1. **Title and Introduction:**

The title of dataset is Prediction on enrollment of courses. This is an uncleaned dataset containing the information of a couple of thousand course dataset in Udemy online platform. The dataset contains 2004 entries having 16 attributes. In this project, data is used which is already cleaned previously and after cleaning it contains 1455 data entries having 16 features again. As some features are dropped and some dummy variables are added in it. In this project, 26 different models are going to fit and then the best one is going to be chosen.

1. **Dataset Analysis:**

In this dataset, in the starting, all features contained the null values especially the price attribute which had 1345 null values. Then after checking the null values, all rows are dropped where all data is missing. After that, we left with only 1970 rows. There two columns price and price2. One of the columns contains more null values. So, we fill the missing values of attribute price2 with price and then dropped the attribute price. After there are some attributes, whose name contains space. That space is replaced with “\_”. Then, the special characters such as ‘$’, ‘%’, ‘CA’ are replaced with ''. Attribute duration values are also changed into hours. The changes in last updated attribute, also took place.

The data type of columns discount, price2 and last updated changes into float64 data type.

To reduce the number of unique values in course category attributes some values are combined. Then, all the unwanted columns were dropped.

After cleaning the dataset, explanatory analysis was performed. The summary of the dataset was checked. The total number of rows that were left were 1936. This analysis showed that there is huge increase in the values from 75% to max value. Similar case with some other features. Then the percentiles for 0.01, 0.25, 0.5, 0.75, 0.99 were also checked. The maximum value for the enrollment in courses is 1112443 and lowest was 36.

After that, courses were checked that had the highest enrollment. According to our analysis, development course has the highest enrollment followed by it software. However, business and education course category have the lowest enrollment. Moreover, there were 20 courses whose enrollment is more than its 99th percentile which is 308696.65 in our case. Moreover, there are 8 courses whose average rating is higher than 99% percentile. However, 16 courses have average rating less than 1% percentile.

During the correlation analysis, we found that number of ratings has high positive correlation with the enrollment. In addition to this, inst student and inst review also has moderate positive correlation with enrollment. Moreover, average rating and inst rating has positive correlation which is not that strong. During the univariate analysis, we can see that the enrollment distribution (after we removed some outliers)still has a longer tail on the right but is better than the previous one.

However, average rating distribution follow some normal distribution as compared to other attributes. During the multivariate analysis, there is positive relationship between number of ratings and enrollment. We have higher number of counts for the business and education course category. we have higher number of enrollments for business and education followed by development and then It Software than 75% percentile of its value. We have lower number of average of ratings for business and education followed by development and then It Software than 25% percentile of its value.

Again, the correlation is performed, and we found that number of ratings and enrollment has high positive correlation which is 85.36% in our case.

Then some feature engineering is done, where we transformed the course category into dummy values and also dropped the unwanted columns. Then at the end , we were left with 1455 rows having 16 attributes.

1. **Exploratory Data Analysis:**

Univariate Analysis: **Distribution and histogram plot of the enrollment**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Distribution plot after removing some outliers

Distribution plot with outliers

We can see that the enrollment distribution (after removing some outliers) still has a longer tail on the right but is better than the previous one.

**Distribution and histogram plot of the average rating**:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Distribution plot after removing some outliers

Distribution plot with outliers

The average rating distribution somehow follow some normal distribution.

Chart

Description automatically generated with low confidence

During the correlation analysis, we found that number of ratings has high positive correlation with the enrollment. In addition to this, inst student and inst review also has moderate positive correlation with enrollment.

1. **Feature observation and hypothesis:**

We can see that number\_ratings has high correlation with enrollment. inst\_student has moderate positive correlation with enrollment followed by inst\_review. Moreover, inst\_rating and avg\_rating has high positive correlation with each other. It is better to consider one of them for analysis. Duration and lectures also have positive high correlation. Inst\_review and inst\_student also has high positive correlation.

1. **Simple Linear Regression Report:**

In this, I am going to use all features using no feature transformation, polynomial feature transformation, log transformation, no feature transformation using MinMax scaling, polynomial feature transformation with MinMax scaling and log transformation with MinMax scaling.

Moreover, I will use the same procedure with manual selection by selecting three features such as number of ratings, inst\_student and inst\_review. In addition to this, I will again follow the same procedure with Variance threshold by using threshold as 0.2 and select k best by selecting 3 features.

In Manual selection, I chose number of ratings, inst\_student and inst\_review as features because they have high correlation with enrollment.

In Variance threshold selection, I chose threshold as 0.2, so 9 features are selected and those are number of ratings, last updated, lectures, duration, price2, discount, inst\_review, inst\_student and inst\_Course.

In SelectKBest method, I chose k = 3, so the selected features are number of ratings, inst\_student and inst\_review.

1. **Linear Regression with Lasso/Ridge Report:**

In lasso regression, alpha is chosen small which is 0.0001. It is not normalized. A dataframe is created using alphas, root mean square error and r square. The coefficient of the best result have RMSE as 4450.05 and R2 as 0.778. The selected features are 15 in number.

In ridge regression, alpha is 0.001. It is not normalized. A data frame is created using alphas, root mean square error and r square. The coefficient of the best result have RMSE as 4446.89 and R2 as 0.778 at index 7. The selected features are 15 in number.

1. **Summary:**

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, scatter chart

Description automatically generated

According to me, Ridge regression model performs better as compared to other models. I choose this, as its R square is 77.82% and root mean square error is also 4446.89 which is less as compared to all models. After this, if we want to choose other models then we can choose Lasso regression model as their root mean square error is too low as compared to others. From the actual vs predicted values, we can see there is positive relationship and after 35000 there could be outliers. We can remove those in order to improve our machine learning model.