Cinematic Intelligence: Using Data Science to Predict Movie Success

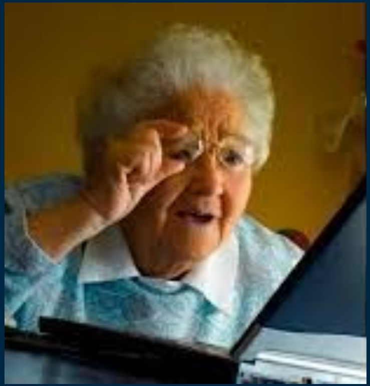
By: Jace Aung Kaung Kaung, Joel Lee Pak Xin,
Ye Wint Myint Myat

Lab group: A124_Team 2

TABLE OF CONTENTS

Problem Definition 01

Problem Definition



The Question:

How can we estimate a future movie's gross revenue, profit and budget needed for movie?

Data Set:

From Kaggle movie csv data (Over 5000 data)
Includes director, star, writer, budget, genre, IMDB score, runtime and votes.

Practical Motivation

For us:

- Movie fanatics
- Movie industry getting bigger and bigger
- Curious if bigger budget movies would do better

For real world applications:

- Challenging difficult for film companies to set aside how much budget needed to film
- Profit-seeking companies will want to know if a movie is profitable which actors to hire and the genre before filming.

Exploratory Data Analysis

02

Cleaning our Data

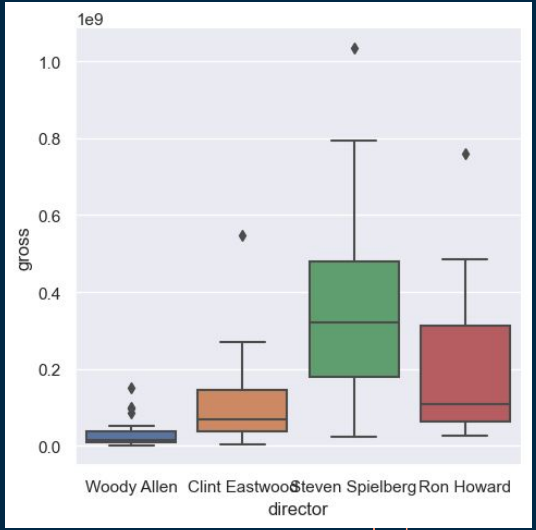
- Remove NaN values
- Add ‘profit’ column
- Convert Categorical values to Numerical values

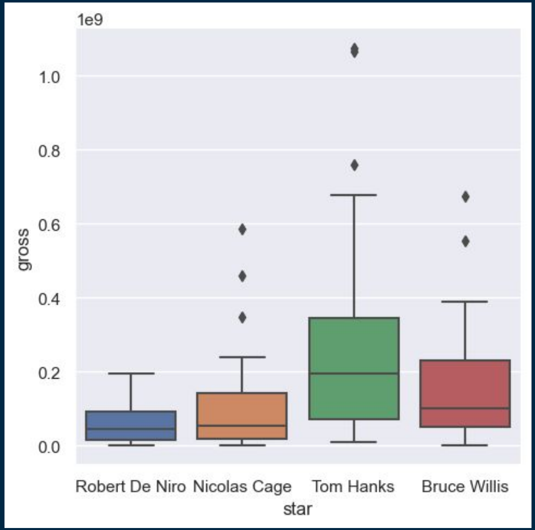
	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
16	Fame	R	Drama	1980	May 16, 1980 (United States)	6.60000000	21,000.00000000	Alan Parker	Christopher Gore	Eddie Barth	United States	NaN	21,202,829.00000000
19	Stir Crazy	R	Comedy	1980	December 12, 1980 (United States)	6.80000000	26,000.00000000	Sidney Poitier	Bruce Jay Friedman	Gene Wilder	United States	NaN	101,300,000.00000000
24	Urban Cowboy	PG	Drama	1980	June 6, 1980 (United States)	6.40000000	14,000.00000000	James Bridges	Aaron Latham	John Travolta	United States	NaN	46,918,287.00000000
25	Altered States	R	Horror	1980	December 25, 1980 (United States)	6.90000000	33,000.00000000	Ken Russell	Paddy Chayefsky	William Hurt	United States	NaN	19,853,892.00000000
26	Little Darlings	R	Comedy	1980	March 21, 1980 (United States)	6.50000000	5,100.00000000	Ron Maxwell	Kimi Peck	Tatum O'Neal	United States	NaN	34,326,249.00000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.10000000	18.00000000	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7,000.00000000	NaN
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.70000000	36.00000000	Dusty Dukatz	Lisa Huston	Michael Saquella	United States	NaN	NaN
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.70000000	29.00000000	Nkanya Nkwai	Lynno Lovert	Onyama Laura	United States	58,750.00000000	NaN
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall	Christina Roz	United States	15,000.00000000	NaN
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.70000000	7.00000000	Pereko Mosia	Pereko Mosia	Siyabonga Mabaso	South Africa	NaN	NaN

