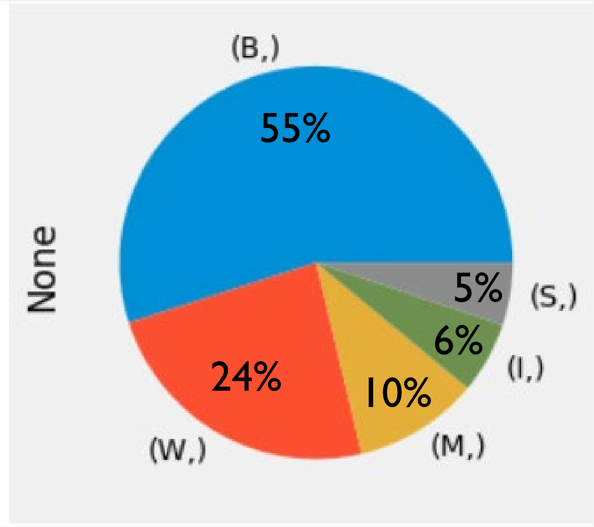
PROPOSAL

To	MoneyLion
From	Morris Lee
Date	11/11/2021
Subject	To Propose a Solution of Predicting and Mitigating Loan Risk

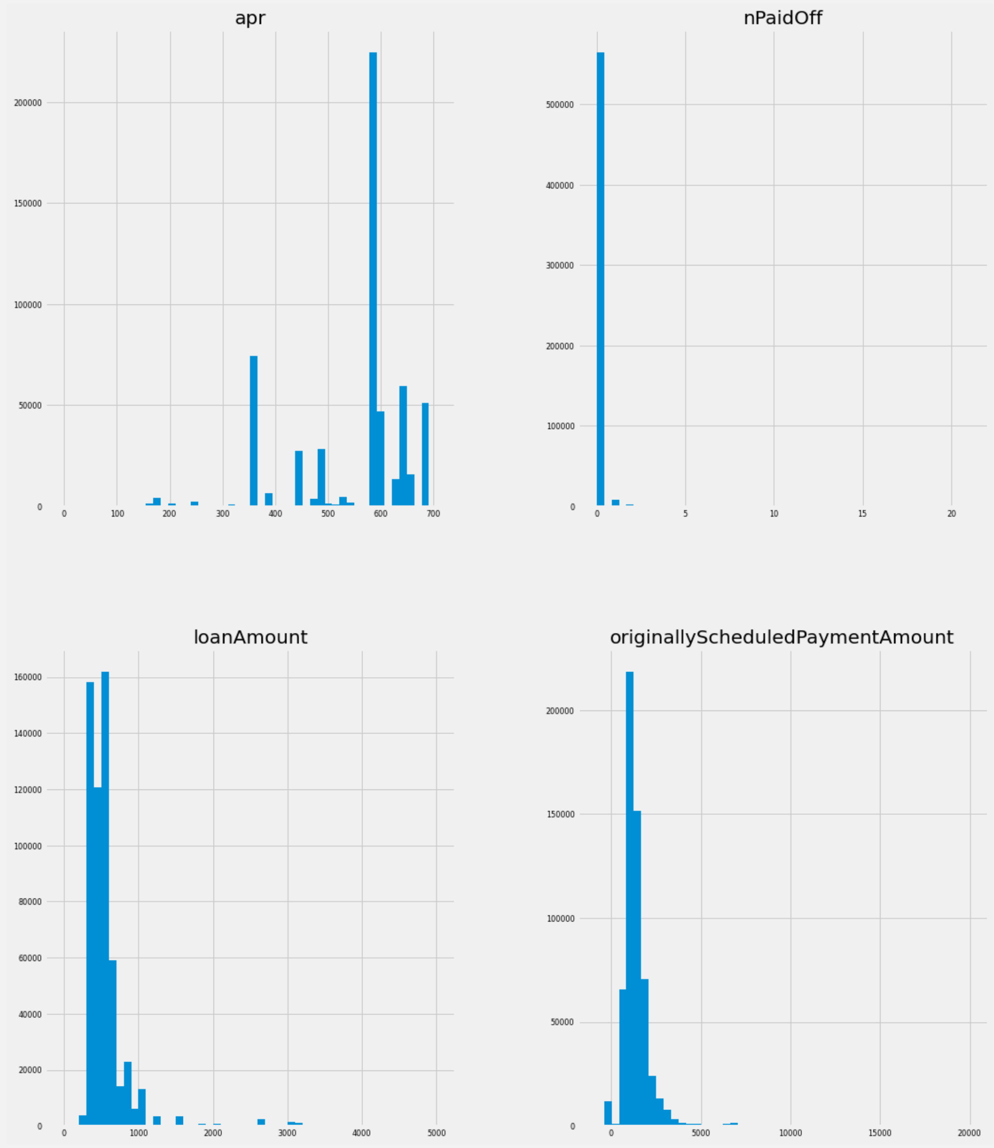
EXPLORATORY DATA ANALYSIS

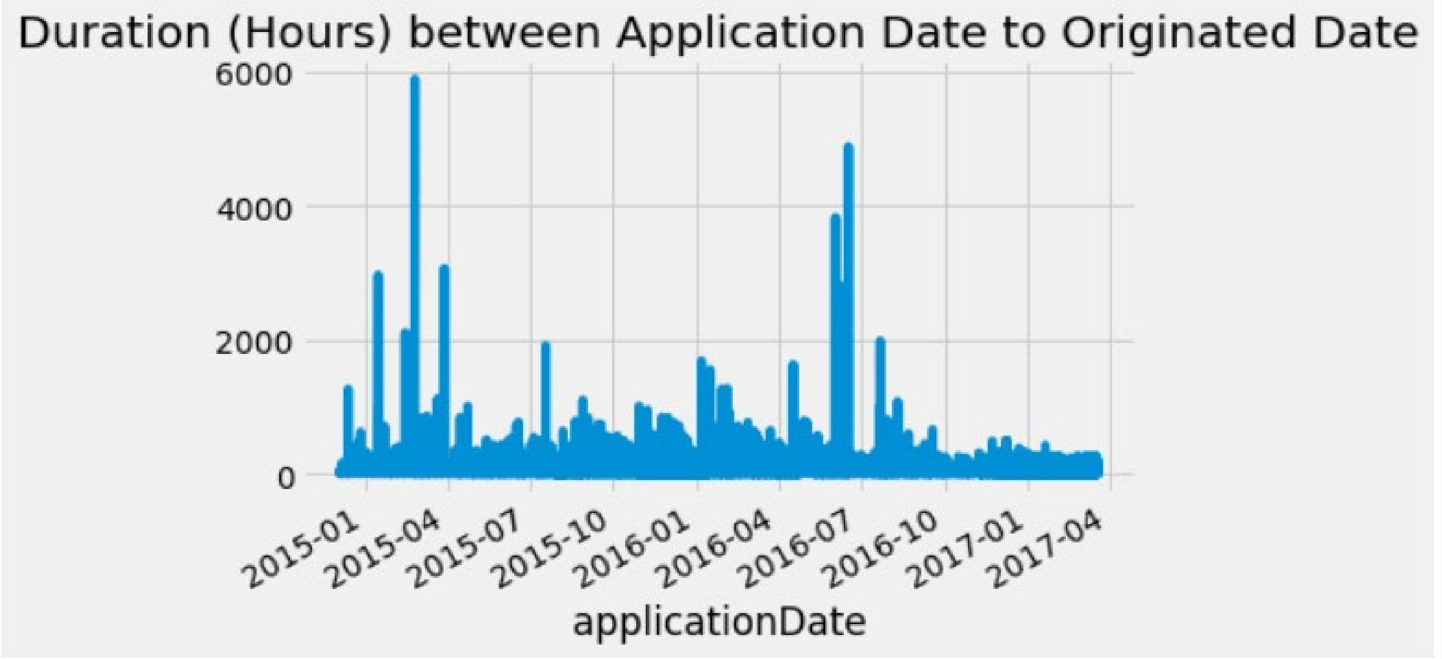
Repayment Loan Frequency



Biweekly
Irregular
Monthly
Semi monthly
Weekly

Distribution of Loan Dataset





INSIGHTS SUMMARY

1.Is there any loan approved but not funded? If yes, how many of them?

Yes, There are 1054 of loans approved but not funded, amounting to 0.0263% of overall approved loans

2.What is the most common of loan status? Ratio of it?

There are 450984 of withdrawn loans, which means that applicant notified the lender they no longer want their loan application processed, amounting to 78% of overall loans

3.How many of the loan status are paid off loan? Ratio of it?

There are 11427 of paid off loans, which means that applicant, amounting to 1.98% of overall loans

4.How is the ratio of Custom Collection?

There are 13895 of payment required custom collection, amounting to 0.1529% of overall payment

5.How is the percentage of payment checked?

There are 209621 of payment checked (successful), amounting to 30.4079% of overall payment

6.How many of it has listed paymentReturnCode?

There are 31533 of payment has return code (failed), amounting to 4.7935% of overall payment

7.If loan is not funded, is there payment checked?

