Osemekhian Ehilen
Intro to Data Mining – DATS 6103
12/6/2021

INDIVIDUAL FINAL REPORT – GROUP 4

INTRODUCTION

The problem of predicting the chance of admission for students from a dataset by UCLA Graduates from kaggle, brought about this project. The aim is a regression problem thereby allowing us to dive into multiple machine learning regression solutions.

My workload for this project was broken into four:

- Data Visualization
- Data preprocessing
- Machine learning modeling
- GUI (pyqt5)

WORK DESCRIPTION

Among the above list, I was involved in GUI (pyqt5) and partly machine learning modeling also cross validation.

CROSS VALIDATION

Cross validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set (Allen, David 1974).

I cross-validated the data set with several machine learning models to see which model gives the best result or accuracy. The models used for testing are: Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Ada Boosting Regression, Extra Tree Regression, K-Neighbors Regression, Support Vector Regression.

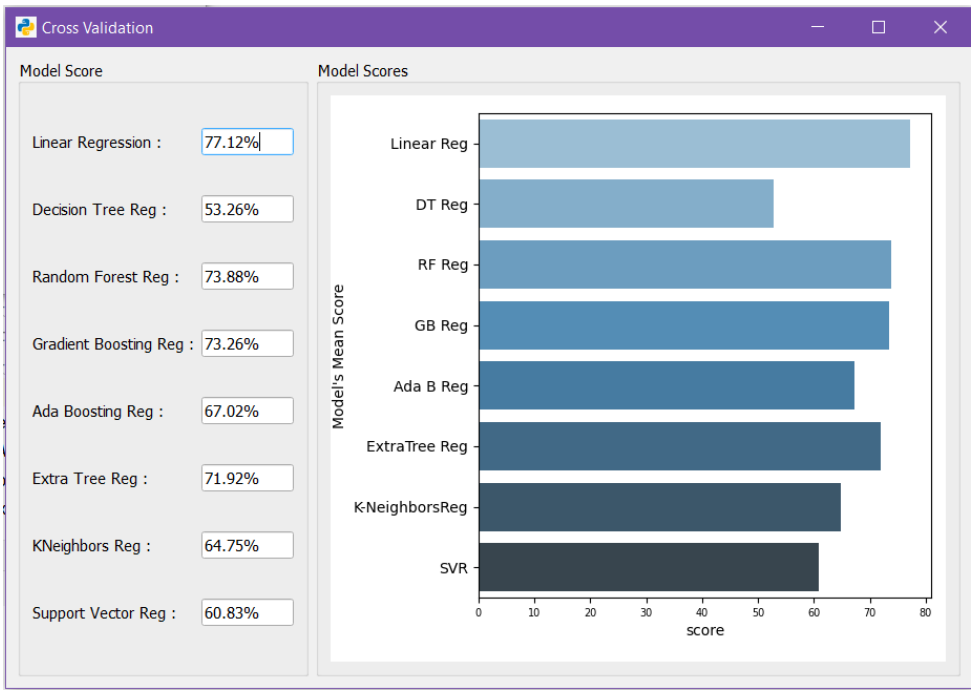

Fig 1

From the figure above, Linear Regression and Random Forest Regression gives the best accuracy score.

LINEAR REGRESSION

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). (Wikipedia)

The linear regression model is given below:

$$y_i = \beta_o + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \epsilon_i, \quad i = 1, \ldots, n.$$

Where, y is the dependent variable, β_o is the intercept, β_n are the coefficients, ϵ_i is the error term.

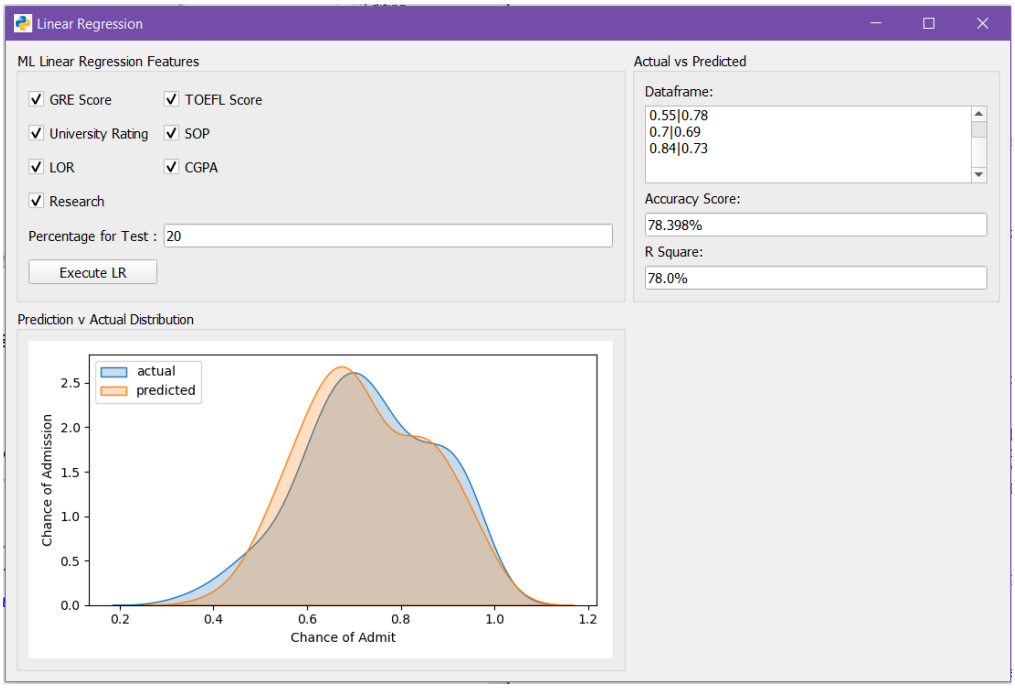

Fig 2

The Linear Regression model was trained with 80% of the data set and tested with 20% of validation/test data split to get an approximate accuracy score of 78%.

The density chart in the above canvas section (down left) shows the distribution of the predicted and actual dependent/target variable.

VOTING REGRESSOR

Getting more than one claim and taking the most claim gives a more confident decision. The Voting Regressor is an ensemble technique that takes the average prediction of multiple regression models and creates a final prediction. But using this did not give a significant difference to what Random Forest and Linear Regression gave as predictions.

PYQT5

The foundation of our pyqt5 was gotten from Professor Amir Jafari (link).

Building on that I made use of the seaborn library for most of the charts. A few fancy layout was also made to display our modeling result.

SUMMARY AND CONCLUSION

In summary, linear regression amongst other regression models performed best with an accuracy of 78%. The R squared also shows that about 78% of the data set was explained by the model.

Linear regression can effectively predict the chance of admission from the data set and possibly new data.

I would suggest for any future research that a ridge regression be experimented to see if the accuracy score would increase as well as a minimal error obtained due to the tuning parameter of ridge model.

REFERENCES

Allen, David M (1974). "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction". *Technometrics*. **16** (1): 125–127. doi:10.2307/1267500. JSTOR 1267500.

https://en.wikipedia.org/wiki/Linear_regression