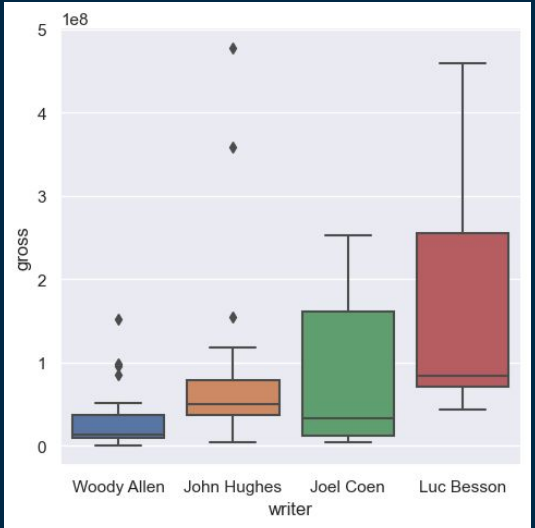
2247 rows × 17 columns

Exploring the Data

Categorical Variable against Gross Revenue

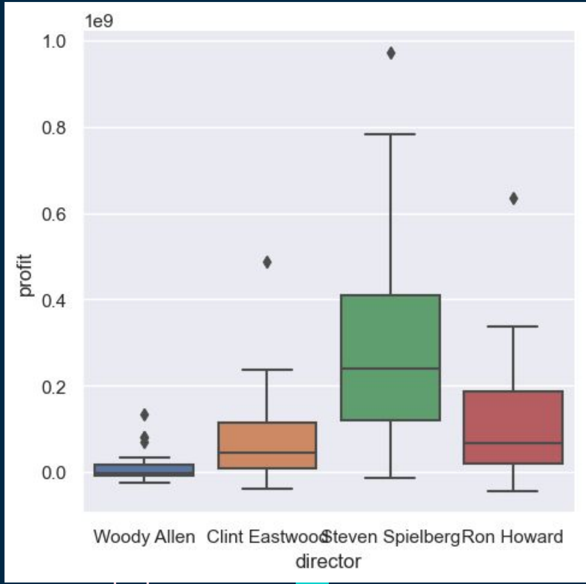


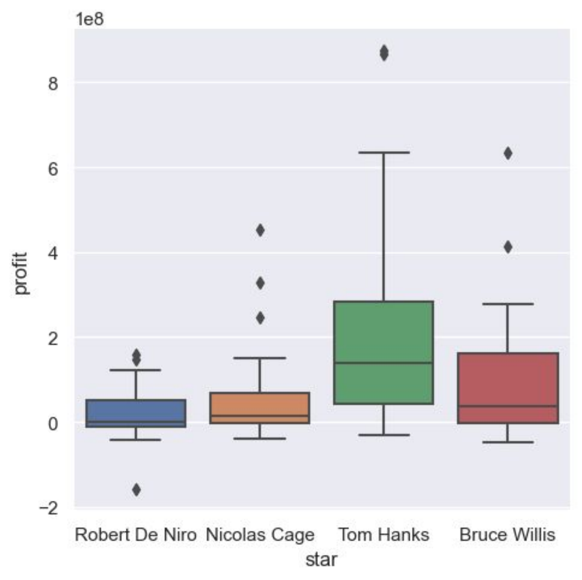


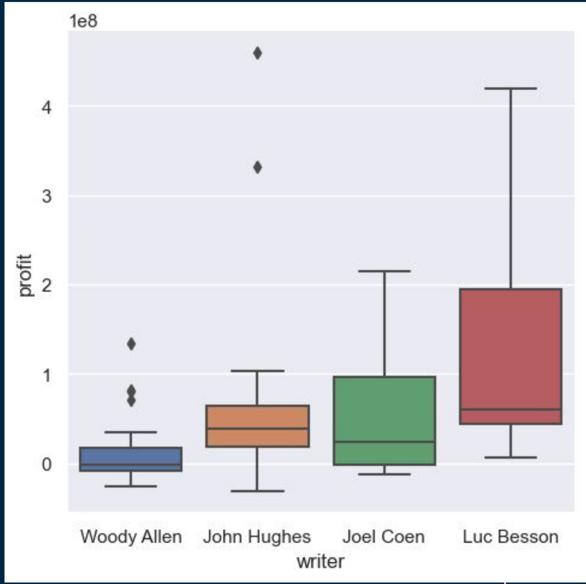


Exploring the Data

Categorical Variable against Profit







Cleaning Our Data

Group Gross Revenue

According to each level of the categorical variable

Calculate Mean Gross of Each Group

Assign this to be numerical value of the level

Higher Mean Suggests Higher Performance

The person/genre is able to bring about higher gross revenue on average

Cleaning our Data

Before

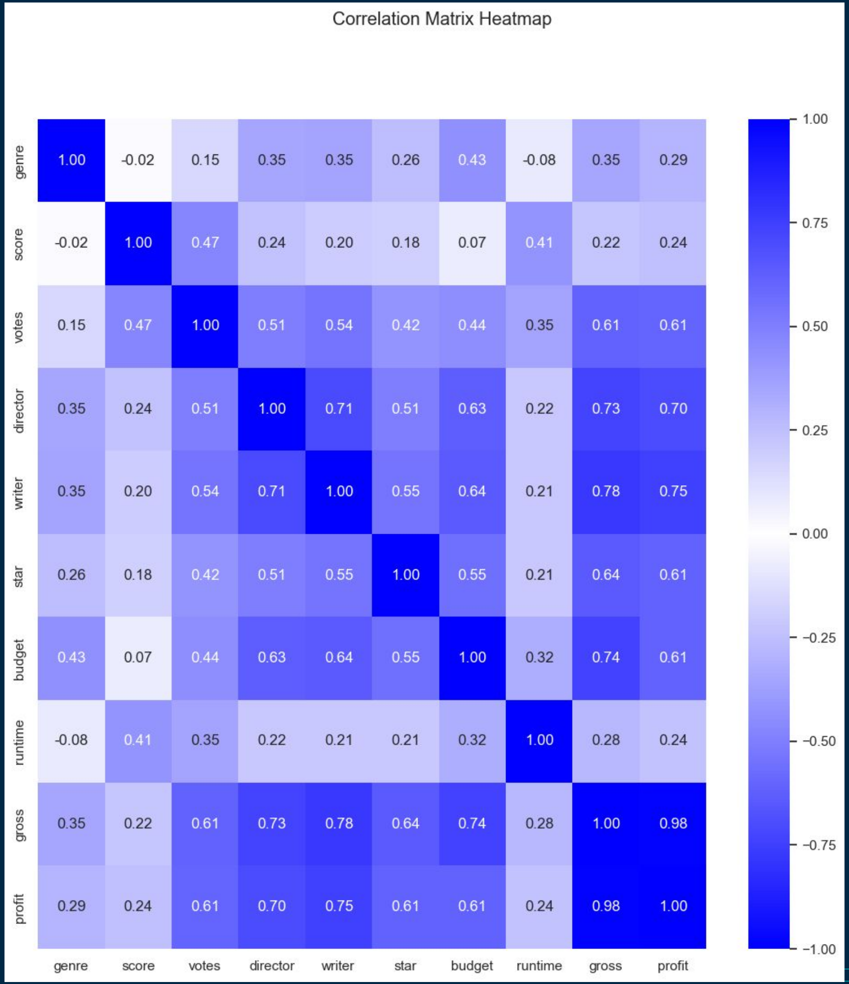
	genre	score	votes	director	writer	star	budget	runtime	gross	profit
0	Drama	8.40000000	927,000.00000000	Stanley Kubrick	Stephen King	Jack Nicholson	19,000,000.00000000	146.00000000	46,998,772.00000000	27,998,772.00000000
1	Adventure	5.80000000	65,000.00000000	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	4,500,000.00000000	104.00000000	58,853,106.00000000	54,353,106.00000000
2	Action	8.70000000	1,200,000.00000000	Irvin Kershner	Leigh Brackett	Mark Hamill	18,000,000.00000000	124.00000000	538,375,067.00000000	520,375,067.00000000
3	Comedy	7.70000000	221,000.00000000	Jim Abrahams	Jim Abrahams	Robert Hays	3,500,000.00000000	88.00000000	83,453,539.00000000	79,953,539.00000000
4	Comedy	7.30000000	108,000.00000000	Harold Ramis	Brian Doyle-Murray	Chevy Chase	6,000,000.00000000	98.00000000	39,846,344.00000000	33,846,344.00000000

After

	genre	score	votes	director	writer	star
0	60,369,136.46465817	8.40000000	927,000.00000000	46,678,224.00000000	56,264,777.93103448	83,348,568.77777778
1	133,268,232.13455658	5.80000000	65,000.00000000	42,718,332.75000000	30,830,480.00000000	15,088,310.50000000
2	168,023,228.81060070	8.70000000	1,200,000.00000000	213,163,027.00000000	538,375,067.00000000	506,740,622.00000000
3	59,167,658.83689839	7.70000000	221,000.00000000	85,045,841.00000000	87,596,763.00000000	42,724,366.66666666
4	59,167,658.83689839	7.30000000	108,000.00000000	65,592,597.88888889	25,822,323.00000000	37,010,279.87500000

