Assessing Movie Profitability

Tommaso Ghisini, Pietro Marini, Gaspar Dugac,
Georgi Angelchev, Matthew Cernicky, Usama Muhammad

See the Project GitHub

Agenda

https://github.com/tommasoghisini/MovieRevenuePrediction

1) Introduction
2) Univariate Analysis
   I. Numerical Variables
   II. Categorical Variables
3) Data Preprocessing
4) Feature Engineering
5) Bivariate Analysis
   I. Categorical Variables
   II. Continuous Variables
6) Models
   I. Logistic Regression
   II. Decision Tree
   III. Random Forest
   IV. Gradient Boosting
7) Performance
8) Implications

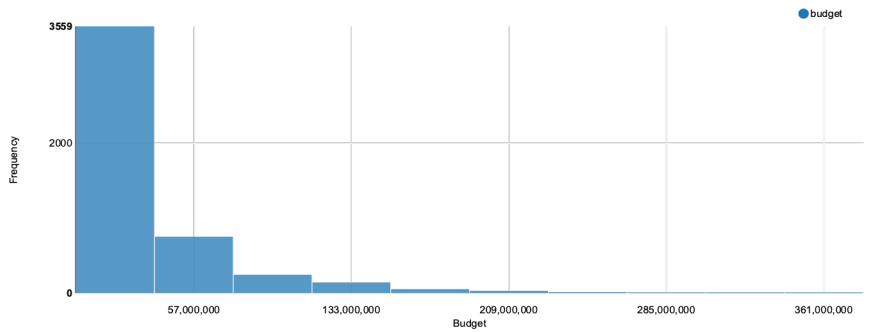
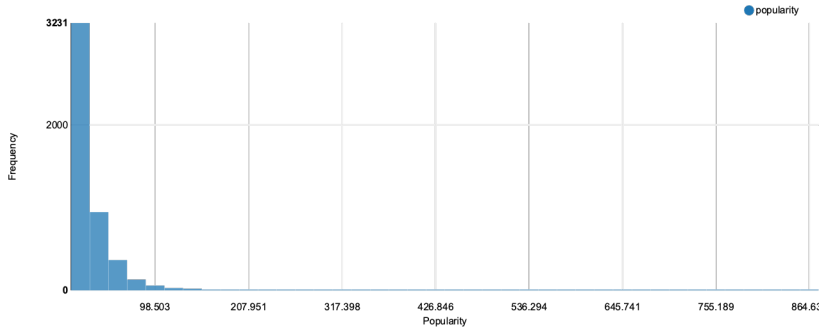
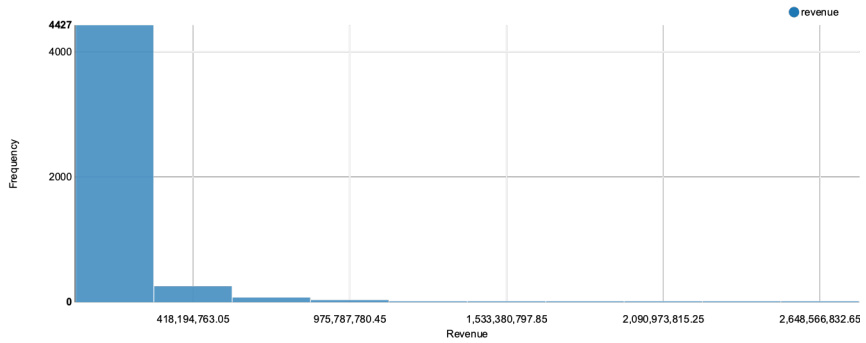
Introduction

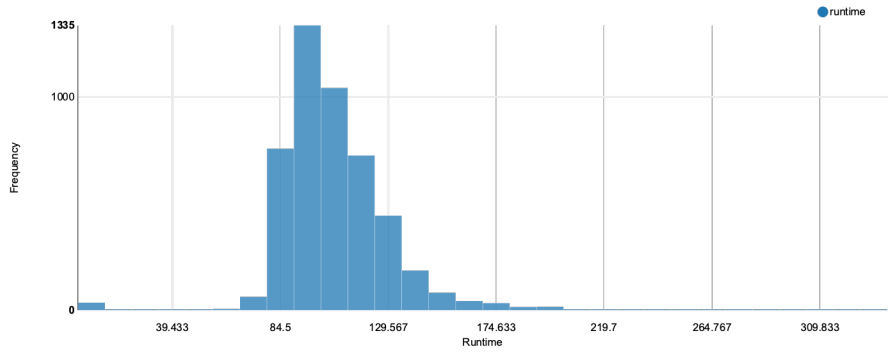
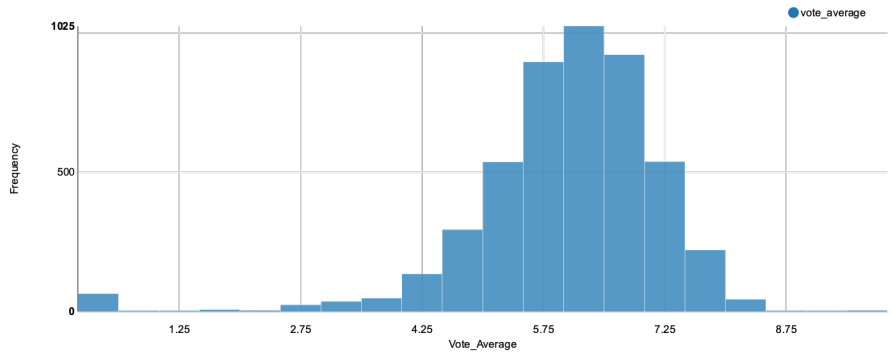
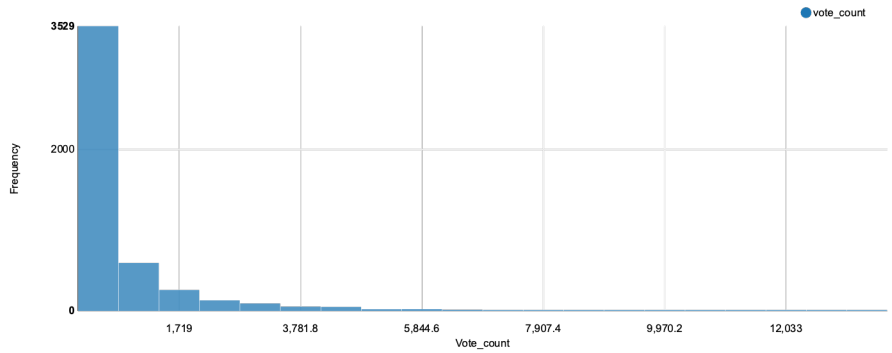
In this project, we are taking on the role of the **Data Science Team** of a movie production company. We want to predict if a movie will be profitable, given a variety of factors: popularity of actors, budget, country of production, etc.

Production companies are responsible for funding movies which they believe will return a profit. This decision is discretionary, which leads to potential risks of human error and bias. In order to minimize risk and remove the possibility of human error when it comes to assessing potential profitability of a proposed movie, we constructed a model designed to analyse the features of a proposed movie and assess its potential profitability.

The data for this project is collected from The Movie Database (TMDb), source - Kaggle. The data is in two CSV files. The main dataset contains 5000 movies and 20 columns, while the second file contains information on movie cast and credits. The process of observing the dataset, manipulating it, and building a predictive model is described in the following slides.

Univariate Analysis of Numerical Variables

Variable	Descriptive Statistics	Detail	Insight
budget	Numerical, continuous **Mean:** 29m, **Std. Deviation:** 40.7m **Quantiles:** 0 Min 780k 25% 15m 50% 40m 75% 380m Max		The number of high-budget movies is lower than the number of low-budget movies, meaning high-budget movies are a rarity.
popularity	Numerical, continuous **Mean:** 21.5, **Std. Deviation: 31.8** **Quantiles:** 0 Min 4.67 25% 12.9 50% 28.4 75% 876 Max		This variable shows how popular a movie is based on the views of pages related to it. This score is calculated after movie publishment date.
revenue	Numerical, continuous **Mean:** 82.3m **Std. Deviation:** 163m **Quantiles:** 0 Min 0 25% 19.2m 50% 92.9m 75% 2.79b Max		Total revenue resembles an exponential distribution. The number of high-grossing movies is significantly lower than the number of low-grossing ones.

Variable	Nature	Detail	Insight
runtime	Numerical, continuous Mean: 107 Std. Deviation: 22.6 Quantiles: 0 Min 94 25% 103 50% 118 75% 338 Max		Most movies have a runtime between 85 and 130 minutes. The median runtime is equal to 103 minutes. Recent movies often have a longer runtime.
vote_average	Numerical, continuous Mean: 6.09 Std. Deviation: 1.19 Quantiles: 0 Min 5.6 25% 6.2 50% 6.8 75% 10 Max		The distribution of average votes resembles a normal distribution. The median - 6.2 implies the votes are slightly left-skewed, otherwise the median would be ~5. This variable is recorded after publishment.
vote_count	Numerical, continuous Mean: 690 Std. Deviation: 1.23k Quantiles: 0 Min 54 25% 235 50% 737 75% 13.8k Max		The distribution resembles an exponential one. The majority of movies have less than 1000 votes. This means only a few movies have a high audience interaction. This variable is recorded after publishment.

