Project Luther:

Model the Influence of Factors on Skillshare Class Enrollment

# Summary

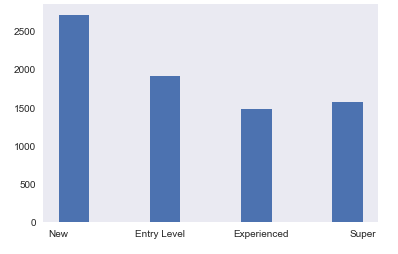
I used Python web scrapping libraries such as Selenium and BeautifulSoup to scrap about 9000 classes from Skillshare.com, then used Pandas and Sklearn to construct a Linear Regression model to predict class enrollment.

# Data Scrapping

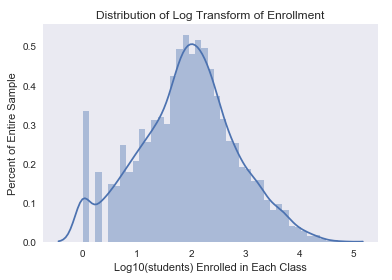
I used selenium to interact with Skillshare.com, navigating to desire pages and closing any popups. Browser was automated to select the class type and to scroll to the end of the page for more classes. Then class URLs were parsed using BeautifulSoup and were saved into a dictionary with class types as keys and lists of class URLs as values. Finally, data of each class was scraped from HTML and embedded JavaScripts. Exception handling was included to handle errors to avoid interruption. Data was saved into a dictionary at the end of parsing one class type or every time when the program stops due to errors. Detail codes are documented in Project 2 - Scraping Data from Skillshare.com.ipynb.

# Linear Regression

Before modeling, comes the data cleaning and feature engineering. I was able to extract most features with simple processing like regex, summing characters, and etc. They are number of videos, summed minutes of all videos, length of class description, whether images included in class description, whether links included in class description, whether sample project provided by the teacher, whether premium membership required, class start date, class types, and number of tags used. However, it is unreasonable to turn teachers into dummy variables. Model would overfit. Therefore, I decided to group teacher into four groups. After grouping, below is a graph showing number of classes taught by teachers with different experience.



What I was trying to predict - the number of students enrolled for each class needs transformation as well. Since I was not interested in predicting the exact number, but rather the order of magnitude. Thus, I did a log10 transform on the labeled data, and the transformed data has a fairly normal distribution. And then I started fitting regression models.

. 

The first model I fit was a first-degree linear regression with class start date as the only feature. This model yields a 0.3293 R square score. Then I added all features and fit a first-degree linear regression first without regularization, then with L1 and/or L2 regularization, further more I repeated same steps for 2nd degree polynomial regression models. 5-fold cross validation was used to find the best alpha if regularization was opted in. R squared scores are listed below. Although 2nd degree polynomial regression provides better R square score, but since the improvement was not outstanding, linear regression without regularization was chosen for the ease of interpretation.

# Results

When interpreting final fitted model, I noticed that creative and design classes are most popular classes on Skillshare.com. The highest positive coefficient that is not a class type is whether sample project is provided by teachers. It shows if teachers want to attract more students they should showcase their work. The most significant negative coefficients tech and finance related class types. It suggests that if Skillshare.com want to expand their market, they should improve their tech and finance related classes to attract more new students.