Exploring the Dara

Correlation Matrix Heatmap



Machine Learning Technique

03

Using Linear Regression to Solve Our Problem

Serves as a Efficient Prediction Model

Widely applicable to most films in general

Ensure Reliability of Prediction Model

Goodness of fit can be compared quantifiably

Identification of Significant Factors

Smaller p-value implies that a factor is more significant

Determine Underlying Trends

Quantitative analysis of Relationships

Our Linear Regression Process to Solve Our Problem

Dropna, numericalise categorical variables

Ensure data is ready for use

1. Data Cleaning

2. Fitting the Model

Supervised Machine Learning

Compute the coefficients of the models using training sets

Recognising significant factors

Interpretation of Results

3. Analysis

4. Evaluation

Validity and Usefulness

Determine the reliability and accuracy of our models in solving our problem

Our Linear Models and Their Intended Purpose

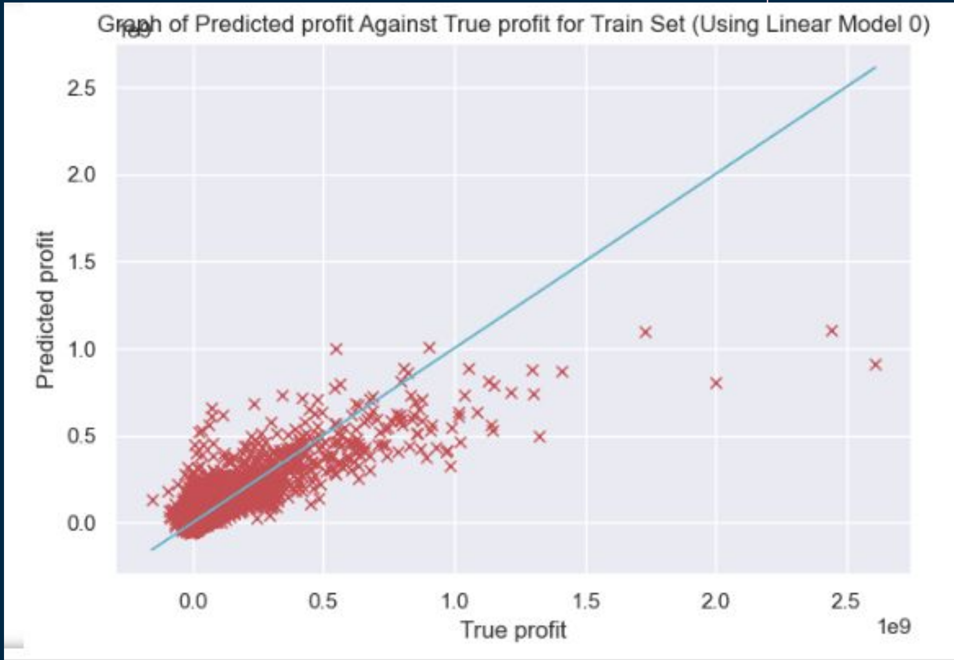
Linear Model	Dependent Variable			Independent Variable								
	profit	gross	budget	genre	score	votes	director	writer	star	budget	runtime	profit
0	✓			✓	✓	✓	✓	✓	✓	✓	✓	
1	✓				✓	✓				✓	✓	
2	✓						✓	✓	✓			
3		✓		✓	✓	✓	✓	✓	✓	✓	✓	
4		✓			✓	✓				✓	✓	
5		✓					✓	✓	✓			
6			✓	✓	✓	✓	✓	✓	✓		✓	✓
7			✓	✓	✓	✓					✓	✓
8			✓				✓	✓	✓			✓

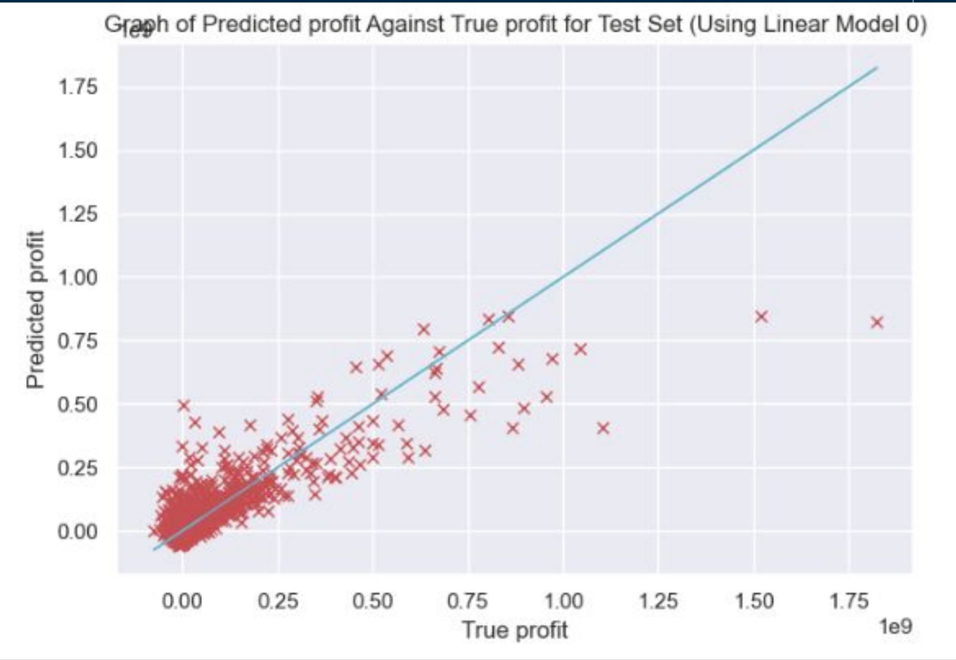
Our Linear Models and Their Intended Purpose

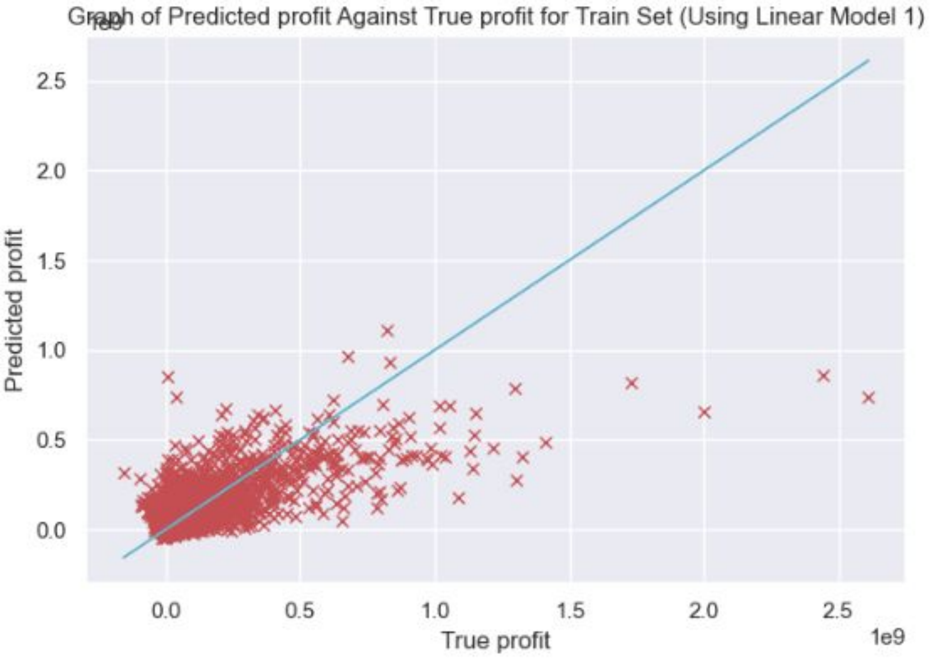
Linear Model	Purpose
0	to predict profit using all independent variables
1	same as Linear Model 0 but excluding categorical variables
2	same as Linear Model 0 but excluding numerical variables
3	same as Linear Model 0 but predicting gross revenue instead of budget
4	same as Linear Model 1 but predicting gross revenue instead of budget
5	same as Linear Model 2 but predicting gross revenue instead of budget
6	to estimate budget needed to attain a desired profit
7	same as Linear Model 6 but excluding categorical variables
8	same as Linear Model 6 but excluding numerical variables except profit

