COURSES REASONABLE DISCOUNT PERCENTAGE PREDICTION

PREPEARD BY:
RANYA ALKHTANI

EMAIL & GITHUB ACCOUNTS:
RANYAALKHTANI@GMAIL.COM

REGRESSION PROBLEM
PROJECT REPORT

ABSTRACT

The goal of this project was to make Linear Regression on web scribed data in order to predict reasonable discount percentage for a course. I worked with Udemy courses Dataset which I scribed. after exploring the data I come up with a target and an idea of model. I wanted to make prove of concept using Udemy courses Dataset. the Linear Regression model that I chosen as the Best model, have been going through a lot of experiences with different features.

DESIGN

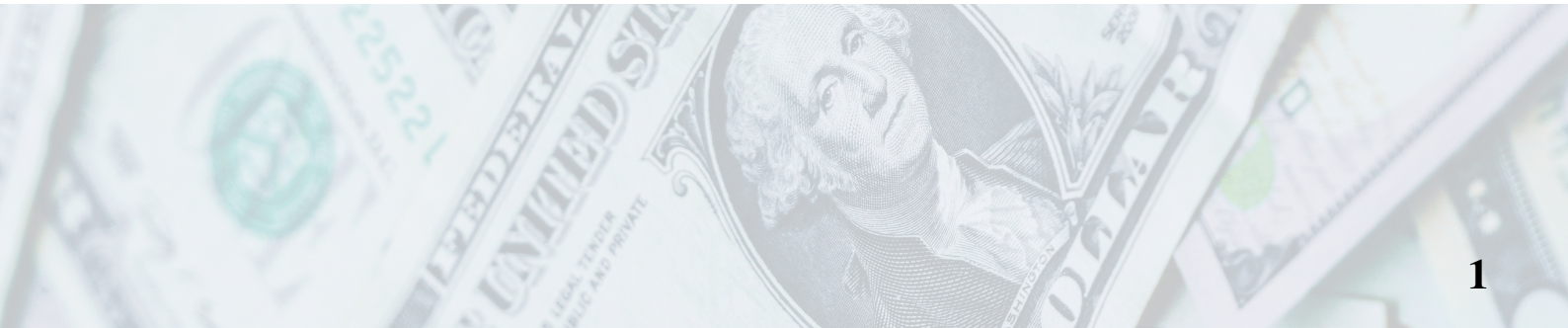
This model is to help the course's instructors by predicting the reasonable discount percentage. The predication is based on the nature of the course. So, the instructors can gain the students attraction, trust, and money. Also, many students feel like cheated when subscribing to specific courses, such analysis could help them to be aware of the market value that the courses have and choose what suits them.

DATA

The data obtained by performing web scarbing on (https://www.udemy.com) website which include many courses with 65 languages and covering almost all subjects that anyone needs. The dataset consists of approximately 20,000 entries linked to 11 features. The features we concern with and get from this website for the courses are ["Title", "Instructor", "Description", "Rating", "Participants", "Original_Praice", "Discount_Price", "Hours", "num_lectures", "Level", "Catogory"].

ALGORITHMS

- **Data Collection:** web scribing.
- **EDA: plotting :** null, correlation.
- **Data Cleaning:** Handling null.
- **Pre-Processing:**
  - **Feature preparation:** dummy values generation.
  - **Feature Engineering:** the target, scaled features, ploy features.
  - **Feature Selection:** Lasso.
- **Modeling:** Linear Regression.



TOOLS

Technologies: python, Google colab.
Libraries: SKlearn, StatesModel, BeautifulSoup, Selenium, NumPy, Pandas, Matplotlib, Seaborn, LinearRegression, time, os, Request, IPython.core.display, plotly , missingno,Pipeline ,MinMaxScaler, PolynomialFeatures , train_test_split, cross_val_score ,KFold, GridSearchCV, Lasso.

COMMUNICATION

in this part of the report I will show and document all of communication work has been shown both of visualisations and slide show: