NOW SHOWING

SC1015 - B135 TEAM 2 - MINI PROJECT

CRACKING THE CODE TO BOX OFFICE SUCCESS

STARRING:

LEE CHENG YAO, YAU JUN HAO, ZANE YEE SUN

Industry Statistics

Problem Statement

Dataset

Data Cleaning

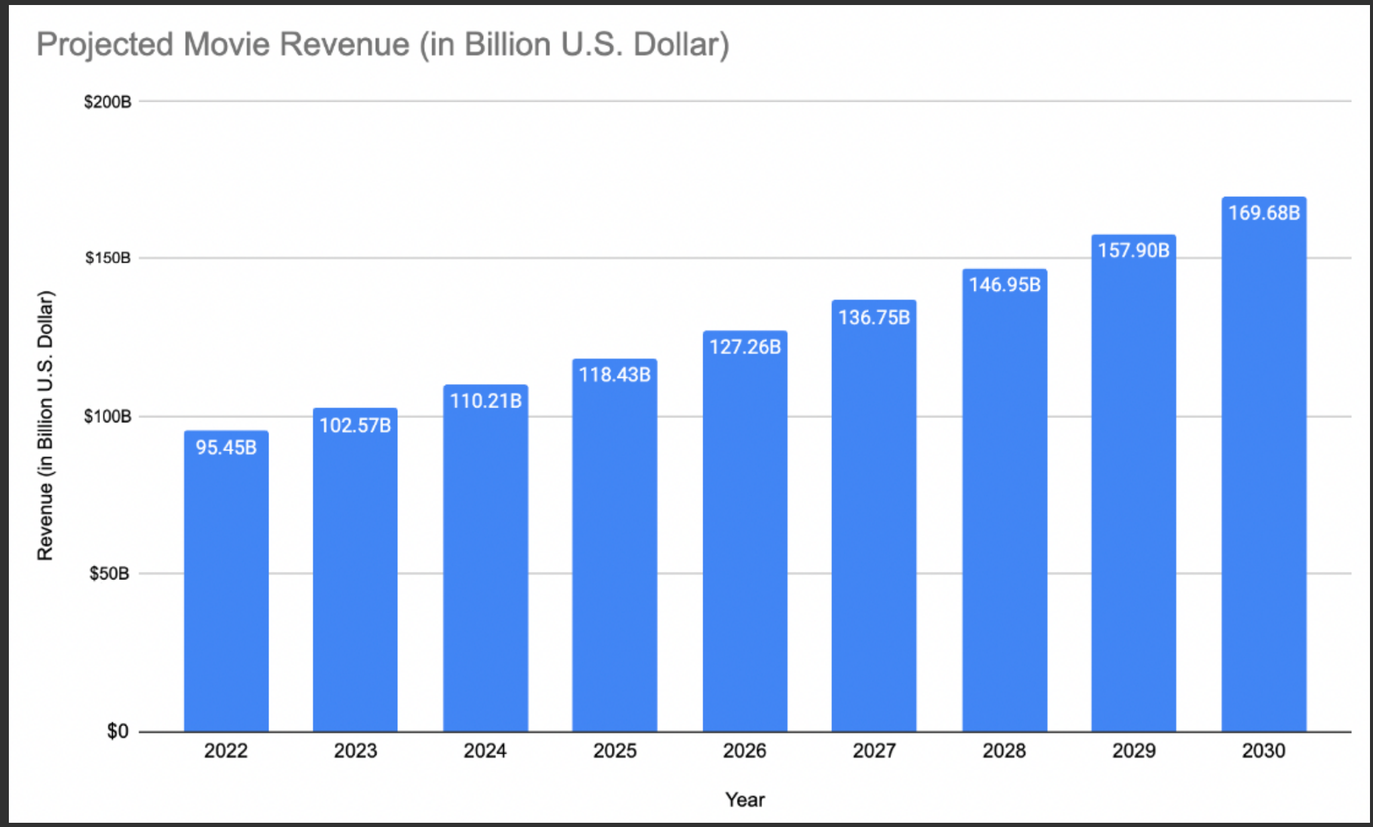
Movie Industry Statistics

- Enormous industry with global market size of USD$267 billion
- General increase in movie released since 2000s
- Peaked in 2019 with USD$42 billion in revenue

No. of movie released per year (2000 - 2022)

Movie Industry Statistics

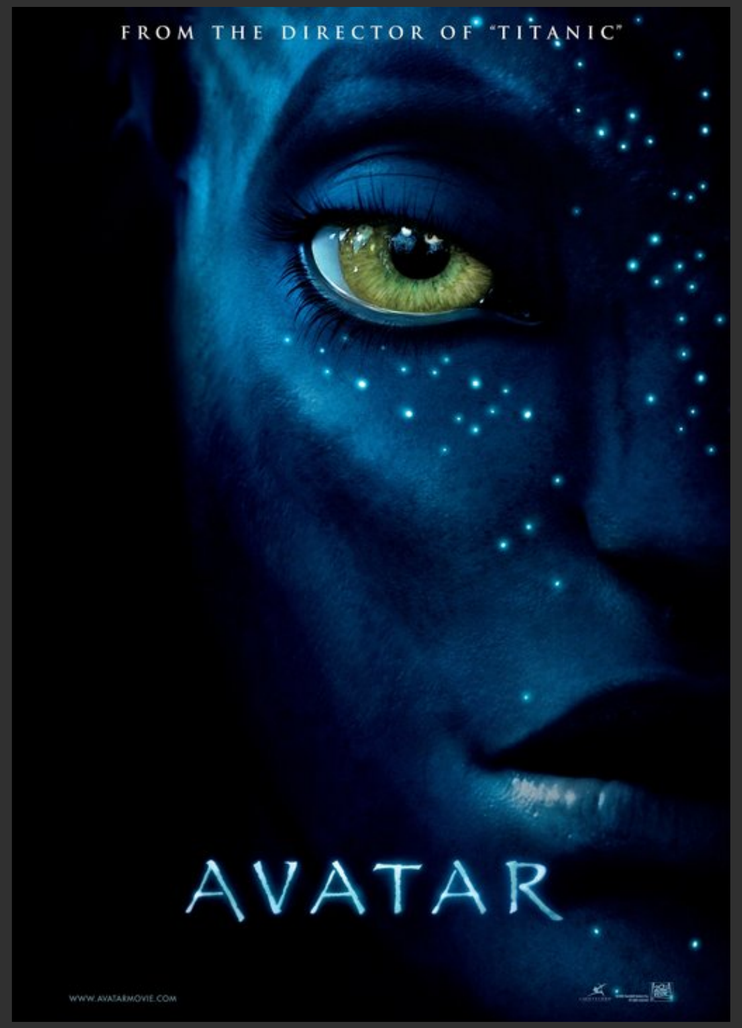
- 4.18 billion individuals have access to the internet
- Increasing video media entertainment consumption
- Expected large growth for film market by 2030



Projected Yearly Revenue in the Film Industry (2022 - 2030)

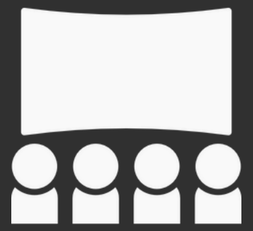
Movie Industry Statistics

- Top 50 highest-grossing films earns > USD$1 billion
- Gross revenue is influenced by different factors and preferences



PRACTICAL MOTIVATION

SAMPLE COLLECTION



Movies Go-ers will ask:
What movies should I watch?

Actors will ask:
What is the potential success of this movie?

Directors and Producers will ask:
How much revenue will my movie generate?

Problem Statement:

How can we create the
MOST successful movie?

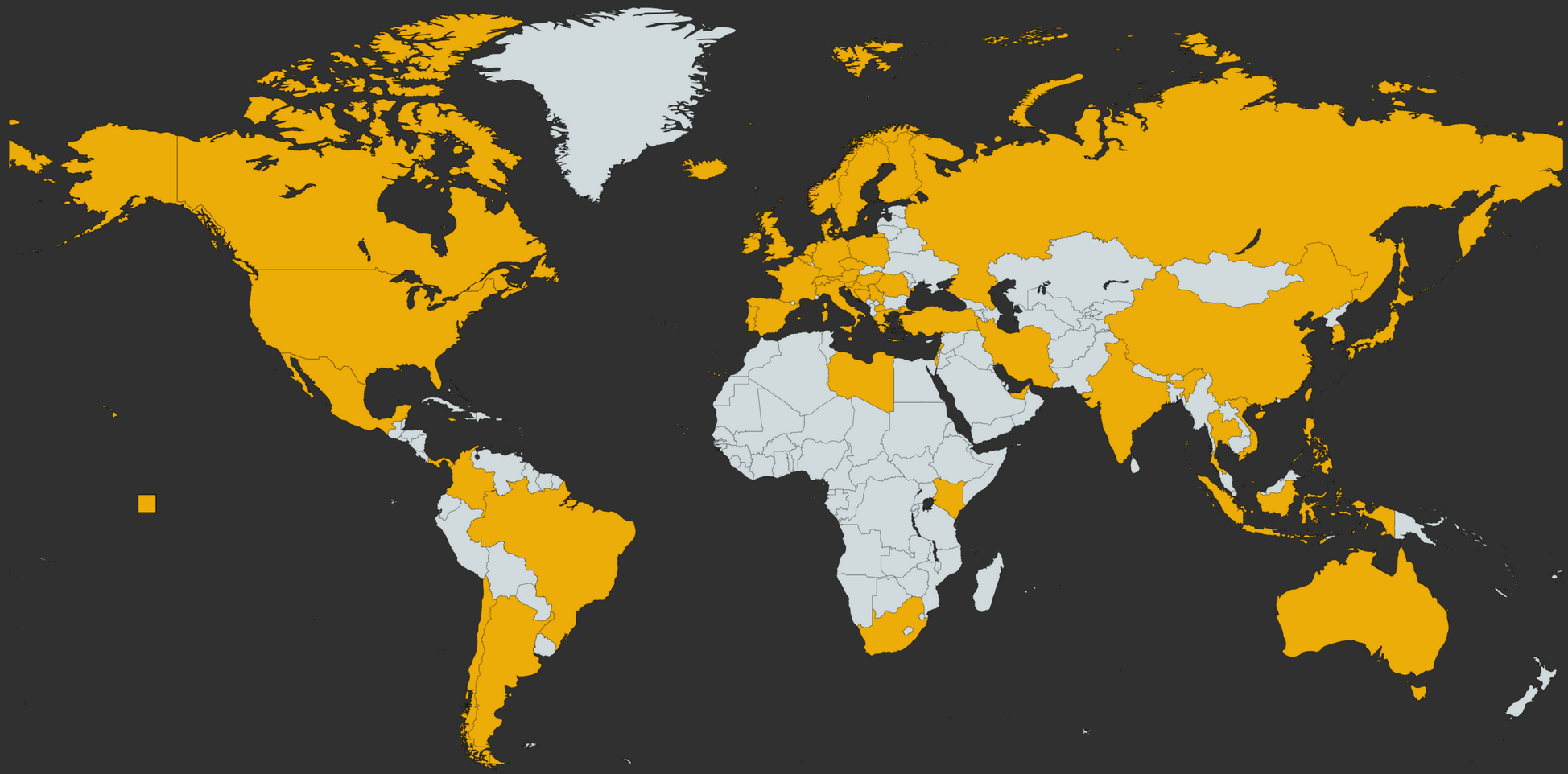
Dataset Used

IMDb

7668 MOVIES

1986 — 4 DECADES — 2020

59 COUNTRIES



Dataset Used

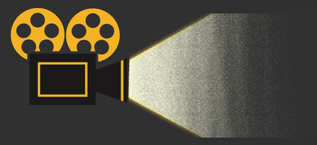
15 Variables

Variables		
Name	Rating	Genre
Year	Released	Score
Votes	Director	Writer
Star	Country	Budget
Company	Runtime	Gross

SUCCESS = REVENUE

The "gross" column represents revenue earned by a movie

#	Column
0	name
1	rating
2	genre
3	year
4	released
5	score
6	votes
7	director
8	writer
9	star
10	country
11	budget
12	gross
13	company
14	runtime



PROBLEM FORMULATION

DATA PREPARATION

Numerical Data	Categorical Data	
Score	Name	Director
Votes	Rating	Writer
Budget	Genre	Star
Runtime	Year	Country
Gross	Released	Company

Data Cleaning

No. of NULL Values

name	0
rating	77
genre	0
year	0
released	2
score	3
votes	3
director	0
writer	0
star	0
country	0
budget	2171
company	17
runtime	4
gross	189

Multiple columns with null values

1. Numeric (red box)
    a. Replaced with median value
2. Categorical (blue box)
    a. Removal of row (if < 2% of rows)
    b. "Rating" - replaced with "Not Rated"

Reserving Test Samples

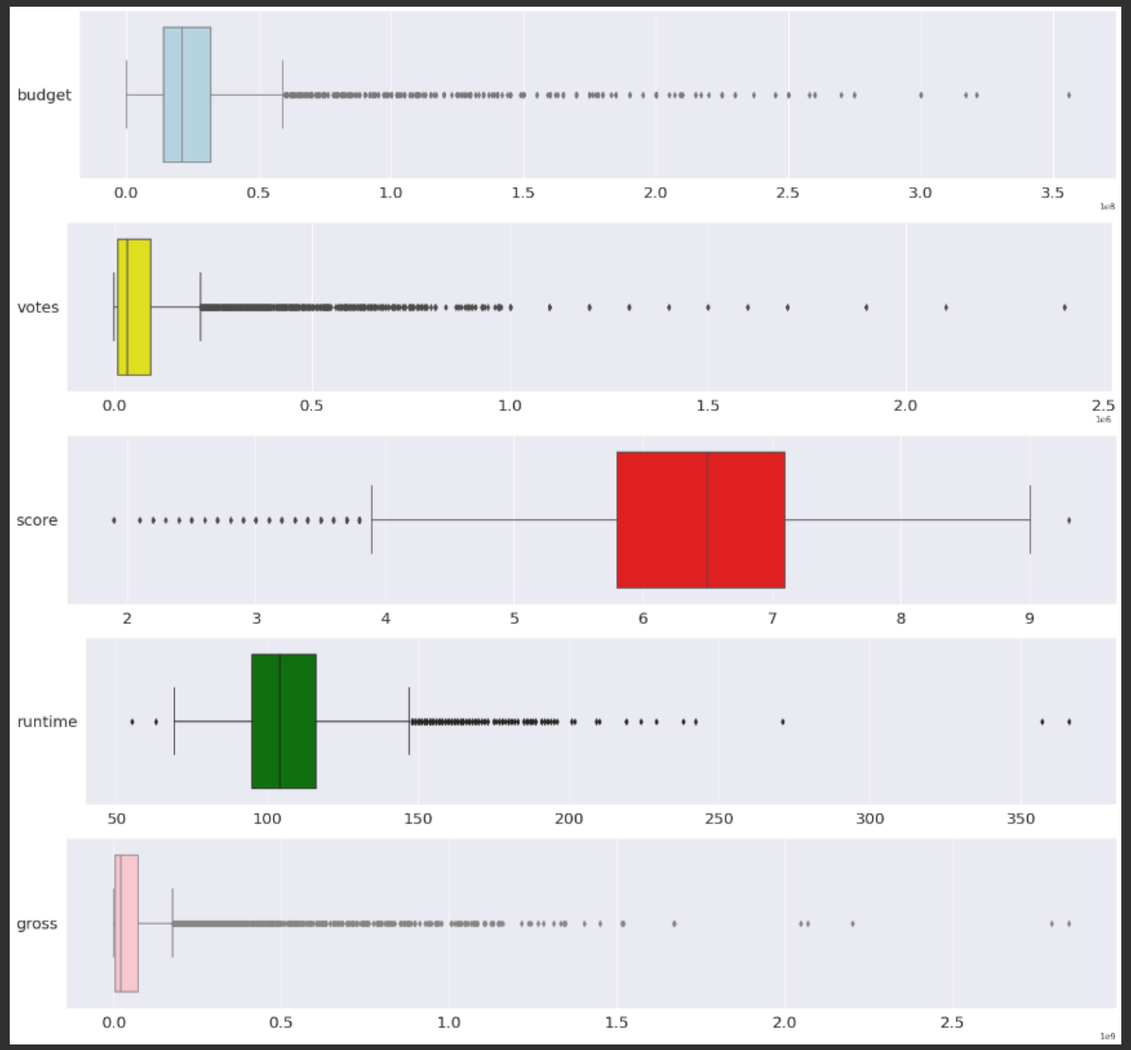
	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	company	runtime	gross
7280	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	Lucasfilm	124.0	538375067.0
1100	The Hunter	R	Adventure	2011	October 6, 2011 (Australia)	6.7	38000.0	Daniel Nettheim	Alice Addison	Willem Dafoe	Australia	21000000.0	Porchlight Films	100.0	1680778.0
1271	Videodrome	R	Horror	1983	February 4, 1983 (United States)	7.2	86000.0	David Cronenberg	David Cronenberg	James Woods	Canada	5952000.0	Filmplan International	87.0	2120439.0

