أكاديمية سدايا
SDAIA Academy

Regression Project

Elyas Maghrabi

Abstract

Udemy is one of the most popular E-learning platforms in the world. As mentioned on their website, the platform has over 75,000 instructors, 150,000 courses, 250 million enrollments and 33 million minutes worth of content.

Udemy is a massive online open course (MOOC) platform that offers both free and paid courses. Anybody can create a course, a business model by which allowed Udemy to have hundreds of thousands of courses.

Objective

The objective is to analyze dataset based on several variables, determine what variables affect courses Subscribers the most, then build a model that can predict the Subscribers of a courses.

Dataset

The dataset was collected by using web scraping methods like:

- Selenium
- Beautiful soup

EDA:

- Pandas
- Matplotlib
- NumPy

Dataset contains :

- There are 9899 rows and 12 columns in the records of courses from 4 Levels (Beginner, Intermediate, Expert and All Levels) taken from Udemy.
- Subscribers column is the column represents how many people have subscribed to each course (target).

Time Of Scraping:

Around 4 - 5 hour

Challenges of Web Scraping

- In the beginning one of the challenges, I come across while scraping information from websites is the various structures of websites. (Meaning, the templates of websites will different and unique)
- Getting Banned (web scraper bot sends multiple parallel requests per second or unnaturally high no of requests)
- Need to use try and catch to avid loss your work.

