# Movie Advisor

**Predicting upcoming movies' box office revenues in Taipei for theater managers to plan released weeks and halls for new movies.**

Data Mining Team 3
Jessica Deng, Jenny Wang, Sam Wang, Sean Xie
**2016/1/12**

**Executive Summary**

1. Summary

Our primary stakeholder is theater managers, an important role in theater who have to arrange released weeks and halls for each new movie. Therefore they have a potential need of knowing how new movies will perform on box office revenues. However, there's a gap among the box office revenues in Taipei and in US and other movie features, and cause the prediction difficult. Hence, our business goal of this project is to allow managers knowing how new movies will perform on box office revenues in Taipei in advanced.

To use data mining method to achieve our business goal, first, we turn the business goal into data mining goal. Now our data mining goal is to predict box office revenues in Taipei and the outcomes managers will get are box office revenues of each movie in Taipei. This project is then an ongoing project, which means the managers can use this model repeatedly once they have new movie record.

The data we have consists of movie features such as budget, movie type, IMDB rating, release date in US, and box office revenues in US. The time period is from 2010 to 2015, 2,632 movies in total initially. After handling the missing values and outlier values, we have around 560 record that are accessible. We did data preprocess (e.g. dummy variables) for certain variables such as movie types and movie rating, then partition it. All we did before building the model were aim to make our project more accurate. We choose XLMiner and R as our data mining tools. The data mining method we used was linear regression. Our client can predict whether a movie will have great box offfice revenues in Taipei or not by the ultimate linear regression model, and they will get predicted box office revenues as the outcome.

2. Recommendation

Although our outcome of the model has a huge rate of error, it's still much better than the average prediction. So those important predictors we mention in this report are credible.

For further works, we suggest the managers of theaters to collect more accesible data records and more valuable data dimensions like past released weeks and halls since the biggest weakness in our project is the data size. On the other hand, we suggest further studies should take environment changes and number of rating people into account. In addition, we also suggest further studies to try classification as the data mining method if our client require only level of box office revenues but not exact numbers.

## I. Business Goal and Humanistic Evaluation

Our main client, theater managers, have to arrange released weeks and halls for each new movie. Therefore they have potential needs of knowing how new movies will perform on box office revenues, and use the information to develop right strategies, increase profits, and reduce the costs.

## II. Analytics/Data Mining Goal

Our data mining goal is to to predict box office revenues in Taipei based on movie features and some movie information in USA such as box office revenues there. The project is an ongoing, predictive, and supervised task, and the main outcome variable is box office revenues in Taipei.

## III. Data

We've captured the data from 2010 to 2015, total of 2,632 movies which have been released in Taipei. We extracted about nine columns from Yahoo! Movie, by python, and for rest of the columns we collected manually through True Movie, Atmovies, PTT (the biggest bulletin board system in Taiwan), Dorama, YouTube and IMDB. As mentioned above we have 2,632 records initially, however after removing records which have too many missing values and of other situation (will be further discussed later in data pre-processing), only 560 records are left.

As for the variables, we have 21 attributes originally, after adding one more column DF that equals to the day difference between Taiwan and US released date, creating dummy variables for movie types/ratings, and going through other processes (will also be talked about later), we have 42 dimensions, sample data is pasted in appendix (1), and some of the variables are shown as following:

| **Dimension** | **Description** |
|---|---|
| Name_CN | The Chinese title of the movie |
| Name_EN | The English title of the movie |
| Date_TW | The release date in Taiwan (i.e. 2010/12/11) |
| Length | The length of the movie (i.e. 134 mins) |
| Agent | The movie agents in Taiwan. (i.e. CatchPlay) |
| Expectation | The audience expectation from Yahoo movie. (i.e. 0.95) |
| Production | The original production corporation of the movie (i.e. Warner bros.) |
| Country | The country of movie production (i.e. Japan) |
| Language | The main language of the movie |
| Date_US | The release date in US (i.e. 2010/11/20) |
| Budget | The budget of the movie (i.e. 12,000,000 USD) |
| Box office_USD | The box office revenues in the US (i.e. 1,700,000 USD) |
| IMDB | The customer rating (i.e. 3.5) |

| | |
|---|---|
| Youtube | The page views of the trailer on YouTube (i.e. 2,645) |
| DF | Date difference of release date between US and TW (i.e. -49) |
| Type | The type of the movie, and it's transformed to dummy (i.e. Action, Adventure) |
| Movie Rating | The movie rating in Taiwan, and it's transformed to dummy (i.e. Restricted) |
| Box office_TW | The box office revenues in Taipei city (i.e. 30,000,000 NTD) |

## IV. Data Preparation

We first created dummies for movie Rating and Type for prediction. Next, we removed movies that were released earlier in Taiwan than in USA, because we can't get box office revenues in US then. We also removed the records including incomplete data, for example, the records with 0 expectation or were just released in Taiwan. Finally, we found that there were several missing values in Budget column. We tried the model with the records possessing complete Budget data firstly, and we noticed that Budget was an important predictor for prediction. Therefore we did clustering and get the median for each cluster as the Budget values to fill in with. We also explored our data by visualization with Tableau, finding some insights and thoughts for later works. Details are shown in appendix (2).

