MA515 Project Report

by Tanmay Aeron(2019CSB1124)

Submitted in Partial Fulfillment for Course on

Foundations of Data Science (MA515)

~ Submitted to Dr Arun Kumar

Department of Mathematics

Indian Institute of Technology, Ropar

Ropar – 140001

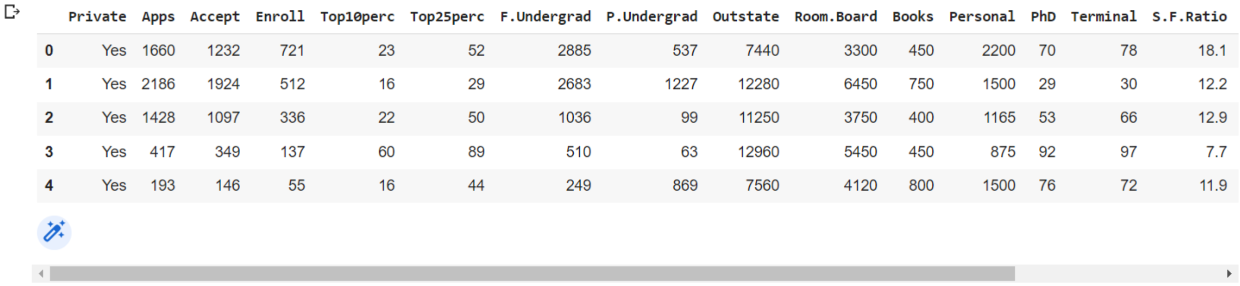
2022

Problem Statement:

Do exploratory data analysis on the data. Using the multilinear regression predicts the number of applications accepted. Further use ridge regression technique.
Dataset - college dataset

Data Description

Data Link - https://www.kaggle.com/datasets/faressayah/college-data



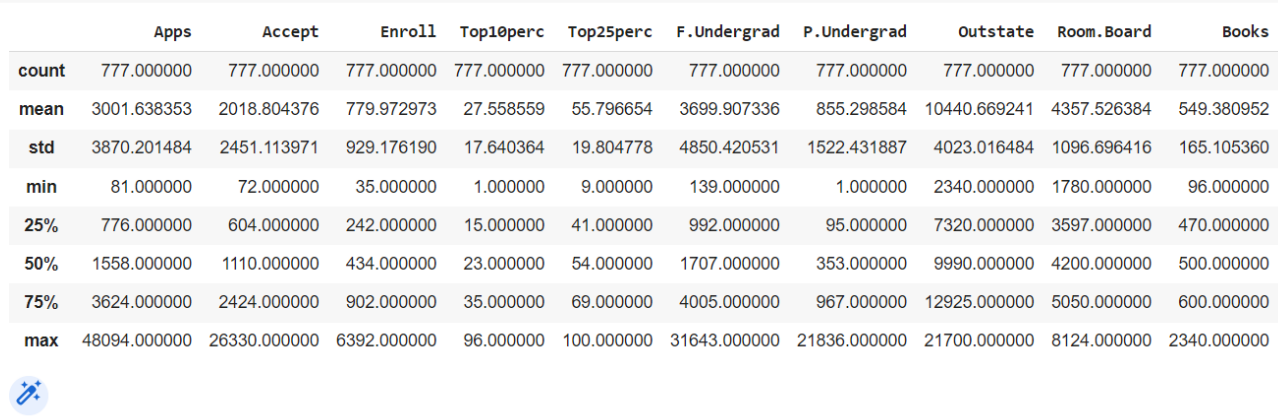
	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
0	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1
1	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2
2	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9
3	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7
4	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9

