Movie Grossing Success Prediction with Convolutional Neural Networks on Graphs

A Network Tour of Data Science (EE-558) - Project

Oriol Barbany - Manuel Cherep - Natàlia Gullón - Carlos Medina

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Table of Contents

1. Introduction

Goal: Predict movie's success before its release

Motivation: Build a product that allows producers to determine whether a movie is feasible or not

Output: Reduce movie to binary classification task. "Will the movie revenue be higher than the budget?"

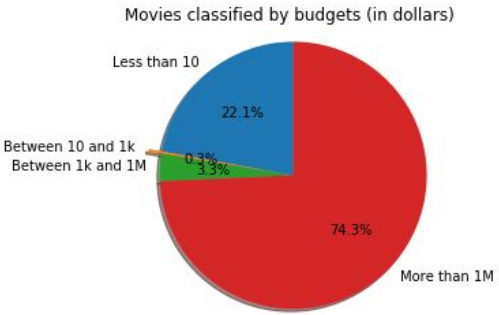
Approach: Predictions made with Convolutional Neural Networks on Graphs

Table of Contents

2. Data exploitation

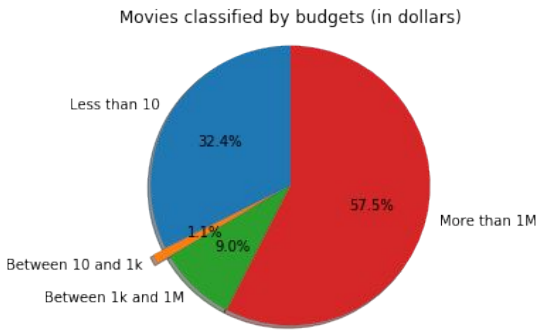
2.1. Data acquisition

Original Kaggle Data



Use TMDb API

Collected Data



Total of 4,803 movies, 3,712 usable

Total of 10,000 movies, 6,651 usable

2. Data exploitation

2.2. Preprocessing and feature engineering

Delete features unknown before release (e.g. vote average, vote count, popularity...)

Process non-numerical features (e.g. cast, director, genres mixture...)

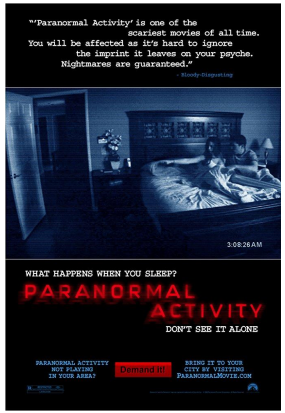
$$r_i = \left(\frac{1}{N_i} \sum_{n=1}^{N_i} r_{n,i} \right) \cdot \left(\frac{N_i}{\max_i N_i} \right)^{\beta}$$

β: Ponderate number of movies done

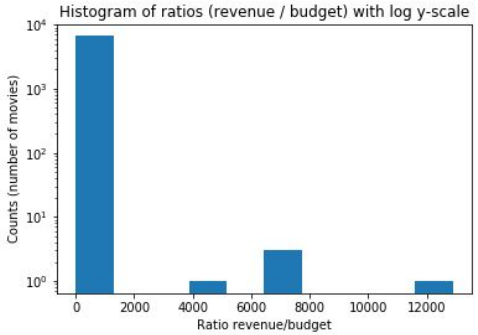
2. Data exploitation

2.3. Handling of outliers

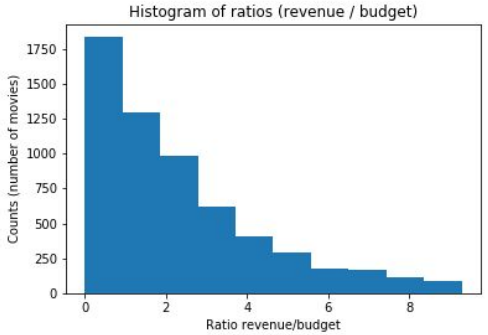
Delete outliers with inner fence approach (Interquartile Range)



