University of Essex

Department of Computer Science and Electronics Engineering

Report: Part II B

Subject: CF969 Big-Data for Computational Finance

Registration number: 2201829

Supervisor: Dr. Panagiotis Kanellopoulos

Date of submission: 21/04/2023

CF969 Assignment - Report Part 11B

Introduction

The main objective of this document is to summarize the parameter selection, results and analysis obtained from the below 3 classifications – Linear Regression, Logistic Regression, and Neural Network for the given dataset.

1) Linear Regression

Generally, linear regression is not ideal for binary classification as it is commonly used for regression problems. Binary classification has been implemented here by converting the result (which is float values) to binary by mapping the result to a pre-defined threshold. In this case, I have chosen the threshold as .6, in such a way that if the value obtained from prediction is greater than or equal to 0.6, it will be considered as 1 and if less than 0.6 then it will be converted to 0. Our aim is to classify whether the firm is investment grade or not.

i) **Parameters Selection: '**alpha' (regularization strength) is the parameter used for the hyperparameter tuning. It controls the amount of regularization that can be applied to the model. 'alpha' value is smaller means less regularization and a larger value means more constraints on the model coefficient. I have chosen the alpha value as .0001, after comparing different matrices like accuracy, roc value and f1 score from the classification report. I couldn't find any change in the value of evaluation matrices for results obtained using ridge regularization, but Lasso shows significant improvement after fine-tuning with different values of alpha, after .0001 the ROC value start decreasing so chose .0001as the value for alpha. Table 1 contains the result of Lasso regularization with different alpha values.

Alpha	Accuracy	ROC
.1	.752	.45
.01	.758	.58
.001	.767	.623
.0001	.7735	.6258
.00001	.7735	.6251

Table .1 Accuracy and ROC score for different values of alpha

ii) **Result and Analysis**: We got an accuracy of 77%, indicating that both models are correctly predicting the overall class labels for approximately 77% of the validation data. For Ridge and Lasso regularization, a maximum ROC value of .6258 indicates that the model's performance is better than a random chance and the model has the capability to discriminate the negative and positive classes. But it shows poor performance in the case of F1 score. Figure-1 contains the ROC curve for binary classification with Ridge and Lasso.

