CREDIT RISK ANALYSIS USING DATA MINING TECHNIQUIES

Final Project Report
ISM6136.006F22 - Data Mining

Submitted by

Uma Srikanth Reddy Koduru	U94125452
Mohammed Zubair Shaik	U43640420
Dinesh Reddy Naredla	U96484815
Arun Kumar Pathipati	U37833602
Manikantha Varaprasad Inakollu	U55983409

Major in

BUSINESS ANALYTICS AND INFORMATION SYSTEMS

Under the guidance of

DR. MOHAMMADREZA EBRAHIMI

MUMA COLLEGE OF BUSINESS

UNIVERSITY OF SOUTH FLORIDA

Table of Contents

1. Introduction to Credit Risk Analysis

- Analyzing credit risk helps lenders evaluate if they want to extend credit based on a borrower's capacity to repay their debts.
- Before extending trade credit, you can determine the probability that a customer would overdue a payment by performing a credit risk analysis.
- High credit risk can have a negative effect on the lender by increasing collection costs and causing irregular cash flows.

Due to the banks' poor lending practices, the world experienced a global financial crisis in 2008. Customers who were unable to repay the loans were given credit by banks, which caused the bubble to grow and burst, which in turn caused a global economic slowdown. The "BASEL norms" are credit standards that have been agreed upon by all central banks worldwide in order to curb such risky lending conduct. Before issuing loans to retail and institutional clients, financial institutions are required by law to conduct an independent analysis of credit risk. The study is carried out by looking at the client's prior history.

Retail customers are the only ones included in our analysis. The lending financial institution performs the actual customer risk assessments, and independent financial agencies as Moody’s and S&P Global perform the fictitious ones. The capacity of the customers to repay is a main factor in retail risk. This study considers the repaying capacity (called independent variables). This study's conclusion classifies customers into good (timely loan repayment) and bad (late loan repayment) categories (defaulting type).

2.Purpose of credit risk analysis

To forward the objectives of the lenders, credit risk analysis seeks to assume an acceptable degree of risk. Goals may include things like business expansion, profitability, and qualitative elements.

Default risk is just one entity-specific risk element, even though credit analysis may rank risks and estimate the chance of default. When deciding whether the predicted outcomes are acceptable to their business and financial exposure, lenders take costs and advantages into account holistically.

Lenders use a variety of information from the borrower, the lender, and third sources like credit agencies to determine the cost of risk. Some metrics, including credit scores and credit risk analysis models, are instruments that let lenders calculate their projected loss by considering the likelihood of default.

3.Exploratory data analysis

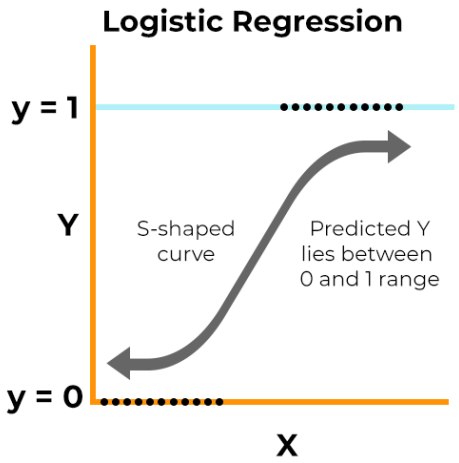
Exploratory data analysis is a crucial procedure that entails performing early investigations on data to find patterns, identify anomalies, test theories, and validate assumptions with the aid of statistical measures and graphical representations.

4. Algorithms

This machine learning algorithms are supervised learning algorithms, since the target variable is a labelled variable (either good-1 or bad-0)

