EXPLORATION AND CLASSIFICATION OF CREDIT RISK PROFILE CLASSIFIER

Ankit Swarnkar
Department of Data Science
Indiana University
Bloomington, IN
ankswarn@iu.edu

Karun Dhingra
Department of Data Science
Indiana University
Bloomington, IN
dhingrak@iu.edu

Abstract

Predictive analytics and classification are now an integral part of credit management. There is a heavy research going on in this domain to be able to accurately predict and classify the credit profile of the customer which can be beneficial to the credit business. Working in this context, we try to explore some of insights from the credit data and build model which would be able to classify customer as future defaulter or non-defaulter. We extracted useful pattern using exploratory data analysis. The data mining techniques used in this project are Logistic Regression, SVM and random forest.

Keywords: credit defaulter algorithms, regression, defaulter classification, credit profile analyzer

1. Introduction

Credit card delinquency has become increasingly commonplace over the past few years. With increase in amount of data and advance machine, it is now possible to get nearly accurate prediction. Because there is lots of information of customer, it is now possible to implement effective model which can blend all the information regarding the customer to help Business with statistical analysis. Extracting behavioral insight from the customer transactions, customer business and personal profile can help us classify them as future non-defaulter and defaulter. Data Mining along with Business Intelligence solution has revolutionized financial domain. The main idea of this project is to present a model which can use to determine probability of default for credit card and determine the best statistical methods of prediction.

The main idea of this project is to combine recent machine learning and data mining techniques through which we can classify the customer as credible or not credible clients. We used different classification techniques for our study. We have performed a classification analysis of the different classification algorithms and selected the best model for our final classifier. The method we used are 1. Logistic Regression 2. Support Vector Machine and 3. Random Forest

The paper is structured as follow: In Section 2 we provide the data models and some of the interesting insight from the data. In section 3, we applied supervised machine learning approach and provide statistical analysis on prediction. Section 4 shows the conclusion.

2. Data Analysis Model and Insight

Classification refer to prediction of class group and grouping them based on outcome predicator. In our project, we are grouping the customer whether the customer is credible or non-credible. Using the covariates of previous month credit history, demographic information and personal data we are trying to build a function which will take all these information as input and classify the customer. These functions are then deployed to application which can help the business to know their customer in a better way.

The data we can used for the classification can either be labelled or non-labelled dataset. The dataset we used are labelled imbalance data set which contain anonymized details of the customers. We used most popular classification techniques for our study: Logistic Regression, Support Vector machine and Random Forest.

2.1 Data

The dataset used for this project is based on financial dataset of default payments in Taiwan from 2009. We acquired the data from UCI[3] repository to classify Taiwan credit card holders. The dataset is already normalized as a binary variable for outcome variable. Integer number coding is done in categorical variables which are as follow:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

The dataset used for our project is imbalanced dataset as shown in figure 1. This is generally the case while dealing with defaulter dataset. The data set for non-defaulter in our case is 79.1% and defaulter tends to 19.9% of the total dataset. To handle these imbalanced dataset frameworks known as ‘Sampling Methods’ are generally used. These methods aim to modify an imbalanced data into balanced distribution using some mechanism of over, under or creating synthetic dataset. We have used the over sampling methodology so we don’t miss out important samples and can mitigate the bias caused by the imbalance.

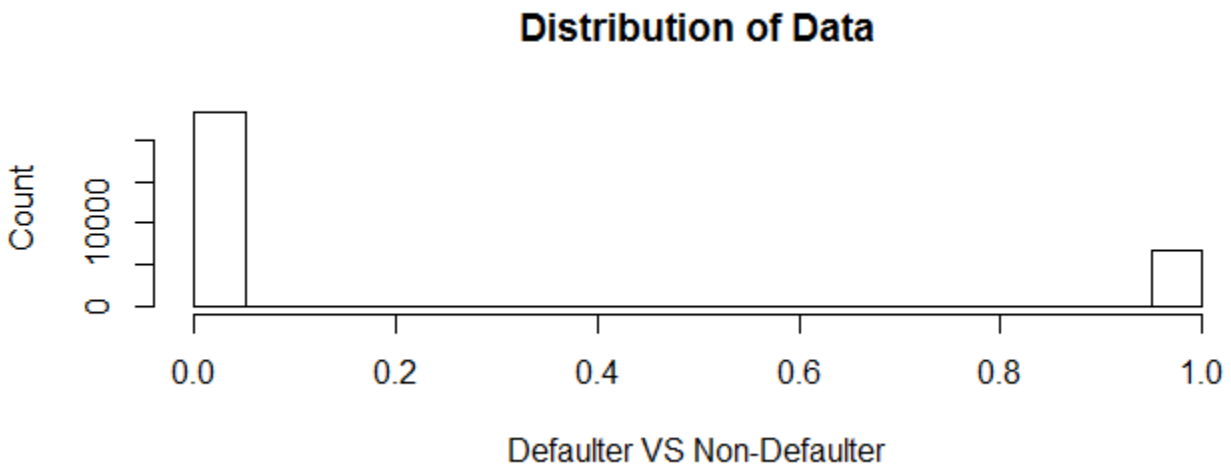


Figure 1: Distribution of data as defaulter and non-defaulter.

