Movie's success: prediction of the revenue after the release

Introduction

When a movie is released, the most important question for the production company is: how much will I earn from this? This project aims to answer this question: using some information available when a movie is released, we are going to predict its revenue.
In the end, we will summarize the main findings and limitations of our analysis: even if the quality of our predictions wasn't extremely good, we'll try to find out the most relevant features in estimating the economic success of a movie.

Model goals

prediction: build a predictive machine able to provide a sufficiently good estimate of the movie's revenue. This task will be assessed looking at two metrics: *coefficient of determination* (R^2) to measure how much the model is able to explain the data variability and *Root Mean Squared Error* to estimate model's variability in test predictions.
interpretation: understand which are the main features needed to obtain good earnings. To do so, we'll evaluate the magnitude, sign and statistical significance of the estimated regression coefficients and the feature importance extracted from the ensemble models.

General methodology

data manipulation: all models have been trained on scaled data to obtain more comparable values for coefficients and RMSE (anyway, we have also computed RMSE on its original scale), reduce multicollinearity issues in models with interactions (which can arise from huge differences in scale) and not influence penalties assigned by shrinkage techniques.
cross validation procedure: 10-fold cross validation has been performed for each model, in order to avoid particularly "lucky" or "unlucky" splits that can influence our results. Thus, when we talk about the R^2 or the RMSE of a model, we mean the average score obtained in cross validation, instead plots, coefficients and feature importance are related to the best split found for each model (the split with the highest R^2).

Pre-processing and preliminary analysis

We have used two of the datasets that can be found at `https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv`: `movies_metadata.csv` and `credits.csv`. These information have been taken from `https://www.themoviedb.org/`, an interactive site where users can vote their favourite movies and TV series.
The datasets were very big with a lot of information that were not useful for our goal. First of all, since movies from remote past suffers of sparsity for some important variables and since they are strongly different from more recent movies, we decided to focus the analysis only on productions released between 1980 and 2017. Later on we have transformed some of the variables to make them more adapt for our purposes:
`original language:` we have grouped the languages into five different categories: English, Europe (which contains all the languages spoken in Europe except English), India, Asia and Others (which contains some African languages and one artificial language). We have made the distinction between India and the rest of Asia because India has a very fruitful movie production that is quite different from the rest of the continent;
`production companies:` first of all, to make the model easier to read we have decided to take just the first company cited in the observation since it is the most important one in the production process. After this operation, we found out that a big part of the movies released in the world in the last years are produced by five big production companies, that are called Five Big Studios: NBCUniversal, Paramount Global, The Walt Disney Studios, Warner Bros. Entertainment and Sony Pictures Entertainment. So we have decided to group our production companies in six categories: the Big Five Studios and "Smaller company", which contains all the movies that are not produced by the other five. After looking at revenue and budget values for the observations with a missing value, we have decided to add them to Smaller company category since the means were significantly lower than the ones for the Big Five;
`famous count:` this variable was created starting from the `credits.csv` dataset. We have isolated the actors for each movie, then we have decided to call "famous" all the actors that were in more than 10

movies. After that, for every movie we have counted the number of famous actors present in it;
`genre`: in this variable are stored the genres which the movie belongs to. Since every movie can have more than one genre, we decided to create a separate dataframe with 20 dummy variables, one for each genre, so that every movie (identified by the original title and an id) have a 1 for each genre it belongs to.

Exploratory data analysis

Variable	Description	Distribution
`revenue`	Target variable that shows how much money a movie earns. Numeric, expressed in US Dollars. Missing values were removed.	Not normal. Outliers for `revenue` $>$ 1.500.000.000$
`belongs to collection`	Boolean variable, True if the movie belongs to a saga.	About 23% of the observations are True.
`budget`	How much money has been used to make the movie. Numeric, expressed in US Dollars. Very important predictor so missing values were removed.	Not normal. Outliers for `budget` $>$ 250.000.000$
`original language`	Spoken language in the movie. Categorical with 5 categories.	Around 88% of the movies are in English.
`production companies`	Principal studio that has produced the movie. Categorical with 6 categories.	Every studio of the Big Five has around 10% of the movies. The rest is produced by smaller companies.
`runtime`	Length of the movie. Numeric, expressed in minutes.	Not normal. Outliers for `runtime` $>$ 300 minutes.
`famous count`	Number of famous actors in a movie, numeric.	Not normal.
`genre`	This is composed of 20 dummy variables, one for each genre to which a movie can belong.	

