Alec Strauss

Final Paper

Introduction

People all around the world dream about being famous in Hollywood. Now most people want to be the leading actor, and others want to be the director. There are a select few out there that might even want to write movies. All these professions are extremely difficult and take a long time to reach the pinnacle. However, there might be a way to "make it" in one of these professions, writer.

In the world of analytics there are many different methods and models that can do many different things. The goal here is to use various analytical techniques to develop a Hollywood movie that is sure to succeed.

Analysis and Models – About the Data

The data contained over 4,000 movies taken from the IMDb movie database. There were 22 different variables within the data set, but for this analysis not all were used. These are how the variables were used for analysis.

Budget and revenue were numerical variables ranging from $1,000 to $2,787,965,087.00. Whenever these variables were used, they were transformed into categorical variables. This was done so that models that required a factor to make a prediction could do that. Below is a table breakdown of how exactly the budget and revenue variables were transformed.

Budget Categories	**min**	**max**
Lowest Budget	$7,000	$5,000,000
Very Low Budget	$5,200,000	$12,000,000
Low Budget	$12,300,000	$20,000,000
Average Budget	$20,500,000	$30,000,000
High Budget	$30,250,000	$50,000,000
Very High Budget	$50,100,000	$80,000,000
Highest Budget	$80,341,000	$380,000,000

Revenue Categories	**min**	**max**
Lowest Revenue	$1,036	$7,022,209
Very Low Revenue	$7,022,728	$19,701,164
Low Revenue	$19,777,647	$39,421,467
Average Revenue	$39,438,674	$70,992,898
High Revenue	$71,000,000	$122,126,687
Very High Revenue	$122,195,920	$231,411,584
Highest Revenue	$231,449,203	$2,787,965,087

Another variable that had to be transformed due to its numerical nature was the movie's runtime. Below is a table breakdown of how the runtime variable was transformed.

Runtime Groupings	Runtime Range
1	25-86
2	87-89
3	90-92
4	93-94
5	95-96
6	97-98
7	99-100
8	101-102
9	103-105
10	106-107
11	108-109
12	110-112
13	113-115
14	116-118
15	119-121
16	122-125
17	126-130
18	131-136
19	137-148
20	149-338

That is the end to how variables were transformed. The only other modification to note pertains to a movie's release date. This was simply truncated to just be the release month.

Analysis and Models – Models

As promised, there will be many analytical methods and models used to develop this newest Hollywood success. To begin association rule mining will be used to generate which combination of genres yield the greatest revenue for a movie.

Once the genre is decided on, the movie specifications will be the next focus. This will involve using decision trees to decide things like the movies budget and runtime.

The most important aspect of this new movie will be its overview. Text mining will be used to see which words are used in the highest revenue and lowest revenue movies. Once the overview is written, a Multinomial Naïve Bayes model will be developed to predict the revenue of the new movie based on the overview that was written for it.

Results – Association Rule Mining

Association rule mining is going to give this new movie a genre. Rules will be created that say which genres are associated with the highest revenue. Below are two tables depicting the top five rules generated with the highest support, lift, and confidence.

Lhs	Rhs	Support	Confidence	Lift	Count
{Adventure, Drama, Fantasy, Romance}	{Highest Revenue}	0.0011471179	0.8	5.624194	4
{Animation, Family, Fantasy, Music}	{Highest Revenue}	0.0008603384	1.0	7.030242	3
{Animation, Comedy, Fantasy, Music}	{Highest Revenue}	0.0008603384	1.0	7.030242	3
{Comedy, Family, Fantasy, Music}	{Highest Revenue}	0.0008603384	1.0	7.030242	3
{Animation, Comedy, Family, Fantasy, Music}	{Highest Revenue}	0.0008603384	1.0	7.030242	3