Yes, There is 349 payment checked, although loan is not funded

THE PERCENTAGE OF "PAYMENT CHECKED OF FUNDED LOAN" ACCORDING TO DIFFERENT LOAN STATUS

Percentage of Payment Checked of Funded Loan

loan_status	loan_isFunded	isFunded_pay_checked_percentage
Settlement Pending Paid Off	1.0	84.615385
Paid Off Loan	11427.0	59.005073
Settlement Paid Off	708.0	47.235398
Pending Paid Off	169.0	44.045858
New Loan	8112.0	24.849738
Settled Bankruptcy	325.0	24.083770
Returned Item	1182.0	20.643767
External Collection	11334.0	17.391805
Charged Off Paid Off	159.0	16.112431
Internal Collection	5564.0	15.663995
Charged Off	1.0	0.000000

Above Average

Average 30%

Below Average

PROBLEM STATEMENT

As a result, we can know that a loan is **risky** when the loan payment checked percentage **below 30%.**

This is because the **average** payment checked percentage of overall payment is **30.4079%**.

Therefore, the loan status such as **"Internal Collection"**, **"Charged Off Paid Off'**, **"External Collection"**, "Returned Item" are considered **risky**.

EXPLANATION

Regarded as **Non Risky Loan** Status, as the percentage of payment checked are **above average**

Settlement Pending Paid Off	84.615385
Paid Off Loan	59.005073
Settlement Paid Off	47.235398
Pending Paid Off	44.045858

Regarded as **Risky Loan** Status, as the percentage of payment checked are **below average**

New Loan	24.849738
Settled Bankruptcy	24.083770
Returned Item	20.643767
External Collection	17.391805
Charged Off Paid Off	16.112431
Internal Collection	15.663995
Charged Off	0.000000

It indicate that that the payment to be received by company is getting difficult and require more time

PROPOSED SOLUTION

It is clear that we need to reduce the likelihood of unfavorable loan status before it has happened. **But how can we do that?**

The method proposed here is to use supervised learning **classification method** to predict the **loan status** given the loan data.

So that company can take initiative to **reject high risk loans before the loans are funded**.

PRE-PROCESSING STEPS THAT HAVE IMPLEMENTED

- Drop unwanted attributes
- Filter the loan status we would like to predict
- Label encode categorical data
- Change data type to the correct one
- Imputation to handle missing values
- Standardisation – AB testing a new dataset, df1
- Normalisation – AB testing another new dataset, df2
- Train Test Split

MODELS TRAINING

- Gaussian_Naive_Bayes and Bernoulli_Naive_Bayes are trained

In [61]:
```
naive_bayes('accuracy')
```

```
Naive Bayes Classification is executed
Found the best parameters and best score with GridSearchCV
Gaussian_Naive_Bayes df0 accuracy : 0.6709920413978319
Gaussian_Naive_Bayes df1 accuracy : 0.3057435124837607 5
Gaussian_Naive_Bayes df2 accuracy : 0.2699854801908318
Bernoulli_Naive_Bayes df0 accuracy : 0.8144739571392702
Bernoulli_Naive_Bayes df1 accuracy : 0.813181366088057
Bernoulli_Naive_Bayes df2 accuracy : 0.8144739571392702
Executed time is 7.899519443511963 seconds
Values are stored into a dataframe
```

Out[61]:
```
{'datasets': ['df0', 'df1', 'df2'],
 'Gaussian_Naive_Bayes': [0.6709920413978319,
  0.30574351248376075,
  0.2699854801908318],
 'Bernoulli_Naive_Bayes': [0.8144739571392702,
  0.813181366088057,
  0.8144739571392702]}
```

MODELS TESTING

```
In [64]:  1  predict(models)

df0 Test_Predict_BernoulliNB (accuracy): 0.8150376859186543
df1 Test_Predict_BernoulliNB (accuracy): 0.813395750181224З
df2 Test_Predict_BernoulliNB (accuracy): 0.8150376859186543

Out[64]: {'datasets': ['df0', 'df1', 'df2'],
          'Gaussian_Naive_Bayes': [0.6709920413978319,
           0.30574351248376075,
           0.269985480190831８],
          'Bernoulli_Naive_Bayes': [0.8144739571392702,
           0.813181366608805７,
           0.8144739571392702],
          'Test_Predict_BernoulliNB': [0.8150376859186543,
           0.813395750181224３,
           0.8150376859186543]}
```

RESULTS OF THE MODEL ACCURACY

	datasets	Gaussian_Naive_Bayes	Bernoulli_Naive_Bayes	Test_Predict_BernoulliNB
0	df0	0.670992	0.814474	0.815038
1	df1	0.305744	0.813181	0.813396
2	df2	0.269985	0.814474	0.815038

Given the loan dataset df0, we can use **BernoulliNB** to make **classification** prediction of different loan status with a Test Predict accuracy of **0.815**.

CONCLUSION

Therefore, by using the machine learning model, company is able to decide in advance which loan to approve or reject by knowing the predicted loan status in order to mitigate the loan risk.

Future work: One can also try out non-parametric models such as Decision Tree, Random Forest, SVM and Ensemble Learning to identify their performance on classification. From the exhaustic search, we can better identify which model works the best.