PySpark Project: Graduate Admissions

Mike Christison

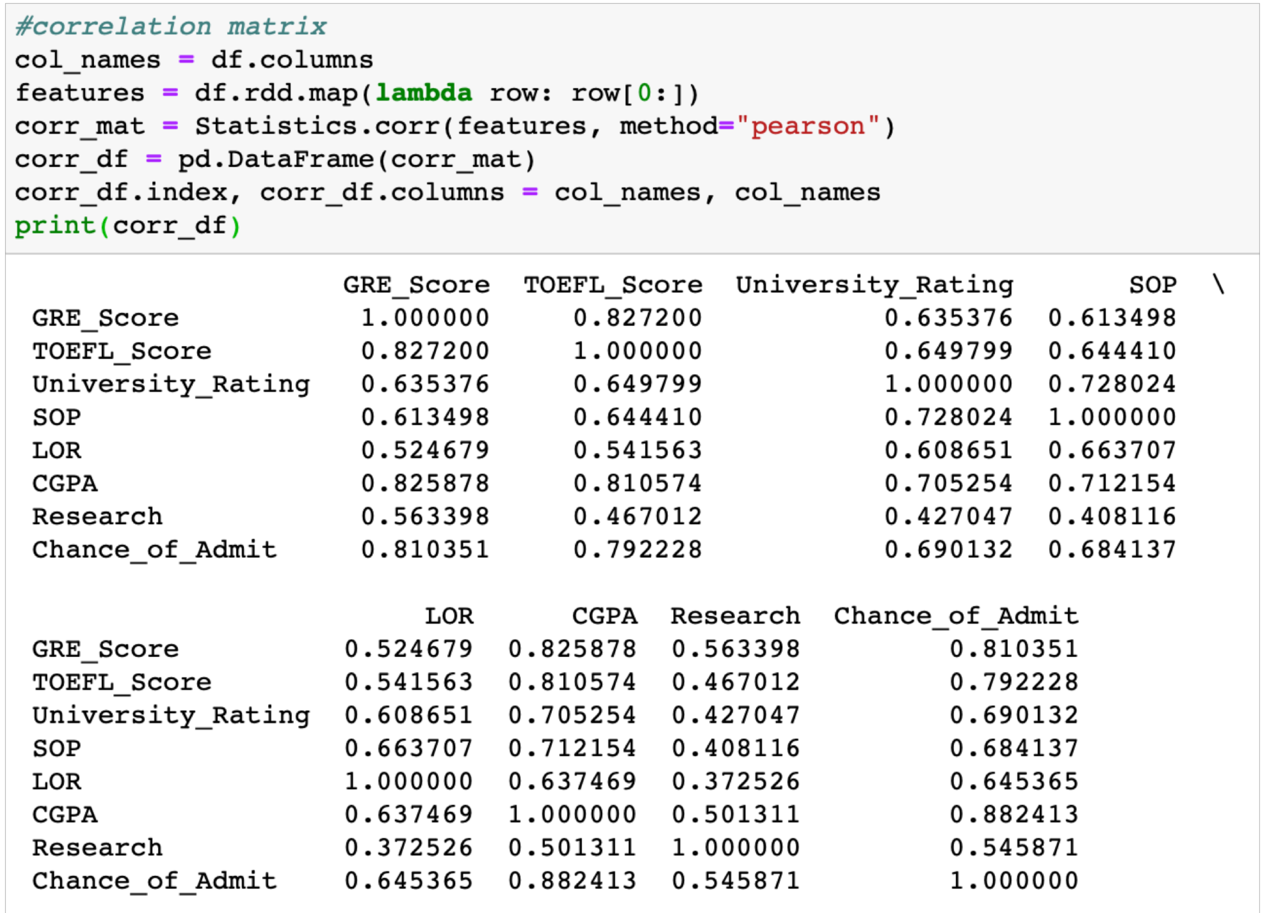
The objective of this project was to examine the results of an existing predictive model and use regression to extract insights on feature importance using linear regression.

The source of the data was from Kaggle, provided by Mohan S Acharya. The goal of the original model was related to predicting graduate admissions to give prospective students a sense of understanding their chances of getting into graduate school.

https://www.kaggle.com/mohansacharya/graduate-admissions/home

Analysis: Correlation

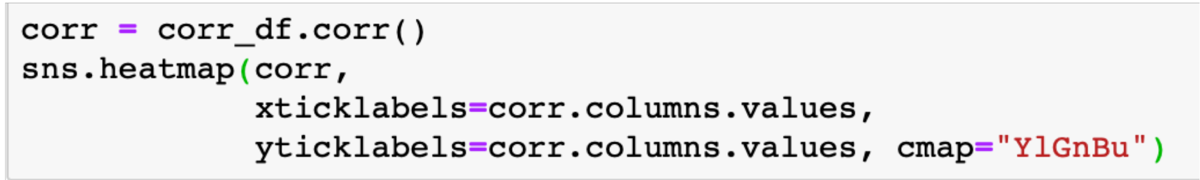
A Pearson correlation matrix revealed some sense of potential interactions between feature variables. GRE, TOEFL, and CGPA were all fairly strongly correlated with one another. That was in addition to being strongly correlated with the y variable, “Chance_of_Admit.”



```
#correlation matrix
col_names = df.columns
features = df.rdd.map(lambda row: row[0:])
corr_mat = Statistics.corr(features, method="pearson")
corr_df = pd.DataFrame(corr_mat)
corr_df.index, corr_df.columns = col_names, col_names
print(corr_df)
```