2.2 Data Insights

We tried to explore some of the hypothesis to learn more about our data and extract any hidden information using data mining visualization techniques. We explored hidden pattern to answer following three hypotheses:

Hypothesis 1: Whether there is any pattern between previous transaction month on the outcome variable.

Hypothesis 2: Whether married man tend to be more defaulter.

Hypothesis 3: Whether there is any strong correlation between independent variables used in the analysis.

To answer first hypothesis, I used time series plotting framework and tried to plot the frequency of previous default months for both defaulter and non-defaulters. The below figure show that there is some trend which can be explored. The red line indicates trend of the defaulters and blue line indicate the trend of non-defaulters. We can see that not all the months plays a significant role here. The defaulting cases for August (shown as 2) are high for both defaulters and non-defaulter however the October month tend to have lower defaulting case for non-defaulters and high defaulting cases for defaulters. It can be a sampling variation but we can test this hypothesis using a richer data set.

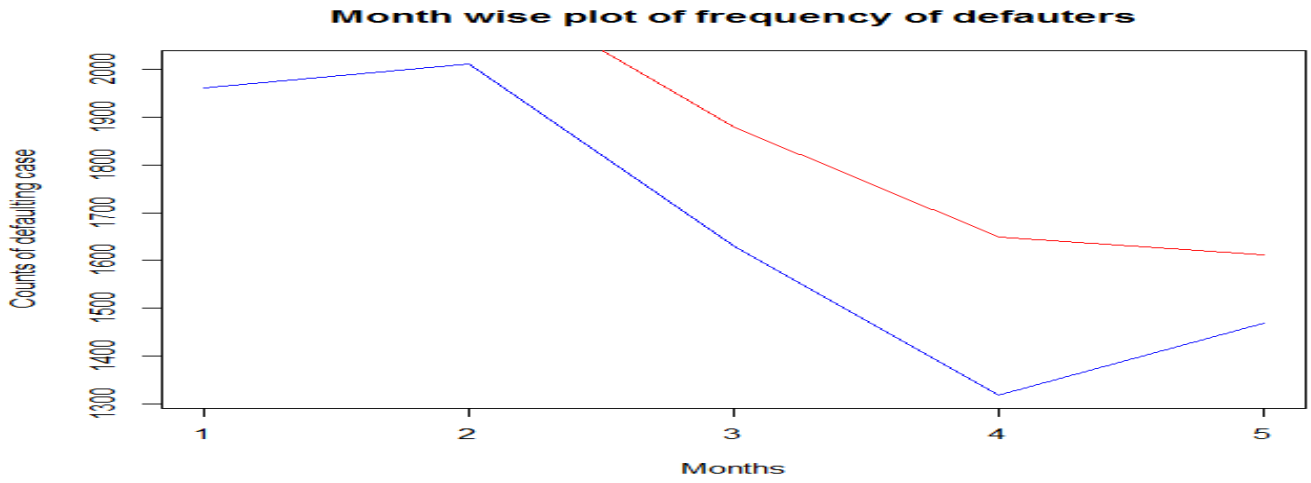
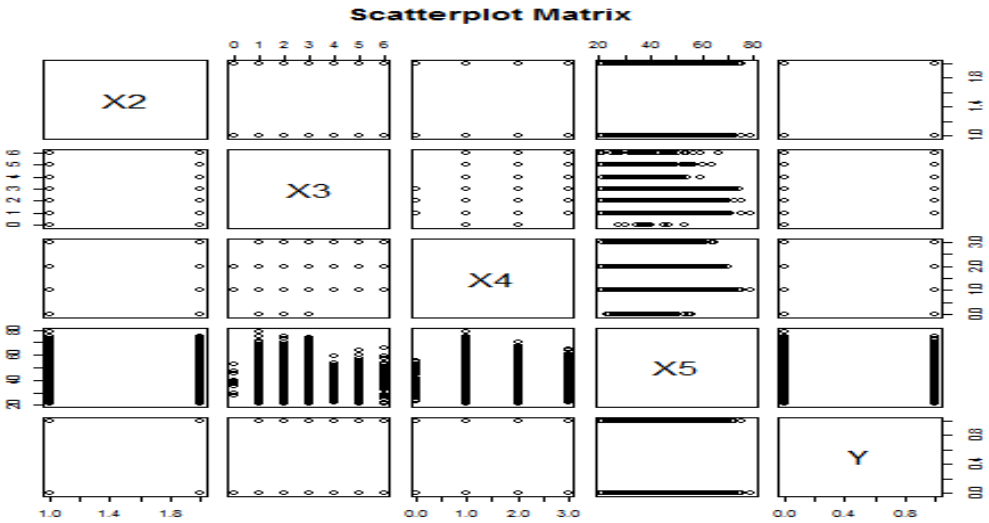


Figure 2: Time trend lines for defaulters and non-defaulters across previous months

For second hypothesis, we created a new variable which capture whether the customer with sex as male is married or not. We plotted the data as a pie chart where we can see the distribution of men among three group as single, married or others as shown in figure 3 and can see that some difference between the married men and the single person. Single tend to be more prone then the married person as seen from the visual data however again it can be a sample variation caused by the small data set used by used. The amount of data for the others is not significant to draw conclusion. We also tested similar hypothesis for the women and found that result was similar in case of women, marriage is not driving factor in women defaulter case. The graph of women case is also shown in the figure 3.

