ANALYSIS OF GERMAN CREDIT DATA USING MICROSOFT AZURE MACHINE LEARNING

N.Venkatesh, Snehika Pandey
Python Developer, Business Analyst
Modak Analytics, Hyderabad, India

Abstract**—*****Banking Industry is a vital supply of Finance in any country. Credit Risk Analysis could be an essential and decisive task in banking sector. Loan Sanction procedure is always preceded by the credit risk analysis of the client. Thus, credit risk needs to be effectively compared. In this paper, we aim to build a classification model (Logistic Regression) using a dataset available in UCI repository. Data set is processed prior to modelling. Model is trained, Test dataset is used for predicting the outcome needed for Business Decision.***

Index Terms**—*****German Credit Data, Azure Machine Learning***

OBJECTIVE

The Main objective of this project is to classify the individuals as good or bad credit risks using a group of attributes. Building a Model to predict the credit risk related to a client, supported by its profile attributes.

INTRODUCTION

Whenever bank receives a loan application, based on the applicant's/application profile bank decide whether to go ahead with the loan approval or not. Basically, we have two types of risks associated with the bank's decision: -

- If the applicant has a good credit risk, i.e. is likely to repay the loan, then not approving the loan of the person will results in loss of business to the bank.
- If the applicant has a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person will results in financial loss to the bank.

ANALYSIS

MINIMIZATION OF RISK AND MAXIMIZATION OF PROFIT ON BEHALF OF THE BANK.

To minimize loss from the bank's perspective, the bank needs a Decision Rule, regarding whom to give approval of the loan and whom not to. A candidate's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

German-Credit Data contains information through twenty- one attributes, and the model classifies whether an applicant lies in the category of Good or a Bad Credit Risk out of 1000 Applicants. A predictive model developed using Attributes is anticipated to guide whether to approve loan of a prospective individual supported his/her profile.

Attribute and their information

Variable name	Description
Creditability	Label
Balance	Account balance
Duration	Duration of credit in months
History	History of previous credit
Purpose	Purpose of credit
Credit amount	Numerical
Savings	Value in savings or stocks
Employment	Length of current employment
Insper	Installment percent
Sex married	Sex & marital status
Guarantors	Guarantors or proof or asset
Residence duration	Duration in current address
Assets	Most valuable available assets
Age	Age(years)
Conc credit	Concurrent credit
Apartment	Type of apartment
Credits	No of credits at this bank

Occupation	Current status of the beneficiary
Dependents	No of dependents
Telephone	Had/has phone
Foreign	Foreign worker

LOGISTIC REGRESSION

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured by a dichotomous variable (in which there are only two possible outcomes).

Logistic Regression Model Using Microsoft Azure (Machine Learning):

Here we are using **Two-class logistic regression** model for building a logistic model which can be used for predicting two (and only two) outcomes.

Logistic regression is a prominent method in statistics that is used to predict the likelihood of an outcome and is exclusively common for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function.

CONFIGURING TWO-CLASS LOGISTIC REGRESSION IN AZURE (Machine Learning):

For training the model we must provide a dataset which contains a class column or label, this model is intended only for two class problems where label should have exactly two values.

For example, the possible values of label variable can be “Yes” or “No”, “High” or “Low”, “Win” or “Loss”, “1” or “0”

Implementation Steps - logistic regression

- Drag and drop the Data set (German credit data/german_credit.csv)
- Click on output circle and then visualize
- Check out the column names
- Find out the dimensions of the dataset
- Drag-and-drop 'select column from dataset' and select Creditability
- Search for Logistic Regression', drag-and- drop it, into the canvas
- Click on Logistic Regression' make sure that in properties window 'Ordinary Least Squares' is selected for solution method
- Search for 'Train Model', drag-and-drop into the canvas
- Connect the output of Logistic Regression' to left input of the 'Train Model' 'select column from dataset' to right input of the 'Train Model'
- Click on 'Train Model', select launch column selector in the properties window
- Select the column(Creditability) for which the prediction has to be done
- Drag-and-drop 'Score Model' from left pane and uncheck the 'Append score column' in properties window
- Connect the output of 'Train Model' to left input of the 'Score Model' 'select column from dataset' to right input of the 'Score Model'
- Drag-and-drop 'Evaluate Model' from left pane
- Connect the output of 'Score Model' to the input of 'Evaluate Model'
- Click on Run
- After execution click on the output circles of 'Train Model', 'Score Model' and 'Evaluate Model' to see the value of R-squared
- In the experiment click ag click on "set up web service" in the bottom pane
- Select Retraining Web Service
- Click on "RUN" to run in the bottom pane
- After execution again click on "set up web service" in the bottom pane and select Predictive Web Service
- Again, click on "RUN" to run in the bottom pane
- After execution click on "Deploy web service" it will deploy and take you to the web service page
- Click on the Test button, enter data to be predicted
- Predict Window will open with the Output

