied for by the borrower. |
| term | chr | The number of payments on the loan. Values are in months and can be either 36 or 60." |
| installment | num | The monthly payment owed by the borrower if the loan originates. |
| grade | chr | LendingClub assigned loan grade. |
| sub\_grade | chr | LendingClub assigned loan subgrade. |
| emp\_title | chr | The job title supplied by the Borrower. |
| emp\_length | chr | Employment length in years. |
| home\_ownership | chr | The home ownership status provided by the borrower during registration or obtained from the credit report. |
| annual\_inc | num | The self-reported annual income provided by the borrower during registration. |
| verification\_status | chr | Indicates if income was verified by [Lending Club], not verified, or if the income source was verified. |
| issue\_d | time | The month which the loan was funded. |
| purpose | chr | A category provided by the borrower for the loan request. |
| title | chr | The loan title provided by the borrower. |
| dti | num | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| Int\_rate | num | Interest Rate on the loan. |

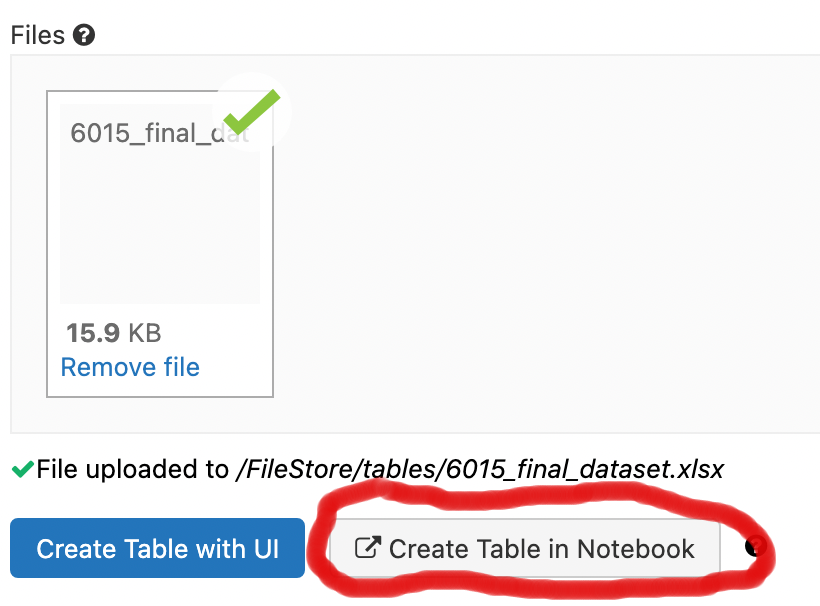
**Comments**

**1.Data Inputing**

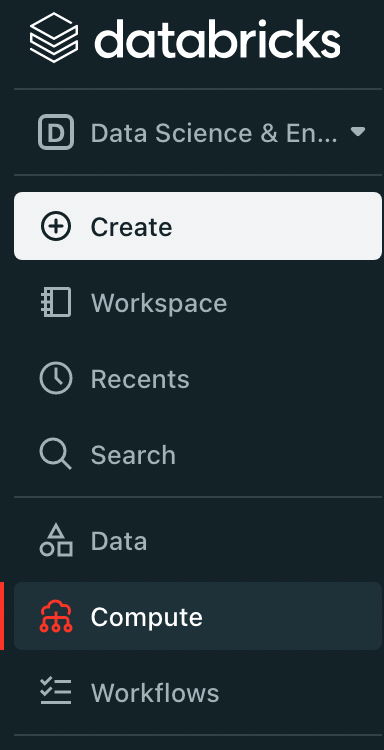
Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

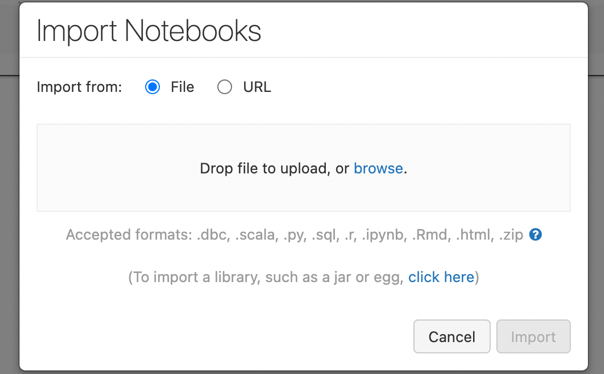
First we enter the databricks interface home page, click on the left toolbar we can find the "Data" column, after clicking on it, you can see a Create Table button in the upper right corner. excel file into the download task, wait for about 10 min to complete the data set input. After the input, we can notice that Create table in notebook”, the system will automatically return you to a notebook, which means you can finish the data analysis and calculation in one interface. But at this point we still need to create the cluster.

Graphical user interface, text, application

Description automatically generated

After we finish entering the dataset, we also need to create a cluster to support the subsequent data analysis and computation, which is equivalent to a server and gives you a place in the huge cloud server. First, click the "Compute" button on the left toolbar, then click Create Cluster, enter the cluster name and the desired databricks cluster version (Runtime: 7.3 LTS) because there is no limit for this version, other versions have a limit of 2.