Removing 3 random samples from main DataFrame for **Final Testing**

Uni-Variate Analysis

Bi-Variate Analysis

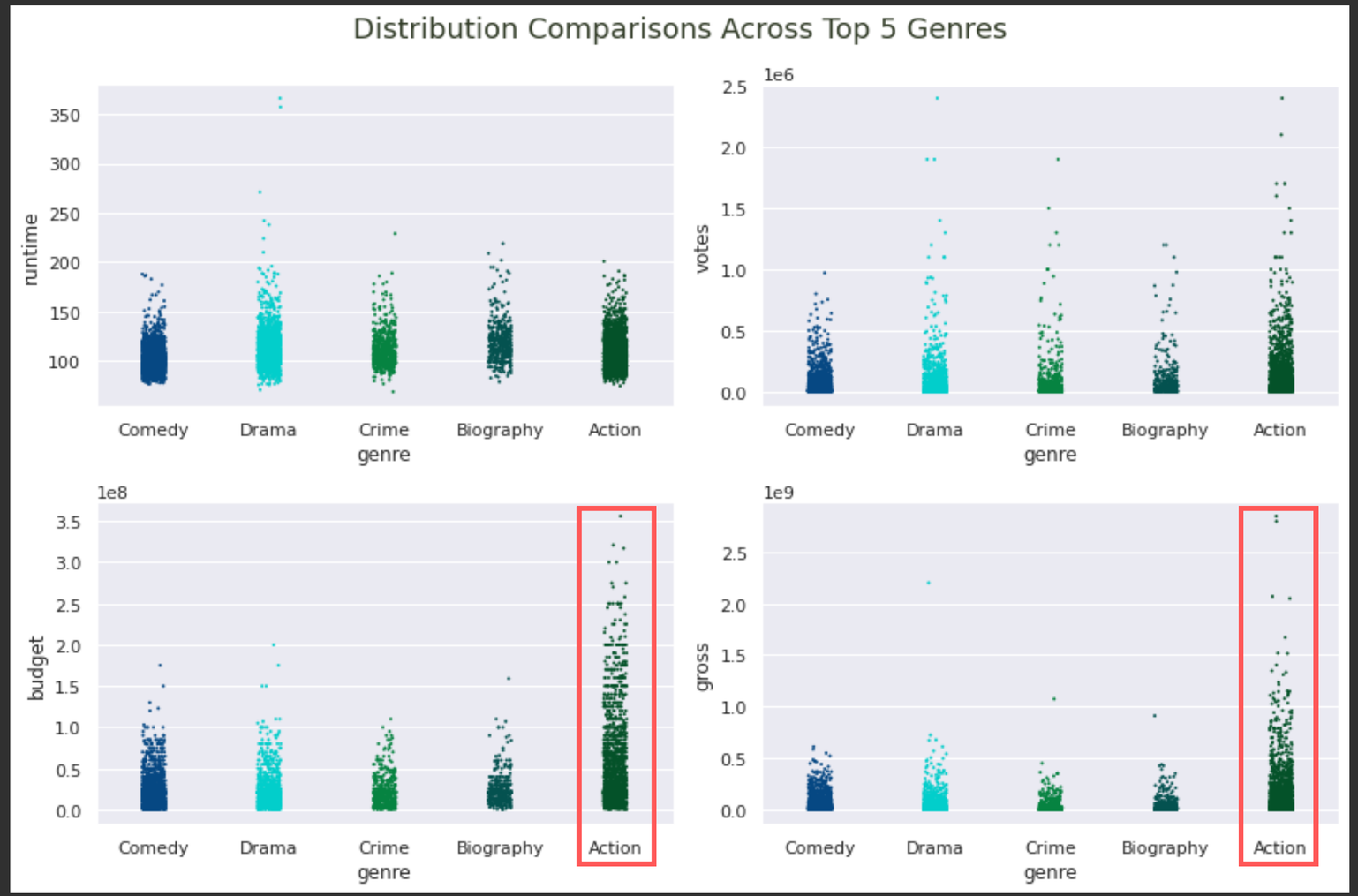
Uni-Variate Analysis: Boxplot



Identifying patterns and distributions

- Heavily right-skewed
- A large number of outliers

Bi-Variate Analysis: Strip-Plot



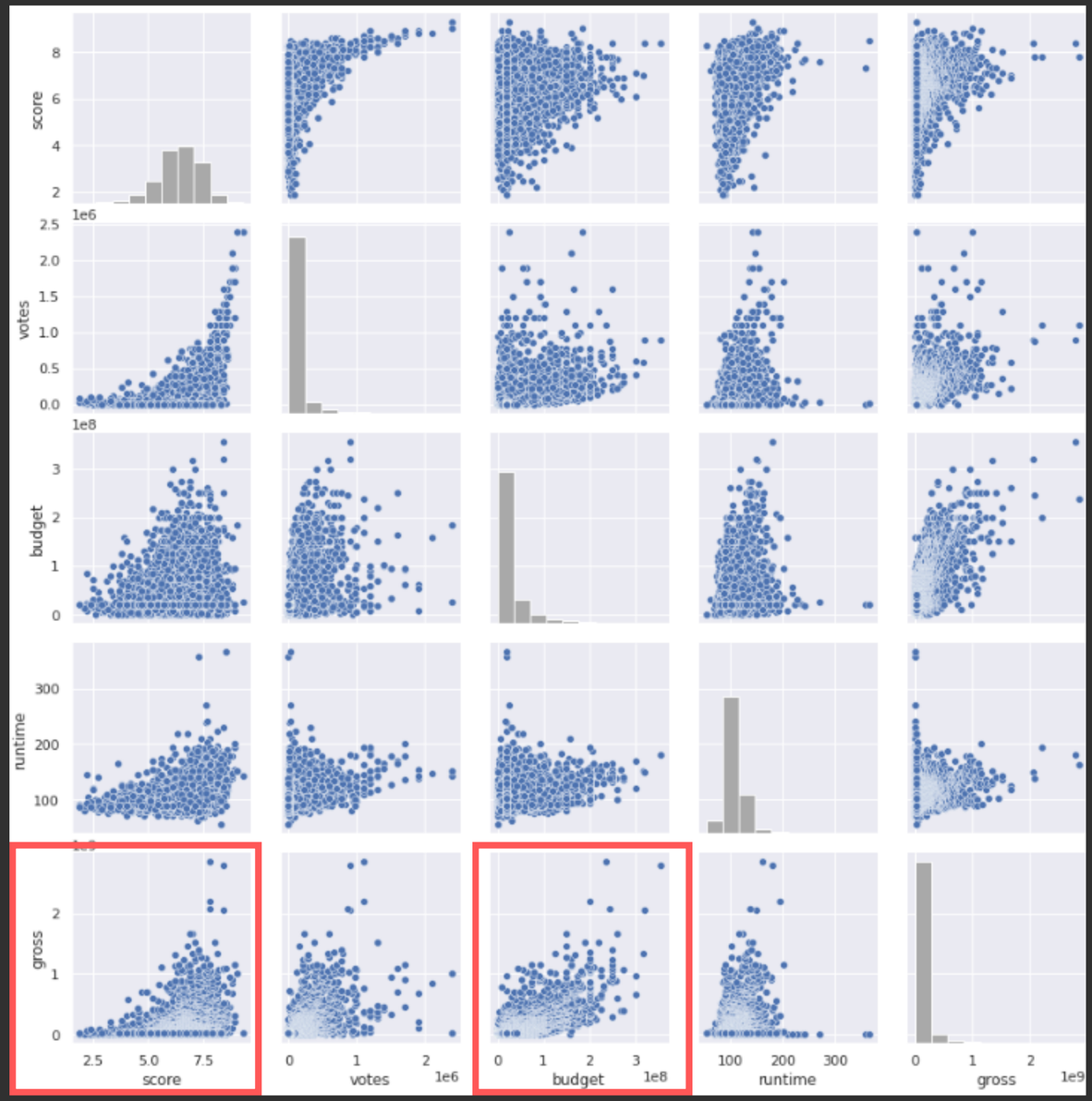
Identifying patterns

- Insights into industry trends
- Guide decision making

Pair Plots

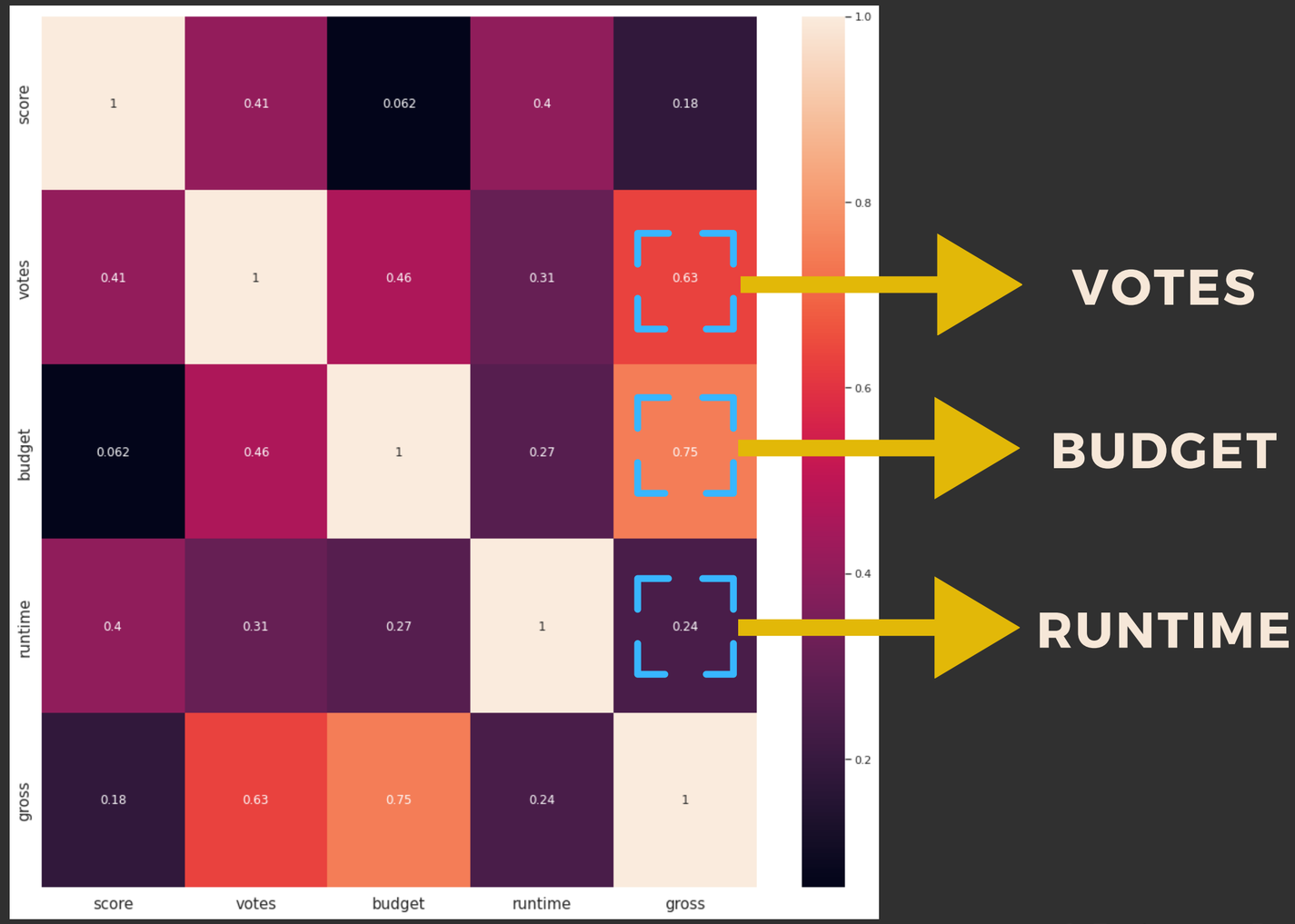
Relationship between variables and revenue (gross)

- Most plots show strong correlation with gross
- Some show non-linear relationship



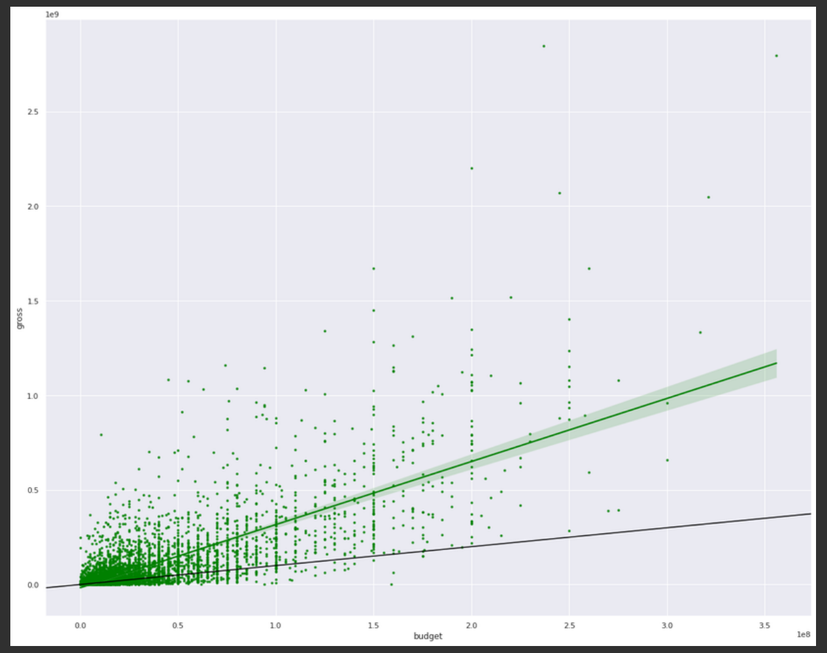
Correlation Matrix

- Deeper analysis of more prominent predictors
- Highest correlation with "gross"
- Relatively strong positive correlation between "gross" and predictors