Logistic Regression: Training Experiment

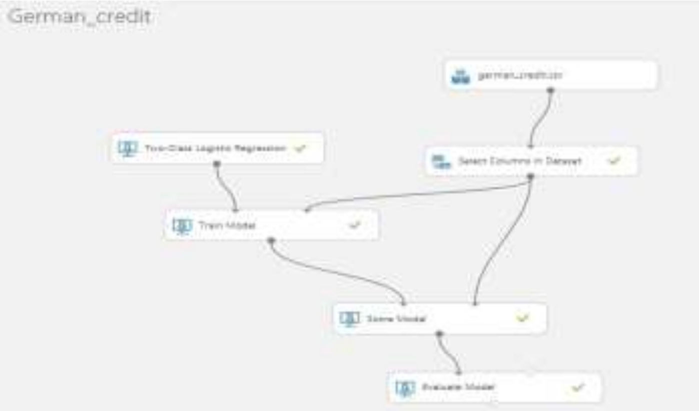
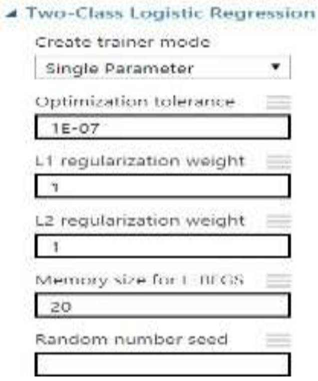


Figure 1: Training Experiment for German Credit Data (Logistic Regression)



Two-Class Logistic Regression

Create trainer mode: Single Parameter

Optimization tolerance: 1E-07

L1 regularization weight: 1

L2 regularization weight: 1

Memory size for L-BFGS: 20

Random number seed:

Figure1.1: (Logistic Regression) Parameters

Logistic Regression: Predictive Experiment:

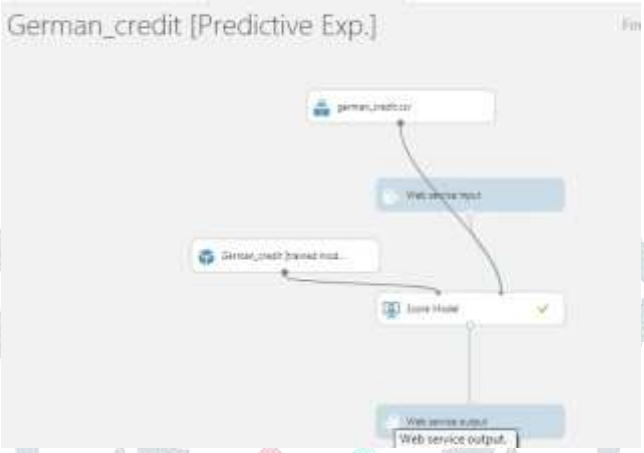


Figure 2: Predictive Experiment for German Credit Data (Logistic Regression)

WHAT TRAINED MODEL DO IN AZURE MACHINE LEARNING

Training a classification model is a kind of *supervised machine learning*. Which means you should provide a dataset that comprises historical data from which to acquire patterns. The data must contain both the outcome which we are trying to predict, and connected factors (variables). The machine learning model uses the data to retrieve statistical patterns and build a model.

HOW TO TRAIN MODEL?

- Add **Train Model** module to the experiment. We can find this module in Azure Machine Learning Studio under the **Machine Learning** category. Expand **Train**, and then drag the **Train Model** module into your experiment.
- On the left input, attach one of the classification models providing in Azure Machine Learning Studio.
- Attach a training data to the right-hand input of **Train Model**.
- For **Label column**, you should identify a single column that contains outcomes of the model which we can use for training. Now Click **Launch column selector** and choose the column in the dataset that contains the values you want to predict. For classification scenarios, the label column should contain categorical values or discrete values. Some samples might be a yes/no rating, a disease classification code or name, or an income group. If you choose a non-categorical column, the module will return an error during training.
- Run the experiment. If you have a lot of data, this can take a while.

SCORE MODEL?

- We can use score model for generating predictions by using trained classification model, predicted values can be in various formats which depends on model and input which we provide.
- If we are going with regression models, the score model generates a predicted value for the class, and probability of the predicted value.
- In case of regression models score model generates just predicted numeric value.
- After generating set of scores by using score models. We can connect scored dataset to evaluate model for finding metrics and performance of the model.

MODELS WHICH ARE NOT SUPPORTED BY SCORE MODEL:

If we are using different types of models we should use the following any one of the score model.

- Score a cluster model – Allocate to cluster (deprecated).
- Creating Recommendations-Score matchbox.

