CLOUD APPLICATION AND INTEGRATION OF COBRA BUILDS

A Project Report Submitted

for the Course

MA691

by

Ansh Bhatt(180123005)
Damayanti R Sambhe(180123010)
Manav Chirania(180123026)
Priya Gulati(180123034)
Satyadev Badireddi(180123041)
Tanmay Jain(180123050)

to the

Department of Mathematics

Indian Institute of Technology Guwahati

Guwahati - 781039, India

November 2021

Disclaimer

This is to certify that the work contained in this project report entitled "**Cloud application and integration of Cobra Builds** " submitted by Ansh Bhatt(180123005), Damayanti R Sambhe(180123010), Manav Chirania(180123026), Priya Gulati (180123034), Satyadev Badireddi(180123041) and Tanmay Jain(180123050) for the course of MA691.

This work is for learning purpose only. The work can not be used for publication or as commercial products etc without mentor's consent.

Guwahati - 781 039
November 2021

Prof. Arabin Kumar Dey
Project Supervisor

ABSTRACT

We used a German Credit dataset with details of applicants applying for a loan classified as good/bad customers. We aim to classify applicants into good/bad customers for loan lending organizations based on applicants' details by using exploratory data analysis and machine learning algorithms. Results were improved by implementing the classifier COBRA aggregation scheme by using Logistic regression, SVM, K-Nearest Neighbors, Gaussian NB and Decision tree classifier. Accuracy of 74% was obtained and recall for bad customers was improved. Later, cloud integration of the trained model was done.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are given a german data set with details of applicants applying for a loan classified as good/bad customers. A good customer is an applicant who has repaid their outstanding loan, whereas a bad customer has defaulted. We use this data to train the machine learning models to classify an applicant into good/bad customer prior to lending loans to help a loan lending organization make better decisions.

As we proceed, we will get an idea about how real business problems are solved using Exploratory Data Analysis (EDA) and Machine Learning (ML). We further developed a platform to ease the prediction process.

1.1 Problem Statement

An applicant's profile is considered when any loan lending organization makes a decision for loan approval. Primarily, two types of risk are associated with lending loans, namely,

1. Not approving the loan in the case when the applicant is likely to repay. This might result in a loss of business for the company.

2. Approving the loan in the case when the applicant is not likely to repay, *i.e.,* is likely to default. It may lead to a financial loss for the organization.

The data contains information about past loan applicants and whether they were "defaulted" or not. The main aim of this project is to classify an applicant into good/bad customer. Companies can use this information to take actions such as denying the loan, reducing the loan amount, lending at a higher interest rate, or to hedge against such defaults. Further, suppose that the borrower is likely to default. In that case, the goal is to be on safer side and prevent the companies losses.

1.2 Data Description

The dataset used was obtained from **UCI**. The data contains details of all the applicants who applied for a loan at a loan lending organization. The dataset has 1000 data points, with 700 good customers and 300 bad customers. The data has 20 attributes in total, 7 numerical and 13 categorical. The default probabilities were considered as 0 and 1, for the ones who paid and those who defaulted, respectively. The data involves the details of the applicants for the loan, such as credit history, purpose of loan, status of checking and savings account etc. The Machine Learning (ML) algorithms and the classifier COBRA aggregation scheme were further applied on this data after data exploration, data pre-processing and feature cleaning. The nomenclature of the data for over 1000 data points is as given below in **Table 1.1**.

Variables	Description
Attribute 1	Status of existing checking account
Attribute 2	Duration in month
Attribute 3	Credit history
Attribute 4	Purpose of the loan
Attribute 5	Credit amount
Attribute 6	Savings account/bonds
Attribute 7	Present employment status
Attribute 8	Installment rate in percentage of disposable income
Attribute 9	Personal status and sex
Attribute 10	Other debtors / guarantors
Attribute 11	Present residence
Attribute 12	Property
Attribute 13	Age in years
Attribute 14	Other installment plans
Attribute 15	Housing
Attribute 16	Number of existing credits at this bank
Attribute 17	Job
Attribute 18	Number of people being liable to provide maintenance for
Attribute 19	Telephone
Attribute 20	foreign worker

Table 1.1: Nomenclature of the Data Considered

Chapter 2

Results and Implementation

2.1 Exploratory Data Analysis

2.1.1 Overall Goal

The goal is to apply a classification model, on the data considered, in order to predict whether an applicant is likely to default on the loan or not. Hence, it is necessary to identify and understand the essential variables, view summary statistics, and visualize the data. Some of the observations are given in next section.

