9/12/2020

PROFITABILITY PREDICTION OF MOVIE PROJECTS

Business Understanding

Team 6
Oguzhan Akan
Darius Hooks
Jaymish Raju Patel
Jason Sabal
Bibinur Zhursinbek

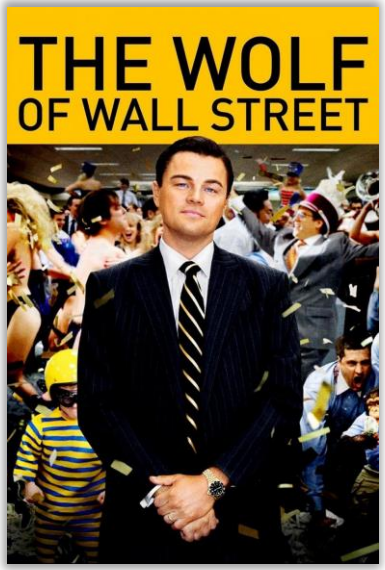
California State University, Northridge

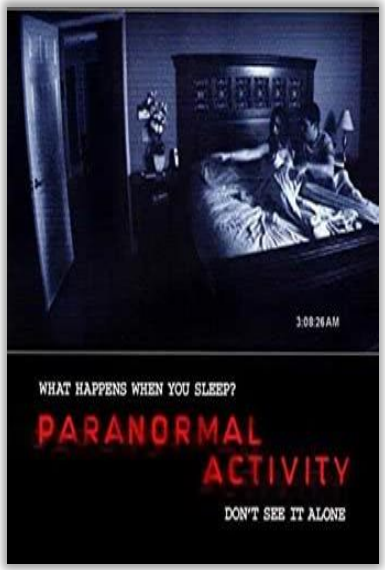
COMP 541 – Data Mining – F2020

Assignment 1

If you were a producer,

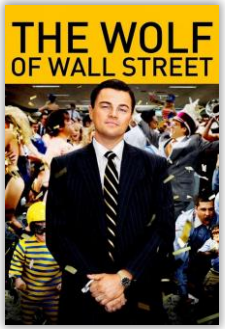
Which movie would you invest in?

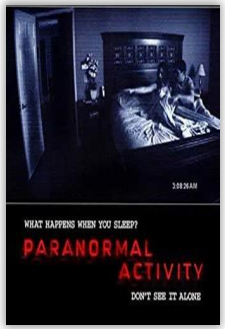




If you were a producer,

Which movie would you invest in?





	The Wolf of Wall Street	Paranormal Activity	Avengers: Endgame
Revenue	$ 392M	$ 193M	$ 2.8B
Budget	$ 100M	$ 15,000	$ 356M
Profitability Ratio	2.9	**12,892 !!**	6.9

1.1. Determining Business Objectives

Background & Objectives

About MovieWorks

- As a movie production company, MovieWorks, our core business is to make investments on movie projects.

Business Objectives

- MovieWorks' goal is to achieve profits by investing in a film. In order to achieve this it is necessary to determine which movie project is going to make more money.

Business Questions

- which actors are going to act
- what is the cost of the movie
- what kind of movies are highly watched
- what movies get highest ratings
- in what locations was the movie released
- in which country which kind of movie is watched maximum

1.1. Determining Business Objectives

Business Success Criteria

Key Performance Indicator: Profitability Ratio

As MovieWorks, our goal is to achieve at least **250%** profitability ratio out of a project we invest in. We define the profitability ratio below.

$$ProfitabilityRatio = \left(\frac{GrossRevenue - Budget}{Budget}\right)$$

1.2. Assessing the Situation

Inventory of Resources

Personnel

- IT Professionals
- Data Scientists
- Data Analysts
- Data Engineers
- Testers

Data

- Types: csv, json, txt, excel
- Tools: Python, R

Hardware & Software

- Computers
- Monitors
- Github
- Anaconda Environment
- Google Drive

1.2. Assessing the Situation

Requirements, Assumptions, and Constraints

Requirements

- Target Group: Film producers, film directors and film studios
- Data Availability