Univariate Analysis of Categorical Variables

GENRE

- 20 unique genres
- Most common: Drama

LANGUAGE

- 37 unique languages
- Most common: English
- Few movies in other languages

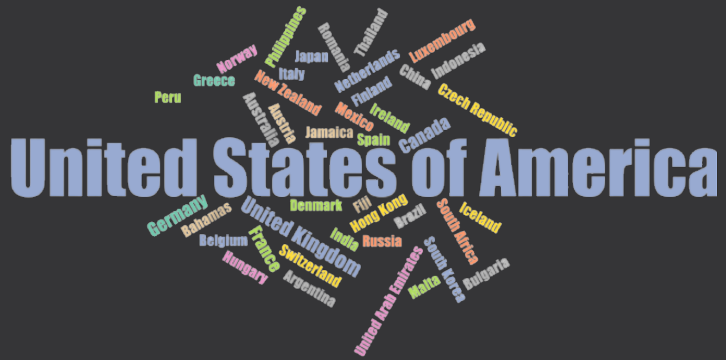
PRODUCTION COMPANY

- 1313 unique companies
- Most common: Paramount Pictures

Univariate Analysis of Categorical Variables

PRODUCTION COUNTRY

- 70 unique countries
- Most common: USA

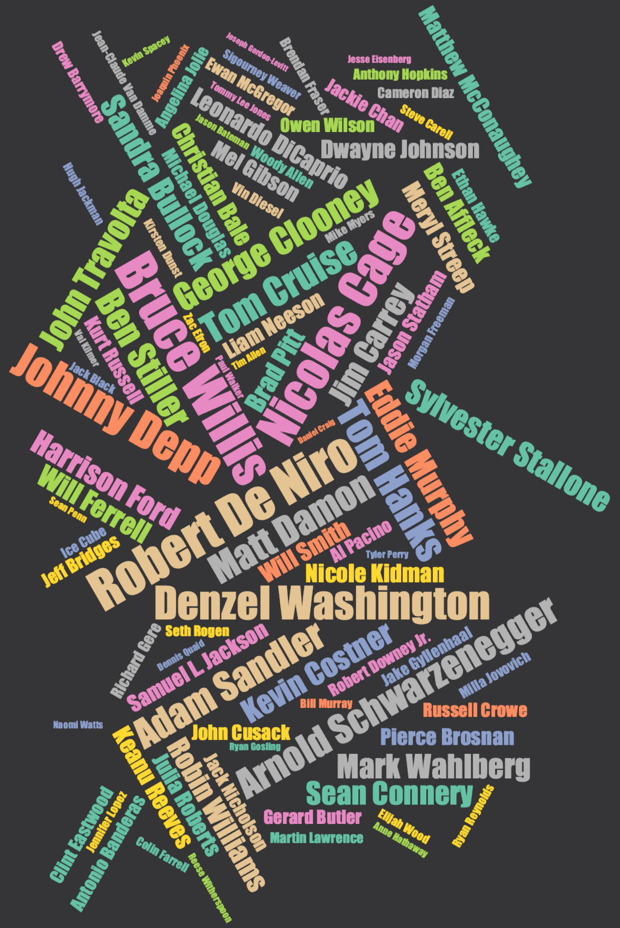


RELEASE MONTH

- Most common: September

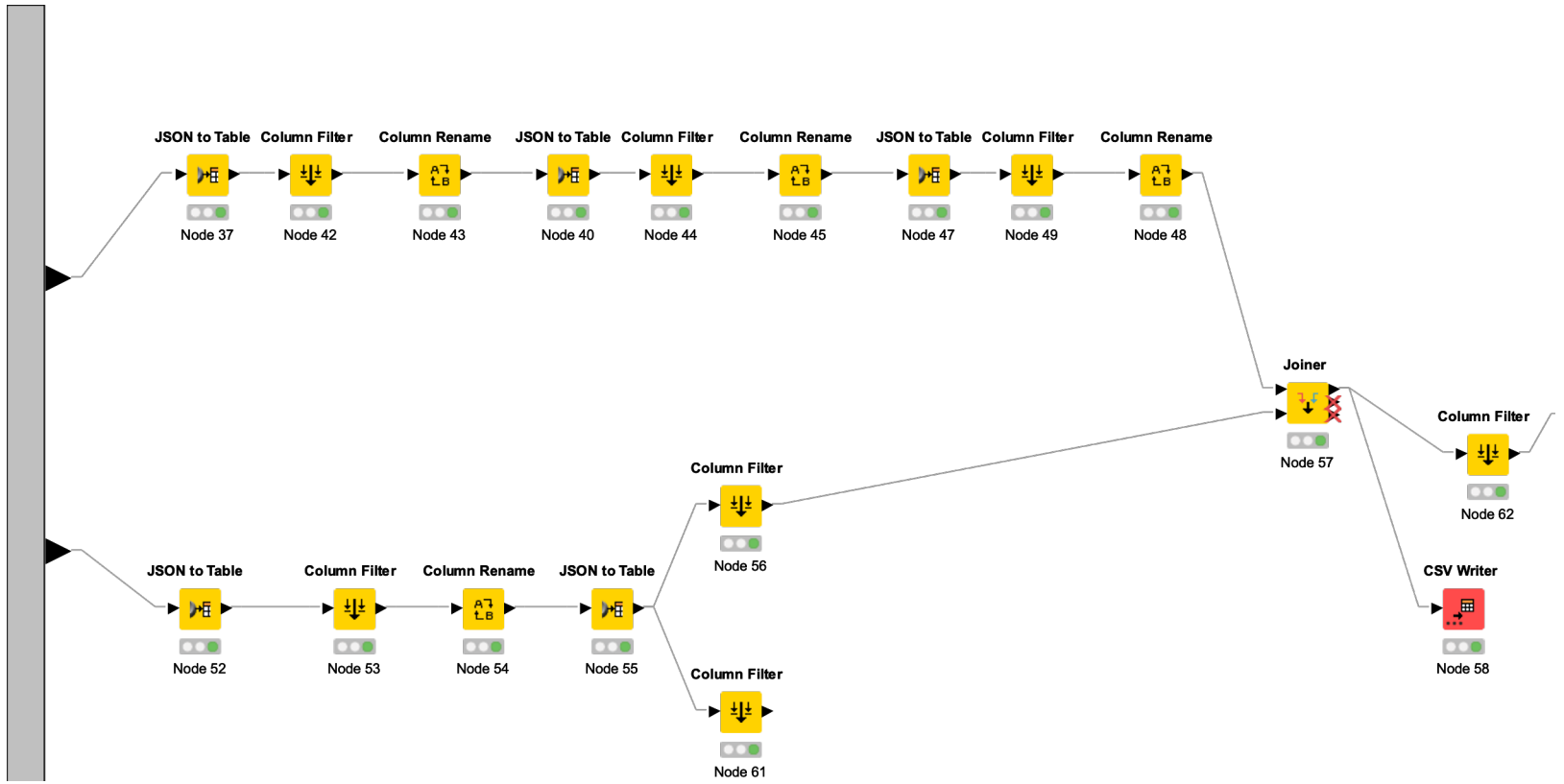
ACTOR 1

- 2095 unique actors
- Most common: Robert De Niro



DATA PREPROCESSING – JSON MANAGEMENT

- Since in both datasets some of the columns are formatted in JSON, we manipulate them to extract single information and keep only:
  - Main genre (**genre_1**)
  - Main actors (**actor_1**, **actor_2**, **actor_3, actor_4, actor_5**)
  - Production company
  - Main production country



DATA PREPROCESSING – FILTERING

We only care about movie markets that are quite 'big', so we are going to filter a few rows that are not relevant in the moviemaking industry

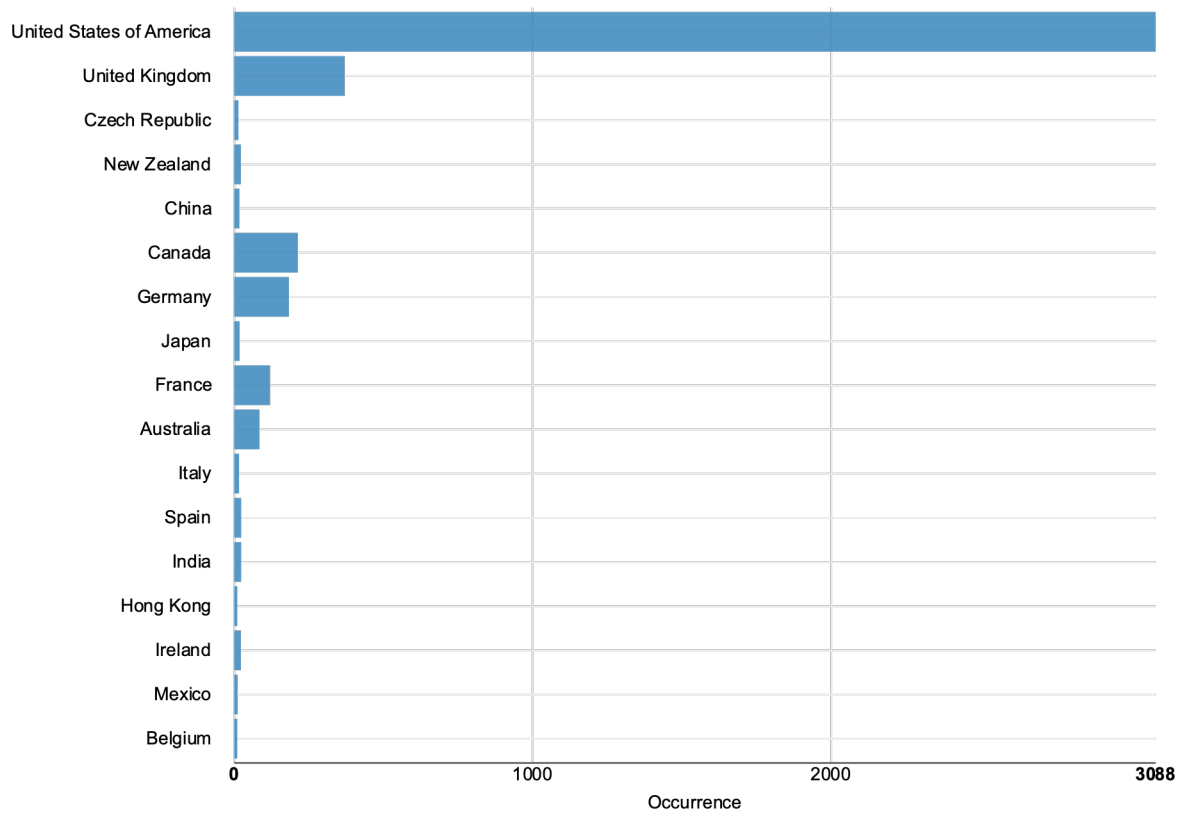
Production Countries:

When it comes to production countries, we remove:

- Countries with less than 10 samples
- Observations with unknown production country

Production Language:

Since the dataset contains few data points for movies in foreign languages, we only kept movies with English as the main language, which are 93.7% of the rows, so we are not losing much data.



DATA PREPROCESSING – OUTLIERS & MISSING VALUES FILTERING

OUTLIERS:

We decided to keep them in the working dataset since the model obtains better results with them included. There is a trade-off between having outliers and losing information. In our model, losing 20% of the dataset due to outliers yields lower results than keeping the outliers.

MISSING VALUES:

We have several movies with budget = 0 and revenues = 0 which is not realistic since it would imply that actors are not being paid and no one is going to see the movie. It also makes it impossible to calculate the profitability variable we created.

The values are probably null because the real budget and revenues are unknown. So, we consider them as missing values.

We remove a movie from the working dataset if it has either:

- Budget = 0
- Revenue = 0

We remove these observations rather than replacing the missing values with budget/revenue estimates/predictions obtained through an auxiliary estimation model.

The reasoning behind this is that replacing a missing budget/revenue value with, for example, the mean of the other observations can severely misrepresent the movie and lead to inconsistencies in the model. In addition, after removing observations with missing values we are left with a considerable amount of data points which are enough to obtain significant results.

FEATURE ENGINEERING – PROFITABLE MOVIE

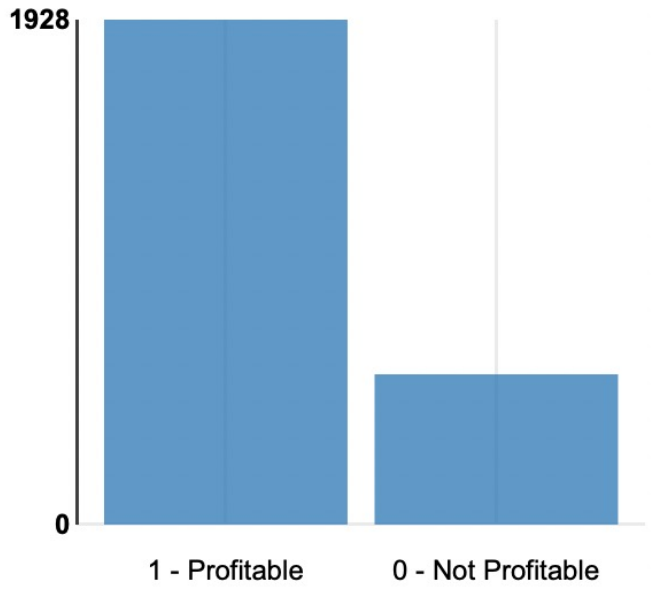
PROFITABLE MOVIE DUMMY:

A movie production company is interested in knowing if a movie will generate positive net profits, relative to the production budget.

$$\text{Profitable} = \begin{cases} 0 & \text{if} \quad \text{Revenue} - \text{Budaet} < 0 \\ 1 & \text{if} \quad \text{Revenue} - \text{Budget} > 0 \end{cases}$$

As we can see from the bar chart on the right, most movies are profitable, and only 22 % of them are not profitable.
This is not a perfect representation of the industry as production companies tend to produce an even number of profitable and not profitable movies. In our dataset this is not the case because most of the movies are well-known ones.

To deal with the unbalanced data (78% / 22%) we decided to oversample the minority class, to train a more sensitive model.

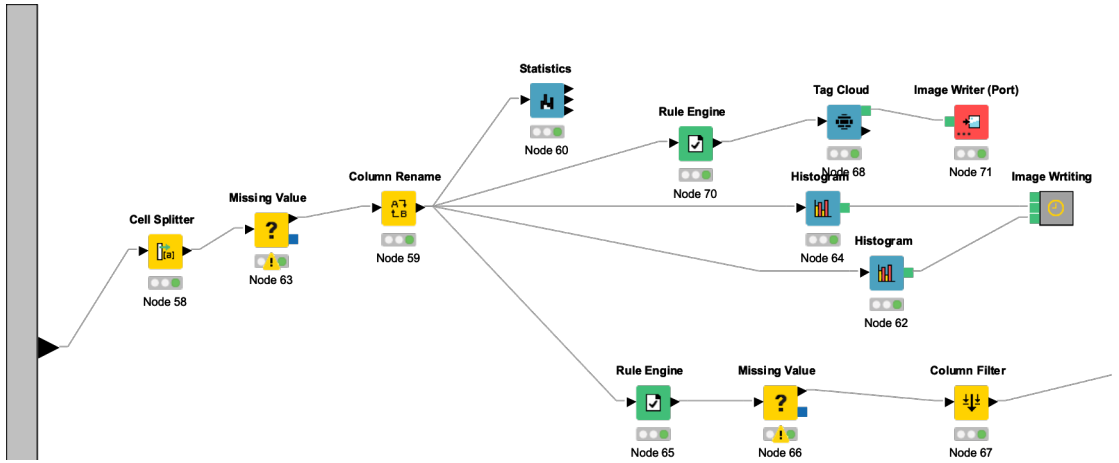


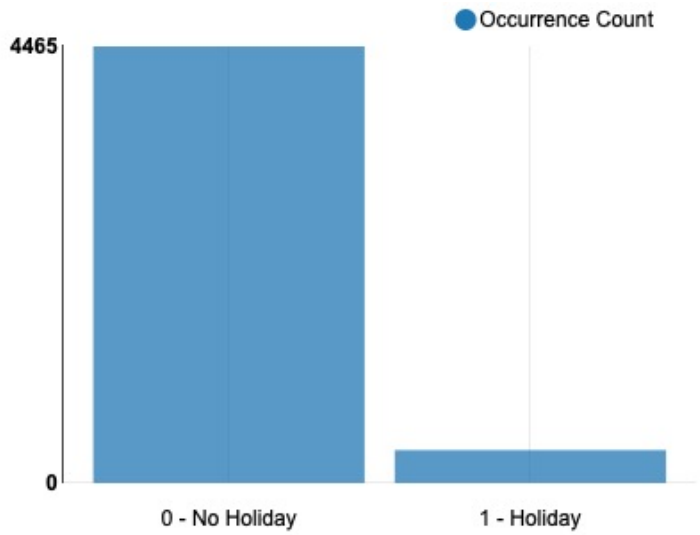
FEATURE ENGINEERING – HOLIDAY DUMMY FOR DATES

- **HOLIDAY DUMMY:**
  - We split the release date column into day, month and year.
  - We use this data to create a Holiday dummy variable which takes value = 1 if the movie was released during the holiday period (December 15th – January 7th).

