DSC 680

Inman, Gracie

Project White Paper

04/07/24

Loan Default Prediction

Loan default is an issue that can be reduced through methods such as loan default prediction. Banks lend out money frequently with the presumption that the money will be returned with interest. When customers default on the loan the bank takes a loss. Being able to predict if a person will default on a loan, can allow banks to deny loans to customers not likely to pay them back. This will reduce risk and losses for banks, but it could also prevent negative impacts on consumers.

Defaulting on a loan can have negative consequences for the borrower such as “damage to your credit score, foreclosure or repossession, collection calls, and even a lawsuit” (Luthi, 2024). According to Luthi from Experian, default is when the borrower has missed one or more payments. (Luthi, 2024). When a borrower defaults on their loan it also negatively impacts the bank. Banks have methods of estimating if someone is likely to default such as “EAD is exposure at default and represents the value of a loan that a bank is at risk of losing at the time a borrower defaults on their loan. Loss given default is the value of a loan that a bank is at the risk of losing, after taking into proceeds from the sale of the asset, represented as a percentage of total exposure.” (Tuovila, 2023). Preventing extending loan offers to people who are likely to default will benefit both the loan companies and potential borrowers.

The loan dataset was obtained from Kaggle (Tse, 2020). The dataset consists of many attributes such as annual income, age, employment length, loan amount, credit length, whether

the person defaulted, and more. The data was imported to Jupyter Notebooks to perform analysis. The columns loan grade and loan interest rate were dropped because they would not be known at the time of application. Each column was visualized using histograms and no outliers were apparent. Each column was checked for unique and missing values. There were 895 rows that contained missing values in the employee length columns. Due to only 3% of rows containing missing values, the missing values were dropped. The columns used in the analysis are age, income, employment length, loan amount if the person defaulted, the percent of the loan compared to income, credit history length, home ownership type, previous default, and the loan intent. The categorical values were encoded for analysis and the data was split (70% training, 30% testing) and the random forest, logistic regression, and Naïve Bayes models were trained.

The metrics used to evaluate the models were accuracy, precision, and a confusion matrix. Table 1 shown below shows the accuracy and precision scores for each model. Random forest had the highest accuracy and precision of the models tested.

Table 1: Accuracy and Precision Scores for Models

Model	Accuracy	Precision
Logistic Regression	0.8116	0.7390
Random Forrest	0.8883	0.8943
Naïve Bayes	0.8004	0.6217

Figures 1-3 shown below illustrate the confusion matrix for each model. Looking at the confusion matrices, the random forest had the highest number of true positives and the lowest number of false negatives. Logistic regression had the highest true negatives and the least number of false positives, but the numbers for random forest are not significantly different. Given the accuracy, precision, and confusion matrix, the random forest model is the best model for the dataset.