Fig. 1 - Some observations of dataset

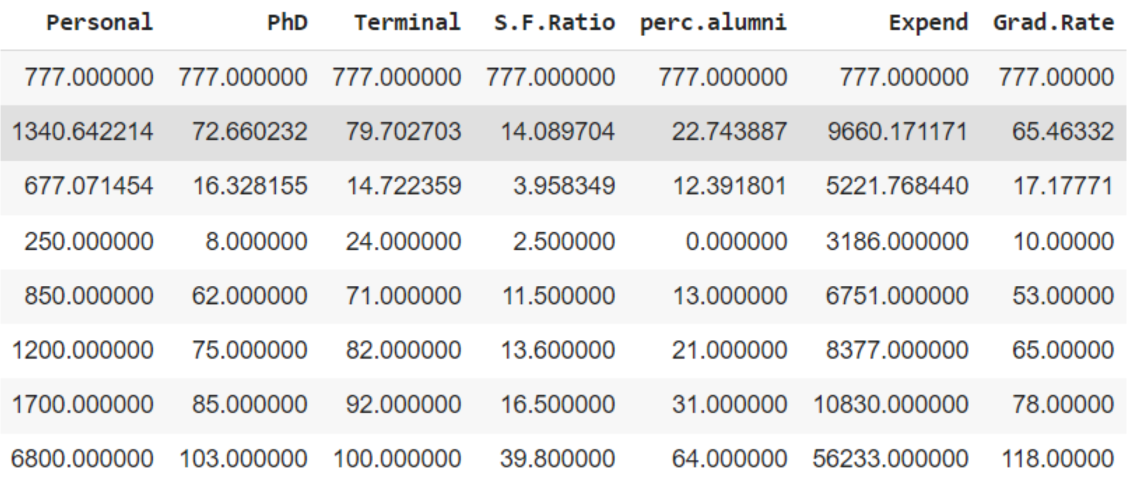
- Dataset has 777 rows and 18 columns

EDA

- Count, mean, min, standard deviation, 25%, 50% and 75% quartile have been calculated
- Private is a categorical variable, hence these calculations have not been done for that variable.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000



Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

Fig. 2 - Various statistical measures on different columns

Correlation matrix

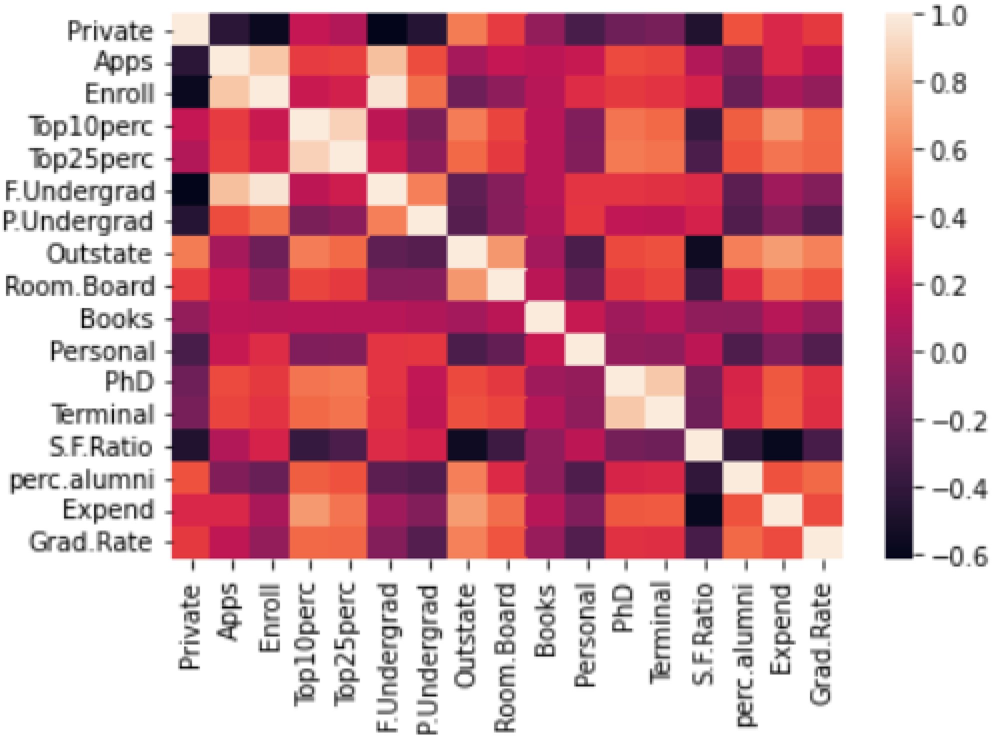


Fig. 3 - Correlation between different features of dataset

We can see that a large number of variables like Enroll and Apps, Top 25 percent and Top 10 percent have high correlation. We need to drop some variables. Apart from that, some variables may not be related to the number of applications accepted.

