Data Science Assignment

Pothula Akash

Saiakash.pothula1@gmail.com

Question: ML model should take director name as input and predict the release year of next movie along with genres

Here we need to do two task with a single ML model

1) To predict release year (a regression task)
2) Get genres of movie (multi label classification)

Since we have to prepare a single model to do both the tasks. We have to consider this problem as Multi Task Learning Problem

Dataset preprocessing:

1) **Dataset has duplicates delete the duplicates.**
2) **Created new columns called next_title_year and next_genre that store next movie details of director as shown in the fig below**

```
sorted_df.head()
```

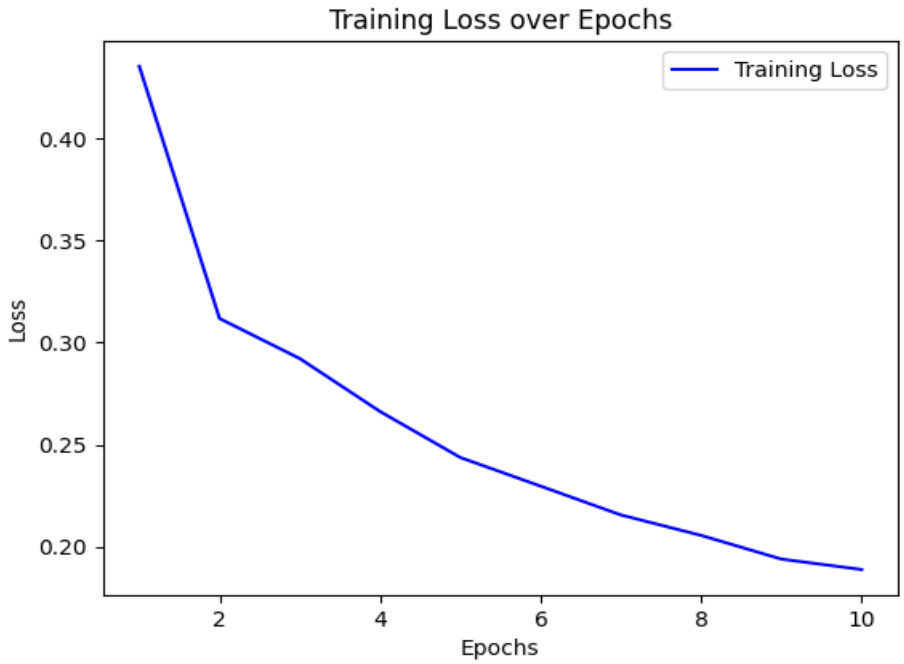
	director_name	genres	title_year	next_title_year	next_genres
0	Adam McKay	Comedy	2004.0	2006.0	Action|Comedy|Sport
1	Adam McKay	Action|Comedy|Sport	2006.0	2008.0	Comedy
2	Adam McKay	Comedy	2008.0	2010.0	Action|Comedy|Crime
3	Adam McKay	Action|Comedy|Crime	2010.0	2013.0	Comedy
4	Adam McKay	Comedy	2013.0	2015.0	Biography|Comedy|Drama|History

3) **Now split the string in the next_genres column and apply one hot encoding result will be as show in fig below.**

	director_name	title_year	next_title_year	Documentary	Action	Family	Film-Noir	Animation	Musical	News	...	Romance	Mystery	Comedy	History	Sport	Drama	War	Adventure	Thriller	Biography
0	Adam McKay	2004.0	2006.0	0	1	0	0	0	0	0	...	0	0	1	0	1	0	0	0	0	0
1	Adam McKay	2006.0	2008.0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
2	Adam McKay	2008.0	2010.0	0	1	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
3	Adam McKay	2010.0	2013.0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
4	Adam McKay	2013.0	2015.0	0	0	0	0	0	0	0	...	0	0	1	1	0	1	0	0	0	1

In addition to director name here we have to input his previous movie release date to predict his next movie release date year and genres.