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, application

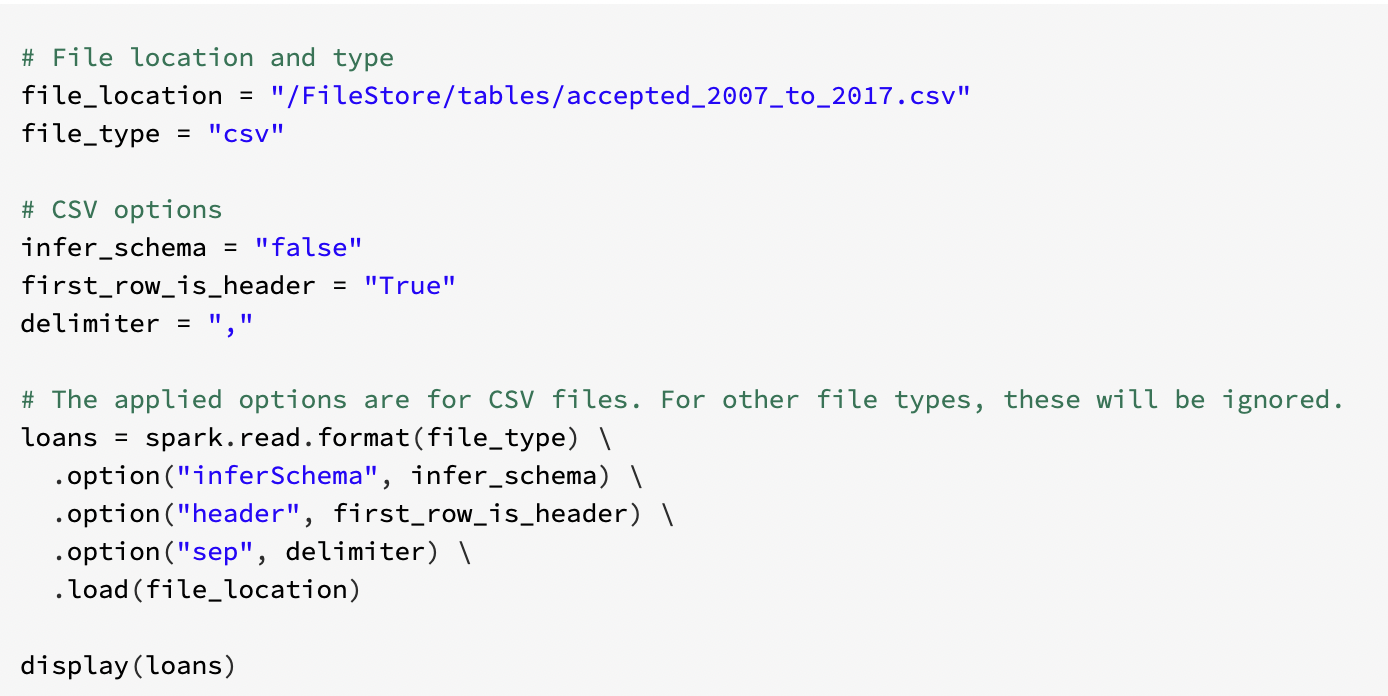
Description automatically generated

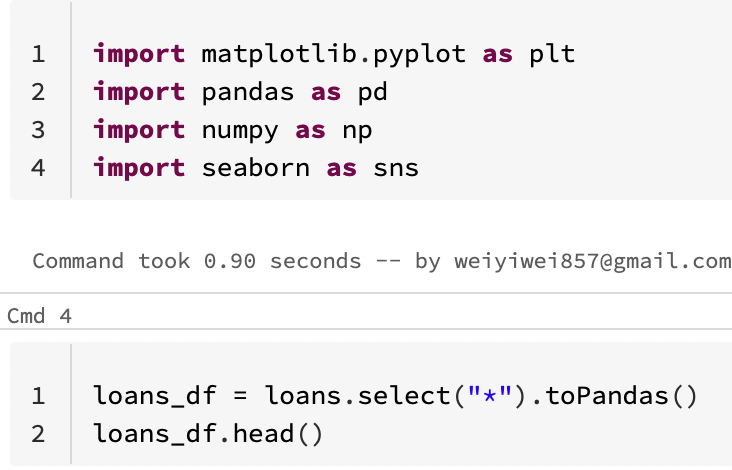
After creation, for professor operation, you just need to go back to the notebook page, click Import notebook inside file, import our .ipynb file, the transfer time is about 1-2min.

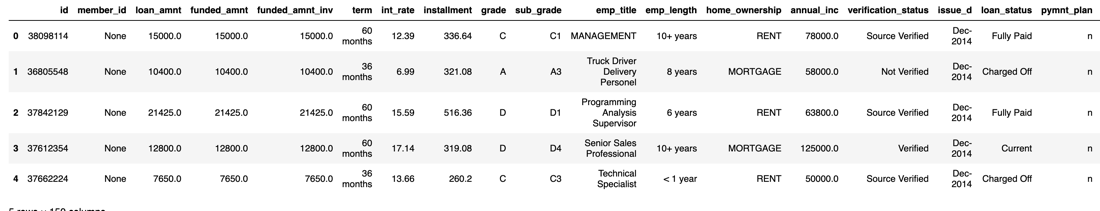
**Graphical user interface, text, application, chat or text message

Description automatically generated**

After the file transfer, click on the cluster you just created, then click on run all and all the codes will be run. Due to the large dataset, the entire code runs in about 5-8 min. Next is the data cleaning part.



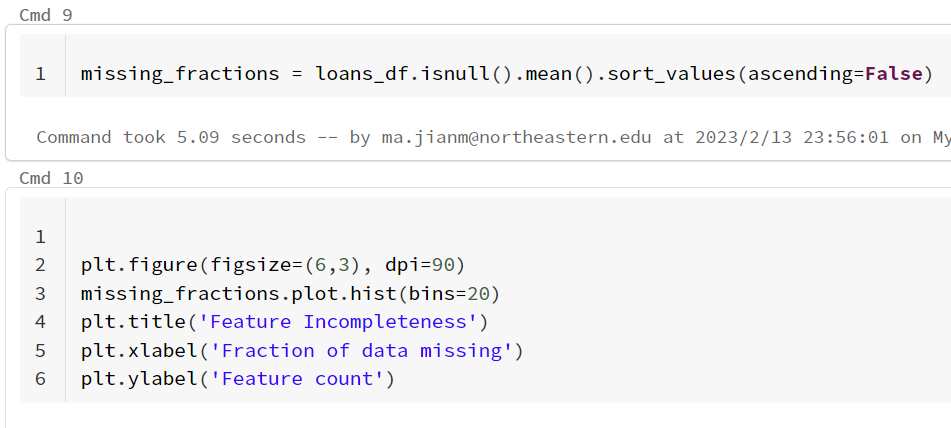




In this case, you can notice that there are a bunch of codes that Databricks has runed, and also we can see the data table.

**2. Data cleaning**

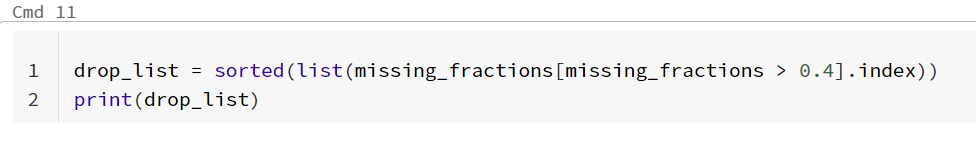
2.1 Missing value checking



Chart, histogram

Description automatically generated

Missing value is the most common unit that needs to be cleaned in our daily life. Without reasonable cleaning of missing value, our results are likely to be influenced by it and produce scenarios that deviate greatly from the actual results. First we use the isnull function to create a variable for the percentage of missing values, and then visualize this variable to come up with this scenario in the figure below. In this case, we decided to remove the features with a missing proportion greater than 0.4, which can relatively reduce the number of variables we need to process. After this operation, the number of features has been reduced from 150 to 92.





2.2 Duplicate value checking

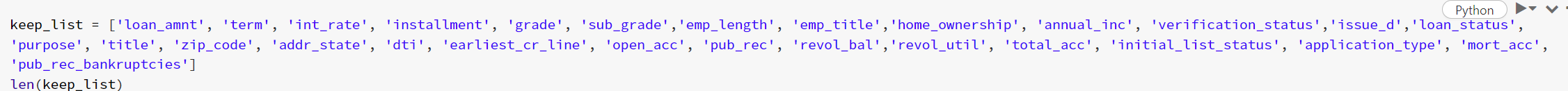
Graphical user interface, application

Description automatically generated

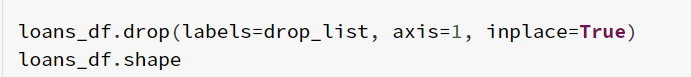
Duplicate value is also a unit that we need to clean, if we do not clean up the duplicate value in time, it will have a great impact on the subsequent analysis and modeling. After checking, we found that there were no duplicate values, so we did not process them.

