CSE 343: Machine Learning Interim Report
Analyzing the influence of different parameters over the revenue of a movie

Ananya Singh 2019144 | Devika Mittal 2019463 | Manas Gupta 2019368 | Prashasti Agarwal 2019075

1. Abstract

The film industry is a billion-dollar industry which has been entertaining the masses for decades. Analysing success of a movie is an exciting topic to gain insights in human psychology and behavior for anthropologists. Our studies may also prove helpful to respective professionals in increasing movie returns, by strategic content creation and advertisement that gains more traction. This may be beneficial in predicting box office revenues, affecting the show business industry. Link to Github repository.

2. Introduction

The aim of our project was to predict revenue of a movie and study how different parameters like genre, title, description keywords, run-time, etc, affect its success. Through several machine learning techniques, we have tried to verify our preconceived notions about various factors which might be important in determining revenue. We have tried several hyperparameter tuned models to finally reach the one that best models our data in predicting revenues.

3. Literature Survey

1. This research outlines how the box office performance of a movie is an amalgamation of various movie features. They used Linear, Logistic regression, Simple Logistic, Multilayer Perceptron, J48, Naive Bayes, PART classifiers to implement and test their data. Accuracies: J48 (82.42 %), Linear (84.34 %), Neural Network (79.07%).[1]
2. They use Logistic Regression, Support Vector Machine and Multi Layer Perceptron to predict the box office revenue of a movie based on the data available before the release of the movie. Maximum Accuracy: Multi Layer Perceptron (85.31 %). [2]
3. This study examines the role of genre, holiday release, production cost, critics rating and sequel in determining the revenue generated by a film. Using a Multiple Regression Model, they conducted a simple correlation test of the correlation coefficients between the global box office revenue and the explanatory variables. [3]

4. Dataset Used

We used Kaggle's The Movies Dataset for our analysis. This is a labelled dataset containing entries of about 45,000 movies listed at IMDB. We have used the following 3 of the 7 files in the dataset for our analysis:

- movies_metadata: contains 45,466 entries each containing 24 features namely - adult, belongs_to_collection, budget, genres, homepage, id, imdb_id, original_language, original_title, overview, popularity, poster_path, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, video, vote_average, vote_count
- keywords: contains movie IDs and the keywords used to describe their plot.
- credits: contains movie IDs with the cast as well as crew involved in the movie.

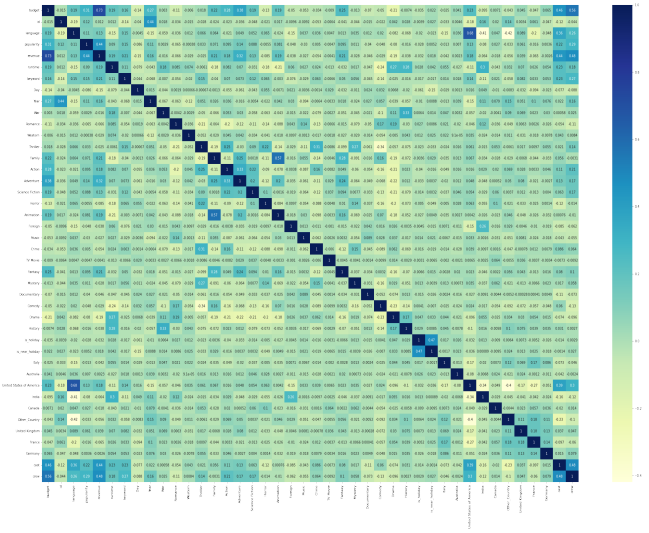
4.1. Feature Selection

☐ The following features were removed from the dataset:

- adult, video, status: had boolean values that were highly biased.
- homepage poster_path: had links to images of which around 50% of the entries were null.
- overview, tagline: had long descriptive texts.
- original_title, title: These were uncorrelated to revenue.
- spoken_languages, production_companies, belongs_to_collection: had too many unique categories and sparse values.
- vote_average, vote_count: statistics of votes are determined only after a movie is released which isn't helpful in predicting revenue.

☐ Using Correlation Matrix Fields with correlation values ≥ 0.6 were dropped with the help of Correlation Matrix [Figure 1.]

Figure 1. Correlation Matrix of all the features



4.2. Data Pre-processing

- Datatype processing: All attributes of the data were converted into strings, integers, floats, dictionaries and lists from the default object datatype.
- Unfilled values removed: Data containing null, empty strings or empty lists along with 0 revenues were identified and removed.
- Further steps are illustrated in [Figure 2].

4.3. Dataset Visualisation

The following observations are prominent in Figure 3

(a) The pi-chart helps in recognizing countries that have produced significant number of movies. We observed most movies were produced by 8 countries.

(b) English dominates distribution of language with roughly 89.3% of movies in English.

(c) On an average, animation movie makes highest average revenue while foreign movie make least revenue.

The following observations are prominent in Figure 4

(a) Budget of a movie does not independently determine revenue as evident from the movies with similar budget generating various revenues.

(b) Releases made early in the week(from Monday) tend to earn more revenue.

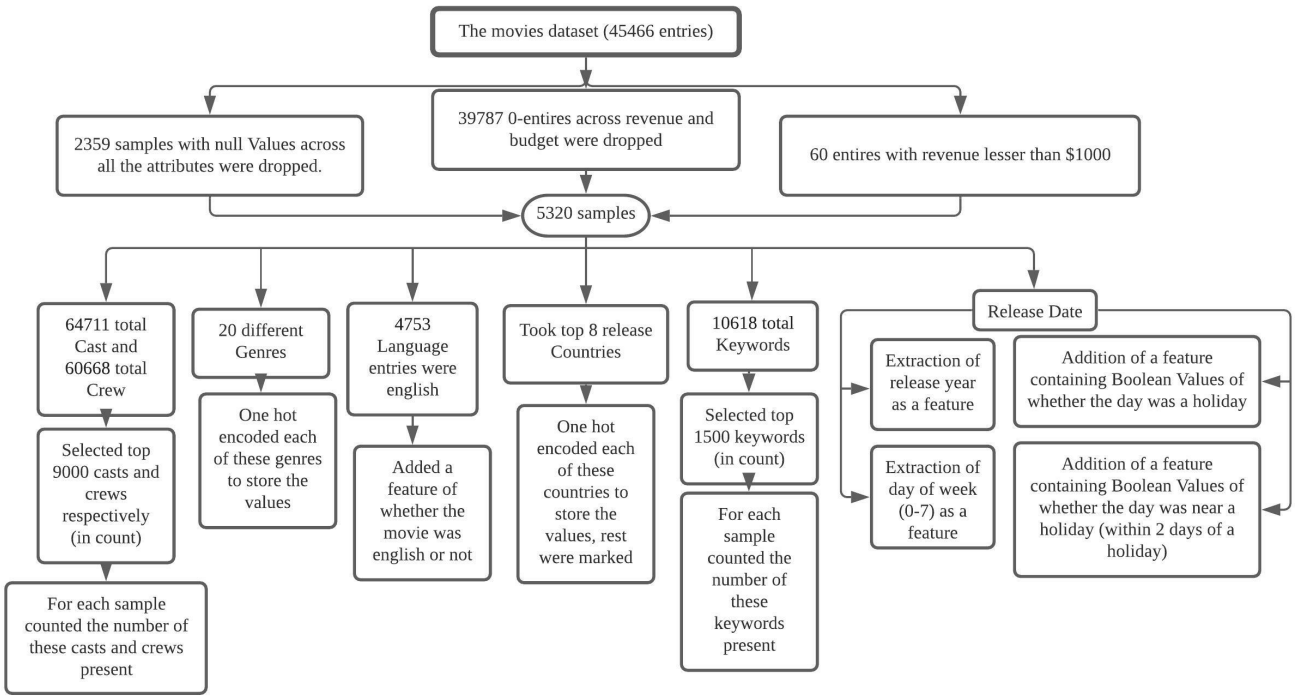
(c) Movies released during holidays after 2000s, tend to make more revenue.

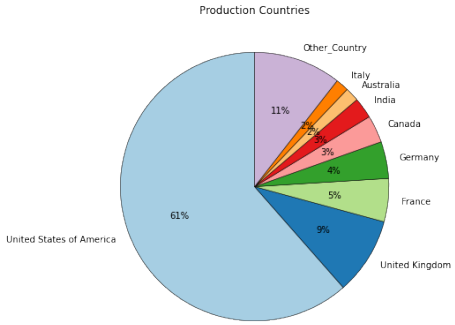
5. Methodology

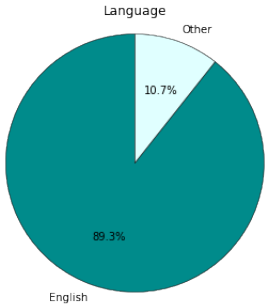
5.1. Evaluation metric used

The R^2 score measures how well a model fits our dataset by comparing it to the mean of y. R^2 values closer to 1, depict better prediction. A negative R^2 value implies, the

Figure 2. Processing genre, cast, crew, keywords, language, release country and release date







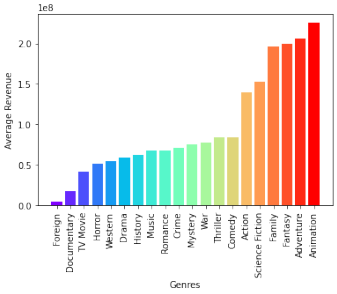
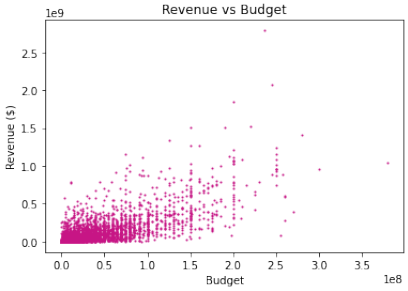
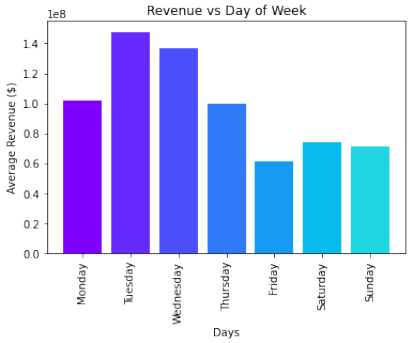


Figure 3. (a) Distribution of production countries (b) Distribution of languages (c) Average revenue v/s Genres





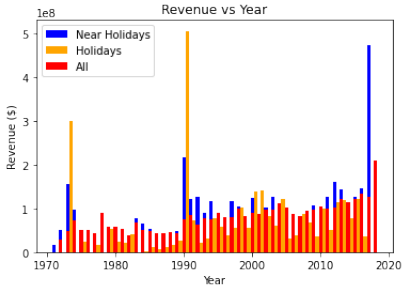


Figure 4. (a) Revenue v/s budget (b) Average revenue on different weekdays (c) Average revenue v/s year

model is performing worse than a constant mean prediction.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y_i})^2}$$

5.2. Model Details

Several regression models and techniques were employed to best fit the data. SKLearn's GridSearchCV was used to find best hyperparameters suitable.

1. **Linear Regression**
   After training a linear regression model, we plotted the coefficients found to interpret the effect on revenue.
2. **Polynomial Regression**
   A degree 2 polynomial regression model was trained after doing a dimension reduction through Principal Component Analysis, SK-Learn's PCA(0.9).
3. **Regularization Techniques**
   Following regularization methods were also employed:
   - Lasso Regression: with $\alpha = 0.001$.
   - Ridge Regression: with $\alpha = 0.1$
4. **Decision Tree Regression**
   Hyperparameters used: max_depth = 7 and min_sample_leaf = 18
5. **Ensemble Techniques**
   - Bagging on decision trees - n_estimators = 150
   - XGBoost on decision trees - Hyperparameters used: max_depth = 4, learning_rate = 0.01, n_estimators = 700
   - AdaBoost on decision trees- Hyperparameters used: learning_rate = 0.01, n_estimators = 300
   - Random forest regressor - Hyperparameters used: max_depth = 10, min_sample_leaf = 4, n_estimators = 800. Then we plotted the effect of increasing the depth on R^2.
   - AdaBoost on random forest - Hyperparameters used: learning_rate = 0.001, max_depth = 10, n_estimators = 150, min_samples_leaf = 4, n_estimators(2) = 100
6. **Neural network**
   Hyperparameters used: hidden_layer_sizes = (100, 50), solver = lbfgs, activation = relu, learning_rate = 0.001, alpha (L2 regularization coefficient) = 1, max_iters = 100
7. **Kernel Regression**
   Hyperparameters used: alpha = 0.01, kernel = laplacian
8. **Stacking**
   Stacking XG Boosted Decision Trees and Kernel Re-

gression as Base Estimators with Linear Regression as final estimator to combine the base models.
Hyperparameters:

- XGboosted Decision Tree): XGBRegressor(max_depth = 4, learning_rate = 0.01, objective = reg:squarederror, n_estimators = 700
- Kernel Regression : KernelRidge (alpha = 0.01, kernel = laplacian)
- Final Model = Linear Regression()

9. **K Nearest Neighbours Regression (KNN)**
Hyperparameters used: n_neighbors = 8, weights = uniform

6. Results and analysis

6.1. Model performance

Linear Regression along with regularization didn't perform well. The model had significant bias and hence we had to resort to non-linear techniques to model our data. Using Polynomial regression of degrees 2 and 3 still had a poor performance implying the data is more complex than the lower degree polynomials.
To tackle this non-linearity, we performed a Kernel Regression i.e. mapped our data to some other higher dimensional space and used a linear regression there. Use of kernels gave us a significant improvement in our previous R2 score.

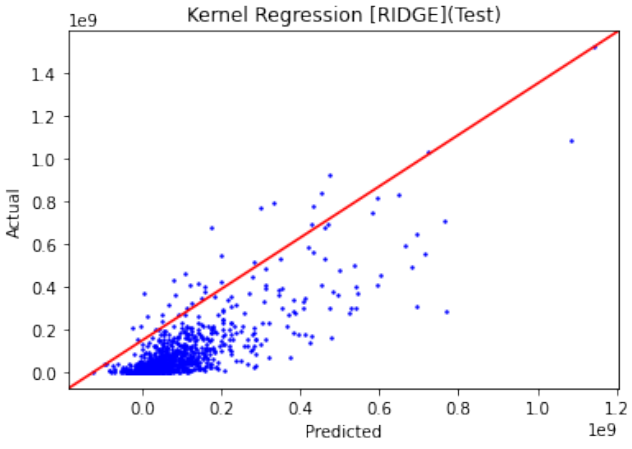


Figure 5. Actual vs Predicted Plot of Test Set using KRR

We then suspected if movies with "similar" features have similar revenues and hence we made use of the K-Nearest Neighbour algorithm. The model gave us decent results. We then decided to use tree techniques like Decision trees and Ensemble methods. Training a decision tree resulted also had a fair score, however it had high bias and variance. To tackle this, we decided to use Random Forests which gave us a lower bias and a lower variance.

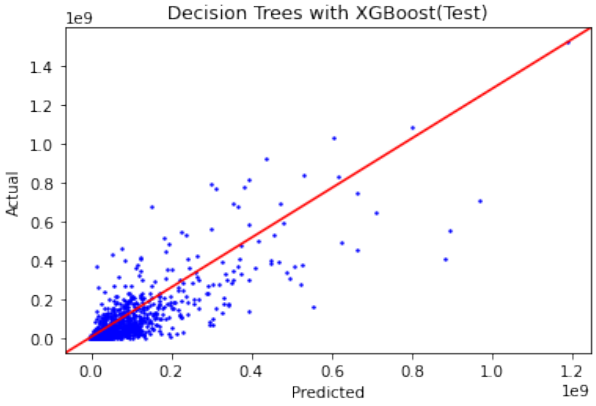


Figure 6. Actual vs Predicted Plot on Test Set using XGBoosted Decision trees

To further improve our model, especially for bias, we used boosting techniques like XGBoost and AdaBoost, which showed improvement in both bias and variance. After this, boosting on a random forest was tried which also showed improvements in test score. After the previous methods, we suspected that our dataset has a very complex distribution and to tackle that we resorted to using Neural Networks to have a model of very high complexity. The network didn't perform too well as we hoped, as it was easily overfitting the training set.
We wanted to test if both of our highest predictive scoring models weren't able to fit on the entire data and leave different aspects of it. Thus in order to harness the strong aspects of each of our best performing models we used stacking. Stacking provided a slightly better R2 score than each of the individual base models.

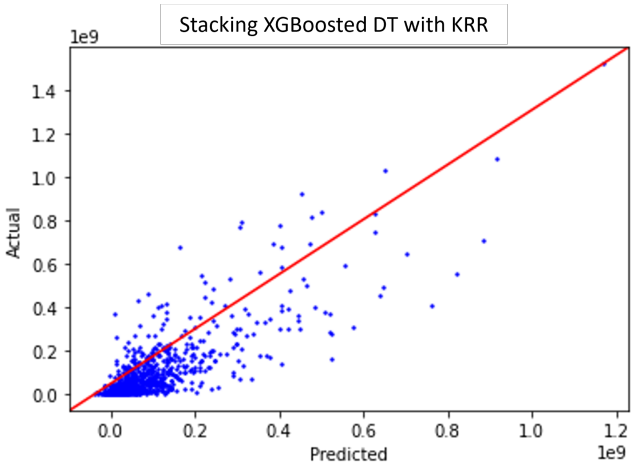


Figure 7. Actual vs Predicted Plot on Test Set using Stacking

Our best performance on testing data was obtained using Stacking of XGBoosted Decision tress with Kernel Ridge Regression with Linear Regression as the final model to combine them.

Figure 8. Result across various models

Model Name	Train Score	Test Score	CV Score
Linear Regression (LR)	0.630	0.559	0.613
Lasso Regularization	0.630	0.559	0.613
Ridge Regularization	0.630	0.558	0.616
Polynomial Regression + PCA	0.553	0.403	0.450
Decision Tree (DT)	0.715	0.584	0.599
Random Forest (RF)	0.847	0.653	0.672
Bagging on Decision Tree	0.720	0.625	0.639
XGBoost on Decision Tree	0.850	0.653	0.676
ADABoost on Decision Tree	0.687	0.536	0.612
ADABoost on Random Forest	0.844	0.654	0.675
Neural Networks	0.739	0.561	0.607
Kernel Ridge Regression (KRR)	0.843	0.630	0.645
Stacking (XG boosted DT and KRR) with LR as final estimator	0.856	0.668	0.680
KNN	0.563	0.287	0.398

6.2. Analysing the features

6.2.1 Coefficient analysis

From the plots of coefficients obtained from linear regression against features, we observe the importance of the features in predicting revenue.

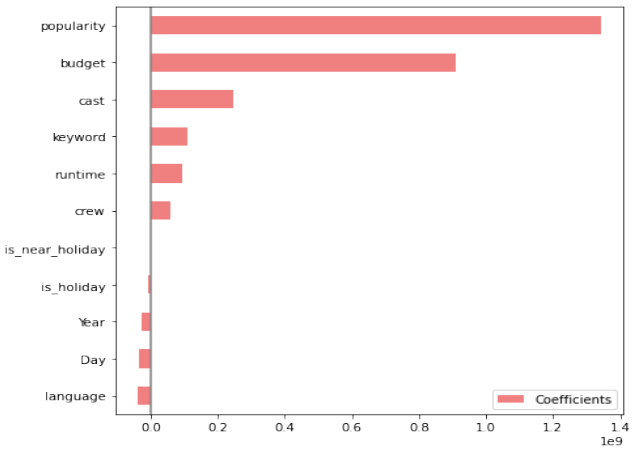


Figure 9. Coefficients of various features after Linear Regression

6.2.2 Feature importance

We used feature importance to analyze decision trees and random forest. It is calculated as the decrease in node impurity weighted by the probability of reaching that node. The higher the value, the more important is that feature.

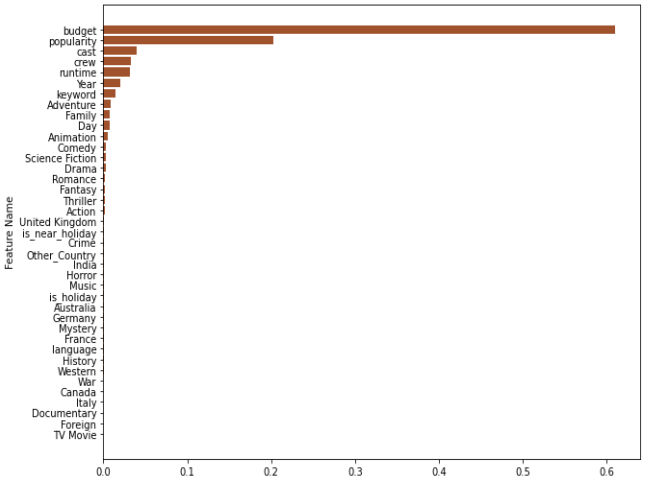


Figure 10. Feature importance using Random Forest

6.2.3 Permutation importance

Permutation importance determines which features are important by systematically removing them (or more accurately, replacing them with random noise) and measuring how this affects the model's performance. This is helpful for models with low interpretability like neural networks. We took the average of the results from permutation importance for the models MLP, KRR, and Stacking.

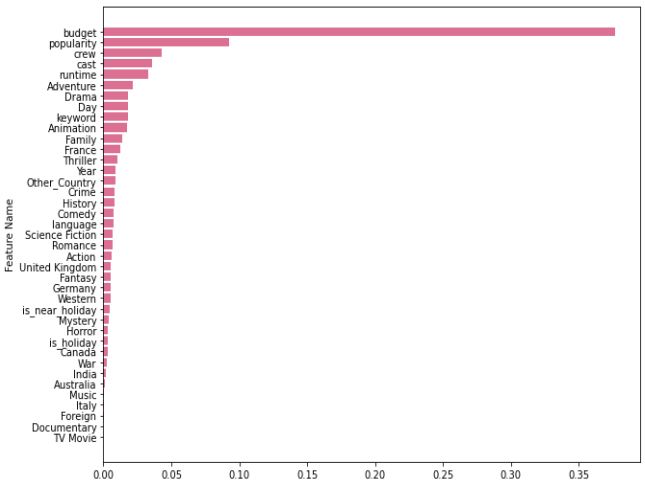


Figure 11. Permutation Importance (average over MLP, KRR, Stacking)

6.2.4 Influence on Revenue by varying Budget and Popularity

As the budget of a movie increases with other features as constant, revenue initially increases. After a certain budget, revenue remains more or less similar. Hence in effect,

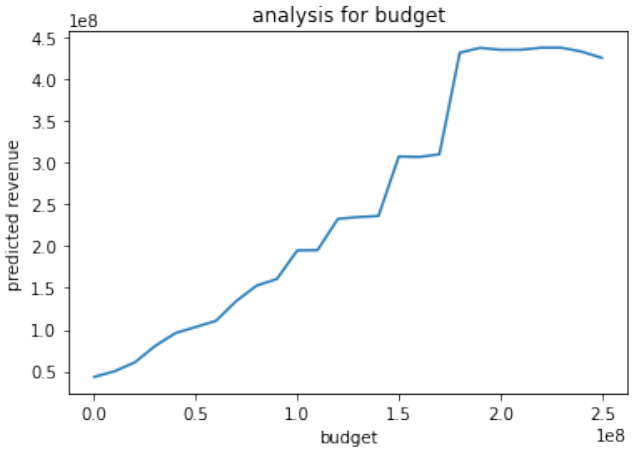


Figure 12. Variation of Revenue with Budget

spending money on an excessive budget is likely to not increase the revenue of the movie.

The curve of popularity initially witnesses a sharp increase in revenue with increasing popularity. However it reaches a peak and notices a decline in it's revenue after which it remains stagnant.

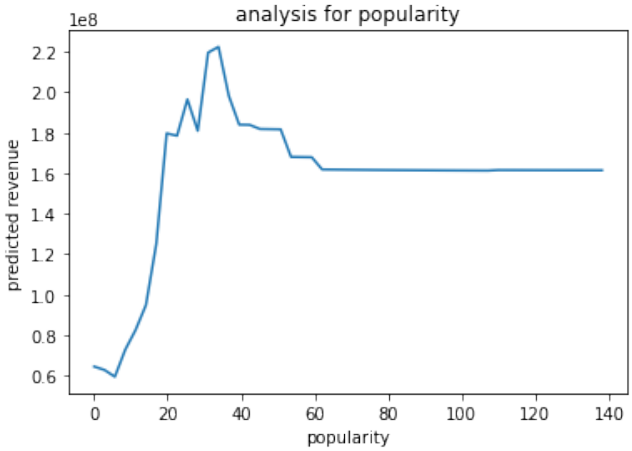


Figure 13. Variation of Revenue with Popularity

6.2.5 Final Analysis

Through feature importance and permutation importance, we observed that budget influences revenue the most. Popularity, crew, and cast are the top features after budget, which significantly influence revenue. Runtime, day of release, and movie description keywords were also significant factors. Looking at the genres, the most revenue generating one is adventure followed by drama, animation, family, and thriller. War, history, and musicals on the other hand, had a lesser influence on revenue. Surprisingly, a movie being released on or near a holiday did not have much impact.

7. Conclusion

Our project so far has allowed us to get hands-on experience in data collection, pre-processing data and analyzing them using various machine learning models ranging from linear and polynomial regressors to decision trees, random forests and neural networks, K nearest neighbours. We have also learned about different techniques like feature selection, regularization, bagging, boosting and stacking.

Our analysis shows that our dataset has several complex features that simple linear models failed to capture. Other models such as Support Vector Regression failed while KNN did not perform too well. We were able to model our data using Random Forests and other ensemble methods over Decision Trees. Kernel Regression also had a good performance. Lastly we combined the best aspects of our top models by using Stacking.

We found that features like popularity, budget, runtime, keyword, cast and crew were very significant in deciding revenue. Among all the genres, adventure, drama, animation and family movies generated the highest revenue meanwhile musicals, historical and war based movies contributed least to the total revenue.

References

[1] Hammad Afzal and Muhammad Hassan Latif. Prediction of Movies popularity Using Machine Learning Techniques . *International Journal of Computer Science and Network Security*, 16(8):127–131, 2016.

[2] Quazi Ishtiaque Mahmud, Nuren Zabin Shuchi, Fazle Mohammed Tawsif, Asif Mohaimen, and Ayesha Tasnim. A Machine Learning Approach to Predict Movie Revenue Based on Pre-Released Movie Metadata. *Journal of Computer Science*, 16(6):749–767, 2020.

[3] N. A. Pangarker and E.v.d.M. Smit. The determinants of box office performance in the film industry revisited. *South African Journal of Business Management*, 44(3):47–58, 2013.