Linear Regression Model to Predict “inst review”

1. **Introduction:**

The goal of this project is to develop a linear model for predicting the "inst\_review" based on different features. The project involves exploring various combinations of data preprocessing techniques, including standard scaling and second-degree polynomial transformations. Different data frames will be created using feature selection methods such as KBest, Variance Threshold, and Correlation.

These data frames will undergo the preprocessing techniques mentioned earlier to find the best combination that yields the most accurate predictions for the instruction review. The ultimate objective is to identify the optimal model that can effectively predict the instruction review based on the selected features and preprocessing methods.

1. **Dataset Analysis:**

The dataset used in this analysis is called "instruction review" and it is stored in a CSV file. It contains various attributes related to an educational institution. The top rows of the dataset can be viewed in the figure below.A screenshot of a computer

Description automatically generated with low confidence

Figure 1. df.head()) for ‘Lab02\_prepared.csv

The data in the table underwent a cleaning process to ensure the suitability of the data for building a linear regression model. The column information was inspected using the .info() function, which provided an overview of the columns and their data types. Based on the results, several data manipulation techniques were applied to prepare the data for analysis.

A screenshot of a computer program

Description automatically generated with medium confidence

The measures implemented include:

* Changing column names
* Checking for nulls and drop or fill them
* Applying lower case to every value of columns and replacing whitespaces to underscores
* Split values from text and make them numerical
* Reduce the number of unique values in the category and generate dummy\_features

Finally, after extensive cleaning the final data set that was obtained (figure 2) is presented in the following figure which have 1904 rows and 18

A screenshot of a computer code

Description automatically generated with low confidence

Figure 2. Final data frame ‘Lab02\_prepared.csv

1. **EDA**

The correlation between features is shown in the following heat map plot (figure 3). The features which log1p() is applied to transform, seems to have high correlation and so we should be careful about multicollinearity

A screenshot of a graph

Description automatically generated with low confidence

Figure 3. Correlation heat map.

The following plot (figure 4) will show the normal distribution of log ins student in the dataset. Plot for log\_inst\_student shows that we have left tail. We have some outliers from 4 to 7. Most of the value is gathered between 10 and 12.

A picture containing diagram, plot, screenshot, line

Description automatically generated

Figure 4. plots to show unvariate analysis for log ins student

Our target inst\_review and ins\_student have the highest correlation with each other. Points are concentrated and form a linear. There seems to be a trend so I would say that there is a strong correlation between the two variables.

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 5. plots to show correlation among higher score features in dataset

1. **Feature Observation and Hypothesis**

Here observation with selecting target feature as "log\_inst\_review"

1. "log\_inst\_student": The high correlation of 0.94 suggests a strong positive relationship between the logarithm of institutional review scores and the logarithm of the number of students in the institution. This supports the hypothesis that larger institutions may have higher review scores.
2. "log\_enrollment": The correlation of 0.64 indicates a positive relationship between the logarithm of institutional review scores and the logarithm of the enrollment size. This aligns with the hypothesis that institutions with a larger enrollment may receive higher review scores.
3. "log\_number\_ratings": With a correlation of 0.61, there is a positive relationship between the logarithm of institutional review scores and the logarithm of the number of ratings. This suggests that institutions with more ratings tend to have higher review scores, potentially indicating greater popularity or student satisfaction.

log\_inst\_review 1.000000

log\_inst\_student 0.948666

log\_enrollment 0.646487

log\_number\_ratings 0.613965

df.corr()["log\_inst\_review"].abs().sort\_values(ascending = False)

1. **Simple Linear Regression report**

The feature selection method in the datasets were:

* **Correlation based selection** – In this analysis, the correlation between features was used to identify the most representative variables for the linear model. The selection process involved setting a threshold of correlation coefficient greater than 0.2, indicating that features with a strong positive or negative correlation to the target variable were selected. The final set of features chosen based on this criterion includes:

A screenshot of a computer

Description automatically generated with low confidence

* **Variance Threshold** – In this approach, the variance of each feature was computed, and a threshold was set to select features based on their variance. The chosen threshold for selection was 0.4, aiming to filter out features with low variance. By applying this threshold, features with relatively higher variances were retained, indicating their importance in the linear model.

