Price Optimization Analysis Project

Name: Gyan Anand

Internship: Data Science Intern, InternsElite

Batch: September 2024

---

1. Introduction

In today's competitive e-learning market, the pricing of online courses plays a crucial role in determining a course's success. By using data science and machine learning models, we aim to find the optimal price for Udemy courses to maximize revenue and improve customer satisfaction.

The core function of this project is to create a **Random Forest Regressor** model to predict the price of courses and to simulate various pricing scenarios to find the optimal price. The predictions from the model will be used to help educators adjust their pricing strategy and increase revenue while maintaining the course's value.

---

2. Problem Statement

The objective of the **Price Optimization Analysis Project** is to implement a model that predicts the optimal price of a course based on course-related features. By simulating different price points, we aim to identify the best price that maximizes revenue.

Given the dataset, we aim to use machine learning techniques to answer the following questions:

- What is the optimal price for a course to maximize revenue?
- How do other features, such as the number of lectures, content duration, and number of reviews, affect the optimal price?

---

3. Dataset and Features

The dataset contains multiple attributes that are related to courses:

- **price_per_hour**: Price per hour of course content.
- **num_lectures**: The number of lectures in the course.
- **content_duration**: Total duration of the course in hours.
- **is_paid**: Whether the course is paid or free.

- **num_subscribers**: The number of students enrolled in the course.
- **num_reviews**: The number of reviews received.
- **lectures_per_hour**: The number of lectures per hour of course content.

The target variable is **price** (course price).

---

4. Model Building

To achieve the objective of price prediction, a **Random Forest Regression** model was used. Other models such as Linear Regression, Ridge, Lasso, and Gradient Boosting were also tested, but Random Forest yielded the best performance.

Steps for Model Development:

1. **Data Preprocessing**:
    - Handling missing and infinite values.
    - Creating new features such as price_per_hour and lectures_per_hour.
2. **Model Training**:
    - The data was split into training and testing sets with an 80-20 ratio.
    - A **Random Forest Regressor** was trained using the selected features.
3. **Model Evaluation**:
    - The model was evaluated using **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared (R^2)**.
    - Random Forest outperformed other models with an R^2 of **0.9948**.

---

5. Price Simulation

After building the model, a pricing simulation was conducted to determine the optimal pricing strategy. The simulation explored a range of prices and evaluated the predicted revenue based on different numbers of lectures, content durations, and prices per hour.

Simulation Parameters:

- **Price per hour**: $5 to $100
- **Number of lectures**: 1 to 50
- **Content duration**: 0.5 to 10 hours

Optimal Pricing Scenario:

The simulation identified the optimal price point where revenue was maximized:

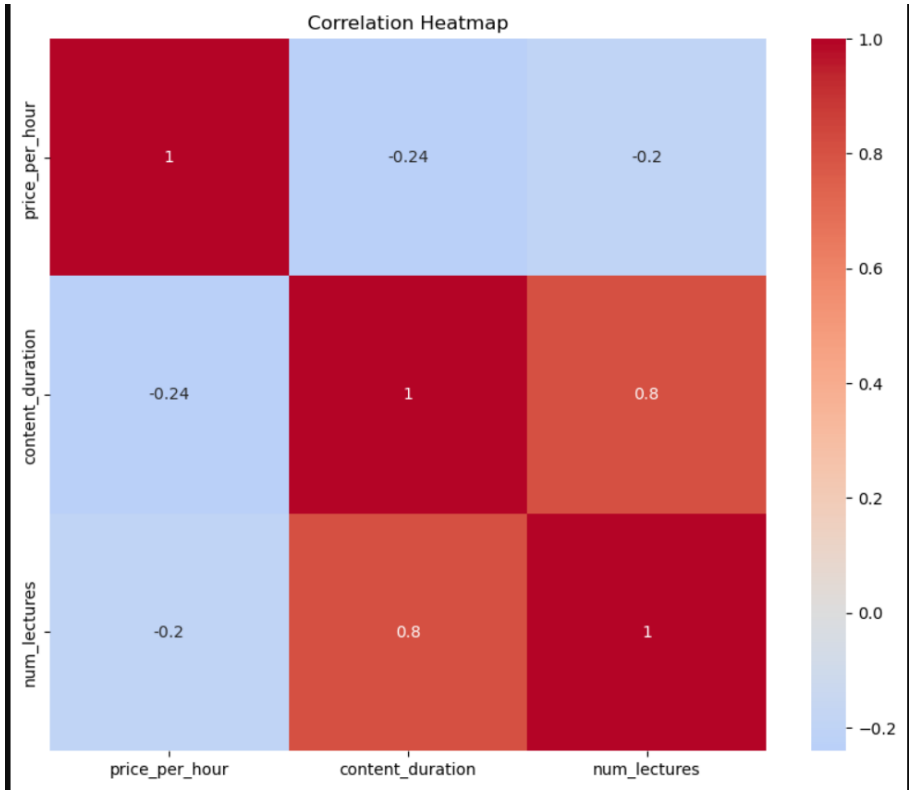
- **Price per hour**: $50
- **Number of lectures**: 30
- **Content duration**: 8 hours
- **Predicted price**: $150
- **Expected Revenue**: $4500