Outlier analysis

Univariate outliers: thresholds enlightened in the table above have been set manually looking at the boxplots;
Multivariate outliers: we found five observations (enlightened in Figure 1) considered as multivariate outliers computing the squared Mahalanobis distances between all the quantitative variables, including `revenue`.

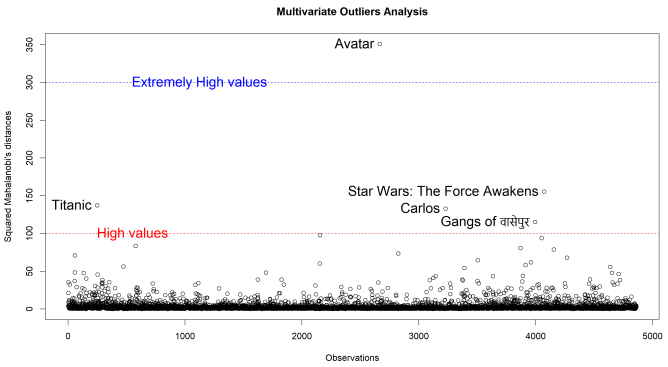


Figure 1: *Squared Mahalanobis distances of the observations*

Linear regression

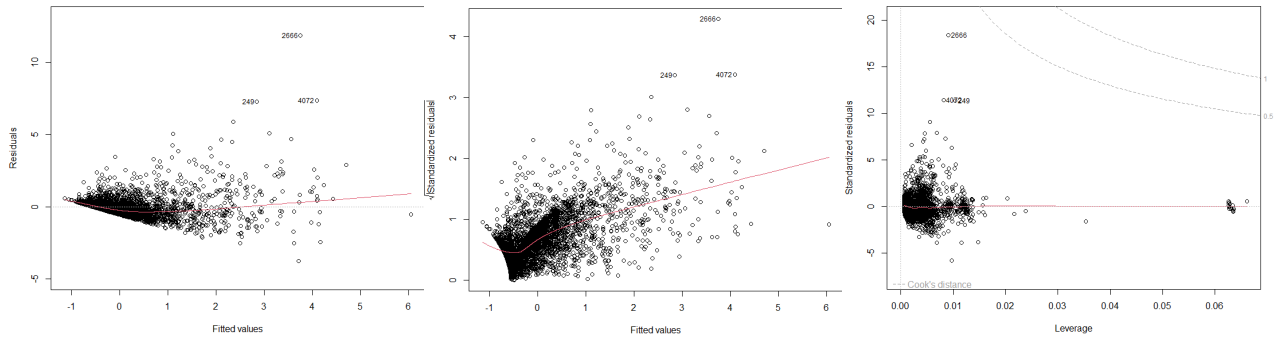
The first predictive model that we examined is the linear regression. The possible existence of a linear relationship between target and predictors has been suggested by the quite relevant Pearson correlation coefficients of `revenue`: $\simeq 0.73$ with `budget`, $\simeq 0.44$ with `famous count`, $\simeq 0.22$ with `runtime`.

Model construction

Base model

We started from an initial Linear Regression and then we applied some changes in data, predictors and model structure trying to improve it. The initial model has been fit without considering the genre information to see if we could reach an acceptable explanation of our phenomenon with less complexity. As a result, we obtained an average R^2 of 0.594 with a RMSE (on scaled data) around 0.651. These scores aren't

particularly good, especially regarding the quality of predictions, since such a value for the RMSE means that the average residual in predictions is around 112.000.000$. Anyway, the level of the R^2 isn't totally bad. Displaying the diagnostic plots in Figure 2, we detect some problems:
independence of the error components: in Figure 2a the error terms tend to show a pattern since we detect higher residuals as the predictions deviate from the mean value of revenue. Instead, the fitted values that are closer to 0 (which means, closer to the average value of `revenue`) have smaller residuals, forming a sort of "U"-shape;
homoschedasticity: Figure 2b underlines that the residuals don't preserve a constant variance, indeed the root standardized residuals grow as we consider predictions that are higher than the mean value of `revenue`;
strange observations: finally, the model doesn't seem to suffer the presence of influential points, since all the leverages and the Cook's distances are behind their "dangerous" thresholds (Figure 2c).
Most of these irregularities are shared by all the different models trained and thus they will be faced in a separate section.