Assumptions

- Accurate and based on yearly information
- Available in various formats

Constraints

- Limited hardware due to COVID
- Limited prediction accuracy

1.2. Assessing the Situation

Risks and Contingencies

Risk	Type	Contingency Plan
Too few projects to prefer so that even though our prediction works fine, we may not find a good project to invest in.	Business Risk	Select multiple low-profit movie projects
Computer hardware may not be strong enough to mine huge amounts of data.	Technical Risk	Filter the data before mining it so hardware can handle.
Data may be incorrect	Data Related Risk	Switch to new datasets
Data may be protected	Data Related Risk	
Mining environment may not be compatible among team members	Technical Risk	Setup an external environment before starting to project that is available for every member's use

Terminology

Business Glossary

Term	Description
Profit	A general term that indicates the difference between a revenue amount and a cost amount.
Profitability Ratio	The rate at which the revenue increases with respect to the budget.

Data Mining Glossary

Term	Description
Accuracy	Refers to the degree of fit between the model and the data.
Precision	A measure of how variable the estimate would be over other similar data sets.
Clustering	Clustering algorithms find groups of items that are similar.
Classification	Problem of attempting to predict the category of categorical data by building a model based on some predictor variables.

Costs and Benefits

Cost Type	Description	Estimated Amount	Is it a variable or fixed cost?
Hardware	Servers & Telecommunications & Maintenance	$200/mo	variable
	Desktops/Laptops	$5000	fixed
Labor	Development & Management	$400k/yr	variable
Software & Data Related	Github Subscription	$20/mo	variable
	Data Collection	$0	none
	Model Deployment	$50/mo	variable

Primary benefit of the project will be providing management better decisions on the investment strategy and eventually help the company grow!

Data Mining Goal and Success Criteria

Data Mining Goal

- MovieWorks' data mining goal is to accurately and precisely predict the profitability ratio of a proposed movie project.

Data Mining Success Criteria

- Classification Model:
  - **75% Accuracy**
  - **40% F-1 Score**
- Regression Model:
  - **Root Mean Square Error (RMSE)**

Project Plan

1.4. Project Plan

Project Plan

Phase	Duration	Resources	Inputs	Outputs	Dependencies	Risks
Business understanding	1 week	Business analysts	Company's background, Data	Report	Workforce, customers' preferences	Business risk, no clear vision *Recommendations:* have a few ideas for the project
Data understanding	3 week	Analysts	Data	Reports: Initial data collection report, Data description report, Data quality report	Business understanding	Problems with data, problems with technology *Recommendations:* backup data, find different resources
Data preparation	5 week	Analysts	Raw data	Dataset description report, Final dataset	Data understanding	Problems with data, problems with technology *Recommendations:* check data regularly

1.4. Project Plan

Project Plan

Phase	Duration	Resources	Inputs	Outputs	Dependencies	Risks
Modeling	3 week	Data Scientists	Data	Report, Model	Data preparation	Data problems, building a poor model, which does not highlight important parts *Recommendations:* return back to the Data preparation model and create model
Evaluation	1 week	Data scientists, Management	Model	Report	Modeling	Having small budget and less resources *Recommendations:* update information about budget and resources throughout the process
Deployment	1 week	Customer, Data Analysts	Data, model	Report	Evaluation	The project does not meet initial goals *Recommendations:* return to previous phases and follow the process once again

1.4. Project Plan

Tools & Techniques

Phase	Tools	Techniques
Data Understanding	Python, Web Drivers and Browsers	Web scraping for data collection, pivoting for data exploration
Data Preparation	Python	Filtering, Replacing, Insertion, Transformation, Merging
Modeling	Python	Correlation Analysis, Feature Selection, Dimensionality Reduction, Splitting & Shuffling Data, Model Selection, Parameter Optimization
Evaluation	Python	Receiver Operating Characteristics Calculation (ROC), Confusion Matrix Analysis
Deployment	Python, PowerPoint	Documentation

Thank you for your attention.