2.3 Feature selection

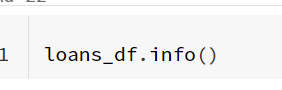
Since the number of variables is 92, it is still a large number. For the subsequent data analysis, I need to combine the actual business scenarios and official recommendations to select indicators within 30. According to the importance of indicators in actual business scenarios, we selected 28 valuable indicators.



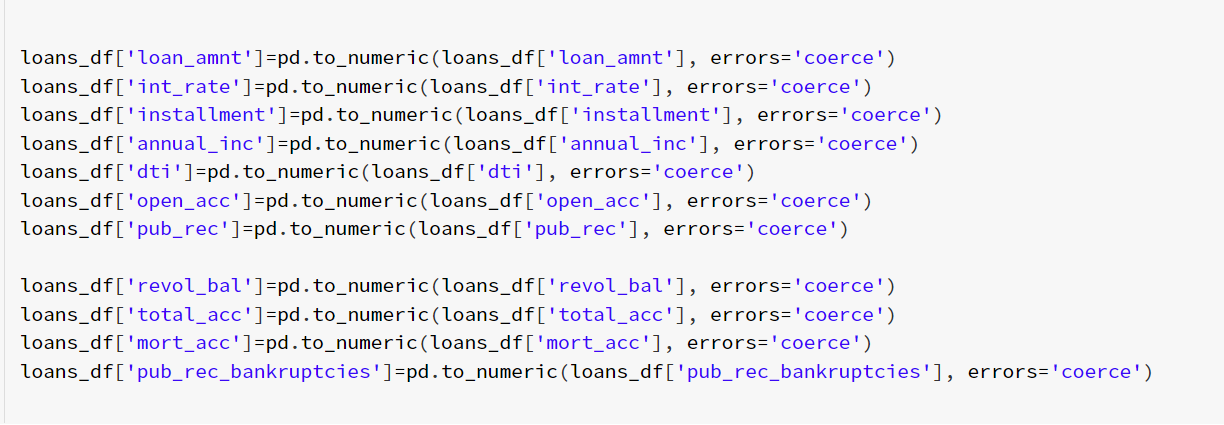




2.4 Data types checking



Table

Description automatically generated

Checking data types is the most important part of data cleaning. If the data type of a variable does not match its nature, the computer may misidentify it, which will then have a great negative impact on subsequent calculations and modeling. First of all, we use info() to check the data types and find that all variables are object, which is obviously wrong. We picked out the features that belonged to numerical variables according to the actual business scenario, and then converted them to data types to reach the following changes.

Table

Description automatically generated

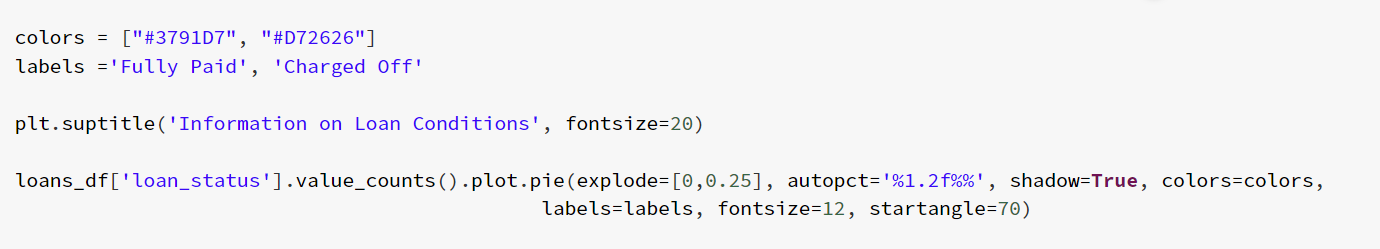
**3. Explortary analysis**

In this section, we will expand on the relationship between several important variables and response variable. Most importantly, we have created a visualization funtion in order to achieve suitable plots in a easy way. Not only applied for continuous data but also for categroical data. Nearly all variables will be concentrate on the loan status. Here is the funtion screenshoot.

Graphical user interface, text, application, email

Description automatically generated

3.1 Response variable



Chart, pie chart

Description automatically generated

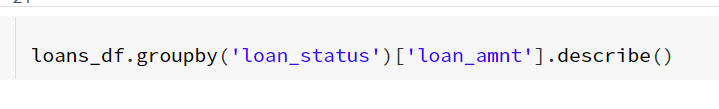
Since the purpose of the modeling project is to predict the probability that a loan will be paid or charged off, our response variable is loan\_status, which includes many categories, such as Current, Fully paid, Charged off, and Charged off. For example, Current, Fully paid, Charged off, Late (31-120 days) and several other categories. Considering the actual business situation, we will only consider the completed loan transactions, so we will only consider the Fully paid and Charged off categories, which also simplifies our model. Below is a chart of the composition of these two categories, with Fully paid people accounting for up to 80% and Charged off people for about 20%. Since our outcome is a categorical variable, in the next analysis, if the feature is a numerical variable, we will use the distribution of numerical variables, the boxplot of numerical variables and categorical variables as the display relationship graph, and also import more detailed statistical summary of the description, which will allow us to analyze in both macro and micro perspectives This will allow us to analyze what is happening at both macro and micro levels.

3.2 Loan amount

Data Dictionary: “The amount the borrower promises to repay, as set forth in the loan contract.” [2]