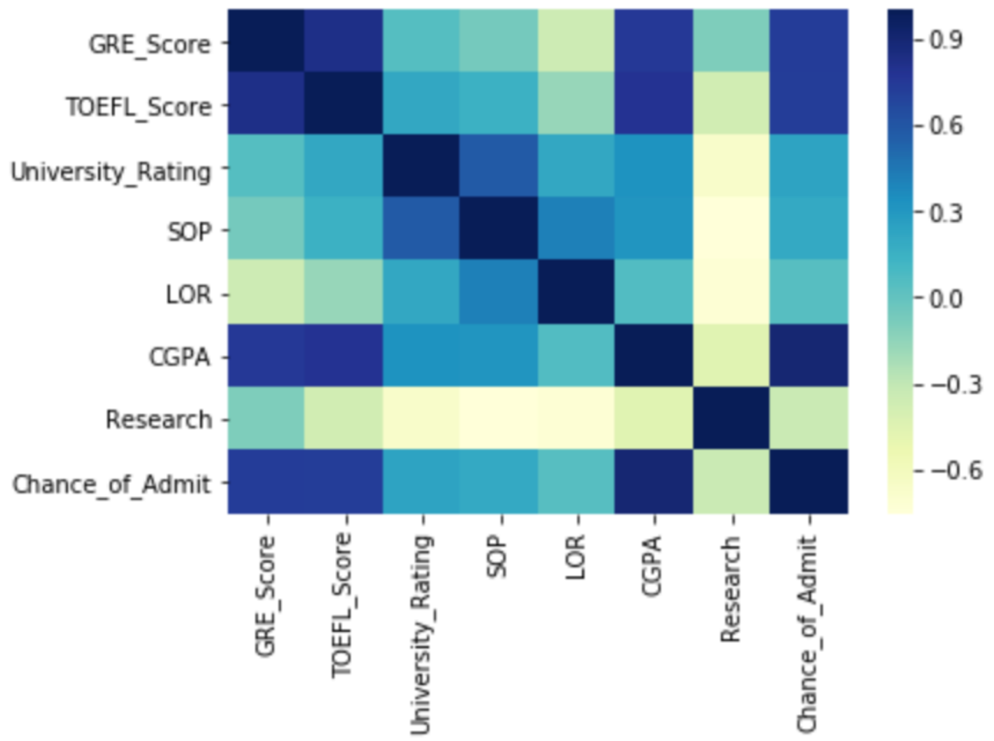
```
                   GRE_Score  TOEFL_Score  University_Rating       SOP  \
GRE_Score           1.000000     0.827200           0.635376  0.613498
TOEFL_Score         0.827200     1.000000           0.649799  0.644410
University_Rating   0.635376     0.649799           1.000000  0.728024
SOP                 0.613498     0.644410           0.728024  1.000000
LOR                 0.524679     0.541563           0.608651  0.663707
CGPA                0.825878     0.810574           0.705254  0.712154
Research            0.563398     0.467012           0.427047  0.408116
Chance_of_Admit     0.810351     0.792228           0.690132  0.684137

                        LOR      CGPA  Research  Chance_of_Admit
GRE_Score          0.524679  0.825878  0.563398         0.810351
TOEFL_Score        0.541563  0.810574  0.467012         0.792228
University_Rating  0.608651  0.705254  0.427047         0.690132
SOP                0.663707  0.712154  0.408116         0.684137
LOR                1.000000  0.637469  0.372526         0.645365
CGPA               0.637469  1.000000  0.501311         0.882413
Research           0.372526  0.501311  1.000000         0.545871
Chance_of_Admit    0.645365  0.882413  0.545871         1.000000
```



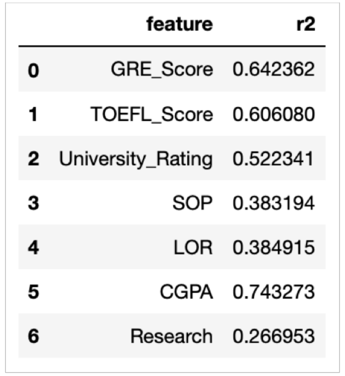
```
corr = corr_df.corr()
sns.heatmap(corr,
            xticklabels=corr.columns.values,
            yticklabels=corr.columns.values, cmap="YlGnBu")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f61830f6c88>
```

