Northeastern University
College of Professional Studies

ALY 6040 Report – Data munging and Data Wrangling, building regression model

Group Project

By –

Dishali Sonawane

Mrigank

Pranav Khanna

Swapnil Lokhande

Sakshi Dalicha

ALY6040 Data Mining Applications SPRING 2019 CPS

Instructor: Justin Grosz

Due Date: June 9, 2019

Modelling technique selection

We have selected the multiple linear regression techniques for modelling our dataset. Multi- Linear regression usually examines the relationship between two or more variables. This technique helps in in examining how the independent variables influence the dependent variables.

- Dependent variable is the target variable that we are trying to predict.
- Independent variables are used to predict the dependent variable.

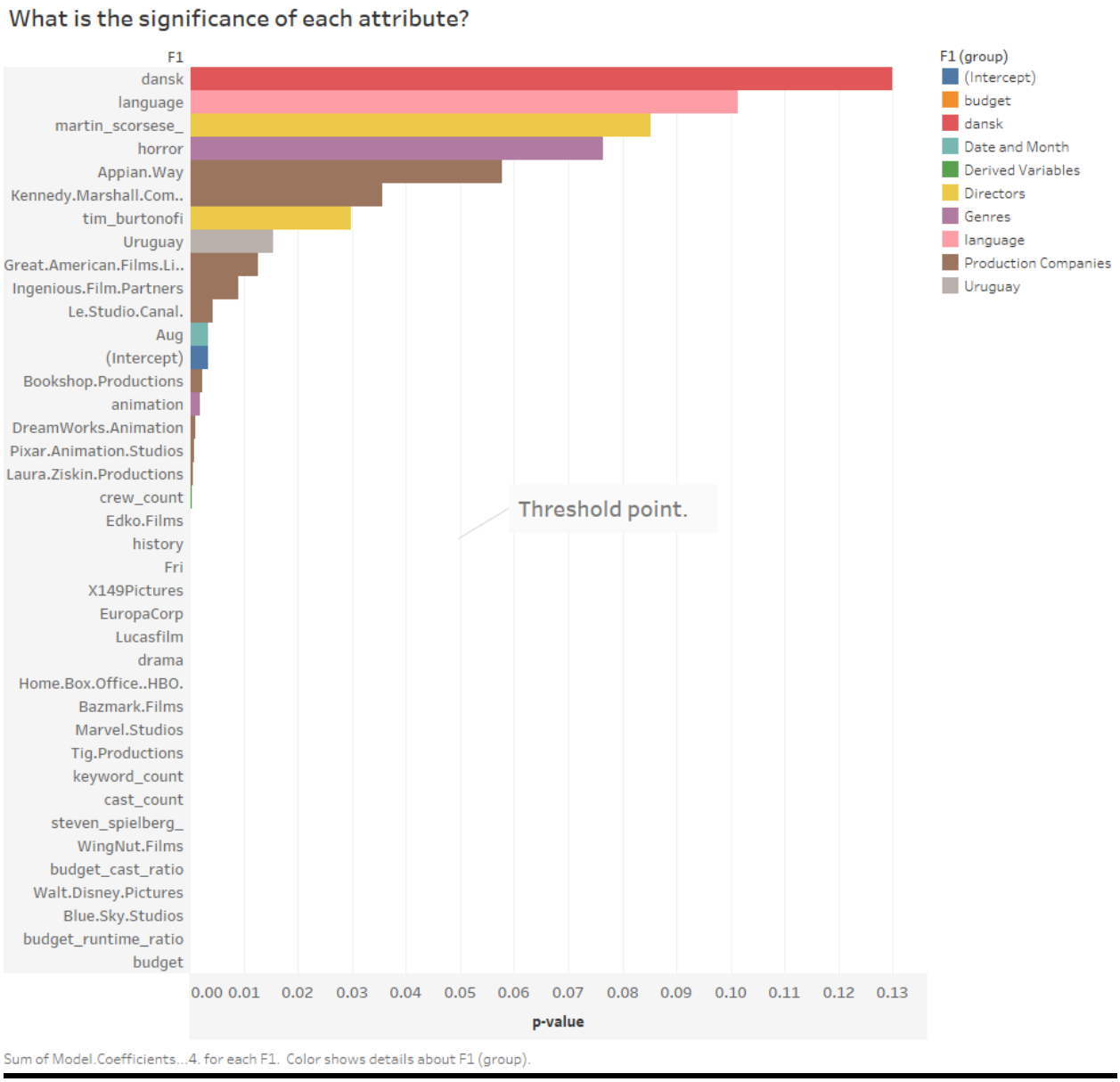
To begin with the investigation, we must find the dependent and the independent variables. We discover that ‘revenue’ is the dependent variable in other words the target variable and the other variables like the genres, budget, runtime, production company, production country, language of the movie, date, cast, crew etc. are the independent variables. In our case we must understand what factors affect the **‘revenue’** collection of the movies. Particularly we are curious about how the independent variables affect the revenue collection of a movie.

