



BOX OFFICE REVENUE PREDICTION USING LINEAR REGRESSION IN MACHINE LEARNING



A MINI PROJECT

- *PRESENTED BY*

S.THARUN 231801506

V.VARUN 231801185

SYNOPSIS

- Introduction
- What is Box Revenue ?
- Understanding Linear Regression
- Data Collection and Pre-Processing
- Feature Selection
- What is Genre ?
- Model Training and Evaluation
- Know about Random Forest Algorithm
- Challenges and Future Improvements
- Conclusion

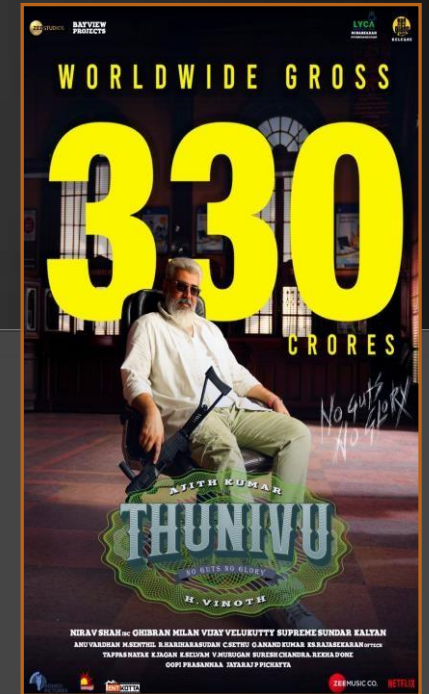
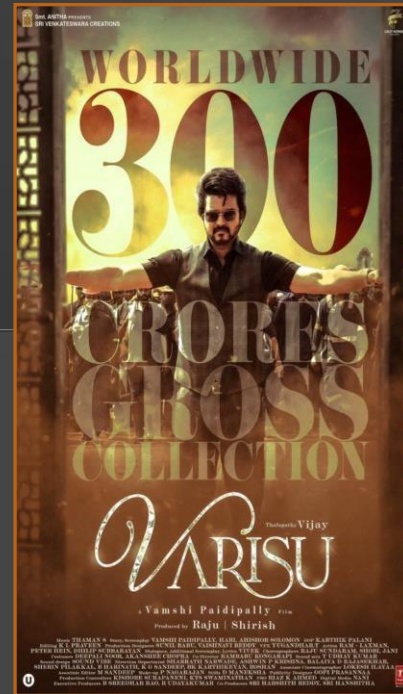
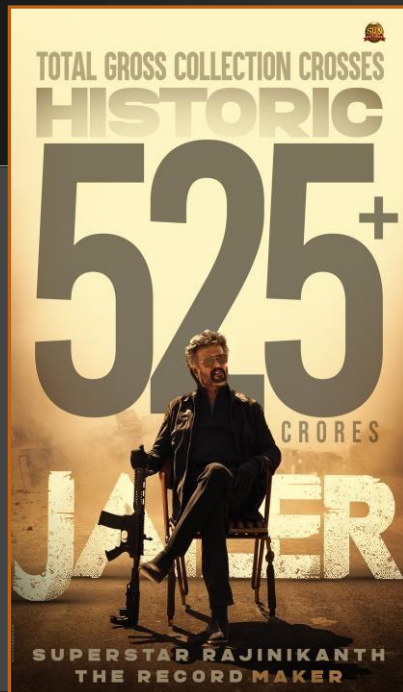
INTRODUCTION

- Welcome to the presentation on Box Office Revenue Prediction using Linear Regression in Machine Learning.
- In this presentation, we will explore how linear regression can be used to predict box office revenue for movies.
- We will discuss the importance of accurate revenue forecasts and the role of machine learning in improving predictions



WHAT IS BOX REVENUE ?