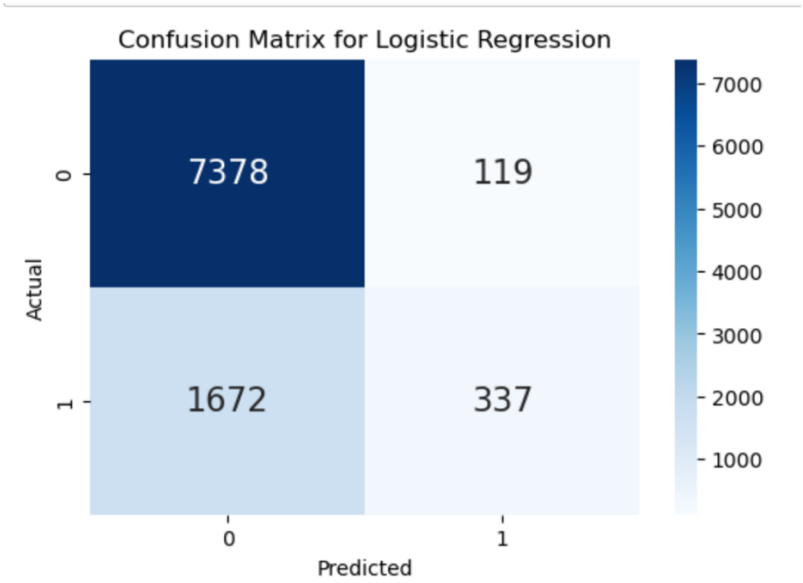


Figure 1: Logistic Regression Confusion Matrix

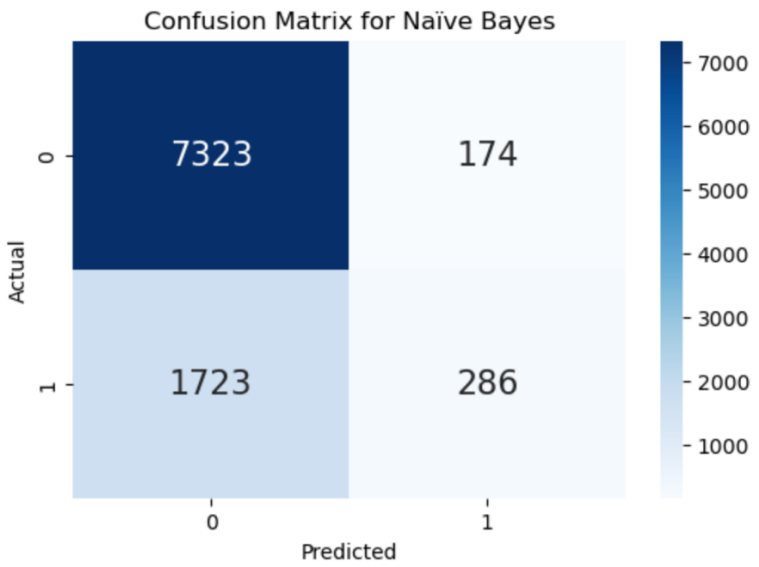


Figure 2: Naive Bayes Confusion Matrix

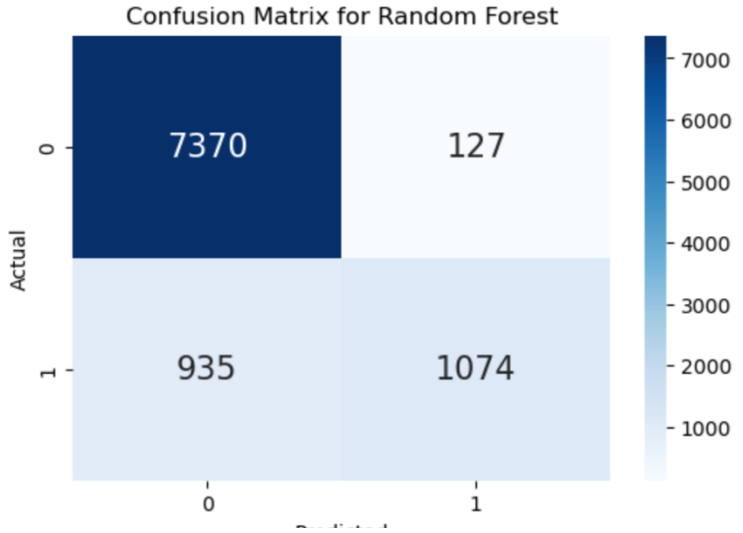


Figure 3: Random Forest confusion matrix.

Random forest was chosen as the best model of the ones tested to represent the data. Feature importance was performed to determine what the most important feature is in determining if someone will default. The results of feature importance are plotted below in

Figure 4. According to the figure, the three most important features are the percent of the loan requested amount compared to income, the person's income, and the loan amount.

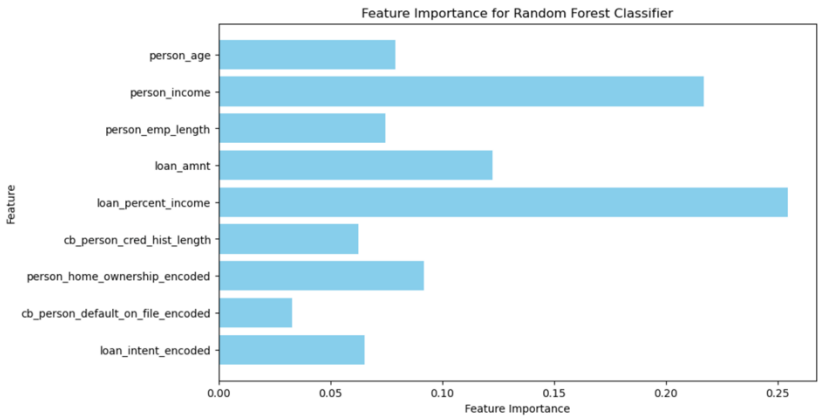


Figure 4: Bar chart of feature importance.