Histogram of Features

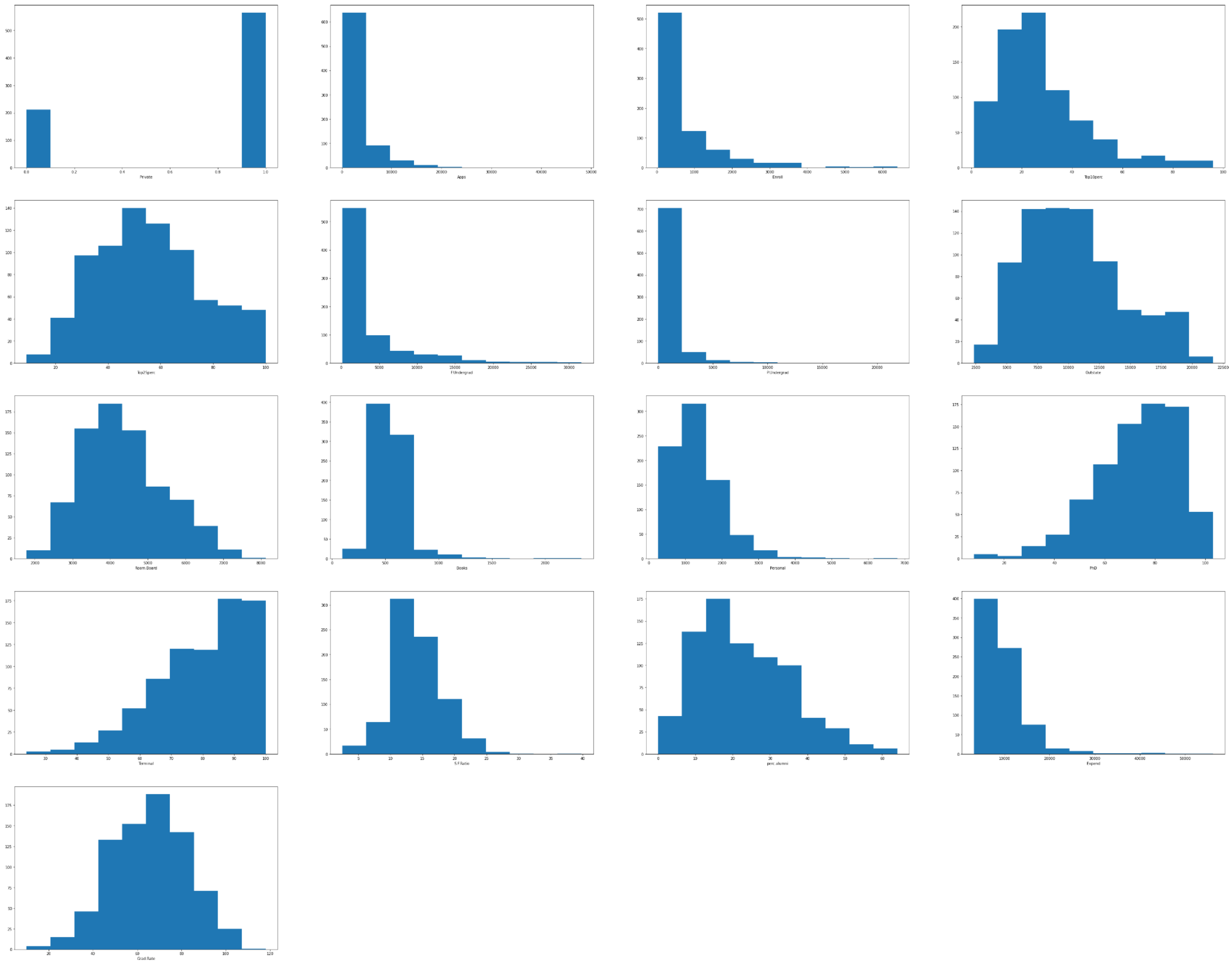


Fig. 4 - Histogram of all features of dataset

Data-preprocessing

Label Encoding

In the dataset there is a categorical variable ‘Private’ which takes two values Yes and No.Now to train the model on the dataset, it should only consist of numerical values. Hence, we encode them as 0 - ‘No’ and 1-’Yes’ with the help of sklearn.LabelEncoder.

Scaling of Features

Since Ridge Regression requires features to be centered, I scaled features to have mean 0 and standard deviation 1.

Test Training Split

Dataset was split into training and test data in 75:25 ratio. Training data contains 582 observations while test dataset contains 195 observations.

Model Results and Interpretation

MultiLinear Regression

Since there was high correlation between several variables, we needed to drop some variables. I decided to use the forward selection method to select my model. I kept the adjusted coefficient of determination threshold on the training dataset to be 0.95. Trained model included only four features :

1. Apps
2. Enroll
3. Top10perc
4. Outstate

Intercept - 2028.2577319587629
Coefficients on scaled features - [1681.83686391, 964.32427461, -382.18327051, 198.18932927]

Mean Square Error on Training Dataset - 303442.5901404387
Mean Square Error on Testing Dataset - 332967.6259252047

R square value on Testing Dataset - 0.9346767732742691
Adjusted R square value on Testing Dataset - 0.9333015474484642

Ridge Regression

Ridge Regression method is used when data has very high correlation. Therefore, I didn't drop any column for ridge regression.
I varied the value of lambda from 0 to 100 increasing it 0.1 every time.

Then I selected the lambda for which mean square error on test data was minimum.

