MKT 740 HW 6: Predicting Box Office Success

Shifath Hafeez Summaya
April 5, 2025

Objective

In this assignment, I built a multiple linear regression model using Python and the `scikit-learn` package. The goal was to predict how much money a movie would earn at the box office based on three features: its production budget, the popularity of its genre, and the star power of its main cast.

Code Implementation and Understanding

I followed these steps to build and understand my multiple linear regression model in Python:

- **Imported necessary libraries:** I imported the libraries needed to load the dataset, perform calculations, and create the model.

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
```

- **Loaded the dataset:** I used `read_excel()` from pandas to load the data and previewed the first few rows with `head()`.

```
movieData = pd.read_excel("HW 6 Data.xlsx")
print(movieData.head(11))
```

- **Selected input and output variables:** I assigned the input features to X and the target variable to y.
  - inputFeatures = [ProductionBudget, GenrePopularity, CastStarPower]
  - targetRevenue = BoxOfficeRevenue

```
inputFeatures = movieData[['ProductionBudget', 'GenrePopularity', 'CastStarPower']]
targetRevenue = movieData['BoxOfficeRevenue']
```

- **Created and trained the model:** I created an instance of the LinearRegression model and trained it using the fit() function.

```
regressionModel = LinearRegression()
regressionModel.fit(inputFeatures, targetRevenue)
```

- **Printed the coefficients and intercept:** I printed the intercept and coefficients to see how much each input feature affects the prediction.

```
print(f"Intercept: {regressionModel.intercept_:.2f}")
print(f"Coefficient for {feature}: {coef:.2f}")
```

- **Made a Prediction:** I used the trained model to predict the revenue for a new movie with the following input values:
  - Production Budget = 120
  - Genre Popularity = 8
  - Cast Star Power = 6

```
newMovieData = pd.DataFrame([[120, 8, 6]],
    columns=['ProductionBudget', 'GenrePopularity', 'CastStarPower'])

predictedRevenue = regressionModel.predict(newMovieData)[0]
print(f"\nPredicted Box Office Revenue for input [120, 8, 6]: ${predictedRevenue:.2f} million")
```

Understanding the Equation

The regression model learned from the data can be represented in this general form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{1}$$

Where:

- y is the predicted Box Office Revenue
- β_0 is the intercept (baseline revenue when all features are zero)
- x_1, x_2, x_3 are the input variables:
  - x_1 = Production Budget
  - x_2 = Genre Popularity
  - x_3 = Cast Star Power
- $\beta_1, \beta_2, \beta_3$ are the coefficients learned from the training data using the `model.fit()` function in scikit-learn

Insights and Interpretation

- **Coefficients and Their Interpretation:** After training the regression model, I printed the intercept and coefficients to understand how each feature affects the predicted Box Office Revenue. The results are:
  - ▷ Intercept (β_0): -103.86
  - ▷ Coefficient for Production Budget (β_1): 3.60
  - ▷ Coefficient for Genre Popularity (β_2): 9.80
  - ▷ Coefficient for Cast Star Power (β_3): 13.56

  Interpretation:
  - ▷ For every additional $1 million in production budget, the predicted revenue increases by approximately $3.60 million.
  - ▷ For each 1-point increase in genre popularity (1–10 scale), revenue increases by about $9.80 million.
  - ▷ For each 1-point increase in cast star power, revenue increases by approximately $13.56 million.
- **Complete Regression Model:** Using the intercept and coefficients from the trained model, the complete regression equation is:

$$\text{BoxOfficeRevenue} = -103.86 + (3.60 \times \text{ProductionBudget}) + (9.80 \times \text{GenrePopularity}) + (13.56 \times \text{CastStarPower})$$

  This equation allows us to plug in any combination of input values to predict the expected box office revenue for a movie.

- **Prediction for Given Values:** I made a prediction for a new movie with the following input values:
  - Production Budget = 120
  - Genre Popularity = 8
  - Cast Star Power = 6

  Predicted Box Office Revenue:

$$\text{BoxOfficeRevenue} = -103.86 + (3.60 \times 120) + (9.80 \times 8) + (13.56 \times 6)$$

$$= -103.86 + 432 + 78.4 + 81.36 = \boxed{487.90 \text{ million USD (approx.)}}$$

  This shows how the model can be used to estimate movie performance based on known features.

Submission Links

The primary implementation is hosted on Google Colab as required. Additionally, a GitHub repository is included for version control and easier access to all related files.

Platform	**Link**
Google Colab Notebook	`https://colab.research.google.com/drive/16qo0oP3FsYKIKrO42Z1eJ0ykfC7b6fav`
GitHub Repository	`https://github.com/sh2794s/boxofficerevenue`

Sample Result

To visualize the working of my regression model, I have included a sample result below. This shows the prediction output for a movie with the following inputs and the model predicted a revenue of approximately $487.90 million.

- Production Budget = 120
- Genre Popularity = 8

- Cast Star Power = 6

.

```
D:\>python D:\work\MKT-740\Summ\3\MovieRevenue.py
Dataset Preview:
    BoxOfficeRevenue  ProductionBudget  GenrePopularity  CastStarPower
0                500               100                8              9
1                150                40                6              5
2                300                70                7              8
3                450                85                9              9
4                200                55                5              7
5                100                20                4              6
6                600               130               10             10
7                250                60                6              7
8                350                75                8              8
9                 50                10                3              4
10               400                90                9              9

Model Summary:
Intercept: -103.86
Coefficient for ProductionBudget: 3.60
Coefficient for GenrePopularity: 9.80
Coefficient for CastStarPower: 13.56

Predicted Box Office Revenue for input [120, 8, 6]: $487.82 million

D:\>
```

Figure 1: Box Office Revenue Prediction