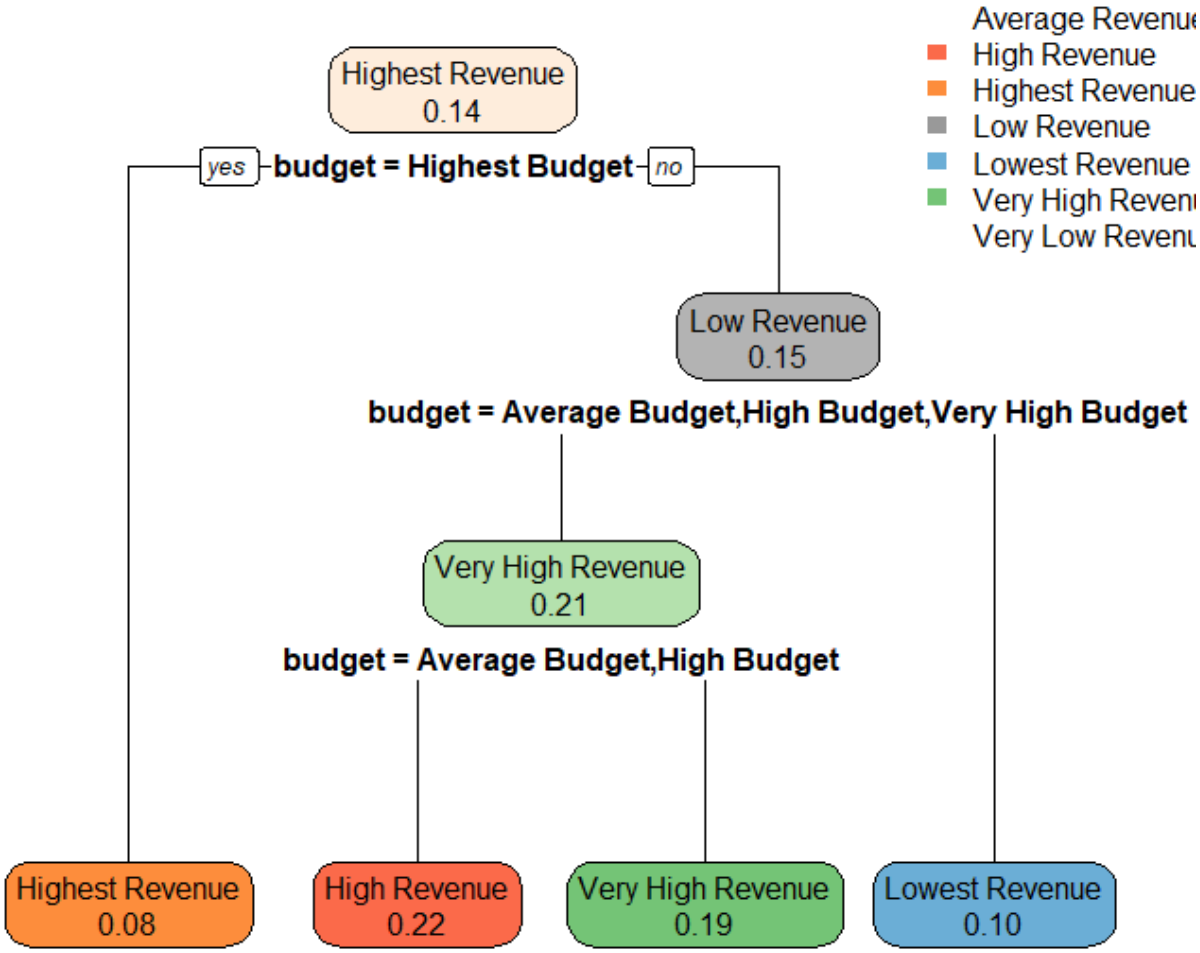
Lhs	Rhs	Support	Confidence	Lift	Count
{Sciece Fiction, Western}	{Highest Revenue}	0.0002867795	1.0	7.030242	1
{Animation, Mystery}	{Highest Revenue}	0.0002867795	1.0	7.030242	1
{Adventure, Science Fiction, Western}	{Highest Revenue}	0.0002867795	1.0	7.030242	1
{Action, Science Fiction, Western}	{Highest Revenue}	0.0002867795	1.0	7.030242	1
{Comedy, Science Fiction, Western}	{Highest Revenue}	0.0002867795	1.0	7.030242	1