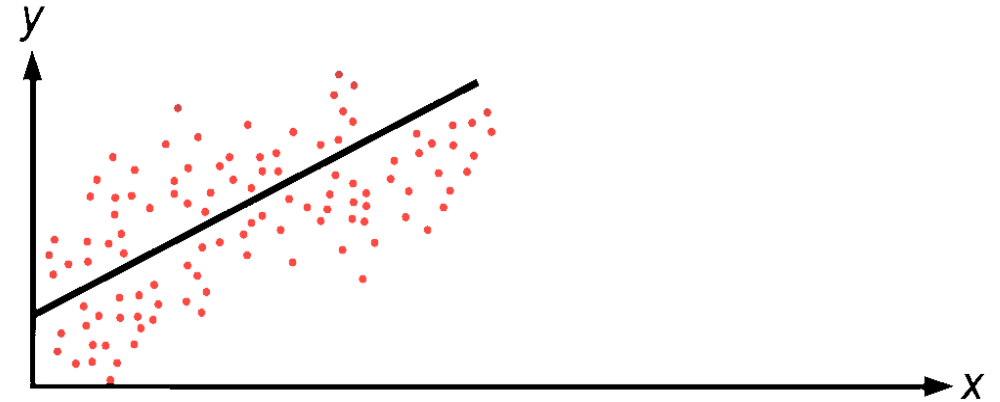
- Box office revenue refers to the total amount of money generated from ticket sales for a particular movie. It is a crucial metric in the film industry as it reflects the popularity and financial success of a movie.



UNDERSTANDING LINEAR REGRESSION

- Linear regression is a “Supervised Machine Learning” algorithm used to model the relationship between a dependent variable (box office revenue) and one or more independent variables (e.g., marketing budget, release date, genre).
- By fitting a line to the data, we can estimate future revenue based on these variables.

Linear Regression



DATA COLLECTION AND PRE-PROCESSING

- Collecting relevant data for box office revenue prediction involves gathering information on movies' attributes, marketing campaigns, and release details. This data needs to be cleaned, normalized, and transformed into suitable formats to ensure accurate predictions.

```
Loading Dataset

In [2]: df = pd.read_csv('boxoffice.csv',
                        encoding='latin-1')
df.head()

Out[2]:
```

	title	domestic_revenue	world_revenue	distributor	opening_revenue	opening_theaters	budget	MPAA	genres
0	Star Wars: Episode VIII - The Last Jedi	\$820,181,382	\$1,332,539,889	Walt Disney Studios Motion Pictures	\$220,009,584	4,232	\$317,000,000	PG-13	Action,Adventure,Fantasy,Sci-Fi
1	The Fate of the Furious	\$228,008,385	\$1,236,005,118	Universal Pictures	\$98,786,705	4,310	\$250,000,000	PG-13	Action,Adventure,Thriller
2	Wonder Woman	\$412,563,408	\$821,847,012	Warner Bros.	\$103,251,471	4,165	\$149,000,000	PG-13	Action,Adventure,Fantasy,Sci-Fi,War
3	Guardians of the Galaxy Vol. 2	\$389,813,101	\$663,756,051	Walt Disney Studios Motion Pictures	\$146,510,104	4,347	\$200,000,000	PG-13	Action,Adventure,Comedy,Sci-Fi
4	Beauty and the Beast	\$504,014,165	\$1,263,521,126	Walt Disney Studios Motion Pictures	\$174,750,616	4,210	\$160,000,000	PG	Family,Fantasy,Musical,Romance

```
In [3]: df.shape

Out[3]: (2694, 10)
```


WHAT IS GENRE IN MOVIES ?



Comedy



History



Sci-Fi



Romance



Thriller



Mystery



Drama



Horror



War



Action



Musical



Superhero



Animation



Road Movie



Fantasy



Movie

FEATURE SELECTION

- Choosing the right features is crucial for accurate predictions.
- We will explore various factors such as budget, cast, crew, genre, release date, and marketing spend.
- Through feature selection techniques, we can identify the most influential variables for our linear regression model.

