**The Analysis of Movies’ Global Box Office and Oscar Nominations**

Group A

Tianyi Deng

Biran Li

Jiyang Liu

Yu-lin Shih

***Abstract***

We studied global box office of major movies, and analyzed the key attributes to be considered in Oscar nominations. We acquired the dataset of IMDB 5000 from kaggle.com, the movies and ratings from grouplens, and the Academy Award records. The data set size is around 510MB. We predicted the global box office by using Linear Regression model and Gradient Boosting Tree model. For Oscar nominations, we used Decision Tree and Logistic Regression as classification methods. We ran these models on both Azure ML and Spark ML, then compared them using cross validation to find out which algorithm would produce the most accurate forecast.

***Introduction***

The movie industry is one that yields more than 11 billion dollars annually, and the box office of one popular movie could be over $500 million. The numbers are very attractive to conglomerates as well as independent studios, as they hope to scope some profits in the pool. The barrier to entry for this industry is high, and the ever developing technologies used in movies amplifies the competition even more. To help companies in this industry better understand what attributes are contributing most to the box office, we categorized all the relevant factors and studied each one in terms of their importance. Also, most filmmakers (if not all of them) dreams about winning the Oscars, while the result of the final winner could be hard to predict, we decided to analyze the nominations and reached a conclusion on which elements should be emphasized if one were to join the chase of the golden statue.

***Related Works***

Jason and Bridge from Stanford University forecasted the gross movie box office in 2013. They used Linear Regression and Logistic Regression generalized for multiple labels, and they also used the K-means clustering to increase efficiency. Michael and Kang from Iowa University predicted movie success, in which they used Artificial Intelligence Algorithm. We chose Linear Regression and Gradient Boosted Tree Regression, which differ from their techniques. We also used Spark machine learning on cloud computing, and cross validated the models to compare their performance. This process enhanced the accuracy of our research, thus differentiated our work from theirs.

As for the Academy Award, Atakancetinsoy from Bigml used Decision Tree and Logistic Regression in his analysis, which were the same as ours; but our movie dataset was much more comprehensive -- we used the data between 1927 to 2015, whereas he used data from 2000 to 2015. We focused on the forecast of the nominations, and he predicted the winner of each category.

***Background***

Lots of research has been done on movie industry already, and we built our project based on the existing resources available online. We gathered our data from the following three sources: GroupLens Research for Movielens Lab Dataset (<https://grouplens.org/datasets/movielens/>), IMDB 5000 Movie Dataset (<https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>), and The Academy Awards(1927-2015) Dataset (<https://www.kaggle.com/theacademy/academy-awards>).

Movielens dataset contains 20 million movie ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users. IMDB dataset provides 28 variables for 5043 movies , spanning across 100 years in 66 countries. The Academy Award dataset has the official record of past Academy Award winners and nominees from 1927 -2015.

***Methods***

From the data sources mentioned above, we selected over 500MB data, which seemed overwhelming. In order to elevate data quality, we first took a few steps to clean and prepare the data. Then we trained and tested them using various algorithms, and finally compared the metrics of those algorithms to determine which one had better performance.

**Movies Global Box Office**

We used the GroupLens Research and IMDB data for our global box office research. First of, we used Jupyter to change and add columns, and changed data type from string to integers.Then we uploaded the files to Azure Machine Learning, where we joined data, selected columns in dataset,

cleaned missing data, edited metadata, and normalized data. After we are done with the initial state in data preparation, we pinpointed the following to further understand the data: the target column was global gross sales, which had the minimum value of $423 and maximum value of $2,783,918,982; the number features were budget, domestic gross sales, global gross sales, movie duration and imdb score, whereas the categorical features included language, country, and major genres.

With these information identified, we chose 2 algorithm to use: the Linear Regression Model and the Gradient Boosted Tree Regression Model.

In Azure ML, we split the data, used 70% of them to train and 30% to test. The metrics in evaluation showed that the Root Mean Squared Error is 52,050,094.7637 for Linear Regression and 63,194,986.7556 for Gradient Boost Tree Regression; the Coefficient of Determination is about 90.6% for Linear Regression and 86.2% for Gradient Boosted Tree Regression.