4.1 Multiclass Logistic regression

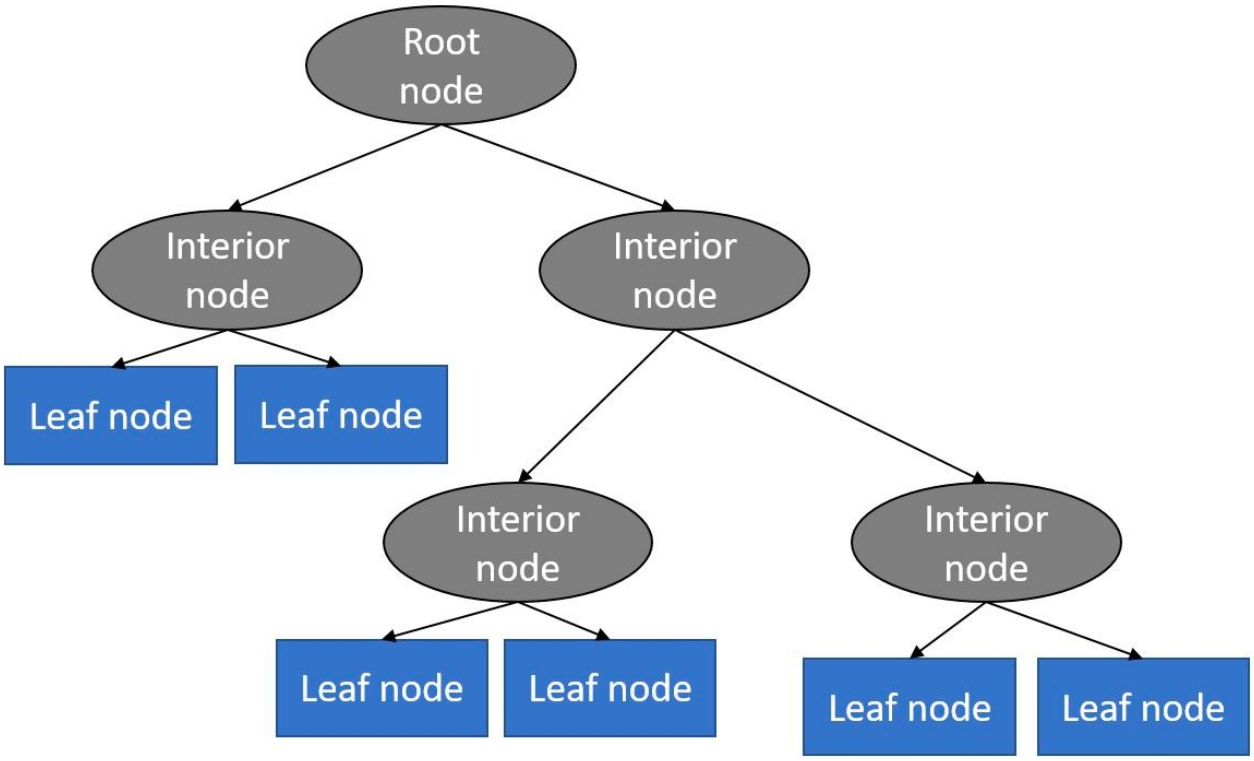
Logistic regression is a statistical analysis method to predict the probability of occurrence of an event. In our case, “0” indicates bad customer and “1” indicates good customer. Since the probability cannot be greater than 1 the predicted probability oscillates between 0 and 1. In the logistic regression model the sigmoid function is used. The independent variables are fed into this sigmoid function to get the output probability. The “x” in the sigmoid function represents the sum independent variables with individual beta coefficients.



$$f(x) = \frac{1}{1 + e^{-(x)}}$$

4.2 Multiclass Decision Forest

Decision forest or Random Forest is a combination of multiple decision trees, these combinations give higher accuracy in classification. The root node is split into multiple decision nodes, which further split and this process goes on and the split reaches a final point where the nodes does not split further the resulting node is called Terminal node.



5. Data Modelling

Data Set: Customer credit data set

Independent variables

1	months_since_earliest_cr_line	account open date
2	emp_length_int	employment length in years
3	term_int	number of payment installments
4	grade	loan grade
5	sub_grade	loan sub grade
6	home ownership	owned /rental / mortgage
7	verification status	verification of income
8	loan status	approved or declined
9	purpose	home / credit card/ car
10	addr_state	state

Dependent variable/Target variable

When the customer repays the loan will be termed as a good customer and vice versa. The Bad customer is indicated as “0” and the good ones as “1”. The output of the model explains the probability of “0” class or “1” class. The variable is termed “Good_Bad”.

5.1 Data pre-processing with python

Number of training Examples: 24,999 and Cleaning of data has been done using Python

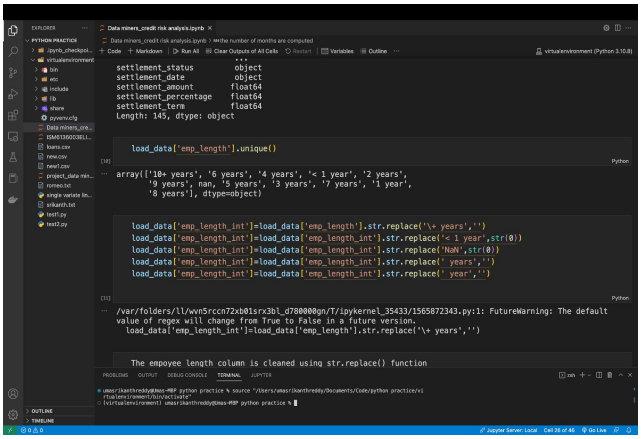


fig: python code

The original data set(loans.csv) has 2.2 million rows, so to make it easier to handle this data, a sample of first 25 thousand rows are taken for further preprocessing in Azure Ml.

5.2. Model Building with Azure ML

The output from the python script (a 25k training example set) is loaded into the Azure ML Now data is further cleaned (all the missing rows and columns are replaced with the MICE probability model). Next the dependent and independent variables are selected from the data set. Next two machine learning models (Logistic Regression, Random Forest) are selected. The data is split into 70% (for training) and -30% (testing). The remaining of the data is supplied to train model del one of Azure ML, for both the models. Now the “score model” component is added to the process flow (to confirm the results obtained). Finally, the results are visualized using the “Evaluate model” component.

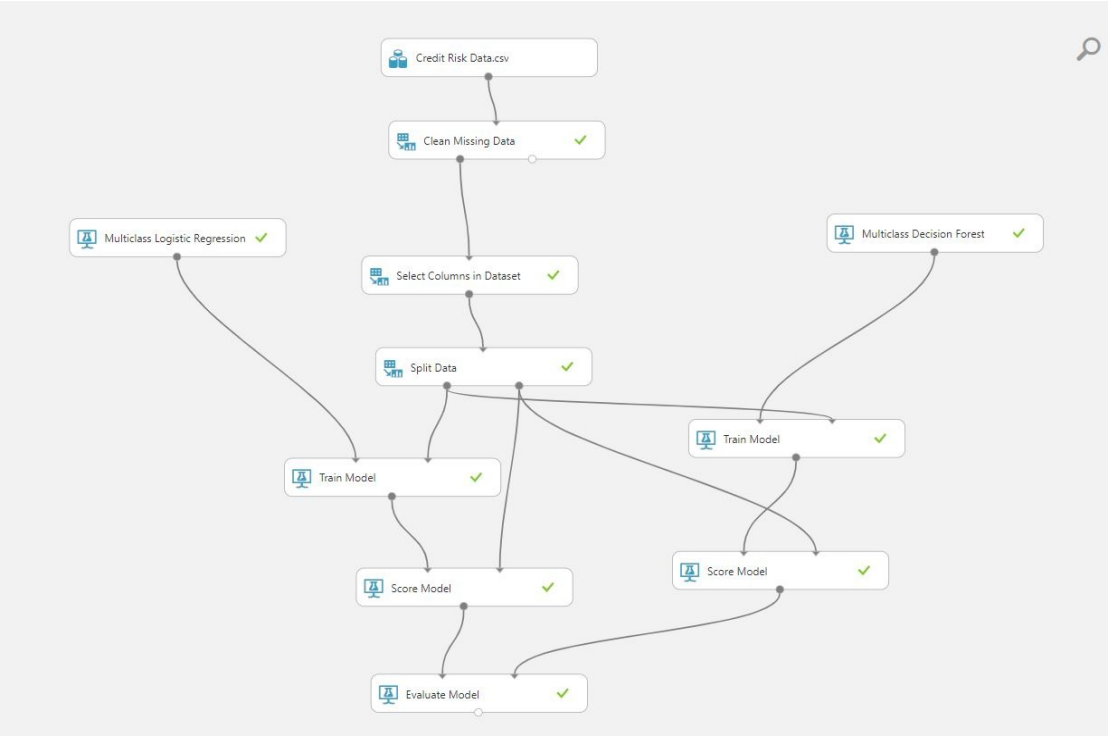


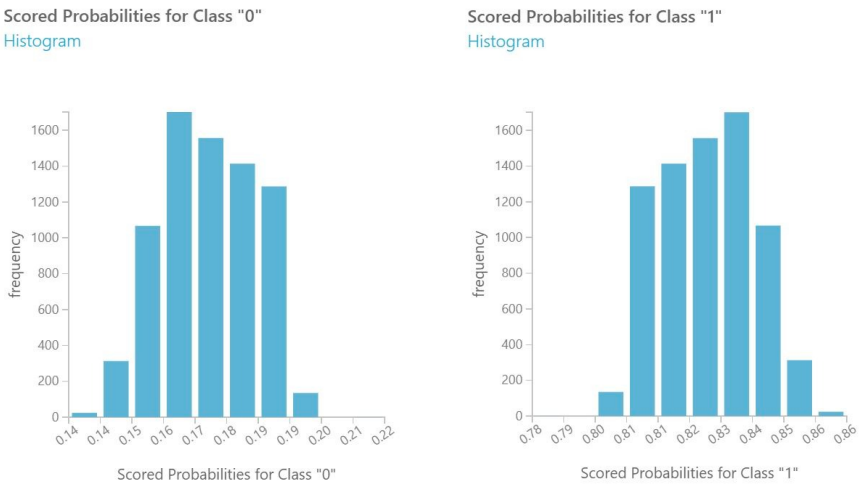
fig: azure ml model

6. Results

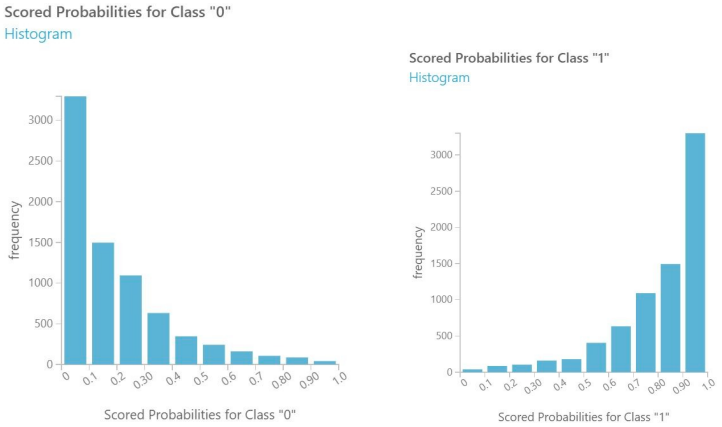
6.1 Model Analysis

Below are the metrics from the score model for the probabilities of class 0 and 1

a) Logistic Regression



b) Decision Forest



6.2 Final Model Evaluation

For the below metrics we consider Multiclass decision forest is predicting some risk metrics, as most of the data is skewed it may predict 100 percent to differentiate, we used below models.

