Muhammad Ahsan Syed
ms02743

Data Science Final Project Report

Introduction:

This project will try to gain some insights on movies from the dataset of movies we obtained from grouplens.org and TMBD. We will be trying to answer some broader questions like what are common features within movies, how does budget play role, how does popularity come into account etc. We will then final try to predict the revenue of a movie based on the rest of its features.

Who:

This dataset was collected from TMBD which is a popular database for movies. A major chunk of this dataset also comes from Grouplens, which is a social computing research lab.

Need:

Primarily, this data was collected so that an exploratory data analysis can be carried out narrate story and the history of cinema throughout time

We will look at impact of actors and directors on movies.
We are going to also analyze the impact of budget and genre on revenue
We will look at different movies of franchises comparing it to movies that don't belong to any franchise
We will look at popularity and voting of movies and try to contrast it with its success
Will try to get insights on the best time period for a movie to be released
We will also try to test whether longer running time movies are successful or not
We will test some other hypothesis as well like about Walt Disney studio

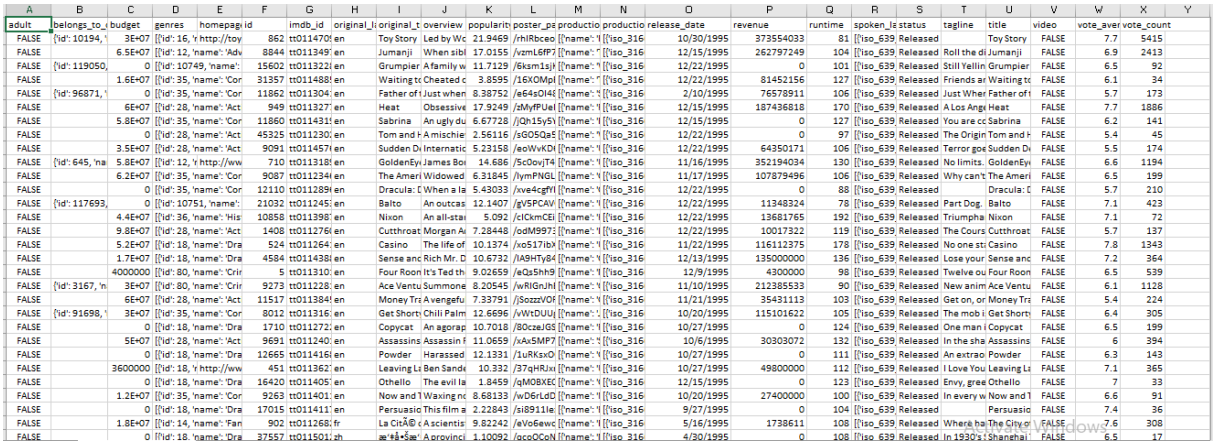
Through this dataset it is also possible to predict movie revenue/success as well as a classifier for letting us know whether a movie will generate a profit or loss.

Requirement and Resources needed:

For this I will be using Python for data analysis. I will be using libraries like pandas, matplotlib, seaborn, numpy, sklearn etc. I will be using a local jupyter notebook environment to run this project.

Dataset Description:

The dataset that we will be looking at was obtained through the TMDB API. The movies available in this dataset are in correspondence with the movies that are listed in the MovieLens Dataset so this is the resulting dataset we got:



adult	belongs_to_c	budget	genres	homepage	id	imdb_id	original_la	original_t	overview	popularity	poster_pa	productio	productio	release_date	revenue	runtime	spoken_la	status	tagline	title	video	vote_aver	vote_count
FALSE	{'id': 10194, '	3E+07	[{'id': 16, '	http://toy	862	tt011470	en	Toy Story	Led by Wc	21.9469	/rhIRbceo	[{'name': '	[{'iso_316	10/30/1995	373554033	81	[{'iso_639	Released		Toy Story	FALSE	7.7	5415
FALSE		6.5E+07	[{'id': 12, '	name': 'Adv	8844	tt011349	en	Jumanji	When sibl	17.0155	/vzmL6fP7	[{'name': '	[{'iso_316	12/15/1995	262797249	104	[{'iso_639	Released	Roll the di	Jumanji	FALSE	6.9	2413
FALSE	{'id': 119050,	0	[{'id': 10749, 'name':		15602	tt011322	en	Grumpier	A family w	11.7129	/6ksm1sjK	[{'name': '	[{'iso_316	12/22/1995	0	101	[{'iso_639	Released	Still Yellin	Grumpier	FALSE	6.5	92
FALSE		1.6E+07	[{'id': 35, '	name': 'Cor	31357	tt011488	en	Waiting to	Cheated c	3.8595	/16KOMpI	[{'name': '	[{'iso_316	12/22/1995	81452156	127	[{'iso_639	Released	Friends ar	Waiting to	FALSE	6.1	34
FALSE	{'id': 96871, '	0	[{'id': 35, '	name': 'Cor	11862	tt011304	en	Father of	Just when	8.38752	/e64sOI4K	[{'name': '	[{'iso_316	2/10/1995	76578911	106	[{'iso_639	Released	Just Wher	Father of t	FALSE	5.7	173
FALSE		6E+07	[{'id': 28, '	name': 'Act	949	tt011327	en	Heat	Obsessive	17.9249	/aMyfPUeI	[{'name': '	[{'iso_316	12/15/1995	187436818	170	[{'iso_639	Released	A Los Ange	Heat	FALSE	7.7	1886
FALSE		5.8E+07	[{'id': 35, '	name': 'Cor	11860	tt011431	en	Sabrina	An ugly du	6.67728	/jQh15y5y	[{'name': '	[{'iso_316	12/15/1995	0	127	[{'iso_639	Released	You are cc	Sabrina	FALSE	6.2	141
FALSE		0	[{'id': 28, '	name': 'Act	45325	tt011230	en	Tom and H	A mischie	2.56116	/sGO5Qa5	[{'name': '	[{'iso_316	12/22/1995	0	97	[{'iso_639	Released	The Origin	Tom and H	FALSE	5.4	45
FALSE		3.5E+07	[{'id': 28, '	name': 'Act	9091	tt011457	en	Sudden De	Internatic	5.23158	/eoWvKDI	[{'name': '	[{'iso_316	12/22/1995	64350171	106	[{'iso_639	Released	Terror goe	Sudden De	FALSE	5.5	174
FALSE	{'id': 645, 'na	5.8E+07	[{'id': 12, '	http://ww	710	tt011318	en	GoldenEy	James Bo	14.686	/5cOovjT4	[{'name': '	[{'iso_316	11/16/1995	352194034	130	[{'iso_639	Released	No limits.	GoldenEy	FALSE	6.6	1194
FALSE		6.2E+07	[{'id': 35, '	name': 'Cor	9087	tt011234	en	The Ameri	Widowed	6.31845	/lymPNGL	[{'name': '	[{'iso_316	11/17/1995	107879496	106	[{'iso_639	Released	Why can't	The Ameri	FALSE	6.5	199
FALSE		0	[{'id': 35, '	name': 'Cor	12110	tt011289	en	Dracula: D	When a la	5.43033	/xve4cgfY	[{'name': '	[{'iso_316	12/22/1995	0	88	[{'iso_639	Released		Dracula: D	FALSE	5.7	210
FALSE	{'id': 117693,	0	[{'id': 10751, 'name':		21032	tt011245	en	Balto	An outcas	12.1407	/gV5PCAV	[{'name': '	[{'iso_316	12/22/1995	11348324	78	[{'iso_639	Released	Part Dog.	Balto	FALSE	7.1	423
FALSE		4.4E+07	[{'id': 36, '	name': 'His	10858	tt011398	en	Nixon	An all-sta	5.092	/cICkmCEi	[{'name': '	[{'iso_316	12/22/1995	13681765	192	[{'iso_639	Released	Triumpha	Nixon	FALSE	7.1	72
FALSE		9.8E+07	[{'id': 28, '	name': 'Act	1408	tt011276	en	Cutthroat	Morgan A	7.28448	/odM997S	[{'name': '	[{'iso_316	12/22/1995	10017322	119	[{'iso_639	Released	The Cours	Cutthroat	FALSE	5.7	137
FALSE		5.2E+07	[{'id': 18, '	name': 'Dra	524	tt011264	en	Casino	The life of	10.1374	/xo517ibX	[{'name': '	[{'iso_316	11/22/1995	116112375	178	[{'iso_639	Released	No one st	Casino	FALSE	7.8	1343
FALSE		1.7E+07	[{'id': 18, '	name': 'Dra	4584	tt011438	en	Sense and	Rich Mr. D	10.6732	/lA9HTy8	[{'name': '	[{'iso_316	12/13/1995	135000000	136	[{'iso_639	Released	Lose your	Sense and	FALSE	7.2	364
FALSE		4000000	[{'id': 80, '	name': 'Crir	5	tt011310	en	Four Room	It's Ted th	9.02659	/eQs5hh9	[{'name': '	[{'iso_316	12/9/1995	4300000	98	[{'iso_639	Released	Twelve ou	Four Room	FALSE	6.5	539
FALSE	{'id': 3167, 'n	3E+07	[{'id': 80, '	name': 'Crir	9273	tt011228	en	Ace Ventu	Summone	8.20545	/wRIGnJh	[{'name': '	[{'iso_316	11/10/1995	212385533	90	[{'iso_639	Released	New anim	Ace Ventu	FALSE	6.1	1128
FALSE		6E+07	[{'id': 28, '	name': 'Act	11517	tt011384	en	Money Tra	A vengefu	7.33791	/jSozzzVO	[{'name': '	[{'iso_316	11/21/1995	35431113	103	[{'iso_639	Released	Get on, or	Money Tra	FALSE	5.4	224
FALSE	{'id': 91698, '	3E+07	[{'id': 35, '	name': 'Cor	8012	tt011316	en	Get Short	Chili Palm	12.6696	/vWtDUU	[{'name': '	[{'iso_316	10/20/1995	115101622	105	[{'iso_639	Released	The mob i	Get Short	FALSE	6.4	305
FALSE		0	[{'id': 18, '	name': 'Dra	1710	tt011272	en	Copycat	An agorap	10.7018	/80czeJG5	[{'name': '	[{'iso_316	10/27/1995	0	124	[{'iso_639	Released	One man i	Copycat	FALSE	6.5	199
FALSE		5E+07	[{'id': 28, '	name': 'Act	9691	tt011240	en	Assassins	Assassin f	11.0659	/xAx5MP7	[{'name': '	[{'iso_316	10/6/1995	30303072	132	[{'iso_639	Released	In the sha	Assassins	FALSE	6	394
FALSE		0	[{'id': 18, '	name': 'Dra	12665	tt011416	en	Powder	Harassed	12.1331	/1uRKsxO	[{'name': '	[{'iso_316	10/27/1995	0	111	[{'iso_639	Released	An extrao	Powder	FALSE	6.3	143
FALSE		3600000	[{'id': 18, '	http://ww	451	tt011362	en	Leaving La	Ben Sande	10.332	/37qHRJxI	[{'name': '	[{'iso_316	10/27/1995	49800000	112	[{'iso_639	Released	I Love You	Leaving La	FALSE	7.1	365
FALSE		0	[{'id': 18, '	name': 'Dra	16420	tt011405	en	Othello	The evil la	1.8459	/qM0BXEC	[{'name': '	[{'iso_316	12/15/1995	0	123	[{'iso_639	Released	Envy, gree	Othello	FALSE	7	33
FALSE		1.2E+07	[{'id': 35, '	name': 'Cor	9263	tt011401	en	Now and T	Waxing nc	8.68133	/wD6rLdD	[{'name': '	[{'iso_316	10/20/1995	27400000	100	[{'iso_639	Released	In every w	Now and T	FALSE	6.6	91
FALSE		0	[{'id': 18, '	name': 'Dra	17015	tt011411	en	Persuasio	This film a	2.22843	/si8911le	[{'name': '	[{'iso_316	9/27/1995	0	104	[{'iso_639	Released		Persuasio	FALSE	7.4	36
FALSE		1.8E+07	[{'id': 14, '	name': 'Fan	902	tt011268	fr	La CitÃ©	A scientis	9.82242	/eVo6ewc	[{'name': '	[{'iso_316	5/16/1995	1738611	108	[{'iso_639	Released	Where ha	The City o	FALSE	7.6	308
FALSE		0	[{'id': 18, '	name': 'Dra	37557	tt011501	zh	æ·±å¤	A provinci	1.10092	/qcoOCoN	[{'name': '	[{'iso_316	4/30/1995	0	108	[{'iso_639	Released	In 1930's	Shanghai	FALSE	6.5	17