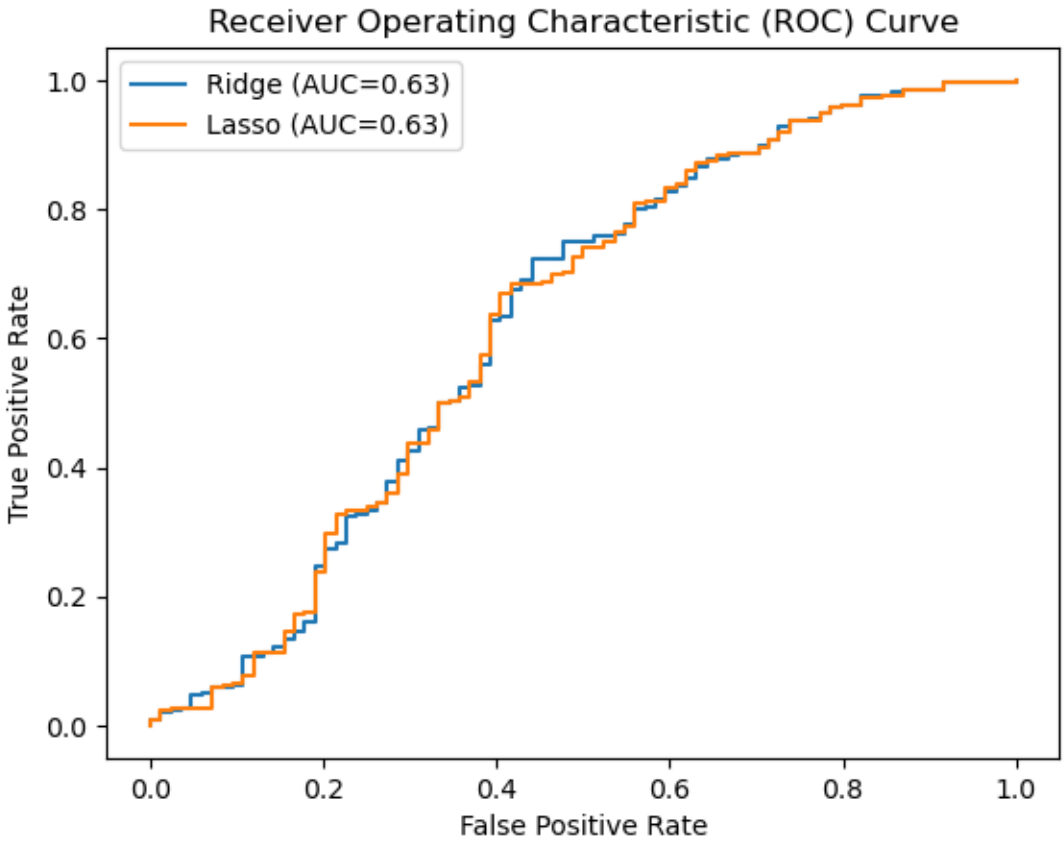

Figure .1 ROC curve of linear regression Lasso and Ridge

Below is the classification report obtained from linear regression using ridge and lasso:

```
Classification Report with Ridge regularization:
              precision    recall  f1-score   support

           0       0.73      0.13      0.22        84
           1       0.78      0.98      0.87       256

    accuracy                           0.77       340
   macro avg       0.75      0.56      0.54       340
weighted avg       0.76      0.77      0.71       340

Classification Report with Lasso regularization:
              precision    recall  f1-score   support

           0       0.73      0.13      0.22        84
           1       0.78      0.98      0.87       256

    accuracy                           0.77       340
   macro avg       0.75      0.56      0.54       340
weighted avg       0.76      0.77      0.71       340
```

iii) **Suitability of Approach:** Linear regression is generally not suitable for binary classification, but the use of regularization using Ridge and Lasso gave a good

accuracy of 77%. A low F1 score and recall for '0' indicates that model is less capable of classifying 0 both in the case of Ridge and Lasso regularization, saying still there is scope for improvement. Tried to train the model by removing outliers and normalization, but performance was poor. Also tried balancing with oversampling but there is an improvement in accuracy.

2) Logistic Regression

Logistic regression is the most used classification method, especially for binary classification. Here logistic regression with Ridge and Lasso has been implemented and details of implementation are described below:

i) **Parameter Selection:** The parameters for logistic regression has been selected using GridSearchCV and the best parameter alpha we got after grid search is, for Lasso - .01 and for ridge - .001.

ii) **Result and Analysis:** After classification, we got an accuracy of 75% and ROC value of .50 and.51 for logistic regression with ridge and lasso. Compared to linear regression, logistic regression shows poor performance. In most cases, the model failed to classify the values for 0. The ROC value of.50 indicates that the result is closer to random guessing, Following are the confusion metrics obtained after the classification.

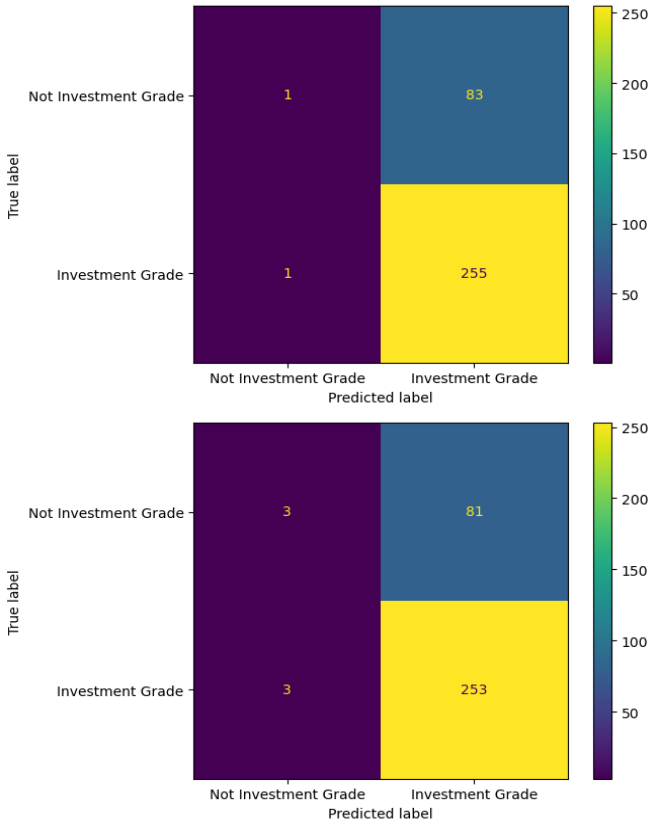

Figure 2. Confusion matrix of Logistic Regression using Lasso and Ridge

iii) **Suitability of Approach:** From the result, it is clear that the model has difficulty in classifying values of 0. Data is imbalanced in the training dataset, which might be the reason for this. Tried oversampling for balancing the data, along with normalization,

and scaling using power transformers and also, trained the model by removing outliers. But didn't get any significant improvement.

