CS 556 – Mathematical Foundations of Machine Learning

Group 24 - College Admissions Predictor

(a) CWID / Student Name:

1. 20011413 / Abraar Mohammed Malik Subhan
2. 20012507 / Kaarthik Senthil Kumar
3. 20015520 / Vrund Patel

(b) LINEAR REGRESSION – Mean Square Error

Training Dataset	Test Dataset
0.004371959793442482	0.003008975420241682

PCA with LINEAR REGRESSION – Mean Square Error

Training Dataset	Test Dataset
0.004588610261602332	0.004674900094626036

(c) SALIENT PROJECT FEATURES:

- The college admission predictor requires us to predict the chances of admission, which is a dependent variable based on independent variables such as GRE, TOEFL, CGPA etc., so we have chosen the **Linear Regression Model**
- Data visualisation techniques such as scatter plot, histogram, bar chart and pie chart are used to analyse the features individually and their influence on the chances of admission
- A correlation heatmap is used to illustrate the dependency strength between all the dataset features as this helps us understand the influence of features on each other
- After visualisation and analysis, we begin processing of data for linear regression by dropping the feature 'Chance of Admit' as that is considered as the target variable
- The data was split into training and test dataset (80% and 20% respectively) and scaled using StandardScaler to achieve standardised values for the features
- Linear regression models were developed and trained on different datasets comprising of a variety of feature combinations and the respective Mean Square Error (MSE) was calculated for each model's training and test datasets, in order to learn the features that lower MSE
- Lowest MSE was obtained when the dataset for the linear regression model did not include the "Research" feature. Now, the dimensionality of the dataset was reduced to '2' and transformation was performed using the PCA model. Post which, the data was split into training, testing dataset and PCA along with Linear Regression was performed to compare the mean square with basic linear regression. PCA with Linear Regression was performed on other datasets which achieved lower MSE in order to analyse the variance.
- Finally, a scatter plot is used to visualise the results of both, Linear Regression model and PCA with Linear Regression model. In conclusion, it is understood that Linear Regression by removing 'Research' feature is the best performing model in comparison with PCA with Linear Regression model after dimensionality reduction on the same dataset.