Budget:$15,000
Gross: $193,355,800
Ratio: ~12,800



Histogram of movie ratios with outliers



Histogram of movie ratios without outliers

Table of Contents

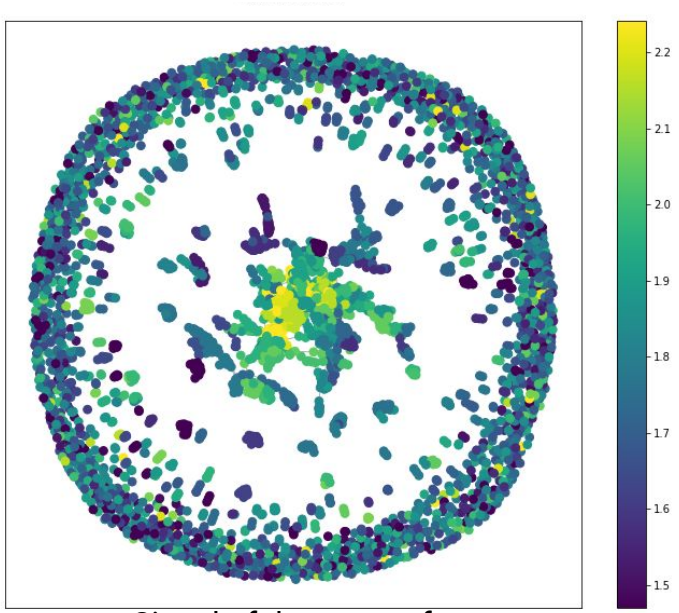
3. Graph

3.1. Creation

Adjacency matrix with Gaussian Kernel (Euclidean distance)

Resulting graph:

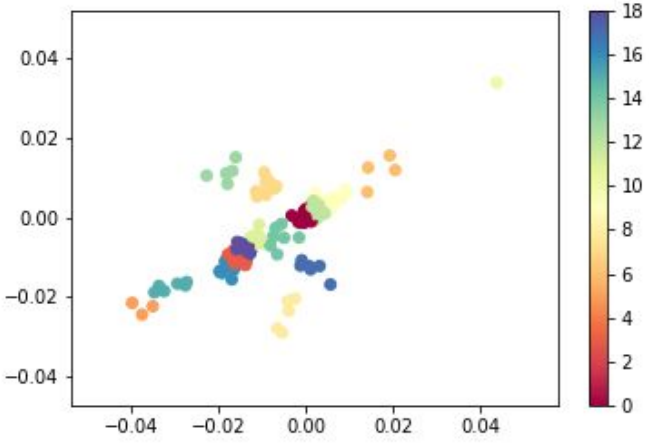
- 5,589 nodes (number of movies)
- 25,519 edges (after forcing sparsity)
- Undirected graph



Signal of the *genres* feature

3. Graph

3.1. Properties



Zoomed in cluster assignment (K-Means)

Feature ranging from ~0.74 to ~2.24

Movie	Genres' ratio	Genres
The Specialist	1.87	Action and Thriller
Argo	1.99	Drama and Thriller
Wish I Was Here	2.20	Drama and Comedy
Gigli	2.20	Drama and Comedy
Erin Brockovich	2.16	Drama

Movies assigned to the same cluster

Table of Contents

4. Data exploitation

Graph over the features (as suggested in *Defferard 2016*): 30 nodes

Upsample to balance classes (from 67% successful to 50%)

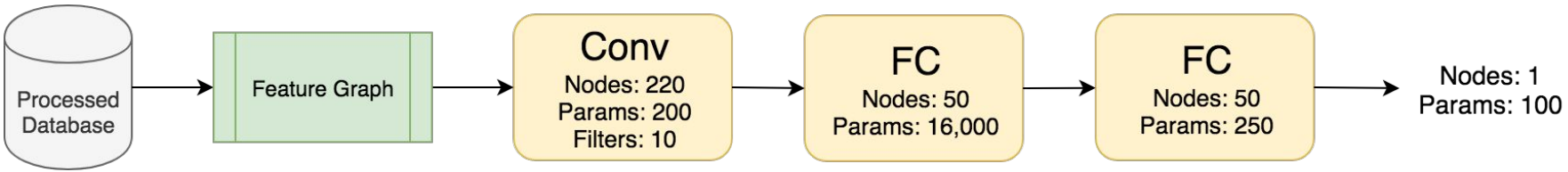


Table of Contents

5. Results

Model	Training accuracy	Test accuracy
GCNN	99.85 ± .07	85.75 ± .64
FC-NN	99.72 ± .05	85.52 ± .76
Logistic Regression	77.56	74.41

Proof of concept

Demo

Table of Contents

6. Conclusions

A future movie’s success can be predicted from its meta-data

Graphs are a powerful tool that allow to study relations among movies and exploit its similarities

Given our prediction rate, our product could be used in a real environment

Importance of treating outliers and process features

Q&A

Thank you

Oriol Barbany - Manuel Cherep - Natàlia Gullón - Carlos Medina