Movie analysis with ML

Today's Agenda

Questions to answer

1. What is the average Run Time for movies in each Genre?
2. Explore the descriptive statistics of Movie Ratings for different Certification categories
3. Which features are correlated?
4. Is there a correlation between the MetaScore and the Movie Rating?
5. What combination of factors leads to high Votes for a movie?

Bonus

6. Null hypothesis " Movies directed by different directors have the same average gross earnings"
7. Creating Movie Recomender.

Introduction

What is IMDb?

Internet movie data base

Source for movie, TV and celebrity content

Dataset

10 000 rows

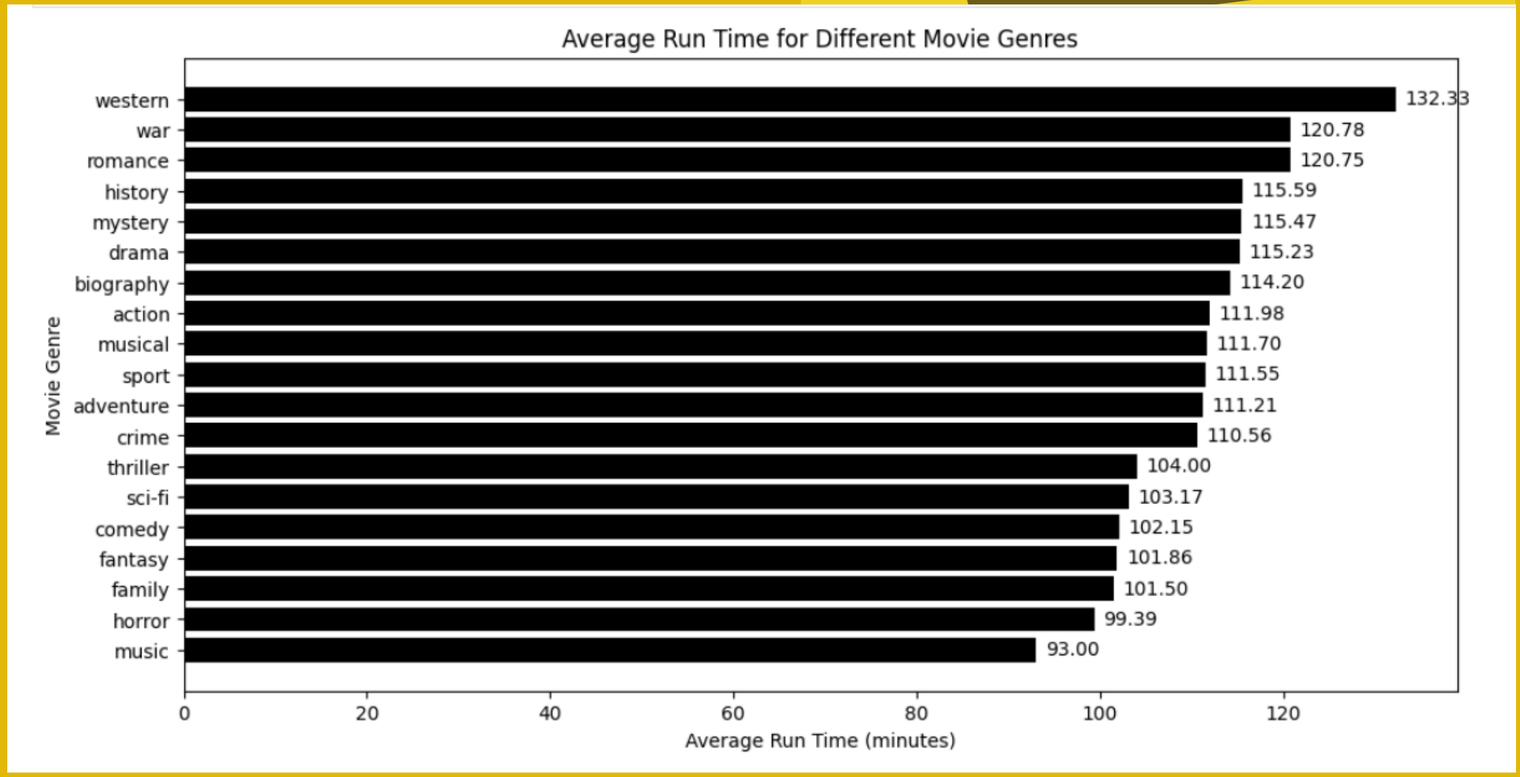
Missing values

Skewed distribution

	Movie Name	Year of Release	Run Time	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	Actors	Description	Certification Group	Rating Group
0	Life Is Beautiful	1997	116	8.6	723773	59.0	57600000.0	['Comedy', ' Drama', ' Romance']	PG-13	['Roberto Benigni']	['Roberto Benigni', 'Nicoletta Braschi', 'Gior...	['When', 'an', 'open-minded', 'Jewish', 'waite...	Teenagers and Above	Excellent
1	Psycho	1960	109	8.5	699966	97.0	32000000.0	['Horror', ' Mystery', ' Thriller']	R	['Alfred Hitchcock']	['Anthony Perkins', 'Janet Leigh', 'Vera Miles...	['A', 'Phoenix', 'secretary', 'embezzles', '$4...	Restricted	Excellent
2	Cinema Paradiso	1988	155	8.5	274382	80.0	11990000.0	['Drama', ' Romance']	R	['Giuseppe Tornatore']	['Philippe Noiret', 'Enzo Cannavale', 'Antonel...	['A', 'filmmaker', 'recalls', 'his', 'childhoo...	Restricted	Excellent
3	Once Upon a Time in the West	1968	165	8.5	342065	82.0	5320000.0	['Western']	PG-13	['Sergio Leone']	['Henry Fonda', 'Charles Bronson', 'Claudia Ca...	['A', 'mysterious', 'stranger', 'with', 'a', '...	Teenagers and Above	Excellent