Loss curve:



Results:

	Train	Test
Genres Accuracy	**43.1 %**	**32%**
Mean Absolute Error of next release year	**0.0407**	**0.038**

Accuracy is less because of less data! This result is obtained after rigorous Fine tuning

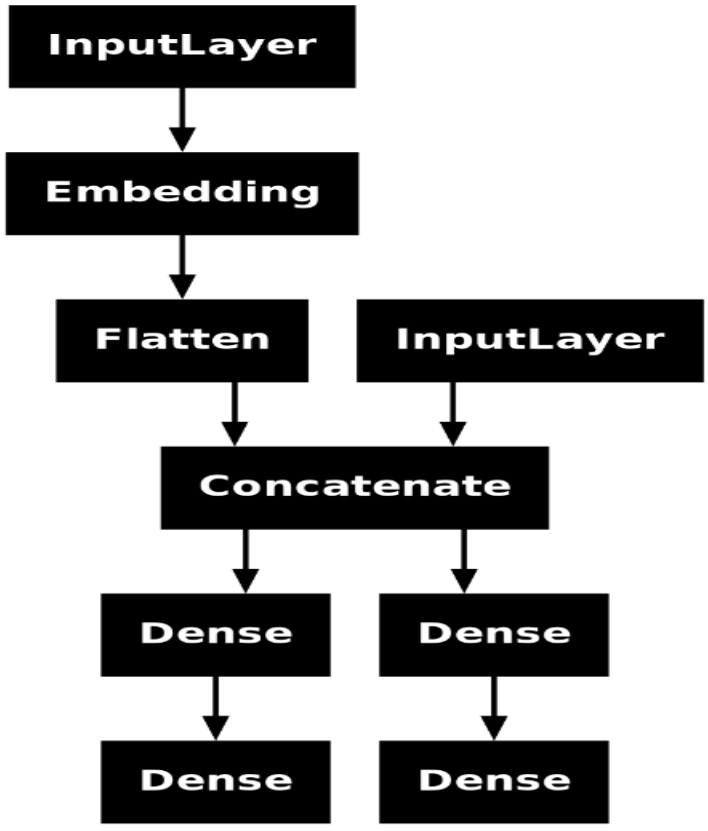
Example Prediction:

Input: director name: “James Cameron” and director’s last movie release date: 1993

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 23ms/step
Predicted Genres:
['Action', 'Sci-Fi', 'Adventure']