Loan default is a business problem around the world and preventing it could benefit both consumers and lenders. Based on the analysis, it was determined that the random forest model was the best model for the dataset. This was determined because the random forest model had the highest accuracy and precision as well as a significantly higher number of true positives. According to the feature importance, the most important feature is the percent of income that the requested loan amount is for.

Ethical concerns and assumptions are always important to take into consideration. One concern is I do not know how this data was obtained and do not have a clear way to show that the data was obtained ethically. Since I do not know where the data was obtained, I also do not know how accurate the data set is. A concern with the model is assuming someone will not pay back a loan based on someone with similar features even if they actually would. The concern is that this can be seen as discrimination based on the feature that flagged them. By using this

dataset, I assumed the data was accurate and ethically obtained. I also assume that this model will be used ethically. When dropping columns, I assumed the columns I dropped would not be obtained through the application process.

As with concerns with ethics and assumptions, there are also concerns with the limitations of the dataset. Since there is no known source for the data, it will not be possible to obtain new data to improve and continually test the model. It is also significantly likely that the model will not apply to other datasets and/or companies.

A challenge with this dataset will be improving the model. While the model is performing well, it has the potential to be improved. The recommended next step would be improving the model if possible. The boosting technique was not effective with this model and reduced the number of true positives. The dataset should be tested further with unseen data and once it is confirmed the model is performing well implement the model.

References:

Luthi, B. (2024, January 24). What happens if I default on a loan? Experian. https://www.experian.com/blogs/ask-experian/what-does-it-mean-to-default-on-a-loan/#:~:text=When%20you%20default%20on%20a%20loan%2C%20it%20could%20trigger%20a,where%20it%20may%20be%20unavoidable.

Tuovila, A. (2023, June 28). Loss given default (LGD): Two ways to calculate, plus an example. Investopedia. https://www.investopedia.com/terms/l/lossgivendefault.asp#:~:text=Loss%20given%20default%20(LGD)%20is,borrower%20defaults%20on%20a%20loan.

Tse, L. (2020, June 2). Credit risk dataset. Kaggle.

https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Appendix

A. Ten questions:

1. Why choose a dataset with no known source?
    i. Although the data had no known source, it had the features desired to test this concept. The model should be tested with data from a known source and modified as necessary.
2. What is the significance of true positives/negatives and false positives and negatives?
    i. The significance of these values is they show in greater detail how the model is classifying each instance versus how it should have been classified. The significance can vary based on the project goal. However, in this case, the goal is to get the greatest number of true positives and negatives. This will show the model is working as intended.
3. If logistic regression had more true negatives and less false positives, why was it not chosen as the best model?
    i. This is because it had a significantly smaller number of true positives than the naïve Bayes model which has similar true negative and false positive numbers. It is important to consider all of the numbers in the confusion matrix.
4. Why is feature importance so important?

    i. Feature importance is so important because it shows what feature is most important in deciding whether or not the customer will churn or not. This can aid in reducing churn by working on the specified features.

5. Why was analysis performed in Jupyter Notebooks?
    i. It is a personal preference. I like the output and user-friendliness of Jupyter Notebooks. The same analysis could have been performed by another IDE such as PyCharm.
6. Why did you visualize the unique values for each column?
    i. I like to do this because it is a quick way for me to see what different values are located within the column. I can check for typos and have a plan for mapping values.
7. Why create a model for something that already has a structure in place to prevent this?
    i. There are always ways to improve current structures, you'll never know if you don't try.
8. What is something you could do differently?
    i. If I could do something differently, I would use data from a known source. This will cut down on ethical concerns and allow for continued testing of the model. I also could have set aside some of the data for further testing.
9. Were your results what you expected?
    i. I would say so, although I did not expect that boosting the model would affect it negatively.
10. Why did you choose the models you did?

i. I chose the models I did because they are all known to work well will classification tasks.