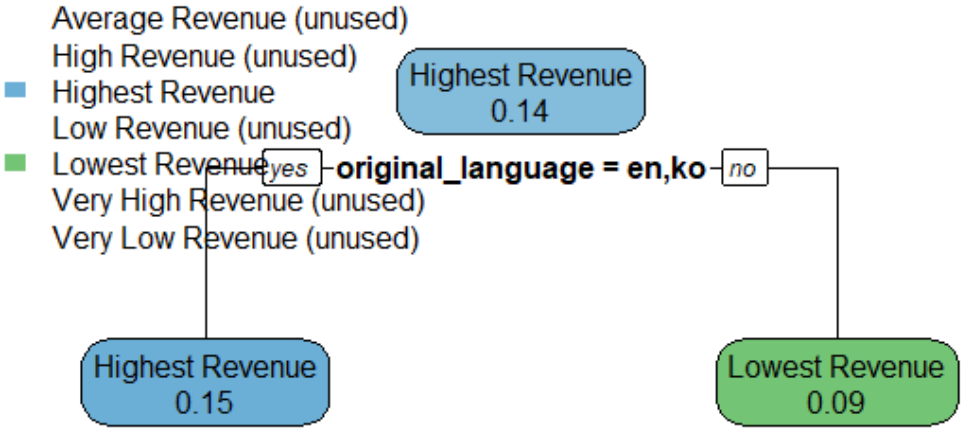
The first table is showing the five rules that had the highest support. The second table is showing the five rules that had the highest confidence and lift. In fact, there was a total of 45 rules generated in this process and 44 of them had the exact same confidence and lift. So instead of showing all 44 here are five random ones. To find the one rule different from the rest look at the first row in the table sorted by support. Since all these rules were extraordinarily strong, it seemed like a fine idea to pick the rule different from the rest as it has a uniqueness that could be promising for the final movie.

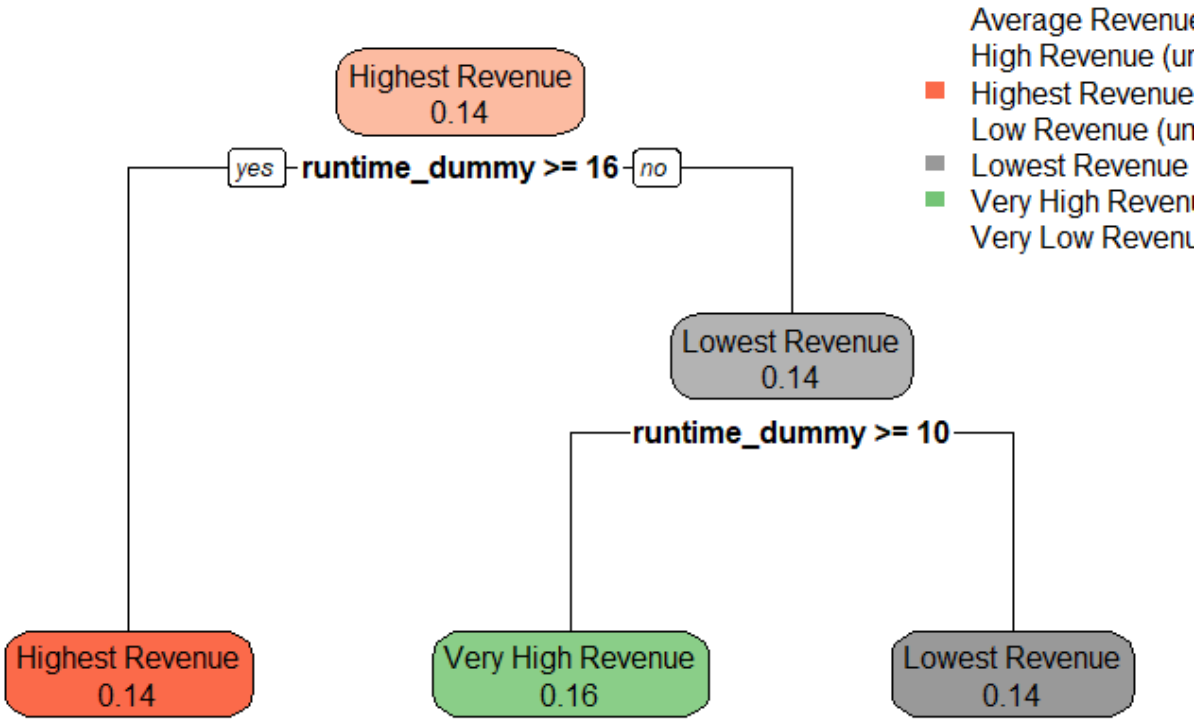
The genre that the new movie will have will be adventure, drama, fantasy, and romance. Also, of note this rule had four occurrences and these four movies happen to all be a part of the Twilight saga. So, now the new movie has a genre, but it also has a universe it can join to help boost its revenue.

