Minal Dhansukhlal 300318615

# Udemy Course Enrolment Prediction using Linear Regression

## Introduction

Udemy Course Enrolment Prediction is project to analyse the data and predict the enrolment of the students based on the features selected during feature engineering. The dataset originally had 2004 rows and 16 columns. It is a supervised learning as the dataset is labelled and linear regression model will be used to find enrolment of the students. Features are selected using three techniques such as Manual selection, Select K-Best and Variance Threshold. Polynomial transformation, Log1P transformation and MinMax scaling is used on the features selected to normalize data. Ridge and Lasso regression are used as modelling algorithms to overcome overfitting of basic linear regression algorithm.

## Data Analysis

Initially the dataset had many null values, unnecessary columns, special characters, and many unique features. The figure below shows a glimpse of unclean data.

Graphical user interface, text, application, email

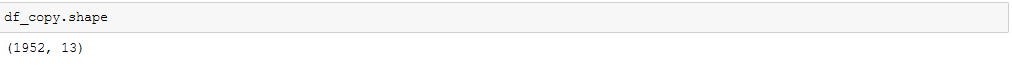
Description automatically generated

Beginning with **removing null values** from the dataset, I removed rows that had all null values in its row. Thereafter I replaced the null values in Price2 with the values in Price since Price had more null values than Price2 and dropped Price column. There were few rows with null values in its rows therefore I removed those rows from the dataset.

Price and Discount had special characters CA$ and % respectively, I eliminated special characters to remove string values which will hinder modelling. Duration has ‘h’ and ‘m’ therefore replaced those letters and converted into decimal format. Split last\_updated date format and converted into decimal format. Reduced course\_category unique values from 14 to 5 categories. Dropped course\_name and course\_sub\_category columns. The figure below shows a glimpse of clean data.

Graphical user interface

Description automatically generated with low confidence



## Exploratory Data Analysis

I had eliminated some of the outliers in enrollment, number\_ratings and inst\_student and remained with 1401 rows. Yet the tail on the right side is evident on those three distribution curves. Avg\_rating has a tail on the left side of the distribution curve, it is skewed to the left. Indicating that most courses had average rating of between 3.5 to 5.

Chart, histogram

Description automatically generated

From the scatter plot and linear regression line between enrolment and number\_ratings, we can see positive relationship between the two variables. Most of the datapoints are on the left bottom of the scatter plots. The scatter plot shows that increase in number\_ratings will increase the enrolment.

Chart, scatter chart

Description automatically generated

There is some positive correlation between enrollment and inst\_student but it is not clear because the data points are very scattered. The data points are denser at the lower part of the plot. From the linear regression model line, it shows that there is positive relation between the two variables thus increase in one variable leads to increase on the other one.

Chart, scatter chart

Description automatically generated

## Feature Observation and Hypothesis

From the correlation table obtained, very few features had strong relationship with enrolment. **Avg\_rating** and **inst\_rating** having similar correlation coefficient of 0.25, **inst\_review** and **inst\_student** having 0.3 and 0.43 respectively. **Number\_ratings** have the strongest relationship with enrolment. Therefore, we can conclude that categories of course independently do not affect enrolment. Based on the correlation table regression model will obtain high accuracy with the five features mentioned.

1. Courses with high average rating (avg\_rating) will have more enrolment. Students usually enroll to courses that has good rating from previous students. Therefore, the higher the average rating, the higher the enrolment of students the two variables have positive relationship
2. Number of ratings received in a course affect student enrolment. The higher the number of ratings received by a course, the higher the enrolment in the course. Students enroll in courses that has been done by many other people previously.
3. The number of reviews affect the number of enrolments. The higher the number of reviews received by a course, the higher the enrolment. Like second point, it gives an understanding that many people have previously taken the course.
4. Inst\_review and number\_ratings can be related in the case that most people who gives rating to the course also leaves a review for the course
5. Courses with very high price will have lower number in enrolment of the course.

## Simple Linear Regression Report

#### Feature Selection

1. **Manual Selection**

There are very few features which are strongly correlated to enrolment therefore I chose features which had correlation coefficient greater or equal to 0.25 with enrolment. 5 features were selected: Avg\_rating, Inst\_rating, Inst\_review, Inst\_student and Number\_ratings.

1. **Variance Threshold Selection**

Using variance threshold value of 0.5, I obtained 9 features. The variance of the features is calculated and features having variance of 0.5 or greater are selected. The features selected are number\_rating, last\_updated, lectures, duration, price2, discount, inst\_review, inst\_student and inst\_course

1. **Select K-Best**

In Select K-Best, we are using f\_regression as scoring function and will rank features in the same order if all the features are positively correlated with the target. K is assigned 7 therefore 7 best scored features are obtained from the algorithm. The features obtained are avg\_rating, number\_ratings, inst\_rating, inst\_review, inst\_student, cat\_development and cat\_it software

## Linear Regression with Ridge/Lasso Report

#### Lasso Regression

All features are used in the model and **alpha** of 33.5698 is selected as it has the lowest RMSE

Graphical user interface, text, table

Description automatically generated Text

Description automatically generated with medium confidence

**Performance Metrics:** R2 = 0.7208535413204171 (72.09%) and RMSE = 3880.9266989993607

#### Ridge Regression

All features are used in the model and **alpha** of 183.2981 is selected as it has the lowest RMSE

Text

Description automatically generated Text

Description automatically generated

**Performance Metrics:** R2 = 0.7201261111612257 (72.01%) and RMSE = 3877.7464233245064

## Summary

Graphical user interface, table

Description automatically generated

Chart, scatter chart

Description automatically generated

Regression model with Select K-Best selection with Polynomial Transformation and MinMax scaling has the highest R square and lowest RMSE thus the best model to choose. To improve the model, I would try to use other feature selection algorithms as we have just tried with 3 methods only. I would have also tried different scaling algorithm instead of MinMax scaling.