Result Analysis

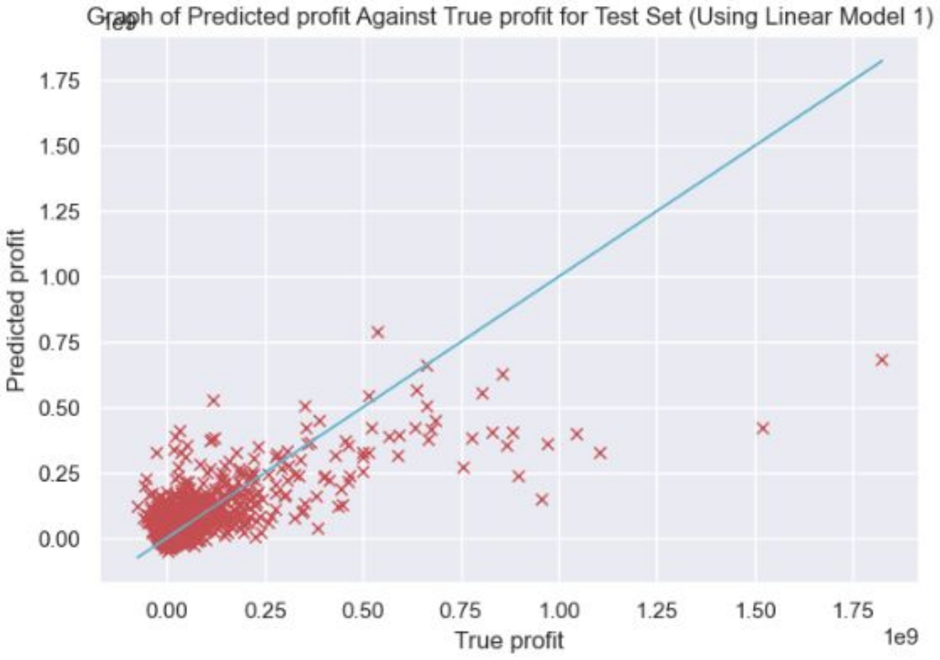
04


Graph of Predicted profit Against True profit for Train Set (Using Linear Model 0)
1e9
Predicted profit
0.0
0.5
1.0
1.5
2.0
2.5
True profit
1e9

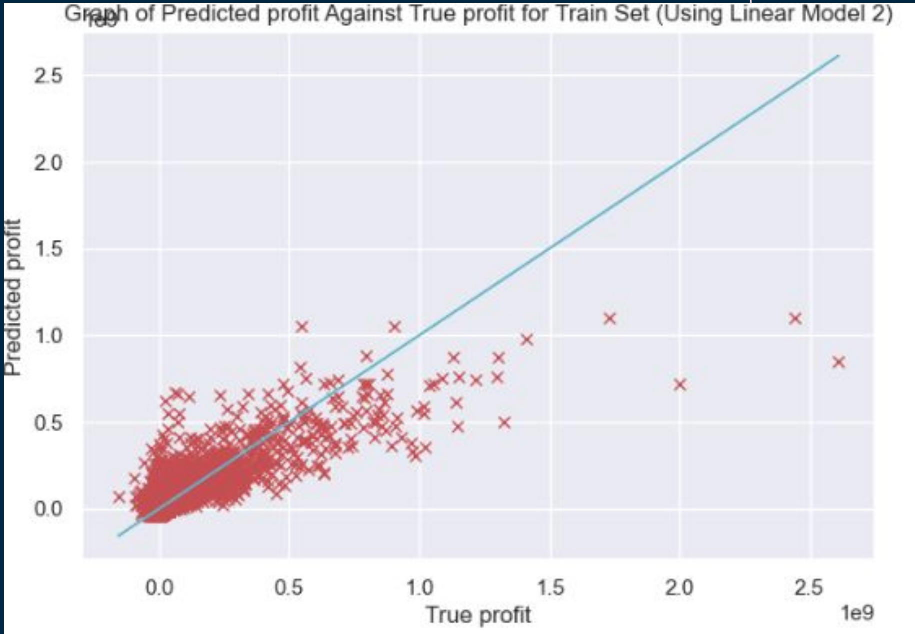

Graph of Predicted profit Against True profit for Test Set (Using Linear Model 0)
1e9
Predicted profit
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
True profit
1e9



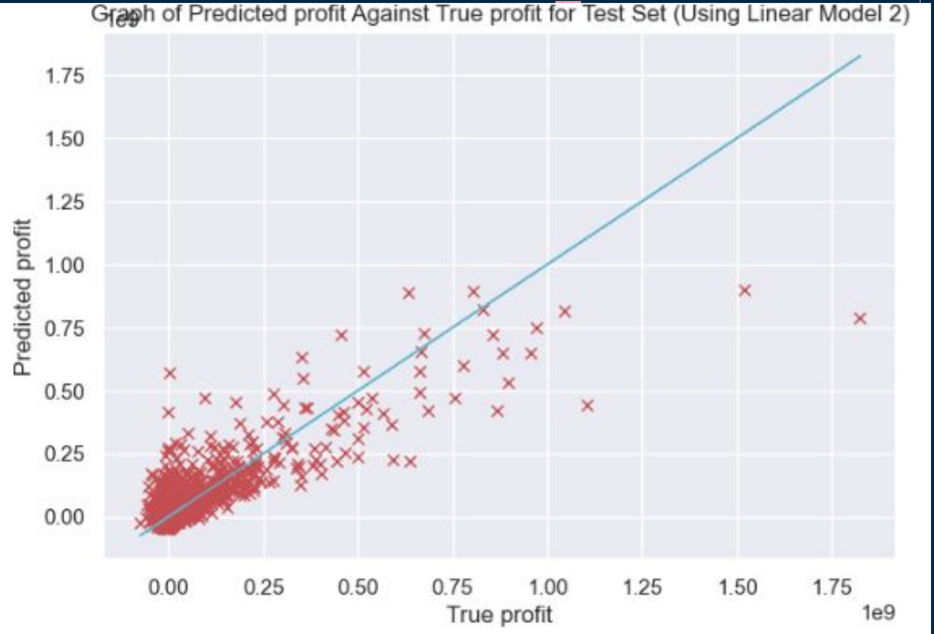
Graph of Predicted profit Against True profit for Train Set (Using Linear Model 1)
1e9
2.5
2.0
1.5
1.0
0.5
0.0
Predicted profit
0.0
0.5
1.0
1.5
2.0
2.5
True profit
1e9