Figure 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45466 entries, 0 to 45465
Data columns (total 24 columns):
adult                    45466 non-null object
belongs_to_collection    4494 non-null object
budget                   45466 non-null object
genres                   45466 non-null object
homepage                 7782 non-null object
id                       45466 non-null object
imdb_id                  45449 non-null object
original_language        45455 non-null object
original_title           45466 non-null object
overview                 44512 non-null object
popularity               45461 non-null object
poster_path              45080 non-null object
production_companies     45463 non-null object
production_countries     45463 non-null object
release_date             45379 non-null object
revenue                  45460 non-null float64
runtime                  45203 non-null float64
spoken_languages         45460 non-null object
status                   45379 non-null object
tagline                  20412 non-null object
title                    45460 non-null object
video                    45460 non-null object
vote_average             45460 non-null float64
vote_count               45460 non-null float64
dtypes: float64(4), object(20)
memory usage: 8.3+ MB
```

Figure 2

Looking at the raw file of the dataset in figure 1 as well as a description of it, we can get an idea about the data. There are 24 columns in the our dataset with over 45,000+ values. Almost all the columns, as we can see from figure 2, are of type object whereas columns like revenue, runtime, vote_average, vote_count are float. We also see that most columns have quite less NaN values and is seen mostly in homepage and tagline columns.

	revenue	runtime	vote_average	vote_count
count	4.546000e+04	45203.000000	45460.000000	45460.000000
mean	1.120935e+07	94.128199	5.618207	109.897338
std	6.433225e+07	38.407810	1.924216	491.310374
min	0.000000e+00	0.000000	0.000000	0.000000
25%	0.000000e+00	85.000000	5.000000	3.000000
50%	0.000000e+00	95.000000	6.000000	10.000000
75%	0.000000e+00	107.000000	6.800000	34.000000
max	2.787965e+09	1256.000000	10.000000	14075.000000

Figure 3

In case of these float values, figure 3 gives us an idea about the disparity of values among these.

