Udemy Course Rating Prediction

By:

- Rahaf Alqahtani
- Rawabi Alharbi

Abstract

The goal of this project is to apply Linear Regression on data scraped from Udemy website, to predict the rating of courses based on the features of those courses (prices, reviews, total hours, etc.). I worked with data that I have scraped from Udemy.com using Selenium and BeautifulSoup. After getting the data we start to clean them remove nulls values, fill null discount with zeros and extract numbers from string. After that we started to explore the dataset and split the data into %60 train/%20validation, and %20 for the test .Finally we tried different types of models to find the best one that fit our data which is Polynomial.

Data set

The data that will be used in this project has been extracted from Udemy website (https://www.udemy.com/courses/it-and-software/). It includes data such as: (rating, prices, trainers, etc.) for each course. It includes 10 features and mor than 9K rows .Here is a description of each one:

Columns	Description
Title	Course title
Description	Description of course content
Price	The original price
Discount	The price after discount
Rating	Number of rating
Reviews	Number of reviews
Trainer	Trainer name
Total_hours	Duration of the course
Total_lectures	Number of lectures
Level	Course level

Algorithms

Our steps in this project was:

1. Problem understanding
2. Data gathering by scrapping
3. Data exploration and visualization
4. Feature engineering
5. Starting training and validation our data on different models.

Tools

- Python and Jupyter Notebook
- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting visuialization
- BeautifulSoup and selenium for web scraping
- Sklearn for ML algorithms
- HTML \ CSS

Communication

In addition to the slides and visuals presented, we will share our work on our Github accounts

- https://github.com/RawabiKhalaf/curses-rating-prediction/blob/main/Source_code.ipynb
- https://github.com/rahaftech/T5_LinearRegressionProject