Also, linear regression is very flexible in terms of the number of attributes we fetch to the model. It can convert the categorical variable for computation and gives the significance of each attribute in the model.

By applying the linear regression modelling technique on the dataset, we discovered that the following attributes are significant

Data Modelling

Our initial model had **1225** variables, after applying linear regression to our model we were able to generate a list of **38** most significant variables in our data based on lower p-values.



Derived Variables

We have derived the following variables after cleaning the dataset

- **budget_cast_ratio**

There are a lot of movies that have much bigger budget and cast, which overshadows rest of the data. So, as to normalize the data we have derived another variable name budget to cast ratio.

- **budget_crew_ratio**

There are a lot of movies that have bigger budget and larger crew in the data. So, in order to normalize the data for all movies we have derived another variable named budget to crew ratio.

- **budget_runtime_ratio**

We derived this variable as the budget varies with the runtime, for example a long runtime movie doesn't generate a lot of revenue. In order to normalize the values, we created the new variable.

Original Variables

- **Budget**

The budget is total amount of money that is spent making the movie, it is a crucial factor and may include the following while formulating the budget.

- Pay for the Actors
- Special visual effects
- Music
- Crew salaries
- Travelling cost
- Promotion cost

So, a high budget film with popular actors, good music and marvelous visual effects will generate high revenue. Also, if the movie is an international release the promotion might help generating more revenue.

- **Genres**

Genres classify the movies and help the audience to identify their choices. It is noted that in our data drama is the most popular genre, followed by crime, thriller and action. It is observed that movies with lots of drama collect more revenue. Also, it is observed that documentary and music related movies have collected the least revenue. Genres play an important role in the revenue collection.

- **Directors**

Directors contributes towards script, screenplay and the overall development of the movie. He guides the crew and the cast member to fulfill the vision. Clarity of vision and a passion to portray the ideas

and visuals into realistic movie are few qualities that matter in directing a movie. Good direction can lead to a good movie and in turn collect a good revenue.

- **Crew Count**

The production company hires the crew for the purpose of maintaining a good working environment. They look over the management, good and support crew adds up to the success of the film, resulting good revenue.

- **Production companies**

The movie produced by big production companies like Walt Disney attract a huge audience and will impact the revenue collected by the movie.