```
: Index(['adult', 'belongs_to_collection', 'budget', 'genres', 'homepage', 'id',
       'imdb_id', 'original_language', 'original_title', 'overview',
       'popularity', 'poster_path', 'production_companies',
       'production_countries', 'release_date', 'revenue', 'runtime',
       'spoken_languages', 'status', 'tagline', 'title', 'video',
       'vote_average', 'vote_count'],
      dtype='object')
```

Figure 4

Figure 4 shows us the name of each column in the dataset.

adult Indicates if the movie is X-Rated or Adult. **belongs_to_collection is** a stringified dictionary that gives information on the movie series the particular film belongs **Budget** column refers to the budget of movie in dollars. **Genre** is stringified list of dictionaries that list out all the genres associated with the movie **homepage** gives the Homepage of the move. I**D** is the unique identifier of the move. **Imdb_id** is the he IMDB ID of the movie. **Original_language** shows the native language of the movie. **Original_title** is the original title of the movie. **Overview** is a brief blurb of the movie. **Popularity** shows the popularity Score assigned by TMDB. **poster_path** is the URL of the poster image. **production_companies** is a stringified list of production companies involved with the making of the movie. **production_countries** is A stringified list of countries where the movie was shot/produced in. **release_date** is the theatrical Release Date of the movie. **Revenue** shows the total revenue of the movie in dollars. **Runtime** gives the runtime of the movie in minutes. **spoken_languages:** A stringified list of spoken languages in the film. **Status** shows the status of the movie. **Tagline** has the tagline of the movie. **Title** has the official Title of the movie. **Video:** Indicates if there is a video present of the movie with

TMDB. **Vote_average** shows the average rating of the movie. **Vote_count** gives the number of votes by users, as counted by TMDB.

Before we move into analysis of our dataset, we first get rid of extra values. We get rid of extra columns that are of no use to us like original_title, imbd_id and adult. We will also fix some values like adding NaN values, fixing datatypes and also adding a new column called 'return' which is basically the ratio between revenue and budget. This return column will actually give us a better picture of the success of any movie. If this return values I positive then it would mean that movie earned some profit. If it is in negative then it would mean that movie is went into loss.

Results/Findings:

1. Do cast and crew play vital roles on the success of movies?

We will now look at the credits dataset which has information about casts and crews.

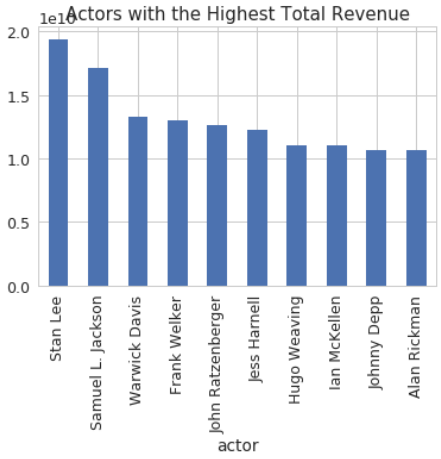


Figure 6

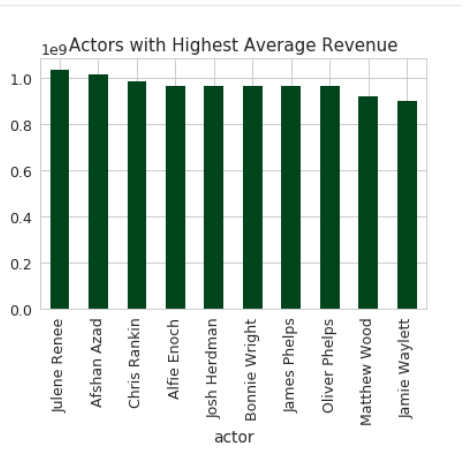


Figure 5

By looking at the actors and the highest revenue the movies generated in which they starred in, we see Stan Lee and Samuel L. Jackson at the top of the list getting over 1.5 billion dollars.

The average of 5 movies gives us a different list however, it may be because these actors belong to a movie franchise that have performed well and are reoccuring in movies.

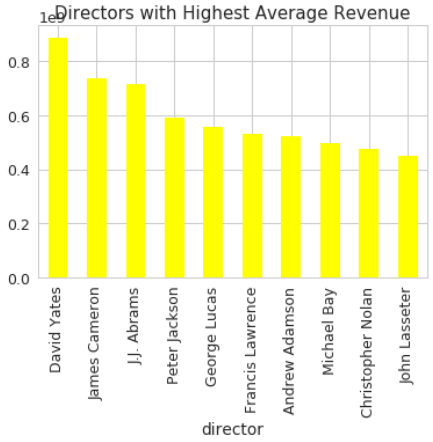


Figure 8

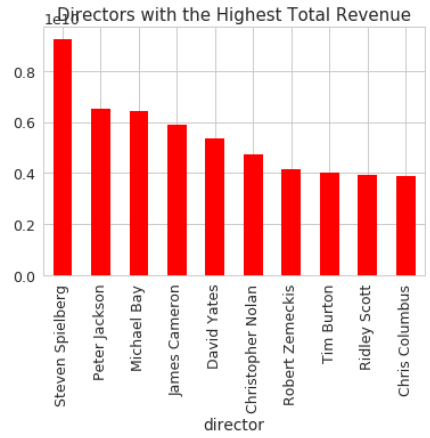


Figure 7

In case of directors, we will look at the highest earnings movies which they have directed as well as the highest average earning ones with at least 5 movies. The average graph will give us more of an idea of how successful directors have been generally. Steven Spielberg, Peter Jackson and Michael Bay appear to be at the top of the list making over half a billion which gives us an idea of whom to look out for, for the next big hit.

2. Does it mean that if a movie has a lot of budget, it is more likely to generate more revenue?

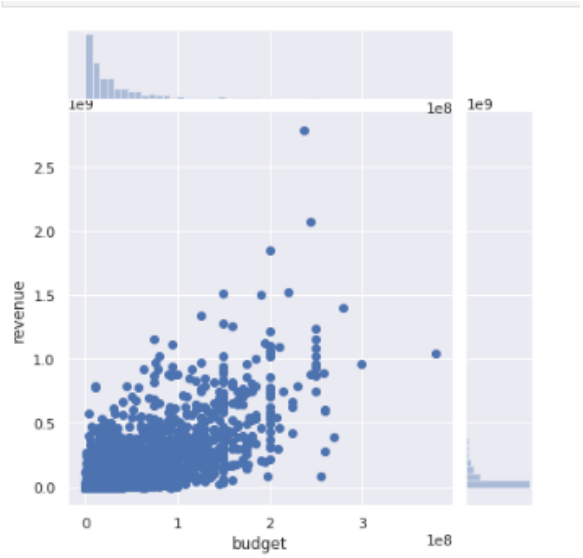


Figure 9