A picture containing icon

Description automatically generated

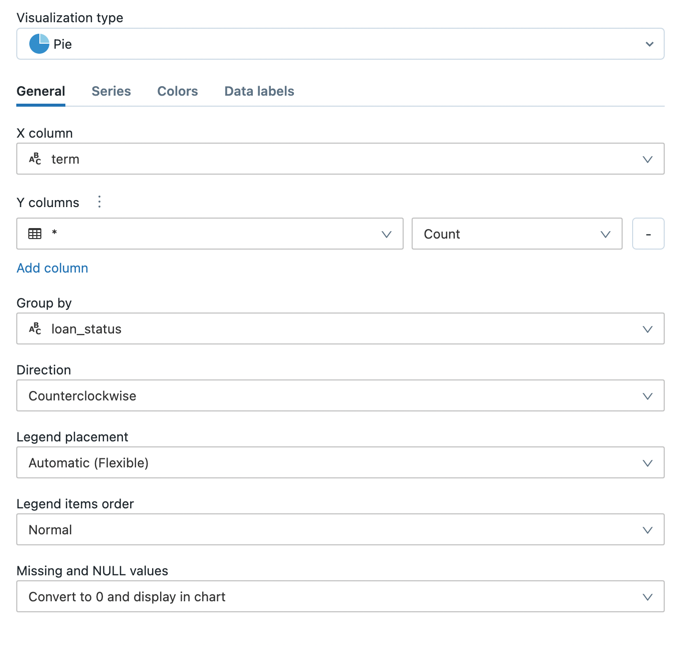
Table

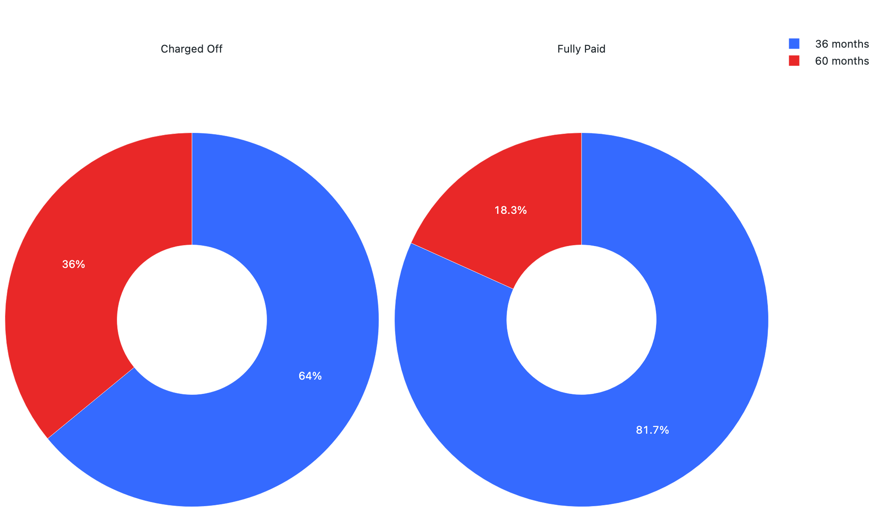
Description automatically generated

Regarding the relationship between Loan amount and loan\_status, first look at the data distribution of Loan amount, approximately Loan amounts range from $500 to $40,000, with a median of $12,000. Then look at the relationship chart, we can find that the charged-off Then we can verify through the statistical table that indeed the average loan amounts of charged-off people are about 1300 higher than those of fully-paid people, and the median is about 2000 higher. Therefore, it can be shown that charged-off loans tend to have higher loan amounts.

3.3 Term

Data Dictionary: "The number of payments on the loan. values are in months and can be either 36 or 60." [3]

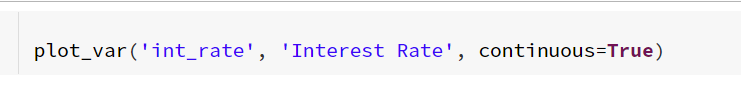




Regarding the relationship between Term and loan\_status, first look at the distribution of Term data, about 76% of people choose 36 months loans and 24% choose 60 months loans. Then after sorting about loan status, it is obvious that those who choose 60 months loan are more likely to charge off, accounting for about 34%, while those who choose 36 months only account for about 16%. I also screened 1000 random records in databricks and used the built-in visualization tool to show this percentage, which is basically the same as the overall results. This suggests that Loans with five-year periods are more than twice as likely to charge-off as loans with three-year periods.

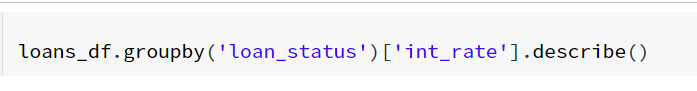
3.4 Interest rate

Data Dictionary: "Interest Rate on the loan." [4]



A picture containing icon

Description automatically generated



Table

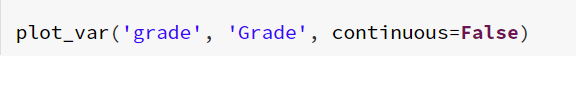
Description automatically generated

First of all, if we look at the data distribution of interest rate, we can see that it tends to be normal, with the mean around 13.5 and the median around 13, ranging from 5.3 to 30. Then, after the classification of loan status, we can find that charged-off people have a significantly higher interest rate. Then, the statistical list verifies that the interest rate of the charged-off people is indeed about 3% higher than that of the fully-paid people, and the median is also about 3% higher. Therefore, it can be shown that the higher the interest rate, the higher the charge-off rate.

3.5 Grade and subgrade

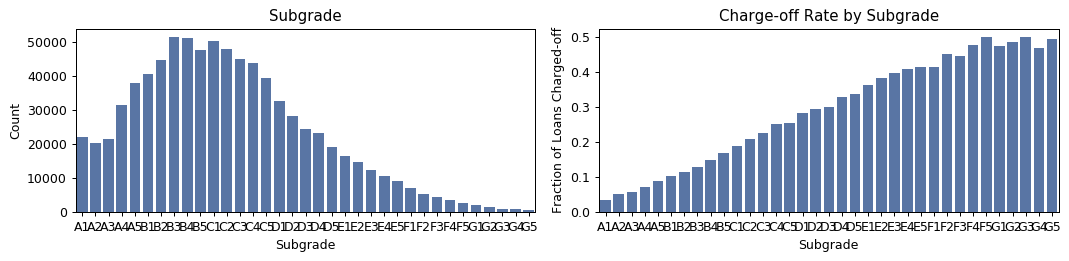
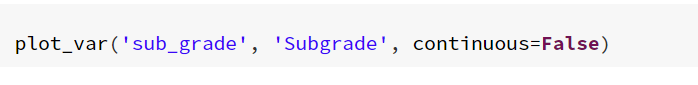
Data Dictionary for grade: "LendingClub assigned loan grade."

Data Dictionary for sub\_grade: "LendingClub assigned loan subgrade."



A picture containing logo

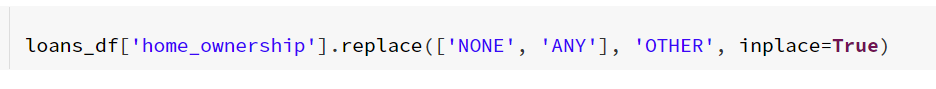
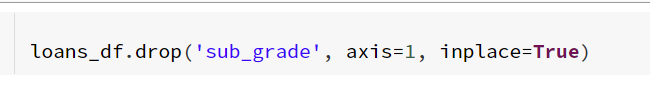
Description automatically generated



Regarding the relationship between Grade and loan\_status, first of all, look at the data distribution of Grade, 55% of people's loan ratings are concentrated in B and C, and only 16% of people have A loan ratings. This percentage can also be shown in the subgrade. Then after the classification of loan status, it is obvious that the lower the transaction rating, the higher the charged-off ratio, which is a very obvious negative correlation. This means that the lender's transaction ratings are high and accurate.

3.6 Home ownership

Data Dictionary: "The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER." [5]





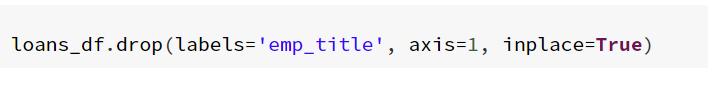
Logo, icon

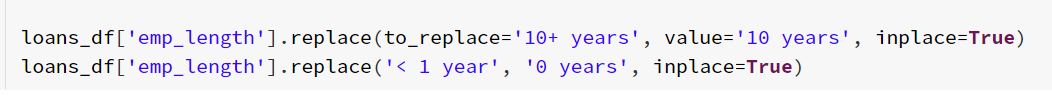
Description automatically generated with medium confidence

First of all, regarding the distribution of home ownership data, it is obvious that mortgage has the highest percentage, about 50%, followed by rental, with a percentage of about 40%. The third is owning a home, with a ratio of only 10%. Next, after the classification of loan status, it can be seen that the probability of charge-off is the highest for those who rent, about 25%; the next probability is the highest for those who own their own homes, about 21%. The charge-off rate for mortgage holders is relatively low, around 17%. Thus, there appear to be large differences in charge-off rates by home ownership status. Renters and homeowners have a higher probability of charge-off.

