**Application of Machine Learning Models for Movie Box Office Prediction and Data Exploration**

Victoria Anorve, Joel Pepera, Benya Piyathamsoontorn, Radha Yadavilli, & Dan Zhou

---01/05/2020

**Keywords**

*Box Office Predictions, Python, SQL, Machine Learning, Model Training, Real-time prediction APP*

Movie box office revenue prediction is an important issue in the film industry. And the application of machine learning models for planning and prediction revenues has become an irrevocable part of data studies aimed at improving the subsequent movie production. Generally, the prediction is based on using data collected from the multiple social media, internet social sources including YouTube, Twitter, Box Office Mojo, IMDb, The Numbers, etc.

Success or failure rate of movie can depend on several features: release period, investment budget, actor/actress selections, Movie rating, Studio, director and so on. Our capstone project aims to develop a best model based upon the data mining techniques that will help in predicting the success of a movie production in advance for the purpose of reduce the risk of movie production.

Throughout this project, we will solve the following questions and hypothesis:

* How much variation is there in box office performance by movie?
* What attributes about the movie show the strongest correlation to sales?
* What level of prediction accuracy is possible with the data we have?
* How much revenue will a particular movie gross in sales?

**Data Collection**

Our dataset are aggregated from Box Office Mojo, Internet Movie Database(IMDb), The Numbers. First, we downloaded the detailed information of movies from IMDb. Then we collect the URL of the 10 web pages with Requests with we constructed loop to go through each of the pages. we scraped Box Office Mojo, IMDb, The Numbers websites to gather the data to a CSV file with Beautiful Soup modules installed in Python 3. The aggregated the metadata contains 73,627 movies released since 1977 to 2018. All collected data are writing to a comma-separated vales (CSV) files. Due to missing or invalid data is impractical for modeling, only around 2,800 movies and 12 variables which have valid box office revenue data from 2000 to 2018 will be used as a means of creating data frame.

https://github.com/georgetown-analytics/Box-Office-Smash/tree/master/notebooks



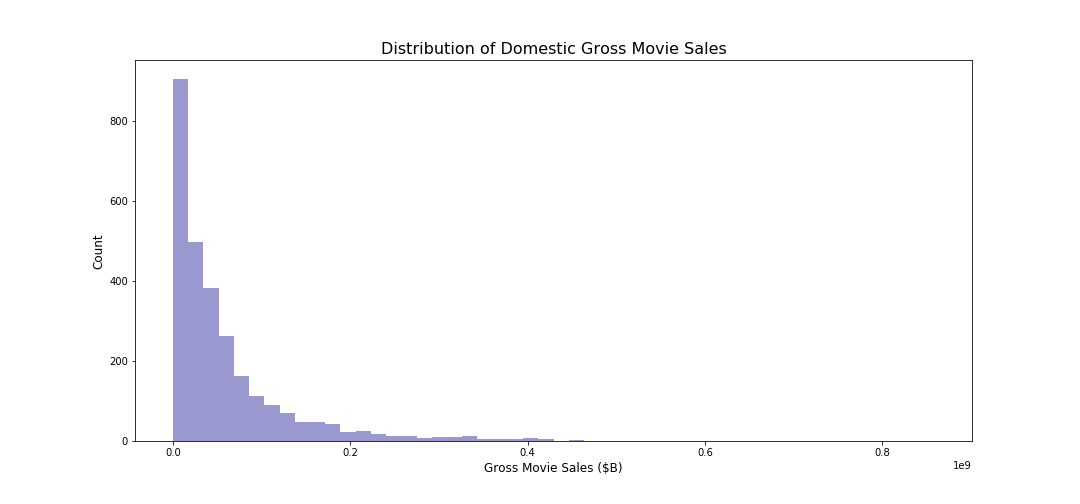
**Exploratory Data Analysis**

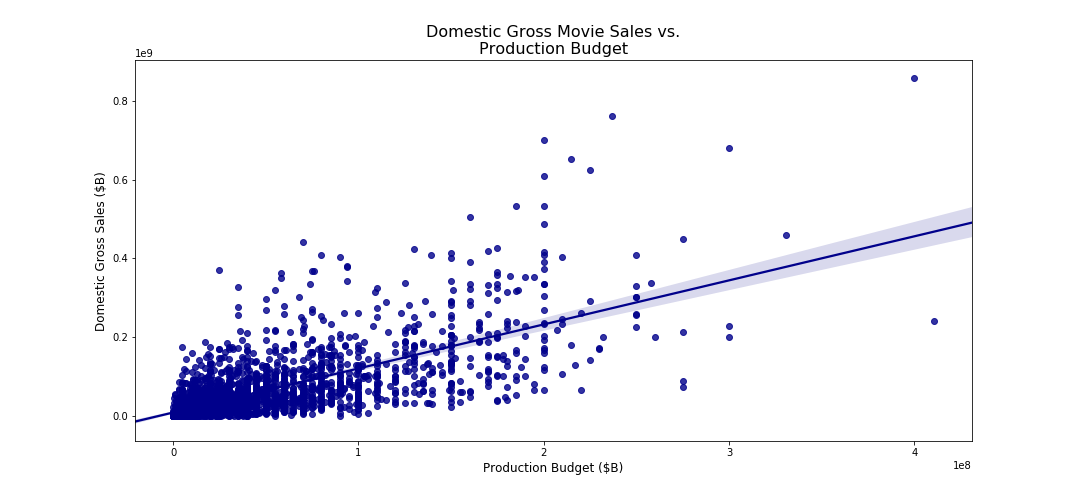
The variables used for input were:

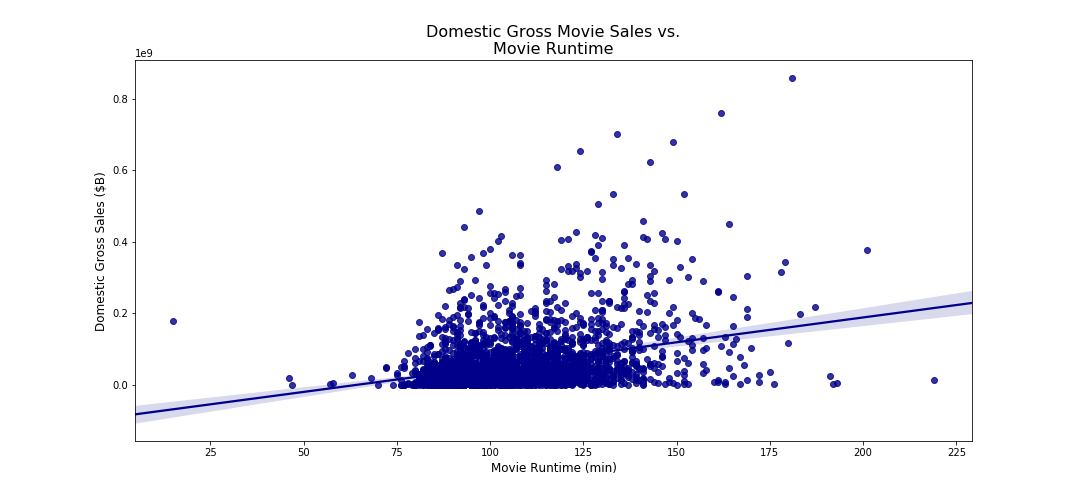
Domestic Gross Box Office Sales (target),Release Date, Production Budget, Movie Rating, Studio, Movie Runtime, genres, Keyword Descriptions, Directors, Writers, Composers, Actors/Actress

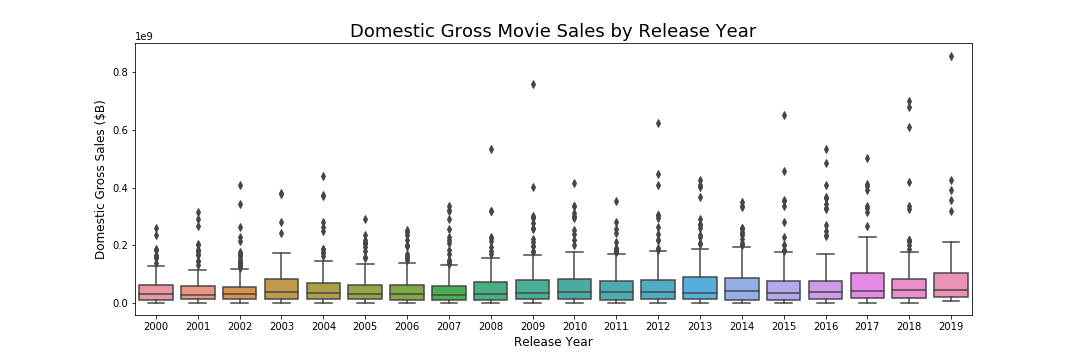
--The gross of the movie sale has high correlation with Production Budget, Movie Runtime, release time, and genres.

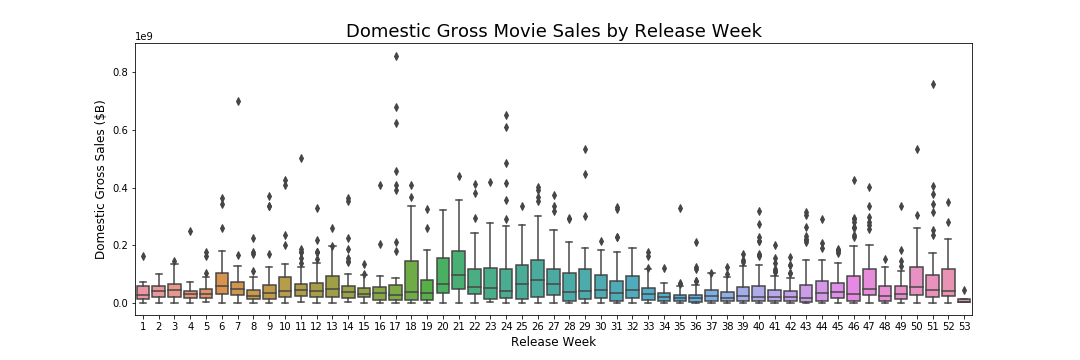
-- The gross of the movie sale has low correlation with actor and actress.