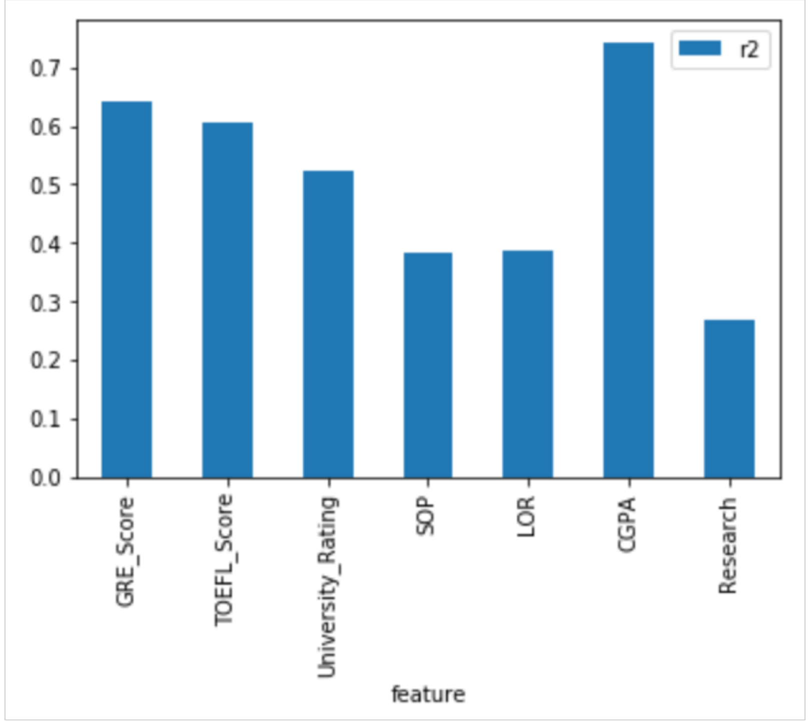


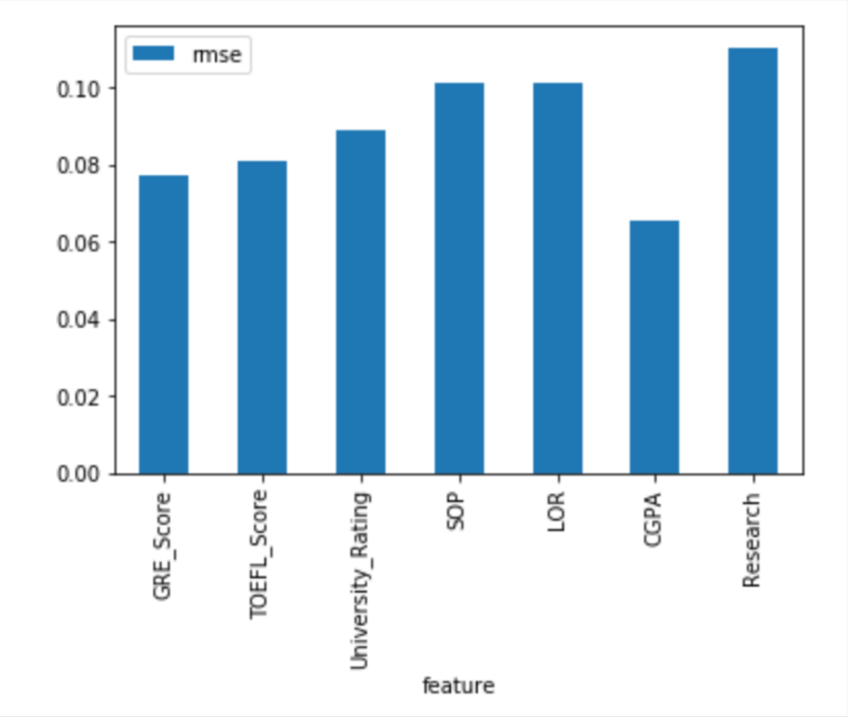
Analysis: R2 and RMSE

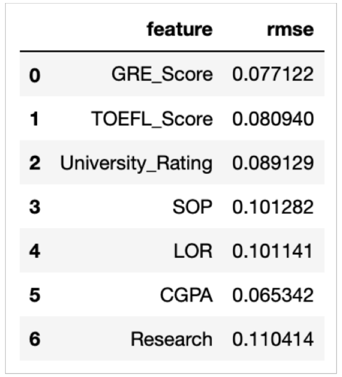
Here are the results of running regressions against the y-variable on each feature individually. R-Squared and RMSE showed fairly similar results to the correlation. Overall, CPGA seems to have the strongest impact on the predictive model, with the highest R-Squared and lowest error, with GRE and TOEFL having fairly strong impacts as well. While all features seem to show some level of importance, apparently it pays to get good grades!



	feature	r2
0	GRE_Score	0.642362
1	TOEFL_Score	0.606080
2	University_Rating	0.522341
3	SOP	0.383194
4	LOR	0.384915
5	CGPA	0.743273
6	Research	0.266953







	feature	rmse
0	GRE_Score	0.077122
1	TOEFL_Score	0.080940
2	University_Rating	0.089129
3	SOP	0.101282
4	LOR	0.101141
5	CGPA	0.065342
6	Research	0.110414