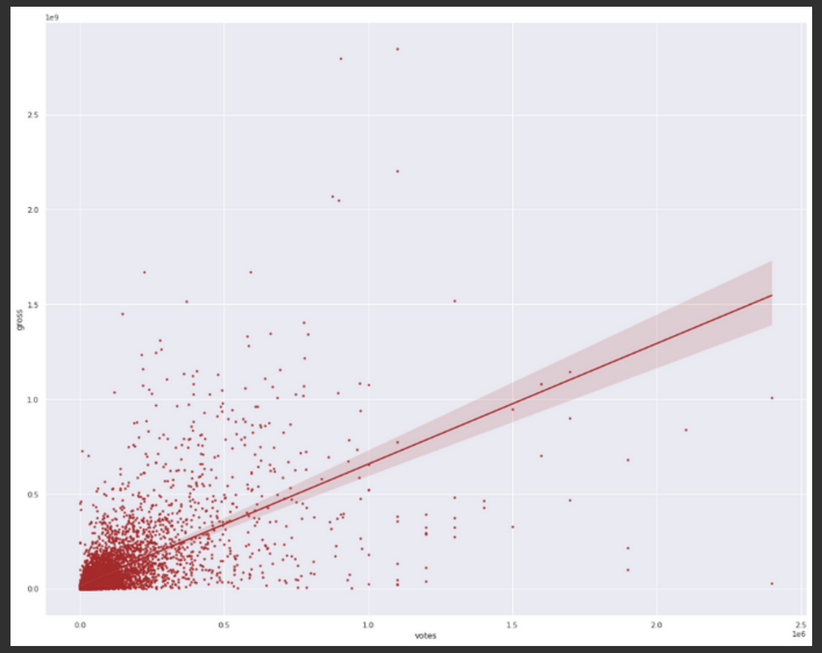
Bi-Variate Analysis

Gross - Budget



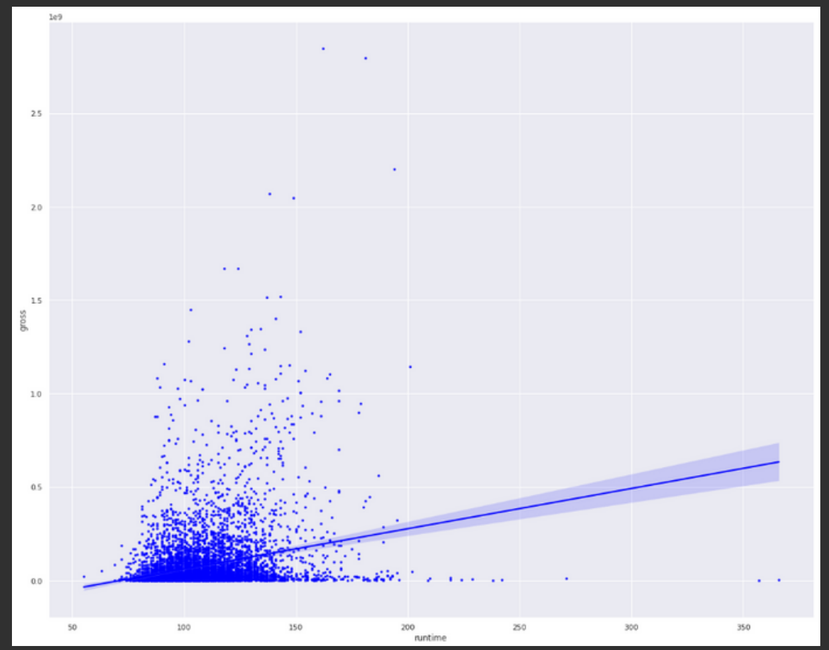
Correlation = 0.75

Gross - Votes



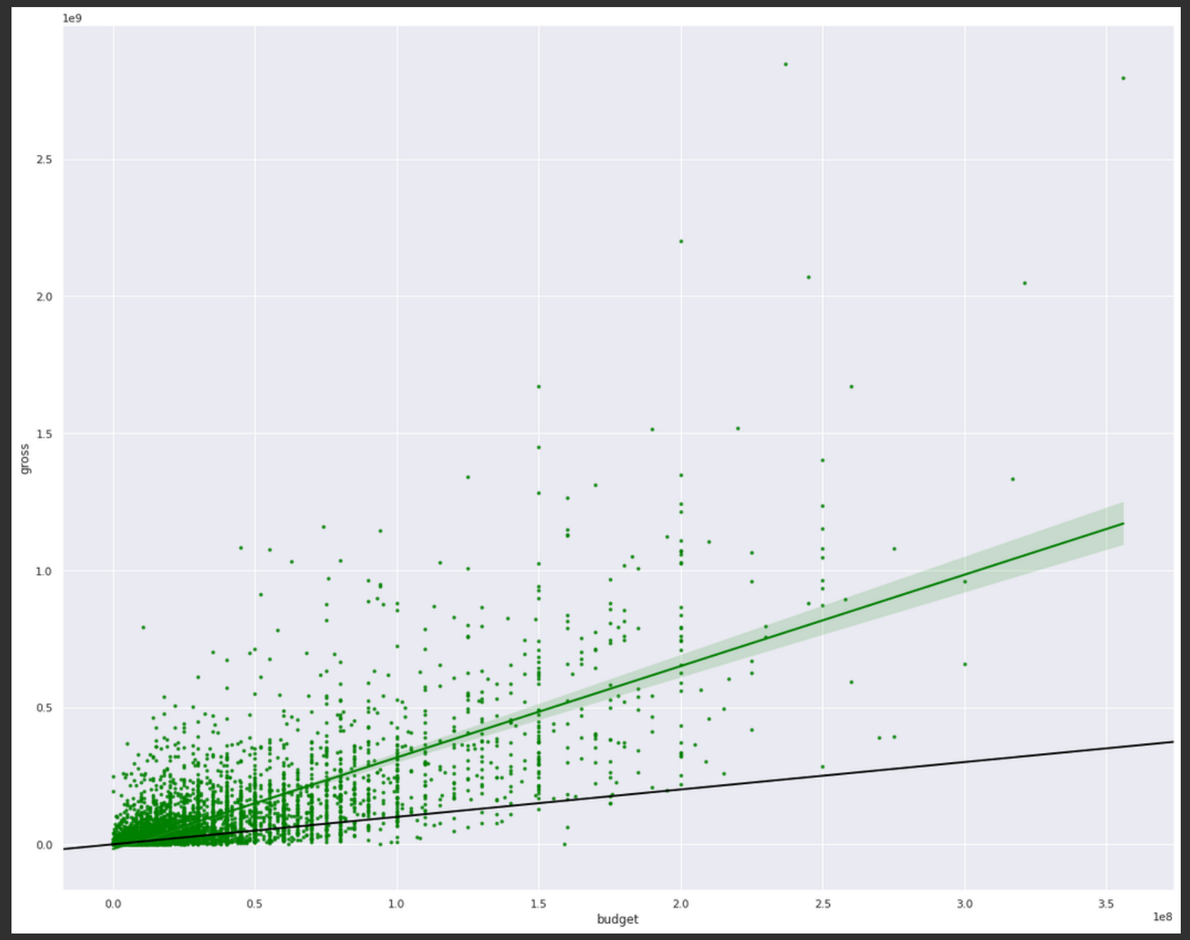
Correlation = 0.63

Gross - Runtime



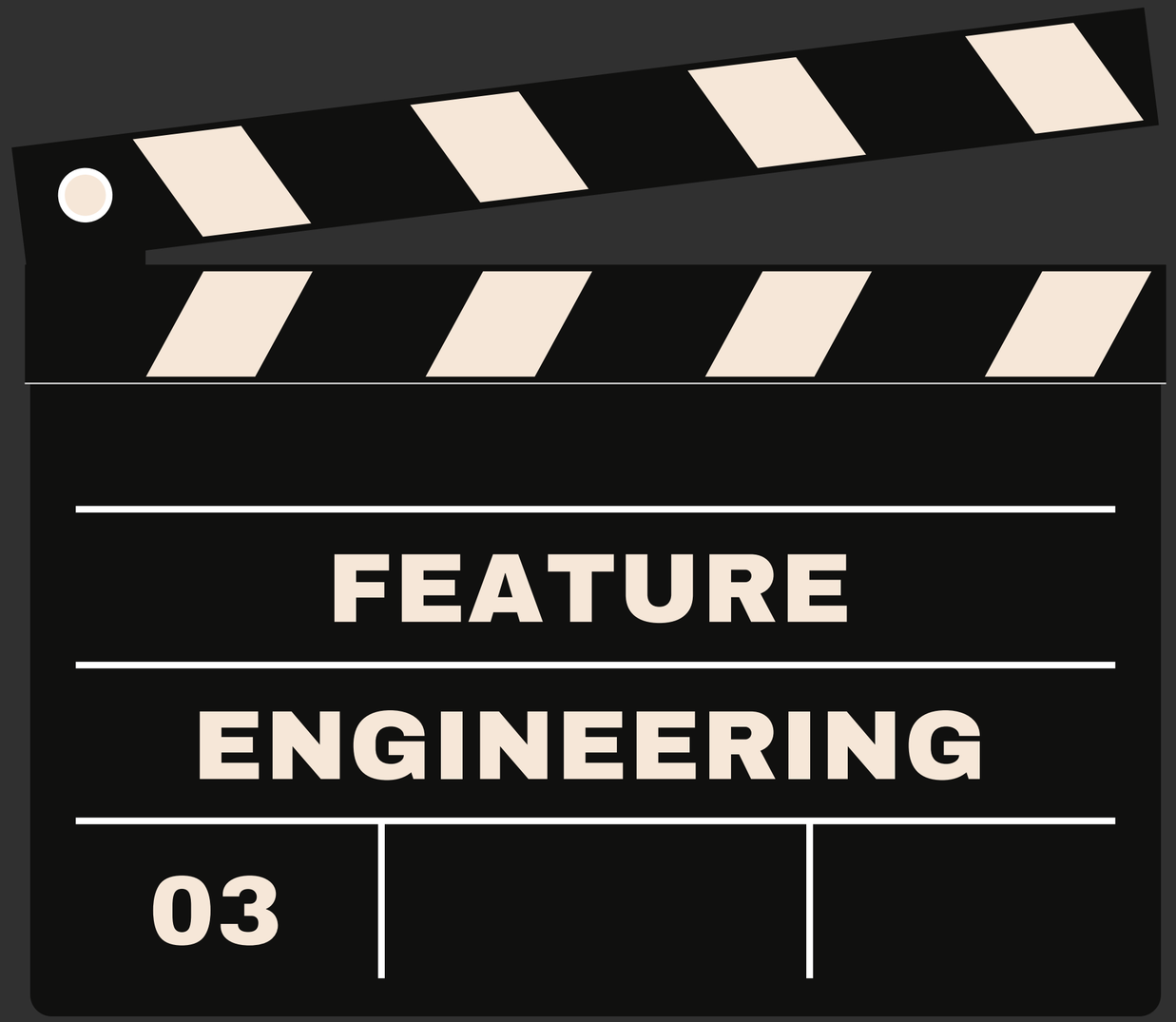
Correlation = 0.24

Bi-Variate Analysis



Gross - Budget

- Black line: Budget = Gross
- Majority of movies are profitable



Skewness Correction

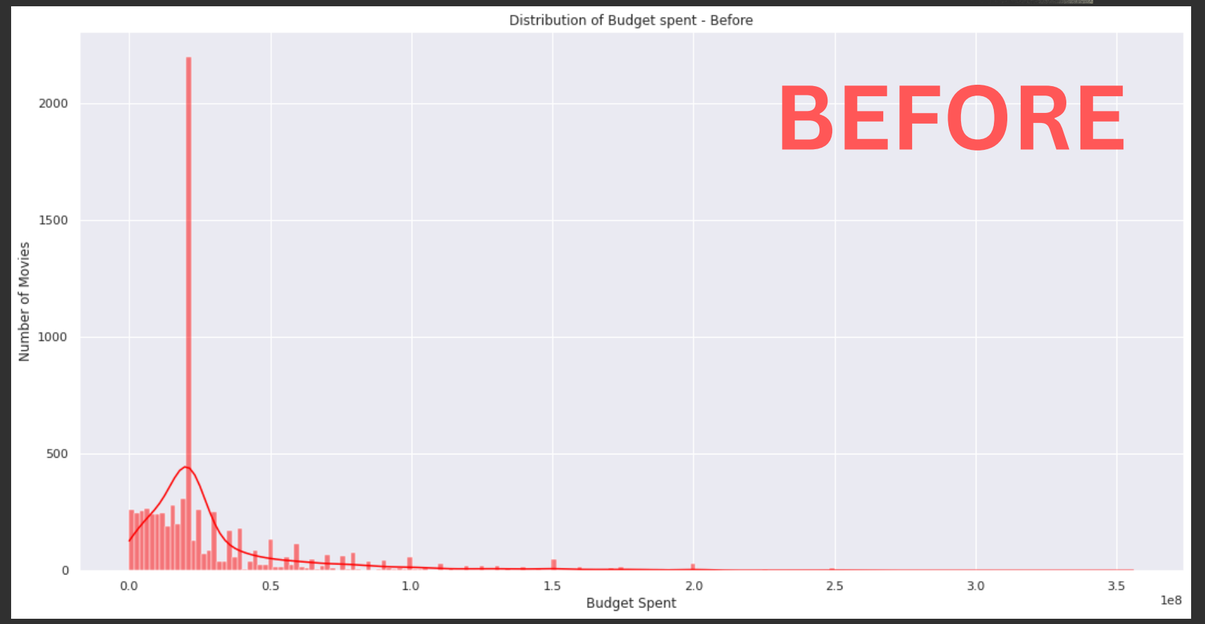
Remove Insignificant Data

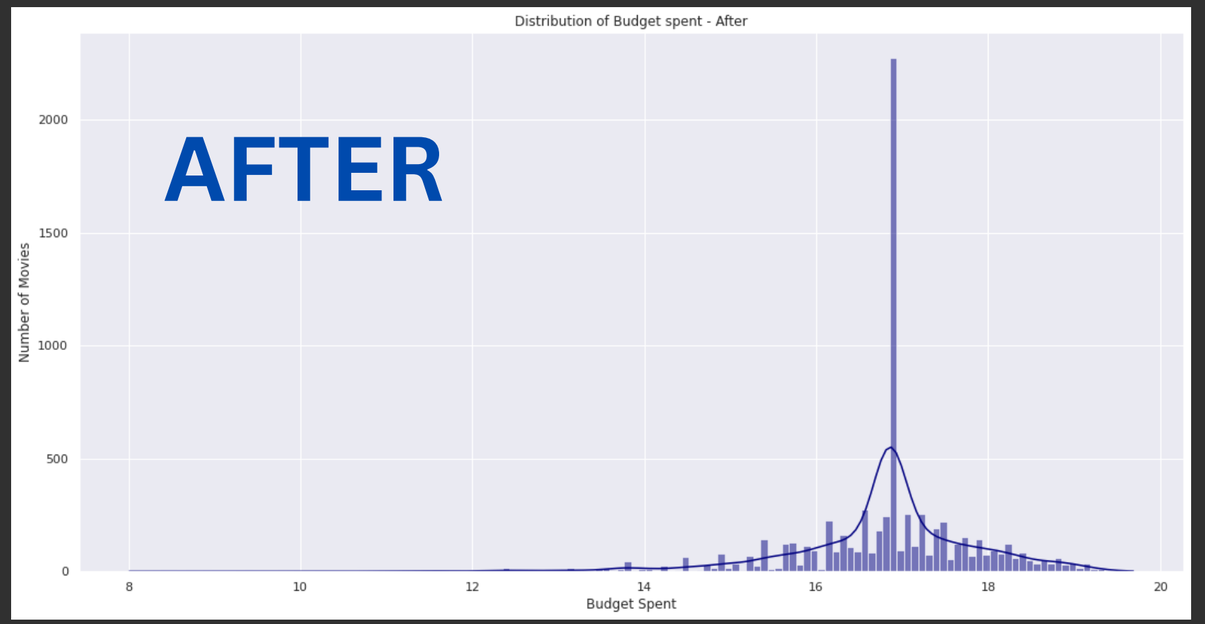
Ordinal Encoding

BUDGET DISTRIBUTION

Correcting Skewness

Logarithmic transformation for non-negative, right-skewed, numeric variables on histogram and KDE line

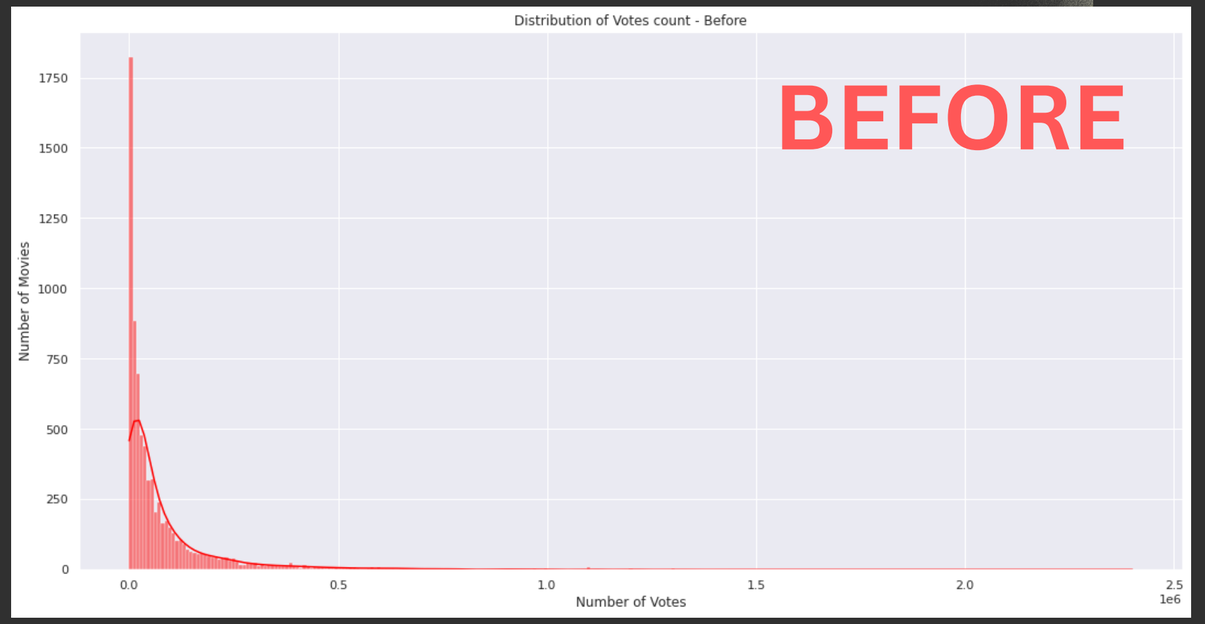


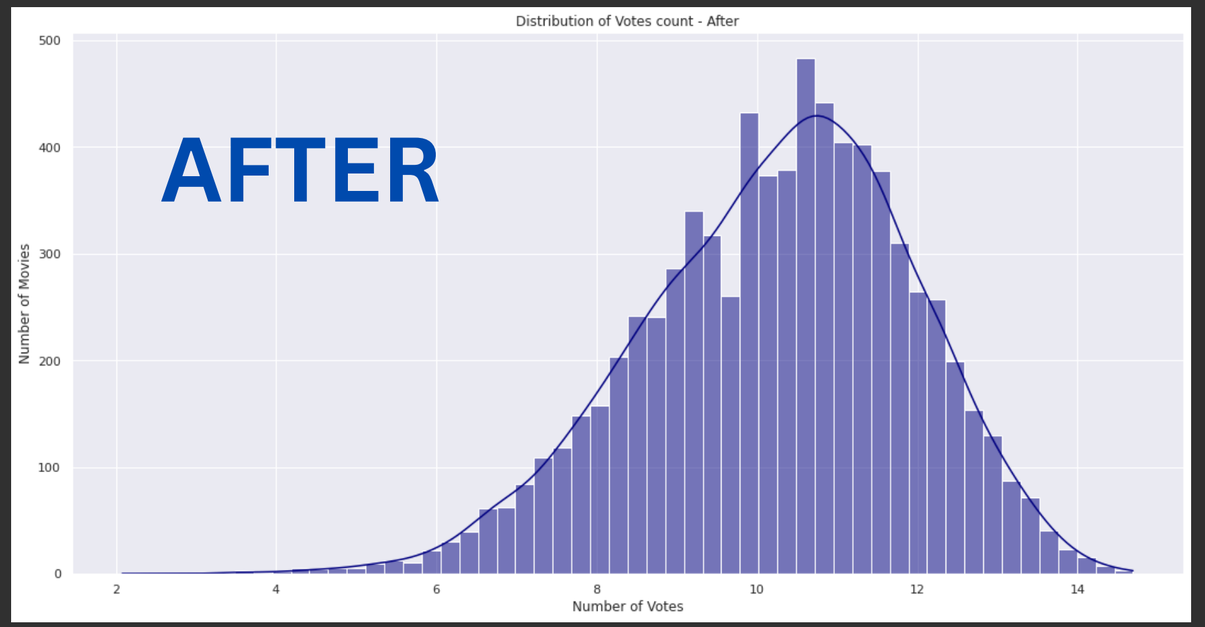


VOTES DISTRIBUTION

Correcting Skewness

Logarithmic transformation for non-negative, right-skewed, numeric variables on histogram and KDE line

