Credit Risk Analysis

Chahhat Jagdish Pandit, Vijay Kinnera Thota, Aatusa Mehdiyan
Computer Science Department
San Diego State University

Introduce Problem

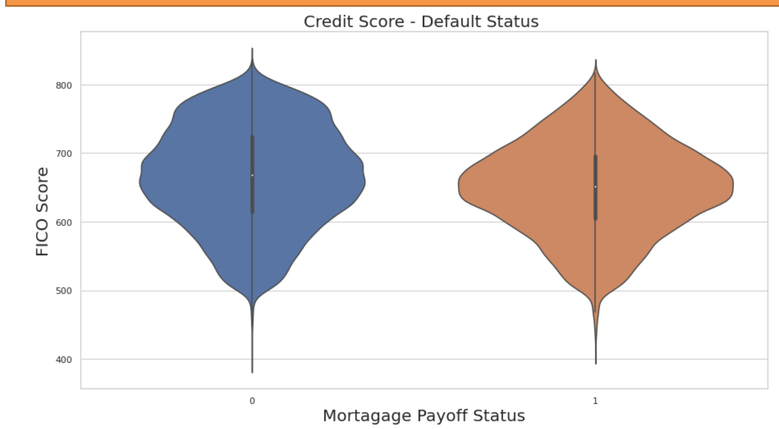
- The objective of the credit risk analysis is to assess the borrowers' creditworthiness by quantifying the risk of loss to which the lender is exposed. The probability of default, loss due to default, and exposure to default are the three measures that lenders use to measure credit risk.
- This is the data set of mortgage observations for 50,000 residential U.S. mortgage borrowers over 60 periods, 600,000 records in total. The data set is a randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios and provided by International Financial Research (www.internationalfinancialresearch.org).
- We have applied SVM, Random Forest, KNN, Logistic Regression Classification models to determine whether the borrower is able pay their loan(0) or not(1) at the end of the maturity period.

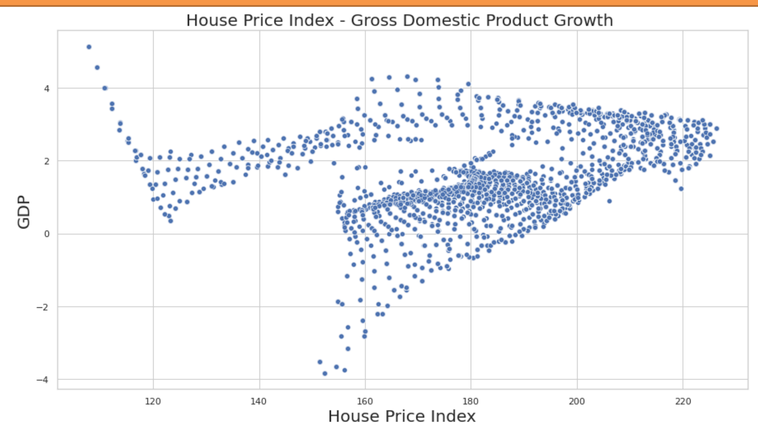
Data Exploration

- In this dataset there are two variables - time: Time stamp of observation and orig_time: Time stamp for origination
- Based on these columns we have derived a column maturity_time_period_days which is difference between these two variables. By adding this column, we can know in how many days the loan has been paid or how many days are there for the loan to reach the maturity.

Further, data exploration is done by visualizing the data using violin plot and scatter plot as shown under the visualization header.

Data Visualization





Model Results

We have built four different models to predict our output variable, that are:

1. SVM(Support Vector Machines)-
The train accuracy is 0.81 and test accuracy is 0.78.
The first graph on the left is the precision recall graph:-

2. Random Forest Classifier-
The train accuracy is 0.82 and test accuracy is 0.79.
The second graph is the precision recall graph:-