Predicted Year:
1966.4742336273193
```

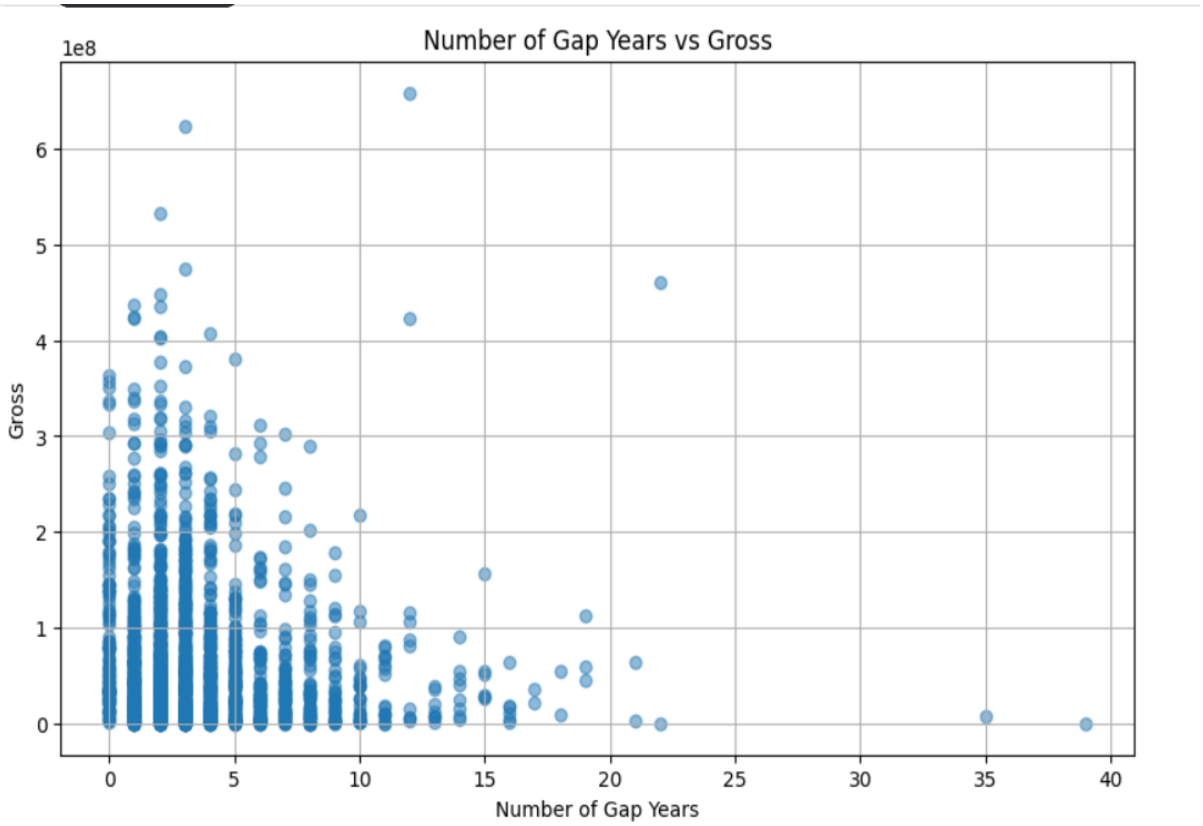
Model used:



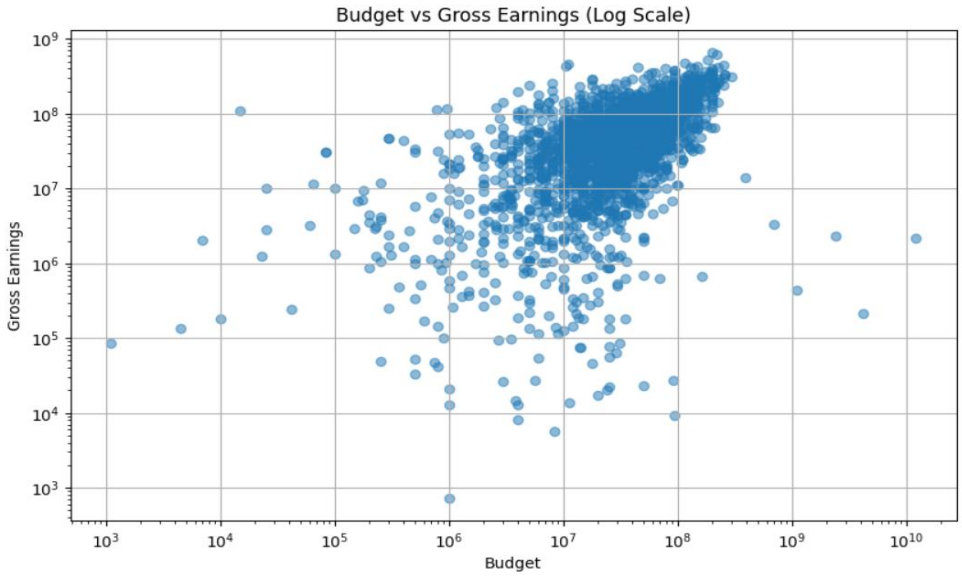
EXPLORATORY DATA ANALYSIS:

We created a column called num_gap_years which tells us number of years director took to make a film

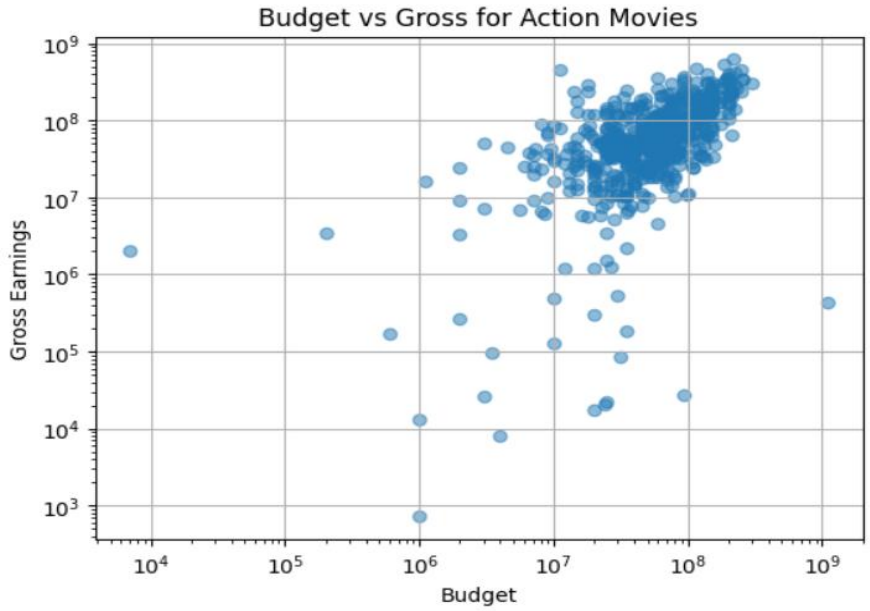
Insight: The Gross collections are higher when the Number