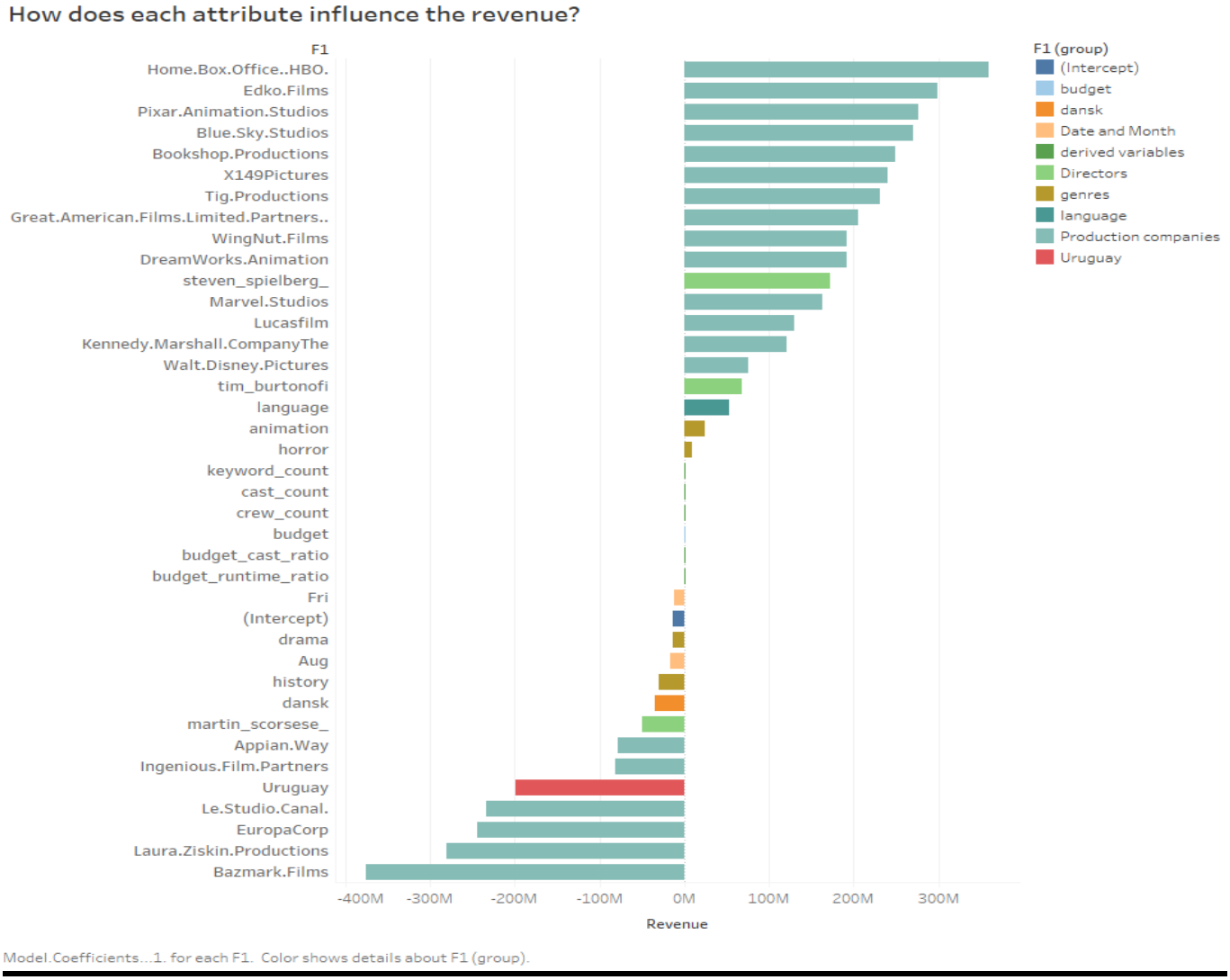
- **Month and Day of the week**

The release data plays a significant role, for instance if the movie is released during vacations and festivals it is more likely that a movie generates more revenue. More over if a movie is released on Friday then over the weekend people tend to watch the movie and adds up to the revenue.

- **Languages**

The language the movie is released impacts the revenue collection, for example if the movie is international release and its in English language it will generate more revenue but, on the other hand if a local language movie is released globally it will not generate more revenue.

Significant attributes



Model Accuracy

Attributes: 1225

Residual standard error: 94270000 on 1809 degrees of freedom
Multiple R-squared: 0.7166, **Adjusted R-squared: 0.5302**
F-statistic: 3.844 on 1190 and 1809 DF, p-value: < 2.2e-16

Next, we reduced the number of attributes based on the p-value of attributes.

Attributes: 38

Residual standard error: 79630000 on 2800 degrees of freedom
Multiple R-squared: 0.687, **Adjusted R-squared**: **0.6648**
F-statistic: 30.89 on 199 and 2800 DF, p-value: < 2.2e-16

RMSE : 8748470

Business learning

From a list of 21 different **Genre**, our model predicted that (1) **animation**, (2) **drama**, (3) **horror**, (4) **history** are the most significant genre when it comes to representing the revenue of the movies.

Business Interpretation:

1. If the movie is of genre history or drama the net revenue of the movie positively influenced.
2. If the movie of the genre horror or animation the net revenue of the movie is negatively influenced.

If we consider the **Director**, we can see that the top three directors are (5) **Steven Spielberg**, (6) **Tim Burton** and (7) **Martin Scorsese**.

Business Interpretation:

1. So, if a movie made by Steven Spielberg or Tim Burton it positively impacts the revenue, is it more likely that it will generate a high revenue as huge number of audiences may be attracted towards these directors.

We also observe that (8) **August month** has a negative influence on the revenue.

Business Interpretation:

1. Try to release movies in other months as there are no festivals in that month, moreover the months like November, December make huge revenues due to Thanksgiving week and New year.

Additionally, we observe that (9) **Friday** has negative influence on the revenue.

Business Interpretation:

1. one reason for this might be as all the movies are released on Friday the audiences are more attracted to other movies that have good review from audiences.

Viewers like to watch movies of the popular **production companies.** From our model we observed that there are **20** production companies that are most significant, out of which 14 production companies have a positive impact on the revenue, whereas remaining 6 negatively influence the revenue.

Business Interpretation:

1. Popular production companies positively impact on the revenue, whereas less popular companies negatively influence the revenue.

Model Optimization Recommendation

We intend to include new attributes in our model. Currently we are only including only most frequent attributes. For instance, out of a list of 3000 directors we have only used 20 most frequent directors, also for production companies we are only including most frequent production company names in the model. Also, there is a thought to include some derived variables to normalize records.
Next, optimization recommendation is to test different machine learning algorithm such as random forest and neural network to find if they provide any improvement.

Conclusion

For this week we added a lot of different elements to our model. First, we generated new variables such as cast_count, crew_count, included the most frequent directors, added most frequent production companies, included budget_to_cast ratio, budget_to_crew ratio. We were able to reduce the number of attributed to 38 from 1225 total based on the p-value generated by the regression model. This optimization helped us to increasing the accuracy of the prediction from 53 percent to 66 percent. Further, we discussed out further modification and optimization that we could add in our model in coming weeks.