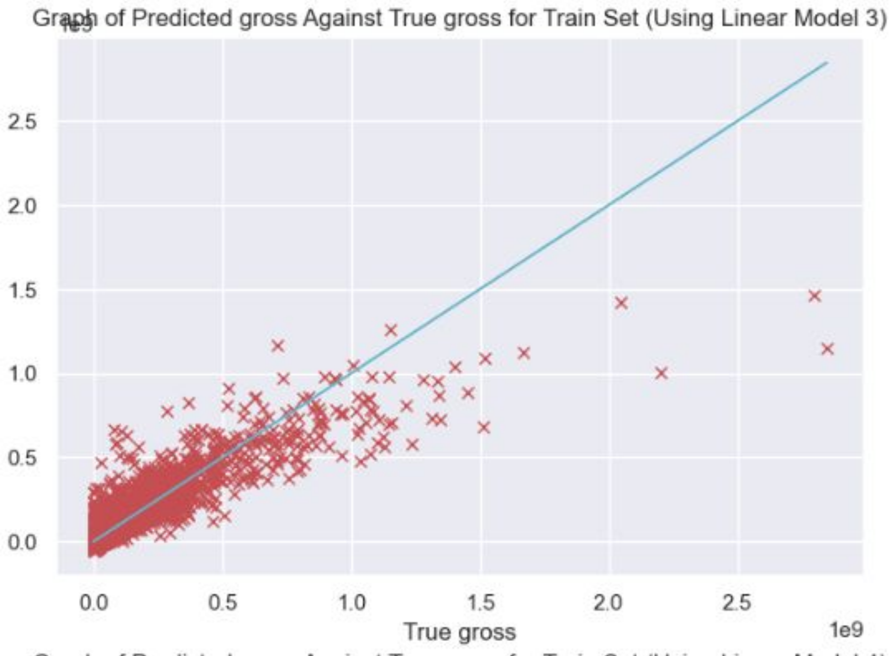
Graph of Predicted profit Against True profit for Test Set (Using Linear Model 1)
1e9
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Predicted profit
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
True profit
1e9



Graph of Predicted profit Against True profit for Train Set (Using Linear Model 2)
1e9
Predicted profit
True profit
1e9
0.0
0.5
1.0
1.5
2.0
2.5

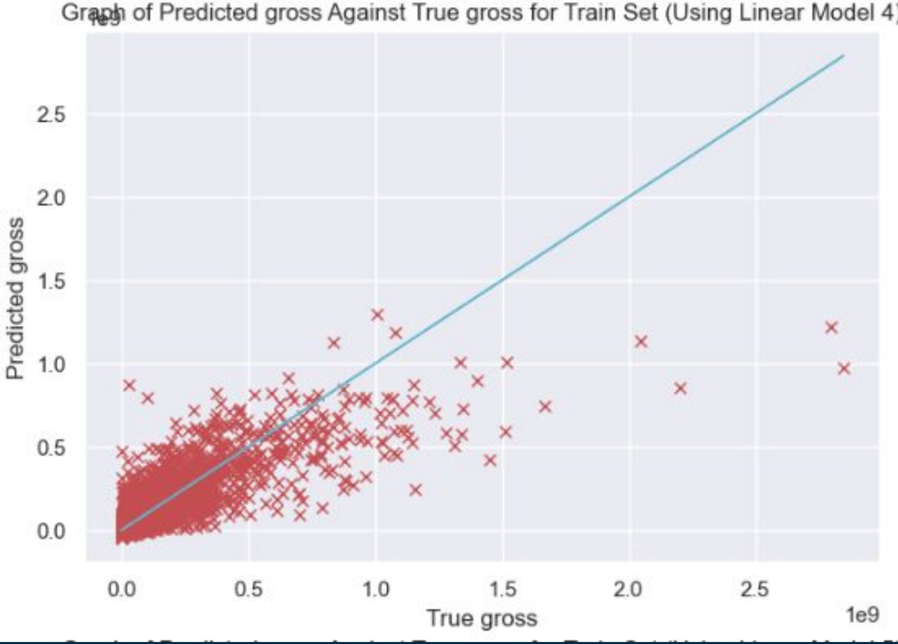


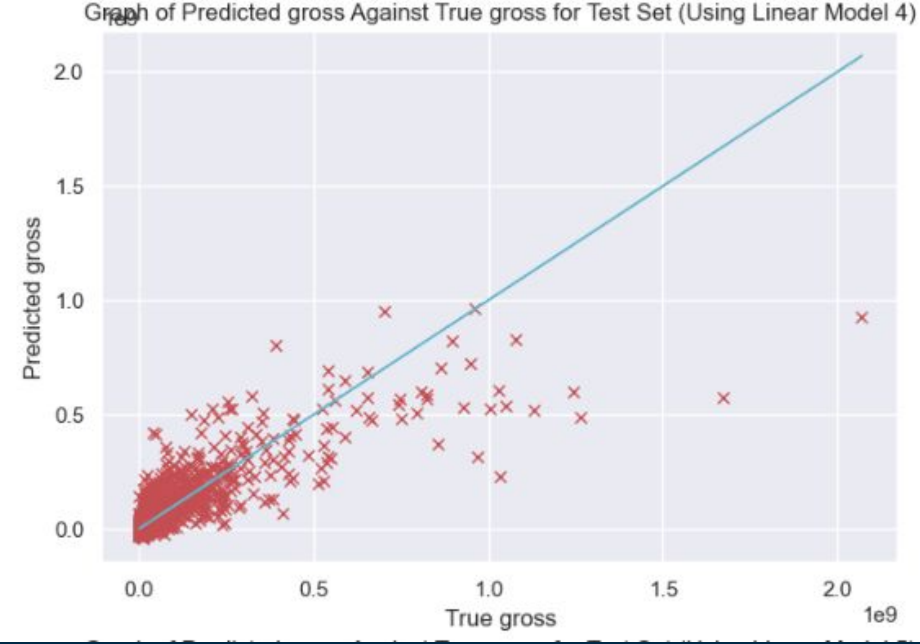
Graph of Predicted profit Against True profit for Test Set (Using Linear Model 2)
1e9
Predicted profit
True profit
1e9
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