From figure 9 we can see that for most of the movies that have a budget of 0 million to 15 million have made revenues around 0 to 0.5 billion approximately. Most of the movies lie in this chunk whereas we see some movies going more than 15 million whose revenue are also increasing. However this are very few examples of it and we still see some movies even in 20 million budget having very less revenue.

3. Which genre has generated the most revenue?

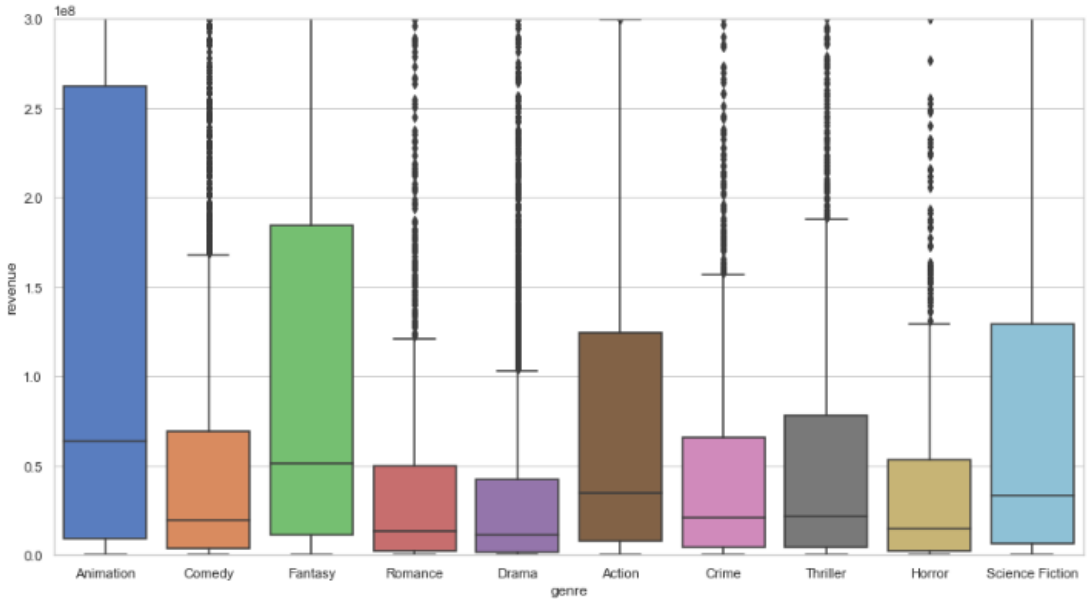


Figure 10

In terms of the dataset that we have, we see that Animation turns out to be the genre that has generated the most revenue in terms of median as well which is followed by Fantasy and Science Fiction.

Animated movies by Pixar and Disney have always been popular as they attract a wide range of audience. There is however one thing missing in figure 10. This is only looking at the total revenue and not taking into account the amount of budget

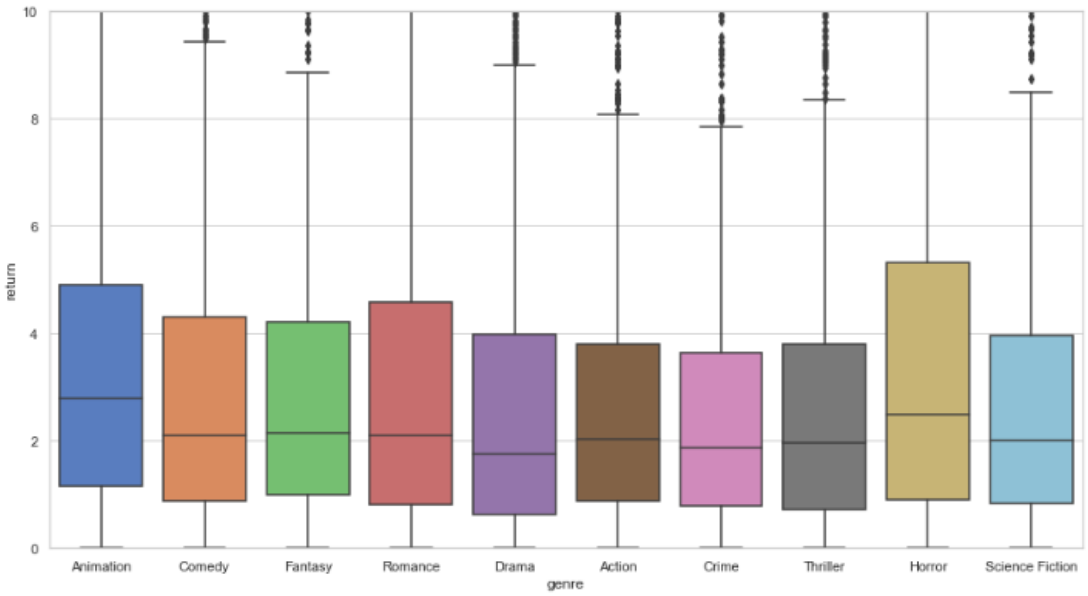


Figure 11

Figure 11 catches a more real world like scenario as we look at return (explained in 'terms' section). Even in this case, on average, Animation still has the most return. Which is then followed by horror. Which makes sense as horror movies don't have big budgets compared to other genres like fantasy/Science fiction etc and end up generating good amount of revenue which yields in a good return score

4. It is popularly believed that the media industry as of now is not driven by individual movies but rather franchises. We will try to answer some of that here

	belongs_to_collection	count	mean	sum
552	Harry Potter Collection	8	9.634209e+08	7.707367e+09
1160	Star Wars Collection	8	9.293118e+08	7.434495e+09
646	James Bond Collection	26	2.733450e+08	7.106970e+09
1317	The Fast and the Furious Collection	8	6.406373e+08	5.125099e+09
968	Pirates of the Caribbean Collection	5	9.043154e+08	4.521577e+09
1550	Transformers Collection	5	8.732202e+08	4.366101e+09
325	Despicable Me Collection	4	9.227676e+08	3.691070e+09
1491	The Twilight Collection	5	6.684215e+08	3.342107e+09
610	Ice Age Collection	5	6.433417e+08	3.216709e+09
666	Jurassic Park Collection	4	7.578710e+08	3.031484e+09
1111	Shrek Collection	4	7.389518e+08	2.955807e+09
1359	The Hunger Games Collection	4	7.360407e+08	2.944163e+09
1352	The Hobbit Collection	3	9.785078e+08	2.935523e+09
1245	The Avengers Collection	2	1.462481e+09	2.924962e+09
1388	The Lord of the Rings Collection	3	9.721816e+08	2.916545e+09
1656	X-Men Collection	6	4.681387e+08	2.808832e+09
112	Avatar Collection	1	2.787965e+09	2.787965e+09
835	Mission: Impossible Collection	5	5.557956e+08	2.778978e+09
1146	Spider-Man Collection	3	8.321155e+08	2.496347e+09
1290	The Dark Knight Collection	3	8.212387e+08	2.463716e+09

Figure 12

Let us first take a look at figure 12 first. With an estimated total revenue of 7.7 billion dollars from just 8 movies, Harry Potter franchise takes the lead with ever-so-famous franchises like star wars and james bond following.

	belongs_to_collection	count	mean	sum
112	Avatar Collection	1	2.787965e+09	2.787965e+09
1245	The Avengers Collection	2	1.462481e+09	2.924962e+09
479	Frozen Collection	1	1.274219e+09	1.274219e+09
446	Finding Nemo Collection	2	9.844532e+08	1.968906e+09
1352	The Hobbit Collection	3	9.785078e+08	2.935523e+09
1388	The Lord of the Rings Collection	3	9.721816e+08	2.916545e+09
552	Harry Potter Collection	8	9.634209e+08	7.707367e+09
1160	Star Wars Collection	8	9.293118e+08	7.434495e+09
325	Despicable Me Collection	4	9.227676e+08	3.691070e+09
968	Pirates of the Caribbean Collection	5	9.043154e+08	4.521577e+09
1457	The Secret Life of Pets Collection	1	8.754579e+08	8.754579e+08
1550	Transformers Collection	5	8.732202e+08	4.366101e+09
1146	Spider-Man Collection	3	8.321155e+08	2.496347e+09
1290	The Dark Knight Collection	3	8.212387e+08	2.463716e+09
1649	Wonder Woman Collection	1	8.205804e+08	8.205804e+08
530	Guardians of the Galaxy Collection	2	8.183724e+08	1.636745e+09
422	Fantastic Beasts Collection	1	8.093423e+08	8.093423e+08
635	Iron Man Collection	3	8.081825e+08	2.424548e+09
1383	The Lion King Collection	1	7.882418e+08	7.882418e+08
309	Deadpool Collection	1	7.831130e+08	7.831130e+08

