Report Paper of Group Project
Data Scraping

Group 8

Team members:

Arman Aghamyan, Liana Minasyan and Hakob Hakobyan

Instractor:

Hrant Davtyan

TA:

Vazgen Tadevosyan

What factors affect on movie's budget covering?

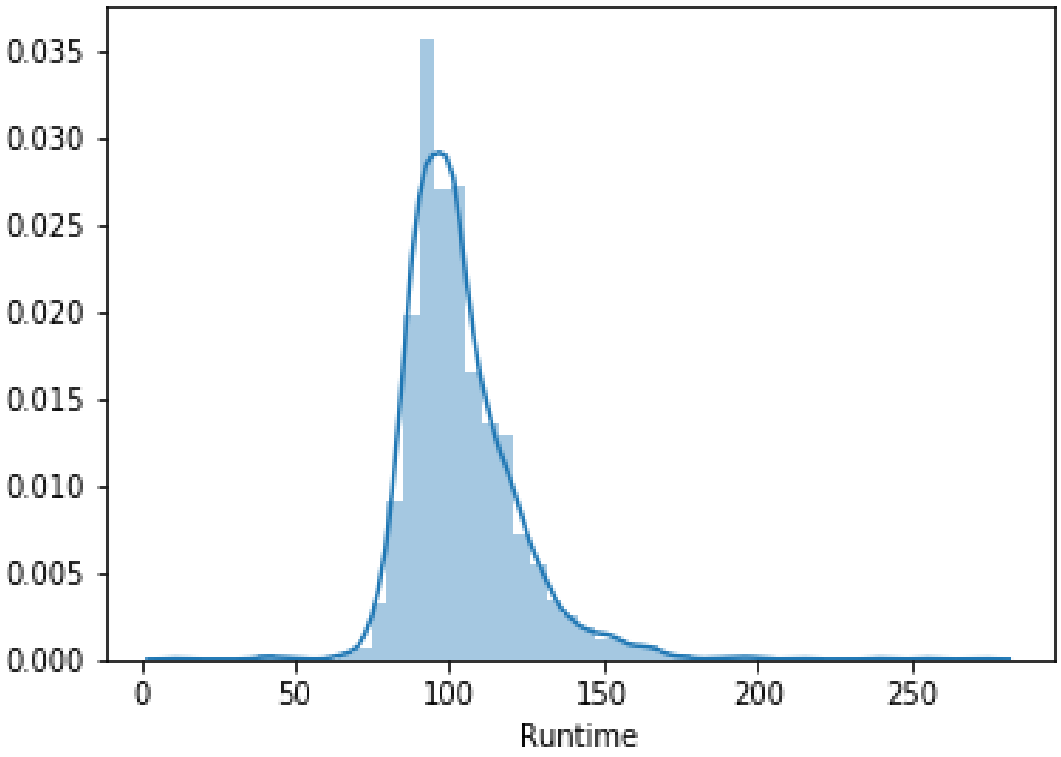
The main idea of the project was to find factors which affect on covering budget of movies. Besides make prediction of genre by movie description. This is steps how we make our project