Figure 3: a) Defaulter men distribution b) Defaulter Women distribution on basis of marital status

Finally, we tried to get hidden relation from the data using correlation framework however not much interesting result were shown in the output have taken the scatter plot of categorical data. The correlation plots is given as following:



2.3 Data preparation for analysis

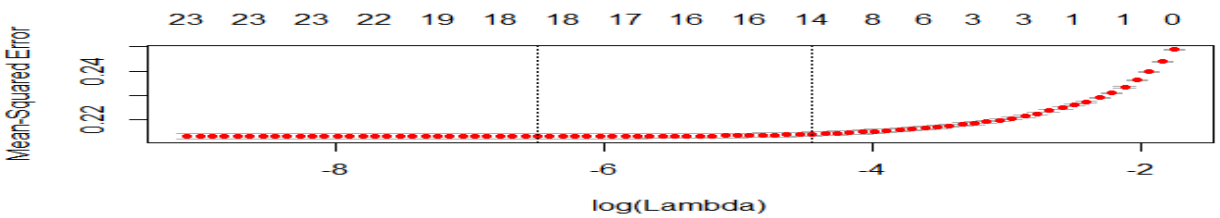
The open data provided by the UCI site was already normalized however there were some error in the data. We removed the missing data in salary using the median of the distribution. Error nous data in categorical data was removed rather than substituting the data with the average term. Finally, to reduce the bias due to high imbalance of the data, we employed SMOTE smoothing algorithm and used over sampling technique. We created a balance dataset of 35000 records. Finally, we created two data set with 75% and 25% split to create training and test data.

3. Result

We ran the three-classification algorithm on the 30,000 data-set which took estimated time of 10 mins to generate the result. The generalized error is enlisted in the table. Our result are significant only if the below assumptions are valid:

- the error terms need is independent and there exist linearity of independent variables and log odds
- The data are independent that data is independent and identically distributed data

For Baseline classifier, we used logistic regression with elastic net regularization. Ordinary logistic regression was overfitting so we deployed L1 and L2 mixed penalized regression on the mode. cross validation can also be used to select lambda as shown in plot. We also conducted Anova also on our model. The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better the model. The accuracy we got was 61.2%.



Next, we used Support Vector Machine for the classification. The model assumes that the data is linearly separable. We got a high accuracy of 71.5%.

Finally, we used the Random Forest which is the most popular ensemble learning method for classification. We got highest accuracy of 93.2 %. We restricted the number of tree to 1000 due to time constraint. The Random Forest also let us know the most notable features. In our model, according to random forest, being defaulter in month of July is the most important predictor.

We tabulated the result are enlisted below:

MODEL	ACCURACY SCORE
LOGISTIC REGRESSION	**61.2%****
SUPPORT VECTOR MACHINE	**71.5%*****
RANDOM FOREST	**93.3% ****

P-value * <0.1 ** < 0.05 *** <0.01

4. Conclusion

4.1 Final Model

We selected Random Forest as our model of selection. Because of the following reasons:

a) **Higher Accuracy:** Random Forest accuracy is significantly better than other two models. It reaches more than 90% with different selection of samples with variation of 5 %.
b) **High Outliers:** In Financial domain credit analysis, the probability of getting an outlier is high and thus Random Forest (using subset of training sets with bagging and subsets of features can help reduce their effect[5] A black swan is an event or occurrence that deviates beyond what is normally expected of a situation and is extremely difficult to predict. As Black Swan Effect is more probable in financial domain, which make it more appropriate to use random forest.

4.2 Summary

For getting insight and classify the credible and non-credible customer, we used data mining techniques and three supervised machine learning algorithms. For the analysis of the default credit card clients. For the extension of the model so that it can be utilized with bigger dataset and from the perspective of the feature data falling on a note of range which is being extended, we avoided normalizing. We did not modify it to a lower level and for the same reason we stick around the techniques which are not only helpful with the current data in its original form and can also be stretch to both ends. Obvious technique under these situations was Random Forest. As anticipated, it brought along the best accuracy out of the three mining techniques. We started with Logistic Regression as our baseline classifier which has in build assumptions of the dataset. We have used SVM as the third technique, if one expect that the data to is reasonably clean and outlier free, structural risk minimization using SVM is a powerful approach, however due nature of financial domain high outlier trend. Thus we present random forest as our final model.

5. Reference

[1] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Application.

[2] Rok Blagus and Lara Lusa, 2013, SMOTE for high-dimensional class-imbalanced data,

[3] http://archive.ics.uci.edu/ml/machine-learning-databases/00350/

[4] https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[5]https://www.researchgate.net/post/Is_random_forest_better_than_support_vector_machines