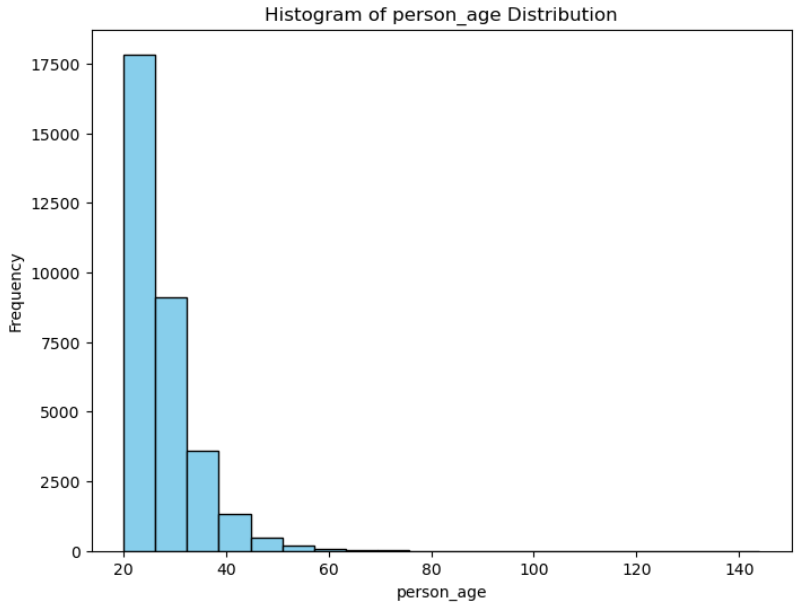
B. Debt recovery

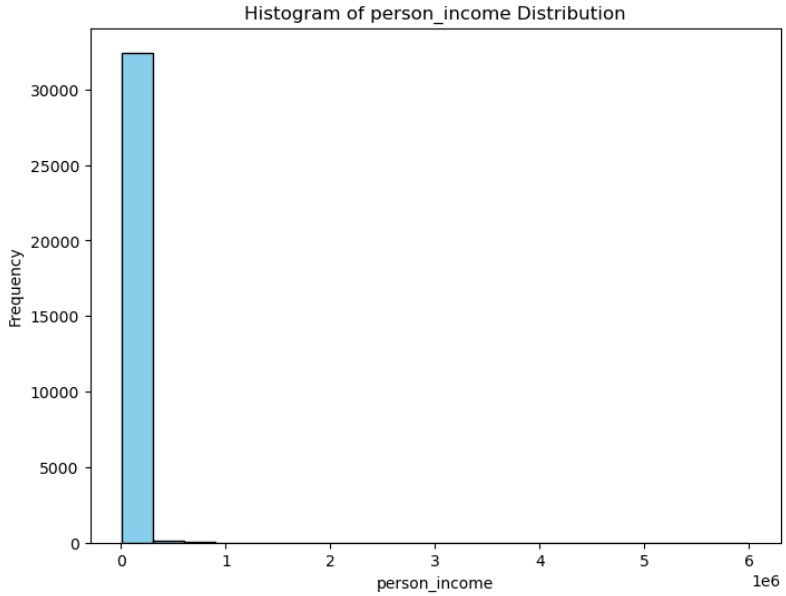
1. It is important to note that the model will not correctly identify every person who will default on their loan. Loan default is inevitable in some cases but is possible to recover.

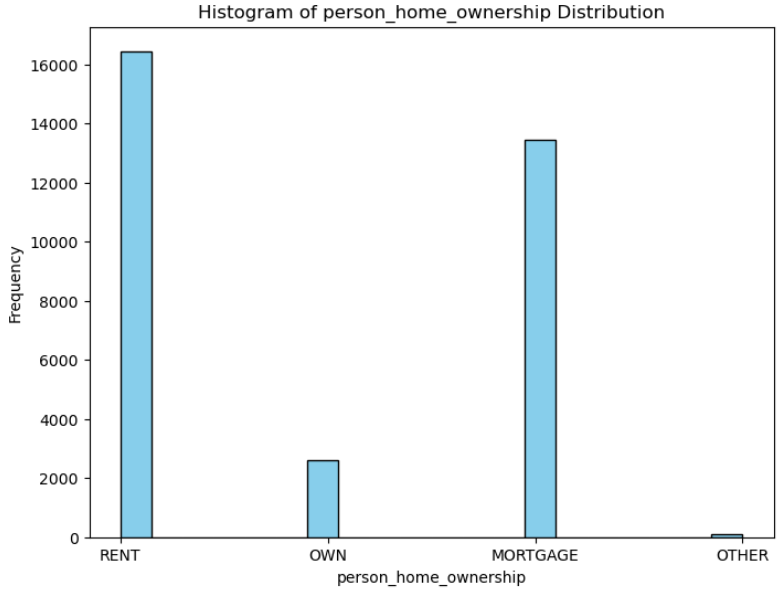
    i. There are many resources to help consumers after default. Debt.org contains resources for debt recovery after defaulting on a loan.