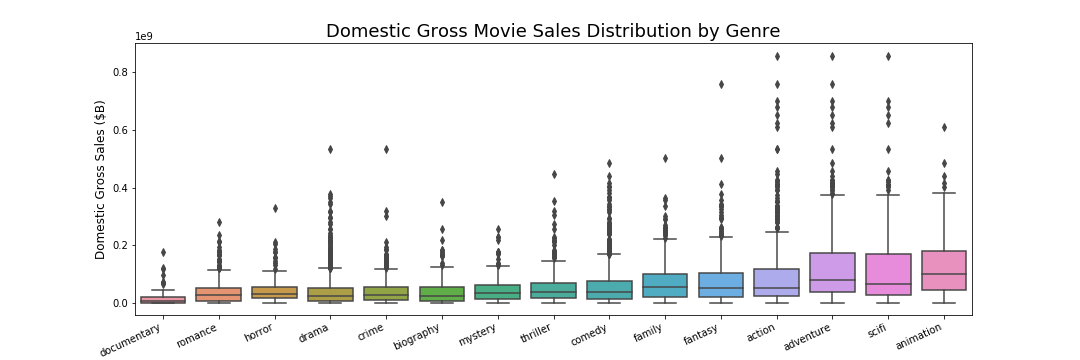
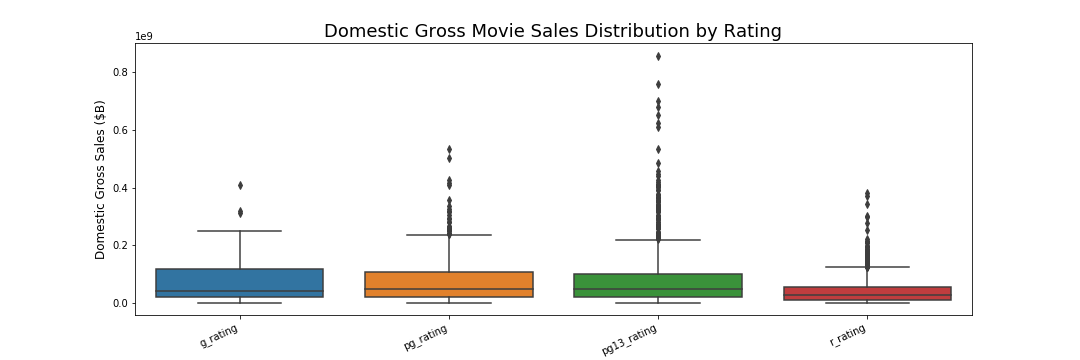