Exploratory Data Analysis

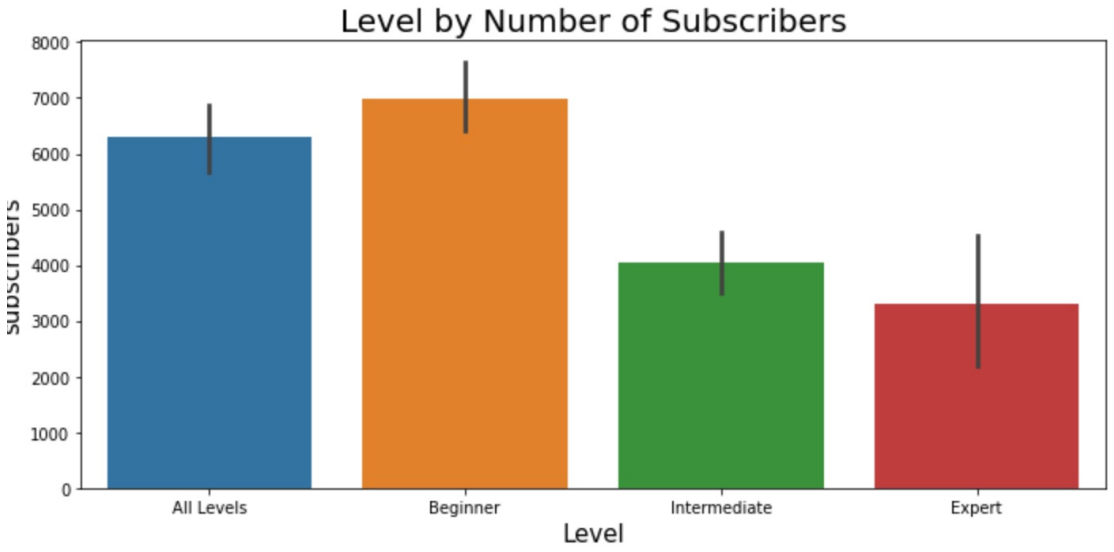
Exploratory Data Analysis

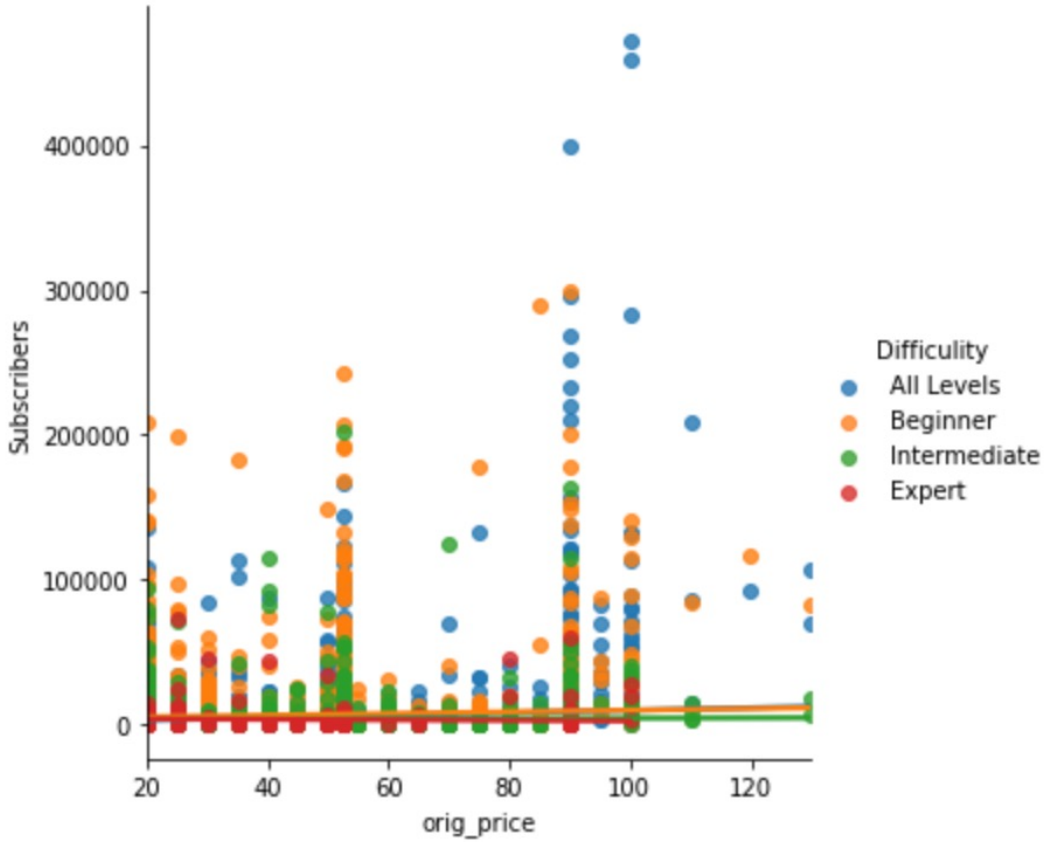
```
udemy.isna().sum()
```

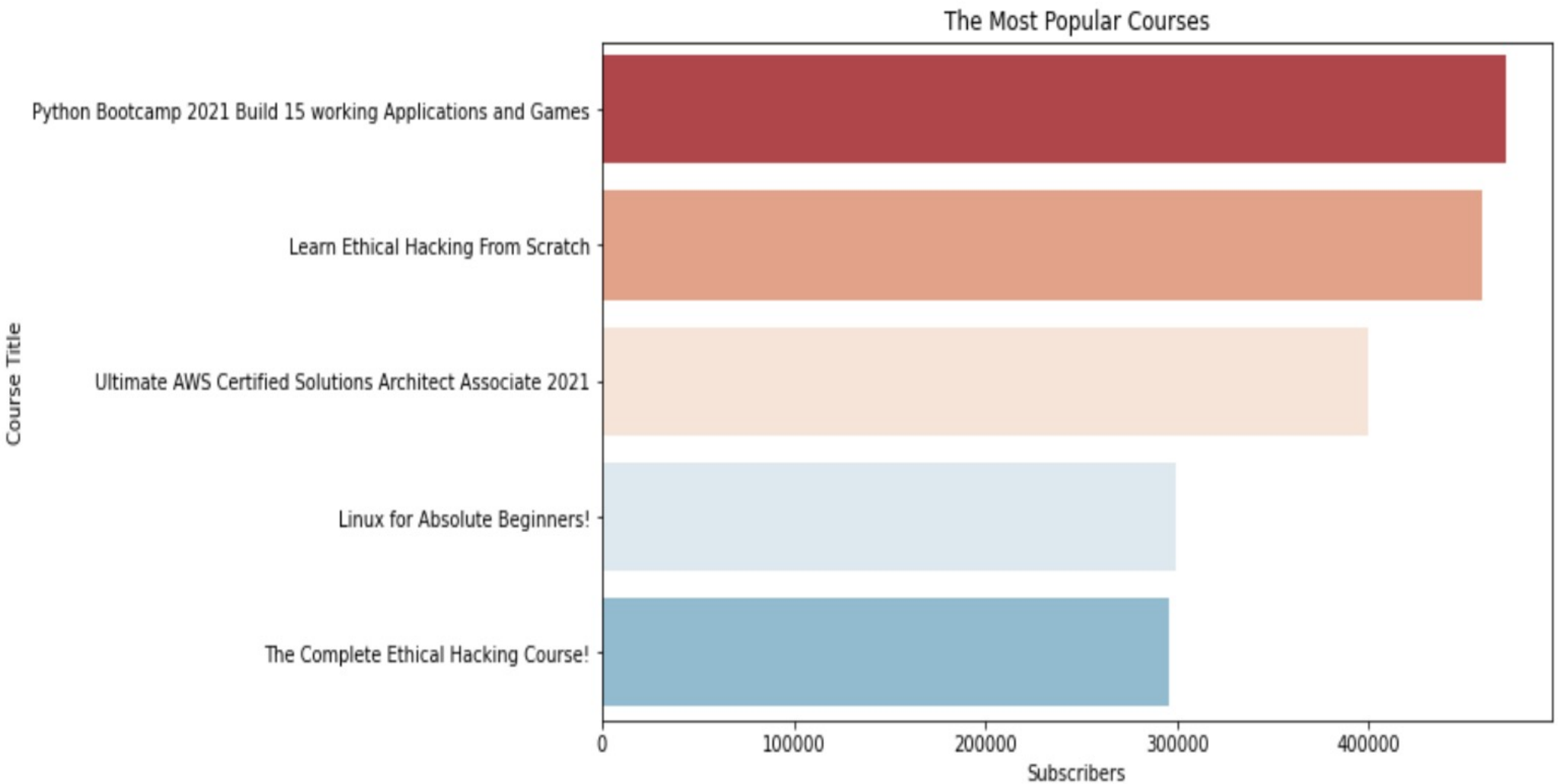
```
url                    2
Course Title           2
Course Headline        2
Instructor             2
off_price              2
orig_price          1403
Rating               605
Number of Ratings    605
Course Length          2
Number of Lectures     2
Difficulity            2
Subscribers            2
dtype: int64
```

```
: udemy.isna().sum()
```

```
: url                  2
  Course Title         2
  Course Headline      2
  Instructor           2
  off_price            0
  orig_price           0
  Rating               0
  Number of Ratings    0
  Course Length        0
  Number of Lectures   0
  Difficulity          2
  Subscribers          0
  dtype: int64
```