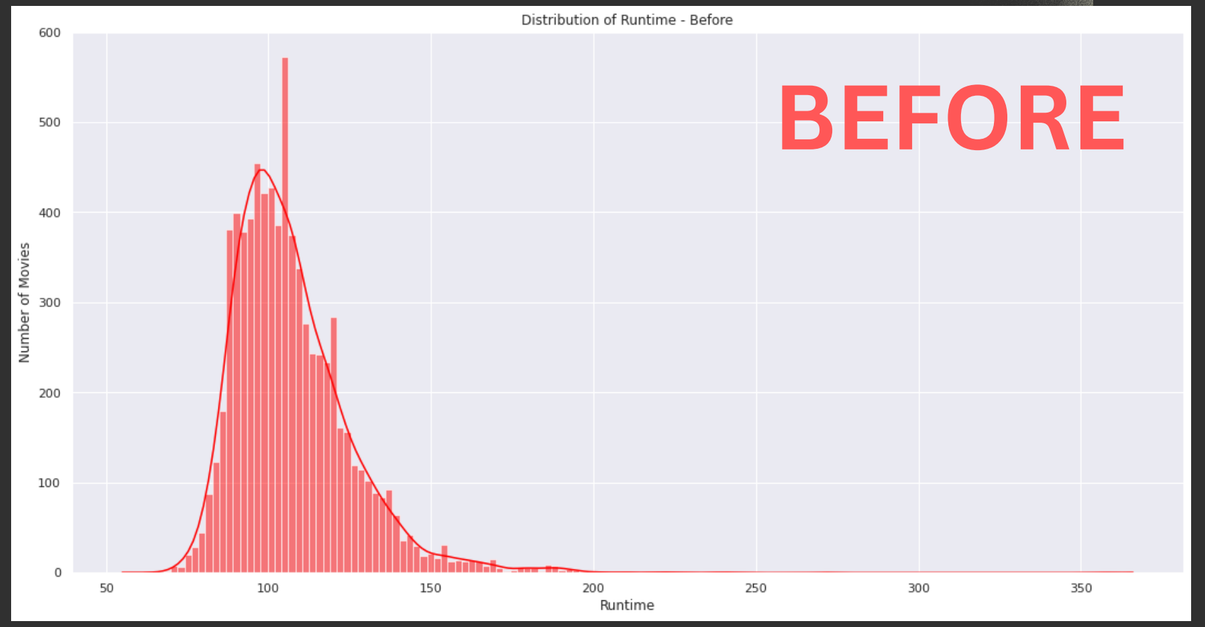


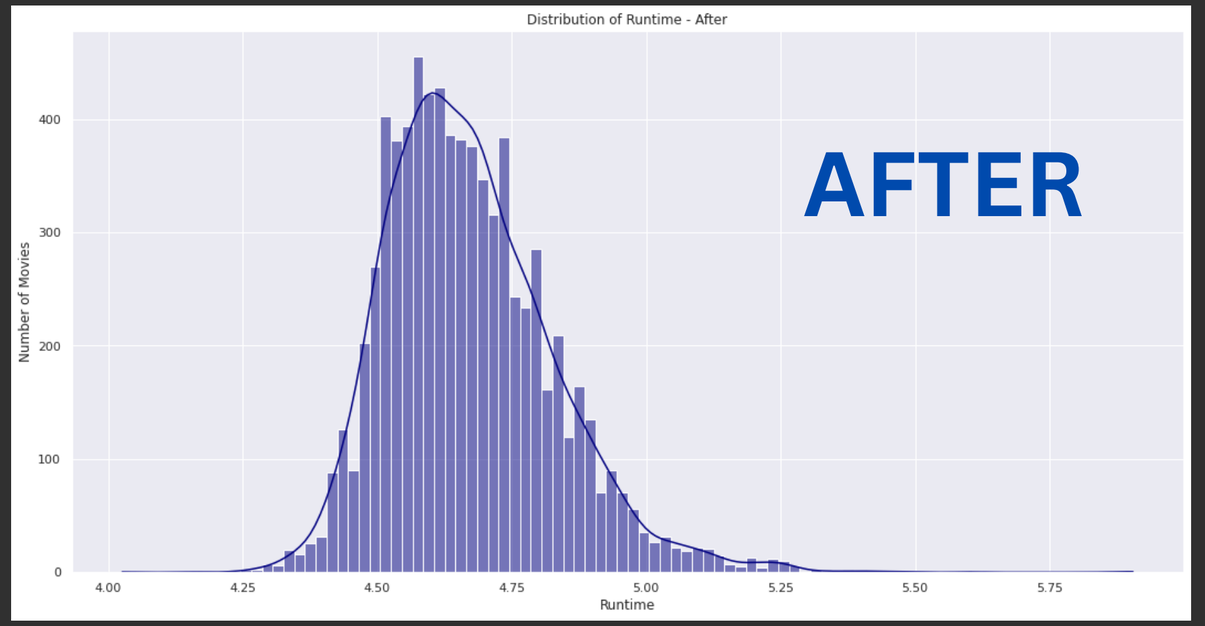


RUNTIME DISTRIBUTION

Correcting Skewness

Logarithmic transformation for non-negative, right-skewed, numeric variables on histogram and KDE line

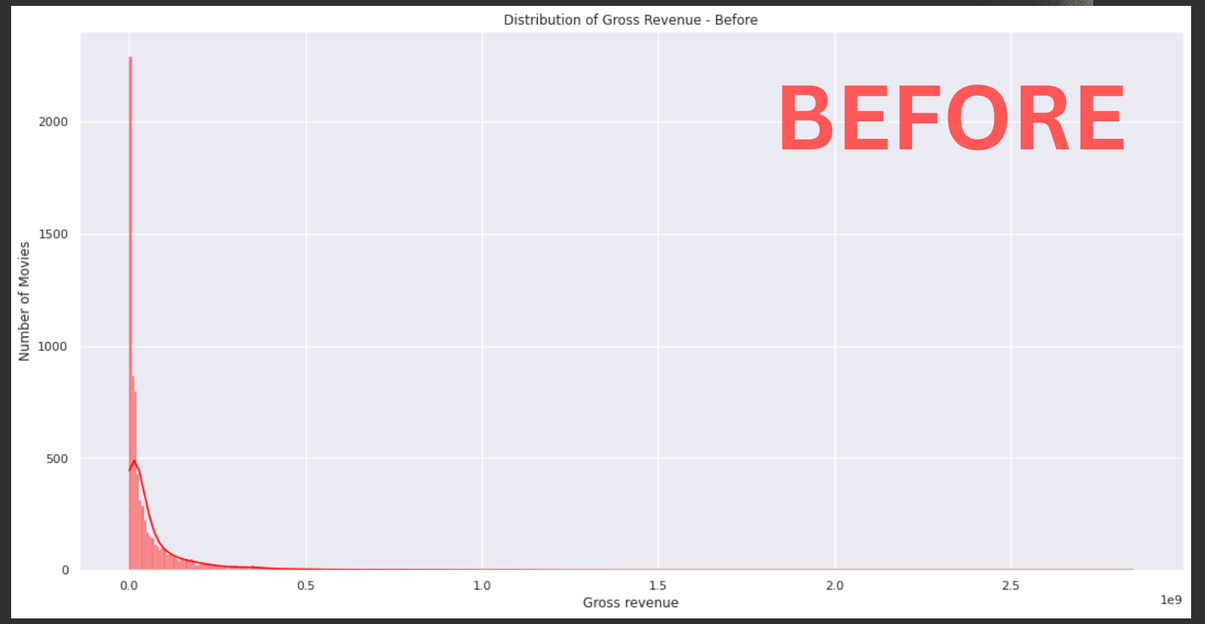


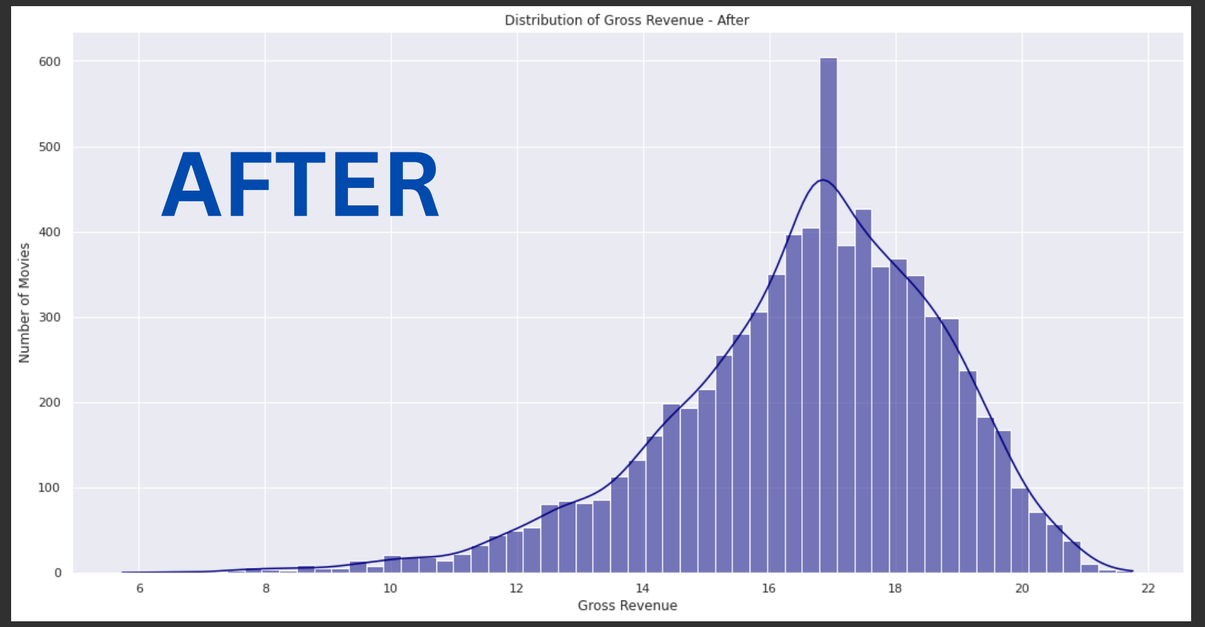


GROSS DISTRIBUTION

Correcting Skewness

Logarithmic transformation for non-negative, right-skewed, numeric variables on histogram and KDE line





REMOVE INSIGNIFICANT DATA

Drop columns that do not add value

- "Released" - very similar to country
- "Year" - uncontrollable variable since we cannot traverse time

We will be removing these columns to prevent inaccuracies

	name	rating	genre	score	votes	director	writer	star	country	budget	company	runtime	gross
0	Trojan War	PG-13	Comedy	5.7	8.665786	George Huang	Andy Burg	Will Friedle	United States	16.523561	Daybreak	4.454347	5.736572
1	Madadayo	Not Rated	Drama	7.3	8.537192	Akira Kurosawa	Ishirô Honda	Tatsuo Matsumura	Japan	16.292049	DENTSU Music And Entertainment	4.905275	6.391917
2	Run with the Hunted	Not Rated	Crime	5.2	6.601230	John Swab	John Swab	Ron Perlman	United States	16.860033	Roxwell Films	4.543295	6.526495
3	The Untold Story	Not Rated	Comedy	5.7	5.771441	Shane Stanley	Lee Stanley	Miko Hughes	United States	16.860033	Visual Arts Entertainment	4.653960	6.673298
4	Love, Honor and Obey	R	Comedy	6.5	8.556606	Dominic Anciano	Dominic Anciano	Sadie Frost	United Kingdom	16.860033	British Broadcasting Corporation (BBC)	4.644391	7.244942

Ordinal Encoder

- Converting categorical variables to numeric indexes
- Better predictions

Director	Rating	Genre
James Cameron	PG-13	Romance
Steven Spielberg	R	Action
Dominic Anciano	PG	Comedy
James Cameron	PG-13	Action

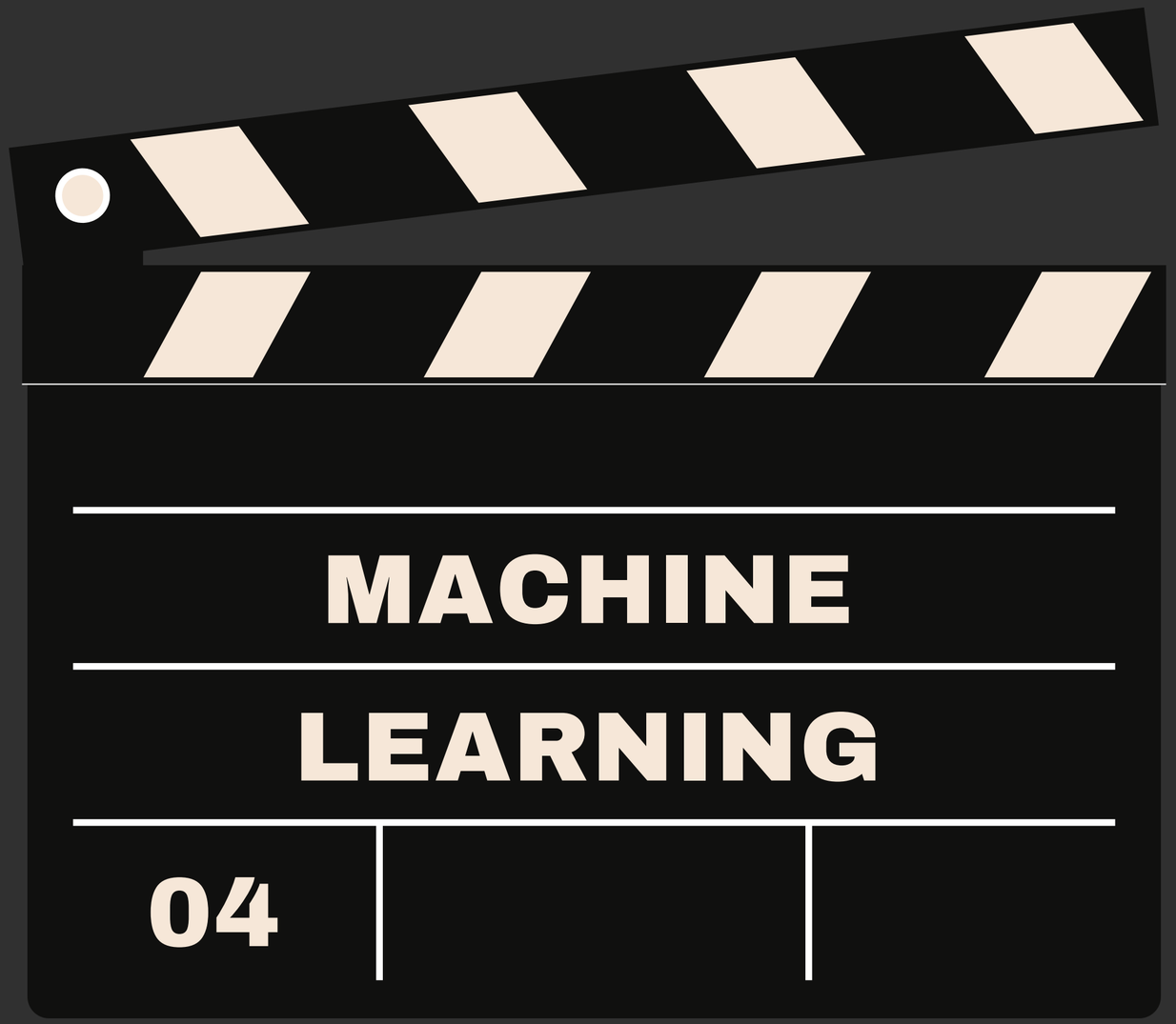
Ordinal Encoder

Director	Rating	Genre
1	1	1
2	2	2
3	3	3
1	1	2

DATA PRE-PROCESSING

These will be the variables passed into our prediction models

Variables						
Name	Rating	Genre	Score	Country	Budget	Company
Votes	Director	Writer	Star	Runtime	Gross	



Random Forest

Gradient Boosting

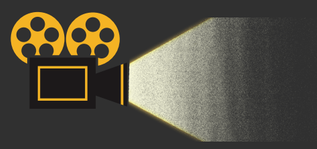
XGBoost

Light GBM

Multiple Linear*

Polynomial*

* will not be covered in presentation



MACHINE LEARNING

ALGORITHMIC OPTIMISATION

Train-Test Split

- Split the dataset into 80:20 ratio
- Response: "Gross"
- Predictors: as shown in table

K-Folds

- Performed 5 k-fold cross-validation

Predictors			
Name	Rating	Genre	Score
Votes	Director	Writer	Star
Country	Budget	Company	Runtime

Response
Gross

What are we looking for?