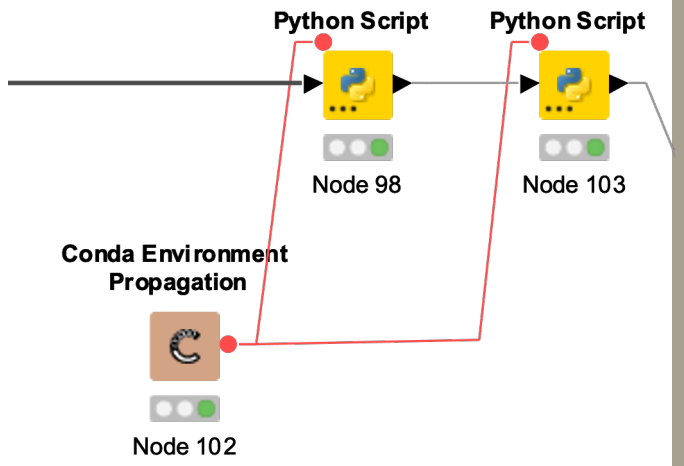
$$\text{Holiday} = \begin{cases} 1 & \text{if date in (DEC 15, JAN 7)} \\ 0 & \text{otherwise} \end{cases}$$

  - We hope this variable will help us capture the effect of releasing a movie during the holiday period and how that affects the movie's box office success.





FEATURE ENGINEERING – CAST QUALITY INDICES



Given the number of unique actors we have and the number of actors participating in each movie, we cannot perform One-Hot Encoding for these categorical variables. This is the reason why we compute numerical variables which represent the importance of actors in some way.

ACTOR POPULARITY given past movies

We create this variable for each actor in the cast. It is the average of the popularity variable of all movies an actor has taken part in. This can prove insightful for the model because actors who participate in popular movies usually bring a lot of attention to them so people would watch a movie for the sole purpose of seeing a particular actor in it. That's why it could affect revenues, thus profitability.

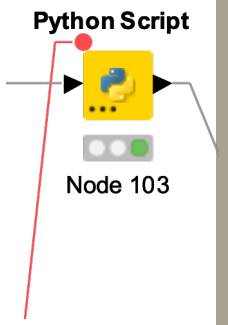
ACTOR BUDGET given past movies

We create this variable for each actor in order to have a numerical variable that reflects how much money were spent on projects an actor took part in. It is the average of the budgets of all movies an actor took part in. Actors who participate in high budget movies are usually popular ones, so they bring more customers to the production which in turn generates higher revenues.

ACTOR REVENUES given past movies

We wanted to have a numerical variable also measuring if an actor has been in the cast of high-revenues movies before the production of the movie considered. We computed the average of the gross income of the movies in which the actor played a role in the past, relative to the movie considered.

FEATURE ENGINEERING – PRODUCTION COMPANY LEVEL



Given the amount of unique Production Companies in the dataset, it would be computationally infeasible to perform One-Hot Encoding, so we compute numerical variables that represent the strength of each Production Company.

PRODUCTION COMPANY POPULARITY given past movies

This is a numerical variable which represents how popular a production company is. We take the average of the popularity indices of all movies produced by a company. This variable can prove insightful for the model since the degree of popularity is an important factor for revenue generation, thus profits.

PRODUCTION COMPANY BUDGET given past movies

This is a numerical variable which represents, on average, how much money is spent on a movie by each production company. We take the average of the budgets of all movies produced by a company. How much a production company spends on a movie is important for revenue generation because higher investments usually result in higher payoffs.

PRODUCTION COMPANY past movies REVENUES given past movies

We create this variable for each production company in order to have a numerical variable that considers average revenues of the movies produced by a production company before the year in which the proposed movie was produced

Bivariate Analysis – Correlation Matrix

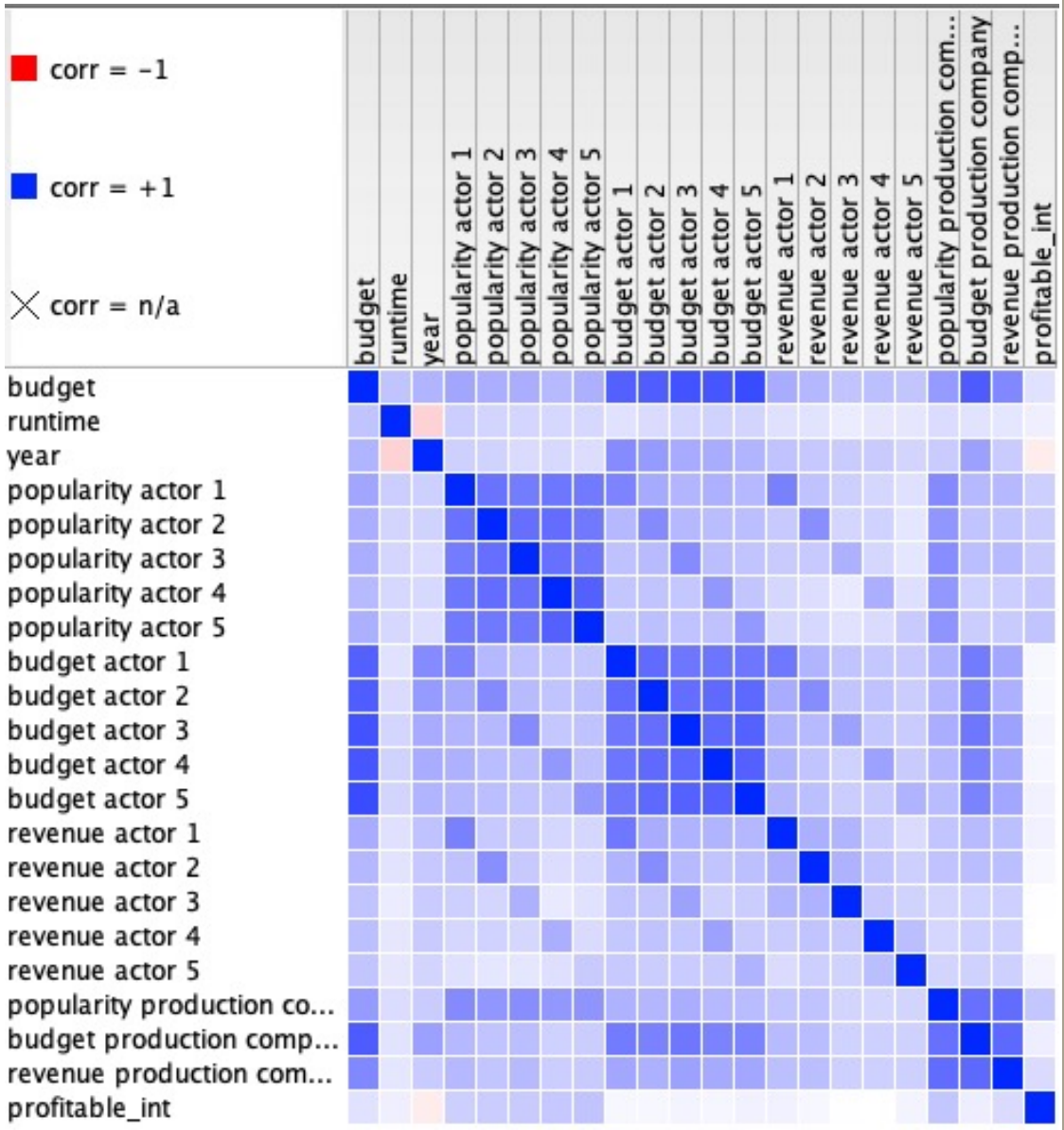
To understand the relationships between the numerical variables we plotted a correlation matrix.

We had to convert the binary dependent variable into a numerical variable, since Pearson correlation is only defined between numerical variables. We could do that since a binary categorical variable can be interpreted as numerical (impossible for multi-class categorical variable).

As we see from the profitable row, the dependent variable has very low correlation coefficients, meaning that the other variables are not very informative on profitability.

The most related variables are popularity variables relative to cast members and production company (around 0.2)

Row ID	budget	runtime	year	popularit...	popularit...	popularit...	popularit...	popularit...	budget a...	budget a...	budget a...	budget a...	budget a...	revenue a...	revenue a...	revenue a...	revenue a...	revenue a...	popularit...	budget p...	revenue ...	profitable...
budget	1	0.24	0.301	0.361	0.32	0.324	0.278	0.318	0.653	0.668	0.713	0.696	0.75	0.33	0.285	0.24	0.253	0.23	0.402	0.68	0.48	0.116
runtime	0.24	1	-0.163	0.196	0.167	0.163	0.157	0.155	0.117	0.145	0.165	0.176	0.168	0.12	0.112	0.073	0.097	0.096	0.142	0.112	0.103	0.062
year	0.301	-0.163	1	0.19	0.177	0.147	0.151	0.137	0.465	0.403	0.337	0.335	0.303	0.244	0.221	0.202	0.207	0.168	0.205	0.383	0.204	-0.072
popularity act...	0.361	0.196	0.19	1	0.573	0.535	0.555	0.54	0.501	0.343	0.296	0.312	0.283	0.514	0.244	0.191	0.161	0.121	0.47	0.284	0.283	0.19
popularity act...	0.32	0.167	0.177	0.573	1	0.588	0.597	0.546	0.282	0.471	0.282	0.267	0.261	0.212	0.453	0.165	0.179	0.102	0.419	0.24	0.23	0.202
popularity act...	0.324	0.163	0.147	0.535	0.588	1	0.582	0.55	0.245	0.274	0.466	0.261	0.239	0.206	0.213	0.315	0.16	0.104	0.457	0.26	0.277	0.206
popularity act...	0.278	0.157	0.151	0.555	0.597	0.582	1	0.652	0.229	0.239	0.225	0.416	0.231	0.162	0.133	0.081	0.323	0.124	0.416	0.183	0.197	0.219
popularity act...	0.318	0.155	0.137	0.54	0.546	0.55	0.652	1	0.225	0.254	0.242	0.247	0.41	0.161	0.156	0.116	0.145	0.209	0.43	0.2	0.209	0.238
budget actor 1	0.653	0.117	0.465	0.501	0.282	0.245	0.229	0.225	1	0.607	0.554	0.561	0.553	0.55	0.302	0.239	0.225	0.214	0.308	0.537	0.356	0.031
budget actor 2	0.668	0.145	0.403	0.343	0.471	0.274	0.239	0.254	0.607	1	0.593	0.611	0.614	0.328	0.461	0.233	0.244	0.2	0.292	0.508	0.315	0.032
budget actor 3	0.713	0.165	0.337	0.296	0.282	0.466	0.225	0.242	0.554	0.593	1	0.615	0.646	0.305	0.28	0.376	0.224	0.203	0.327	0.553	0.377	0.043
budget actor 4	0.696	0.176	0.335	0.312	0.267	0.261	0.416	0.247	0.561	0.611	0.615	1	0.656	0.303	0.241	0.183	0.38	0.207	0.281	0.507	0.34	0.036
budget actor 5	0.75	0.168	0.303	0.283	0.261	0.239	0.231	0.41	0.553	0.614	0.646	0.656	1	0.288	0.256	0.201	0.21	0.305	0.269	0.507	0.355	0.051
revenue actor 1	0.33	0.12	0.244	0.514	0.212	0.206	0.162	0.161	0.55	0.328	0.305	0.303	0.288	1	0.317	0.302	0.202	0.146	0.235	0.273	0.253	0.054
revenue actor 2	0.285	0.112	0.221	0.244	0.453	0.213	0.133	0.156	0.302	0.461	0.28	0.241	0.256	0.317	1	0.31	0.239	0.193	0.244	0.267	0.26	0.031
revenue actor 3	0.24	0.073	0.202	0.191	0.165	0.315	0.081	0.116	0.239	0.233	0.376	0.183	0.201	0.302	0.31	1	0.212	0.178	0.168	0.182	0.188	0.002
revenue actor 4	0.253	0.097	0.207	0.161	0.179	0.16	0.323	0.145	0.225	0.244	0.224	0.38	0.21	0.202	0.239	0.212	1	0.258	0.162	0.187	0.188	-0
revenue actor 5	0.23	0.096	0.168	0.121	0.102	0.104	0.124	0.209	0.214	0.2	0.203	0.207	0.305	0.146	0.193	0.178	0.258	1	0.168	0.182	0.187	0.045
popularity pro...	0.402	0.142	0.205	0.47	0.419	0.457	0.416	0.43	0.308	0.292	0.327	0.281	0.269	0.235	0.244	0.168	0.162	0.168	1	0.583	0.606	0.228
budget produ...	0.68	0.112	0.383	0.284	0.24	0.26	0.183	0.2	0.537	0.508	0.553	0.507	0.507	0.273	0.267	0.182	0.187	0.182	0.583	1	0.614	0.067
revenue produ...	0.48	0.103	0.204	0.283	0.23	0.277	0.197	0.209	0.356	0.315	0.377	0.34	0.355	0.253	0.26	0.188	0.188	0.187	0.606	0.614	1	0.144
profitable_int	0.116	0.062	-0.072	0.19	0.202	0.206	0.219	0.238	0.031	0.032	0.043	0.036	0.051	0.054	0.031	0.002	-0	0.045	0.228	0.067	0.144	1