4	Oldboy	2003	120	8.4	612299	78.0	710000.0	['Action', ' Drama', ' Mystery']	R	['Park Chan-wook']	['Choi Min-sik', 'Yoo Ji-tae', 'Kang Hye-jeong...	['After', 'being', 'kidnapped', 'and', 'impris...	Restricted	Very Good

What is the average Run Time for movies in each Genre?

GENRE	AVERAGE RUN TIME
Action	111.98
Adventure	111.21
Biography	114.2
Comedy	102.15
Crime	110.56
Drama	115.23
Family	101.5
Fantasy	101.86
History	115.59
Horror	99.39
Music	93
Musical	111.7
Mistery	115.47
Romance	120.75
Sci - Fi	103.17
Sport	115.55
Thriller	104
War	120.78
Western	132.33



Descriptive statistics of Movie Ratings for different Certification categories

Certification	count	mean	std	min	25%	50%	75%	max
13+	1.0	7.200000	NaN	7.2	7.200	7.20	7.200	7.2
16+	3.0	7.700000	0.556776	7.1	7.450	7.80	8.000	8.2
Approved	21.0	7.352381	0.634523	6.1	6.900	7.50	7.800	8.3
G	132.0	6.865909	0.701017	4.9	6.400	6.95	7.400	8.3
GP	6.0	7.550000	0.356371	7.0	7.450	7.60	7.600	8.1
M	2.0	7.500000	0.424264	7.2	7.350	7.50	7.650	7.8
M/PG	2.0	7.800000	0.565685	7.4	7.600	7.80	8.000	8.2
NC-17	18.0	6.694444	0.814914	5.0	6.150	6.95	7.400	7.7
Not Rated	268.0	7.073134	0.641413	4.9	6.700	7.20	7.500	8.4
PG	884.0	6.605656	0.779691	4.9	6.100	6.60	7.200	8.4
PG-13	1723.0	6.454382	0.735879	4.9	5.900	6.50	7.000	8.6
Passed	10.0	7.590000	0.479467	6.7	7.250	7.70	7.975	8.2
R	2949.0	6.653306	0.712169	4.9	6.200	6.70	7.200	8.5
TV-14	5.0	6.900000	0.951315	5.3	6.900	7.20	7.300	7.8
TV-MA	7.0	6.800000	1.141636	5.3	5.800	7.10	7.800	8.0
TV-PG	3.0	7.733333	0.208167	7.5	7.650	7.80	7.850	7.9
Unrated	26.0	7.026923	0.551041	5.6	6.825	7.05	7.300	7.9