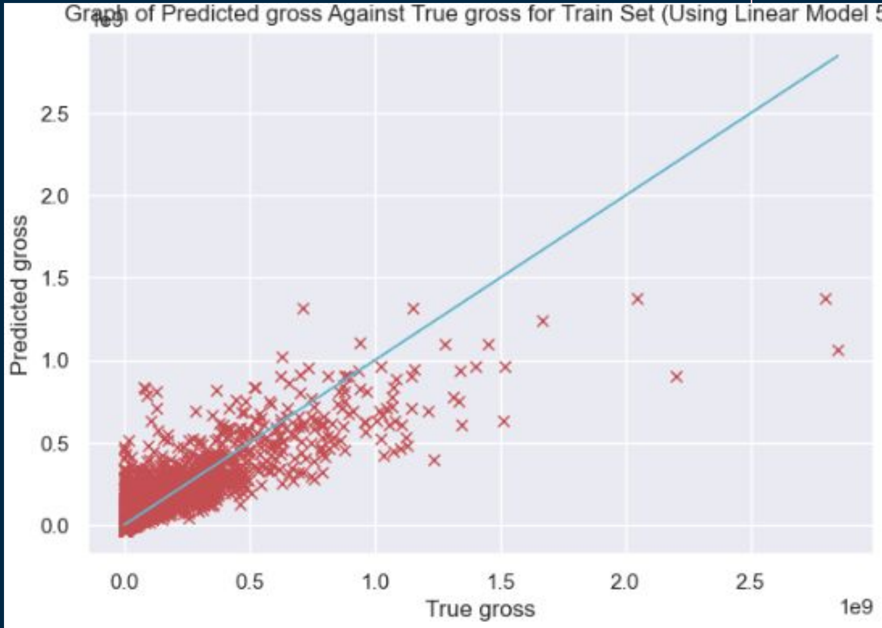


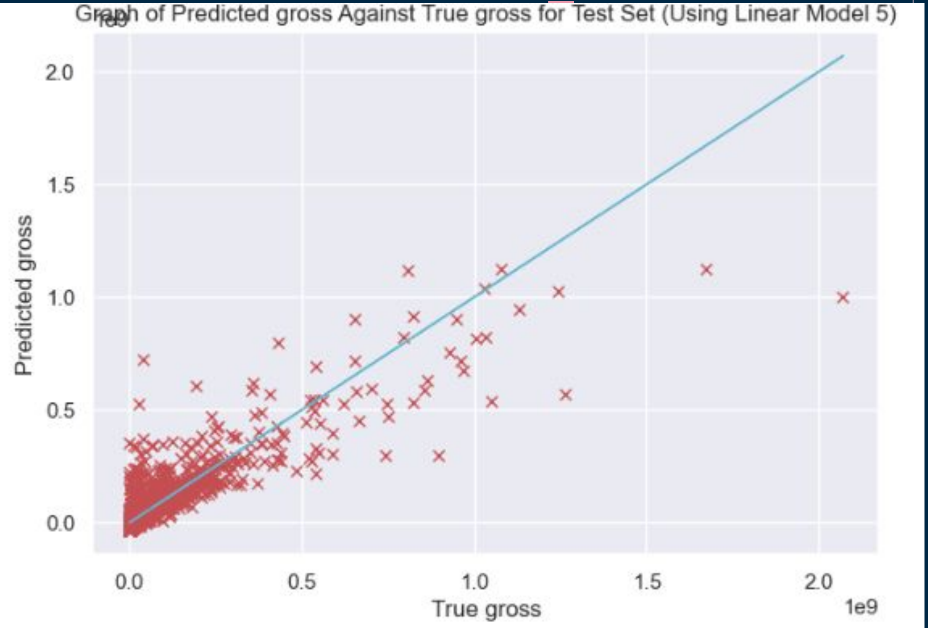
Graph of Predicted gross Against True gross for Train Set (Using Linear Model 3)
1e9
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
True gross
1e9

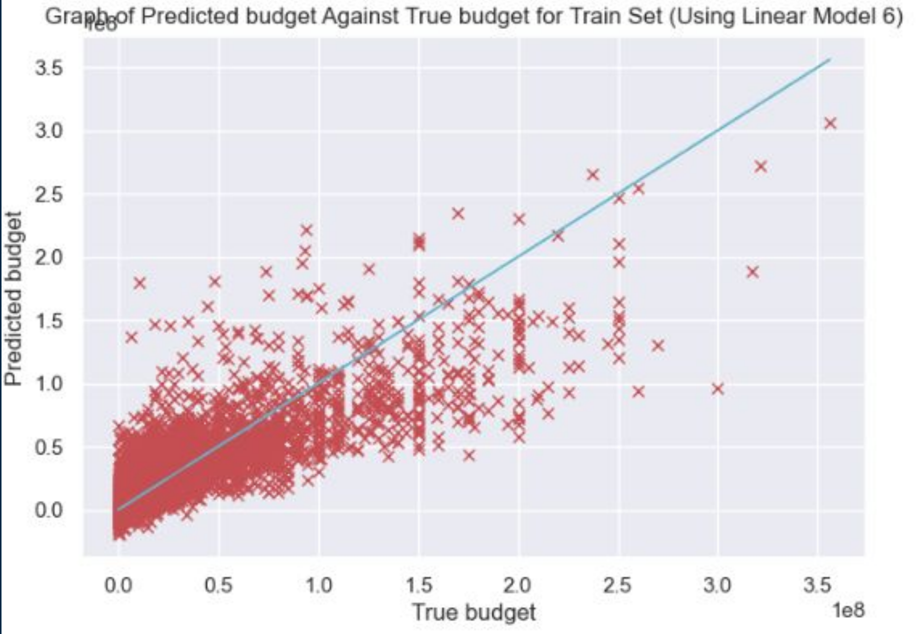
Graph of Predicted gross Against True gross for Test Set (Using Linear Model 3)
1e9
Predicted gross
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
True gross
1e9

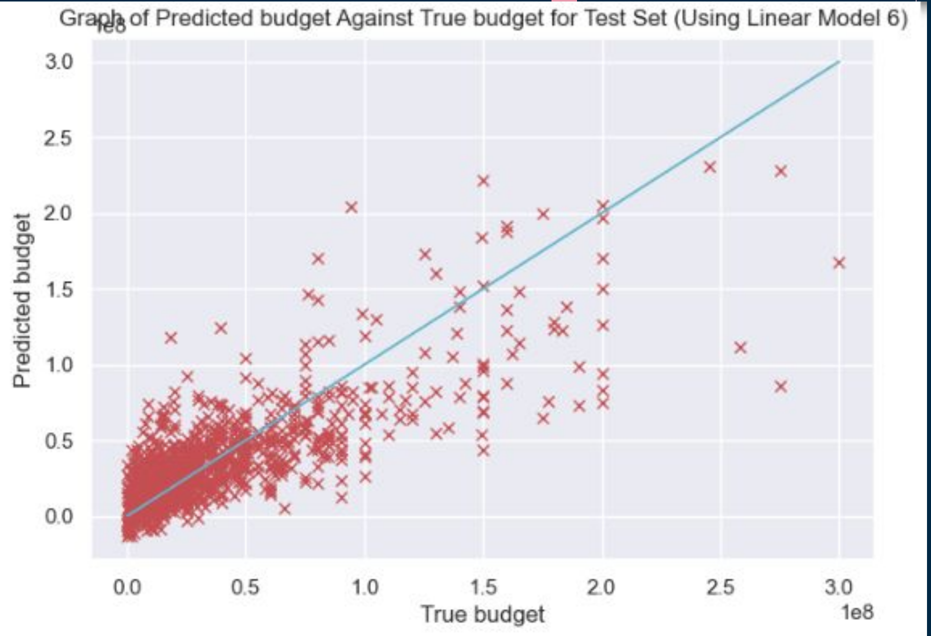

Graph of Predicted gross Against True gross for Train Set (Using Linear Model 4)
1e9
Predicted gross
True gross
1e9
0.0
0.5
1.0
1.5
2.0
2.5

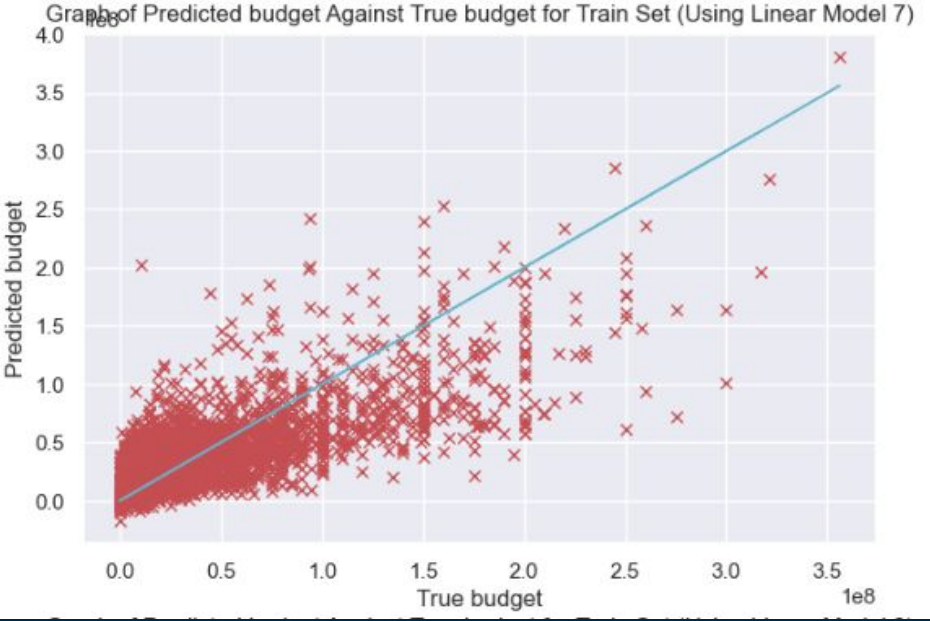

Graph of Predicted gross Against True gross for Test Set (Using Linear Model 4)
1e9
Predicted gross
True gross
1e9
0.0
0.5
1.0
1.5
2.0

