Fundamentals of Data Mining - Practical Midterm

Mohammad Bahrami - 9724133

May 26, 2022

1 Part a

The Problem is to predict whether a credit card applicant should receive one or not. More specifically, Each applicant has a set of attributes at the time of filling their application. The problem to tackle is to decide whether an applicant should receive a credit card given the set of attributes for that applicant. This decision is vital because we can predict whether the applicant is a trustworthy person who will return the credit given by the bank or not.

2 Part b

The Dataset contains 690 samples. Each sample has 14 predictive attribute and one target attribute that is whether a sample with its set of predictive attributes should receive a credit card or not. The predictive attributes are a combination of *Categorical, Integer and Real* values and each numerical value contains a different range of values than others. There are six numerical and eight categorical attribute for each sample.

The attribute names has been removed to keep the confidentiality of the data. Also, a key point to keep in mind is that the categorical values are given numbers instead of the actual string of the category but they should not be treated as numbers and should be counted as categories despite the fact that they are shown by numbers. The dataset has had a few missing values (around 5% of the samples) but they have already been resolved by replacing with the *mode* for the categorical attributes and with the *mean* for the numerical ones. Finally, The target class distribution is 44.4% positive and 55.5% negative class.

3 Part c

For The Numerical Attributes , we can plot the heat map of their linear correlation with each other and with the target value. the result is shown in the part c section of the python notebook. The numerical features don't seem to have any linear correlation with each other and one of them has a minor correlation with the target value.

For The Categorical Values , we can plot their value counts bar plot and overlay the value with target value. This shows the count of each value in each categorical feature clearly but not the portion of the target value in each value. To show the categorical values on the target value and the portion of negative and positives more clearly, the bars are normalized in the second column in the figure (see the part c in the notebook).

4 Part d

For the Preprocessing part, first i have checked for the skewness in the numerical features and have decided to apply the *inverse square root* method as it is known to work best out-of-the-box. Then, I have normalized the numerical columns using the *Z-Score* method because I can use it later on to recognize the outliers. Next, I have searched for outliers in the numerical columns and because I have no information in the financial field, I have remove the samples with outliers.

5 Part e

e1 I chose MLP for the first classifier as it is a general function approximator. For all the e1,e2,e3 parts i use 10-fold cross validation for the validation accuracy and to fine tune the parameters of each model. in the case of mlp, the layers and the regularization.

e2 For the second model i use svm classifier as it is a robust option to chose when choosing models for the fields that we have no idea about. i have tuned the regularization parameter to be on 0.5 and chosen the default rbf kernel.

e3 For the third model I use KNN as a representative of simpler models and have tuned the number of neighbors to be 5 on the validation set.

Also, For the metrics, I have printed the accuracy and the recall and precision and F1-Score. I Also have shown the Confusion matrix.

6 Part f

I have chosen the stacking ensemble with a SVC final estimator. no time to explain more!