3.7 Employment length

Data Dictionary: "Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years."









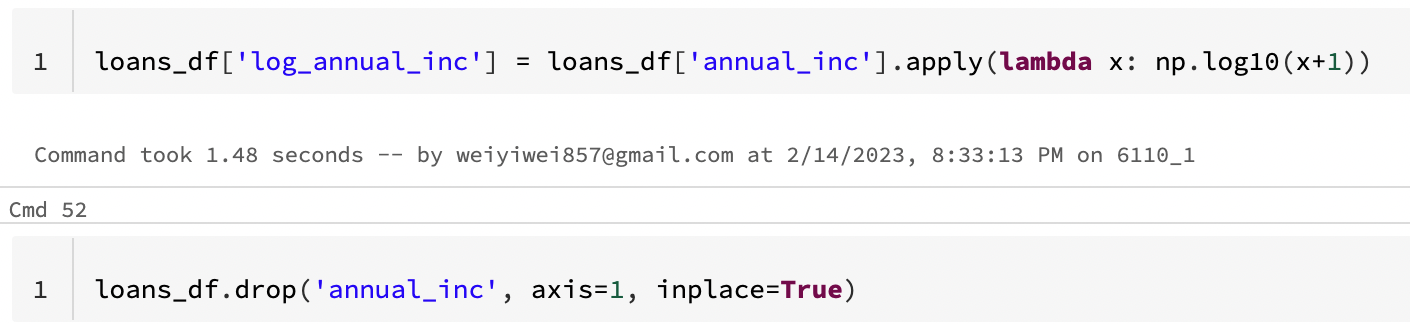
**Icon

Description automatically generated**

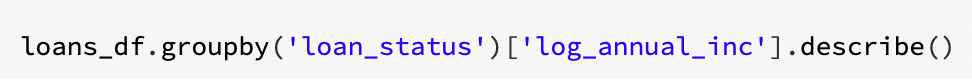
First of all, regarding the data distribution of the length of employment, we can see that in the loan population, most of them are people who have been in the industry for more than 10 years, the proportion is about 32%, followed by people who have no experience in the industry, the proportion is 13%, and the rest of the proportion is almost the same. After the classification of loan status, we can see that the probability of charge-off is the largest for people with no experience, and the probability of charge-off is the smallest for people with more than 10 years of experience, which is the largest group. In other words, the less work experience, the higher the charge-off rate.

3.8 Annual income

Data Dictionary: "The self-reported annual income provided by the borrower during registration."[6]







**Icon

Description automatically generated**

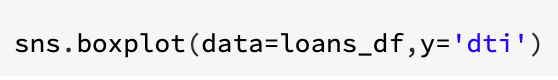
**Table

Description automatically generated**

Since the variable, annual income, is relatively large in number, the overall data distribution will be susceptible to extreme values. Therefore, we use the log function to transform it, and the data distribution is like the above figure. Then, after the classification of loan status, we can find that the charged-off people usually have low annual income, and then we can verify through the statistical table that the average annual income of charged-off people is about 0.4 lower than that of fully-paid people, and the median is also about 0.4 lower. Therefore, the higher the annual income, the lower the charge-off rate.

3.9 Debt-to-Income Ratio

Data Dictionary: "A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income."[7]

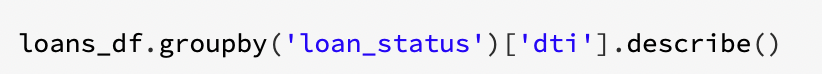


A picture containing chart

Description automatically generated

First of all, looking at the data distribution of dti separately, we can see that it has a lot of extreme values that greatly affect the data distribution. So I decided to check the upper limit of the data (Q3 + 1.5IQR), and also check how many extreme values are above the upper limit. After checking the number of extremes, I decided to remove them.





Chart, histogram

Description automatically generated

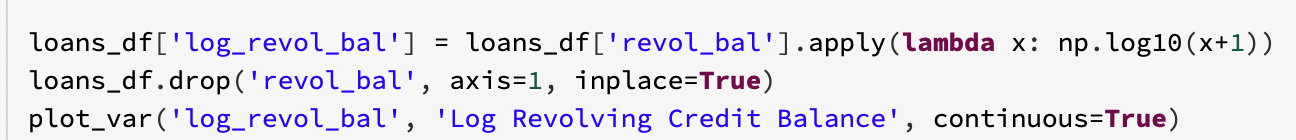
Table

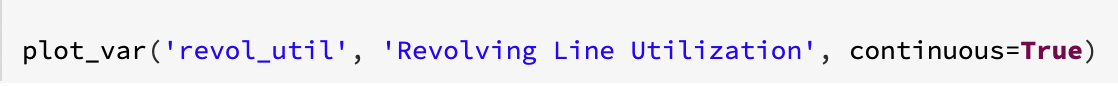
Description automatically generated

After that, the data distribution is shown on the right, and the data tends to be relatively normal. Next, the dti difference between the two types of users was viewed by sorting their loop status through the statistics list. We can clearly see that the charged-off people have a higher dti, with a mean value about 2.7% higher than the fully-paid people and a median value of 3.0% higher. Therefore, there appear to be large differences in charge-off rates by dti. Higher dti have a higher probability of charge-off.

3.10 Revolving balance

Data Dictionary: "Total credit revolving balance."[8]





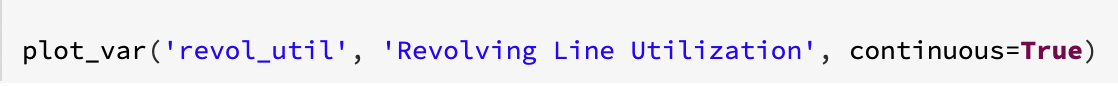
Icon

Description automatically generated with medium confidence

This variable was also transformed into a log function due to its large value. After the transformation, it is classified according to the loop status, and it can be seen that the difference between the two is not very large.