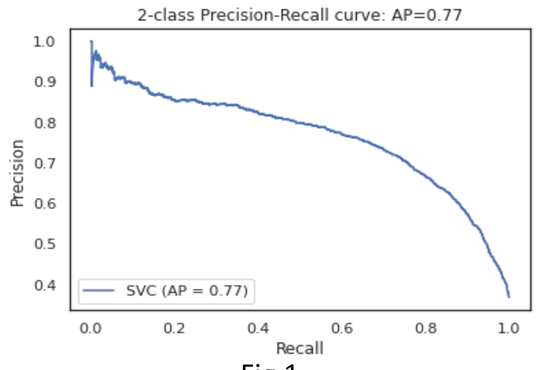


Fig.1

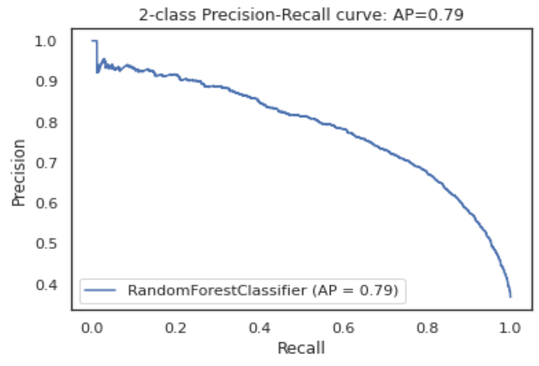


Fig.2

3. KNN-
The train accuracy is 0.79 and test accuracy is 0.78.
The third graph is the precision recall graph:-

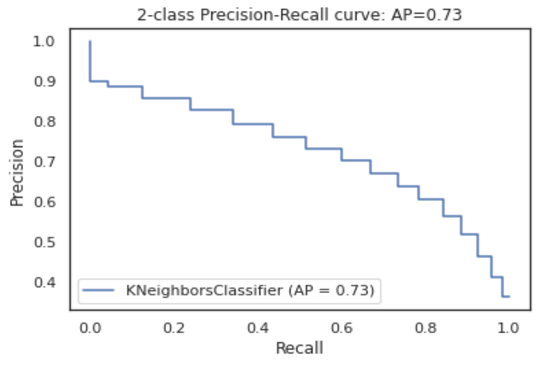


Fig.3

4. Logistic Regression-
The train accuracy is 0.73 and test accuracy is 0.76.

Data Processing

- The variable orig_time has values of when the observation of the loan started. So, we added a column name start_day_observation. This new column contains the subtraction of orig_time and first_time. This new column has the number of day on which the loan observation started during the Maturity Time Period (maturity_time_period_days).
- The dataset has 600,000 records/observation of 50,000 customers. So, in order to obtain a single record for an individual customer we are performing Group By based on id of Customer. So, for this start_day_observation variable we have taken the first value in each group of the customer. The variable start_day_observation contains that day at which the observation of Mortgage started from the total maturity_time_period_days.
- For the variable balance_time we have different amount varying from 0 to 8,70,1859 dollars. So, in order to efficiently analyze our data, we are calculating the percentage difference and standardize the data in a scale. The percentage change is between the beginning balance and ending balance.
- We will use SMOTE to upsample our dependent variables. SMOTE is a method of oversampling which produces synthetic samples from the minority class. It is used to achieve a synthetically class-balanced or almost class-balanced training set, which is then used to train the classifier.

Conclusion

After analyzing all the different learning rates and methods we used, the Accuracy that we got was 79% using Random Forest Classifier. The average precision score we received is 0.79.

References

[1] Trilok N Pandey, M. Suman Kumar, J. Alok Kumar, Satchidananda Dehuri, 2017, Credit Risk Analysis using Machine Learning Classifiers, https://www.researchgate.net/publication/325983636_Credit_risk_analysis_using_machine_learning_classifiers
[2] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, 2007, Credit scoring with a data mining approach based on support vector machines, http://nlg.csie.ntu.edu.tw/~cjwang/paper/Credit%20Card%20Scoring%20with%20a%20Data%20Mining%20Approach%20Based%20on%20Support%20Vector%20Machine.pdf
[3] S.J. Shiv, Srinivasa Murthy, Krishnaprasad Challuru, 2018, Credit Risk Analysis Using Machine Learning Techniques, https://www.researchgate.net/publication/341532931_Credit_Risk_Analysis_Using_Machine_Learning_Techniques