1. Scraping data from Rotten Tomatoes using
    - Scrapy
    - Selenium
    - WPTools API

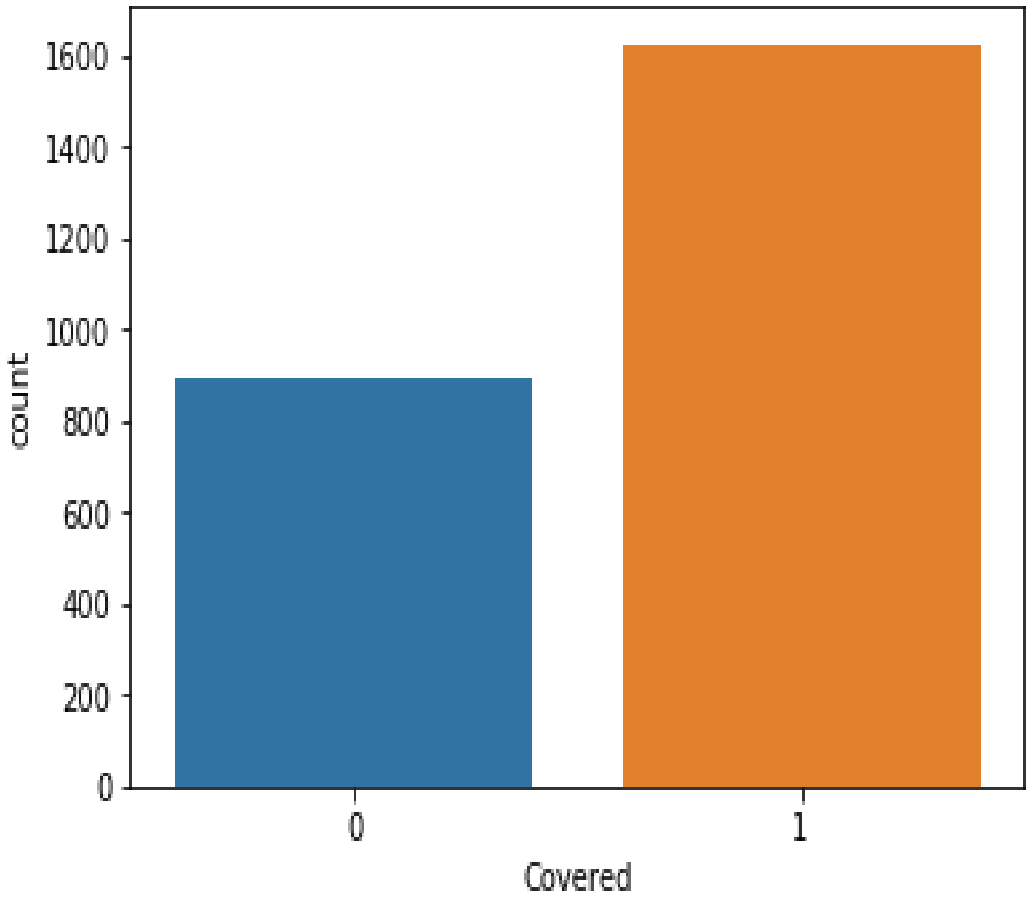
2. Cleaning data with
    - RegEx
    - Pandas

3. Analyzing data
    - Summary statistics and visualizations
    - Logistic Regression
    - Decision Tree
    - Multinomial Naïve Bayes
    - Word Cloud

After scraping data, we have a dirty data with 9500 observations and lot of Na's. We have over 50 columns as info boxes of each movie in Wikipedia contains different information. We chose10 of them that where most common and cleaned each of them individually. We made categorical variables which contained objective type values. And our dependent variable was Covered one. If value is 1 it already covered budget, if not is it 0. This is distribution of Runtime values.



Proportion of movies that covered or not the budget.



Analytical Part of our project.

We run Logistic Regression and Decision Tree using Grid Search CV which is sklear model selection tool to get the best model for this data by automatically changing hyper parameters of models. And we get these results

	precision	recall	f1-score	support
0	0.37	0.52	0.43	209
1	0.70	0.57	0.63	421
avg / total	0.59	0.55	0.56	630