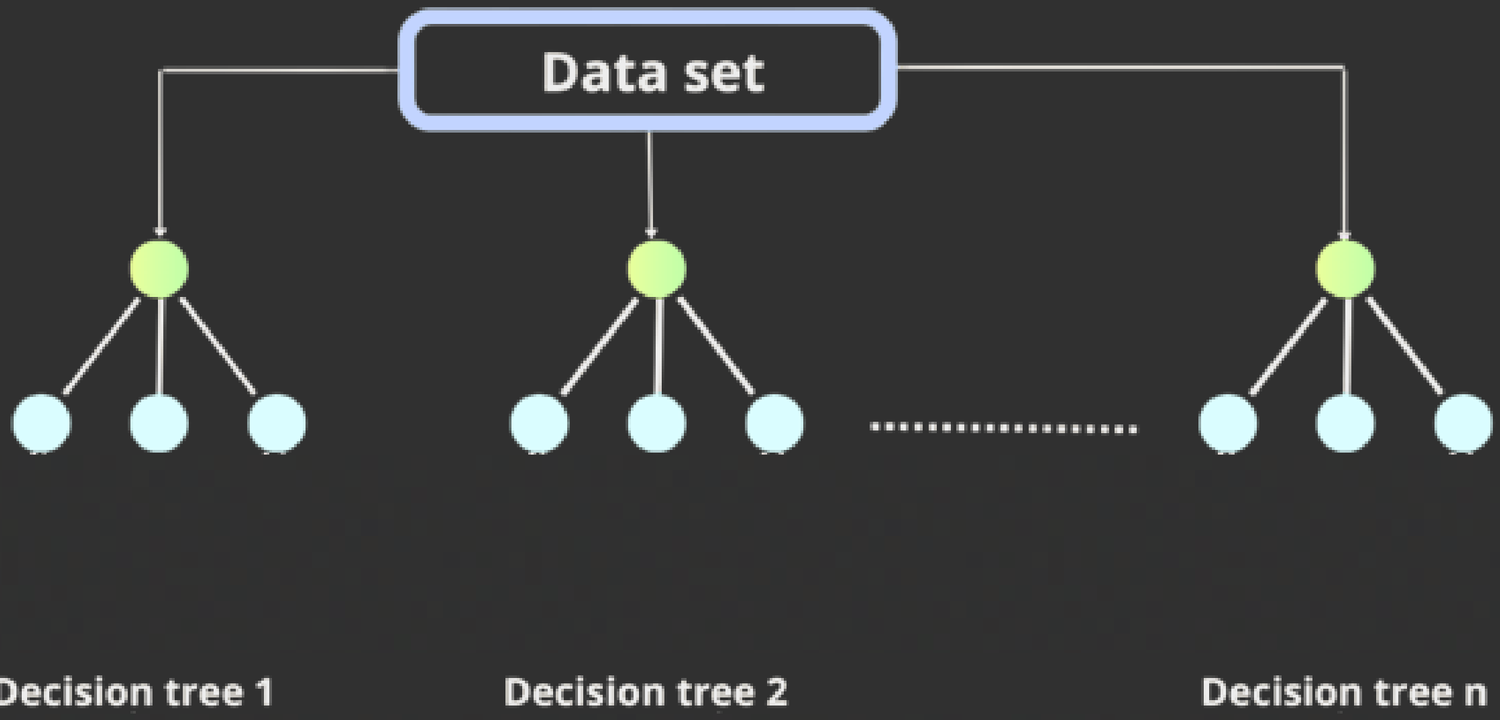
- Root Mean Squared Error (**RMSE**)
  - easily compared across different models
  - response value (gross) has been log-transformed

- Mean Absolute Percentage Error (**MAPE**)
  - provides a measure of relative error
  - easy to interpret - useful when communicating with non-technical stakeholders

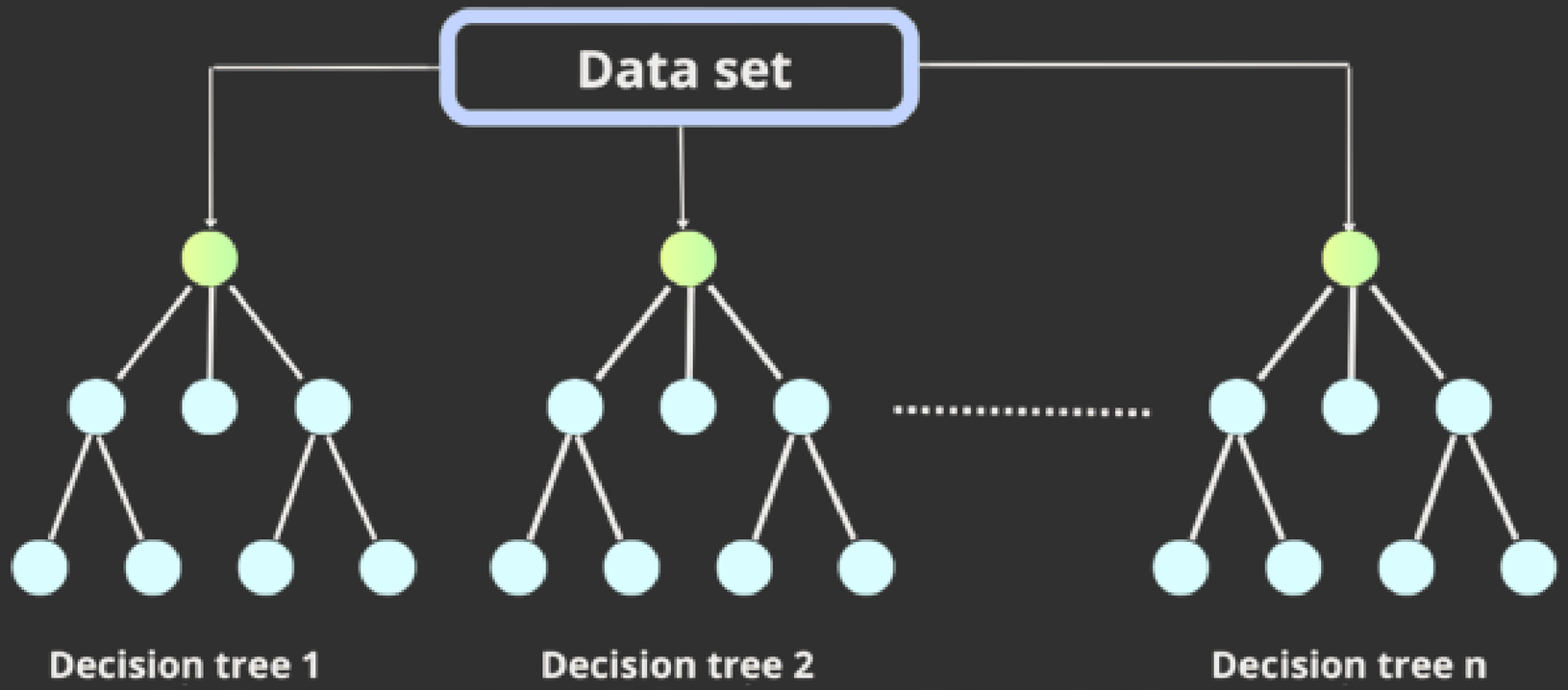
Random Forest

Decision tree 1 Decision tree 2 Decision tree n

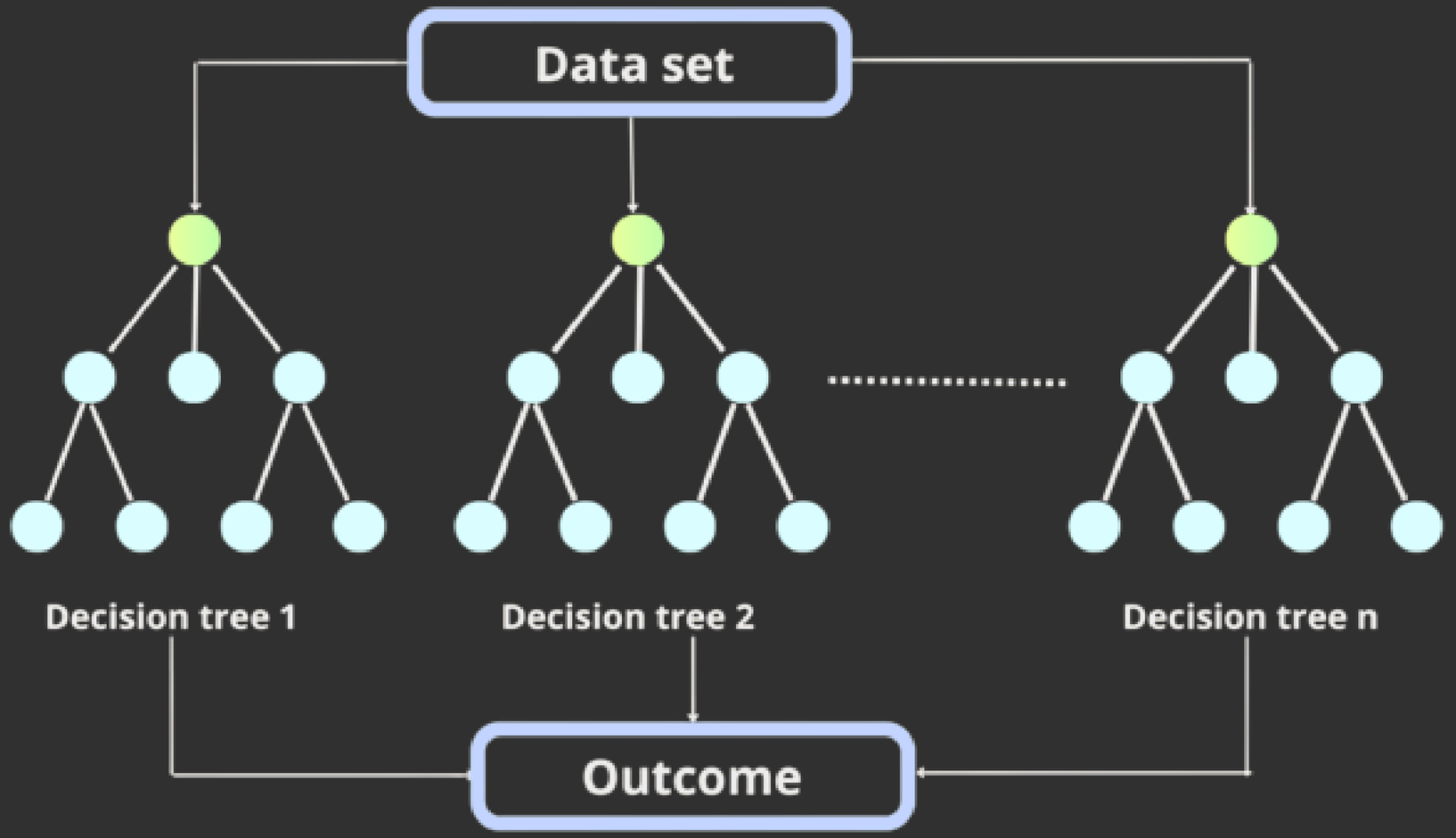
Random Forest



Random Forest



Random Forest



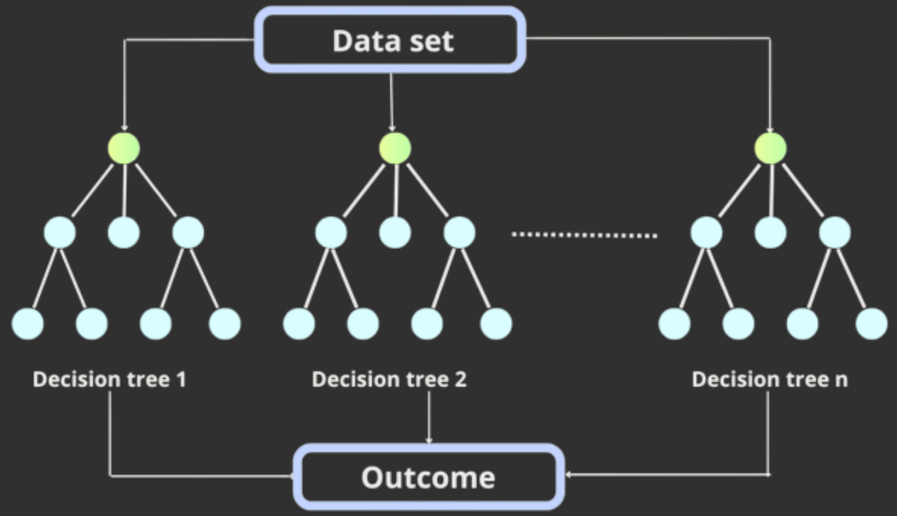
Random Forest

Advantages

- Reduce overfitting
- More **accurate** predictions
- Handles both **numeric** and **categorical** types

Disadvantages

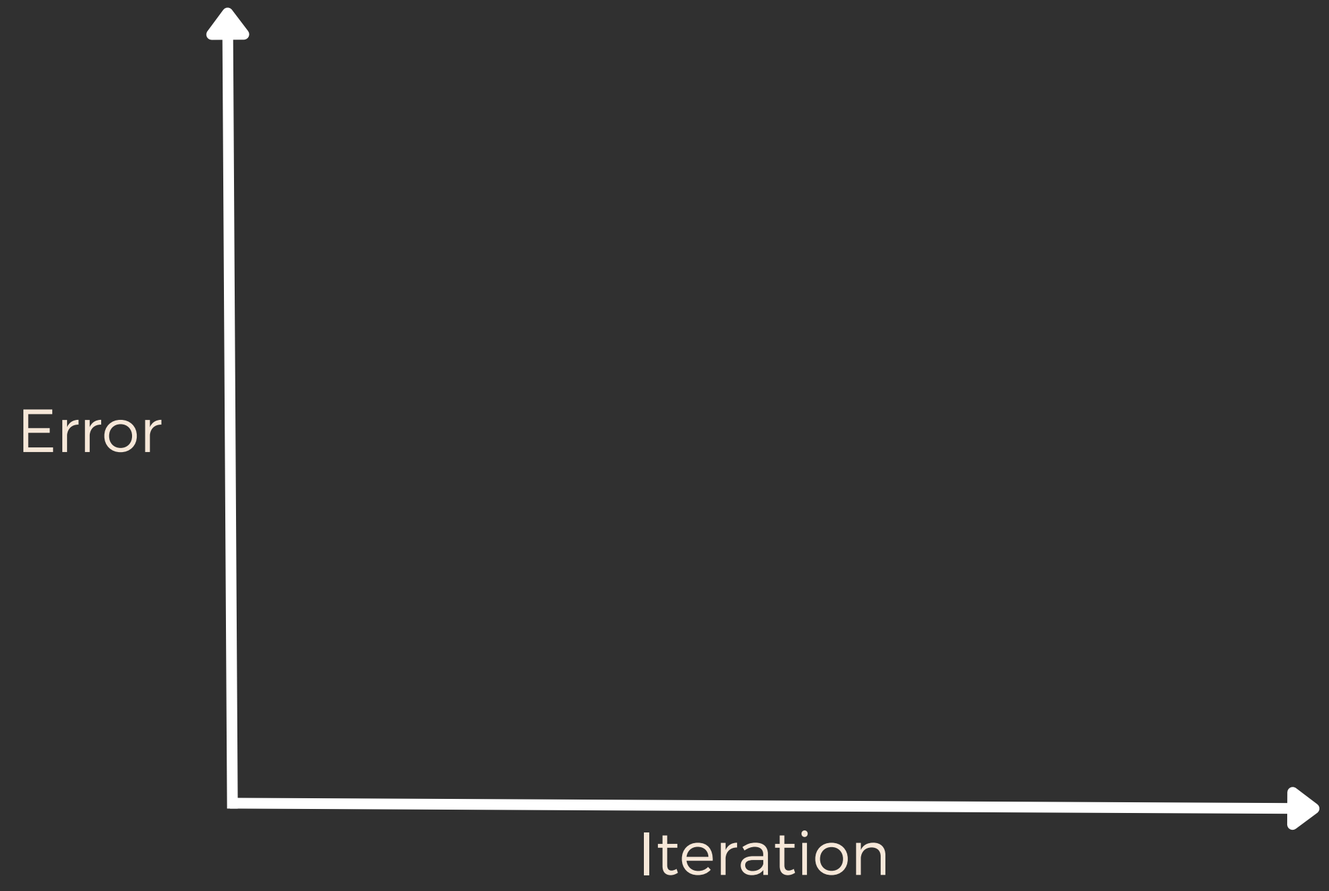
- Large computational power
- **Memory-intensive**, longer execution time



