**1) Title and Introduction**

*Title : Course Review – Regression Project 01*

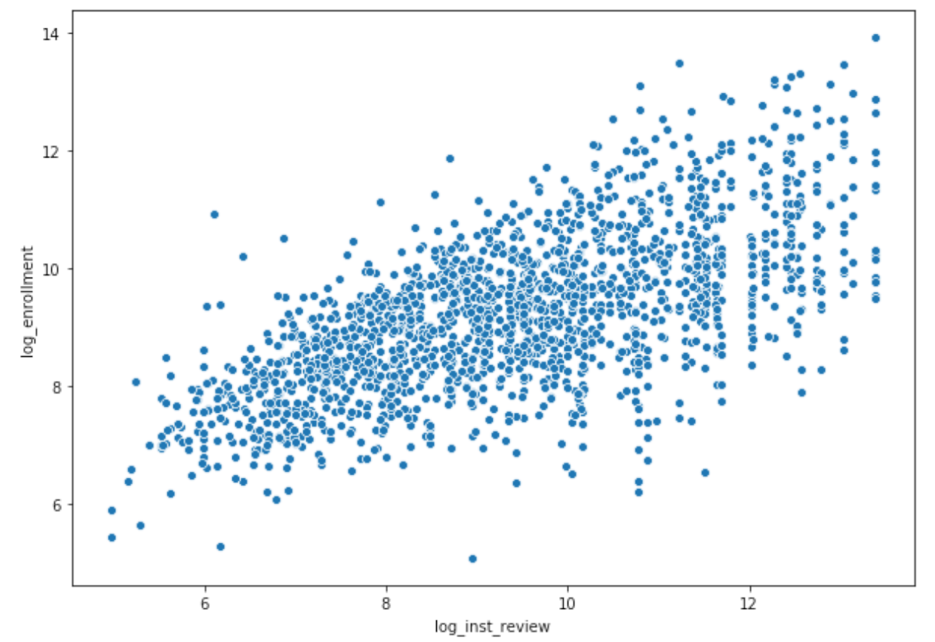
A regression model is used to predict number of reviews a course gets based on various provided factors. We peek, clean, analyze and create a model for it. We compare a number of different regression models and data scaling and data transformation techniques to best fit the regression model.

**2) Dataset Analysis**

The dataset contains 1904 rows of non-null, numerical data from various categories of courses, 11 numerical columns and 7 dummy columns created for different categories. All outliers have been dropped and the data has been log-transformed to bring it closer to a normal distribution. Different Feature selection methods have been carried out with ‘*log\_inst\_review*’ column as the target variable. The final shape of the dataset is

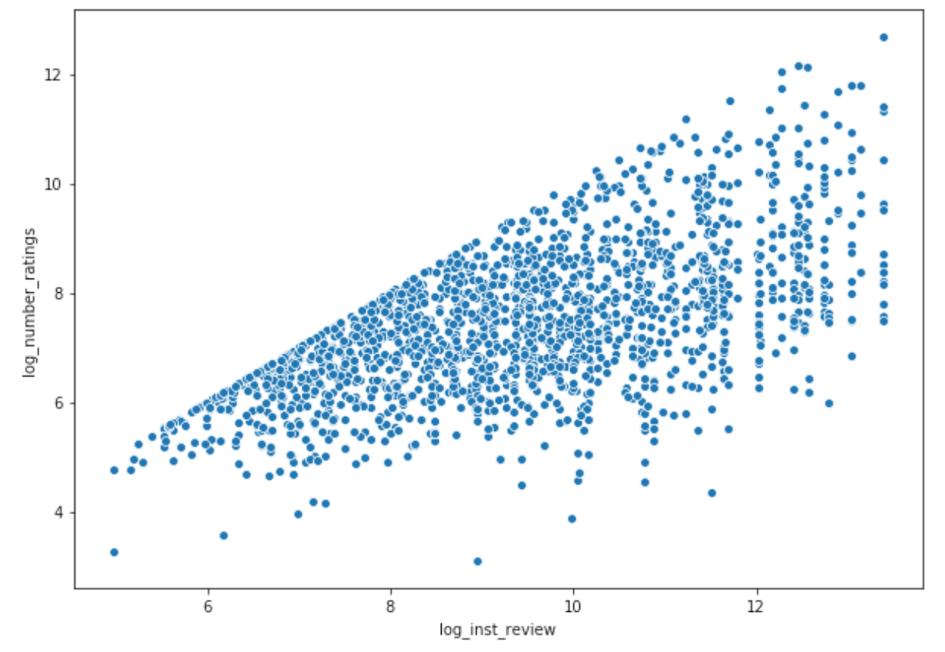
1904 Rows and 18 columns.

