Anomaly Detection Challenge 1

Vishal Bhalla Matriculation No: 03662226
Nithish Raghunandanan Matriculation No: 03667351
German Credit Dataset Classification

November 8, 2015

1 Abstract

The goal of this challenge was to implement a machine learning algorithm (classifier) to classify the users in the German credit dataset. We have implemented a model that uses linear regression to classify the users.

2 Data Pre-processing

The training dataset consisted of 500 samples with binary labels - 0 denoting good user and 1 denoting bad user. The dataset has 20 attributes/features. The attributes in the dataset consisted of both numerical and qualitative/categorical features. The qualitative values were of the form A + attribute number + numerical value. Firstly, we processed these qualitative attributes to remove the prefix A and assign it sequential numbers to make it numerical.

2.1 One Hot Encoding (OHE)

While converting the qualitative features to numerical features using sequential numbers for each feature serves us good for classification but doesn't give good accuracy. This is due to the fact that the numbers are assigned weights based on the values of the numbers we assigned. As such there is no order of importance between these categorical features for us to assign sequential numbers. Hence, we use another technique called as One Hot Encoding (OHE) to assign a binary vector of 0's and 1's for each feature where 1 indicates the presence of that value in that feature.

2.2 Normalizing the numerical features

We normalized the data, especially the numerical features and this normalized data gives us better accuracy as compared to de-normalized data.

3 Model

We spent most of our time in pre-processing the data. We then tried different types of Regression Models to fit our data, namely Linear, Polynomial, Logistic, Bayesian Ridge among others. We also tried SVM with Linear, RBF & Non-linear Kernels. Some of the concerns and issues faced during Model Selection were as below:

1. Dimensionality reduction - We tried to reduce the dimension of data from 59 features to 48 using Principle Components Analysis (PCA) for SVM classifiers in particular which did improve the accuracy for it but not better than that of a Linear Classifier.

2. While fitting a polynomial model, we had to make the optimum degree selection.

3. Cross-Validation - To decide the best model by splitting the training samples in the ratio of 4:1.

4. Regularization - To reduce over-fitting of the models, we tried to assign suitable penalties. But we could not find a suitable penalty for the linear regression models.

4 Results

The performance is evaluated by computing the area under Receiver Operating Characteristics (ROC) curve i.e. termed as (AUC). The results of the different models on the German Credit Dataset are depicted as below:

Table 1: Summary of Area under the Curve (AUC) of different Models

Model	**AUC**
Linear Regression without OHE & Normalisation	0.6125
Linear Regression with OHE	0.7086
Linear Regression with OHE & Normalisation	0.71349
Linear SVM with OHE & Normalisation	0.60735
Non- Linear SVM with OHE & Normalisation	0.54044
Linear Regression with OHE & Normalisation & PCA	0.62816
OHE using Bayesian Regression with Normalization	0.68588

5 Conclusion

By trying out different models on our training set, using de-normalized and normalized data, we infer that the best model is the one incorporating linear regression on normalized numerical feature data and one hot encoding on the categorical features which provided 71.0349% accuracy on the test set.

Regularization would help reduce the over-fitting, but we could not find the best penalty factor for our Linear Regression Model.