MODEL TRAINING AND EVALUATION

- We will delve into the process of model training using linear regression. This involves splitting the data into training and testing sets, fitting the model.
- Evaluating its performance using metrics like mean squared error and R-squared.



jupyter box Last Checkpoint: 15 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run

Model Development

```
In [20]: features = df.drop(['title', 'domestic_revenue', 'fi'], axis=1)
         target = df['domestic_revenue'].values

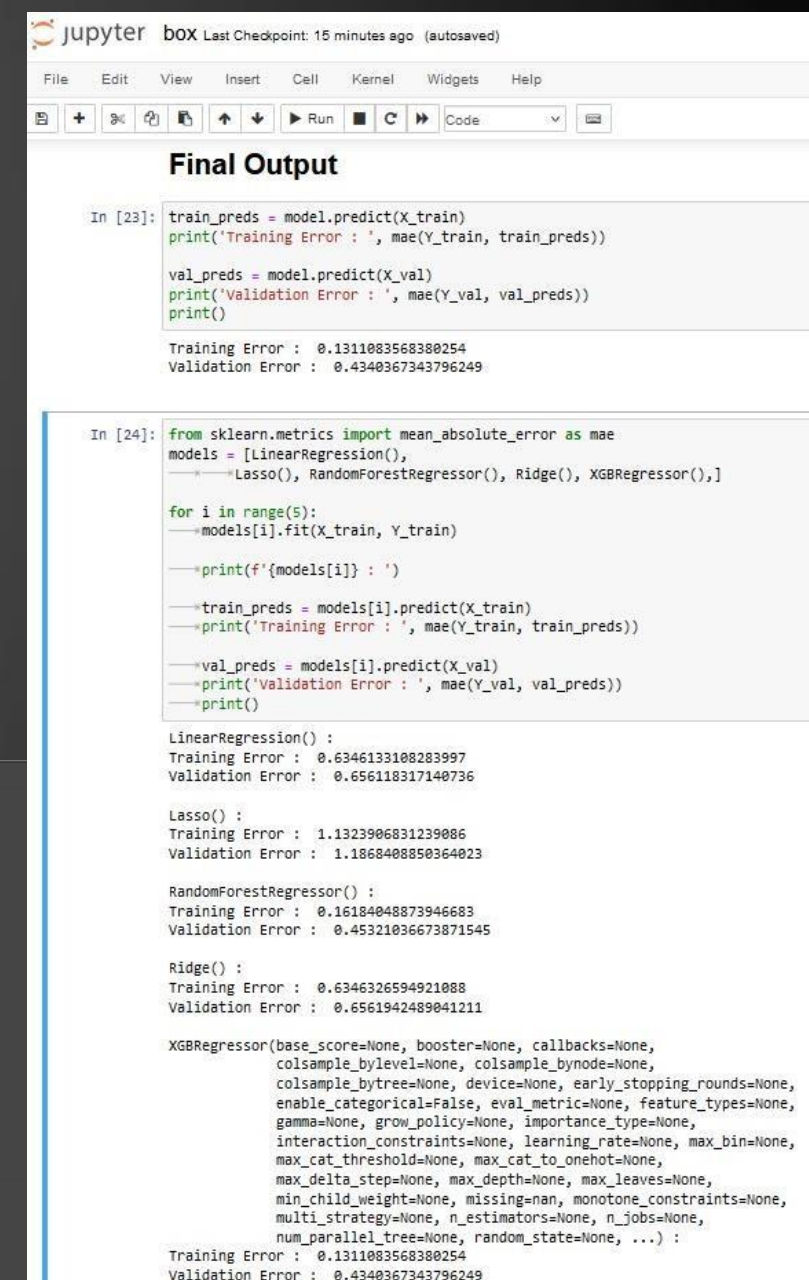
         X_train, X_val, \
         Y_train, Y_val = train_test_split(features, target,
         *-----*-----*-----*-----*-----*-----*-----*-----*-----*
         *-----*-----*-----*-----*-----*-----*-----*-----*-----*
         *-----*-----*-----*-----*-----*-----*-----*-----*-----*
         test_size=0.1,
         random_state=22)

         X_train.shape, X_val.shape

Out[20]: ((2144, 21), (239, 21))
```

RANDOM FOREST ALGORITHM

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.
- Used for both Classification and Regression problems in ML.
- It is based on the concept of **ensemble learning**, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*



The image shows a Jupyter Notebook interface with two code cells. The first cell, labeled 'In [23]:', contains code to train a model and print training and validation errors. The second cell, labeled 'In [24]:', contains code to compare five different models: LinearRegression, Lasso, RandomForestRegressor, Ridge, and XGBRegressor. The output of the first cell shows training and validation errors for a single model. The output of the second cell shows the same metrics for each of the five models, with RandomForestRegressor showing the lowest errors.

```
File Edit View Insert Cell Kernel Widgets Help
In [23]: train_preds = model.predict(X_train)
          print('Training Error : ', mae(Y_train, train_preds))

          val_preds = model.predict(X_val)
          print('Validation Error : ', mae(Y_val, val_preds))
          print()

          Training Error :  0.1311083568380254
          Validation Error :  0.4340367343796249

In [24]: from sklearn.metrics import mean_absolute_error as mae
          models = [LinearRegression(),
                    *Lasso(), RandomForestRegressor(), Ridge(), XGBRegressor(),]

          for i in range(5):
              models[i].fit(X_train, Y_train)

              print(f'{models[i]} : ')

              train_preds = models[i].predict(X_train)
              print('Training Error : ', mae(Y_train, train_preds))

              val_preds = models[i].predict(X_val)
              print('Validation Error : ', mae(Y_val, val_preds))
              print()

          LinearRegression() :
          Training Error :  0.6346133108283997
          Validation Error :  0.656118317140736

          Lasso() :
          Training Error :  1.1323906831239086
          Validation Error :  1.1868408850364023

          RandomForestRegressor() :
          Training Error :  0.16184048873946683
          Validation Error :  0.45321036673871545

          Ridge() :
          Training Error :  0.6346326594921088
          Validation Error :  0.6561942489041211

          XGBRegressor(base_score=None, booster=None, callbacks=None,
                        colsample_bylevel=None, colsample_bynode=None,
                        colsample_bytree=None, device=None, early_stopping_rounds=None,
                        enable_categorical=False, eval_metric=None, feature_types=None,
                        gamma=None, grow_policy=None, importance_type=None,
                        interaction_constraints=None, learning_rate=None, max_bin=None,
                        max_cat_threshold=None, max_cat_to_onehot=None,
                        max_delta_step=None, max_depth=None, max_leaves=None,
                        min_child_weight=None, missing=nan, monotone_constraints=None,
                        multi_strategy=None, n_estimators=None, n_jobs=None,
                        num_parallel_tree=None, random_state=None, ...) :
          Training Error :  0.1311083568380254
          Validation Error :  0.4340367343796249
```

CHALLENGES AND FUTURE IMPROVEMENTS

- While linear regression is a powerful tool, it has its limitations. We will discuss the challenges faced in forecasting box office revenue, such as changing consumer behavior and evolving movie industry trends.
- We will discuss scenarios where linear regression may not be suitable and explore alternative techniques such as polynomial regression and ensemble methods.
- Additionally, we will explore potential future directions, including the integration of more advanced machine learning techniques and big data analysis.



CONCLUSION

In conclusion, harnessing linear regression in machine learning can significantly improve box office revenue prediction accuracy.

By leveraging historical data and relevant movie attributes, we can make informed predictions that benefit the film industry.

The future holds immense potential for further advancements in revenue forecasting using advanced machine learning techniques