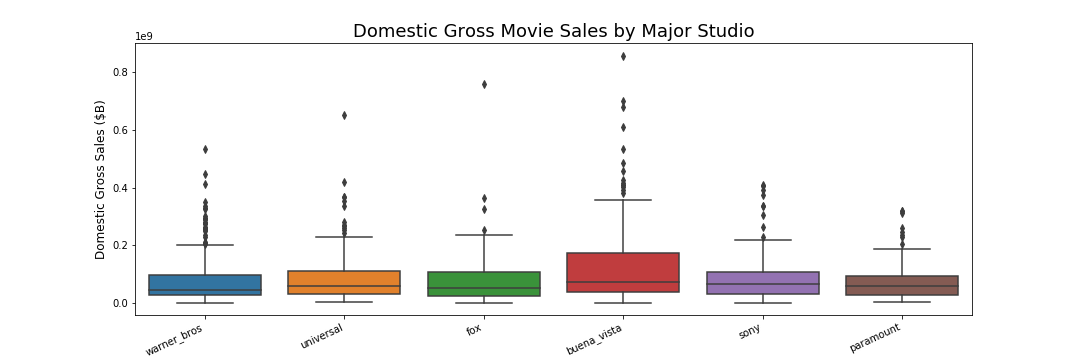


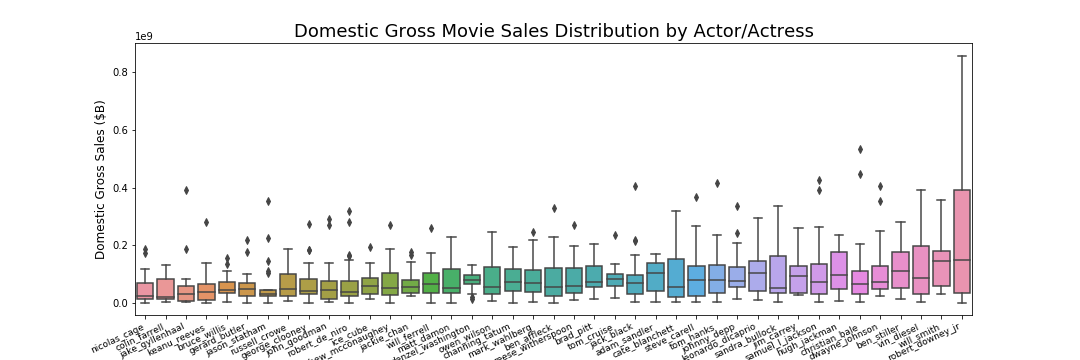


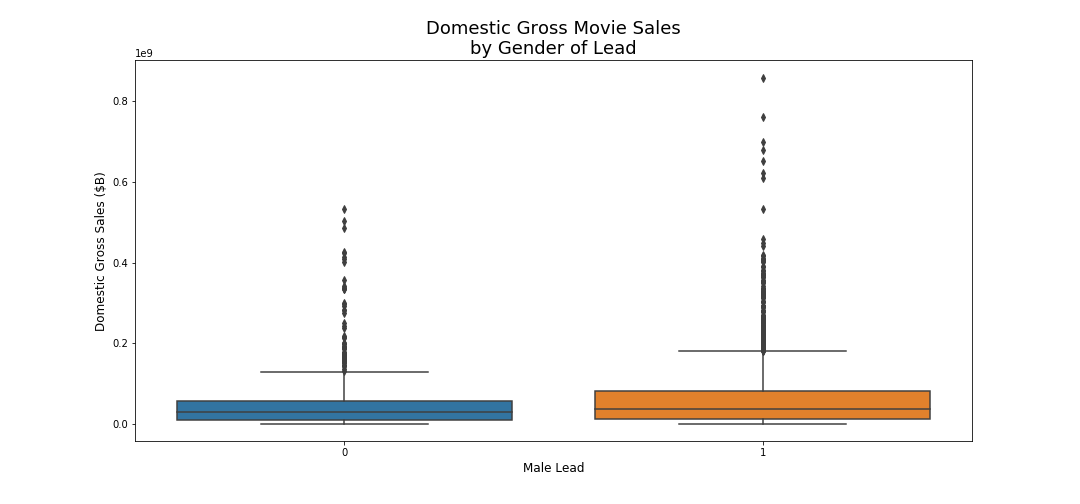


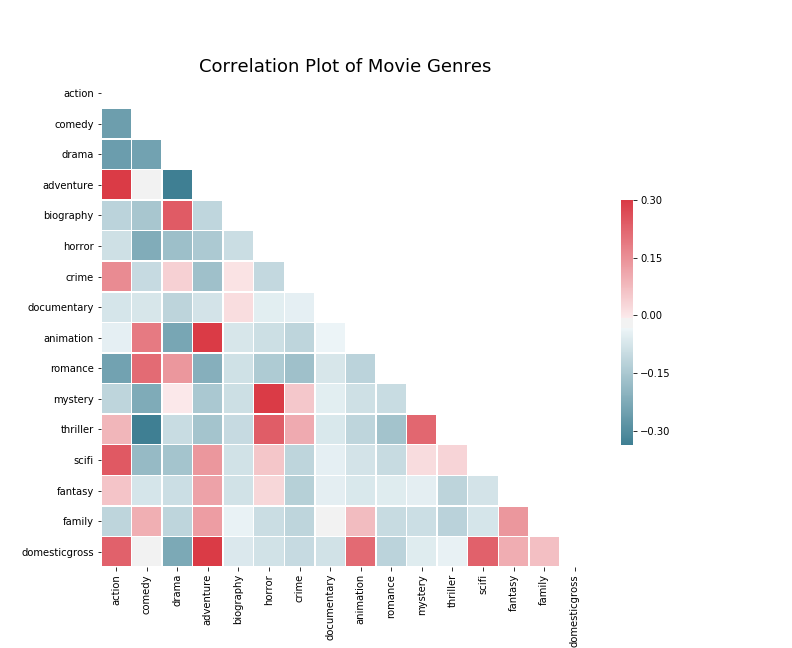
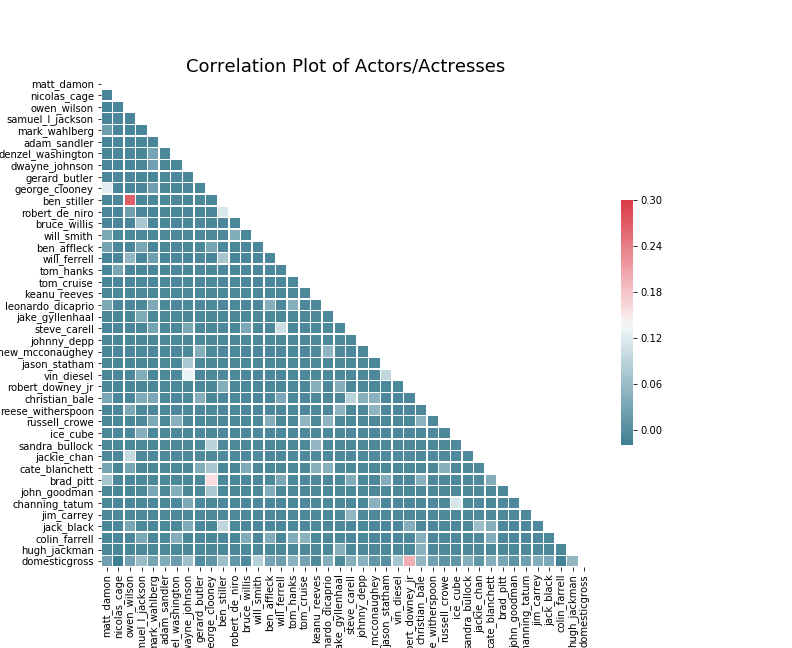


