Case Study 7, Part 1 Homework: Exercises 1-4

Exercise 1

1/1 point (graded)

First, we will import several libraries. **scikit-learn** (`sklearn`) contains helpful statistical models, and we'll use the `matplotlib.pyplot` library for visualizations. Of course, we will use `numpy` and `pandas` for data manipulation throughout.

Instructions

Read and execute the given code, then call `df.head()` to take a look at the data.

Here's the import code:

```
       import pandas as pd
import numpy as np

from sklearn.model_selection import cross_val_predict
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import r2_score

import matplotlib.pyplot as plt

df = pd.read_csv("https://courses.edx.org/asset-
v1:HarvardX+PH526x+2T2019+type@asset+block@movie_data.c
sv", index_col=0)

# Enter code here.
```

What is the title of the first movie in this dataset?

Answer = [Avatar]

Code = [

```
df.head()
```

```
budget genres homepage id keywords original_language original_title overview popularity production_companies ... revenue runtime spoken_languages status tagline title vote_average vote_count movie_id cast
0 237000000 Action, Adventure, Fantasy, Science Fiction http://www.avatarmovie.com/ 19995 culture clash, future, space war, space colony... en Avatar In the 22nd century, a paraplegic Marine is di... 150.437577 Ingenious Film Partners, Twentieth Century Fox... ... 2787965087 162.0 English, Español Released Enter the World of Pandora. Avatar 7.2 11800 19995 Sam Worthington, Zoe Saldana, Sigourney Weaver...
1 300000000 Adventure, Fantasy, Action http://disney.go.com/disneypictures/pirates/ 285 ocean, drug abuse, exotic island, east india t... en Pirates of the Caribbean: At World's End Captain Barbossa, long believed to be dead, ha... 139.082615 Walt Disney Pictures, Jerry Bruckheimer Films,... ... 961000000 169.0 English Released At the end of the world, the adventure begins. Pirates of the Caribbean: At World's End 6.9 4500 285 Johnny Depp, Orlando Bloom, Keira Knightley, S...
2 245000000 Action, Adventure, Crime http://www.sonypictures.com/movies/spectre/ 206647 spy, based on novel, secret agent, sequel, mi6... en Spectre A cryptic message from Bond’s past sends him o... 107.376788 Columbia Pictures, Danjaq, B24 ... 880674609 148.0 Français, English, Español, Italiano, Deutsch Released A Plan No One Escapes Spectre 6.3 4466 206647 Daniel Craig, Christoph Waltz, Léa Seydoux, Ra...
3 250000000 Action, Crime, Drama, Thriller http://www.thedarkknightrises.com/ 49026 dc comics, crime fighter, terrorist, secret id... en The Dark Knight Rises Following the death of District Attorney Harve... 112.312950 Legendary Pictures, Warner Bros., DC Entertain... ... 1084939099 165.0 English Released The Legend Ends The Dark Knight Rises 7.6 9106 49026 Christian Bale, Michael Caine, Gary Oldman, An...
4 260000000 Action, Adventure, Science Fiction http://movies.disney.com/john-carter 49529 based on novel, mars, medallion, space travel,... en John Carter John Carter is a war-weary, former military ca... 43.926995 Walt Disney Pictures ... 284139100 132.0 English Released Lost in our world, found in another. John Carter 6.1 2124 49529 Taylor Kitsch, Lynn Collins, Samantha Morton, ...
5 rows × 22 columns
```

]

Exercise 2

1/1 point (graded)

In Exercise 2, we will define the regression and classification outcomes. Specifically, we will use the `revenue` column as the target for regression. For classification, we will construct an indicator of profitability for each movie.

Instructions

- Create a new column in `df` called `profitable`, defined as 1 if the movie revenue (`revenue`) is greater than the movie budget (`budget`), and 0 otherwise.
- Next, define and store the outcomes we will use for regression and classification.
  Define `regression_target` as the string `'revenue'`.
  Define `classification_target` as the string `'profitable'`.

How many movies in this dataset are defined as profitable (value 1)?

Answer = [2585]

Code = [

```
df['profitable'] = df.revenue > df.budget
df['profitable'] = df['profitable'].astype(int)

regression_target = 'revenue'
classification_target = 'profitable'

df['profitable'].value_counts()
1    2585
0    2218
Name: profitable, dtype: int64
```

]

Exercise 3

1/1 point (graded)

For simplicity, we will proceed by analyzing only the rows without any missing data. In Exercise 3, we will remove rows with any infinite or missing values.

Instructions

- Use `df.replace()` to replace any cells with type `np.inf` or `-np.inf` with `np.nan`.
- Drop all rows with any `np.nan` values in that row using `df.dropna()`. Do any further arguments need to be specified in this function to remove rows with any such values?