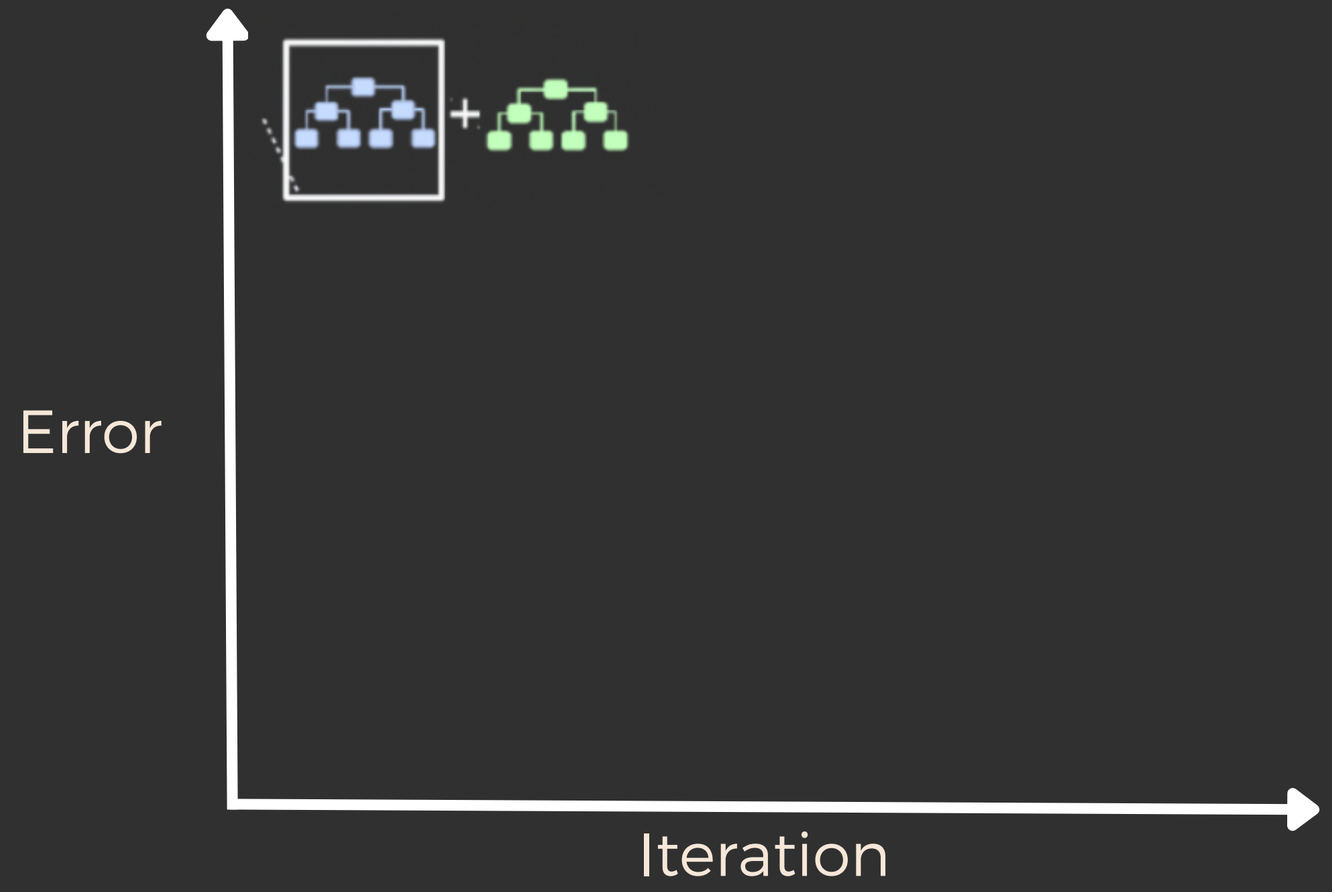
MODEL PERFORMANCE

RMSE - 1.37 MAPE - 6.51%

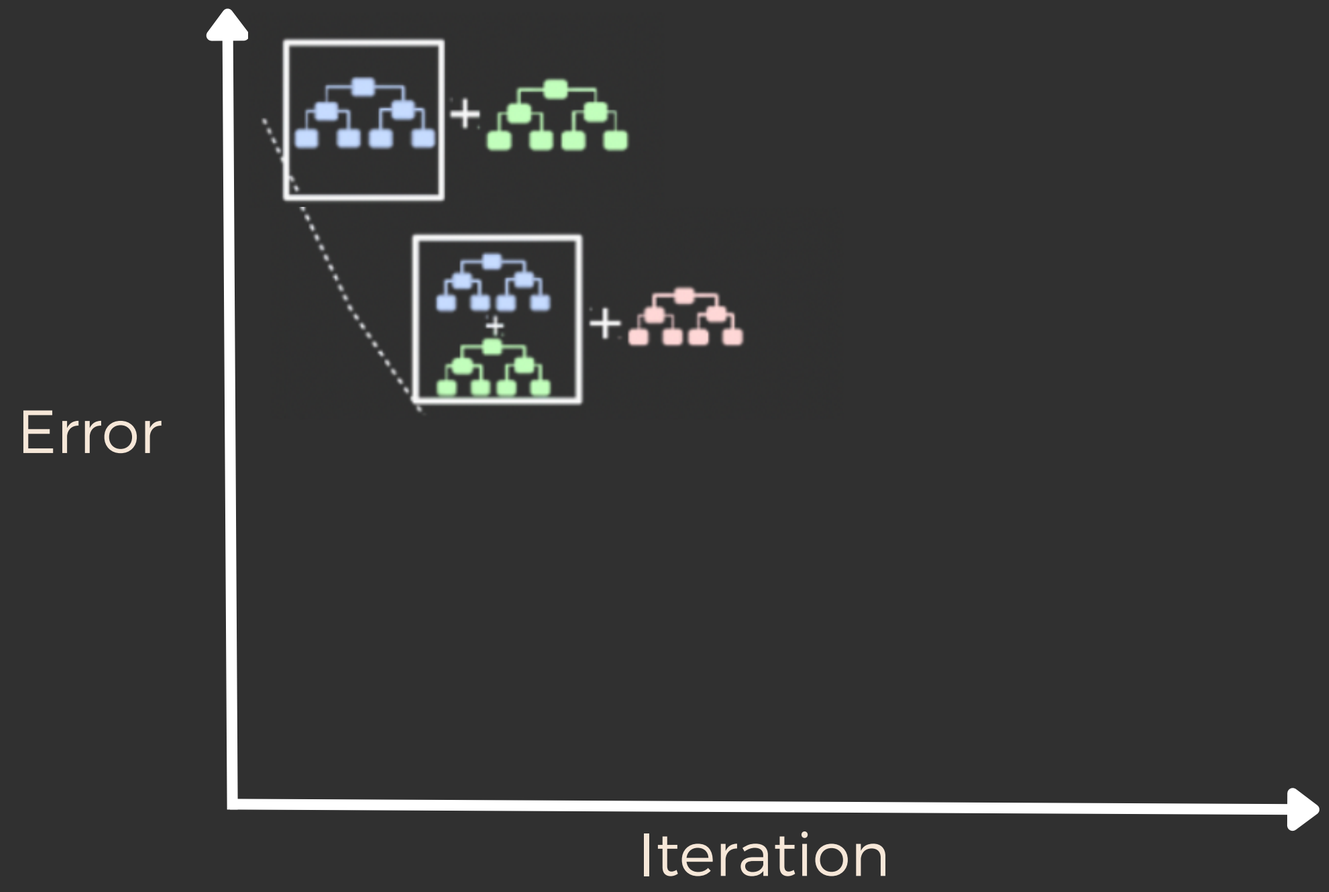
Gradient Boosting



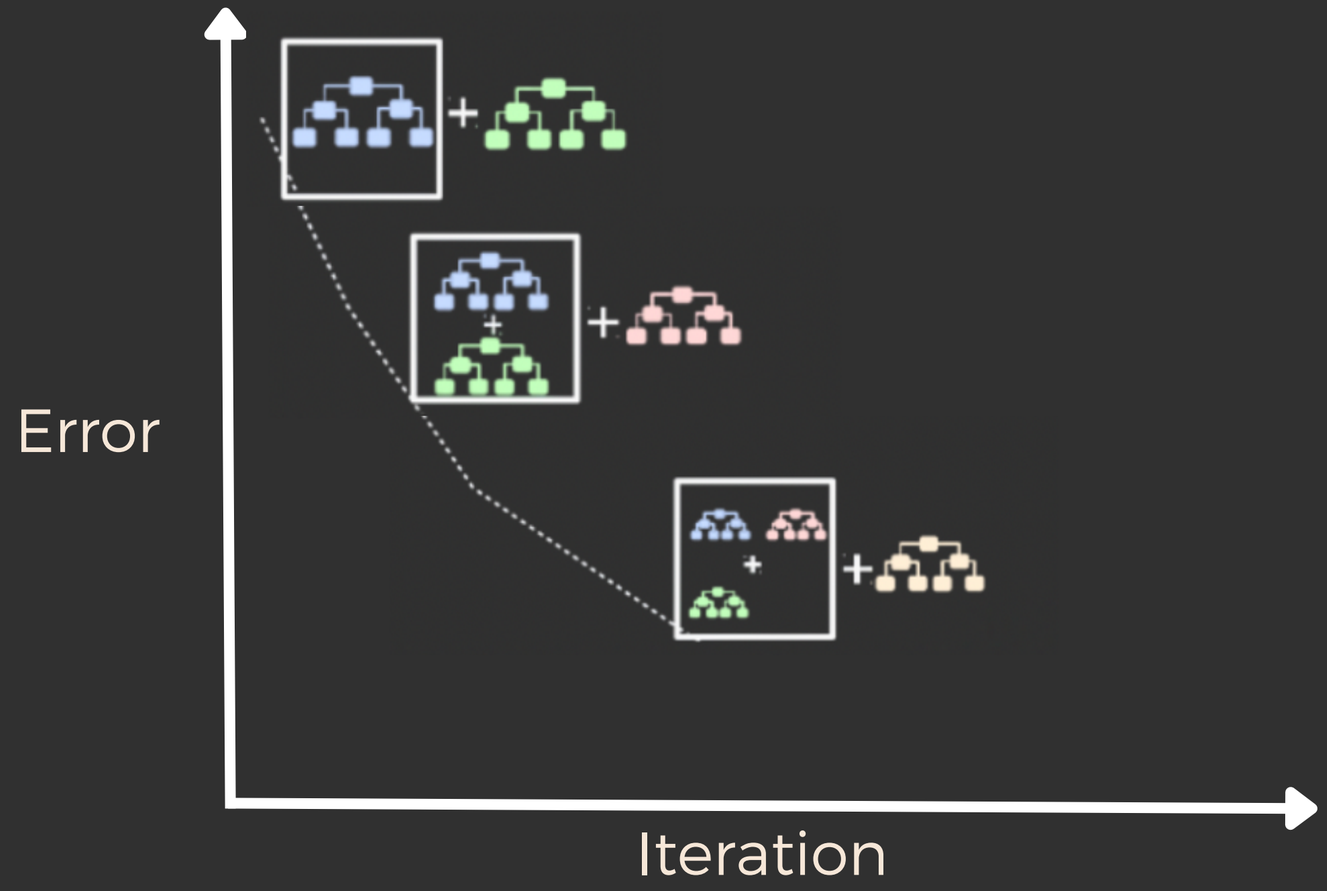
Gradient Boosting



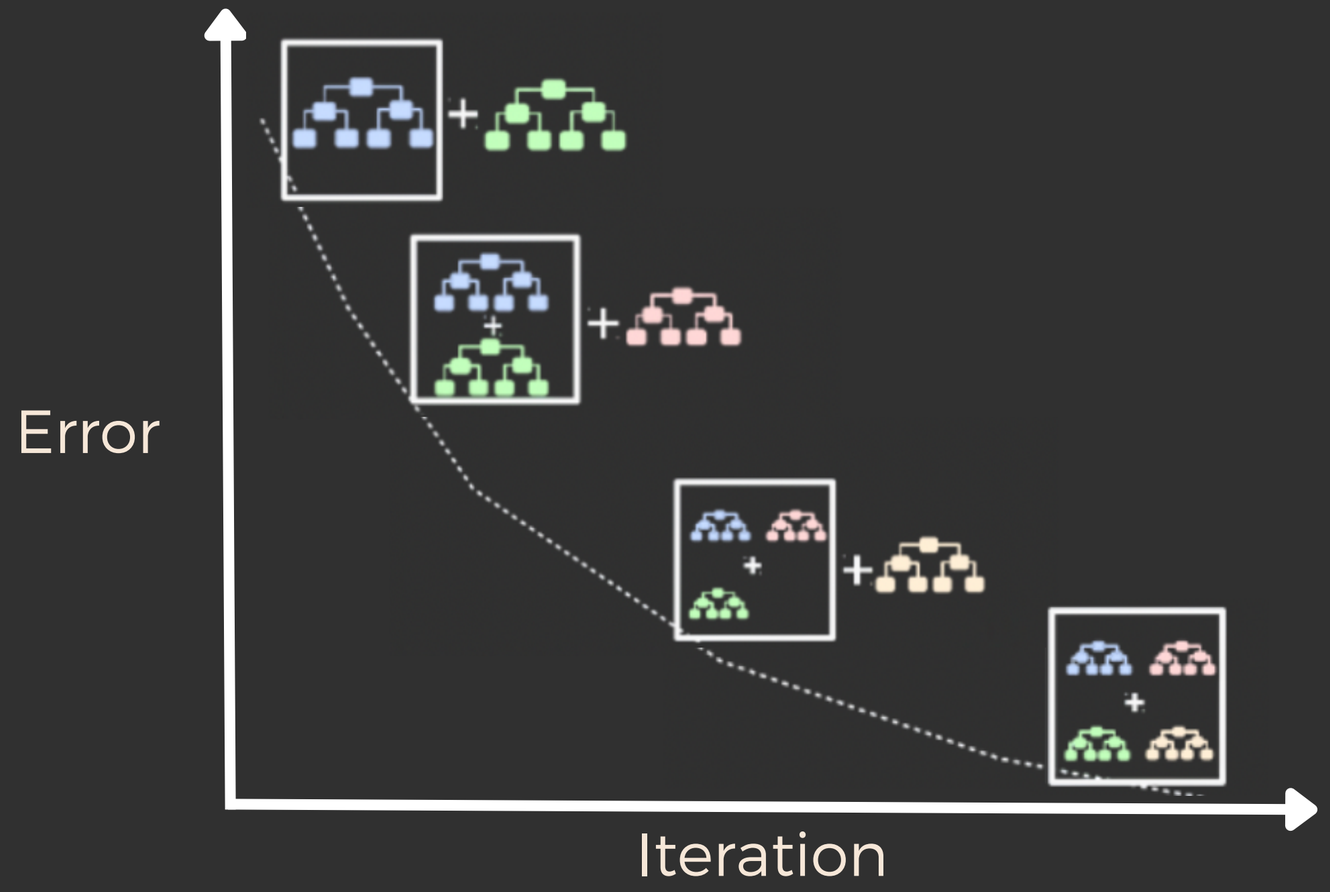
Gradient Boosting



Gradient Boosting



Gradient Boosting



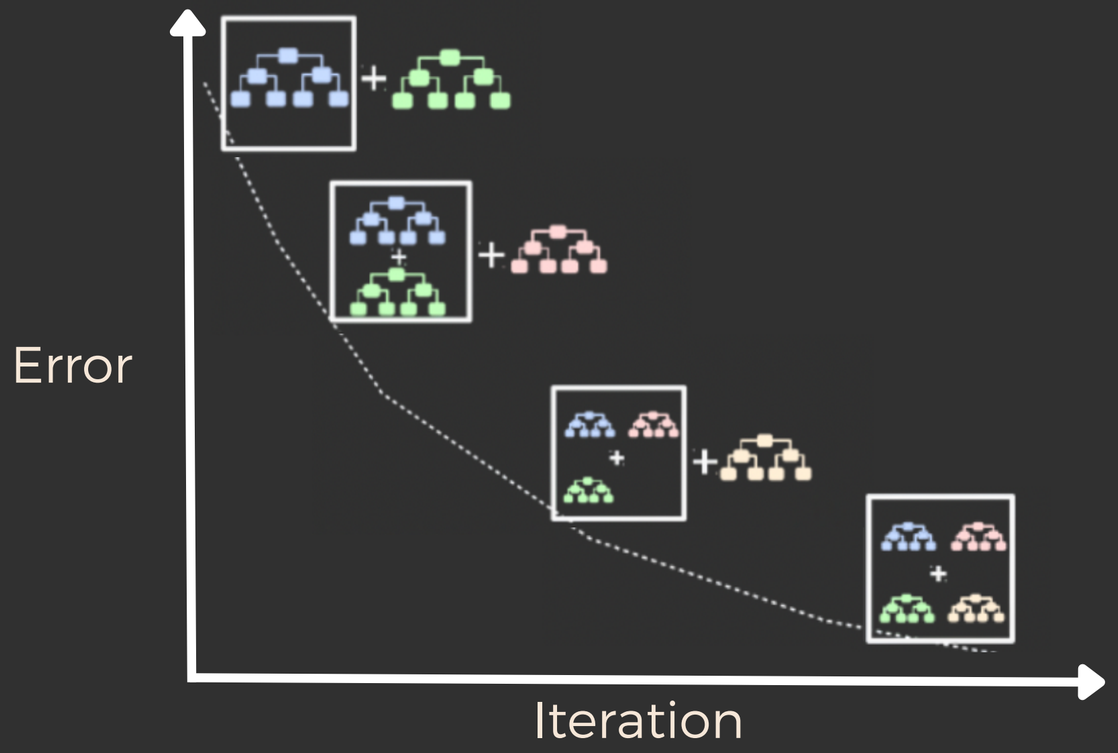
Gradient Boosting

Advantages

- Sequential iteration improves performance
- Able to capture complex patterns in data

Disadvantages

- Requires large processing power
- Prone to overfitting in presence of anomalies



MODEL PERFORMANCE

RMSE - 1.35 MAPE - 6.47%

XGBoost

XGBoost

Advantages

- Extension of Gradient boosting, more robust
- Controls over-fitting
- High performance on medium-structured data