If we were to look at revenues in terms of how each movie made (average), we some changes. From the top 20 lists, we see movies like avatar and frozen right at the top which are, according to the dataset, standalone movies. However we can still see many franchises up there.

	title	sum
3398	Avatar	2.787965e+09
28846	Star Wars: The Force Awakens	2.068224e+09
38665	Titanic	1.849939e+09
4093	Beauty and the Beast	1.689337e+09
2032	Alice in Wonderland	1.597491e+09
30694	The Avengers	1.568143e+09
16855	Jurassic World	1.513529e+09
12321	Furious 7	1.506249e+09
3405	Avengers: Age of Ultron	1.405404e+09
13789	Harry Potter and the Deathly Hallows: Part 2	1.342000e+09
12249	Frozen	1.277285e+09
32603	The Fate of the Furious	1.238765e+09
16034	Iron Man 3	1.215440e+09
33979	The Jungle Book	1.172394e+09
20705	Minions	1.156731e+09
6189	Captain America: Civil War	1.153304e+09
39105	Transformers: Dark of the Moon	1.123747e+09
34599	The Lord of the Rings: The Return of the King	1.118889e+09
27931	Skyfall	1.108561e+09
39104	Transformers: Age of Extinction	1.091405e+09

Figure 13

Finally, we will look at movies alone rather than them being part of collections. Even in this list, there are some anomalies like avatar which is a single movie ruling at the top. But the fact that most of the movies in the top 20 of this list already are from franchises of movies may support the popular belief about franchise movies doing better.

5. Are the movies that generate a lot of money also popular?

We saw many movies who did quite good in terms of revenues. But does that mean they are also popular among the masses?

	title	popularity	year
30700	Minions	547.488298	2015
33356	Wonder Woman	294.337037	2017
42222	Beauty and the Beast	287.253654	2017
43644	Baby Driver	228.032744	2017
24455	Big Hero 6	213.849907	2014
26564	Deadpool	187.860492	2016
26566	Guardians of the Galaxy Vol. 2	185.330992	2017
14551	Avatar	185.070892	2009
24351	John Wick	183.870374	2014
23675	Gone Girl	154.801009	2014
24873	The Hunger Games: Mockingjay - Part 1	147.098006	2014
44274	War for the Planet of the Apes	146.161786	2017
26567	Captain America: Civil War	145.882135	2016
292	Pulp Fiction	140.950236	1994
26560	Pirates of the Caribbean: Dead Men Tell No Tales	133.827820	2017
12481	The Dark Knight	123.167259	2008
536	Blade Runner	96.272374	1982
17818	The Avengers	89.887648	2012
43286	Captain Underpants: The First Epic Movie	88.561239	2017
33361	The Circle	88.439243	2017

Figure 14

The votes from the TMBD definitely show us some other picture. It seems that movies like minions and Wonder Woman are on top of this list, which were not in the list of successful movies by revenue. There are other movies here as well like deadpool, beauty and the beast, Big hero 6, John wick and so on. However there are some movies which come here again like Avengers and Avatar but mostly we can see that movies that have good revenue don't necessarily mean that they are popular.

6. Can the same be said about most voted movies and revenue?

	title	vote_count	year
15480	Inception	14075.0	2010
12481	The Dark Knight	12269.0	2008
14551	Avatar	12114.0	2009
17818	The Avengers	12000.0	2012
26564	Deadpool	11444.0	2016
22879	Interstellar	11187.0	2014
20051	Django Unchained	10297.0	2012
23753	Guardians of the Galaxy	10014.0	2014
2843	Fight Club	9678.0	1999
18244	The Hunger Games	9634.0	2012
26553	Mad Max: Fury Road	9629.0	2015
18252	The Dark Knight Rises	9263.0	2012
2458	The Matrix	9079.0	1999
12588	Iron Man	8951.0	2008
20830	Iron Man 3	8951.0	2013
4863	The Lord of the Rings: The Fellowship of the Ring	8892.0	2001
25084	Jurassic World	8842.0	2015
292	Pulp Fiction	8670.0	1994
19971	The Hobbit: An Unexpected Journey	8427.0	2012
314	The Shawshank Redemption	8358.0	1994

Figure 15

Even in this list, we see many different movies then the revenue ones indicating that more venue making movies are not always the most voted ones.

	title	vote_average	vote_count	year
314	The Shawshank Redemption	8.5	8358.0	1994
834	The Godfather	8.5	6024.0	1972
2211	Life Is Beautiful	8.3	3643.0	1997
5481	Spirited Away	8.3	3968.0	2001
1152	One Flew Over the Cuckoo's Nest	8.3	3001.0	1975