(a) *Residuals vs fitted* **(b)** *Scale location* **(c)** *Leverages*

Figure 2: *Diagnostic plots for the initial model*

Intermediate steps

To improve our model we first tried to remove the outliers (both univariate and multivariate ones) obtaining an increase in the R^2 (now up to $\simeq 0.603$) and a little decrease in the RMSE. Since the number of lifted observations was sufficiently low we decided to keep this update in the following steps.
At this point we included the variables related to the genre information that we have stored in a separate dataset. This modification explained positive effects on both the R^2 and the RMSE, but there is a huge increase in the number of coefficients. So we tried to remove some of the categories that are not so influential in the prediction of revenue. In evaluating the relevance of a genre dummy variable we considered two factors:
percentage variation of `revenue`: for each genre $g = 1, ..., G$, we computed $\overline{Y}_g$, which is the mean value that `revenue` hires on the subset made only of observations that belong to genre g. Then, computing the percentage variations $\Delta_g = \frac{\overline{Y}_g - \overline{Y}}{\overline{Y}}$, we marked as "non relevant" the genres with $|\Delta_g| < 0.1$;
coefficient's relevance: we considered the p-values of the t-statistics for each genre in the model which contains all of them and we marked as "non relevant" all the genres that have a considerably high p-value.
We decided to remove all the genres considered as "non relevant" according to both the above criterion, i.e. `Mystery`, `Thriller`, `Comedy`, `War`, `Music` and `TV Movie`. Thus, we achieved a sort of dimension reduction and also an improvement both in R^2 (now near to 0.613) and in RMSE ($\simeq 0.638$ on scaled data).

Interactions

In the final step, we decided to include some "interaction effects" considering the most logical combinations between our predictors: we suppose that different production companies, according to their dimension and popularity, usually allocate different levels of budget for movies. Same consideration has been done for the number of famous actors. Finally, since famous actors usually require higher wages, we believe in a significant interaction between `famous count` and `budget`. Hence, for each of the 3 interactions, we trained separately the models that included only one of them and then we made a comparison against the model without interaction's coefficients using the ANOVA test: as a result, all the tests rejected the null hypothesis of having interaction coefficients equal to 0. Thus, we decided to include all the interactions in our model, and this led to remarkable improvements: $R^2 \simeq 0.635$ and RMSE $\simeq 0.61$.

Shrinkage methods

Once we have found our "best Linear Regression", we decided to apply shrinkage techniques in order both to reduce variability in prediction and to highlight the main features among all the available predictors and the interpretation. The methods adopted were the Ridge, Lasso and Elastic Net regression: for all of them, we've found the optimal parameters via Cross Validation. The results achieved were not satisfactory since only with Elastic Net regression we obtained a small boosting ($R^2 \simeq 0.639$).

Interpretation of coefficients

Now that we've obtained a model with an acceptable ability in explaining the behavior of the target variable, we can try to point out which are the main factors in supporting movie's revenue, interpreting the coefficients of the best regression found so far, i.e. the one with the interaction coefficients without applying the shrinkage procedures.

Intercept

The intercept shows a small negative value, meaning that when numerical predictors hire their mean values and the categorical ones show their "zero modality", we obtain an estimated revenue which is pretty close to its average (93.889.445$). The high p-value reinforces this idea.

Numerical predictors

`budget:` as expected, the amount of budget invested in a movie impacts highly on revenue, as we can point out from the strong positive value of the coefficient ($\simeq 0.52$);
`famous count:` here we noticed a little magnitude ($\simeq 0.06$) and an high p-value, so at a first sight, it doesn't seem that it strongly influences the estimates;
`runtime:` the running time seems to induce an increase in predictions as the variable hires higher values than its mean. Anyway, the impact on the target is smaller than the `budget` component.

Categorical predictors

`belongs to collection:` the model recognises a remarkable importance to collections. Indeed, if the film belongs to a saga, its expected earnings will be higher than the average by a factor of $\simeq 0.47$;
`original language:` almost all the coefficients have little magnitude and poor significance, meaning that the information provided by the movie's original language isn't crucial in determining its income;
`production companies:` we deduced that movies produced by smaller companies are typically related to earnings smaller than the mean (indeed, the β for this modality is the lowest $\simeq -0.17$). The coefficients for the other studios are still negative but with less magnitude and a higher p-value, so compared to the "zero modality", i.e. NBC Universal, the other big companies have slightly smaller earnings;
`genre:` negative impacts on `revenue` are produced if the movie belongs to `Action`, and/or `History` genre (especially this last one has a remarkable magnitude of $\simeq -0.16$). On the other hand, positive impacts on the target are explained by `Romance` and `Adventure` movies, but the most relevant genre in increasing predicted values is `Animation` ($\simeq 0.14$).