3.11 Revolving line utilization rate

Data Dictionary: "Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit."[9]

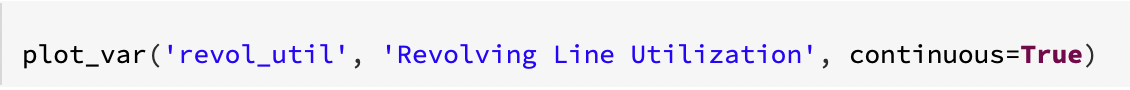


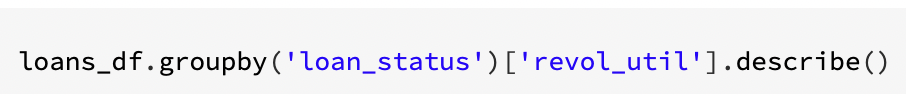
A picture containing icon

Description automatically generated

First of all, looking at the data distribution, we can see that the distribution of the variable is greatly influenced by the extreme values, because we decided to look for the upper limit of its distribution and see that the number of extreme values is only 104, and decided to remove the extreme values







A picture containing icon

Description automatically generated

Table

Description automatically generated

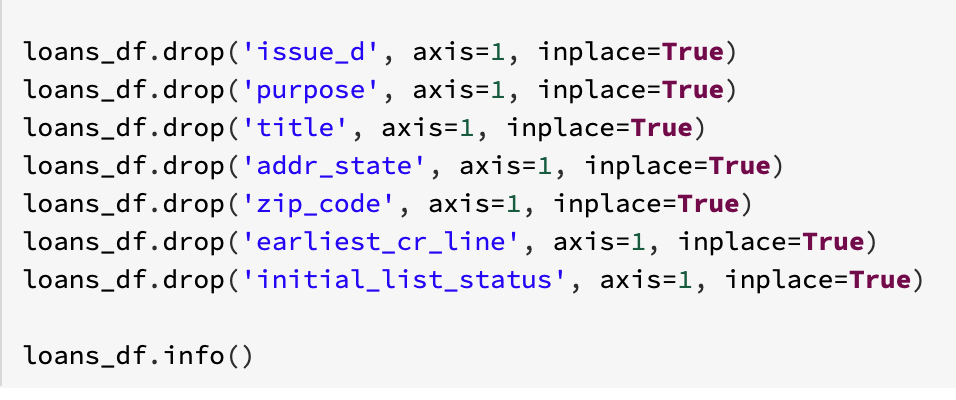
After data cleaning, it can be seen that the data distribution of this variable tends to be normally distributed. Also regarding the classification of loan status, we can find that charged-off people usually have higher revolving utilization rate. then verified by statistical tabulation, indeed the average revolving utilization rate of charged-off people is about 4% higher than that of fully- paid people are about 4% higher and the median is about 5% higher. Thus, it can be shown that there appear to be large differences in charge-off rates by revolving utilization rate. higher revolving utilization rate have a higher probability of charge-off.

**4. Predictive Analysis**

4.1 Feature enigeering

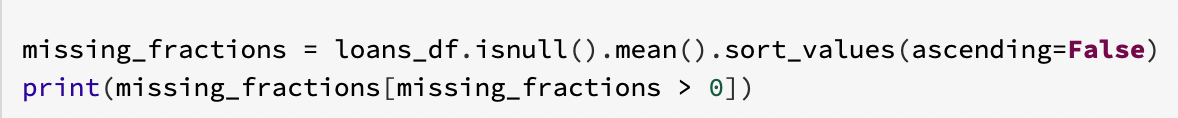
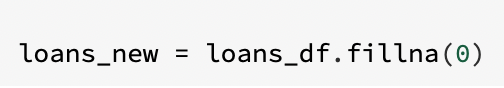
Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features [10].

4.1.1 Removal of useless variables



Since there are too many categories and duplicates, it was decided to delete subgrade, emp\_title and title. due to the platform failure, it was decided to delete issu\_d and earliest\_line. since there is no practical business meaning, it was decided to delete zip\_code. since there is little relevance to the result variables, it was decided to delete initial\_list\_status variable.

4.1.2 Transformation variables

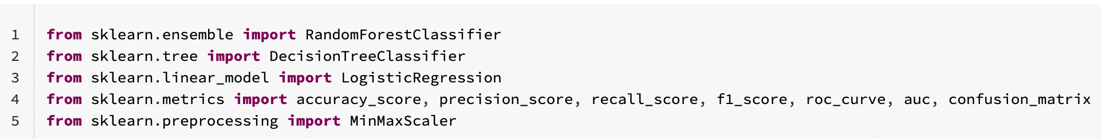


* Since addr\_state has more categories and we can divide it into five regions, namely west, southwest, southeast, midwest and northeast.
* Since the model modeling is sensitive to missing values, by looking at the proportion of missing values, we found that the variables are all numerical variables, so we felt to fill in the missing values as 0.
* Since the values of categorical variables could not be applied to the model, we decided to transform all categorical variables into dummy variables.

In summary, after the above feature processing and cleaning, the shape of the remaining dataset is (814336, 46).

4.2 Pre-modeling work

* Imputing packages and classification evaluation funtion

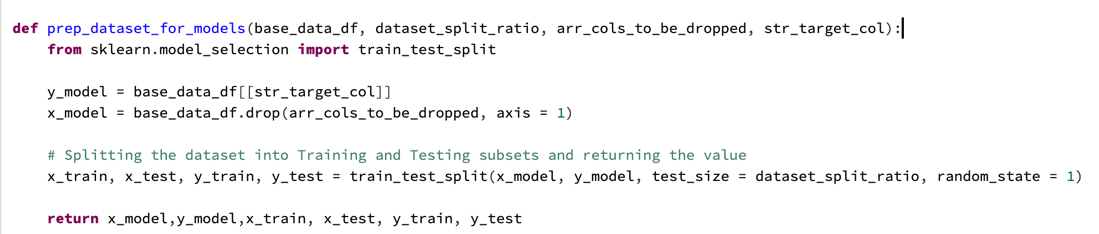


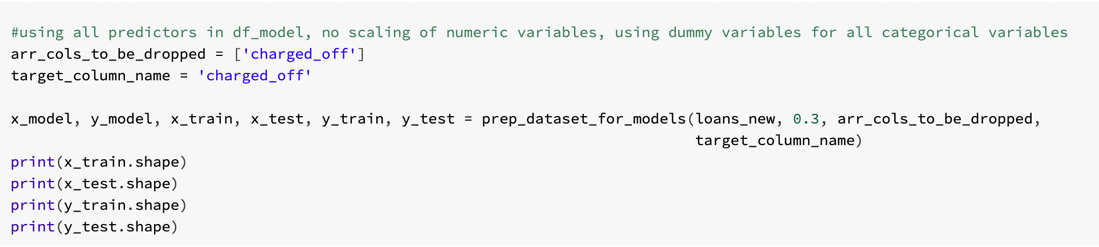
Graphical user interface, text, application, email

Description automatically generated

Since we will then divide the dataset, normalize the data, and use the classification model to model and evaluate the model parameters. Therefore, we need to call the appropriate machine learning libraries, including RandomForestClassifier, DecisionTreeClassifier, LogisticRegression, metrics and MinMaxScaler. Besides, we can also use the eval\_classfication function to meaure model’s performance by computing all factors.

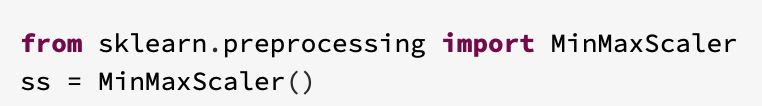
* Train and test data split





This project will randomly assign the training and test sets in a 7:3 ratio.