3) Neural Network

We have to solve two problems here one is to classify the model to predict the investment rating and use that to predict whether it is an investment grade or not. The first part is a multiclass problem while the second part is binary classification. I first predicted the investment rating and using that I predicted the investment grade. the details of the implementation and result are as follows:

i) Parameter Selection – I have tried different approaches for fine-tuning the model.

a) Added different layers with different activation functions such as 'relu', 'sigmoid' and 'tanh' but the current implementation with 'relu' and 'softmax' in the current design gives the best accuracy.
b) Using different learning rates - .001, .0001, .00001,.000001 and .0000001 tried to fine-tune the model. Lr = .0000001 gives better accuracy.
c) Tried different optimizers like Adam, and SGD. Adam optimizer outperformed SGD. Tried adding regularization but accuracy got reduced.

d) **Result and Analysis –** For multiclass classification got an accuracy of 53% and for binary classification got an accuracy of 84%. This model gave better accuracy compared to Logistic regression and Linear regression. Tried with and without removing outliers. After removing outliers' the accuracy and performance of the model improved significantly. Got an F1 score of 66% for 0's and 90% for 1's. This shows a great difference compared to the other two models. Following are the classification report and confusion matrix of results obtained from the binary classification using the result obtained from the investment rating classification.

e) **Suitability of Approach** – This method seems to be more suitable compared to the other model. But memory and time required for training the model will be more compared to the other two. Overall, this method gives the best accuracy and performance.

Classification Report Investment Grade:

	precision	recall	f1-score	support
0	0.72	0.61	0.66	84
1	0.88	0.92	0.90	256
accuracy			0.84	340
macro avg	0.80	0.76	0.78	340
weighted avg	0.84	0.84	0.84	340

Classification Report (Invesment Rating Classification):

	precision	recall	f1-score	support
A1	0.46	0.50	0.48	24
A2	0.42	0.33	0.37	33
A3	0.33	0.20	0.25	5
Aa2	0.67	0.40	0.50	15
Aa3	0.65	0.74	0.69	42
B1	0.67	0.32	0.43	19
B2	0.50	0.60	0.55	5
B3	0.57	0.57	0.57	7
Ba1	0.50	0.17	0.25	6
Ba2	0.62	0.70	0.65	23
Ba3	0.48	0.43	0.45	23
Baa1	0.34	0.48	0.40	31
Baa2	0.60	0.56	0.58	63
Baa3	0.53	0.67	0.59	43
Caa1	0.00	0.00	0.00	1
accuracy			0.53	340
macro avg	0.49	0.44	0.45	340
weighted avg	0.54	0.53	0.52	340

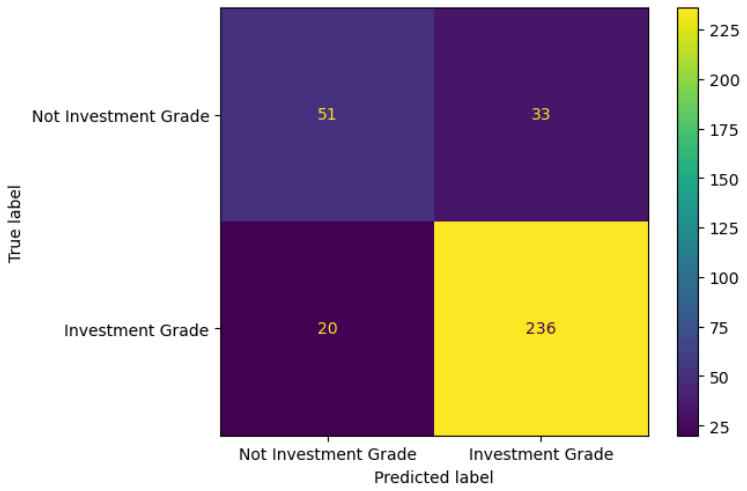

Figure.4 Confusion matrix Investment Grade classification

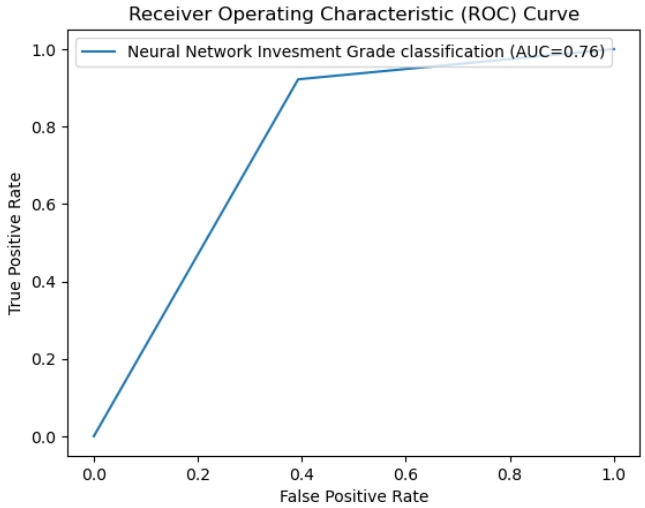

Figure.5 ROC curve Investment Grade classification