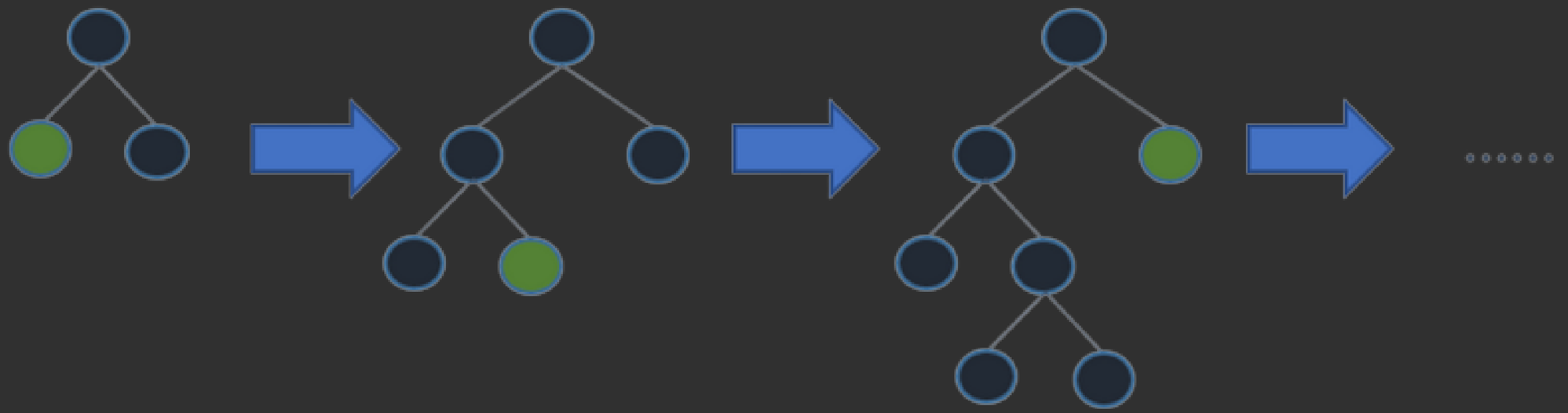
Disadvantages

- Requires large processing power
- Longer execution times

MODEL PERFORMANCE

RMSE - 1.41 MAPE - 6.66%

Light GBM

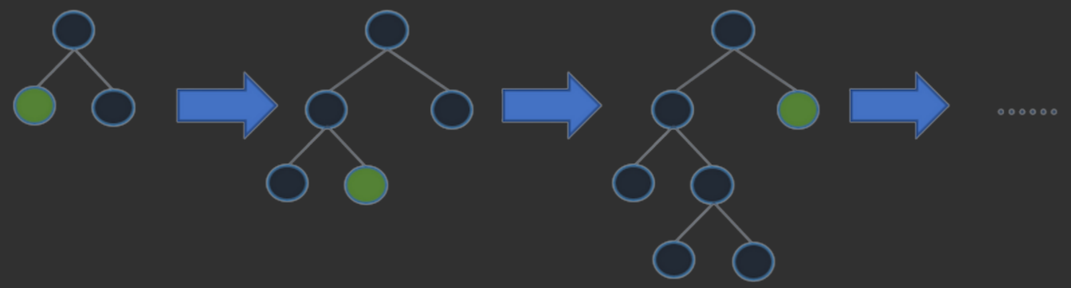


Leaf-wise Expansion

Light GBM

Advantages

- Lower Memory Usage
- Higher computational speeds
- Better training efficiency



Disadvantages

- Sensitive to over-fitting on smaller datasets

MODEL PERFORMANCE

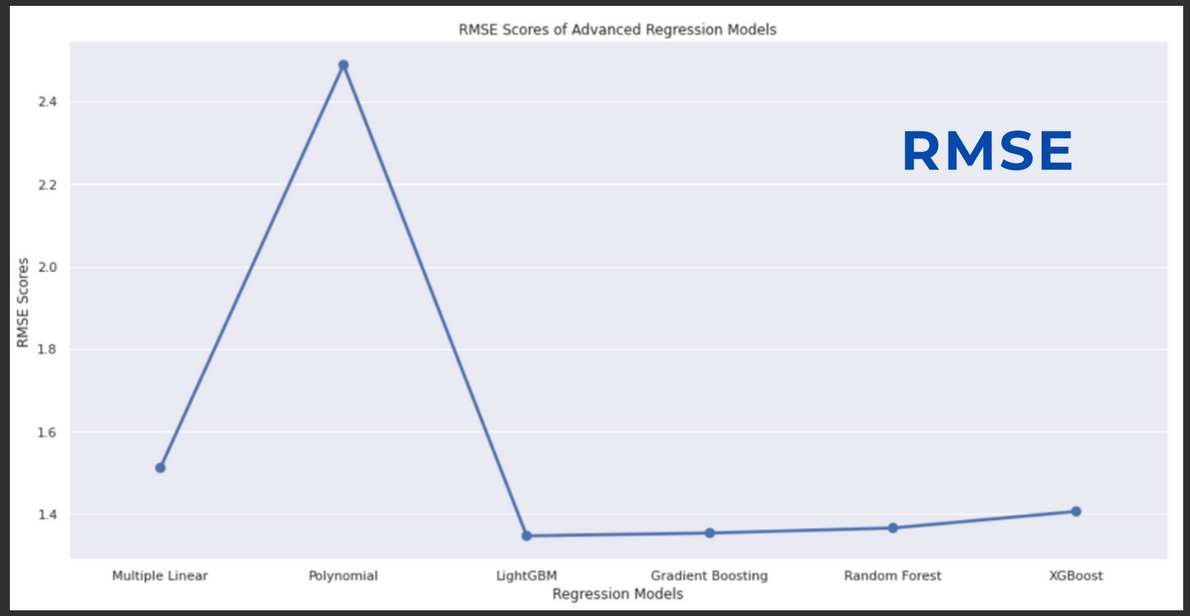
RMSE - 1.35 MAPE - 6.40%

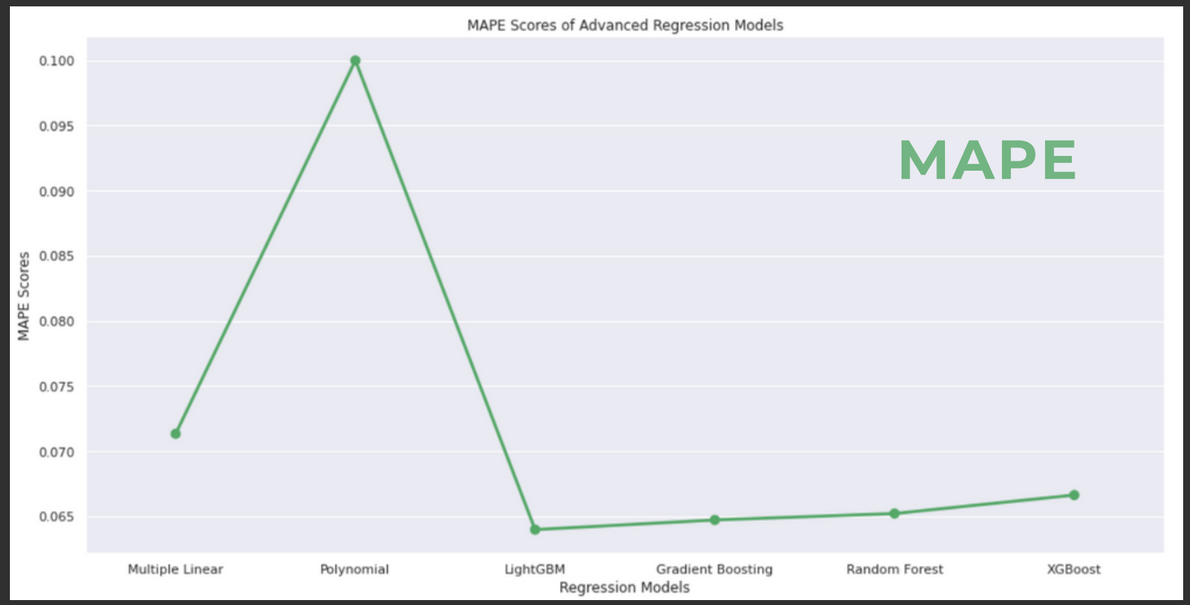
MACHINE LEARNING

ALGORITHMIC OPTIMISATION

Prediction Results

Models	RMSE	MAPE %
Linear Regression	1.51	7.13
Polynomial Regression	2.49	10.00
LightGBM	1.35	6.40
Gradient Boosting	1.35	6.47
Random Forest	1.36	6.52
XGBoost	1.41	6.66





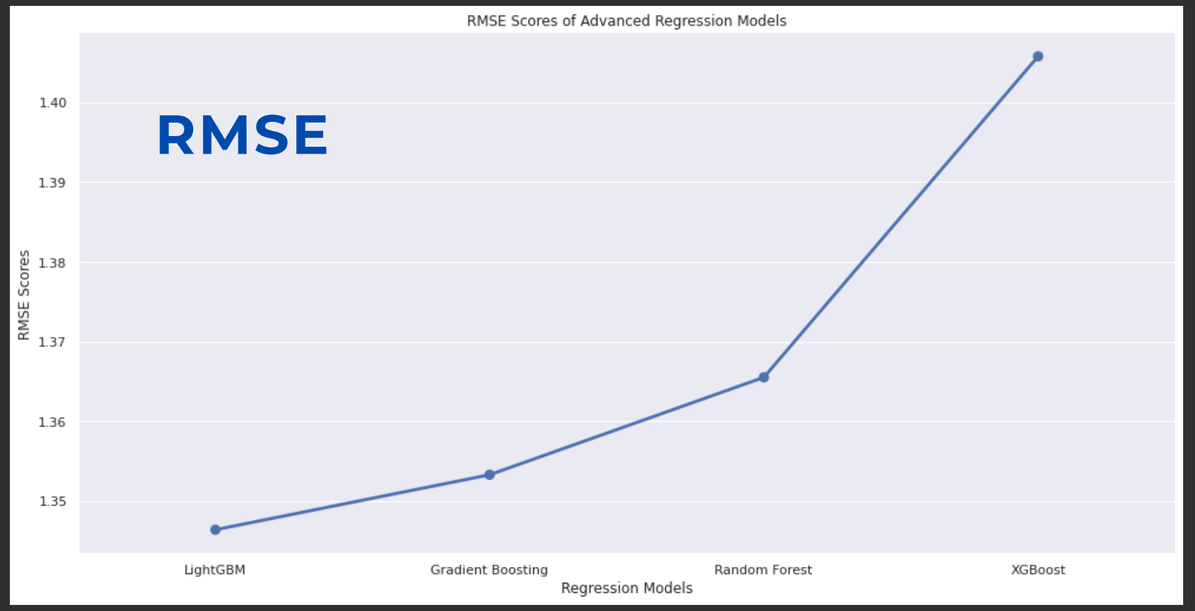
MACHINE LEARNING

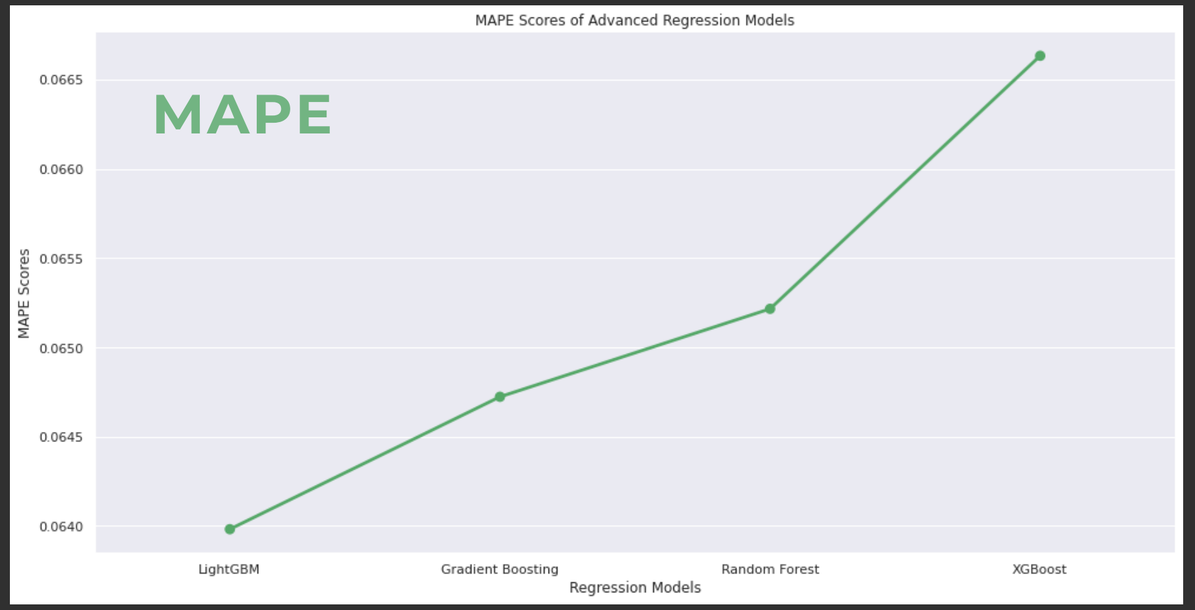
ALGORITHMIC OPTIMISATION

Prediction Results

Models	RMSE	MAPE %
LightGBM	1.346	6.398
Gradient Boosting	1.353	6.472
Random Forest	1.365	6.522
XGBoost	1.406	6.663