Bivariate Analysis – One-Way ANOVA

- Grouping based on profitable variable:
  - Profitable movies
  - Not profitable movies
- We are going to test if difference in mean for the two groups is statistically significant
- The node also performs Levene's test for equality of variance in the groups

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
budget	Between Groups	7.48E16	1	7.48E16	34.131	5.82E-9
budget	Within Groups	5.52E18	2515	2.19E15		
budget	Total	5.59E18	2516			
runtime	Between Groups	4,274.5675	1	4,274.5675	9.557	0.002
runtime	Within Groups	1,124,881.9482	2515	447.2692		
runtime	Total	1,129,156.5157	2516			
year	Between Groups	2,267.3046	1	2,267.3046	13.0774	0.0003
year	Within Groups	436,038.7129	2515	173.3752		
year	Total	438,306.0175	2516			
popularity actor 1	Between Groups	26,399.2679	1	26,399.2679	94.6332	0.0
popularity actor 1	Within Groups	701,594.7298	2515	278.9641		
popularity actor 1	Total	727,993.9977	2516			
popularity actor 2	Between Groups	38,494.4544	1	38,494.4544	106.8917	0.0
popularity actor 2	Within Groups	905,715.9547	2515	360.1256		
popularity actor 2	Total	944,210.4092	2516			
popularity actor 3	Between Groups	48,244.7044	1	48,244.7044	111.8455	0.0
popularity actor 3	Within Groups	1,084,848.7535	2515	431.3514		
popularity actor 3	Total	1,133,093.4579	2516			
popularity actor 4	Between Groups	52,744.6629	1	52,744.6629	127.09	0.0
popularity actor 4	Within Groups	1,043,770.4369	2515	415.0181		
popularity actor 4	Total	1,096,515.0998	2516			
popularity actor 5	Between Groups	69,889.7728	1	69,889.7728	151.4659	0.0
popularity actor 5	Within Groups	1,160,477.7862	2515	461.4226		
popularity actor 5	Total	1,230,367.559	2516			
budget actor 1	Between Groups	1.62E15	1	1.62E15	2.3599	0.1246
budget actor 1	Within Groups	1.73E18	2515	6.88E14		
budget actor 1	Total	1.73E18	2516			
budget actor 2	Between Groups	1.89E15	1	1.89E15	2.6167	0.1059
budget actor 2	Within Groups	1.82E18	2515	7.24E14		
budget actor 2	Total	1.82E18	2516			
budget actor 3	Between Groups	3.94E15	1	3.94E15	4.7084	0.0301
budget actor 3	Within Groups	2.10E18	2515	8.37E14		
budget actor 3	Total	2.11E18	2516			
budget actor 4	Between Groups	2.44E15	1	2.44E15	3.2909	0.0698
budget actor 4	Within Groups	1.86E18	2515	7.41E14		
budget actor 4	Total	1.87E18	2516			
budget actor 5	Between Groups	5.92E15	1	5.92E15	6.4848	0.0109
budget actor 5	Within Groups	2.30E18	2515	9.13E14		
budget actor 5	Total	2.30E18	2516			
revenue actor 1	Between Groups	9.01E16	1	9.01E16	7.4321	0.0065
revenue actor 1	Within Groups	3.05E19	2515	1.21E16		
revenue actor 1	Total	3.06E19	2516			
revenue actor 2	Between Groups	2.78E16	1	2.78E16	2.3942	0.1219
revenue actor 2	Within Groups	2.92E19	2515	1.16E16		
revenue actor 2	Total	2.92E19	2516			
revenue actor 3	Between Groups	1.18E14	1	1.18E14	0.0104	0.9189
revenue actor 3	Within Groups	2.86E19	2515	1.14E16		
revenue actor 3	Total	2.86E19	2516			
revenue actor 4	Between Groups	1.54E12	1	1.54E12	0.0002	0.9891
revenue actor 4	Within Groups	2.08E19	2515	8.28E15		
revenue actor 4	Total	2.08E19	2516			
revenue actor 5	Between Groups	4.14E16	1	4.14E16	5.0737	0.0244
revenue actor 5	Within Groups	2.05E19	2515	8.15E15		
revenue actor 5	Total	2.05E19	2516			
popularity production company	Between Groups	28,617.1512	1	28,617.1512	137.8008	0.0
popularity production company	Within Groups	522,291.2818	2515	207.6705		
popularity production company	Total	550,908.433	2516			
budget production company	Between Groups	7.83E15	1	7.83E15	11.2854	0.0008
budget production company	Within Groups	1.74E18	2515	6.94E14		
budget production company	Total	1.75E18	2516			
revenue production company	Between Groups	4.65E17	1	4.65E17	52.9443	4.56E-13
revenue production company	Within Groups	2.21E19	2515	8.79E15		
revenue production company	Total	2.26E19	2516			

Bivariate Analysis – Categorical Variables

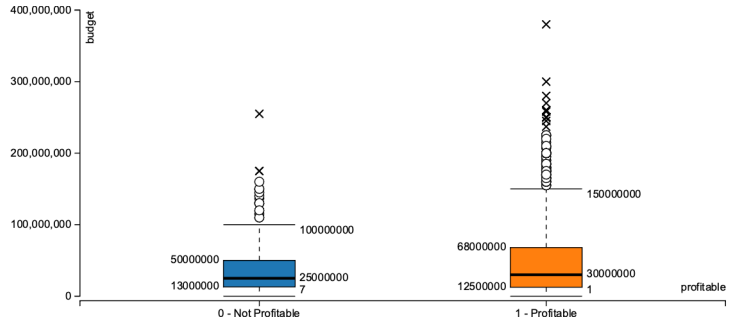
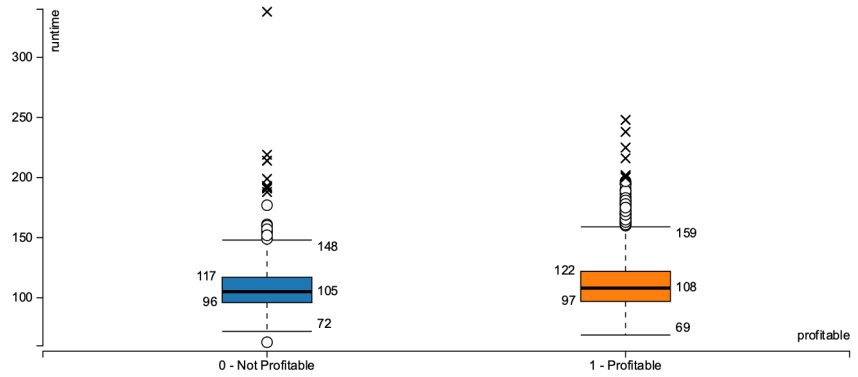
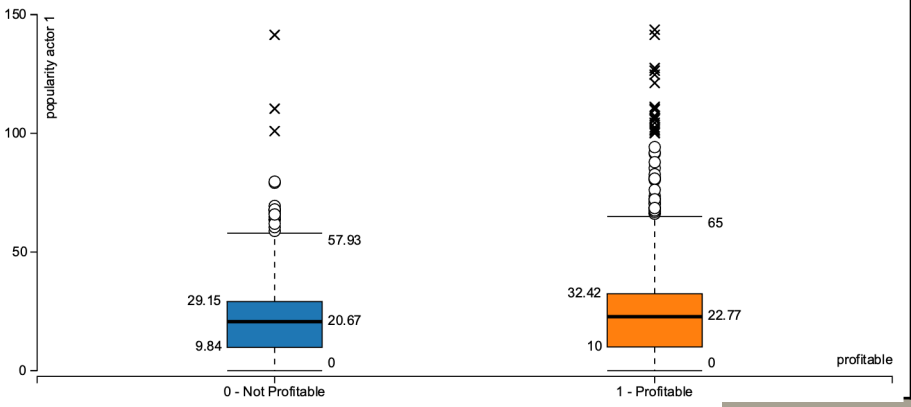
Categorical Variable	Cross Tabulation	Cramer's V	Chi-square test
genre	0 - Not Profitable 1 - Profitable Action 138 394 Adventure 50 205 Animation 14 76 Comedy 102 316 Crime 35 98 Documentary 0 6 Drama 139 356 Family 7 29 Fantasy 11 69 Foreign 1 0 History 3 13 Horror 19 136 Music 5 13 Mystery 6 20 Romance 13 51 Science Fiction 5 61 Thriller 23 80 War 9 7 Western 2 5	> CramerV(tb) [1] 0.1469733	Pearson's Chi-squared test data: data$genre_1 and data$profitable X-squared = 54.37, df = 18, p-value = 1.607e-05
production country	0 - Not Profitable 1 - Profitable Australia 15 37 Belgium 3 2 Canada 34 73 China 0 13 Czech Republic 3 7 France 23 56 Germany 41 77 Hong Kong 1 5 India 2 9 Ireland 5 8 Italy 5 6 Japan 2 8 Mexico 3 5 New Zealand 3 12 Spain 3 8 United Kingdom 42 186 United States of America 397 1423	> CramerV(tb2) [1] 0.1177408	Pearson's Chi-squared test data: data$production_country and data$profitable X-squared = 34.893, df = 16, p-value = 0.004111