The Logistic regression model seems to have a better accuracy over the Random Forest model, but if we look deeper into the regression model it is found that the model does not predict 0 if it were actually 0, in other words the model does not predict a default customer Random forest model predicts the default customer 10.2% of the times (though with relatively lower accuracy).

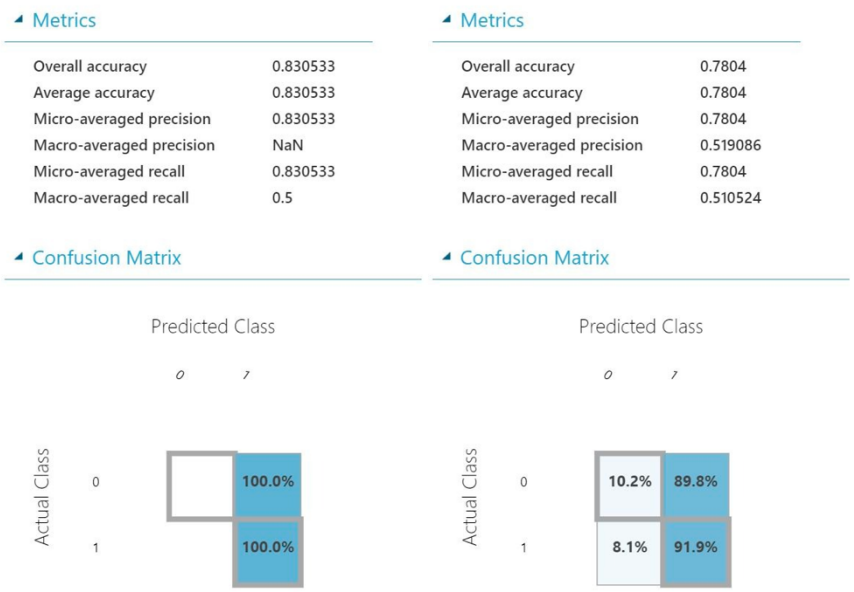


fig: results

7. Conclusion

The true positive in the Random Forest model is 91.9% and true negative is 10.2%. given that “0” means the customer is a bad customer and “1” means good customer. The model predicts 10.2% of the times an actual bad customer and 89.8% of the times the model says the customer is good, but the customer was bad.

Also, the model predicts 91.9% of the times a good customer if the customer was a good one, but it predicts 8.1% of the time a bad one if the customer was originally a good one.

8. References

- https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction/notebook
- https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/