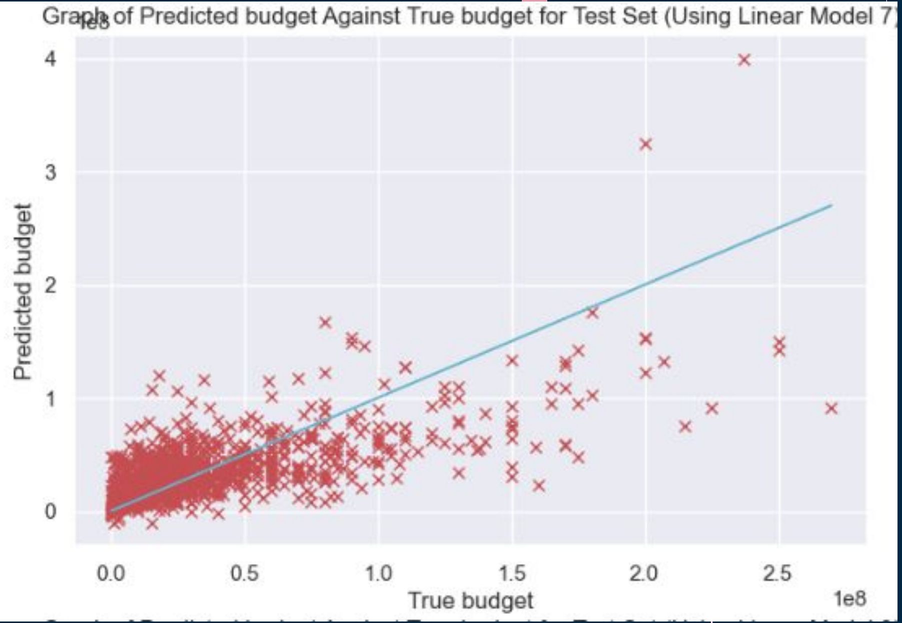

Graph of Predicted gross Against True gross for Train Set (Using Linear Model 5)
1e9
Predicted gross
0.0
0.5
1.0
1.5
2.0
2.5
True gross
0.0
0.5
1.0
1.5
2.0
2.5
1e9

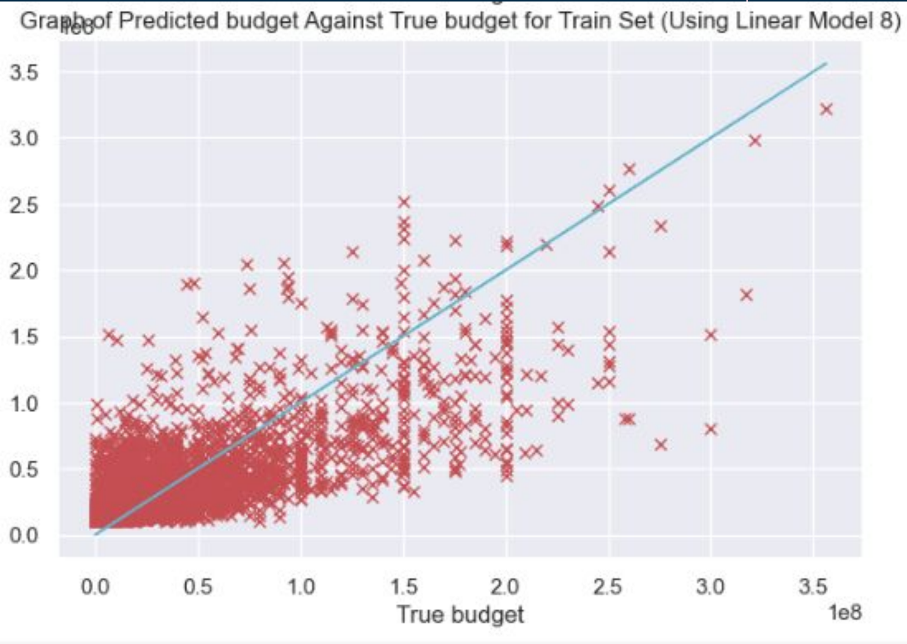

Graph of Predicted gross Against True gross for Test Set (Using Linear Model 5)
1e9
Predicted gross
0.0
0.5
1.0
1.5
2.0
True gross
0.0
0.5
1.0
1.5
2.0
1e9

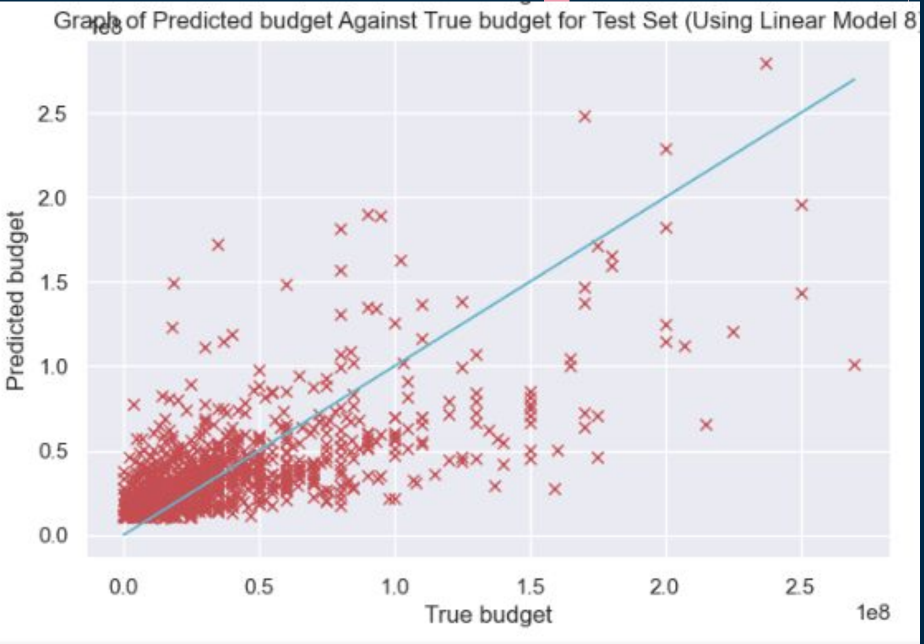

Graph of Predicted budget Against True budget for Train Set (Using Linear Model 6)
1e8
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Predicted budget
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
True budget
1e8


Graph of Predicted budget Against True budget for Test Set (Using Linear Model 6)
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Predicted budget
0.0
0.5
1.0
1.5
2.0
2.5
3.0
True budget
1e8