**3) EDA** *(Exploratory Data Analysis)*

****

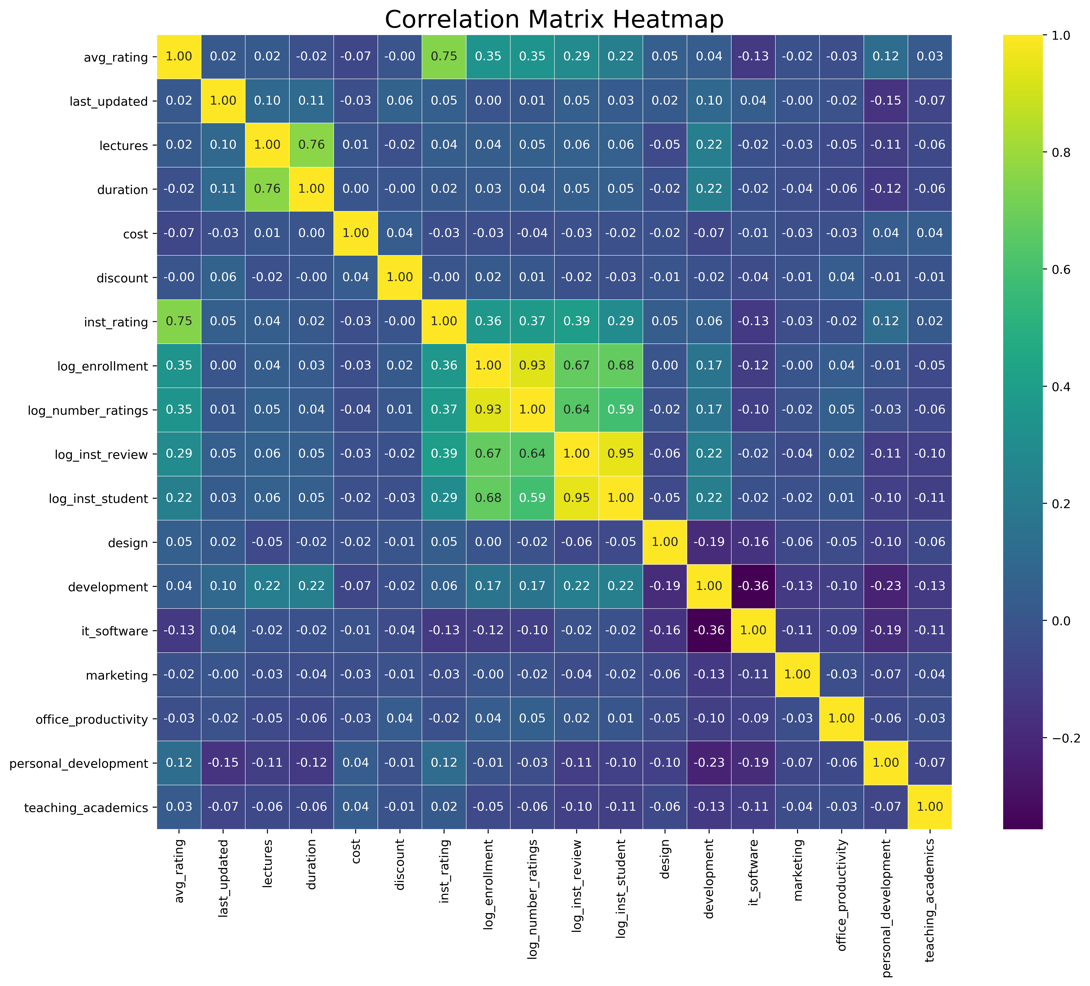
***‘****log\_inst\_review’* vs ‘*log\_enrollment’*

The number of reviews and number of students enrolled to the course has a strong linear relationship.