Interaction's terms

budget coefficient among different production companies: especially in "Walt Disney" and "Warner Bros.", budget's investment higher than the average tends to generate a bigger estimate $\hat{Y}$;
famous actors coefficient among different production companies: effects produced by the number of famous count is mitigated in "Warner Bros." ($\simeq -0.12$) if compared with "NBC Universal";
interaction between budget and famous actors: as predictable, the interaction is positive ($\simeq 0.1$): thus we can say that the growth in `revenue` (with respect to the mean) induced by the increase in `budget` is much more strong as we find more important actors in the movie.

Fixing model's assumptions

In this subsection the goal is to fix the issues raised by diagnostic plots. As mentioned before, all models have shown common problems, so we'll try to solve the violations just for the final one. To do so, we tried to apply some transformations in order to obtain the best configurations for our variables, paying the price of loosing in interpretability of the regression's coefficients.
After several attempts, we applied Box-Tidwell and Box-Cox methods (power transformations) obtaining "more normally distributed" variables. Furthermore, models fit on transformed data achieved a significant improvements as regards residual's distribution (as detectable from diagnostic plots in Figure 3). The improvements in predictions can be seen in Figure 5.

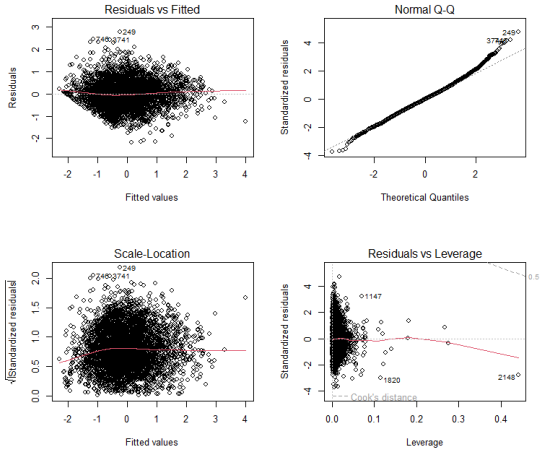


Figure 3: *Diagnostic plots for the model with transformations*

Algorithmic methods

After having fit a proper statistical model, we tried to use algorithmic methods, to see both if they can perform better and if they can suggest other interesting considerations.
We start by fitting a single Decision Tree model to see if it was able to understand the relationship between target and predictors sufficiently well and then we attempt to improve it using two ensemble techniques. Each model has been fit on the dataset configuration which has been found to be the best one in the previous section (i.e. without outliers and considering all the "relevant genres"). We shortly report the definitions of some measures:
deviance: sum of squared residuals in each terminal region;
node purity: related to a predictor, it is the total decrease in residual sum of squares, averaged over all trees in a Forest, when splitting on the considered variable.

Decision Tree

Initially, we fit the Tree with high depth and size (number of terminal regions) in order to build a very low-biased model and then reduce its variability through pruning; in the end we select the optimal number of terminal regions that minimize the model's deviance. As a result, we obtained a Tree with around 11 terminal regions and an average $R^2 \simeq 0.622$ but with an excessive variability (RMSE = 102.802.578$ on the original scale). So, even if the pruning process produced a decrease in the RMSE, the model still achieve poor performance in the test set and thus we decided to apply ensemble techniques to improve the quality of predictions (reducing the variability). Anyway, the Tree has shown an appreciable R^2, so we report the variables used at least by one of the trees fit in cross validation to split the feature space: `budget`, `belongs to collection`, `Animation`, `runtime`, `famous actors` and `production companies`. Hence we can conclude that these are the most relevant predictors in the model's construction.

Ensemble models: Random forests and Bagging

After having fit and compared Bagging and Random Forests, we're going to discuss about the feature importance of our predictors. In the end, we'll provide some possible adjustments in order to reduce the computational time required.

Model's fit

Comparing the results of the two methods, we see that Forests (setting $m = \lfloor \frac{p}{3} \rfloor$ as number of random features considered at each split) outperforms Bagging both in the train set ($R^2 \simeq 0.66$ against $R^2 \simeq 0.67$) but mainly in the test set, scoring an average RMSE of $\simeq 0.563$ against $\simeq 0.598$ (on scaled data). Thus, we conclude that de-correlating the multiple Trees in the ensemble, we reach a remarkable improvement in test's predictions. The values of the predictions compared to the real revenues can be seen in Figure 6. In the following sub section we'll comment the role of the different predictors based on the output obtained from Random Forest.