How many movies are left in the dataset after dropping any rows with infinite or missing values?

Answer = [1406]

Code = [

```
df = df.replace([np.inf, -np.inf], np.nan)
df = df.dropna(how="any")

df.shape
(1406, 23)
```

]

Exercise 4

1/1 point (graded)

Many of the variables in our dataframe contain the names of genre, actors/actresses, and keywords. Let's add indicator columns for each genre.

Instructions

- Determine all the genres in the `genre` column. Make sure to use the `strip()` function on each genre to remove trailing characters.
- Next, include each listed genre as a new column in the dataframe. Each element of these genre columns should be 1 if the movie belongs to that particular genre, and 0 otherwise. Keep in mind that a movie may belong to several genres at once.
- Call `df[genres].head()` to view your results.

How many genres of movies are in this dataset?

Answer = [20]

Code = [

```
list_genres = df.genres.apply(lambda x: x.split(","))
genres = []
for row in list_genres:
    row = [genre.strip() for genre in row]
    for genre in row:
        if genre not in genres:
            genres.append(genre)

for genre in genres:
    df[genre] = df['genres'].str.contains(genre).astype(int)

df[genres].head()
ActionAdventure    Fantasy      Science Fiction    Crime Drama Thriller
Animation    FamilyWestern      ComedyRomance      HorrorMystery      War    History
Music  Documentary  TV Movie     Foreign
0      1      1      1      1      0      0      0      0      0      0      0      0      0
0      0      0      0      0      0      0
1      1      1      1      0      0      0      0      0      0      0      0      0      0
0      0      0      0      0      0      0
2      1      1      0      0      1      0      0      0      0      0      0      0      0
0      0      0      0      0      0      0
3      1      0      0      0      1      1      1      0      0      0      0      0      0
0      0      0      0      0      0      0
4      1      1      0      1      0      0      0      0      0      0      0      0      0
0      0      0      0      0      0      0
          ]
```

Exercise 5

1/1 point (graded)

Some variables in the dataset are already numeric and perhaps useful for regression and classification. In Exercise 5, we will store the names of these variables for future use. We will also take a look at some of the continuous variables and outcomes by plotting each pair in a scatter plot. Finally, we will evaluate the skew of each variable.

Instructions

- Call `plt.show()` to observe the plot generated by the code given below. Which of the covariates and/or outcomes are correlated with each other?
- Call `skew()` on the columns `outcomes_and_continuous_covariates` in `df`. Is the skew above 1 for any of these variables?

Here is the code to get you started:

```
      continuous_covariates = ['budget', 'popularity',
'runtime', 'vote_count', 'vote_average']
outcomes_and_continuous_covariates =
continuous_covariates + [regression_target,
classification_target]
plotting_variables = ['budget', 'popularity',
regression_target]

axes =
pd.plotting.scatter_matrix(df[plotting_variables],
alpha=0.15, \
      color=(0,0,0), hist_kwds={"color":(0,0,0)},
facecolor=(1,0,0))
# show the plot.
```

```
# determine the skew.
```

Which continuous covariate appears to be the most skewed?

`budget`

`popularity`
Correct

`runtime`

`vote_count`

`vote_average`

`revenue`

`profitable`

Exercise 6

0/1 point (graded)

It appears that the variables `budget`, `popularity`, `runtime`, `vote_count`, and `revenue` are all right-skewed. In Exercise 6, we will transform these variables to eliminate this skewness. Specifically, we will use the `np.log10()` method. Because some of these variable values are exactly 0, we will add a small positive value to each to ensure it is defined; this is necessary because log(0) is negative infinity.

Instructions

For each above-mentioned variable in `df`, transform value `x` into `np.log10(1+x)`.

What is the new value of `skew()` for the covariate `runtime`? Please provide the answer to 3 decimal points.

Answer = [0.530]

Code = [

```
 for covariate in ['budget', 'popularity', 'runtime', 'vote_count', 'revenue']:
    df[covariate] = df[covariate].apply(lambda x: np.log10(1+x))

print(df[outcomes_and_continuous_covariates].skew())
budget         -2.816990
popularity     -0.431543
runtime         0.530489
vote_count     -0.677632
vote_average   -1.080038
revenue        -2.177372
profitable     -1.081030
dtype: float64
```

]

Exercise 7

1/1 point (graded)

Now we're going to save our dataset to use in Part 2 of this case study.

Instructions

Use `to_csv()` to save the `df` object as `movies_clean.csv`.

What is the correct way to save the `df` object?

```
pd.to_csv(df)
```

```
df.to_csv("movies_clean.csv")
```

correct

```
pd.to_csv("movies_clean.csv")
```

```
np.full((3,3), dtype=int)
```

Code = [
df.to_csv("movies_clean.csv")
]