****

*‘log\_inst\_review’* vs *‘log\_number\_ratings’*

The number of reviews and number of ratings provided to the course also has strong linear relationship.

****

*Correlation matrix*

Looking at the correlation matrix, all the lighter shades (log-transformed columns) have a very good linear relationship with our target variable.

**4) Feature Observation and Hypothesis**

- Looking at the correlation matrix and the scatter plots, the log transformed columns have a good linear relationship with our target variable...

- Log transformation brings the distribution of values closer to Normal which is why all the transformed columns have a good correlation value...

- None of the other features seem to have a linear relationship with each other, except for 'avg\_rating' and 'inst\_rating', where there is a strong linear relationship...

- Looking at the 3 strongly related features, any model would perform with good accuracy but we cannot know for sure how the model would handle new data...

- But again if the new data also undergoes log transformation, we can assume to a certain extent that the model would perform well...

- We will have to work out a few of the feature selection methods to know which provides the best results...

**5) Simple Linear Regression Report**

We employ 3 different feature selection methods.

The correlation based selection (or) manual selection is looking at the correlation values of all columns with the target column and manually fixing a threshold for a good correlation value and select all columns above that threshold.

The next method is Variance Threshold method, where columns with low variance (almost all values are similar or close-by) specified by a certain threshold value are removed and every other column is selected as a feature.

The next method is Select K-Best method, where the computer generates *n* number of selected features, where the value for *n* is provided.

Even though a strong correlation is a value above 0.5, in the real world it is not always the case. Hence, under the given guidance of selecting 6 to 10 features, I chose the best 6 correlated columns to our target variable.

**6) Linear Regression with Ridge report**

Features : All columns other than target column are fed into the Ridge model as it iterates through to find best case.

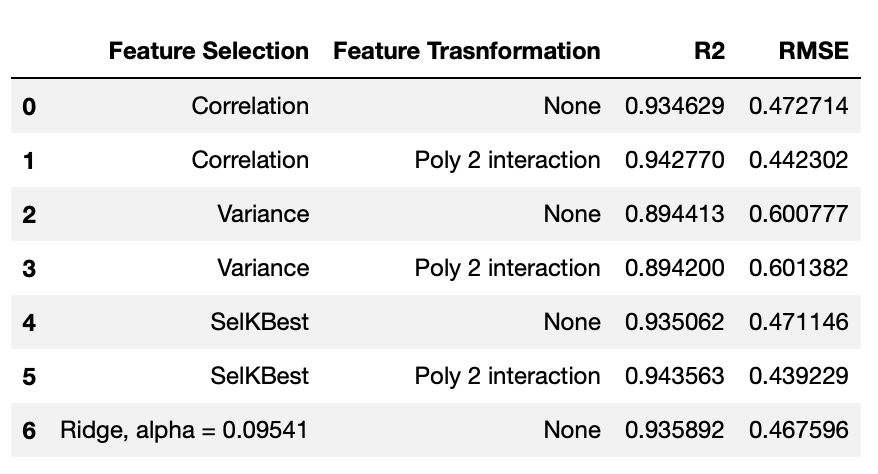
Alpha : Ridge iterates for alpha from 10^n to 1, but we provided 50 equally spaced values from 10^10 to 10^-10.

Maximum number of iterations : We provide the value 10000, but the model stops whenever it finds the best case.