MODEL EVALUATION:

We can use evaluate model for measuring the accuracy of the trained model, the metrics which are generated by evaluate model are purely depend on the type of model that we are evaluating.

Here we have 3 ways to use evaluate model:

- Generate scores on training set.
- Generate scores on model and compare with reserved testing set.
- Compare scores for couple of related models on the same dataset.

Output of logistic regression for German Credit:

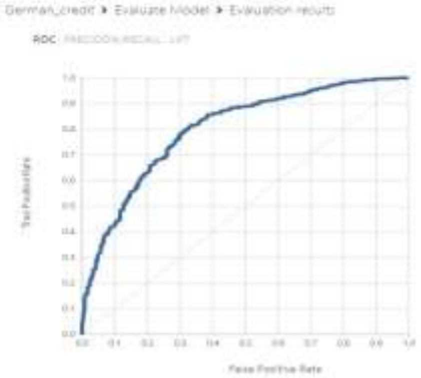
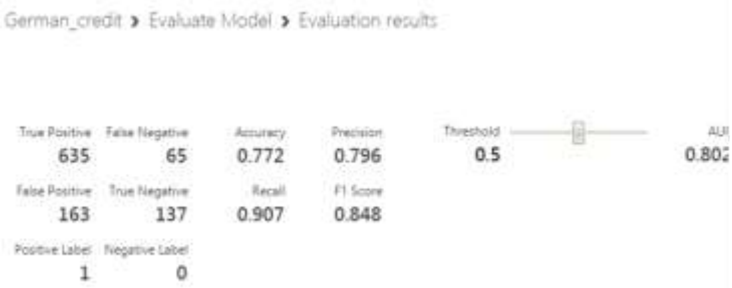


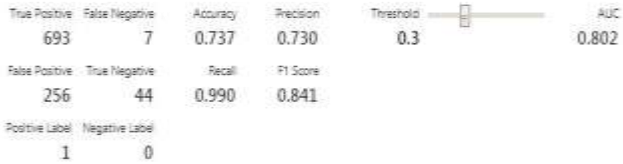
Figure 3: Roc curve



German_credit ▸ Evaluate Model ▸ Evaluation results

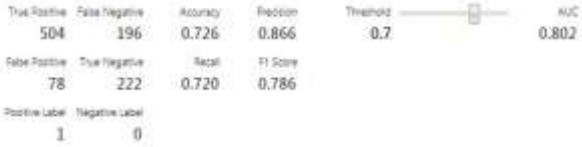
True Positive 635
False Negative 65
Accuracy 0.772
Precision 0.796
Threshold 0.5
AUC 0.802
False Positive 163
True Negative 137
Recall 0.907
F1 Score 0.848
Positive Label 1
Negative Label 0

Figure 4: Confusion Matrix & Accuracy when threshold is 0.5



True Positive 693
False Negative 7
Accuracy 0.737
Precision 0.730
Threshold 0.3
AUC 0.802
False Positive 256
True Negative 44
Recall 0.990
F1 Score 0.841
Positive Label 1
Negative Label 0

Figure 5: Confusion Matrix & Accuracy when threshold is 0.3



True Positive 504
False Negative 196
Accuracy 0.726
Precision 0.866
Threshold 0.7
AUC 0.802
False Positive 78
True Negative 222
Recall 0.720
F1 Score 0.786
Positive Label 1
Negative Label 0

Figure 6: Confusion Matrix & Accuracy when threshold is 0.7

REFERENCES:

[1] Germano C. Vasconcelos, Paulo J. L. Adeodato and Domingos S. M. P. Monteiro, "A Neural Network Based Solution for the Credit Risk Assessment Problem", Proceedings of the IV Brazilian Conference on Neural Networks-IV Congresso Brasileiro de Redes Neurais pp.269274, July 20-22, 1999-ITA, Sao Jose dos Campos-SP- Brazil.

[2] Vincenzo Pacelli and Michele Azzollini, "An Artificial Neural Network Approach for Credit Risk Management", Journal of Intelligent Learning Systems and Applications, 2011, 3, pp. 103-112.

[3] Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi, "A Hybrid Method for Credit Risk Assessment of Bank Customers", International Journal of Trade, Economics and Finance, Vol. 2, No. 2, April 2011.

[4] Hamadi Matoussi and Aida Krichene, "Credit risk assessment using Multilayer Neural Network Models-Case of a Tunisian bank" 2007.

[5] Eliana Angelini, Giacomo di Tollo, and Andrea Roli "A Neural Network Approach for Credit Risk Evaluation", Kluwer Academic Publishers, 2006, pp. 1-22

[6] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen, "Credit scoring using the hybrid neural discriminant technique", Expert Systems with Applications (Elsevier) 23 (2002), pp. 245-254.

[7] Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study", Decision Support Systems (Elsevier) 37 (2004) pp.543-558