What factors affect movie's budget covering?

Students: Arman Aghamyan, Liana Minasyan, Hakob Hakobyan
Instructor: Hrant Davtyan

American University of Armenia
August 2019

1 Introduction

This paper is a result of the project performed for ECON 317 Data Scraping 2019 course at the American University of Armenia. Paper is focused on finding the relationship between movie's box office coverage over budget and it's audience reviews, critics reviews, run-time, distributor company, producer, leading actors, cinematography, editing, and music:

We also performed Multinomial Naive Bayes for genre prediction based on a description of 300 randomly selected movies.

2 Analysis and Conclusion

2.1 Data collection

For the purposes of this project, we used data scraping techniques and got data from Rotten Tomatoes website and Wikipedia. In case of Rotten Tomatoes we directly scraped from the website using Selenium web-driver, while for Wikipedia, we used Wikipedia wptools API.

Collected data have descriptive statistics showed in Figure 1:

	Unnamed: 0	Runtime	Distributor	Covered	Actors	Producer	Cinematography	Editing	Music
count	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000	2517.000000
mean	2617.810489	103.638804	0.288439	0.646007	0.176798	0.061184	0.174414	0.132698	0.274136
std	1428.460309	18.227213	0.453126	0.478302	0.381573	0.239715	0.379540	0.339315	0.446166
min	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1388.000000	92.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2670.000000	100.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	3877.000000	112.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	5006.000000	272.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 1: Statistical characteristics

2.2 Results

After cleaning data we performed logistic regression with results described in Figure 2: As we see, only Distributor and Producer variables are statistically

```
model1.get_margeff().summary()
```

Logit Marginal Effects

Dep. Variable:	Covered
Method:	dydx
At:	overall

	dy/dx	std err	z	P>|z|	[0.025	0.975]
Runtime	0.0006	0.001	1.086	0.277	-0.000	0.002
Distributor	0.0614	0.022	2.819	0.005	0.019	0.104
Actors	-0.0383	0.025	-1.523	0.128	-0.088	0.011
Producer	-0.0881	0.039	-2.256	0.024	-0.165	-0.012
Cinematography	-0.0209	0.026	-0.808	0.419	-0.071	0.030
Editing	0.0009	0.029	0.032	0.975	-0.056	0.058
Music	-0.0036	0.023	-0.157	0.875	-0.048	0.041

Figure 2: Logit Summary

significant based on p-value less than 0.05 (Figure 3): This result is reasonable, as the main driving force for movie's financial success is marketing campaign with its components. For getting sure that considering our data we get the best model we performed greed search and tuned hyper-parameters. In the end, we concluded, that logistics regression best fits the data. Out of our variables, only those two are related to the marketing campaign. So we may conclude that our model is accurate and consistent with theory.

Additional sub-project of genre prediction resulted in word-clouds for drama and documentary. Accuracy of this model is 0.41 based on RMSE metrics.

```
print(model1.pvalues<=0.05)
```

```
Intercept         False
Runtime           False
Distributor        True
Actors            False
Producer           True
Cinematography    False
Editing           False
Music             False
dtype: bool
```

```
#making summary as dataframe to manipulate easily
log_summary=pd.read_html(model1.summary().tables[1].as_html(),header=0)[0]
log_summary=log_summary.rename(columns={"Unnamed: 0":"Parameter"})
log_summary[["Parameter","coef"]][(log_summary["P>|z|"]<=0.05)]
```

	Parameter	coef
2	Distributor	0.2704
4	Producer	-0.3877

Figure 3: Significant variables