of Gap Years is lower. Many of the exceptional high gross collections are happening when the director is taking 2-4 years to make the movie.



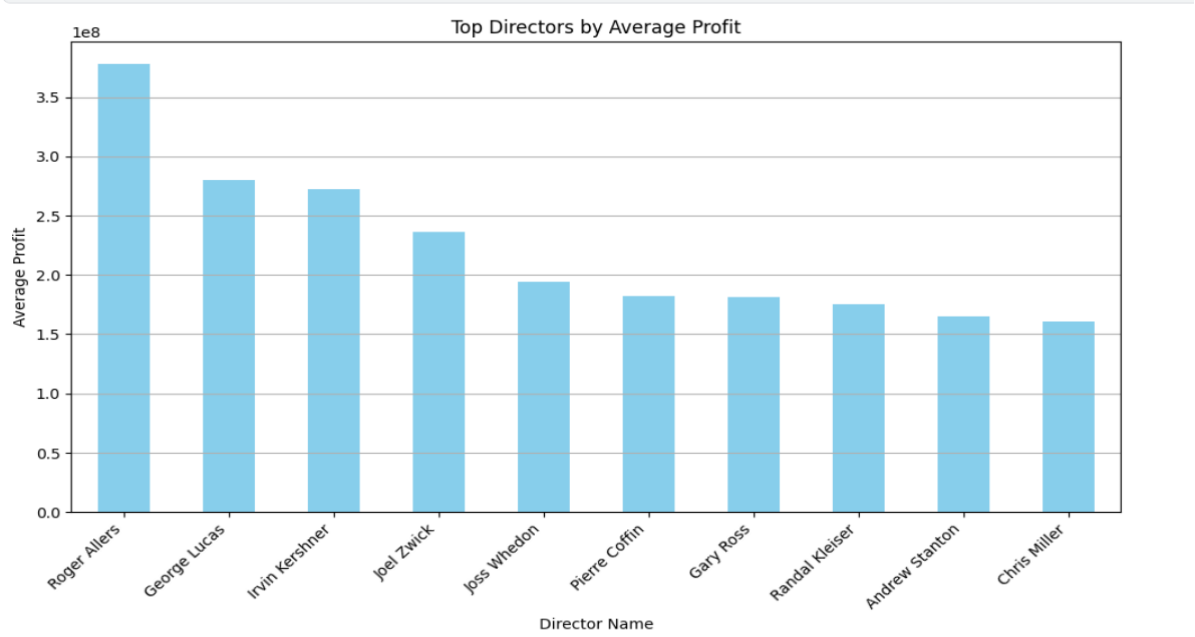
Insight: high budget movies are having high gross



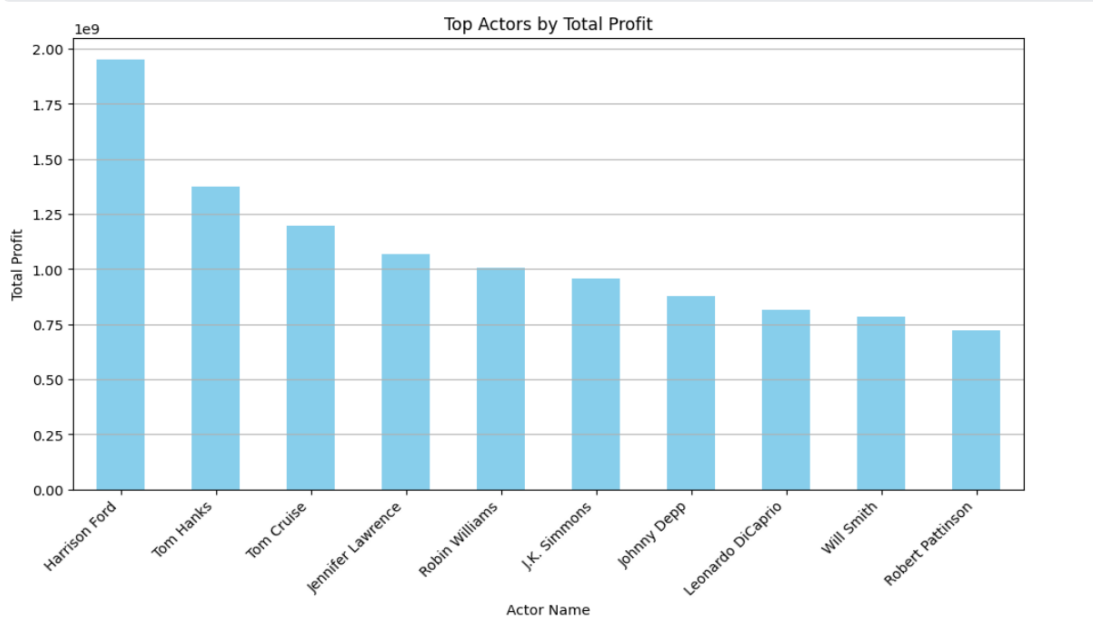
Insight: Low budget Action movies like budget less than 10^7 is not collecting a good