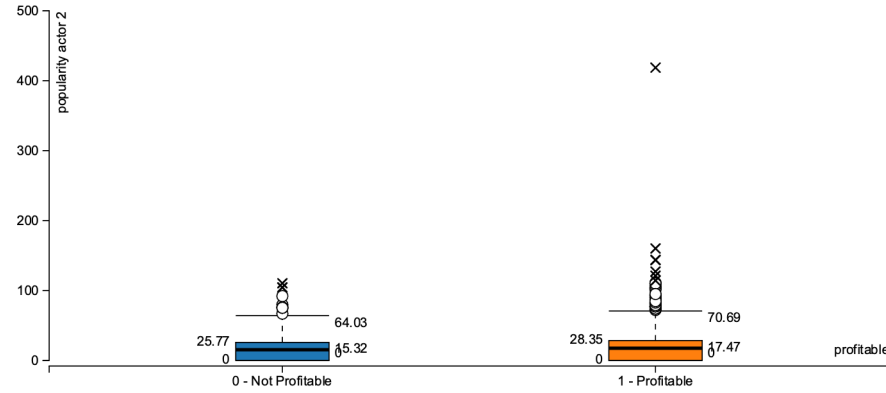
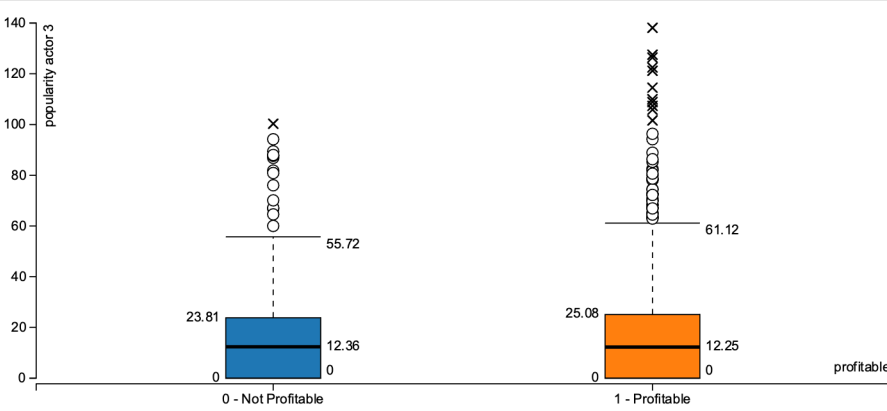
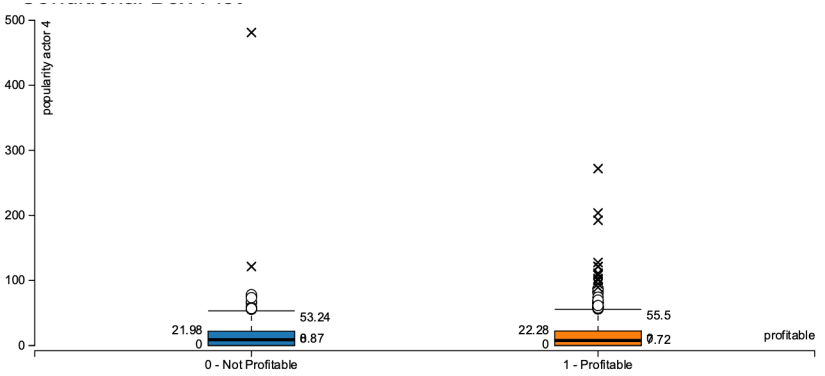
Categorical Variable	Cross Tabulation	Cramer's V	Chi-square test
holiday	0 - Not Profitable 1 - Profitable 0 553 1815 1 29 120	> CramerV(tb3) [1] 0.02177341	Pearson's Chi-squared test data: data$holiday and data$profitable X-squared = 1.1933, df = 1, p-value = 0.2747
year	0 - Not Profitable 1 - Profitable 1985 2 12 1986 3 12 1987 1 18 1988 2 16 1989 0 19 1990 2 19 1991 5 15 1992 4 19 1993 9 21 1994 9 26 1995 15 35 1996 14 42 1997 17 46 1998 21 52 1999 25 61 2000 21 54 2001 31 67 2002 34 79 2003 21 58 2004 30 72 2005 29 83 2006 35 85 2007 22 75 2008 29 82 2009 33 85 2010 32 94 2011 28 100 2012 18 88 2013 18 92 2014 17 82 2015 20 76 2016 8 40	> CramerV(tb4) [1] 0.183546	Pearson's Chi-squared test data: data$year and data$profitable X-squared = 84.796, df = 85, p-value = 0.4858

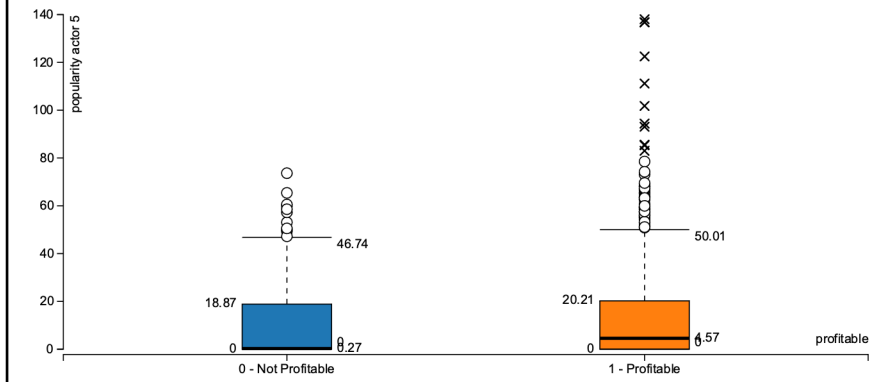
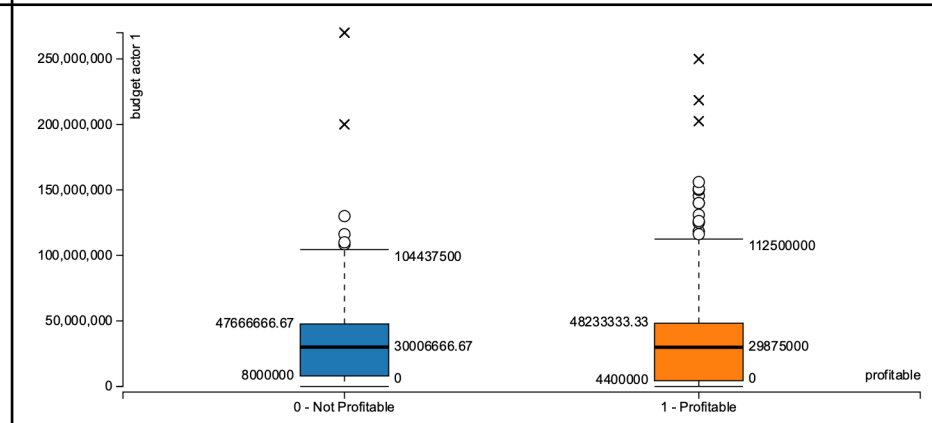
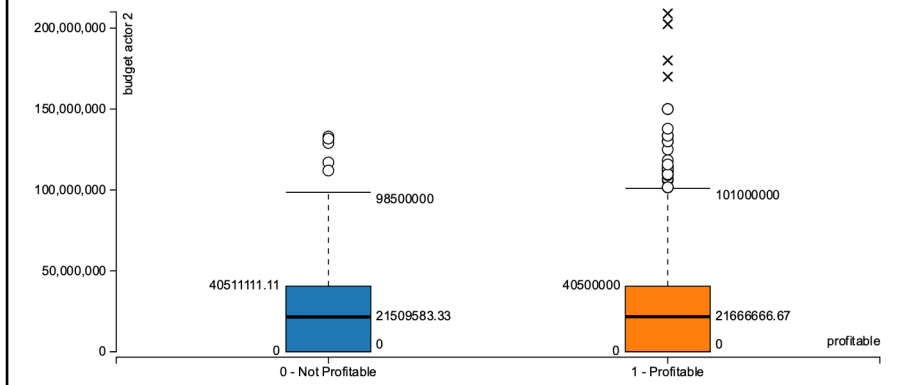
Bivariate Analysis – Categorical

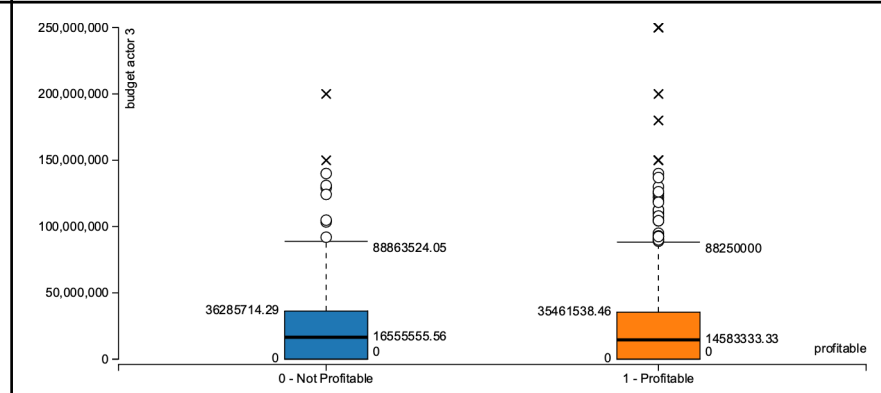
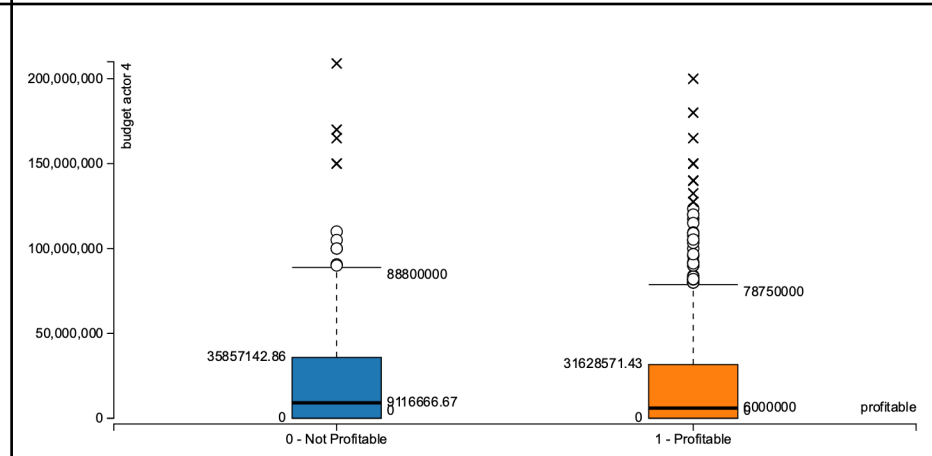
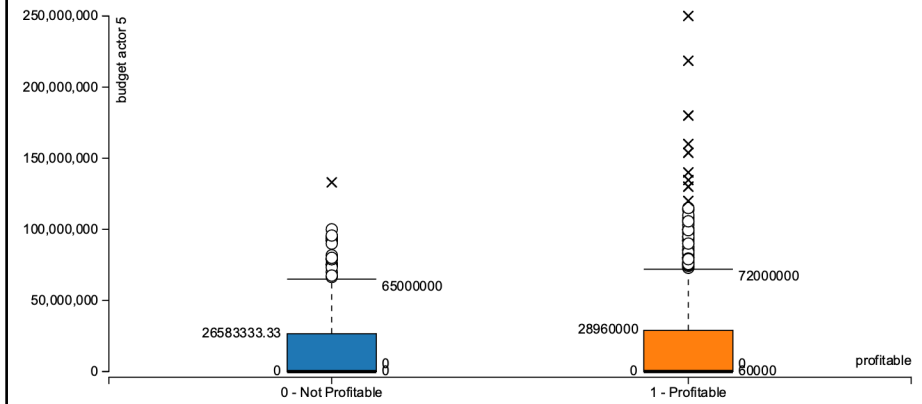
- There is a significant relationship between **profitability** and the variables **genre** and **production country** since the p-value from the Chi-square test are near zero and 0.0041, respectively
- However, Cramer's V is roughly 0.15 for **genre** and 0.12 for **production country** meaning the relationship is not very strong and its impact will be limited
- Unfortunately, the **holiday** dummy we created to inspect the effect of holidays on box office success is not significant since its p-value is 0.27 and Cramer's V is 0.02
- The **year** variable is also not significant with a p-value of 0.49, but it has a Cramer's V of 0.18 which is the highest among the categorical variables
- From the bivariate analysis of categorical variables, we can conclude that none of them will have a strong effect on our model. Specifically, the **genre** and **production country** variables both have a low Cramer's V, while the **holiday** and **year** variables are not significant.
- Nonetheless, in our model we will include the **genre** and **production country** variables while omitting **holiday** and **year**