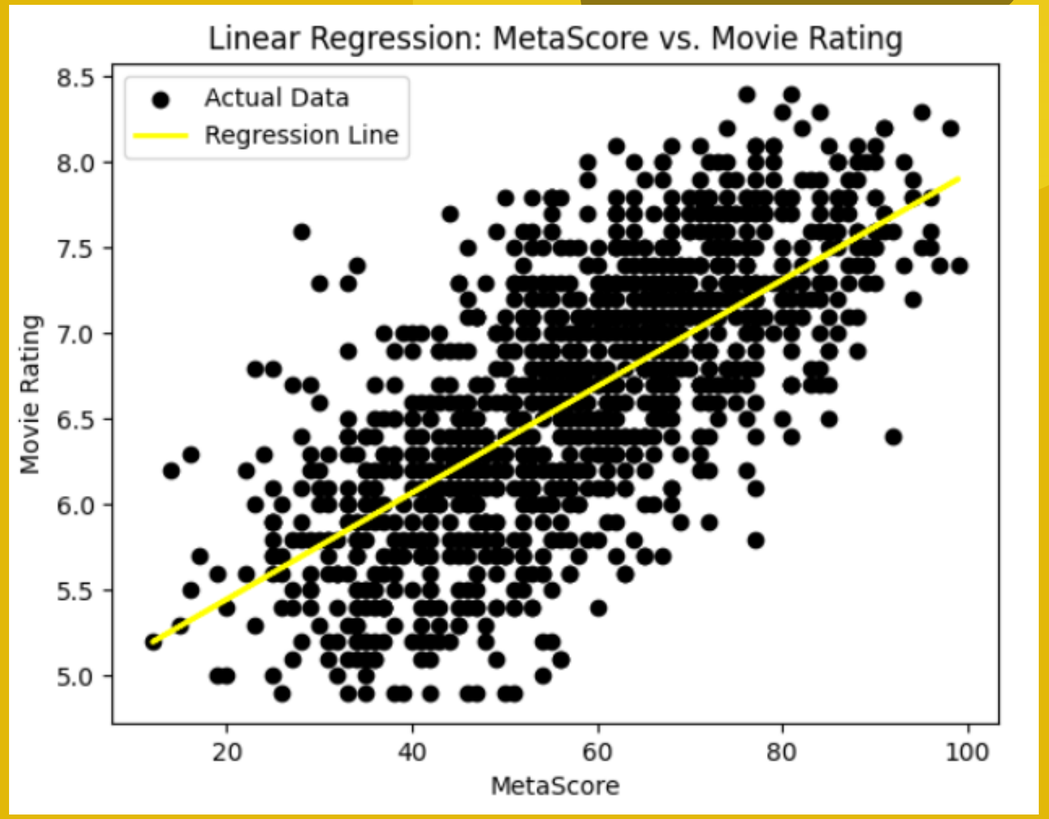
Certification Group	count	unique	top	freq
General Audiences	1016	5	Below Average	626
Not Rated or Rare	319	5	Average	111
Parental Guidance	30	4	Good	15
Restricted	2969	6	Below Average	1866
Teenagers and Above	1728	6	Below Average	1256

Which features are correlated?

	Year of Release	Run Time	Movie Rating	Votes	MetaScore	Gross
Year of Release	1.000000	-0.020851	-0.182270	0.124971	-0.123889	0.040097
Run Time	-0.020851	1.000000	0.337579	0.232321	0.208528	0.147151
Movie Rating	-0.182270	0.337579	1.000000	0.346424	0.706113	0.016331
Votes	0.124971	0.232321	0.346424	1.000000	0.180018	0.616684
MetaScore	-0.123889	0.208528	0.706113	0.180018	1.000000	-0.038226
Gross	0.040097	0.147151	0.016331	0.616684	-0.038226	1.000000

Is there a correlation between the MetaScore and the Movie Rating?