* Data Standardization



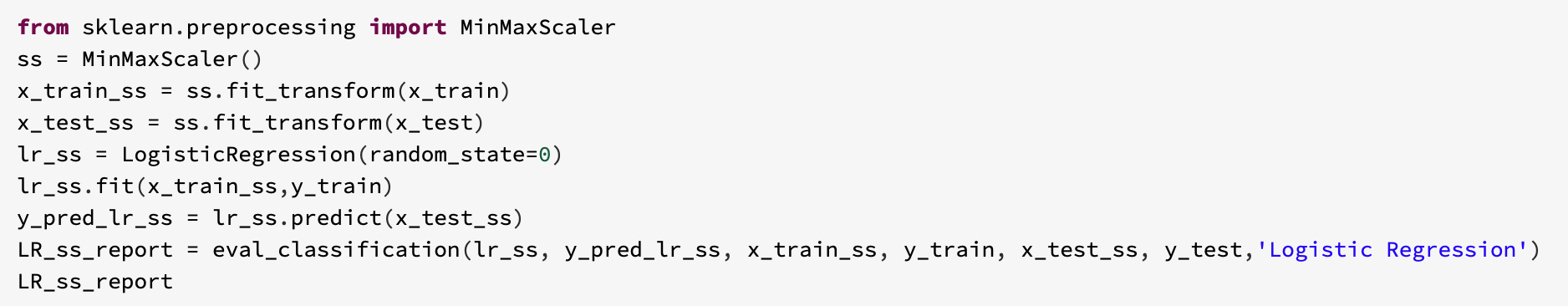
We will use MinMaxScaler() to normalize the data so that data with different specifications are converted to the same specification. Why normalization and standardization? Because the units or sizes of features vary widely, or the variance of a feature is several orders of magnitude larger than other features, which can easily affect (dominate) the target result. Normalizing the data is done so that data with different specifications are converted to the same specification.

4.3 Classification model building

* Model introduction

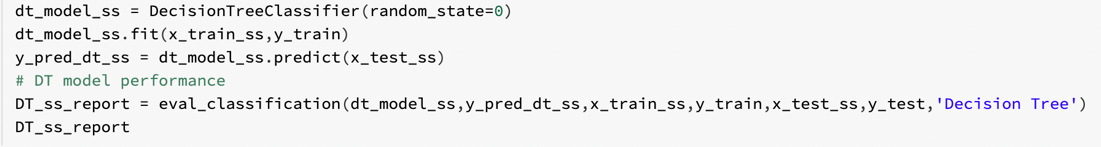
The three models that were chosen were logistic regression, decision tree, and random forest. We chose these because they are all classification models that we have studied and built in other classes. They all have reasonable computation requirements and are more easily interpreted than other models like neural networks. They also all provide information about feature importance, which is useful to be able to understand which independent variables have the biggest impact on the target variable [11]. For each model, we built the simplest version first by using all predictor variables from our cleaned dataset and leaving all parameters at their default values.

*Logistic Regression*



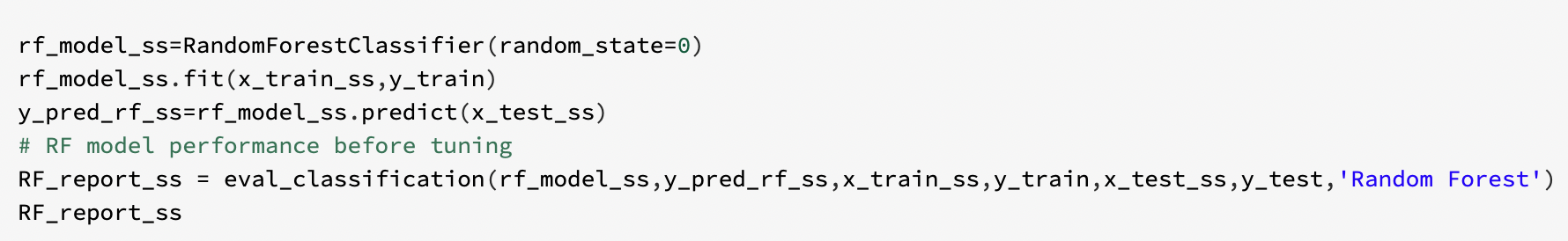
Logistic regression is a type of generalized linear model that is used for binary classification from a set of numeric and/or categorical explanatory variables. Logistic regression fits a logistic sigmoid function to a dataset in order to develop a probability prediction for each data point. It then compares each probability to a predefined threshold probability to make a prediction of 0 or 1 [12].

*Decision Tree*



Decision Tree is a supervised learning technique that can be used for both classification and regression problems, but it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. [13].

*Random Forest*



Random forest is a machine learning algorithm that is used widely in classification and regression problems. It builds many decision trees on different samples from the training data and takes their majority vote for classification and average in case of regression. Each node of a given tree only considers a subset of the features. Due to these attributes, random forest tends to reduce overfitting compared to a single decision tree. Another strength of the random forest algorithm is that it can handle binary, categorical, and numerical features, which minimizes the pre-processing needed [14].

* Model comparison

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Regression | Decision Tree | Random Forest |
| Train score | 0.796 | 1.0 | 1.0 |
| Test score | 0.796 | 0.638 | 0.795 |
| Accuracy | 0.796 | 0.638 | 0.795 |
| Recall | 0.095 | 0.382 | 0.117 |
| Precision | 0.538 | 0.252 | 0.515 |
| F1 Score | 0.161 | 0.304 | 0.191 |
| AUC | 0.537 | 0.544 | 0.544 |

If we only look at accuracy, we are not fully understanding the performance of the model. It is useful to check precision, recall, and AUC space to see whether the model struggles to accurately predict positive or negative cases more. Precision is the percentage of positive predictions that were correct, and the highest score is is 79.6% for logistic regression model. AUC space is the area under the ROC curve is a composite measure of the effect of all possible classification thresholds. One way of interpreting the area under the curve is to view it as the probability that the model will rank some random positive category sample over some random negative category sample. In the actual loan scenario, we need to consider three issues.

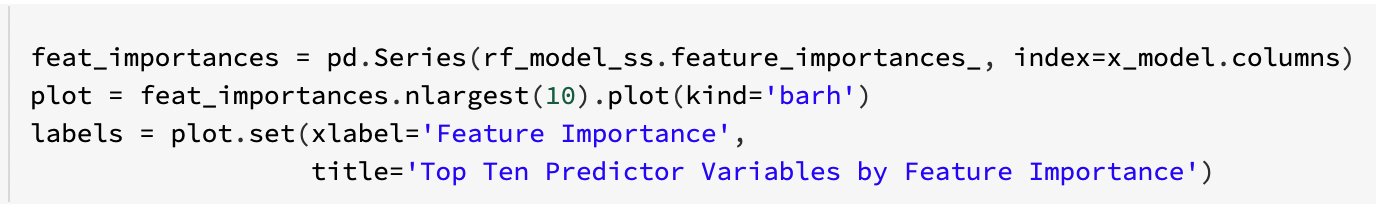
1. The greater risk of identifying the person who is actually going to write off the loan as fully-paid means that the first type of error is even more intolerable.

2. The sample is relatively unbalanced, with about 20% of the percentage of pin payments and 80% of the percentage offully paid.

3. Overfitting is not allowed.

In summary, precision and AUC are the next main metrics we refer to. Base on that, at first, as the train score is 1.0, decision tree model and random forest model are overfitting. Next, in terms of precision, Logistic regression model > Random forest model > decision tree model. Besides, in terms of AUC, three models’ value are nearly same. Therefore, logistic regression model is the best model compared to other two models.