1176	Psycho	8.3	2405.0	1960
2843	Fight Club	8.3	9678.0	1999
1178	The Godfather: Part II	8.3	3418.0	1974
12481	The Dark Knight	8.3	12269.0	2008
292	Pulp Fiction	8.3	8670.0	1994
23673	Whiplash	8.3	4376.0	2014
522	Schindler's List	8.3	4436.0	1993
1154	The Empire Strikes Back	8.2	5998.0	1980
1161	12 Angry Men	8.2	2130.0	1957
1170	GoodFellas	8.2	3211.0	1990
2216	American History X	8.2	3120.0	1998
351	Forrest Gump	8.2	8147.0	1994
289	Leon: The Professional	8.2	4293.0	1994
2884	Princess Mononoke	8.2	2041.0	1997
18465	The Intouchables	8.2	5410.0	2011

Figure 16

Even if we consider average vote. The most average votes gaining movies are also very different from the top grossing movies. However, one can argue that most these movies are quite old.

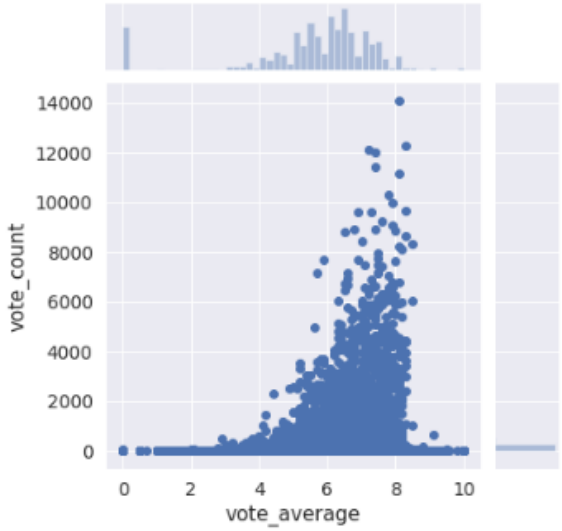


Figure 17

Even when we look at the correlation between average and count of votes, we see that more votes doesn't really mean that that movie has a good score so that should also be take into account.

7. What time periods in the year as well as in months do movies come in more often

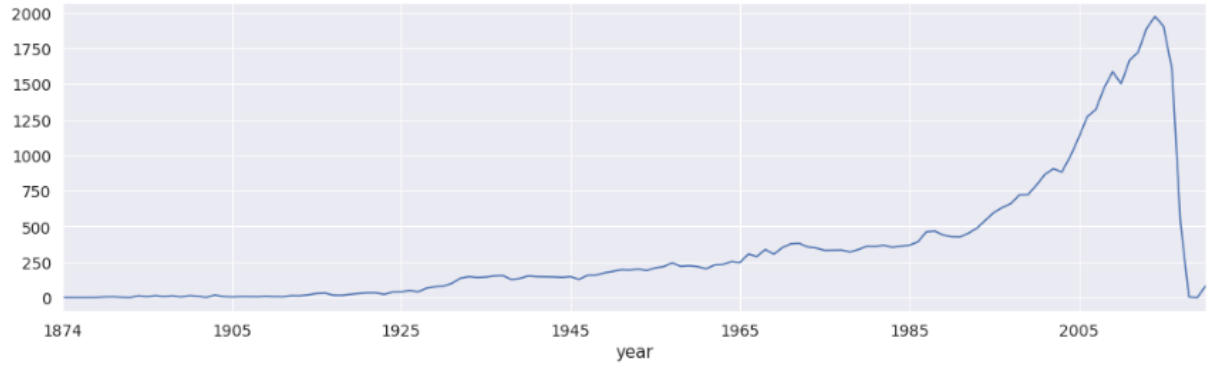


Figure 18

From figure 18 we can see after the year 1985 there is a sudden boom in the count of films in the dataset.

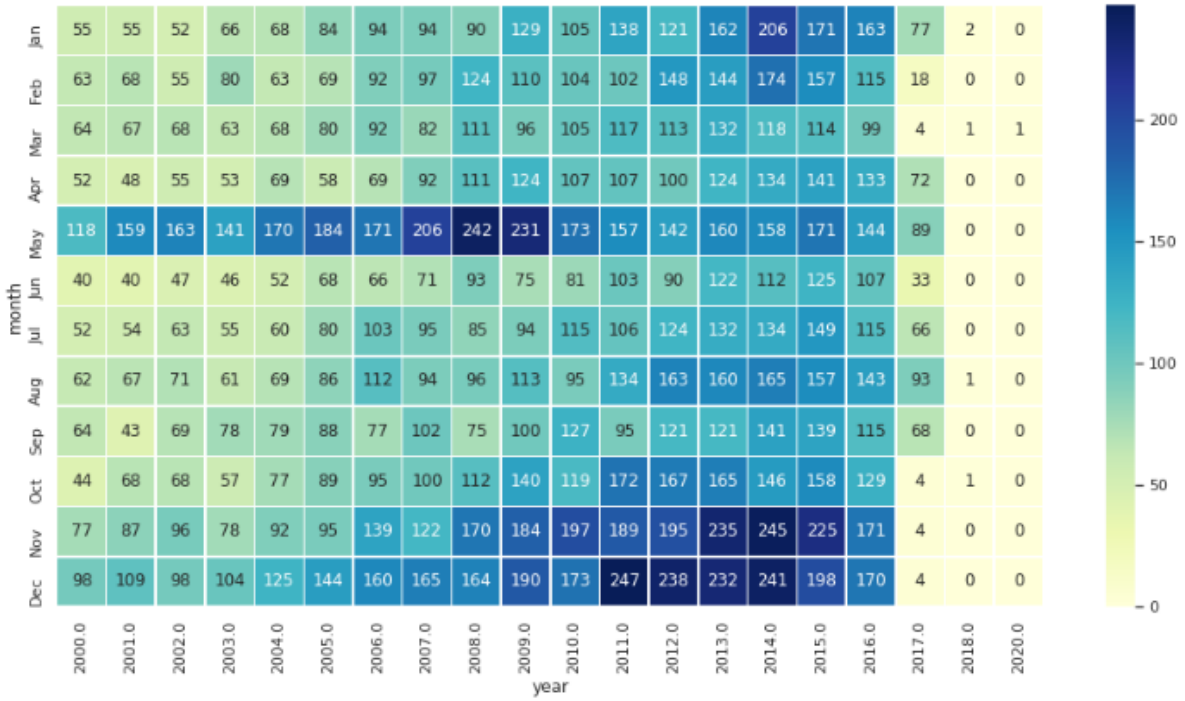


Figure 19

From figure 19 we get more clearer picture of things. The color pallets indicate the count of movies where light yellow means lesser movies and dark blue spots means more count of movies. We see a big chunk of movies in the years 2011, 2012, 2013, 2014 in the month of December.

8. Recently we have seen that movies like avengers have done pretty great even though it is approximately 3 hours long. Does this mean that longer movies are more likely to succeed.

Even though this dataset is updated till the year 2016/2017, we can still try to get an idea

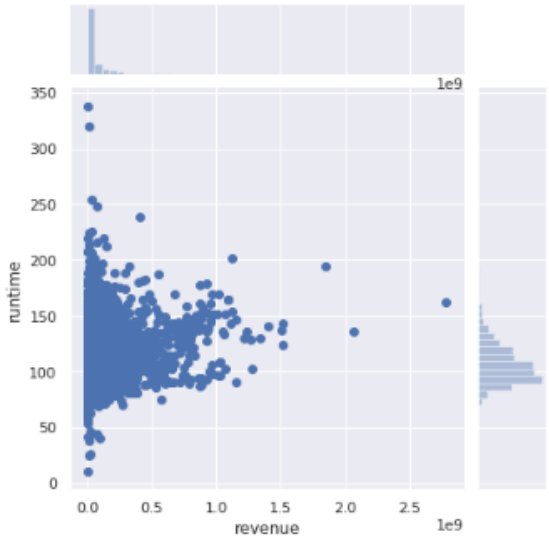


Figure 20

From the looks of it, it seems that many movies having a runtime of 100 minutes to 180 minutes reached a revenue of 1 billion and more which shows that these runtimes may be a good standard if anyone wants to get a lot of revenue.

9. In the recent times Walt Disney Studio has been acclaimed to have taken over the movie industry as they produce movies from animated ones till live actions. Are they really the best studio in terms of top grossing movies?

	Total	Average	Number
Warner Bros.	6.352519e+10	1.293792e+08	491
Universal Pictures	5.525919e+10	1.193503e+08	463
Paramount Pictures	4.880819e+10	1.235650e+08	395
Twentieth Century Fox Film Corporation	4.768775e+10	1.398468e+08	341
Walt Disney Pictures	4.083727e+10	2.778046e+08	147
Columbia Pictures	3.227974e+10	1.367785e+08	236
New Line Cinema	2.217339e+10	1.119868e+08	198
Amblin Entertainment	1.734372e+10	2.550547e+08	68
DreamWorks SKG	1.547575e+10	1.984071e+08	78
Dune Entertainment	1.500379e+10	2.419966e+08	62
Village Roadshow Pictures	1.490470e+10	1.674685e+08	89
Relativity Media	1.460315e+10	1.269839e+08	115
Touchstone Pictures	1.412178e+10	8.937839e+07	158
DreamWorks Animation	1.370752e+10	4.031622e+08	34
Legendary Pictures	1.346866e+10	3.367166e+08	40
Metro-Goldwyn-Mayer (MGM)	1.237679e+10	5.979126e+07	207
Marvel Studios	1.169964e+10	6.157703e+08	19
Columbia Pictures Corporation	1.134909e+10	8.106493e+07	140
Pixar Animation Studios	1.118853e+10	6.215852e+08	18
TSG Entertainment	1.015788e+10	2.745373e+08	37

Figure 21

We are now looking at production companies of movies in terms of how much revenue they have made. We see that WB studios leads in this case having almost 500-movie count.

	Total	Average	Number
Pixar Animation Studios	1.118853e+10	6.215852e+08	18
Marvel Studios	1.169964e+10	6.157703e+08	19
Revolution Sun Studios	8.120339e+09	5.413559e+08	15
Lucasfilm	9.898421e+09	4.499282e+08	22
DreamWorks Animation	1.370752e+10	4.031622e+08	34
DC Entertainment	6.212609e+09	3.882880e+08	16
Dentsu	6.853205e+09	3.807336e+08	18
Jerry Bruckheimer Films	8.957441e+09	3.732267e+08	24
Marvel Enterprises	6.538067e+09	3.441088e+08	19
Legendary Pictures	1.346866e+10	3.367166e+08	40