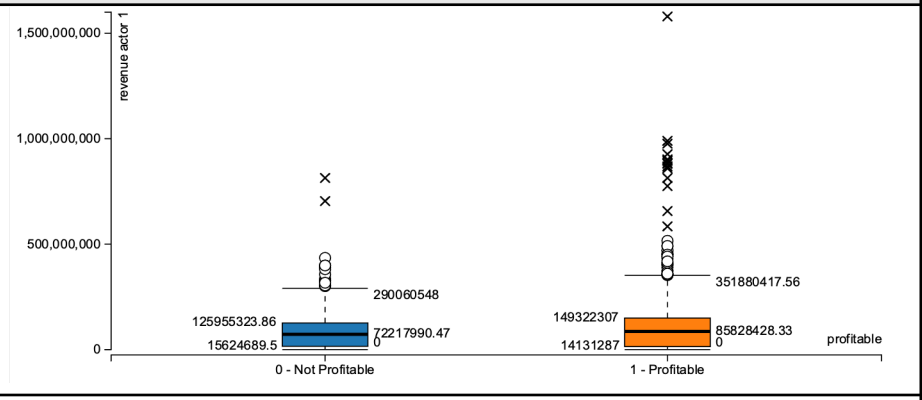
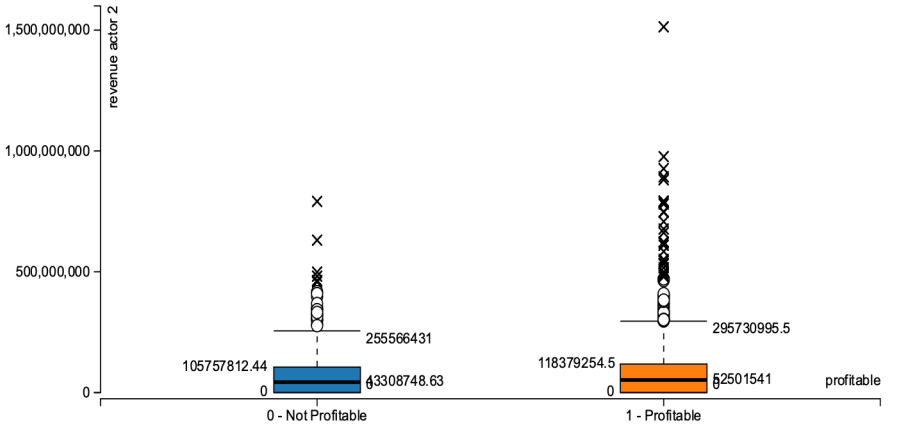
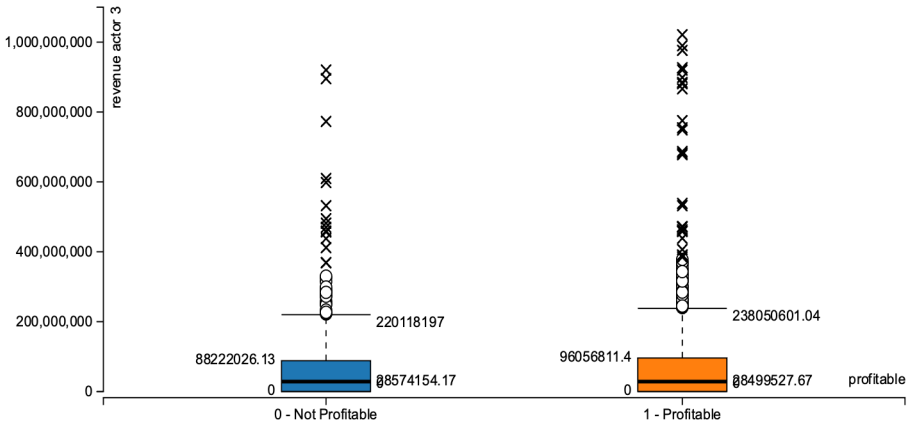
Bivariate Analysis – Continuous Variables

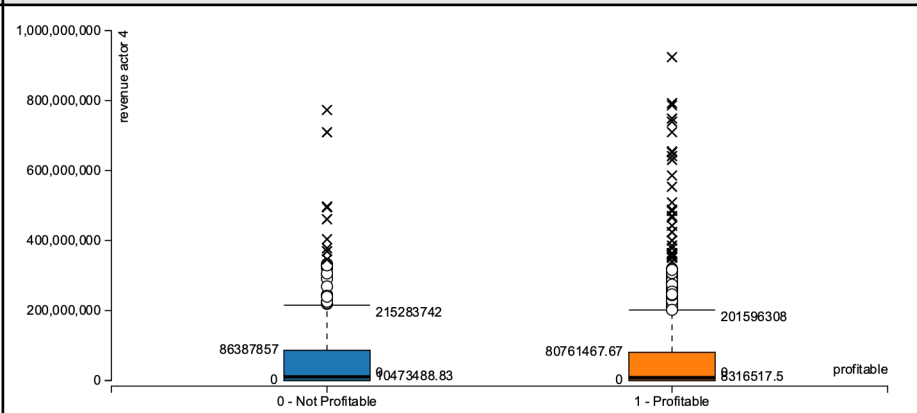
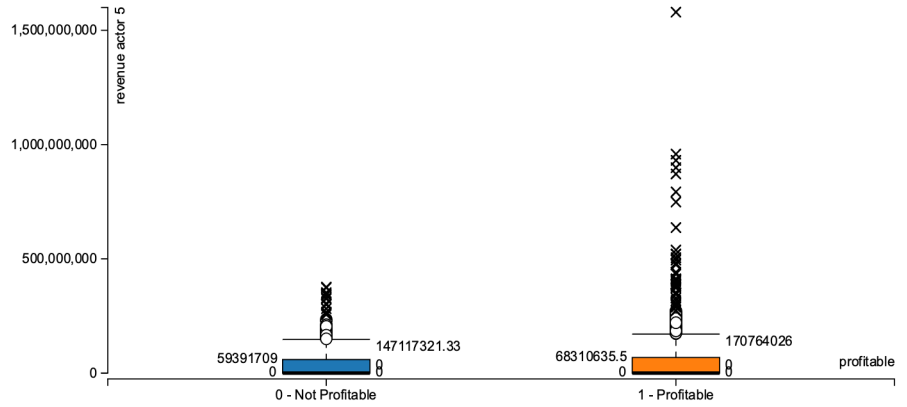
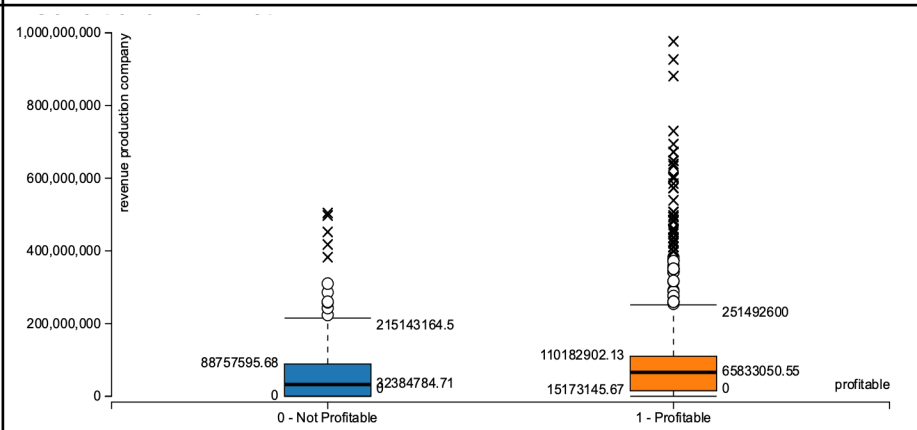
Continuous Variable	T-test	Conditional Boxplot
budget	Welch Two Sample t-test data: data$budget by data$profitable t = -7.4681, df = 1565.8, p-value = 1.345e-13 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -16331058 -9536872 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 35487077 48421042	
runtime	Welch Two Sample t-test data: data$runtime by data$profitable t = -3.0482, df = 937.04, p-value = 0.002367 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -5.080923 -1.100888 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 108.5137 111.6047	
popularity actor 1	Welch Two Sample t-test data: data$popularity.actor.1 by data$profitable t = -2.246, df = 1053.4, p-value = 0.02491 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3.4274853 -0.2311151 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 21.3351 23.1644	

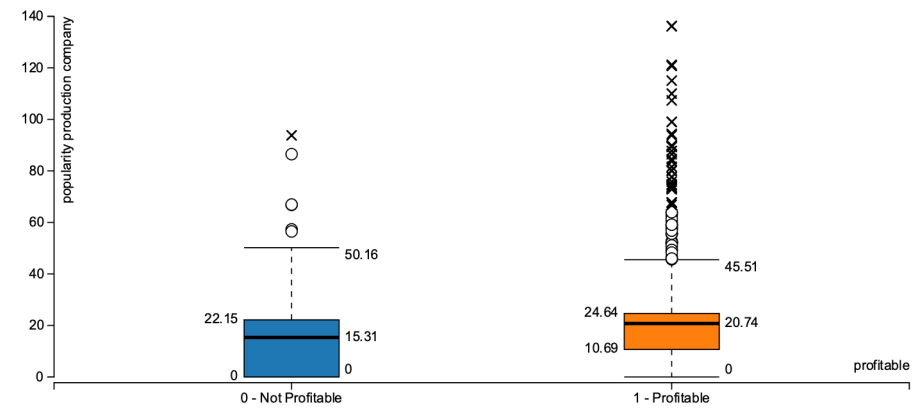
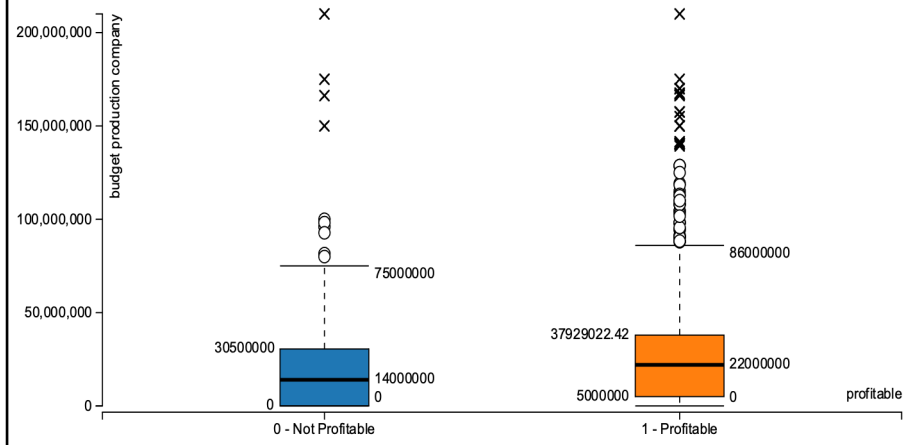
Cotinuous Variable	T-test	Conditional Boxplot
popularity actor 2	Welch Two Sample t-test data: data$popularity.actor.2 by data$profitable t = -2.7275, df = 1197.6, p-value = 0.006474 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3.9484875 -0.6446242 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 17.11816 19.41472	
popularity actor 3	Welch Two Sample t-test data: data$popularity.actor.3 by data$profitable t = -0.40845, df = 1029.5, p-value = 0.683 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -1.884395 1.235069 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 15.31753 15.64219	
popularity actor 4	Welch Two Sample t-test data: data$popularity.actor.4 by data$profitable t = 0.22406, df = 792.77, p-value = 0.8228 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -1.930655 2.428204 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 13.67097 13.42219	

Categorical Variable	T-test	Conditional Boxplot
popularity actor 5	Welch Two Sample t-test data: data$popularity.actor.5 by data$profitable t = -2.7457, df = 1146.1, p-value = 0.006133 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3.1806278 -0.5294482 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 10.19284 12.04788	
budget actor 1	Welch Two Sample t-test data: data$budget.actor.1 by data$profitable t = 0.0061475, df = 951.68, p-value = 0.9951 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -2483447 2499055 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 30637434 30629630	
budget actor 2	Welch Two Sample t-test data: data$budget.actor.2 by data$profitable t = -0.31944, df = 1023.5, p-value = 0.7495 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -2629693 1893377 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 24789976 25158133	

Continuous Variable	T-test	Conditional Boxplot
budget actor 3	Welch Two Sample t-test data: data$budget.actor.3 by data$profitable t = 1.3102, df = 943.09, p-value = 0.1904 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -787650 3952207 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 22473164 20890886	
budget actor 4	Welch Two Sample t-test data: data$budget.actor.4 by data$profitable t = 1.877, df = 881.24, p-value = 0.06085 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -107033.4 4798003.9 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 20255816 17910331	
budget actor 5	Welch Two Sample t-test data: data$budget.actor.5 by data$profitable t = -1.4414, df = 1055.1, p-value = 0.1498 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3632171.1 555791.3 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 15170368 16708558	

Continuous Variable	T-test	Conditional Boxplot
revenue actor 1	Welch Two Sample t-test data: data2$revenue.actor.1 by data$profitable t = -3.0653, df = 1174.2, p-value = 0.002224 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -23276032 -5108376 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 89664392 103856596	
revenue actor 2	Welch Two Sample t-test data: data2$revenue.actor.2 by data$profitable t = -1.6878, df = 1107.8, p-value = 0.09173 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -17045525 1281087 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 73044272 80926491	
revenue actor 3	Welch Two Sample t-test data: data2$revenue.actor.3 by data$profitable t = -0.10167, df = 955.34, p-value = 0.919 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -10426764 9399642 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 65179702 65693263	

Continuous Variable	T-test	Conditional Boxplot
revenue actor 4	Welch Two Sample t-test data: data2$revenue.actor.4 by data$profitable t = 0.013967, df = 996.14, p-value = 0.9889 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -8171610 8288768 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 54263227 54204649	
revenue actor 5	Welch Two Sample t-test data: data2$revenue.actor.5 by data$profitable t = -2.7056, df = 1350.5, p-value = 0.006904 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -16586269 -2643564 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 39125851 48740767	
revenue production company	Welch Two Sample t-test data: data2$revenue.production.company by data$profitable t = -9.0434, df = 1462, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -39246807 -25255640 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 52110280 84361504	

Categorical Variable	T-test	Conditional Boxplot
popularity production company	Welch Two Sample t-test data: data$popularity.production.company by data$profitable t = -8.2815, df = 1109.8, p-value = 3.468e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -6.662996 -4.110479 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 14.33128 19.71801	
budget production company	Welch Two Sample t-test data: data$budget.production.company by data$profitable t = -5.8687, df = 1079.8, p-value = 5.836e-09 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -8869996 -4424939 sample estimates: mean in group 0 - Not Profitable mean in group 1 - Profitable 19233624 25881092	