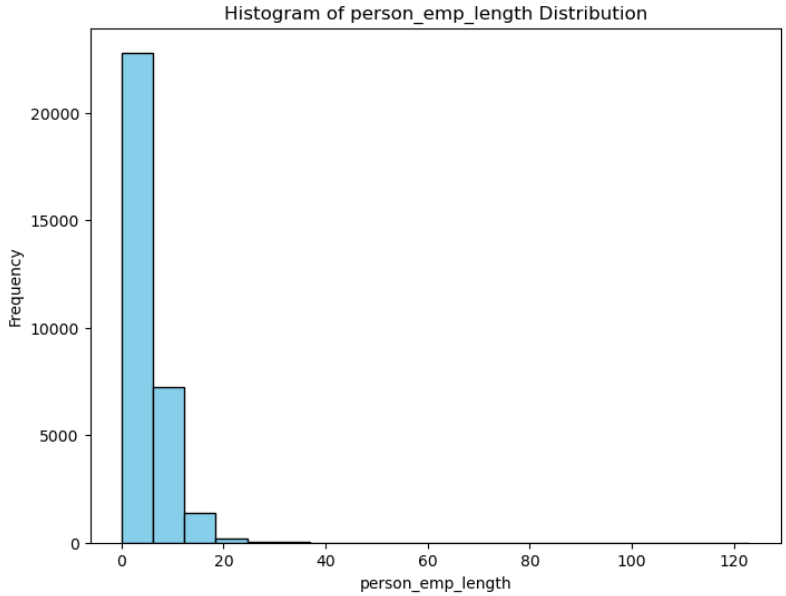
        1. Fay, B. (2023, August 24). *Debt Recovery and the Debt Recovery Process*. Debt.org. https://www.debt.org/advice/recovery/

