Franck Arnaud Olilo

Credit Risk Modeling: Classification vs Regression Models

Introduction

In recent years, finance officers and bankers have faced numerous challenges all over the world on risks they are exposed to, such us compliance, operational and credit risk etc….. The ability to lend money to borrower/ debtor with or without information. Therefore, skills and tools to predict and measure a probability of default or expected loss are being developed by financial institutions. Nowadays financial institutions rely on quantitative analysis and statistical models.
In this paper, we develop a logistic regression in the purpose of predicting and classifying good and bad credit applicants and compare the logistic regression model prediction score with the Naïve Bayes, SVM , Random forest and XGBOOST one.

Data Overview and Descriptive Analysis

For this case study, we will be using the German credit data set, the data was obtained on Kaggle https://www.kaggle.com/datasets/uciml/german-credit .
The data has 1000 observation and 21 variables, each observation represented a person who took credit with a German bank attribute and its creditability score.
Figure 1 and figure 2 show a quick overview on how much the data is balanced (more descriptive figure are included in the python code)

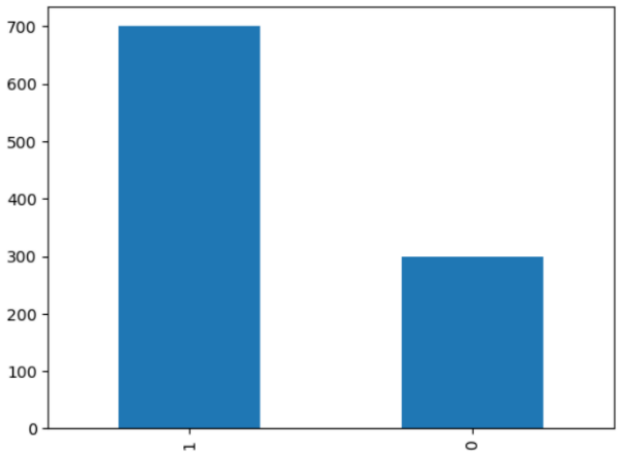

Figure 1

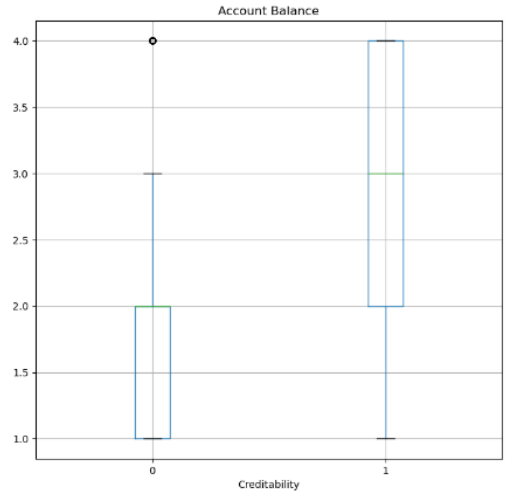

Figure 2

Model Development

A logistic regression model attempts to estimate the occurrence of an event (failure) for randomly selected observations versus the probability of the event not occurring. Suppose you have data for 1000 loans, all predictors, and whether the borrower has defaulted. The probability of failure is here called the response or dependent variable. The default itself is a binary variable. That is, its value is 0 or 1 (0 is not the default, 1 is the default)

Logistic regression can be said to be a classification algorithm used to predict a binary outcome (1/0, default/no default) given a set of independent variables. This is a special case of linear regression when the outcome variable is categorical. Predict the probability that a failure will occur by fitting the logit function to the data.

The logistic model can be expressed as:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_3$$

And the probability that a default happens

$$P(Creditability = 1 \mid x_1, \ldots, x_m) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)}}$$

In this paper used the very useful python package called sklearn package to perform the logistic regression , However before splitting our data on training and testing data, we observed that our data is unbalanced as shown in In figure 1 , which illustrate that 70% of the dataset are “default” and 30% are “no default”, the unbalanced data may lead to wrong prediction s, therefore we decided to resample the data set using the imblearn package and smote function. The new balanced data have equal sample size on default and no default as shown in figure 3. We now splited our data set with 75 % training and 25% testing.

