Assignment 3 – Report

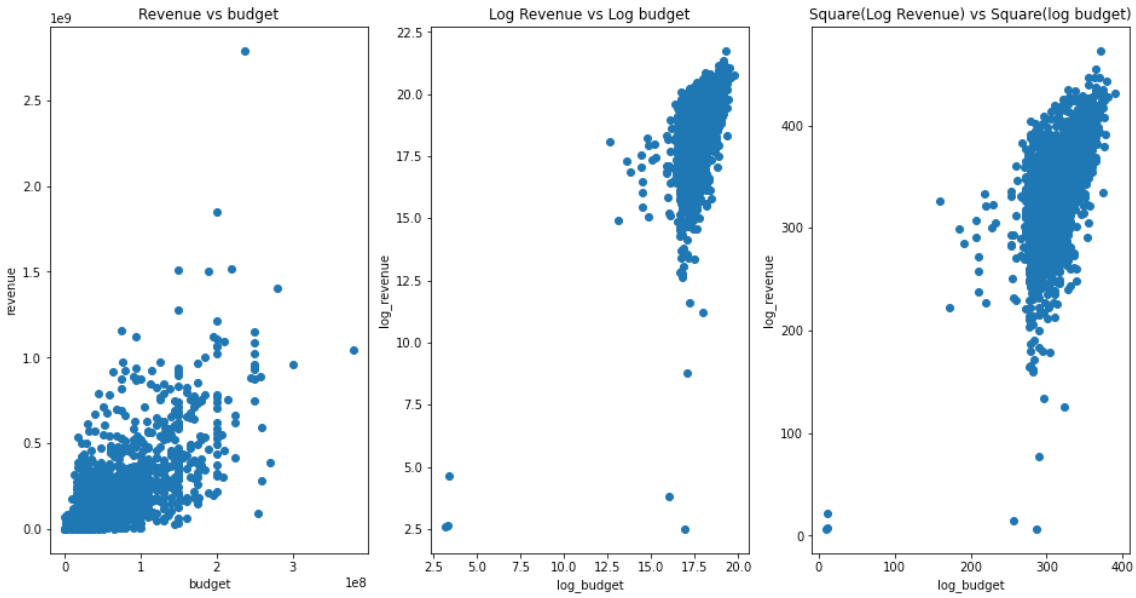
Yao Yuan z5092195

Part-1

Feature Consideration:

For budget

The following scatter plots shows the relationship between Revenue-Budget; Log_Revenue-Log_Budget; Square(Log_Revenue)-Square(Log_Budget) respectively.



It is not hard to see that the second one is too intense and the third one is more suitable for regression model because it is more like to have some mathematical relationships.
After applying all three of them on the same model, as expected, the third one produces the best results.

Plot	Revenue-Budget	Log_Revenue-Log_Budget	Square(Log_Revenue)-Square(Log_Budget)
Coefficient	0.26	0.36	0.42

For “original language”, “genres”, “production company”, “crew”, “keyword”, “cast”

I used the same method “One hot encoding” to utilize these features.

Basically, I count how many times every unique instance appears in the training set for each feature and apply “One hot encoding” to the top 10s of each feature.
For example, for “genres”, I compute how many times each genre appears in the training set. And I pick the top 10 genres to do one hot encoding. In this case, these genres are selected: 'Drama', 'Action', 'Comedy', 'Thriller', 'Adventure', 'Crime','Romance', 'Science Fiction','Family','Fantasy'. The following graph shows each genre’s movie count.

```
Drama              832
Action             756
Comedy             726
Thriller           653
Adventure          568
Crime              348
Romance            333
Science Fiction    329
Family             315
Fantasy            297
Mystery            186
Animation          172
Horror             155
History            102
War                 78
Music               59
Western             30
Documentary          3
dtype: int64
```

For “runtime”

I compared two kinds of processing method.

- The first one is directly using the runtime value
- the second one is dividing runtime into different time range categories which starting from less than 60 minutes, 60 to 80 minutes and so on. Then I apply one hot encoding to these categories.

After validating my results, I find that the second method has better results.
The reason behind this might be that the actual runtime is too sparse and can produce noises.

For “release date”

I convert the time into three categories which are “year”, “weekday”, “month” and “quarter”. And applying one hot encoding again to these categories.

For “Homepage”

From the given training set, I spot that some movies do not have homepage. Therefore, I apply to “one hot encoding” to the homepage. When the movie does not have homepage, I write 0 to that cell, otherwise 1.

Different models:

Model Names	Coefficient
GradientBoostingRegressor	0.43
LinearRegression	0.26
RandomForestRegressor	0.21

Part 2

Feature Consideration:

I applied almost the same method as that in Part1 but different on “Budget”.
Since we are about to do classification problem, for the budget, instead of directly using the value, I used to the same method I apply to “runtime” in Part 1. I split the budget into several range categories and apply one hot coding to them.

Plot	No “one hot encoding”	One hot encoding
Accuracy	0.7075	0.72

Different models:

Model Name	Accuracy
LinearDiscriminantAnalysis	0.71
KNeighborsClassifier	0.7075
SVC	0.6925
LogisticRegression	0.68
DecisionTreeClassifier	0.66
GaussianNB	0.415

I tried using the baseline which is using all 3 as the predicted result and it produces 69.5% accuracy. Therefore, we can see the only the LinearDiscriminantAnalysis and KNeighborsClassifier produce effective results.

Summary

Problem I faced	Solution to it
Find the best math pre-processing method to “Budget” and “Revenue” so that they can have some mathematical relationships.	By trying to apply different math operations and plotting scatter graphs for them and then selecting the best one
Find the best combination of models and features.	Writing two for loops, to loop around different models and different combination of features to find the best one.
Choose which instances should be used when doing one hot encoding.	Apply general knowledges. For example, I can select the instances from “Crew” by looking at the top 10 crews who have highest average revenue in their movies, but it does not make senses. Since if this crew only produces one movie but with high revenue, it cannot guarantee the next movie can also have high revenue.
Possible Data missing	Apply average value to “Budget”, “runtime” if they are null or 0
Feature selection	Apply a flag to indicate which feature is needed in Part1 or Part2