R2 Score	0.5
Mean Absolute Error	0.42
Mean Squared Error	0.28

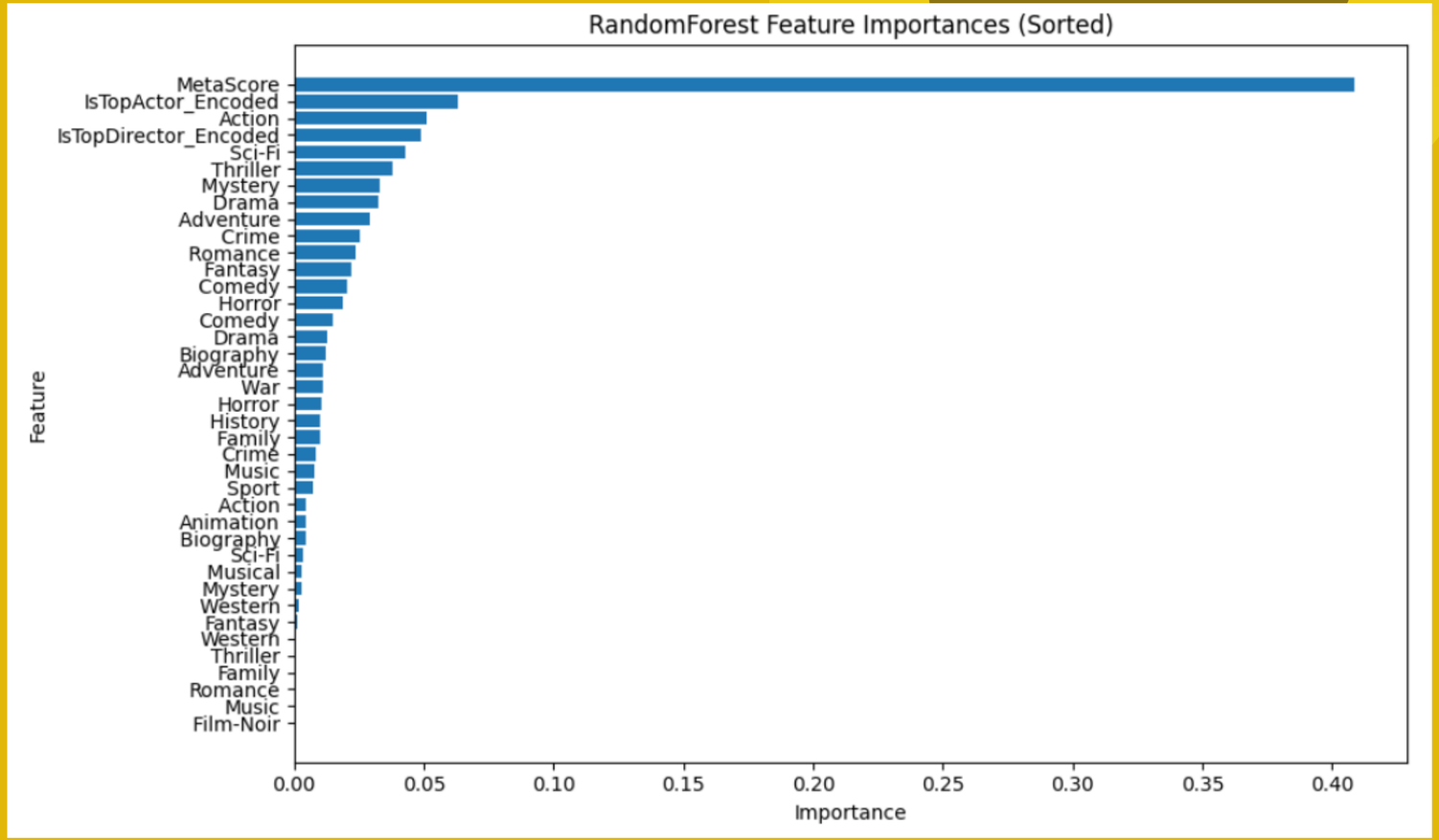


The R2 score of 0.499 indicates that the MetaScore explains about 49.9% of the variance in the Movie Rating. The Mean Absolute Error (MAE) of 0.423 suggests, on average, predictions are off by 0.423 units from the actual values. The Mean Squared Error (MSE) of 0.285 gives the average squared difference between predicted and actual values.

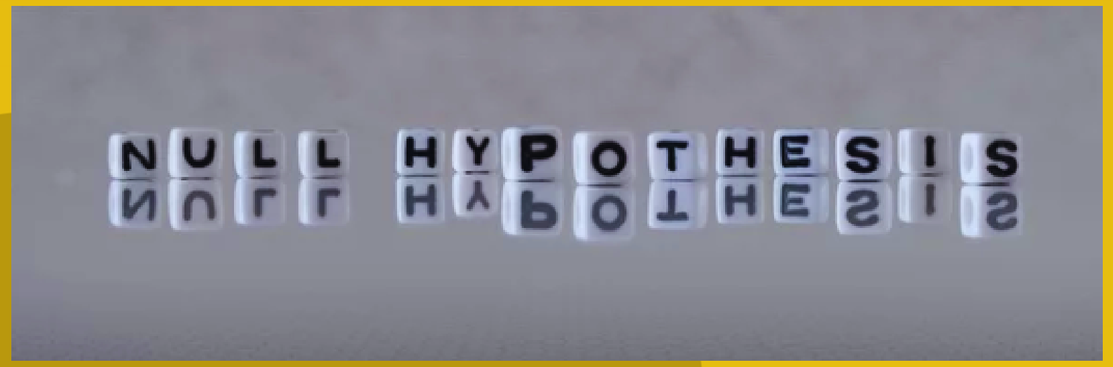
What combination of factors leads to high Votes for a movie?

Model	Result
GradientBoosting - Mean Squared Error	11534313360
GradientBoosting - R-squared	0.19
RandomForest - Mean Squared Error	13572472321
RandomForest - R-squared	0.05

Based on these results, the Gradient Boosting model is providing better predictions and capturing more variance in the target variable compared to the Random Forest model.



Movies directed by different directors have the same average gross earnings



ANOVA Result:	
F-statistic:	1.96
P-value	1.64E-75

Alternative Hypothesis: There are significant differences in the average gross earnings across different directors. The p-value (1.6363626636273887e-75) is extremely small, suggesting that the differences in average gross earnings are unlikely to be due to random chance alone. Therefore, there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. It implies that there are significant differences in the average gross earnings across different directors.

There is significant evidence to reject the null hypothesis.


THANK
YOU!