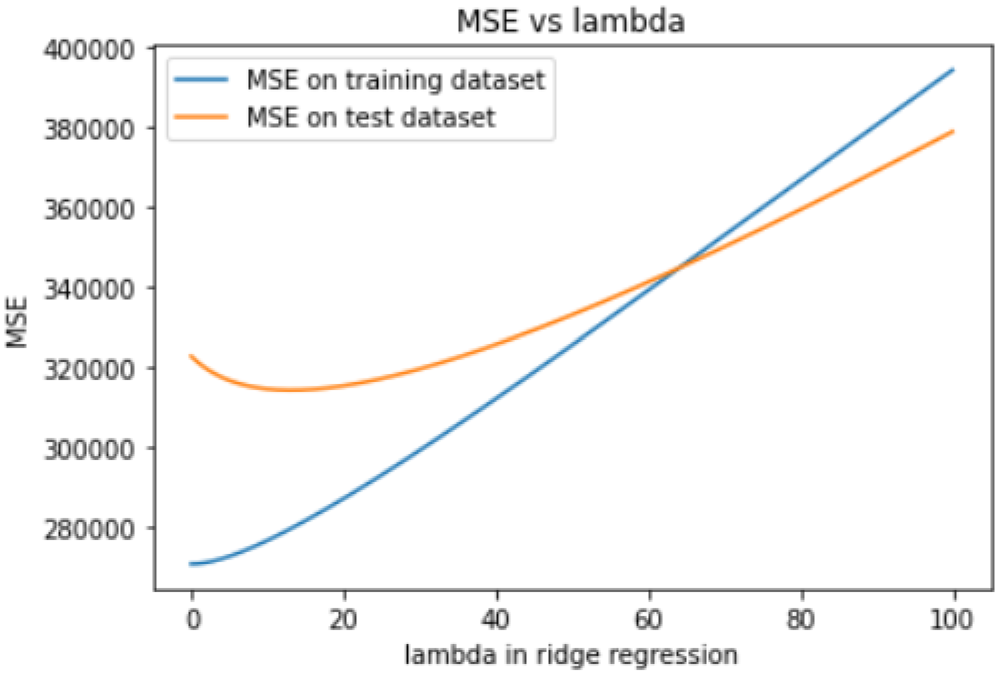


Fig. 5 - Variation of mean square error on training and test data with lambda

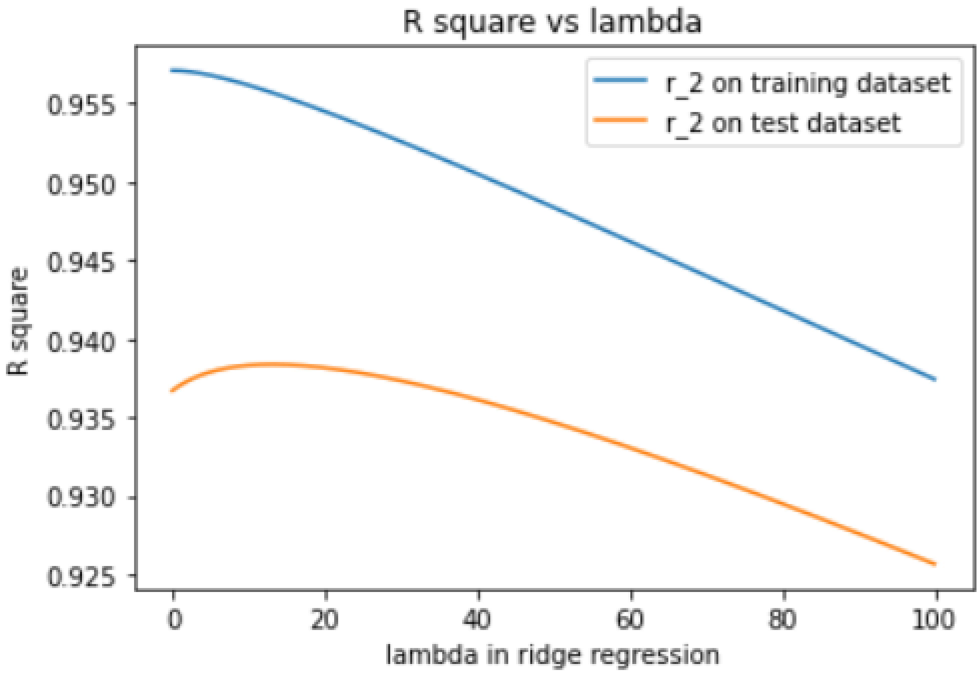


Fig. 5 - Variation of coefficient of determination on training and test data with lambda

Least mean square error on the test dataset was coming for lambda = 13.1.

For lambda = 13.1

Mean Square Error on Training Dataset - 279450.1173674223
Mean Square Error on Testing Dataset - 314168.6692833608
R square value on Training Dataset - 0.9556598810203408
R square value on Testing Dataset - 0.9383648450605581

Comparison of Linear Regression and Ridge Regression

Metric	Linear Regression	Ridge Regression
MSE on training	303442.5901404387	279450.1173674223
MSE on test	332967.6259252047	314168.6692833608
R square on test	0.9346767732742691	0.9383648450605581

Conclusion

From the above analysis we conclude that as data was highly correlated, we needed to drop some columns for multilinear regression or we can use ridge regression on complete dataset as well. For both models we get the value of coefficient of determination close to 0.94.