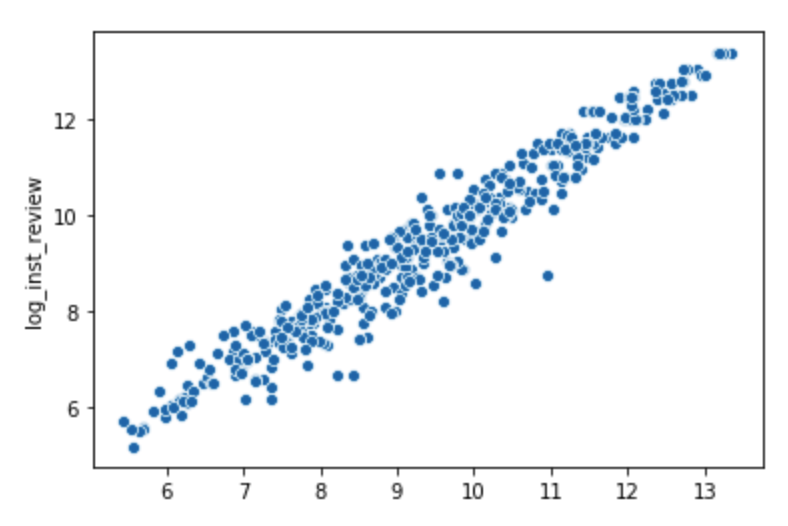
The best performing ridge model had an

**Alpha : 0.09541 Root Mean Squared Error : 0.468 R-Squared : 0.936**

**7) Analysis**

****

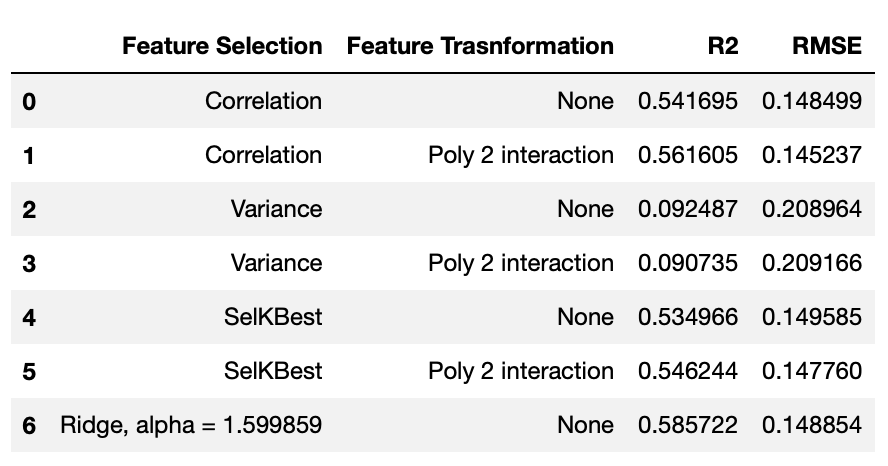


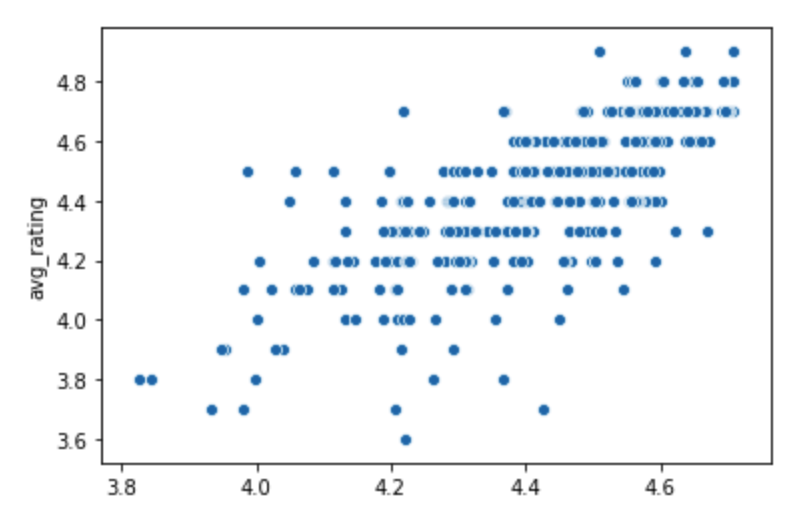
****

The best performing model, after scaling and transformation of features, has very small coefficients of regression. And therefore the margin of error is significantly reduced unlike a model without scaling and transformation. Feature transformation such as polynomial features, helps the regression curve to fit the data better than being a straight line. Feature scaling, brings all features down to a similar scale, so features with different units that vary a lot would look very similar after scaling. We can try to reduce the number of features to ensure all features have strong linear relationship with the target, but that can lead to over fitting. We can also better understand each feature column provided to better select them in regards to the target column.

The model also performs well with the synthetic dataset created with aggregate values from the original. But we cannot be sure about the performance of the model until the model performance is validated with out of the box data.

Bonus Module

****

****

After changing the target variable to ‘avg\_rating’, the linear models seem to not perform well. This is because the target variable does not have much variance in itself and is close to behaving like a categorical column, looking at the multivariate analysis.