Level by Number of Subscribers
Subscribers
8000
7000
6000
5000
4000
3000
2000
1000
0
All Levels
Beginner
Intermediate
Expert
Level


Subscribers
400000
300000
200000
100000
0
20
40
60
80
100
120
orig_price
Difficulity
All Levels
Beginner
Intermediate
Expert


The Most Popular Courses
Python Bootcamp 2021 Build 15 working Applications and Games
Learn Ethical Hacking From Scratch
Ultimate AWS Certified Solutions Architect Associate 2021
Linux for Absolute Beginners!
The Complete Ethical Hacking Course!
Course Title
0
100000
200000
300000
400000
Subscribers

Modelling

Linear Regression

```
Error Table
Mean Absolute Error      :  6028.765450802256
Mean Squared  Error      :  224538896.62211445
Root Mean Squared Error  :  14984.622004645778
R Squared Error          :  0.4920196767585796
```

Polynomial

```
Error Table
Mean Absolute Error      :  5865.699410217118
Mean Squared  Error      :  220094332.94335902
Root Mean Squared Error  :  14835.576596255334
R Squared Error          :  0.502074731487030 2
```

Lasso

```
_____________________________________________
                Error Table
Mean Absolute Error      :  6056.342747368163
Mean Squared  Error      :  225284977.37992412
Root Mean Squared Error  :  15009.496240044971
R Squared Error          :  0.49033179839889307
```

Grid Search

```
_____________________________________________
                Error Table
Mean Absolute Error      :  5970.0340919594255
Mean Squared  Error      :  221802679.64398697
Root Mean Squared Error  :  14893.041316131066
R Squared Error          :  0.49820989326858167
```

Ridge

```
_____________________________________________
                Error Table
Mean Absolute Error      :  6030.1579935394175
Mean Squared  Error      :  224610580.48838136
Root Mean Squared Error  :  14987.013728170845
R Squared Error          :  0.4918575044396396
```

Grid Search

```
_____________________________________________
                Error Table
Mean Absolute Error      :  5891.988805134075
Mean Squared  Error      :  221338707.98876536
Root Mean Squared Error  :  14877.456368235982
R Squared Error          :  0.49925954869550304
```

Comparison Of Performance

	Model	Train Score	Test Score
0	Polynomial	45.78	50.21
2	Ridge	42.18	49.19
1	Lasso	42.06	49.03

Thank You