A picture containing text, screenshot, font, number

Description automatically generated

* **Select K-Best** – This approach involves using a built-in model provided by scikit-learn to select the top k features in a dataframe based on their k-scores. The k-score is calculated using the f\_regression method, which is suitable for regression tasks. In this case, the number of selected features was set to 8, and the target feature for prediction was "log\_inst\_review". By applying this method, the most relevant features were identified based on their k-scores

**This model provided the best results out of all the linear models since it has the highest R2 and the lowest RMSE and thus this model was chosen as the Linear Regression Model to make the final predictions. (Select K-Best)**

A picture containing text, screenshot, number, font

Description automatically generated

1. **Linear Regression with Ridge Model**

The top results for my alpha scores are shown in the following table:

A picture containing text, screenshot, font, number

Description automatically generated

These calculations involved using all the available features in the model and followed the same ratio for training and testing as the other models. The values were chosen based on the alpha parameter that resulted in the lowest RMSE score. After experimentation, an alpha value of 0.316 was identified as the optimal choice.

1. **Analysis**

The obtained results are presented in Figure 6, which includes a summary table of different feature selection combinations and a scatter plot depicting the predicted values versus the actual values for emissions.

A picture containing text, screenshot, font, number

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

Description automatically generated

Figure 6. Summary Analysis and Predicted Price vs Actual Price scatter plot.

Findings on the linear model:

* Based on the above table, I believe the best model to be index 4**: SelKBest with Poly Transform**. Because It has the highest R2 and the lowest RMSE: SelKBest Poly 2 Interaction 0.943891 0.43795

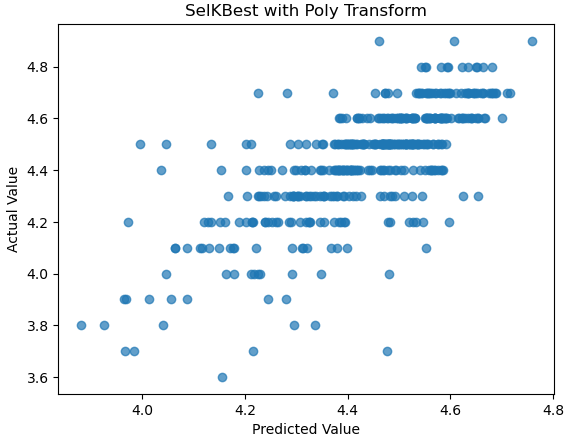
* Coefficients: The feature coefficients provide insights into the relationship between the features and the target variable. In this case, we observe that certain features have a noticeable impact. For example, higher “inst\_rating" and a larger "log\_inst\_student" are associated with higher values of the target variable. On the other hand, higher "avg\_rating" and "log\_enrollment" display negative coefficients of the target variable. These findings suggest that these specific features play a significant role in determining the target variable.

* The linear regression scatter plot seems to have a clear trend and strong correlation between the predicted values (Y\_pred) and the actual values (Y\_test)
* It would be worth trying to improve our prediction to explore and create new features that might have a stronger relationship with the target variable. This could involve combining existing features, extracting relevant information from the existing data.

**Bonus part:**

Changing target as avg\_rating

A picture containing text, screenshot, font, number

Description automatically generated

Findings about selected Ridge linear model:

* Based on the above table, I believe the best model to be **Ridge regression model** is choosen for alpha. Because It has the highest R2 and the lowest RMSE. However, even with best one, accuracy is low
* Coefficients: The feature coefficients provide insights into the relationship between the features and the target variable. In this case, I cannot easily observe any result since that values are so similar and narrow.
* The linear regression scatter plot seems that points are spread out and shows some outliers. There is a weak trend and correlation between the predicted values (Y\_pred) and the actual values (Y\_test)