a closer look at top models

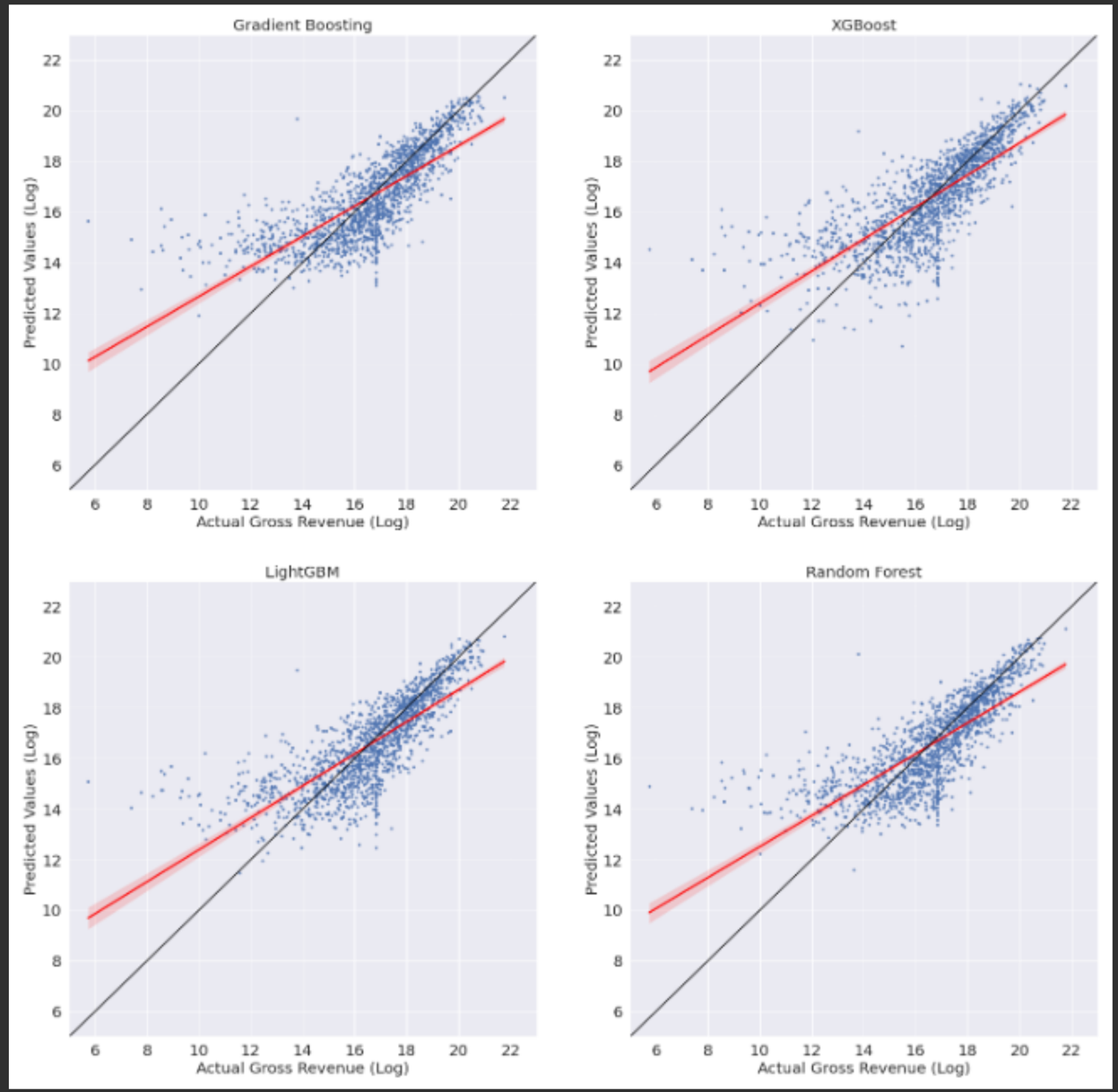




Visualising the Results

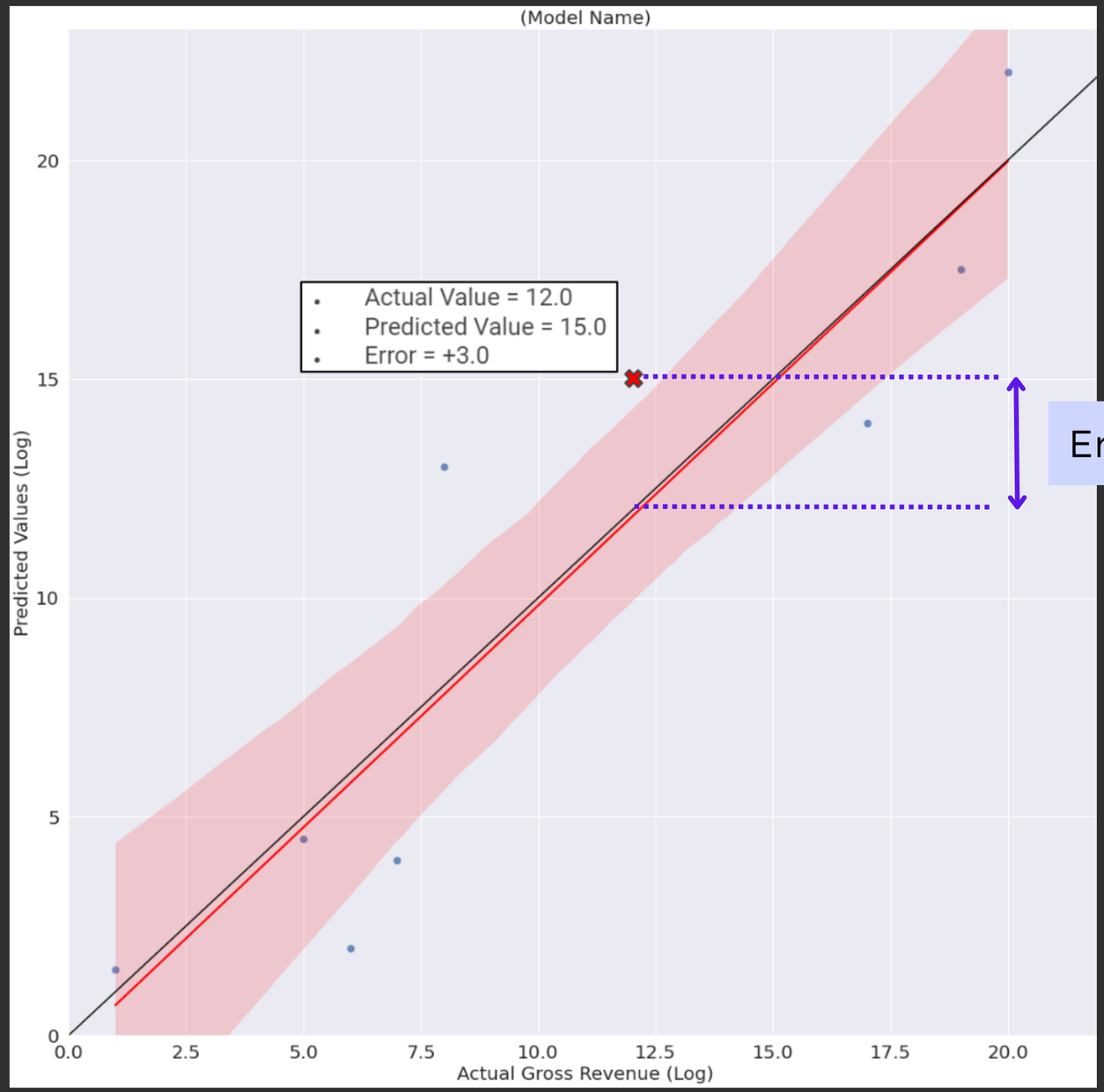
Regression plot:

Predicted vs Actual values



Visualising the Results

- Black line
  - Actual = Predicted
- Red line
  - Best-fit line of model
- Vertical distance = Error



Grid Search Optimisation

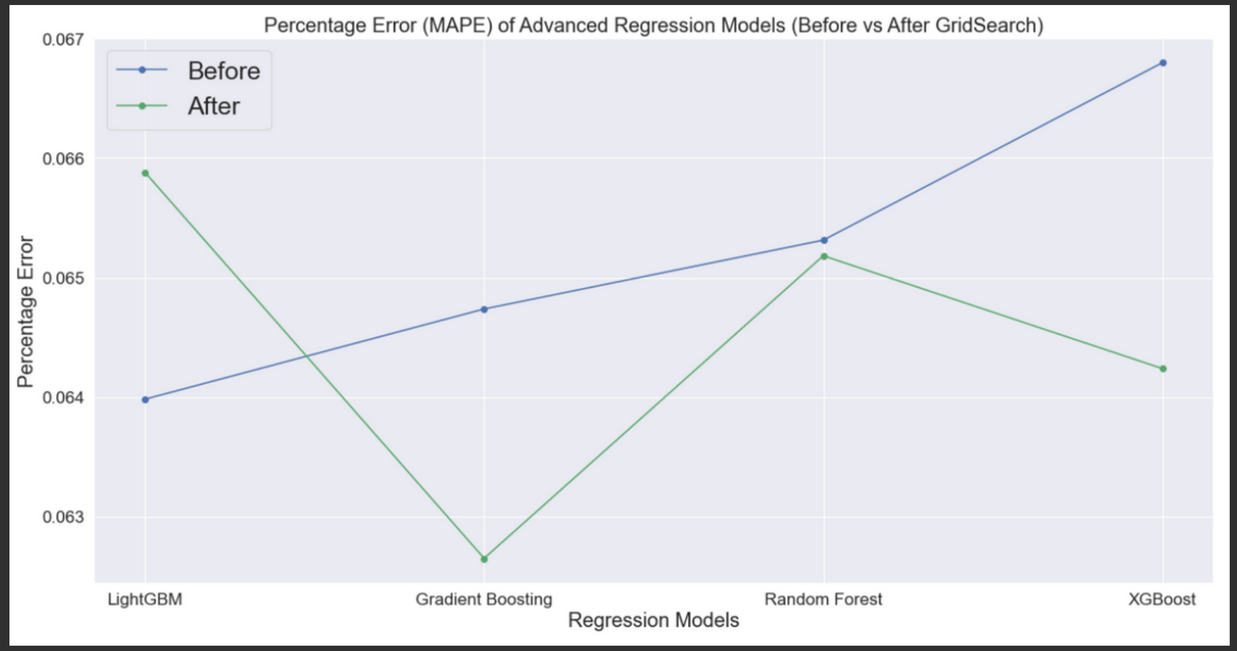
- Iterates through different combinations of specified hyperparameters and their values
- calculate performance per combination
- selects best combination of hyperparameters

Post-Hyperparameter Tuning

Using Grid-Search CV

- Reduced errors in predictions
- Observed anomaly for Light GBM

Model	Gradient Boosting	LightGBM	Random Forest	XGBoost
MAPE (Before)	6.48	6.40	6.51	6.66
MAPE (After)	6.28	6.59	6.31	6.42
Difference (+/-)	-0.20	+0.19	-0.20	-0.24

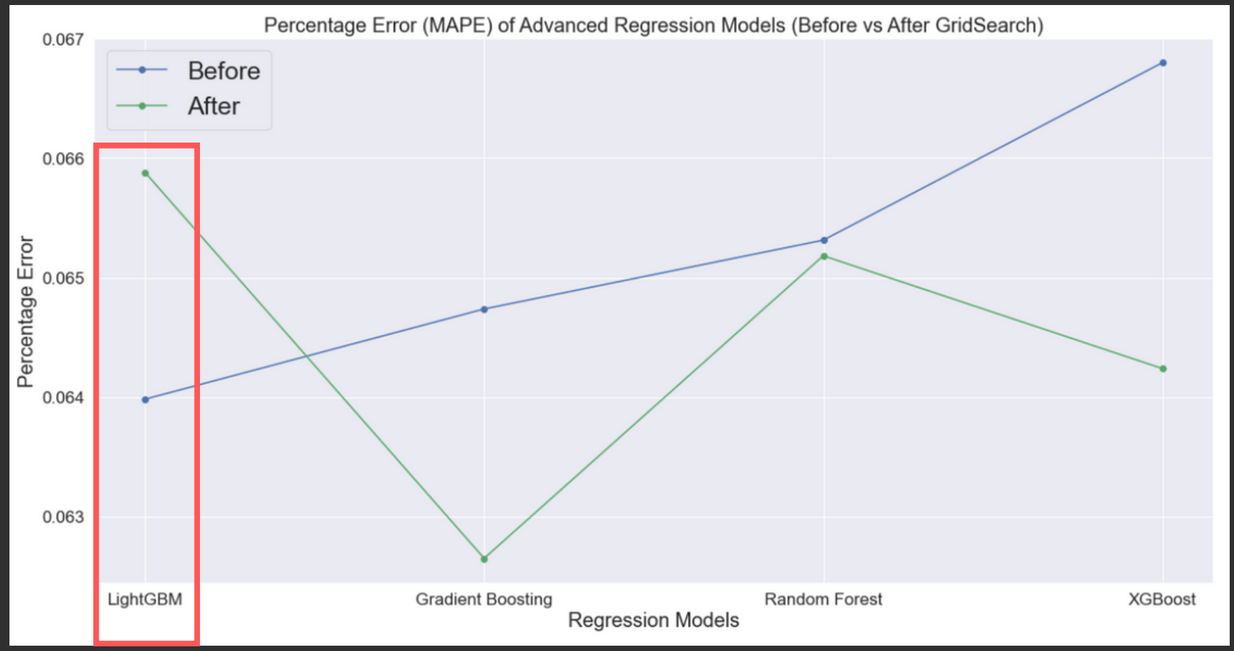


Post-Hyperparameter Tuning

Using Grid-Search CV

- Reduced errors in predictions
- Observed anomaly for Light GBM

Model	Gradient Boosting	LightGBM	Random Forest	XGBoost
MAPE (Before)	6.48	6.40	6.51	6.66
MAPE (After)	6.28	6.59	6.31	6.42
Difference (+/-)	-0.20	+0.19	-0.20	-0.24



Ensemble Learning

Stacking Regressor

- Combination of the top 4 models
- Using respective optimal hyperparameters

Base models: RFR, gradient boosting, XGBoost

Final estimator: Light GBM

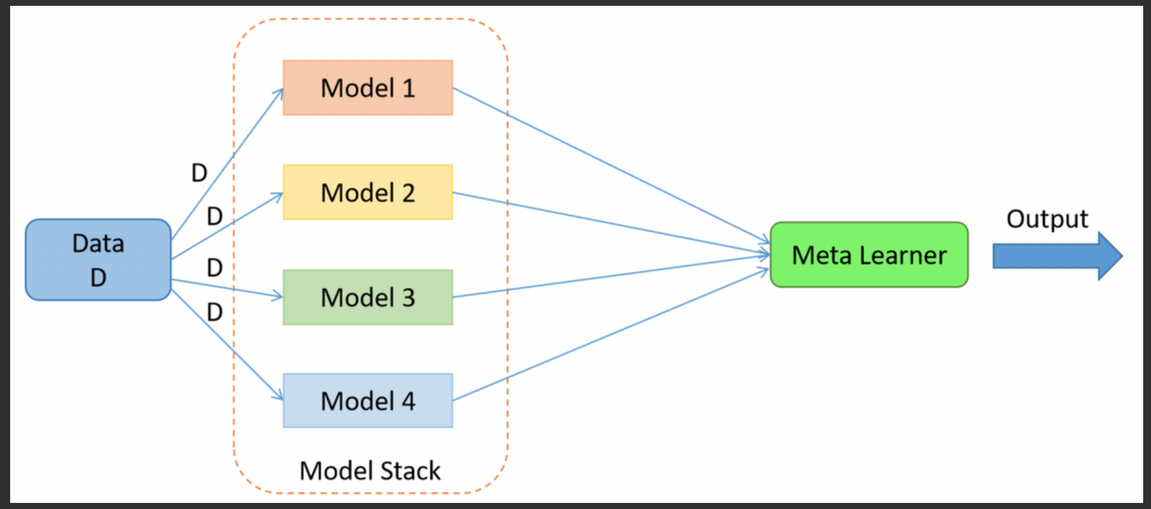
```
# Passing in optimal models as estimators
base_models = [
    ('gbr', gbr_optimal),
    ('rfr', rfr_optimal),
    ('xgbr',xgbr_optimal)
]


# Define the stacking ensemble model
stacking_regressor = StackingRegressor(estimators = base_models,
                                       final_estimator=lgbm_optimal)
```

Ensemble Learning

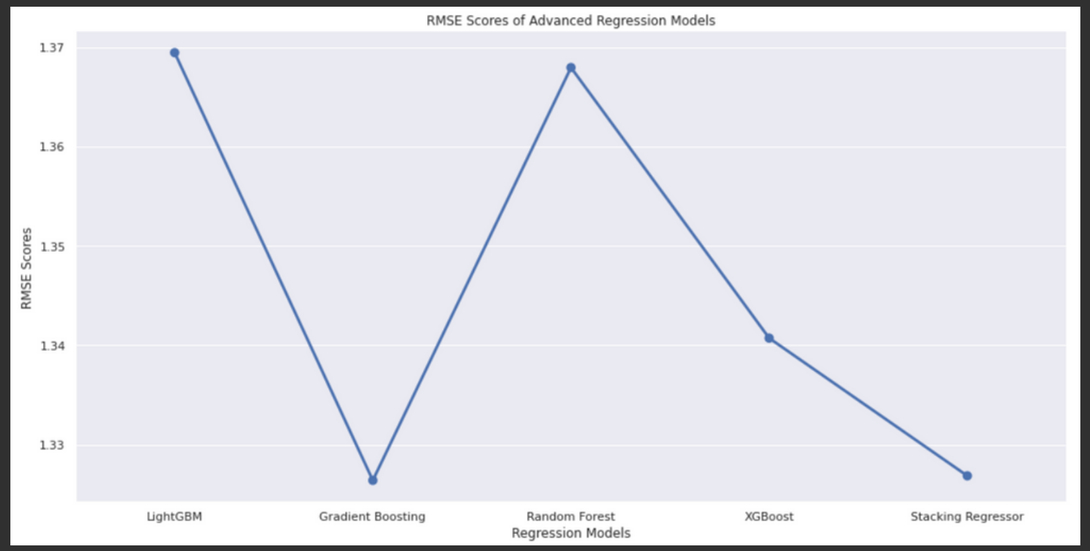
Stacking Regressor

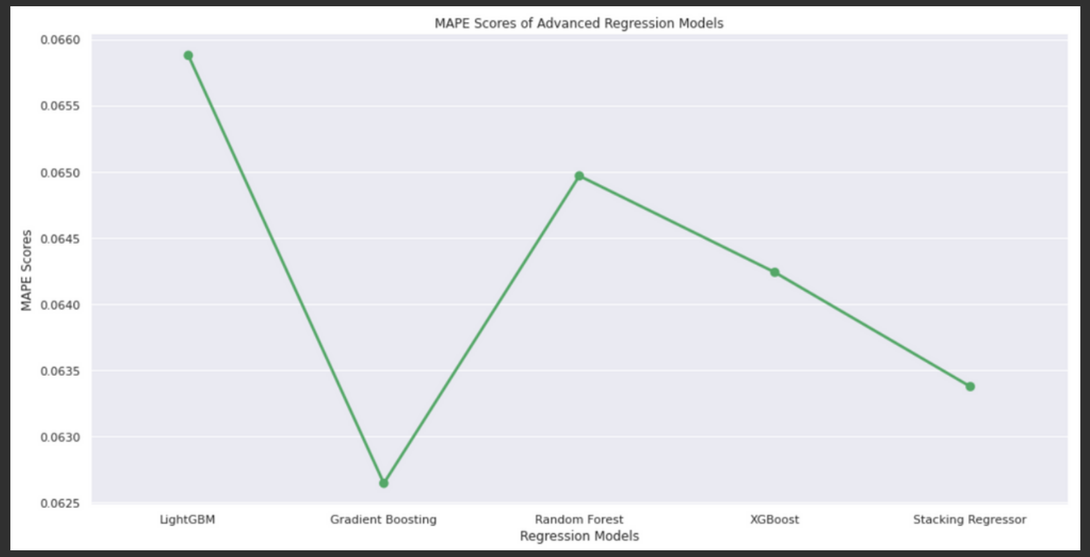
- Meta-Learning to best combine predictions from base algorithms
- Learns from outcome of other Machine Learning models
- Account for and **reduce** anomalies
- Better **reliability** and lower variance of results



MACHINE LEARNING

ALGORITHMIC OPTIMISATION





Post - Optimisation Results

Model	Gradient Boosting	LightGBM	Random Forest	XGBoost	Stacking
RMSE	1.326	1.369	1.368	1.341	1.327
MAPE %	6.265	6.588	6.497	6.424	6.338

Outcome

Recommendations

Techniques Learnt

Conclusion

Outcome

Prediction Accuracy of Gross Revenue

	name	gross	PredGrossRevenue	% Error
7280	Star Wars: Episode V - The Empire Strikes Back	20.1041	18.5917	7.5227
1100	The Hunter	14.3348	15.4086	7.4912
1271	Videodrome	14.5671	16.0538	10.2059

- Model is accurate
- Satisfactory performance - error approximately < 10%
- Components to create a successful movie can be determined

Future Recommendations

- Identify more possible variables
- Consider unconventional or unstructured data
  - Franchise popularity
  - Fanbase size & loyalty
- Diversify revenue or success metrics
  - Platform streaming revenue

Techniques Learnt

- Ordinal Encoding
- Random Forest
- Gradient Boosting
- XGBoost
- Light GBM
- Ensemble Learning - Stacking Regressor

Conclusion - Data Driven Insights

Movie success can be predicted. Additional factors that can impact revenue generated should be considered:

- Seasonal based genres
  - Christmas Comedies
  - Halloween Horrors
- Marketing and Advertising capabilities
- Target demographic trends

THANK YOU