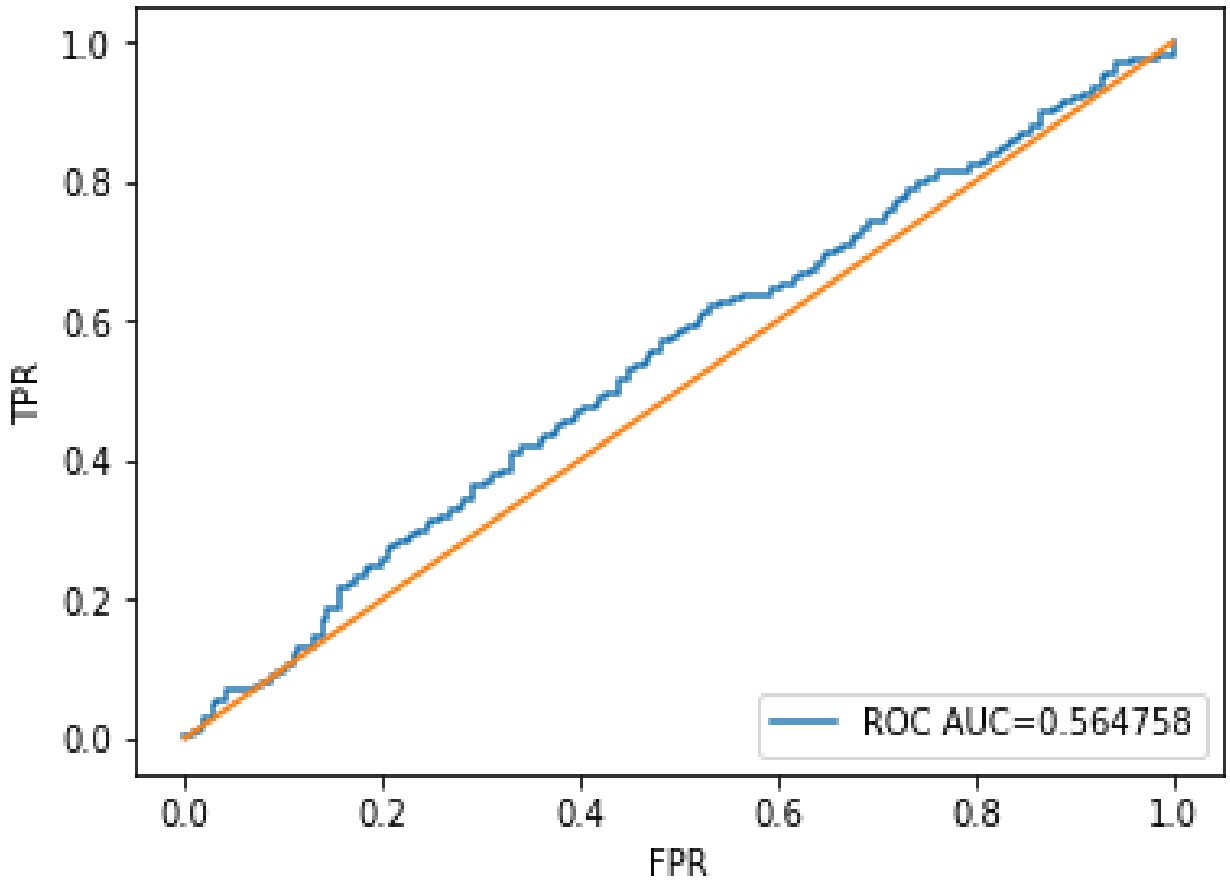
	precision	recall	f1-score	support
0	0.33	0.68	0.44	209
1	0.67	0.32	0.43	421
avg / total	0.55	0.44	0.43	630

First is for Logistic Regression

Second is for Decision Tree

And we can say that Logistic Regression is fitting data better than Decision Tree.

ROC Curve of Logistic Regression.



In conclusion for movie budgets we find that only produsers and distributors affect on budget significantly by

	Parameter	coef
2	Distributor	0.2704
4	Producer	-0.3877

For prediction of genres by movie description we used Multinomial Naïve Bayes model on 300 movie's descriptions which was predicting by 40% by RMSE score.

In future work we should scrape more descriptions and improve predictions.

Word clouds by Drama and Documentary genres.


film
family
time
two
explore
MAY
documentary
American
star
personal
young
death
women
portrait
year
day
first
life
one
world
director
music
band
image
footage
story
history
cultural
performance
extraordinary
feature
America
New
Live
journey
political