Results – Decision Trees

Decision trees were used to decide on the new movie's specifications. Specifically, the specifications were the budget, the language, the release month, and lastly the runtime. Below are the four decision trees that were generated from the analysis.







The first decision tree is deciding the budget. Unsurprisingly the best way to make a movie that ends with a high revenue is to start with a high budget. Upon examination of the other Twilight movies it can be seen that their budget's range from 37 million to 120 million and since only the upper part of this falls within the highest budget range, the new movie will have a budget of 90 million dollars.

The second decision tree deals with what language the movie should be made in. There were two languages that could lead to a highest revenue movie, English and Korean. To remain on par with the other Twilight movies the new one will be made in English as well.

The third decision tree is the release month of a movie. Out of all the decision trees this one had the most options for what could lead to a highest revenue movie. Once again, the other Twilight movies were used to make a final decision on when to release the new movie. Since 4 out of 5 of the Twilight movies have been released in November and November is one of the months that leads to a highest revenue movie, November will be the release month of the new movie.

The final decision tree was the movies runtime. For this a long movie lead to a highest revenue movie. Specifically, a movie that had a runtime between 122 minutes and 338 minutes lead to a highest revenue. The other Twilight movies had an average runtime of 122.75 minutes, so to remain on par with those movies the new movie will have a runtime of 123 minutes.

Results – Text Mining

This new movie is well on its way to becoming a Hollywood success story. The final aspect of the movie is its overview. This is the written text that describes the movie in enough detail so people understand what it is about, but leaves enough mystery to get a potential viewer interested in going to a theatre to watch the movie and find out what happens.

The text mining that was used was word frequency. All 4,000+ other movies overviews were analyzed, and in the end the movies that had the highest and lowest revenues were compared. The goal was to find out which words occurred frequently in a highest revenue movie overview, so those words could be used in the new movie overview. As well, the lowest revenue movie overviews were examined to know which words should not be used in the new movie overview. The table below depicts the most frequently used words of the highest and lowest revenue movies.

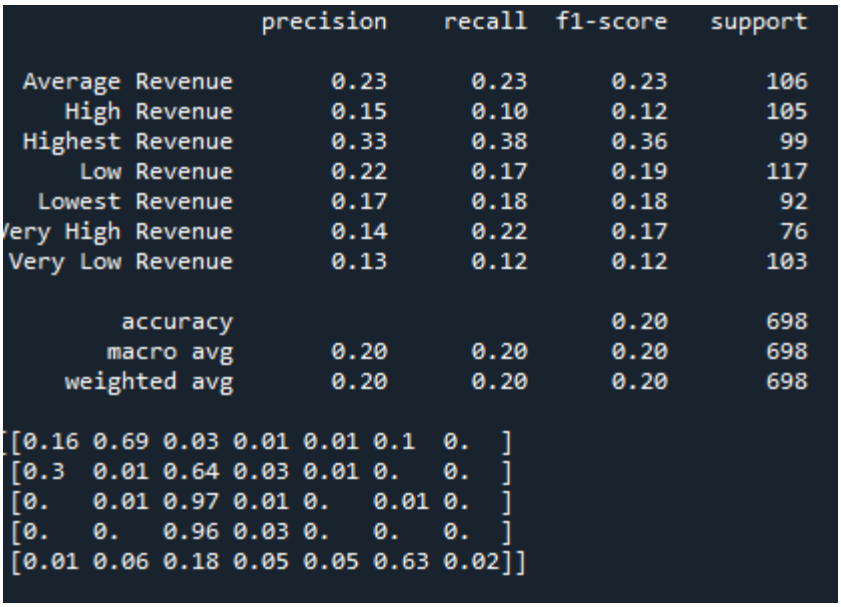
Highest Revenue Words	Word Frequency	Lowest Revenue Words	Word Frequency
World	112	Life	93
New	100	New	73
Life	88	Young	71
Man	56	World	67
Family	50	Man	64
Help	46	Love	60
Earth	45	Family	59
Young	45	Story	49
Father	42	Fil	46
Love	42	Old	46

There are seven words that occur frequently in both the highest revenue movie overviews and the lowest revenue movie overviews. With the goal to not confuse the models these words will not be used unless it is necessary. The words that will be used are Help, Earth and Father as the words occur frequently in highest revenue movie overviews. The words that will not be used are Story, Film and Old as thy occur frequently in the lowest revenue movie overviews. Her is the newest movies overview.