1492 Pictures	5.470574e+09	3.217985e+08	17
Di Bonaventura Pictures	6.428342e+09	3.214171e+08	20
Ingenious Film Partners	8.104726e+09	3.001750e+08	27
Atlas Entertainment	4.781741e+09	2.988588e+08	16
RatPac-Dune Entertainment	5.059895e+09	2.976409e+08	17
Donners' Company	4.741215e+09	2.963259e+08	16
Walt Disney Pictures	4.083727e+10	2.778046e+08	147
TSG Entertainment	1.015788e+10	2.745373e+08	37
Original Film	9.048042e+09	2.585155e+08	35
DC Comics	4.625644e+09	2.569802e+08	18

Figure 22

In figure 22 we have production houses again but now in terms of average revenue. Over here we see pixar and marvel ruling the stage.

Now if we were to look at everything that comes under Walt Disney Studios such as Walt Disney Pictures, Walt Disney Animation Studios, Pixar, Marvel Studios, Lucas film, 20th Century Fox, Fox 2000 Pictures, Fox Searchlight Pictures, and Blue Sky Studios then Walt Disney studio indeed does rule over film industry, especially in case of average movie revenues.

10. Classifying profit or loss

In our dataset, we have a lot of movies that have gone into loss as well. We will make a binary classifier that will tell us whether a movie will go in profit or loss

We are using revenue as the result. As we mentioned before, if revenue < 0 it means that movie went into loss and if revenue > 0 it means movie generated some profit.

We use two classifiers in this case. One is Gradient Boost Classifier while the other is Support Vector Machine Classifier.

GB Classifier gives us an accuracy of 80% whereas SVM Classifier gave us 70%.
This is because GB uses an ensamble method which is like decision trees but it uses a lot trees to make decisions which makes it quite powerful and while SVM is able to project data unto high dimensions it has its limits in the use of kernel.

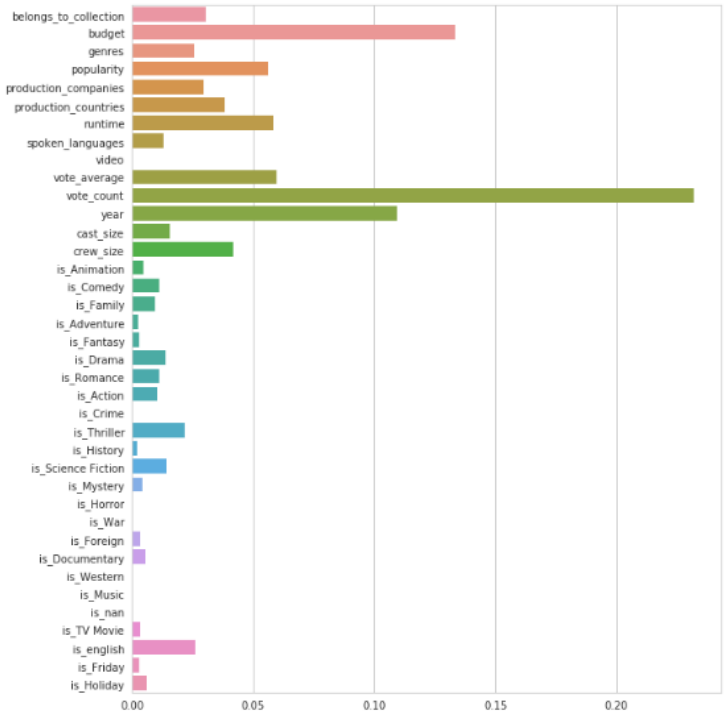


Figure 23

When we look at the features by GB classifier in figure 23, we see the most important features of the model. We see that vote is coming on to be the most important feature. Vote as well as other unimportant features should however be discarded as in real world scenarios, we will not have vote before the revenue has been generated. So ignoring that the most important features comes out to be budget, genre, production house, runtime etc.

11. Predicting revenue

We will now use a regression model to predict the revenue of a movie based on the features of the data that we have.

We already determined how powerful Gradient Boost model turned out to be in classification so we will also use a gradient boost regression model to predict revenue.

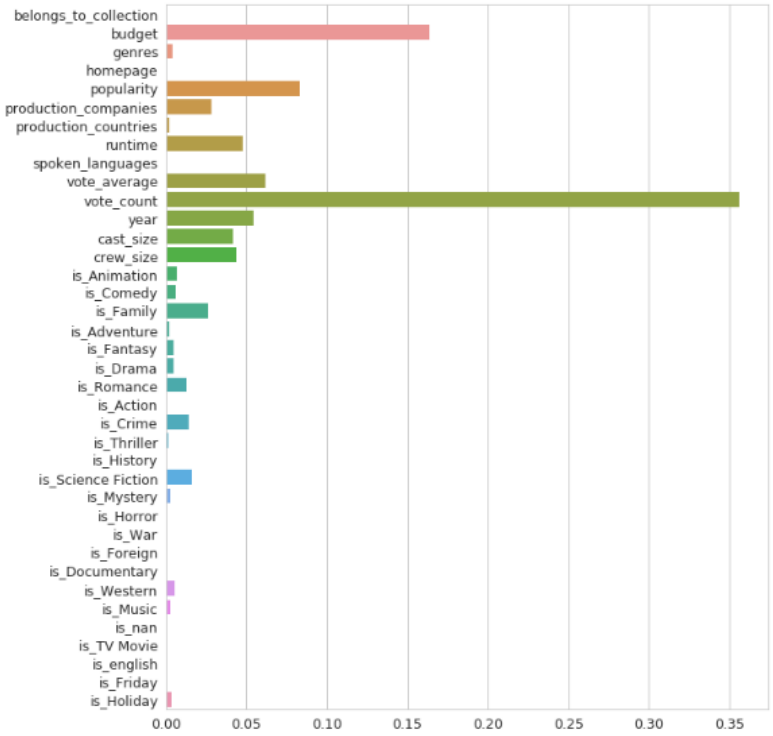


Figure 24

Once we got our model ready and trained. We inserted test dataset to test our model and got an accuracy of 77%, which is quite good. By looking at Figure 24, we see the importance of budget after ignoring votes again because of them being unrealistic in real world scenarios. So budget is a good indicator to know whether a movie will go into profit or loss. We also get other features highlighted this time like year, popularity runtime etc.

Conclusion:

In this project, we looked at the different key players that possibly determine the success of any movie. We also look at how the movie patterns have changed over the years and how the movie industry is totally governed by Walt Disney Studio.
Finally we make a regression model to predict a revenue of movie and a classifier for knowing whether a movie will make profit or go into loss.

Terms:

$$Return = \frac{Revenue}{Budget}$$

Walt Disney Studios:

Walt Disney Studios division are notable film production companies including Walt Disney Pictures, Walt Disney Animation Studios, Pixar, Marvel Studios, Lucasfilm, 20th Century Fox, Fox Searchlight Pictures and Blue Sky Studios.

References:

https://www.themoviedb.org/?language=en-US

https://www.grouplens.org/datasets/movielens/latest/

https://www.springboard.com/workshops/data-science-career-track/

https://en.wikipedia.org/wiki/Walt_Disney_Studios_(division)

Dataset Link:

https://grouplens.org/datasets/movielens/latest/