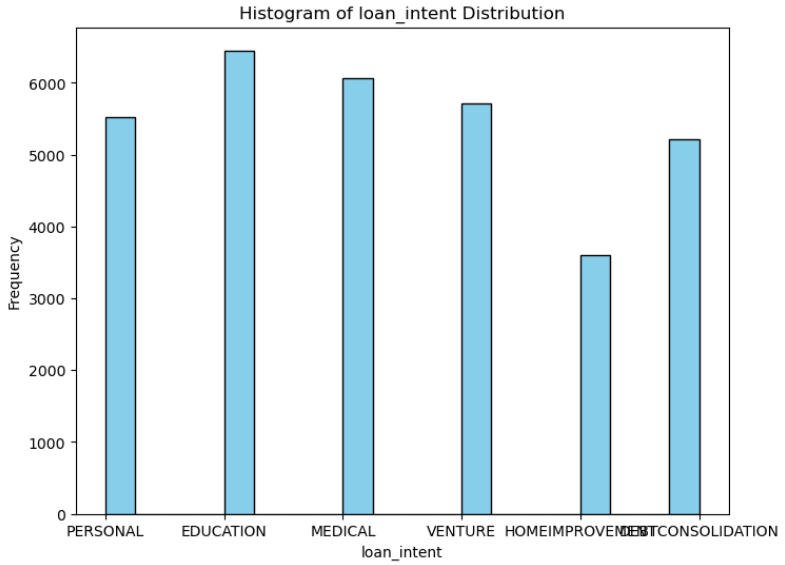
C. Initial Visualizations

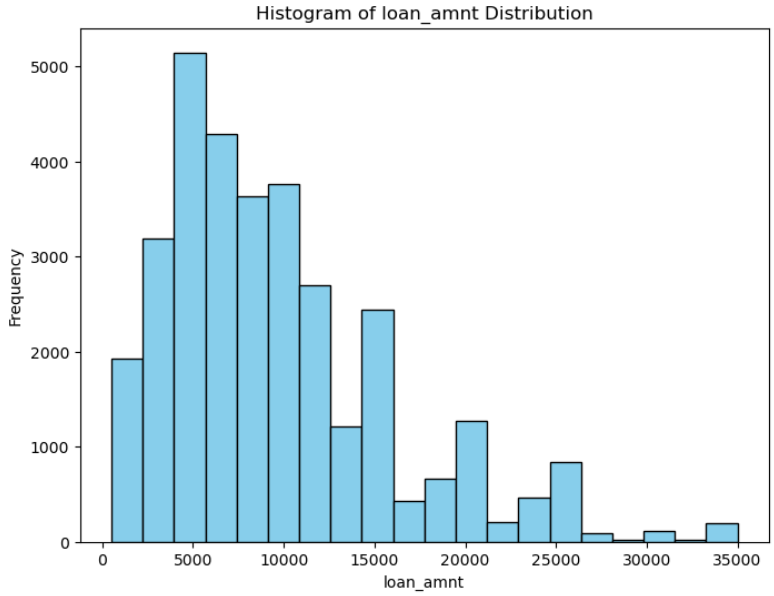


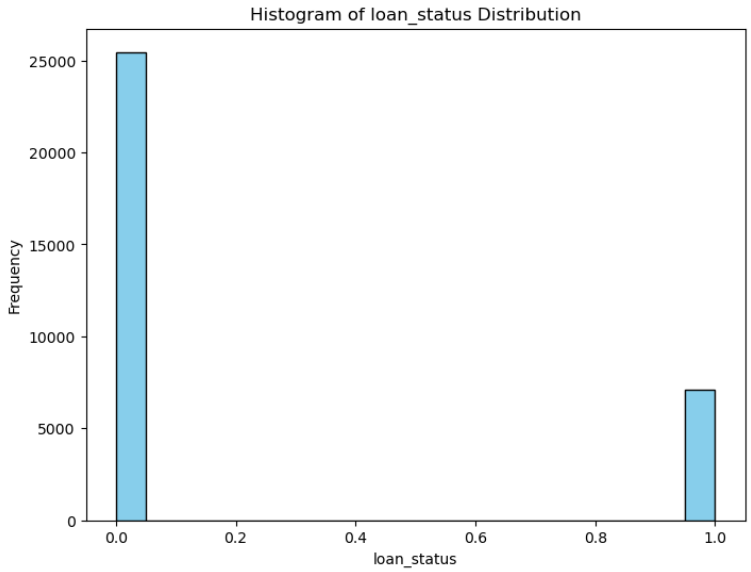

Histogram of person_income Distribution
Frequency
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
6
1e6
person_income

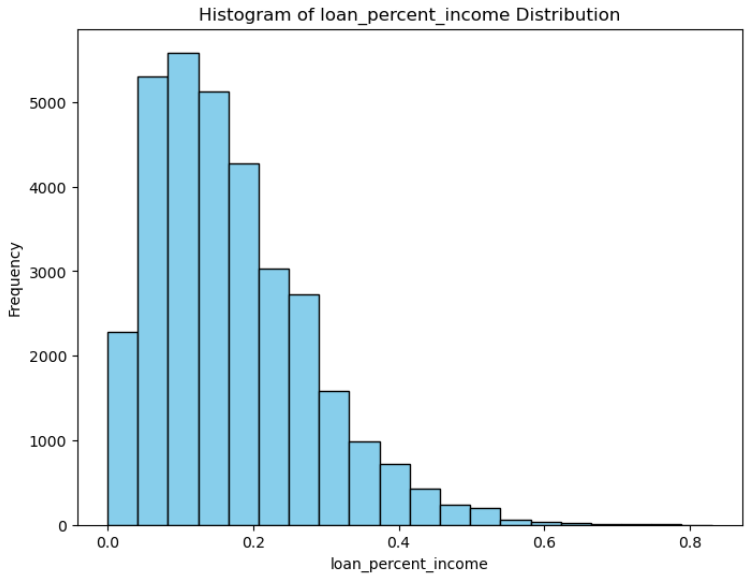

Histogram of person_home_ownership Distribution
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
RENT
OWN
MORTGAGE
OTHER
person_home_ownership