2.1.2 Observations

Feature selection was performed on the considered data to develop a predictive model. It is desirable to reduce the features, and only keep the relevant ones to reduce the computational cost and improve upon the performance of the model. The statistical-based feature selection methods were used to assess the correlation between every input variable and the target variable. The input variables that have the strongest relationship with the target vari-

able were selected. For selecting the relevant input variables, a comparative analysis was performed and some of the observations and the improvements done are noted below:

1. The values of **duration** ranged over a long duration of 4 to 72 months. This attribute was converted into bins of 10 since some values had very few or no entries corresponding to them. Converting to bins of 10, helped in efficient training of the model.

2. Similarly, **Age** attribute was converted into bins of 15.

3. **Installment Rate** attribute was dropped since the results didn't seem to depend on the rate.

4. The data set was imbalanced with the count of 700 good customers and 300 bad customers.

5. The values of **Installment Rates** and number of **Credits at the bank** were limited to 4, whereas the **Number of People Liable** was limited to only 2.

The graph in **Figure 2.1** shows the correlation between various features and loan_status (*i.e.,* whether the debtor will default or not). Positive correlation implies more the value of that particular feature, more is the chance of that debtor to default on their loan.

2.1.3 Data Preprocessing

Data pre-processing was done before applying the classification model to obtain the probabilities of default. This involved removing or filling missing data, removing unnecessary or repetitive features, and converting categorical

Figure 2.1: Correlation Coefficient between Target Variable and Various Attributes

features to dummy variables. This gave us a training set of size 1158 entries. Some of the columns were dropped to improve model performance. These include 'Employment Length', 'Installment Rate', 'Residence Since', 'Telephone' and 'Foreign Worker'. The pre-processed data was split into training set and test set (80-20). Since the dataset was imbalanced, oversampling was done on the training set using SMOTE (Synthetic Minority Over-sampling Technique). The training data was used to train the classification model and the results were evaluated on the test set later, which can be seen in the next section.

2.1.4 Classifier COBRA Aggregation Scheme

PyCOBRA is used for ensemble learning. It serves as a toolkit for regression and classification using ensembled machines, and also for visualisation of the performance of the new machine and constituent machines. ClassifierCobra performs a majority vote among all points which are retained by the COBRA procedure and calculates probability of a point being in a particular class. We split the data into different parts for training machines and for aggregation(50-50 split of initial training data). Five different machines from the scikit-learn library have been used as the constituent machines.

Models Used

- Logistic Regression
- k-Nearest Neighbours
- Gaussian NB
- SVM
- Decision Tree

2.1.5 Results

The Accuracy Score is 74% and the Classification Report for the "Test Set" is given in **Table 2.1**.

	0.0	1.0	Accuracy	Macro Average	Weighted Average
Precision	0.84	0.55	0.74	0.70	0.76
Recall	0.77	0.66	0.74	0.72	0.74
f1-score	0.81	0.60	0.74	0.70	0.75

Table 2.1: Test Evaluation Metrics for Classifier Cobra

The Confusion Matrix is given in **Figure 2.2**.

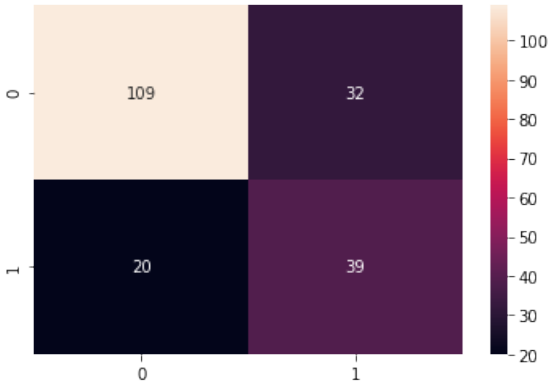

Figure 2.2: Confusion Matrix for Classifier Cobra

This dataset requires use of a **cost matrix** :

$$\begin{bmatrix} & Good & Bad \\ Good & 0 & 1 \\ Bad & 5 & 0 \end{bmatrix}$$

The rows represent the actual classification and the columns the predicted classification. It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1). The cost after applying Classifier Cobra decreased significantly to **132** as compared to > 170 when we applied individual models. We also observed an improvement in recall of bad customers.

Chapter 3

Conclusion

Upon implementing the Cobra classifier on the available data, we observed improvement in the model performance, i.e. reduction in the cost as obtained using the cost matrix mentioned above. Further, the model was deployed using the Django framework on PythonAnywhere.

The github link including the code for all the implementations, including the ML model and the deployment can be found here: **Cobra 10 Github**

The model deployed on PythonAnywhere can be accessed here: **Deployed Model**