**1.Introduction**

This report presents the analysis and modelling of a dataset for a regression task. The goal is to build a linear regression model to predict a target variable based on a set of features The report provides an overview of the dataset, exploratory data analysis, feature selection methods, linear regression models with different feature sets, and an analysis of the results.

**2.Dataset Analysis**

Initially, our dataset had 1889 rows and 18 columns. We dropped unrequired columns that I didn’t find suitable. Later on, we performed univariate and multivariate analysis to observe the relationships between different several features The final dataframe used for modeling has a shape of (N, M), where N represents the number of samples and M represents the number of selected features.

**3.EDA**

**Univariate Analysis**

A picture containing diagram, plot, line

Description automatically generated A picture containing text, screenshot, diagram, line

Description automatically generated

**A picture containing screenshot, text, diagram, line

Description automatically generated**

We performed univariate analysis on some of the features. Here, we can observed that most of Most of the Number of rating are at the range of 0 - 30. Number of ratings where high around 15 and the density value was between 0.08-0.10.

**Multivariate Analysis**

We observed a relationship between cbrt\_NumberRatings and cbrt\_Student Both are directly propotional to each other. As the Number Ratings increase the demand of student also increase .

A picture containing text, screenshot, map, font

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generated

From the second plot, we can see that as Number Ratings increase Enrollment increases.

**4.Feature Observation and hypothesis**

To find the most pertinent characteristics for the regression models, four feature selection strategies are used.

1. **Correlation -Based Selection:** subset of characteristics is manually chosen in accordance with their association with the target variable using the correlation matrix as a guide. Included in this list are the characteristics "NumberRatings," "cbrt\_Enrollment," "Enrollment," "cbrt\_Student," "InstRating," "Student," "Fee," and "Duration."

A picture containing text, screenshot, number, font

Description automatically generated

1. **Variance Threshold Selection**: Variance threshold selection is used to assess the characteristics. Columns that are not numeric are removed, and a variance threshold of 2.5 is used. The chosen features are included in the final DataFrame, df\_vt.

A picture containing text, screenshot, font, number

Description automatically generated

1. **SelectKBest Method :** best K features are selected using the SelectKBest technique, which is based on univariate statistical tests. The train\_test\_split() method divides the feature sets into training and testing sets, and the f\_regression score function creates the SelectKBest instance. The extracted characteristics are put in a DataFrame called df\_selKBest after being extracted.

A screenshot of a graph

Description automatically generated with low confidence

**5. Simple Linear Regression Report**

In this section, several feature selection techniques were used to build linear regression models. The following strategies for feature selection were used:

Selection based on correlation: The characteristics having the highest absolute correlation values with the cbrt\_NumberRatings target variable were chosen. 'NumberRatings', 'cbrt\_Enrollment', 'Enrollment', 'cbrt\_Student', 'InstRating', 'Student', 'Fee', and 'Duration' were the features that were chosen.Variance Features having a variance greater than a threshold value (2.5) were chosen. ChooseKBest approach: Based on the F-regression score, the top 7 characteristics were chosen.We trained linear regression models for each feature selection technique. The R2 score and root mean squared error (RMSE) were used to assess each model's performance. There was also a display of the coefficients for the chosen characteristics.

**6.Linear Regression with Lasso Report:** Lasso regression was applied to the dataset using different alpha values. The performance of the models was evaluated using the R2 score and RMSE. The best alpha value was selected based on the highest R2 score and lowest RMSE.

A screenshot of a computer

Description automatically generated with medium confidence

**7.Analysis**

Below is a table summarising the outcomes of several feature engineering techniques and the corresponding performance metrics:

A picture containing text, screenshot, font, number

Description automatically generated

The SelectKBest approach earned the greatest R2 score and lowest RMSE value, which indicates higher performance when compared to other methods, as can be seen in the table.