Set 10 years after the Twilight Saga ended, Renesmee is now grown up and needs help with a difficult decision. Does she stay with Jacob, the one who imprinted her at birth, even though it goes against here father's wishes? Or does she get to know the new boy in town, who appears to have some secrets of his own?

Results – Multinomial Naïve Bayes

With an overview written the next step is to see what kind of revenue it may lead to. To do this a Multinomial Naïve Bayes models was developed and trained on 80% of the data, this was then tested on the rest of the data and here are the results. Below is a table of those results.



```
                  precision    recall  f1-score   support

 Average Revenue       0.23      0.23      0.23       106
    High Revenue       0.15      0.10      0.12       105
 Highest Revenue       0.33      0.38      0.36        99
     Low Revenue       0.22      0.17      0.19       117
  Lowest Revenue       0.17      0.18      0.18        92
ery High Revenue       0.14      0.22      0.17        76
Very Low Revenue       0.13      0.12      0.12       103

        accuracy                           0.20       698
       macro avg       0.20      0.20      0.20       698
    weighted avg       0.20      0.20      0.20       698

[[0.16 0.69 0.03 0.01 0.01 0.1  0.  ]
 [0.3  0.01 0.64 0.03 0.01 0.   0.  ]
 [0.   0.01 0.97 0.01 0.   0.01 0.  ]
 [0.   0.   0.96 0.03 0.   0.   0.  ]
 [0.01 0.06 0.18 0.05 0.05 0.63 0.02]]
```

This is not the best model that has ever been made, with an accuracy of only 0.20 but it will still be used to predict the revenue of the new movie. Below is a table breakdown of the other four Twilight films in the data set and the new movie, with percentage likelihood predictions for which revenue category the movie will be in based on its overview.

Movie	**Lowest Revenue**	**Very Low Revenue**	**Low Revenue**	**Average Revenue**	**High Revenue**	**Very High Revenue**	**Highest Revenue**
Sunrise	0.01	0.00	0.01	0.16	0.69	0.01	0.03
Twilight	0.01	0.00	0.0	0.30	0.01	0.00	0.64
New Moon	0.00	0.00	0.01	0.00	0.01	0.01	0.97
Eclipse	0.00	0.00	0.03	0.00	0.00	0.00	0.96
Break Dawn – Part 2	0.05	0.05	0.05	0.01	0.06	0.63	0.18

This model correctly predicted the revenue for 3 out of 4 of the other Twilight movies. Increasing the accuracy from 0.20 to 0.75, and thus raising the confidence level for the model's ability to correctly predict how well the new movie will perform. Therefore, seeing a prediction of high revenue for the new movie is awesome and believable.

Having a high revenue will mean the movie will make somewhere between 71 million and 122 million. Which means there is potential that the movie could fall under budget, but therefore analytics was used on more than just the overview. This movie has way more going for it than just the overview, it has all those movie specifications locked in, therefore it is highly likely that the movie will make more than the 90 million dollars and maybe make closer to 120 million dollars so that it is a success and a profit is made.

Conclusion

The journey began with a simple goal, become a Hollywood success story. Writing a movie seemed to be the simplest way to do this, as analytics could be used to develop a movie that would be a success. Along the way many different analytical methods were used, and all but one of these did predict the new movie would fall into the category of a highest revenue movie. The one movie aspect that fell short was the review, which was predicted to only yield a high revenue movie. Even with this minor set back the goal was still met, as the movie could return a 30-million-dollar profit, which is a Hollywood success.

It was a lofty goal, but it does appear to have been achieved. The only way to know for sure is to see audiences in theatres in November of 2022 ready to watch the newest installment of the Twilight saga, Sunrise.