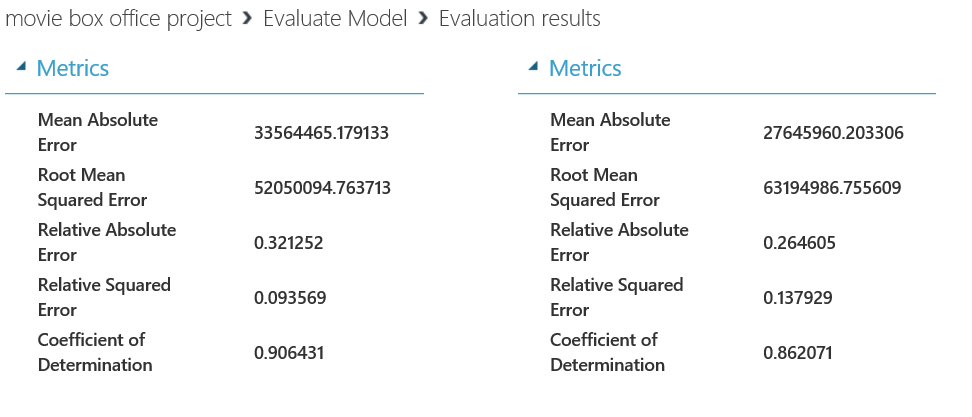


Figure 1: Evaluation results

From the statistics, we could draw the conclusion that Linear Regression had better performance in Azure ML, as the Root Mean Squared Error was lower and the Coefficient of Determination was higher.

We also summarized the importance scores of the features selected earlier: domestic gross sales had the highest score, which means it had the strongest correlation to global gross sales.

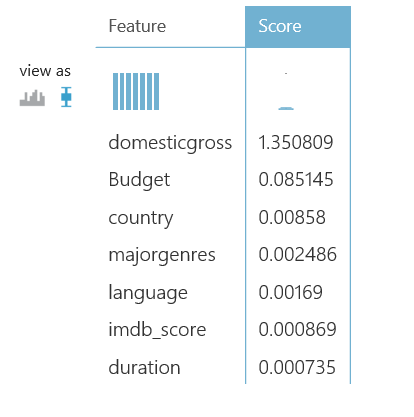


Figure 2: Importance Scores

Then we moved on to conduct the research in Spark Machine Learning: we used Linear Regression on IBM Data Science Experience, and Gradient Boosted Trees in Databricks.

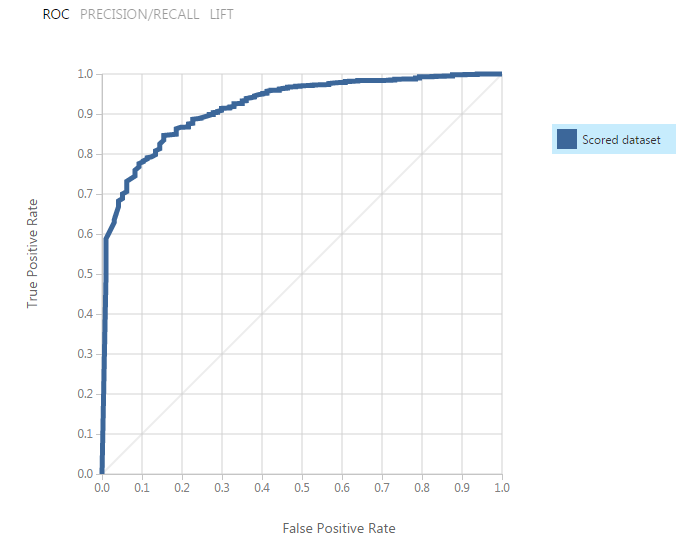
After we loaded and prepared data on these two platforms, we started cross validation: we split data in 10 folds, used 70% for training and 30% for testing. Then we used pipelining to streamline the workflow. Next we trained and tested the data, and arrived at the results below:

Linear Regression had RMSE of 52,699,899.3262, and Gradient Boosted Tree had RMSE of 60,409,700.0000. The absolute value looked high at first glance, however, when compared with the maximum value of global gross sales, it was only roughly 2% of that. Thus, we believe both models, as well as the data and columns we used, were relevant and would provide meaningful insights. According the the smaller RMSE numbers, we concluded that Linear Regression Model was more suitable for our analysis.

**Oscar Nominations**

For Oscar Best Picture analysis, we selected two classification methods -- Decision Tree and Logistic Regression.

We prepared the data by joining IMDB 5000 Movie Dataset and the Academy Award Dataset, which generated a new file called Moive\_Oscar.csv. The features included gross sales, movie budget, director's’ facebook likes, cast’s total facebook likes, IMDB score, movies’ facebook likes, language, and country. The labels were nominations and winners. The final winner was hard to predict, since the decision had great uncertainty and attributes that could not be quantified. We decided to use nominations as our target column to train the models instead. We evaluated both models and received the following results:



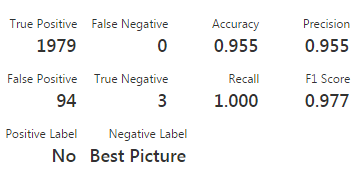
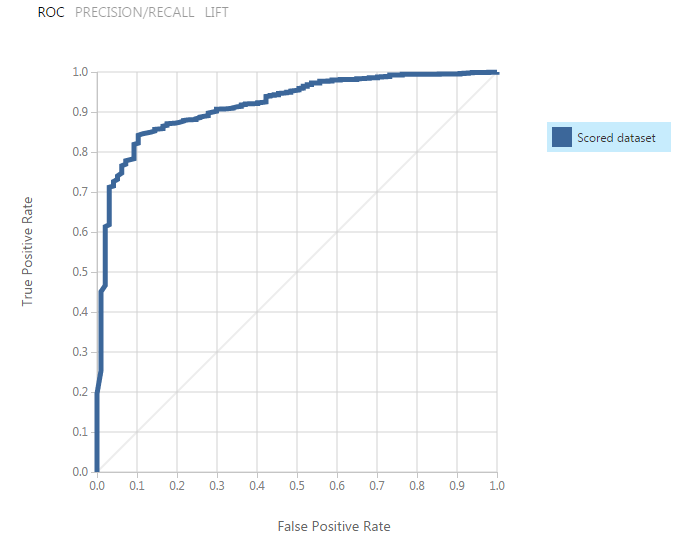


Figure 3: Two-Class Decision Tree



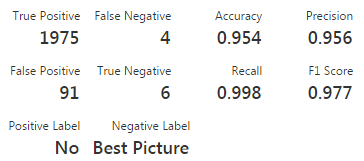


Figure 4: Two-Class Logistic Regression

The accuracy and precision rates in both models were very similar, which means these two models had comparable performances.

To increase the relevance of our data, we also used permutation to rank the importance of the features, and determined the top three were the gross sales, IMDB score, and director’s facebook likes. As we continued our experiment in DataBricks, we chose only these three out of all eight features to filter out the noises. In cross validation, we used 70% and 30% for training and testing, respectively. Then we used pipelining to combine the stages. Last but not least, we used test error, which equals 1 minus accuracy rate, to assess the effectiveness of the two models: the outcomes were identical, confirming that both Decision Tree and Logistic Regression model had great usability for this situation.



Figure 5: Test Error in Decision Tree(0) and Logistic Regression (1)

***Conclusion***

Based on the above study, we compiled our findings and decided that for predicting movie global box office, Linear Regression was better as it had lower RMSE.

|  |  |  |
| --- | --- | --- |
| RMSE | Azure ML | Spark ML |
| Linear Regression | 52,050,094.7637 | 57,188,471.1788 |
| GBT Regression | 63,194,986.7556 | 60,409,700.0000 |

Figure 6: RMSE comparison for movie box office

For Oscar nominations, both Decision Tree and Logistic Regression algorithm were appropriate to use: they had similar precision and accuracy rates in Azure ML, and equal values of test error in Spark ML.

|  |  |  |
| --- | --- | --- |
| Accuracy | Azure ML | Spark ML |
| Decision Tree | .955 | .9542 |
| Logistic Regression | .952 | .9542 |

Figure 7: Test Error comparison for Oscar nomination

***Observation***

We made few observations during the research: Azure ML is good for cleaning missing data during the initial data preparation phase: with the various functions built in including the graphics, it is easy to use. Spark ML on the other hand, is more geared towards advanced graphics where coding is necessary. We also noticed that DataBricks has shorter processing time, whereas IMB data science provides a detailed error log for tracking down the problems.

***References***

1. Microsoft DAT202.3x Implementing Predictive Analytics with Spark in Azure HDInsight
2. Microsoft's DAT203x, Data Science and Machine Learning Essentials
3. Gautam, Geetika, and Divakar Yadav. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis - IEEE Xplore Document. N.p., n.d. Web. 02 May 2017.
4. "MovieLens." GroupLens. N.p., 18 Oct. 2016. Web. 02 May 2017.
5. Concetta A., and Constance H. McLaren. "Movie Data." Journal of Statistics Education,V17n1. N.p., 2009. Web. 02 May 2017.
6. Chuansun7. "IMDB 5000 Movie Dataset." IMDB 5000 Movie Dataset | Kaggle. N.p., n.d. Web. 02 May 2017.
7. "The Academy Awards, 1927-2015." Academy of Motion Picture Arts and Sciences, n.d. Web. 02 May 2017.
8. Merwe, Jason Van Der, and Bridge Eimon. Predicting Movie Box Office Gross (n.d.): n. pag.Http://cs229.stanford.edu/proj2013/vanderMerweEimon-MaximizingMovieProfit.pdf. 12 Dec. 2013. Web. 14 May 2017
9. Patrick Von Sychowski. "AI Algorithm Predicts Box-office Success (Sort Of)." Celluloid Junkie. N.p., 31 May 2016. Web. 14 May 2017
10. "Predicting the 2017 Oscar Winners." The Official Blog of BigML.com. N.p., 25 Feb. 2017. Web. 19 May 2017.