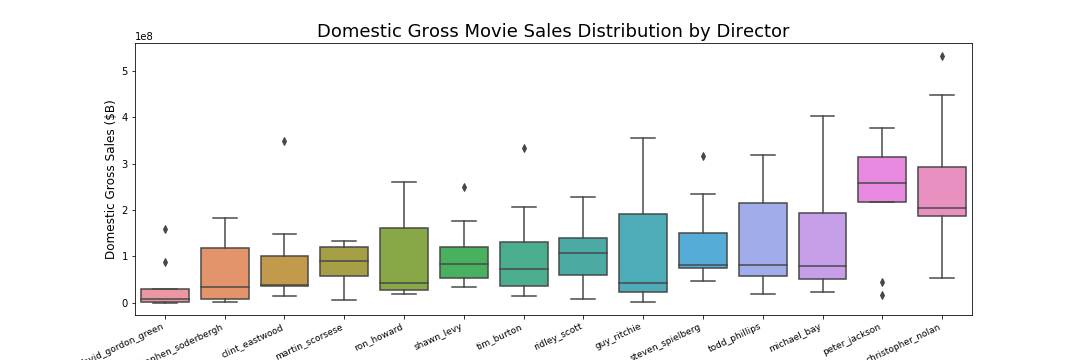


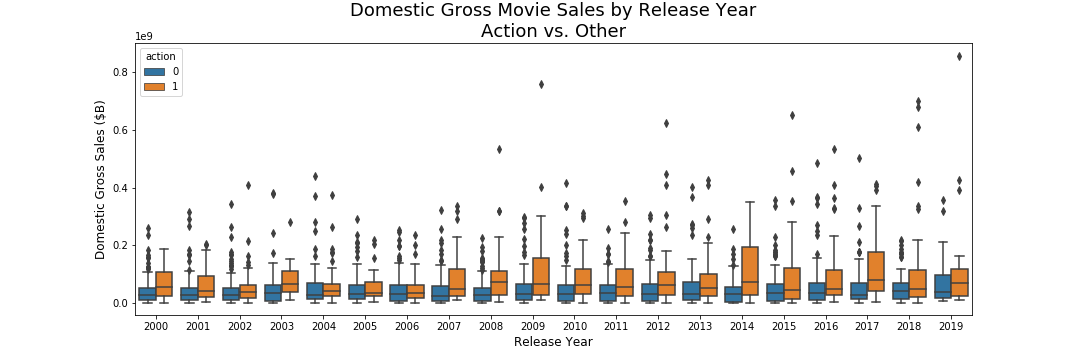


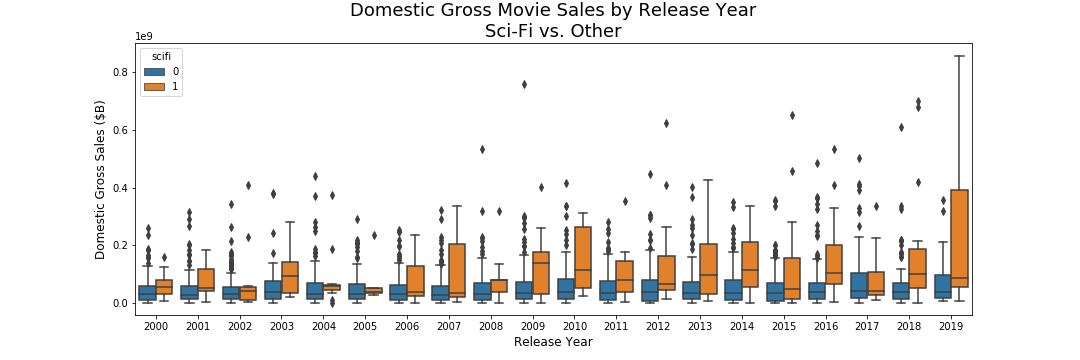


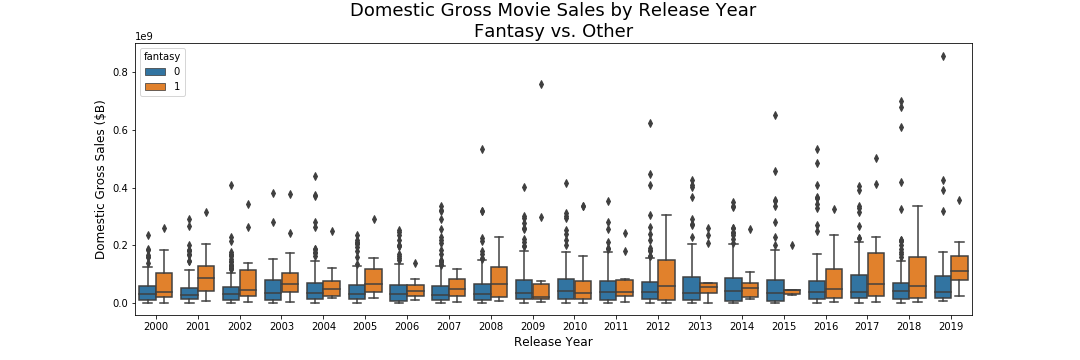


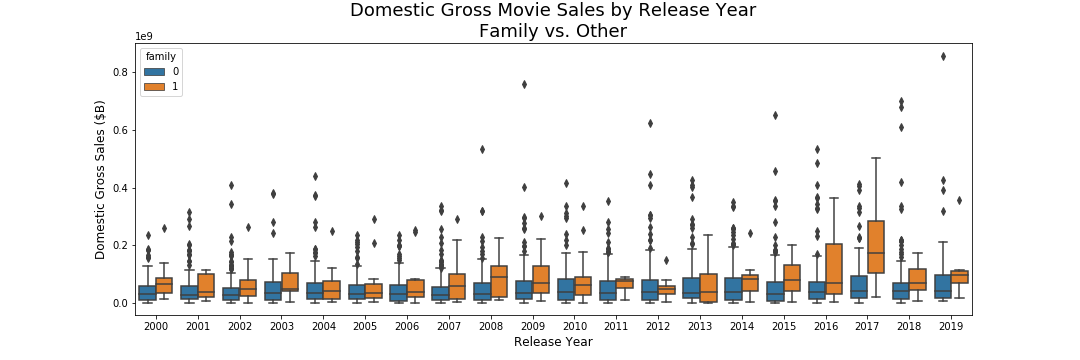


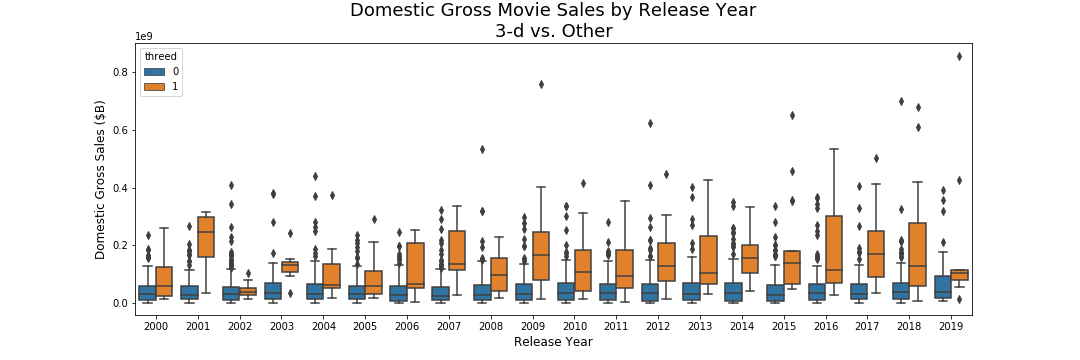








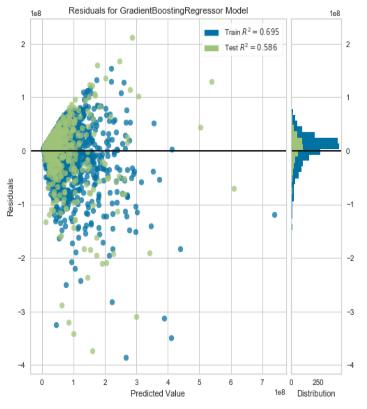
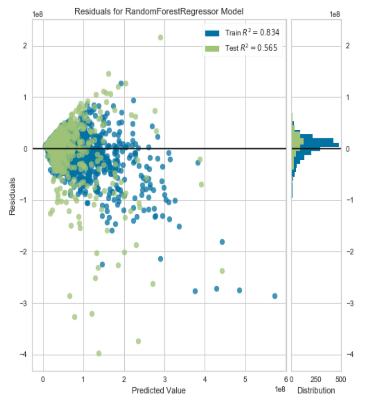
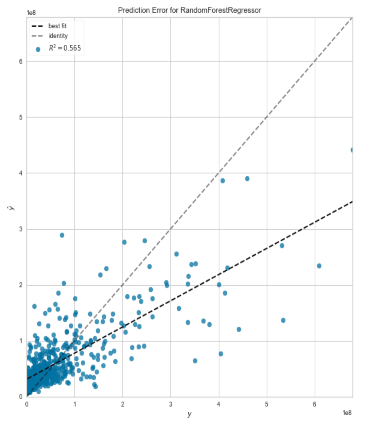
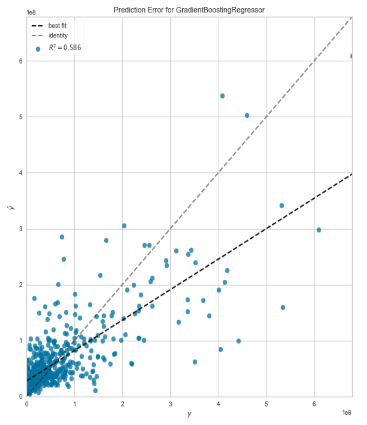
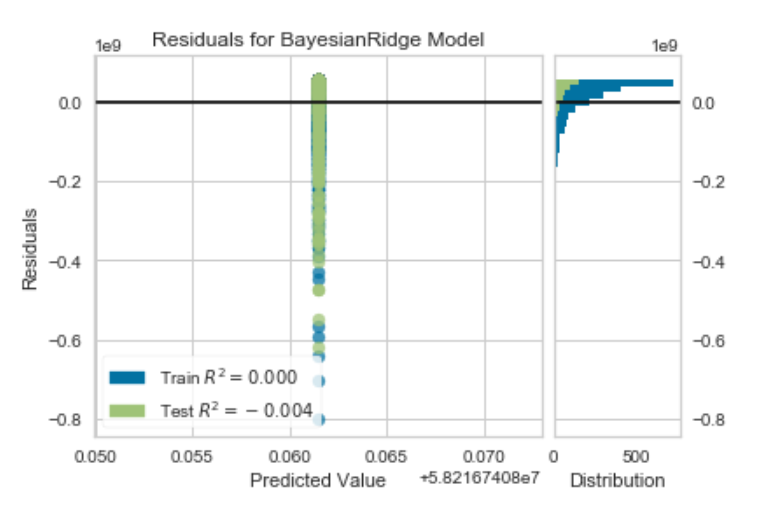
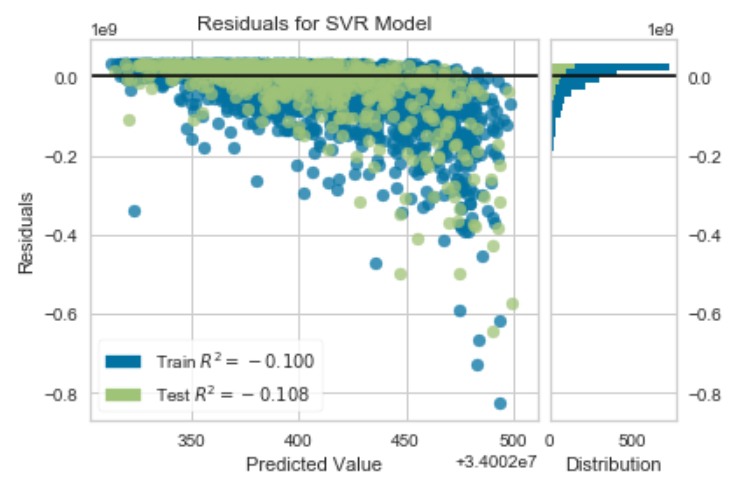
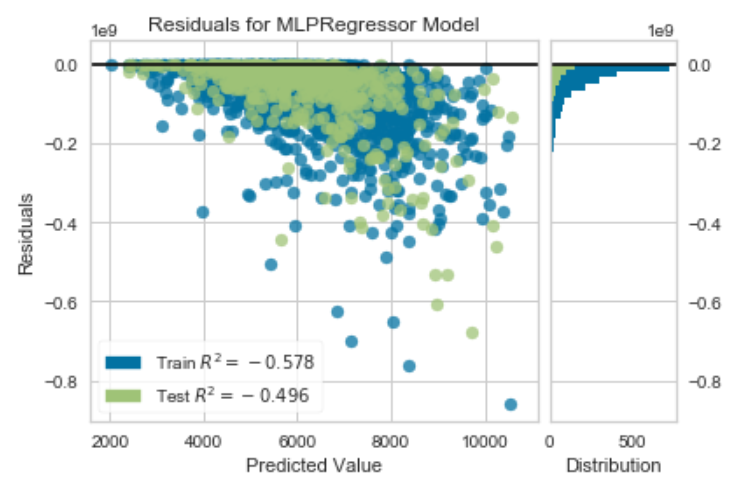
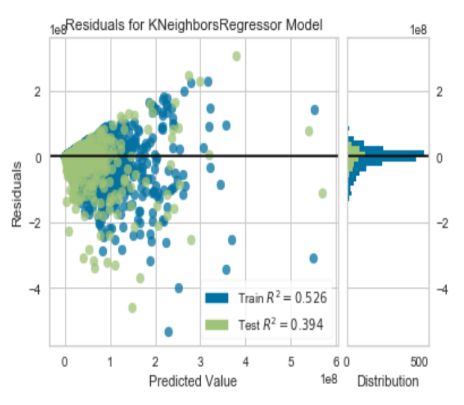
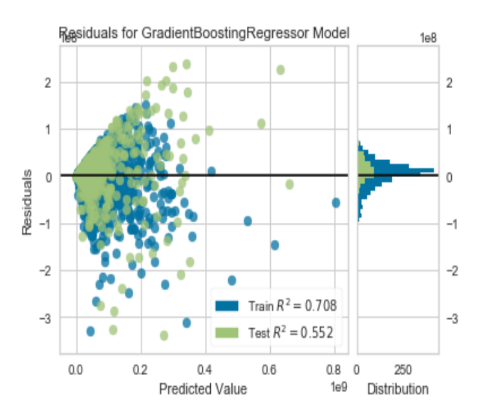
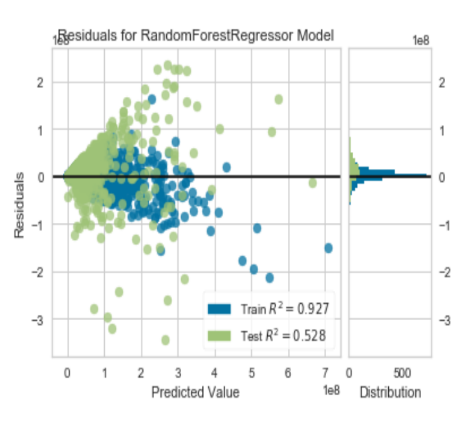
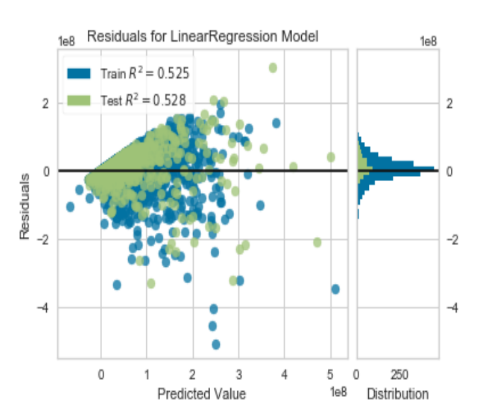
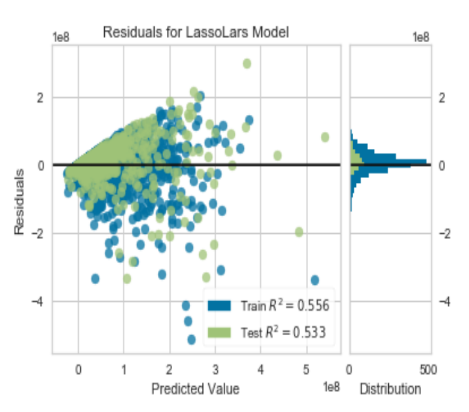




**Build and Validate Model**

In the machine learning step, we implemented and tested several machine learning algorithms with data collected. We initially applied Gradient Boosting Regressor, Multi-Layer Perception, Nearest Neighbors Regressor, Bayesian Ridge, Linear Regression, Elastic Net Regression, Lasso Lars and Random Forest Regressor for prediction of revenues. we explored methods for hyperparameter turning and saw improvement in Gradient Boosting Regressor, Elastic Net Regression, Lasso Lars and Random Forest Regressor model performance. Finaly, the Gradient Boosting Regressor algorithm has the highest performance.

**Fig 1. Target Variable – Gross Domestic Sales**



GradientBoostingRegressor

RandomForestRegressor

**Real-time APP prediction**

<http://ec2-34-238-50-190.compute-1.amazonaws.com/>

**Conclusion**

In our project, predictive models for the box office performance of the movies was trained by factors derived from scraped Box Office Mojo, IMDb, The Numbers websites. According to our models, the Gradient Boosting Regressor algorithm can improve the accuracy of perdition performance.