gross amount. Only high budget action moves are reaching a good gross.



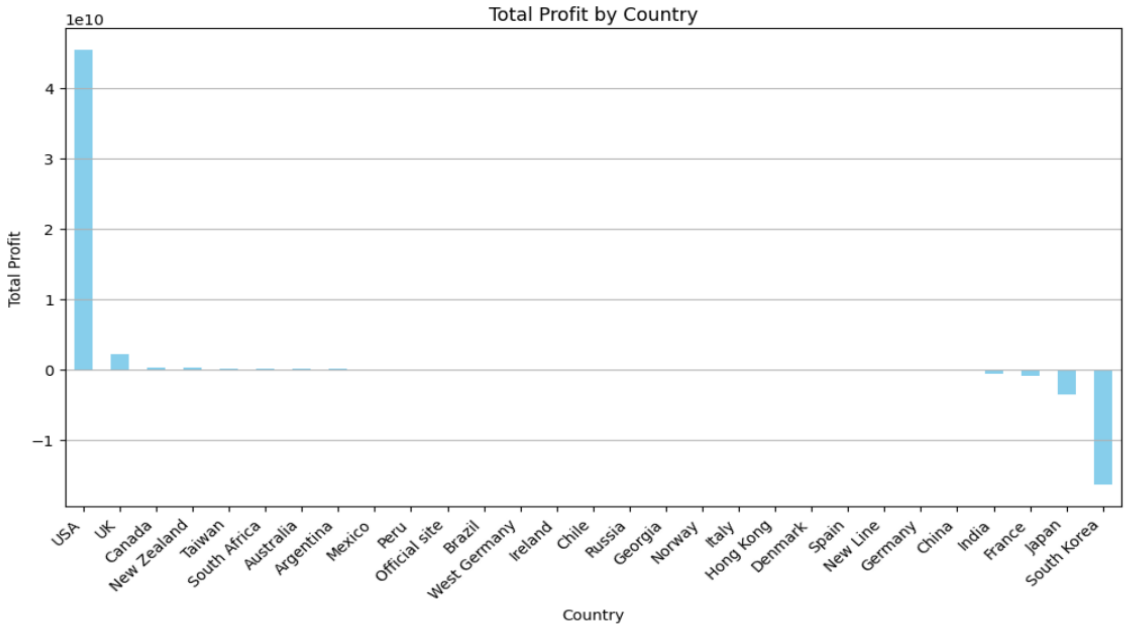
Insight: we created a new column called profit(gross-budget). From the below we can observe that Roger Allers is bringing more average profit



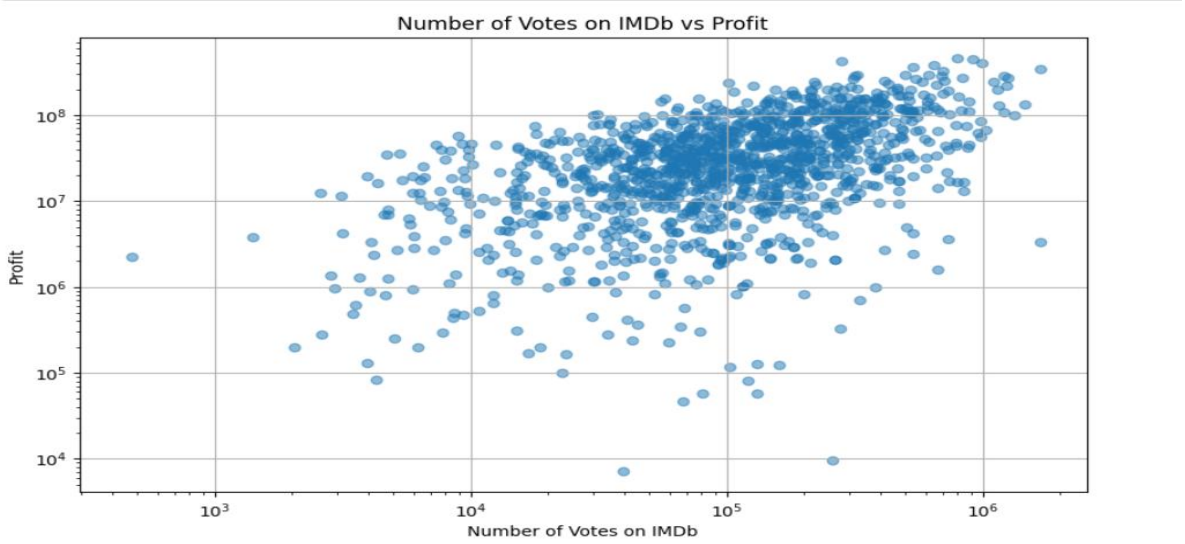
Insight: Top actors who brought most profits till now



Insight: Base country to earn more profits through movies is USA and most loss was incurred in South Korea



Insight: more IMDB rating more profits



Insight: movies with an “M” rating (presumably for mature audiences) tend to have significantly higher average profits compared to other ratings.

movies with an “R” rating (restricted) have the lowest average profits.