---

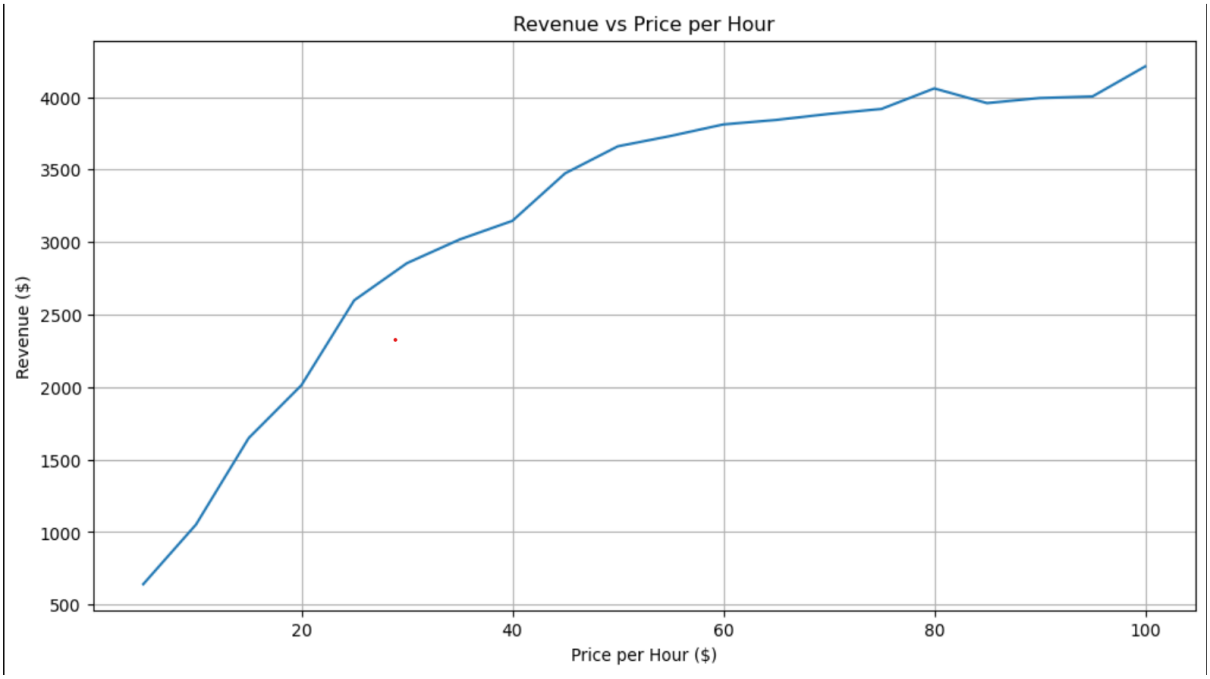
6. Visualizations

Several visualizations were created to better understand the relationship between different variables and revenue:

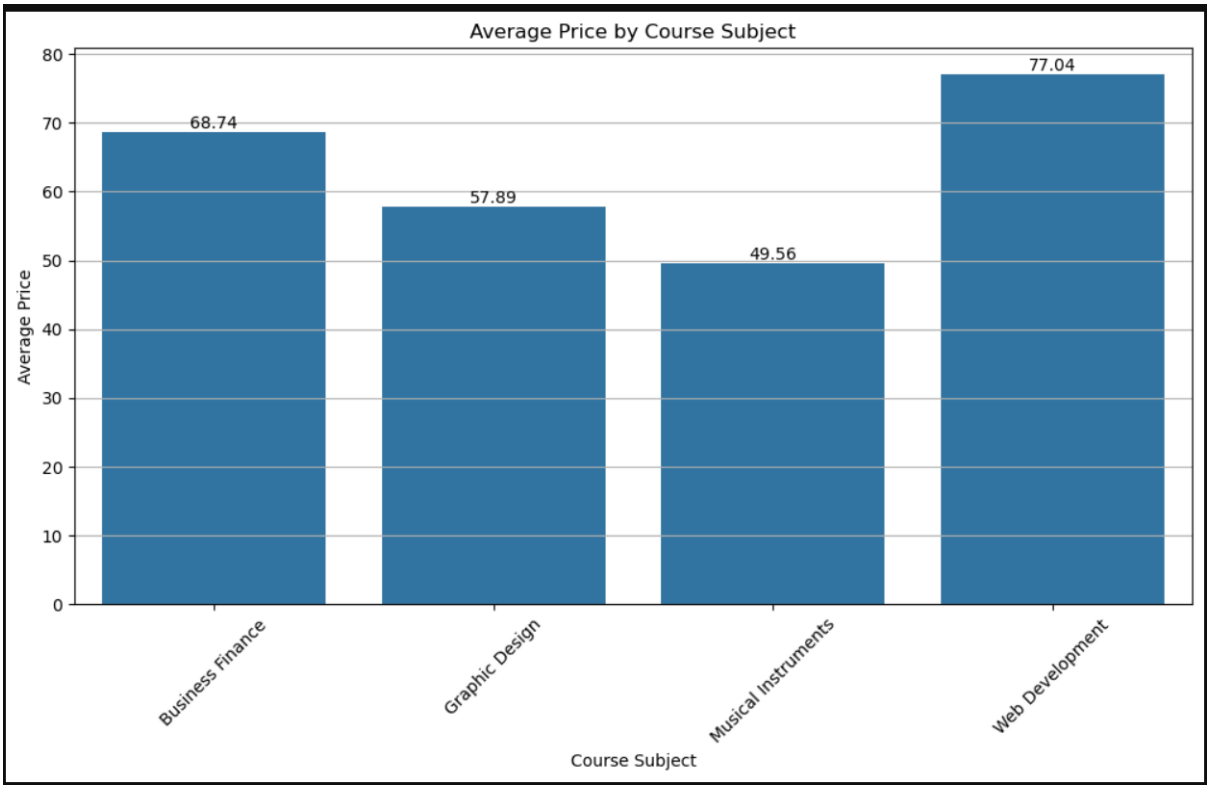
1. **Correlation Heatmap**:
    - A heatmap was created to visualize the correlations between features such as num_subscribers, price_per_hour, and content_duration.



2. **Revenue vs Price per Hour Plot**:
    - A line plot was created to show the relationship between price per hour and expected revenue, with a peak around $50 per hour.



3. **Revenue vs Number of Lectures Plot:**
    - Another plot was created to demonstrate how the number of lectures impacts revenue.



7. Results and Discussion

The analysis demonstrates that the optimal price per hour for Udemy courses lies between $45 and $55, especially for courses with around 30 lectures and 8 hours of content. This pricing strategy is expected to maximize revenue while maintaining course value.

By identifying patterns in the data, this model can assist course creators in setting competitive prices that attract subscribers and increase overall revenue.

---

8. Conclusion

This project successfully developed a price optimization model using Random Forest Regression, allowing us to predict the optimal price for Udemy courses. The simulation provided clear insights into the most profitable pricing strategies. Future work could explore incorporating additional features such as seasonal trends, promotional effects, and student engagement metrics to refine the pricing model