International Journal of Trend in Research and Development, Volume 7(4), ISSN: 2394-9333
www.ijtrd.com

Supervised Machine Learning Modelling & Analysis for Graduate Admission Prediction

Prediction Using Supervised Machine Learning & Exploratory Data Analysis For Graduate Admission In The United States

Sujay S
Department of Electronics & Communication Engineering, Velammal Engineering College, Anna University, Chennai, India

Abstract- Predictive modelling has found its place in this century for providing an in-depth view and in helping humans in their day to day activity. In this paper, I have analyzed and predicted the possibility of a person getting an admit for graduate courses in the United States based on a supervised machine learning algorithm using Python and its various libraries on a Kaggle dataset. After implementing immense research on the dataset, explored the relationship between each factor which contribute in one or the other way to get an admit. Finally, using linear regression, allowed the program to predict the data from the user.

Keywords *- Machine Learning, Linear Regression, Predictive Modelling, Exploratory Data Analysis*

I. INTRODUCTION

The idea is to provide a prediction of a person getting an admit in the US for graduate courses. This can be done by implementing the Linear Regression which is one of the famous statistical methods in linear algebra. The dataset used contains labelled data. So, I have used a supervised machine learning algorithm which is typically used for predicting labelled data. The model trains on the data in the dataset and then predicts the data from the user. I will briefly cover the process in the forthcoming topics.

The dataset contains 400 rows of data and seven columns which are considered for the application for Masters Programs. The factors that contribute to graduate admission are:

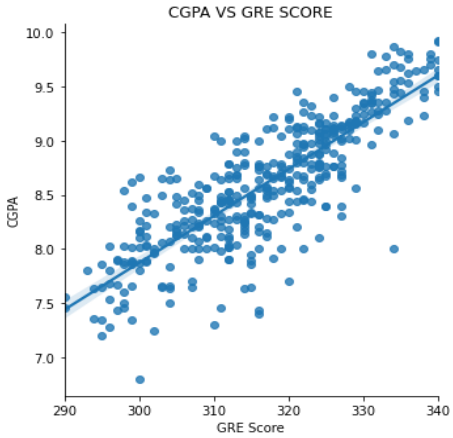
1. GRE score (out of 340)
2. TOEFL score (of 120)
3. Undergraduate University Rating (out of 5)
4. Statement of Purpose and Letter of Recommendation Strength (out of 5)
5. Undergraduate GPA (out of 10)
6. Research Experience (either 0 or 1)
7. Admit Possibility (ranging from 0 to 1)

II. EXPLORATORY DATA ANALYSIS

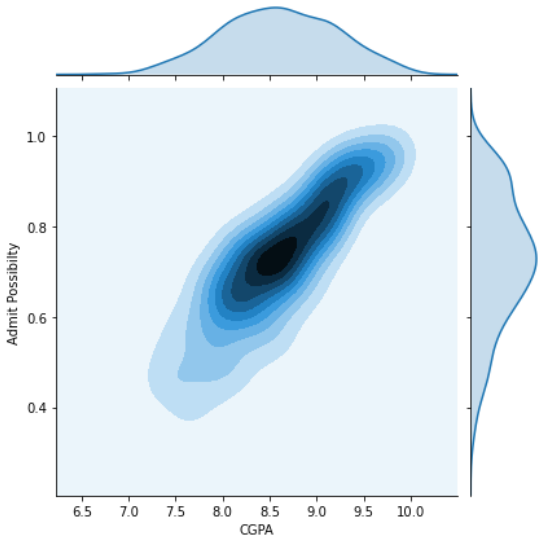
Exploratory Data Analysis unveils itself from the late 1970s developed by John Tukey to display data in such a way that interesting features will be uncovered. Exploratory Data Analysis techniques are used to encourage the data to suggest models that might be appropriate. In this case, it is necessary to choose the right regression algorithm for this data.

After importing the necessary python libraries and finding that there were no null values in the dataset, I have first compared the relationship between the undergraduate Cumulative Grade Point Average and the GRE score.

It is easy to conclude that the candidate's GRE score is in high relationship with his Cumulative Grade Point Average, meaning, the GRE score 's points linearly increases with the Cumulative Grade Point Average.

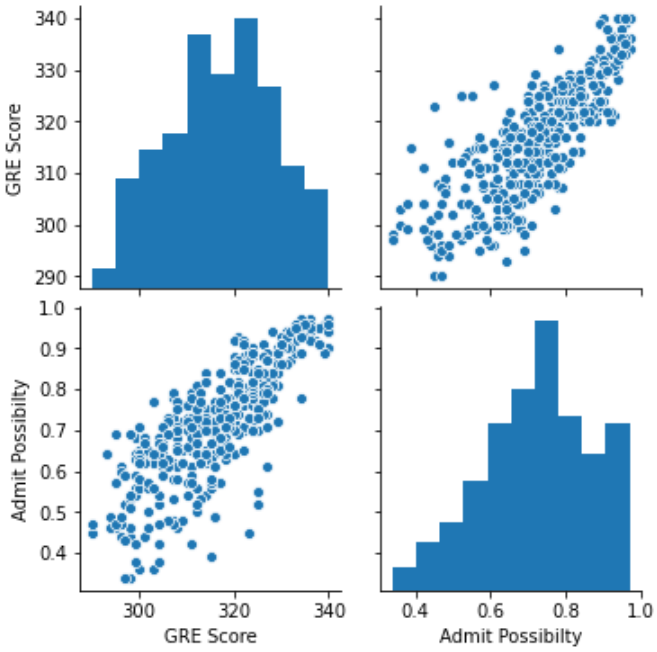


On further proceeding by comparing the Admit Possibility and the Cumulative Grade Point Average using a joint plot, the forthcoming conclusions are made.

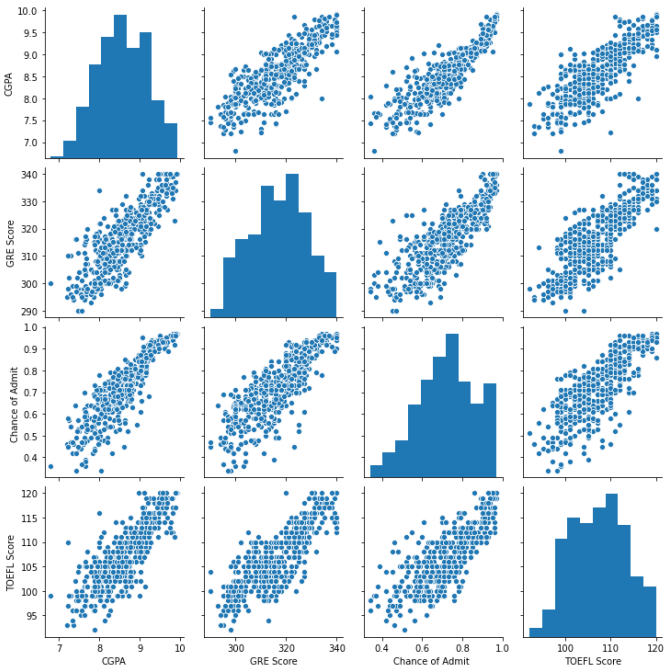


The admit possibility is highly dependent on the Cumulative Grade point average. The chances increase with an increase in the Cumulative Grade Point Average.

Before choosing and fitting the right model for prediction, I have done one final analysis establishing the relationship between admitting the possibility and GRE score using a pair plot.



Now it is evident that the GRE score and the admit possibility are also in high correlation with each other. The pair plot gives a briefer outlook on the correlation.



III. Model Selection and Fitting

As the name suggests, linear regression follows the linear mathematical model for determining the value of one dependent variable from the value of one given independent variable.

To fit a linear regression model, we select those features which have a high correlation with our target variable Admit Possibility. By looking at the Exploratory Data Analysis we can see that the features being GRE scores, Cumulative Grade Point Average, TOEFL scores are in high positive correlation with the target being Admit Possibility.

Import the necessary machine learning libraries and then split the data as x and y where x contains the features except the target and y contains only the target.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
```

```
x=Reading.drop('Admit Possibilty',axis='columns')
y=Reading['Admit Possibilty']
x_train,x_test,y_train,y_test=train_test_split(x, y)
```

Further splitting the data as test and train where the train set contains 80% of data and the test set contains 20% of the data.

A test set should still be held out for final evaluation, but the validation set is no longer needed when doing Cross-Validation. In the basic approach, called k-fold Cross-Validation, the training set is split into k smaller sets.

```
def get_cv_scores(linear_regression):
    scores = cross_val_score(linear_regression,
                             x_train,
                             y_train,
                             cv=5,
                             scoring='r2')

    print('CV Mean: ', np.mean(scores))
    print('STD: ', np.std(scores))
    print('\n')

# get cross val scores
get_cv_scores(linear_regression)
```

```
CV Mean:  0.7802853753428609
STD:  0.04026867116443402
```

```
model = LinearRegression(normalize=True)
model.fit(x_test, y_test)
model.score(x_test, y_test)
```

```
0.8411925676686253
```

The Cross-Validation score says that the model is neither an underfit nor an overfit. Any value between 0 and 1 is good. The Standard Deviation says that the model is not being memorised by the machine.

The model is then subjected to the test set for predicting accuracy. The model is having 84.1% accuracy on the test set.

IV. MODEL PREDICTION

The model finally predicts the data based on user input. This model can help people now and further research can be carried out after it has been deployed as a web application.

```
print('The chance of you getting an admit in the US is {}%'.format(round(model.predict([[3
05, 108, 4, 4.5, 4.5, 8.35, 0]])[0]*100, 1)))
```

```
The chance of you getting an admit in the US is 66.3%
```

V. SOFTWARE USED

Jupyter Notebook: Jupyter Notebook is an excellent open-source web application that allows you to create and share documents that contain live code, equations, visualizations and used for data cleaning and transformation, numerical simulation, statistical modelling, data visualization and machine learning

Python Libraries for Exploratory Data Analysis: NumPy, Pandas, Matplotlib and Seaborn.

Python Libraries for machine learning: Sklearn is an excellent source for implementing machine learning algorithms.

CONCLUSION

Thus with the help of Supervised Machine Learning and Exploratory Data Analysis, the prediction of the possibility of a candidate getting an admit has been successfully implemented.

References

[1] Data set: Mohan S Acharya, Asfia Armaan, Aneeta S Antony: https://www.kaggle.com/mohansacharya/graduate-admissions

[2] Animesh: https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155

[3] Sujay S (My source code): https://www.kaggle.com/sujay12345