Graph of Predicted budget Against True budget for Train Set (Using Linear Model 7)
1e8
Predicted budget
True budget
1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Graph of Predicted budget Against True budget for Test Set (Using Linear Model 7)
1e8
Predicted budget
True budget
1e8
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
2.5


Graph of Predicted budget Against True budget for Train Set (Using Linear Model 8)
1e8
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
True budget
1e8


Graph of Predicted budget Against True budget for Test Set (Using Linear Model 8)
1e8
Predicted budget
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
True budget
1e8

Goodness of Fit and Prediction Accuracy

Linear Model	R-squared value (Goodness of fit)	Mean Squared Error (Prediction Accuracy)
0	0.696	7.75×10^{15}
1	0.528	1.12×10^{16}
2	0.664	8.27×10^{15}
3	0.780	7.75×10^{15}
4	0.659	1.12×10^{16}
5	0.724	9.77×10^{15}
6	0.610	6.75×10^{14}
7	0.529	8.49×10^{14}
8	0.526	8.79×10^{14}

Data-Driven Insights to Our Problem

Highest Goodness of Fit

R-squared = 0.78

High R-squared value suggests that our model 3 is well-fitting and thus reliable

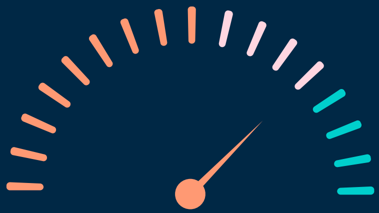
Lowest MSE

MSE = 7.75×10^{15}

Low MSE suggests that our model 3 has high prediction accuracy and thus useful

Reliability

Analysis of variables



As our model 3 is reliable and accurate, we can analyse and infer insights from its results

Data-Driven Insights to Our Problem

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.618e+07	1.21e+07	-2.167	0.030	-4.99e+07	-2.49e+06
genre	-0.0229	0.024	-0.948	0.343	-0.070	0.024
score	-1.891e+06	1.74e+06	-1.088	0.276	-5.3e+06	1.51e+06
votes	178.5961	9.676	18.459	0.000	159.627	197.565
director	0.2444	0.015	16.254	0.000	0.215	0.274
writer	0.3679	0.015	24.980	0.000	0.339	0.397
star	0.3079	0.014	21.786	0.000	0.280	0.336
budget	1.1877	0.051	23.418	0.000	1.088	1.287
runtime	-1.399e+05	9.05e+04	-1.546	0.122	-3.17e+05	3.75e+04

High p-Value

At 5% significance level

Genre, score, runtime do not significantly affect gross revenue

Low p-Value

Close to zero

Votes, director, writer, star, budget are significant factors for gross revenue

Positive coefficients with High Correlation

Corr > 0.7, Coef > 0

writer, budget, director are the factors that are the most positively correlated with gross revenue

Data-Driven Recommendations for Our Problem

Hire Better Writers and Directors

Our Linear Model 3 suggests that quality of writers and directors are the most important factors in determining the film's gross revenue

Allocate Higher Investment Budget

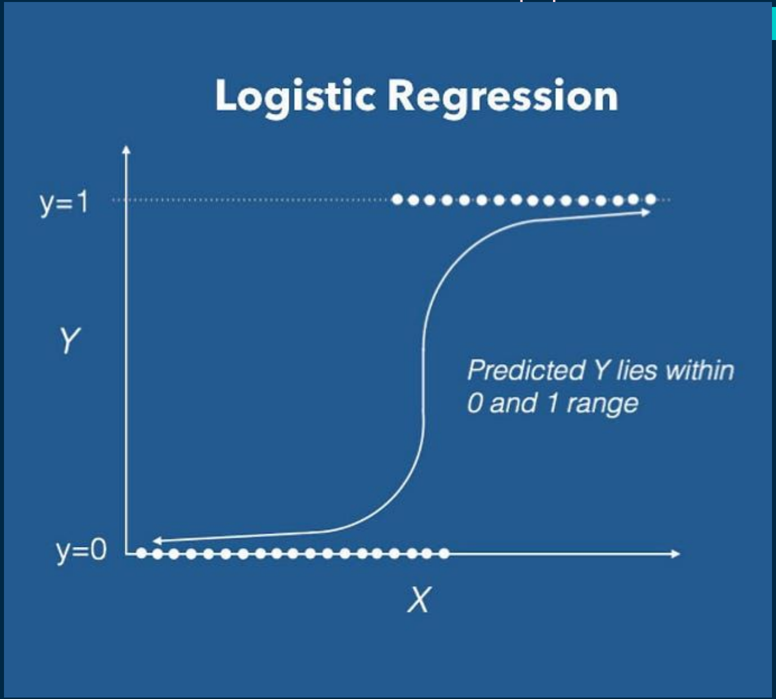
It also suggests that films with higher budget invested in them will return higher gross revenue

Extra

05

Logistic Regression

- Statistical model for binary classification
- Calculates the probability of a particular outcome
- Provides simple output



Logistic Regression Process

1. Defining the ‘success’ variable
2. Create Boolean Variable
3. Split Train/Test
4. Fit Logistic Regression

0.3811

Average Probability of Success of a Movie

Logistic Regression on Genre

Classification accuracy = 0.6934

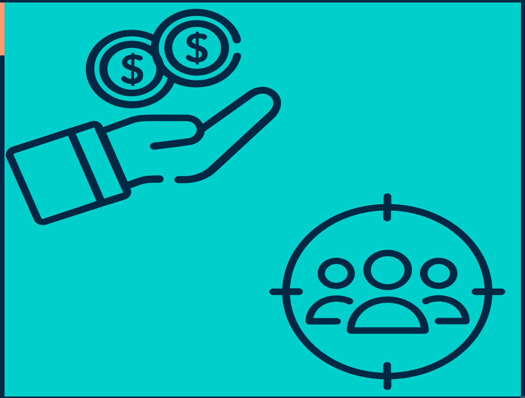
Genre with highest probability: Thriller

```
The overall classification accuracy            :  0.6934441366574331
Probability of profit_3x_budget for genre_Action: 0.3865667383867745
Probability of profit_3x_budget for genre_Adventure: 0.42341974198396365
Probability of profit_3x_budget for genre_Animation: 0.356262444351532
Probability of profit_3x_budget for genre_Biography: 0.449341601188103
Probability of profit_3x_budget for genre_Comedy: 0.45168025465781714
Probability of profit_3x_budget for genre_Crime: 0.45582021780310783
Probability of profit_3x_budget for genre_Drama: 0.45658275927205
Probability of profit_3x_budget for genre_Fantasy: 0.47245975798031176
Probability of profit_3x_budget for genre_Horror: 0.47243549752199093
Probability of profit_3x_budget for genre_Mystery: 0.41939540644517104
Probability of profit_3x_budget for genre_Romance: 0.42701034318203424
Probability of profit_3x_budget for genre_Sci-Fi: 0.46939119933033413
Probability of profit_3x_budget for genre_Thriller: 0.4877441518021466
```

Conclusion

06

Conclusion



Outcome:

1. Prediction model reliable in predicting gross revenue but not profit and budget which solves one of our original problem
2. Top 3 variables to predict gross are director, writer and budget

Interesting finding:

1. Getting a high IMDb score does not mean the movie will lead to high gross revenue.
2. Better writer and directors will drastically improve the movie box office
3. Higher budget can lead to higher revenue due to higher quality workers and performers.

Thank you!