Feature importance

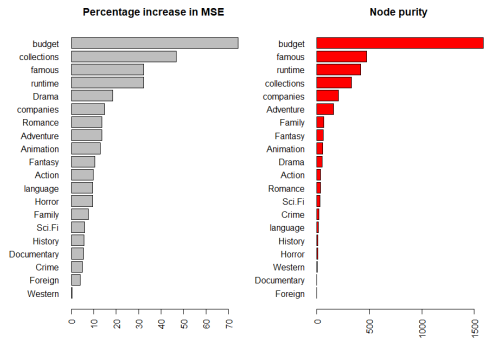


Figure 4: *Feature importance*

Looking at the feature importance plot in Figure 4 for the best Random forest found so far, we can state some considerations:
budget: once again, the amount of investment spent in the movie is the principal feature used by the model. Indeed, removing this predictor cause a strong increase both in the MSE and in the node purity;
number of famous actors and ***running time:*** here these two numerical predictors are considered quite relevant according to both the criteria plotted;
membership to a collection: also the Random Forest algorithm recognizes a certain relevance to this feature since its removal produces a remarkable percentage increase in MSE;
categorical predictors: little importance is assigned to the film's original language, while the production company has a moderate role in decreasing MSE and node purity;
genres: according to the percentage increase of MSE criterion, we can distinguish 3 sub groups of different importance: `Family`, `Scie-Fi`, `History`, `Documentary`, `Crime`, `Foreign` and `Western` don't seem to be crucial, while `Romance`, `Adventure`, `Animation`, `Fantasy`, `Action` and `Horror` show a moderate important. Finally, the only genre which seem to be relevant in reducing MSE is `Drama`.

Conclusions

Generally, we've noticed that a movie's income is a random phenomenon with high variability, influenced by a bunch of factors which makes its correct estimation very difficult to find (indeed it is not unusual that productions unknown before the release obtain unexpected high earnings while eagerly awaited movies fail to meet their expectations). That's probably the main reason why our models weren't able to reach totally accurate estimates (the best RMSE was 88.298.141$, which is not terrible but surely improvable). Table 1 shows the values of the metrics that we have considered for Linear Regression with Interactions, Elastic Net Regression, Linear Regression with Transformations, Decision Tree, Bagging and Random Forest.
Anyway, the general target's trend has been understood. Thus, we try to hypothesize which elements can influence movie's business success. Not surprisingly, high revenue is mainly related to strong investments. Looking at other predictors, we can provide some little suggestions on how to spend the budget: it turned out that producing movie's collections instead of separate films positively affects earnings. Additionally, some genres can remarkably condition the income, particular success is assigned to animation movies, but also dramatic and romantic ones are quite remunerative. Instead, we don't recommend to invest in historical genres, if the goal is just to make money. Finally, the presence of famous actors plays its commercial role: using economical resources to hire popular characters will probably make the investment more valuable.
In conclusion, we claim that our results have been primarily affected by the arbitrary "construction" we've done for some variables, so different aggregations may could lead to other findings. Nonetheless, we believe that including different predictors (technical movie's features, director's name, ...) or adding more observations (especially for ensembles) can produce improvements. Obviously, also trying different models can be a good idea, but we think that optimizing the above models should be the priority.

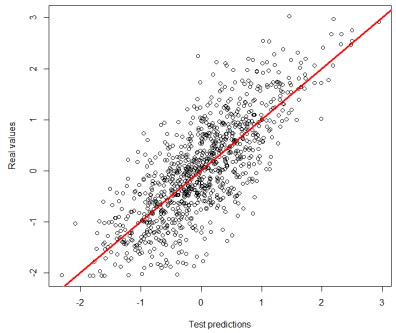


Figure 5: *Real values vs Fitted for L.R. with transformations*

Models	R^2	RMSE
L.R.I.	0.635	0.61
E.N.R.	0.639	0.624
L.R.T.	0.647	0.606
D.T.	0.622	0.655
Bag.	0.660	0.598
R.F.	0.670	0.563

Table 1: *Value of the metrics for different models*

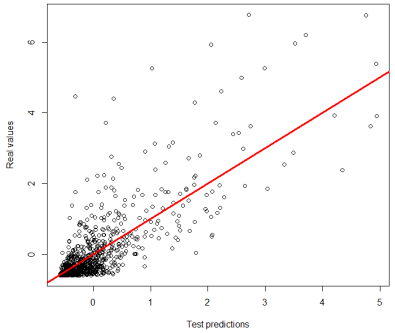


Figure 6: *Real values vs Fitted for random forest*