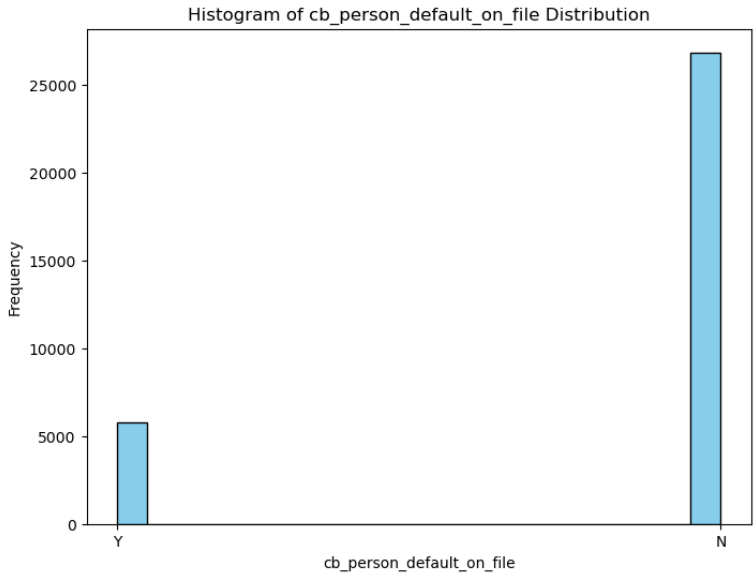

Histogram of person_emp_length Distribution
Frequency
person_emp_length
0
5000
10000
15000
20000
0
20
40
60
80
100
120

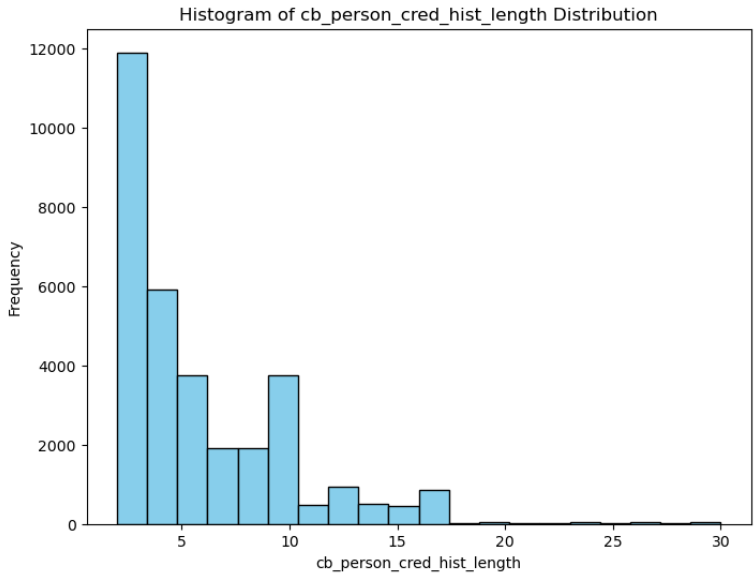

Histogram of loan_intent Distribution
Frequency
loan_intent
0
1000
2000
3000
4000
5000
6000
PERSONAL
EDUCATION
MEDICAL
VENTURE
HOMEIMPROVEMENT
DEBTCONSOLIDATION


Histogram of loan_amnt Distribution
Frequency
loan_amnt
0
1000
2000
3000
4000
5000
0
5000
10000
15000
20000
25000
30000
35000


Histogram of loan_status Distribution
Frequency
loan_status
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of loan_percent_income Distribution
Frequency
loan_percent_income
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8


Histogram of cb_person_default_on_file Distribution
Frequency
cb_person_default_on_file
0
5000
10000
15000
20000
25000
Y
N


Histogram of cb_person_cred_hist_length Distribution
Frequency
cb_person_cred_hist_length
0
2000
4000
6000
8000
10000
12000
5
10
15
20
25
30