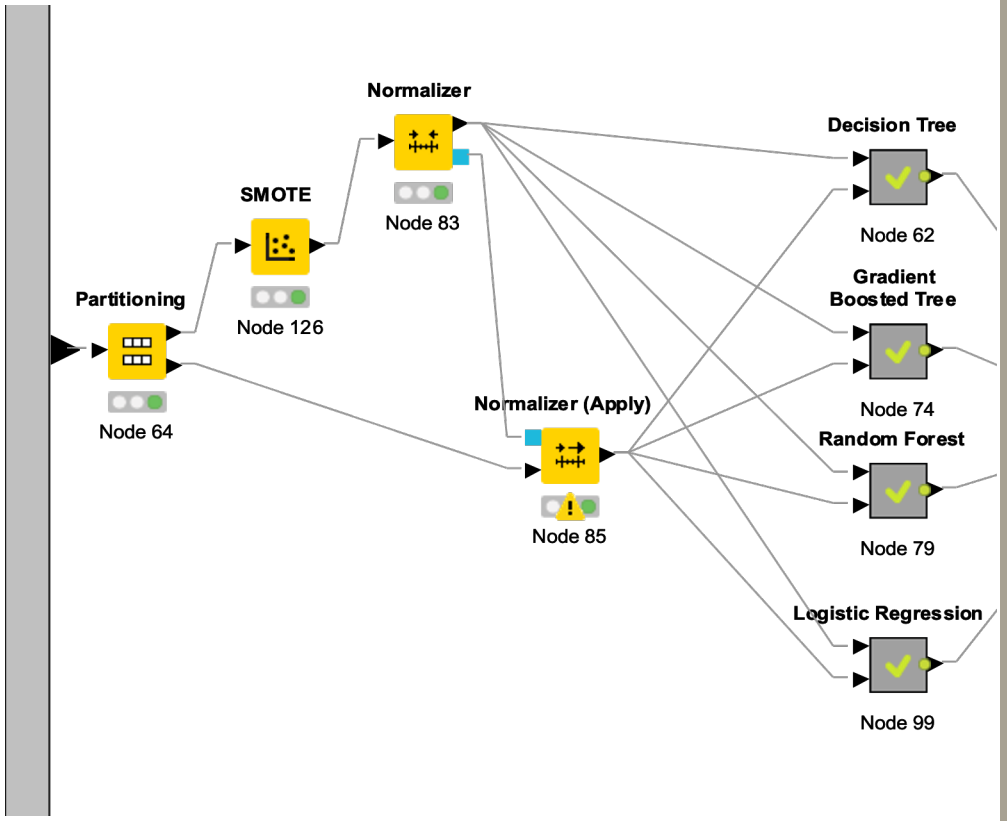
Bivariate Analysis – Continuous

- Looking at the t-tests, **budget**, **runtime**, **popularity actor 1** and **popularity actor 2** all have a significant relationship with **profitability**
- However, **popularity actor 3** and **popularity actor 4** are insignificant with **profitability.** But surprisingly, **popularity actor 5** is significant
- The budget for all 5 actors (**budget actor 1, budget actor 2, budget actor 3, budget actor 4, budget actor 5**) is insignificant with **profitability**
- When it comes to revenue per actor, **revenue actor 1** and **revenue actor 5** have a significant relationship with **profitability**, while **revenue actor 2, revenue actor 3** and **revenue actor 4** do not. This is a similar pattern to the **popularity** variable
- The p-value for **revenue production company** is near zero which means the relationship with **profitability** is highly significant.
- The **popularity production company** and **budget production company** are both highly significant with **profitability**

Data Partitioning, Oversampling and Normalization

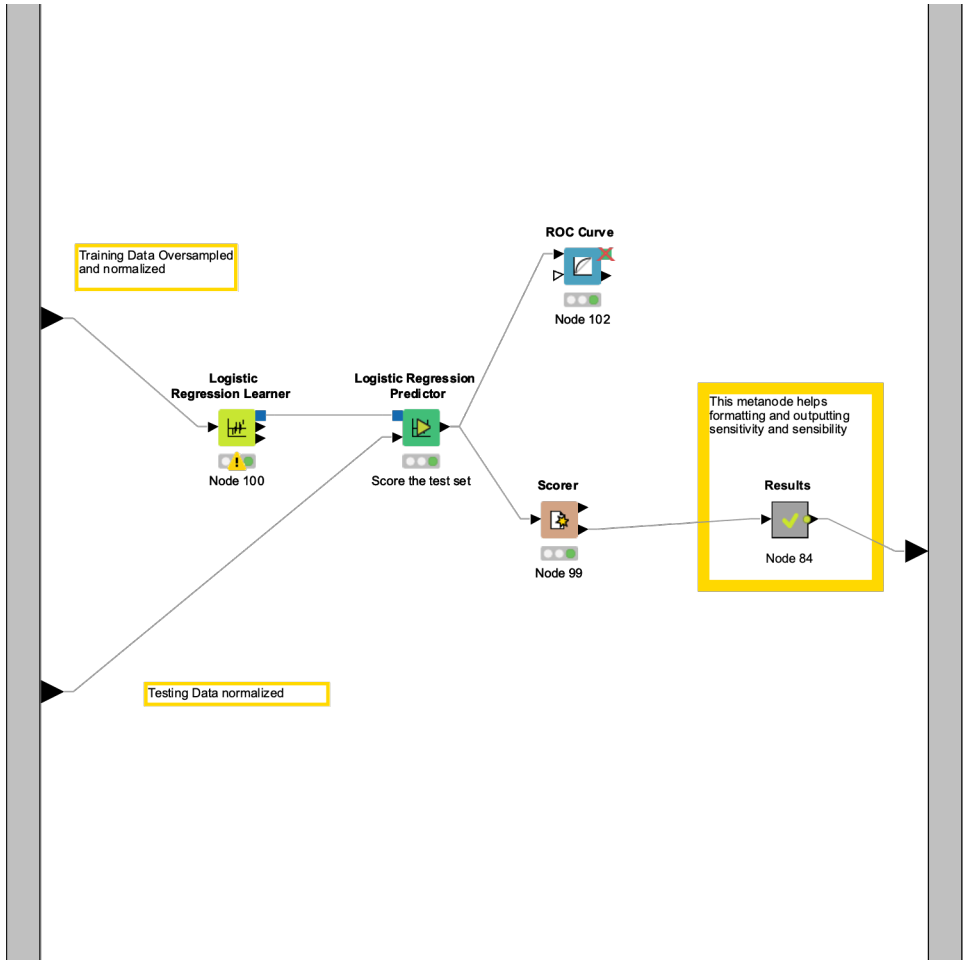
Before model estimation we did:

- Partitioning: we used an 80/20 ratio for the training and testing dataset
- Oversampling: knowing that the relative frequency of the dependent variable was not even (78 % / 22 %), we decided to use SMOTE to oversample the minority class in our target variable (the optimal k parameter for kNN is optimal inside range {3,5}). This way we created 1069 new datapoints.
  The oversampling was applied only to the training dataset
- Normalization: we used a MinMax normalization in the range [0,1] for all the numerical variables



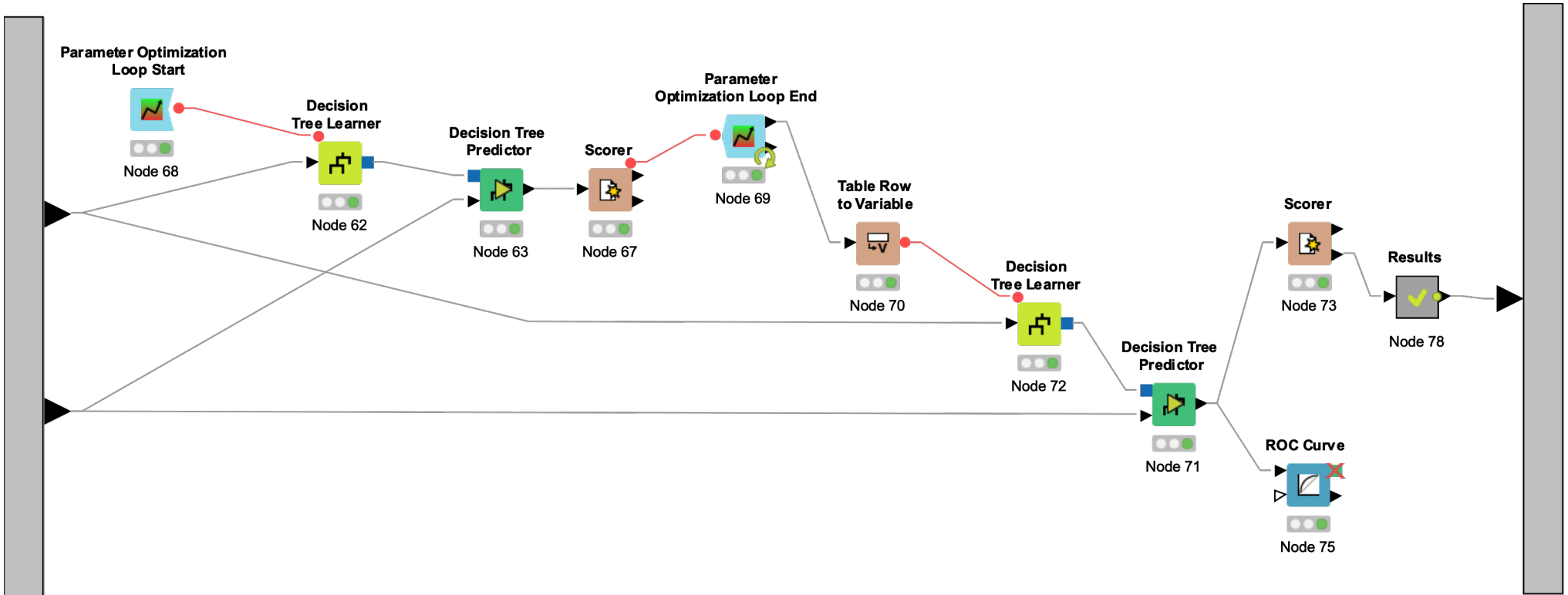
Logistic Regression Model

- A logistic regression model is a classification algorithm which is used when the target variable is categorical. In our model the target variable is the dichotomous state of profitability.
- We tried both the regularizations available, but both performed worse, so we opted for the uniform prior. As a solver we used Stochastic Average Gradient and selected 500 epochs for convergence purposes and used an epsilon of 0.01 as the stopping criteria.



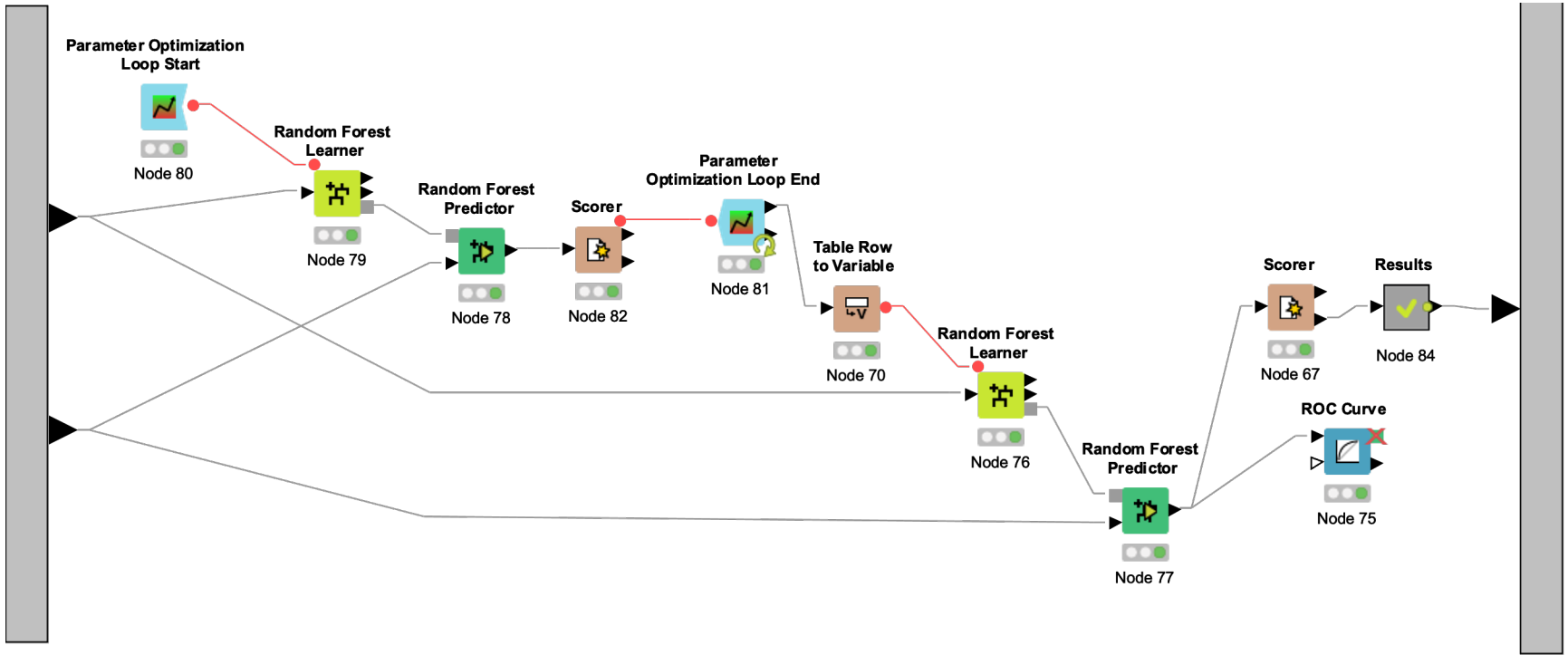
Decision Tree Model

- Setting a class specific sampling rate, we can balance the dataset. This allows us to obtain a higher specificity which allows us to detect unprofitable movies.
- The first step in this Model is the optimization of the hyperparameter. This is done using a loop over a decision tree learner node. For this model we used 100 prediction models as the number of decision trees with a step-size equal to 1.
- Then model is trained with the optimal hyperparameter