* Feature importance



Chart, bar chart

Description automatically generated

This graph shows that dti is the most influential feature and contributes the most to the accuracy of the model. This was the numeric variable that we found to show a significant difference in distribution between charge-off and fully paid during our EDA process. The loan interest rate and revolving line utilization rate variable were also found to be a good quality predictor in our EDA process, so this being in the top four here is in line with our expectations. Feature importance helps us to interpret the meaning of the results in the context of the data, giving us a deeper understanding. It also can help us in terms of feature selection as we build a second model. Fewer, stronger features can improve the efficiency and effectiveness of the model.

**Comments**

Looking back on the whole project process, we learned a lot of new things and encountered a lot of difficulties, and this experience is undoubtedly of high value to our subsequent study and work. First of all, let's talk about our advantages

1. Volume of data and number of variables meet operational criteria

2. The data sources and variables explained by multiple sources of information and high credibility

3. The analytical framework of the project is widely applied and conducive to the accumulation of project experience

4. The Databricks community version requires only a registration, then you will have a Spark cluster environment configured with 6G RAM. As a Spark beginner, you no longer have to worry about configuring your development environment.

Next here are what we learned:

1. How to use big data platform Databricks

Through this project of using big data platform to do data analysis and prediction, we learned about the components of big data platform, such as how to import and manage files, create cluster, etc.

2. Advanced data analysis and prediction capabilities

Through this data analysis of LOAN dataset, we have improved both from data cleaning perspective to data visualization and mastered more skills of cleaning and visualization.

3. Knowledge of loan finance

In order to understand the dataset, you have to go and understand the meaning of each variable clearly so that you can filter and analyze it reasonably. I benefited a lot from gaining some general knowledge about loan knowledge by searching on the lending club official website and Google.

In addition to the advantages and our gains, we encountered a lot of difficulties in the process of doing the project, for example:

1. Bugs in platform

Such as some visualization of the graph can not be loaded out, can not intercept the time format string and other problems, its led to the project stalled in the middle. But our solution is to use statistical description methods instead of graphs, with time variables as we can not handle we can only delete, in addition the variable does not affect us much.

2. Limitations of visualization

Since the visualization that comes with the Databricks platform is limited to only 1000 rows of data, this was very painful for our huge dataset. The solution is that we chose to import the Python open source visualization libraries, matplotlib and Seborn.

3. Variable filtering

This is a complex and large project, we need to combine practical business knowledge and past application cases to filter the critical variables, although there are some more scientific methods to achieve this purpose, such as the statistical significance test. Since the dataset includes 150 columns, there are many unimportant columns in it. Therefore we only keep loan features known to potential investors, and this information needs to be queried by us on their official website.

4. Data cleaning

The most important point about this issue is how to define the data to be cleaned, generally speaking, the null, duplicate and extreme values. But in more detail, what percentage of columns need to be deleted if the percentage of null values exceeds? What is the range of extreme values? All these questions need to be decided in the context of business purposes and practical application scenarios.

**Conclusions summary**

The above can be summarized as the following insights:

* Fully paid vs Charged off = 80% : 20%
* Charged-off loans tend to have higher loan amounts, interest rate, dti and revolving utilization rate, lower loan grade, employment length and annual income.
* Five-year periods are more than twice as likely to charge-off as loans with three-year periods.
* Renters and homeowners have a higher probability of charge-off.
* Base on precision rate and AUC space, Logistic regression model is the best model compared to other two models.
* The most important variables for predicting charge-off are the dti, loan interest rate, revolving balance and evolving line utilization rate.

In conclusion, we have been looking for a long time in kaggle to find a dataset that meets the criteria, and we also wanted to do a data analysis project on loan for a long time, and we think this project is a good opportunity to get a deeper understanding of the financial or insurance industry of loan forecasting and understand the key and hidden variables. The purpose of this project is also very practical and commercial, which is to predict the probability that a loan will be paid or charged off. At the same time, the most important thing is that this project will give us a deeper understanding of the architecture of the Big Data platform and the role of each layer. This will help us to handle very large data sets in the future and to know how to use the Big Data platform to solve problems that cannot be solved by traditional databases. In short, if our group embarks on this project, we can practice many skills, such as Python language, business analysis thinking, information search ability and modeling knowledge expansion. All of these will be very substantial help for our future employment. Meanwhile, thanks to Prof. Valeriy's guidance for more than 1 month, we learned a lot about big data platforms and about computer expertise.

**References**

[1] Nathan George. (n.d.) All Lending Club loan data. Retrived Jan 30, 2022 from[**https://www.kaggle.com/datasets/wordsforthewise/lending-club?select=accepted\_2007\_to\_2018Q4.csv.gz**](https://www.kaggle.com/datasets/wordsforthewise/lending-club?select=accepted_2007_to_2018Q4.csv.gz)

[2] Loan Amount. (n.d.) *The Mortgage Encyclopedia*. (2004). Retrieved February 11 2023 from <https://financial-dictionary.thefreedictionary.com/Loan+Amount>

[3] Term. (n.d.) *Financial Glossary*. (2011). Retrieved February 11 2023 from <https://financial-dictionary.thefreedictionary.com/Term>

[4] Interest rate. (n.d.) *Financial Glossary*. (2011). Retrieved February 11 2023 from <https://financial-dictionary.thefreedictionary.com/interest+rate>

[5] Home ownership. (n.d.) *Farlex Financial Dictionary*. (2009). Retrieved February 11 2023 from <https://financial-dictionary.thefreedictionary.com/Home+Ownership>

[6] Annual income. (n.d.) *Farlex Financial Dictionary*. (2009). Retrieved February 12 2023 from <https://financial-dictionary.thefreedictionary.com/Annual+Income>

[7] DTI. (n.d.) *Farlex Financial Dictionary*. (2009). Retrieved February 12 2023 from <https://financial-dictionary.thefreedictionary.com/DTI>

[8] Fidelity. (n.d.) What is a revolving balance? Retrieved February 12 2023 from <https://www.fidelity.com/insights/personal-finance/glossary-revolving-balance>

[9] Experian. (n.d.) What is a Credit Utilization Rate? Retrieved February 13 2023 from <https://www.experian.com/blogs/ask-experian/credit-education/score-basics/credit-utilization-rate/>

[10] Harshil Patel. (Aug 30, 2021). What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. Retrieved February 13 2023 from <https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10>

[11] Gupta, S. (2020, December 2). *Popular Classification Models for Machine Learning*. Analytics Vidhya. Retrieved February 13 2023, from <https://www.analyticsvidhya.com/blog/2020/11/popularclassification-models-for-machine-learning/>

[12] Kanade, V. (2022, April 18). *What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices*. Spiceworks. Retrieved February 13 2023, from <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

[13] Zhou, V. (2019, March 29). *A simple explanation of Gini impurity*. Victor Zhou. Retrieved February 13 2023, from <https://victorzhou.com/blog/gini-impurity/>.

[14] Kho, J. (2018, October 19). *Why Random Forest is My Favorite Machine Learning Model*. Medium. Retrieved February 13 2023, from <https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706>