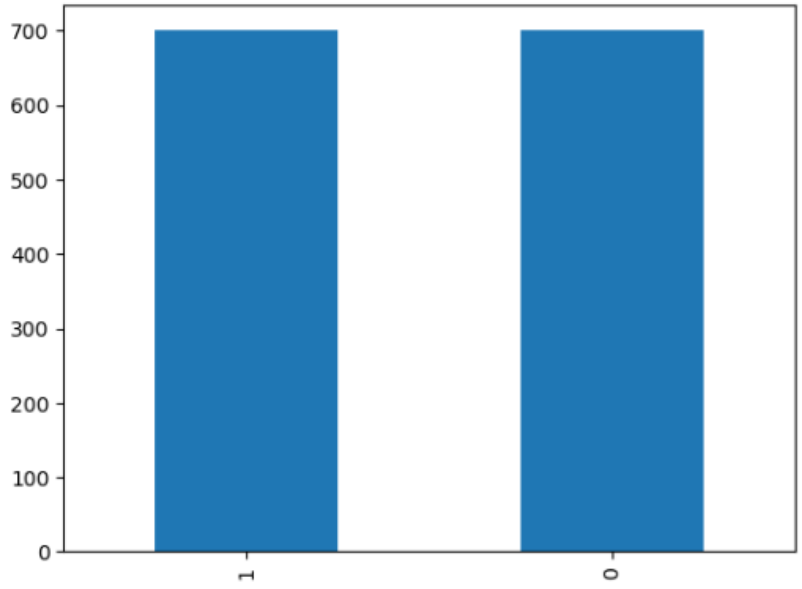

Figure 3

Model Validation

The last part of our paper was to evaluate our logistic regression model and compare it with the other models listed in the introduction.
As quick overview, 1400 data was used I the model development, 1050 were used as training data and 350 as testing data.
To evaluate our model we will use the confusion matrix table (figure 4) and the ROC curve and AUC (figure5).

The confusion matrix measures the overall predicted accuracy and precision of the model

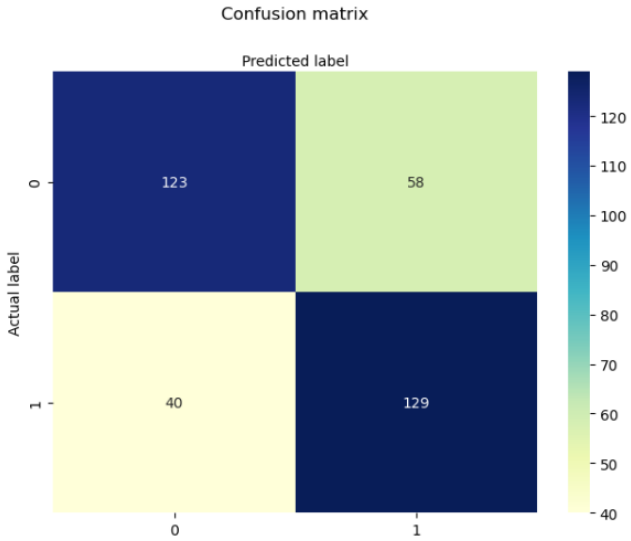

Figure 4

TP= True Positive
FP= False Positive
TN= Ture Negative
FN=False Negative

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$= \frac{123 + 129}{123 + 58 + 40 + 129}$$

$$= 0.72$$

The ROC curve plot the rate of false positive on the x-axis and tue positive rate on y-axis and measure performance. The AUC are located under the curve , a better model tend to have the greaater AUC

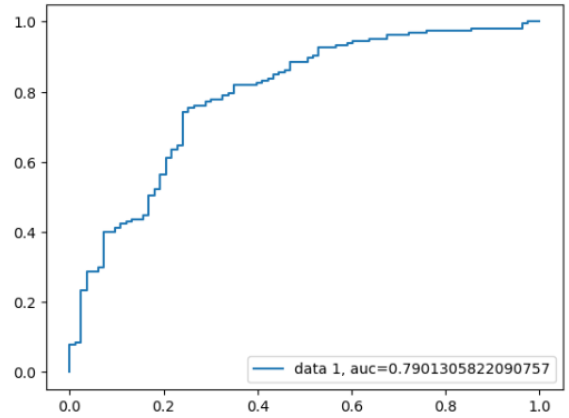

Figure 5

Model Comparsion and Discussion

The last part of our paper is to compare the models , even if those model were not discussed in the paper , the Python code can be viewed in the paper code.
Table 1 illustrate the accuracy of our 6 model , with the best model being the random forest follow by XGBOOST and Logistic Regression and KNN having the lowest

Model	Classification Accuracy
Logistic regression	**0.78**
Naïve Bayes	**0.74**
Support Vector Machine	**0.76**
KNN	**0.6**
Random Forest	**0.8**
XGBOOST	**0.79**

Table 1

In future work of this problem we will try to use an unbalanced data and also trying to select variables by forward or stepwise selection instead of using all given variables

Conclusion

Being able to predict the default or npt of a consumer reduced the risk taken by insruance and/or financial company, while each company keep its model private our choice in this paper is to write a model on different model ,the paper demonstrate that modeling credit risk is not an easy task, as we can have hundred of millions of data set and hundred of differents variables.one may also decided to used less or more variables , unbalanced or balance data set. We also observed that machine learning algorithm (random forest and XGBOOST) have a better classification accuracy than logistic regression

Reference

https://su-plus.strathmore.edu/bitstream/handle/11071/6789/Consumer%20credit%20risk%20modelling%20using%20machine%20learning%20algorithms.pdf?sequence=3&isAllowed=y

https://theses.gla.ac.uk/6592/1/2015shirzadiphd.pdf

https://dspace.spbu.ru/bitstream/11701/5370/1/Thesis_Amir_Azatarrahian_May_2016.pdf

https://www.diva-portal.org/smash/get/diva2:1116036/FULLTEXT01.pdf

https://deepblue.lib.umich.edu/bitstream/handle/2027.42/63707/ajzhang_1.pdf

https://www.ibm.com/docs/en/spss-statistics/saas?topic=regression-using-binary-logistic-assess-credit-risk

https://utstat.toronto.edu/~ali/papers/creditworthinessProject.pdf