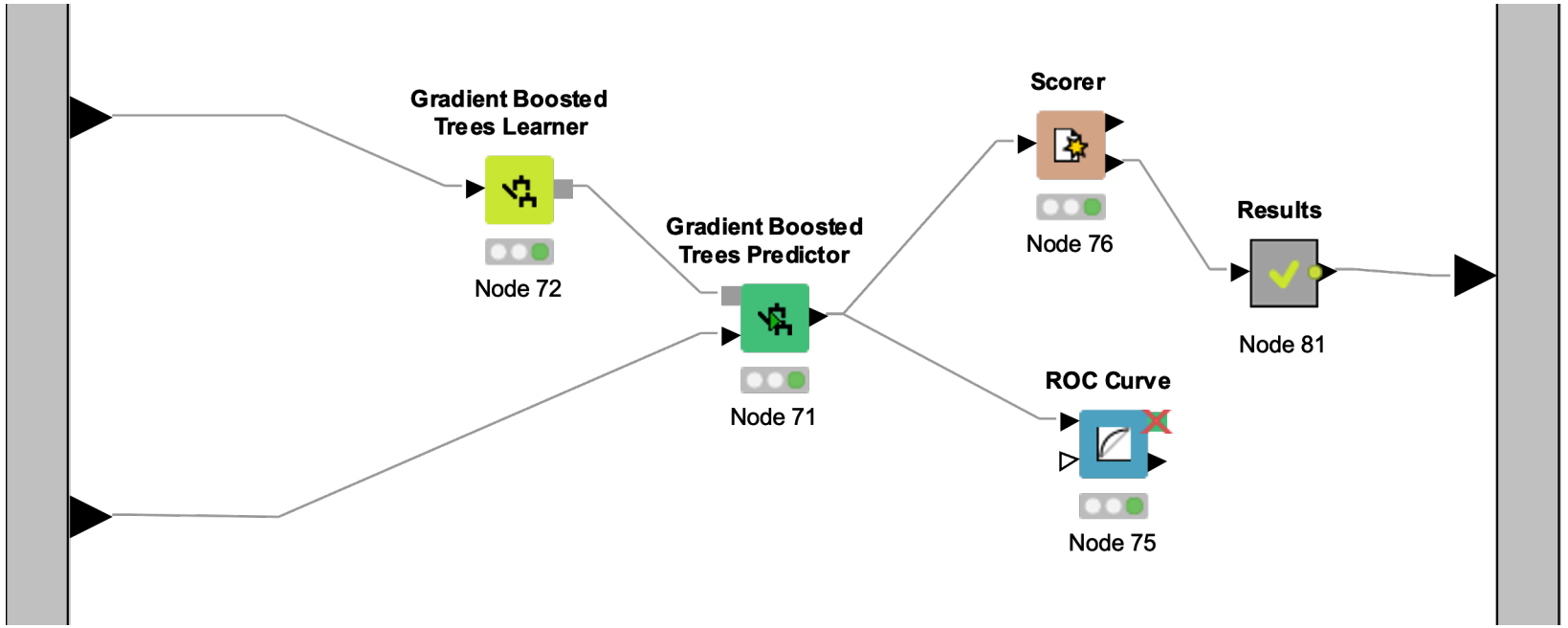
Random Forest Model

- Random Forests ensemble learning is a bagging model in which multiple decision trees give predictions together and the majority vote is the class predicted by the random forest. Random Forests correct the decision trees' tendency of overfitting.
- We run a hyperparameter optimization over the Maximum number of Levels the trees that are bagged in the model, from 1 to 60 with step-size 1. We used this hyperparameter to limit the growth of the trees in the forest, avoiding overfitting.
- The optimal hyperparameter was 42 models, thus the Random Forest Learner was trained with Maximum Tree Depth of 42



Gradient Boosted Tree Model

- Gradient boosting is used in regression and classification models to improve the results of a weak prediction model. A gradient boosted tree model gives a prediction by using smaller decision trees, which act as weaker prediction models in this case. If the smaller decision trees are weak, this algorithm tends to perform better than a standard random forest model.
- We set the limit of tree depth to 10, with a learning rate of 0.1 and 800 models. We used no attribute sampling and showed all attributes to all the trees



Models Performance Metrics

The table belows shows the accuracy statistics for each model

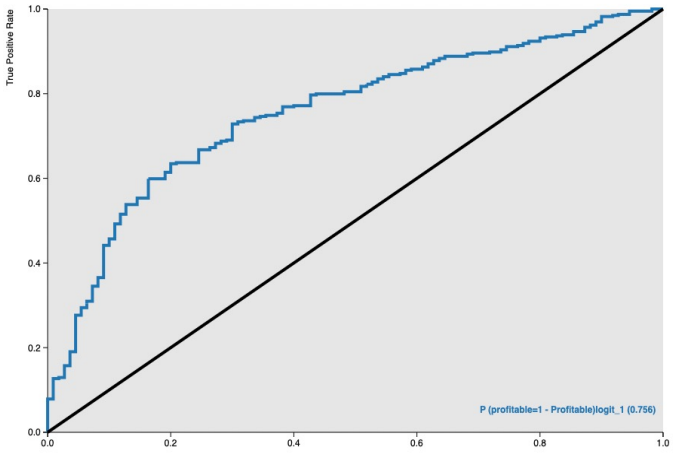
Model Name	Accuracy	Recall	Precision	Sensitivity	Specificity
Decision Tree	0.726	0.751	0.881	0.751	0.636
Gradient Boosted Tree	0.764	0.825	0.867	0.825	0.545
Random Forest	0.762	0.807	0.878	0.807	0.600
Logistic Regression	0.698	0.698	0.893	0.698	0.700

- Random Forest and Decision Tree perform very well, having good values of all the metrics
- Gradient Boosted Tree has the highest Accuracy, Recall and Sensitivity but the lowest Specificity. It has a very competitive and Precision. Overall, it's the model we think performs better.
- Logistic Regression achieves the highest Precision and Specificity but has respectively the lowest values for the other metrics.

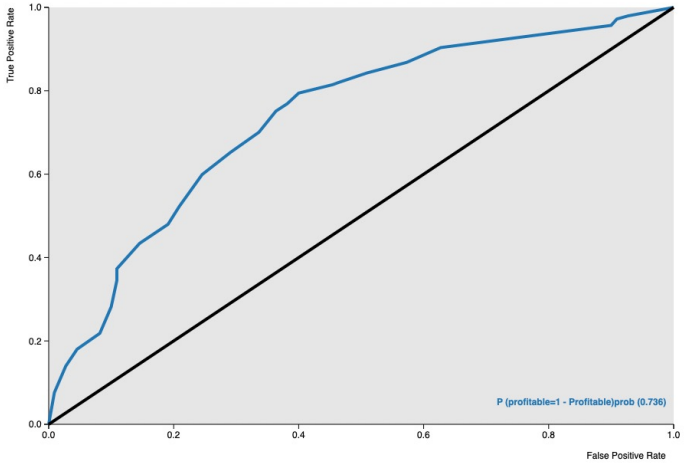
Therefore, none of them stochastically dominates the others. To pick the best we will decide based on **Area Under Curve**.

Performance Metrics: Area Under Curve

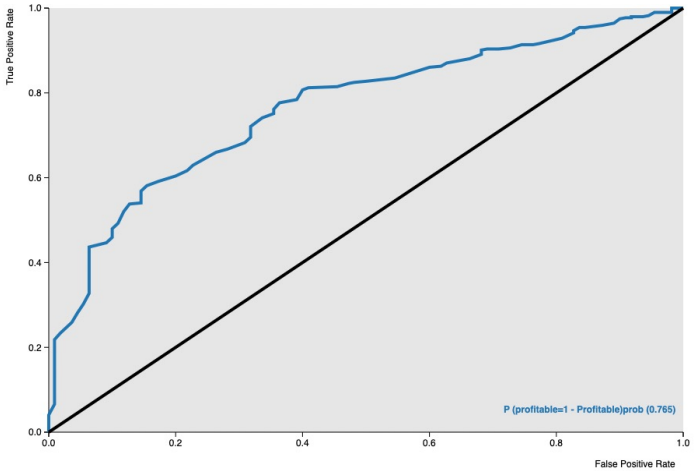
Logistic Regression AUC:0.756



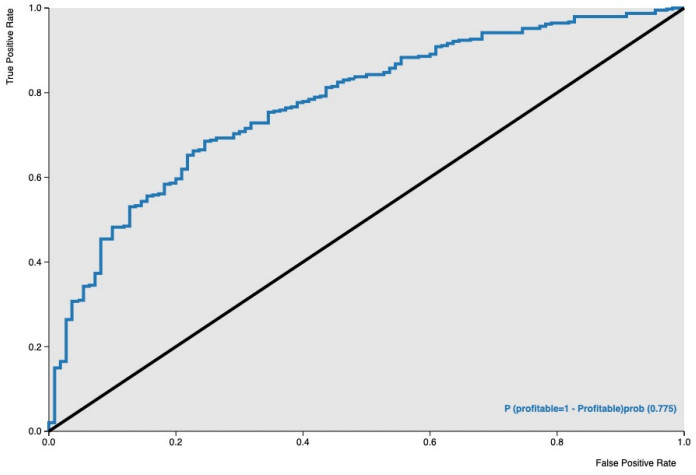
Decision Tree AUC: 0.736



Random Forest AUC: 0.765



Gradient Boosted Tree, AUC: 0.775



The performance metric we took in consideration at the end was **Area Under Curve** in order to deal with a situation with a very skewed sample distribution:

- 78 % profitable
- 22 % not profitable

In this context obtaining accuracy of 78 % could come with a very low specificity, True Negative Rate.

Our first models had very high accuracy but low specificity, it didn't detect True Negatives, unprofitable movies. We cared about that because being able to spot them brings competitive advantage in the movie industry.

We solved the problem with class balancing and obtained still high accuracy (close to 80% for most models) and specificity always higher than 60%

Conclusions about our research

- The movie industry has an expanding global presence, with millions of unique customers all around the world. The movie industry is large; in 2020 the industry was worth around 234.9 billion dollars (Businesswire). Given the size of the industry, the potential to make money is high. As a result, film production companies spend billions of dollars each year financing the production of films. The largest spender on films was Disney, spending around 28.6 billion in 2020.(Televisual)

- Financing a film's production is an investment, and for media companies an integral part of their business models. For a film to be commercially successful it must have higher gross revenues than costs.

- A Gradient Boosted model produced the best results for our classification, returning accurate predictions over 76.4% of the time, with a specificity of 0.545. This is a strong result given the industry is tough to predict, especially amongst the categorical variables. For example, popularity of actors is influenced by how the public perceives them, which may vary yearly. Some studies show up to 80% of films do not return a profit. The benefit to a production company is large, this model would improve the decision-making process for such companies potentially saving them millions each year from the avoidance of flops.

- The database contains more movies which were profitable than not, 1928 to 574. This is an issue for a classification model since the it trains on data which is mostly profitable. One way to improve the outcomes of the model is the inclusion of a human into the decision-making process, human-in-the-loop (HITL). The inclusion of a human can lead to an improvement in specificity, as there is an improved feedback loop.