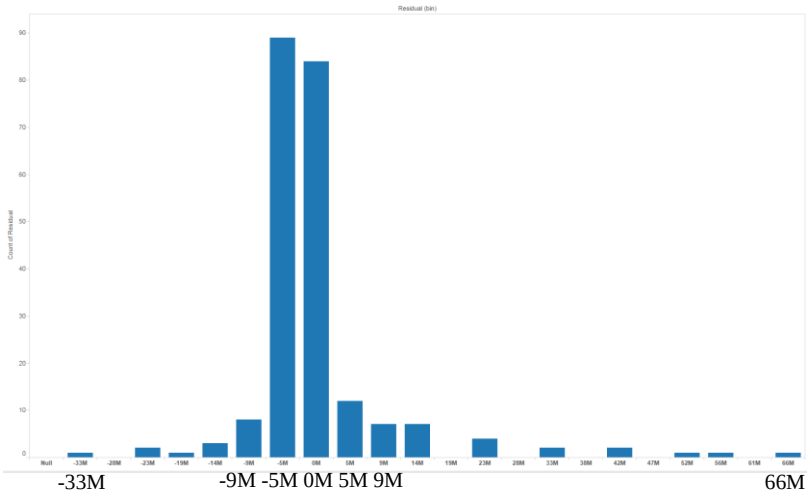
## V. Method

1. Data Partition, Variable Selection, and Linear Regression

We choose XLMiner and R as our data mining tools. First we partitioned dataset into only training and validation dataset because of the small data size, and tried linear regression with all numerical predictors and dummy variables. We found that the regression model with 60/40 partition ratio performs better. On the other hand, since there were several negative prediction values of box office revenues, we decided to take the form of "ln" to make our results positive. The box office revenues were more reasonable and performs better than the former one then. We used the variable selection method Stepwise to find the best subset of predictors, and it performs more accurately than the one with all predictors. Our ultimate linear regression model and its results show as below:

| Input Variables | Coefficient | Std. Error | t-Statistic | P-Value | CI Lower | CI Upper | RSS Reduction |
|---|---|---|---|---|---|---|---|
| Intercept | 2.199266 | 1.126238399 | 1.952753807 | 0.051704 | -0.01635 | 4.414879 | 72912.74 |
| Action | 0.387448 | 0.157891436 | 2.453885862 | 0.014655 | 0.076833 | 0.698062 | 58.22223 |
| Crime | -0.49101 | 0.224720299 | -2.18499959 | 0.0296 | -0.9331 | -0.04893 | 8.635878 |
| Expection_LN | 3.360491 | 0.567218993 | 5.924503816 | 7.96E-09 | 2.24462 | 4.476363 | 351.6101 |
| Budget_USD_ | 0.335688 | 0.062097141 | 5.405859428 | 1.25E-07 | 0.213527 | 0.45785 | 319.4753 |
| USA_USD_LN | 0.224544 | 0.029657868 | 7.571141518 | 3.84E-13 | 0.166199 | 0.282889 | 61.43727 |
| IMDB_LN | 1.012856 | 0.317899764 | 3.186086873 | 0.001582 | 0.387462 | 1.63825 | 2.165651 |
| YouTube Trai | 0.288635 | 0.040603482 | 7.108621398 | 7.44E-12 | 0.208757 | 0.368513 | 57.22768 |
| DF_LN | -0.30287 | 0.050873856 | -5.953358032 | 6.79E-09 | -0.40295 | -0.20279 | 32.21697 |

| Total sum of squared errors | RMS Error | Average Error |
|---|---|---|
| 2.729E+16 | 11037670.5 | 2421600.235 |




2. Neural Network and Regression Trees

We also tried neural network and regression trees for prediction, but they didn't perform better than the regression model. The results of these two methods are put in appendix (3) (4).

3. Ensemble

We found that the residuals of all three models were actually positive correlated so it's not necessary and efficient to do ensemble. Correlation plots are shown in appendix (5).

## VI. Performance Evaluation and New Data Prediction

We finally choose the linear regression as our prediction model. The results including variables, RMSE, and residual histogram are all listed in Method above. We have compared it with Naïve results. Our RMSE performs better, and the error rate is not only better but far beyond.

| Total sum of squared errors | RMS Error | Average Error | Error_rate_regression | Error_rate_naive |
|---|---|---|---|---|
| 7.6901E+16 | 18528520.32 | 433343.0817 | 114.34% | 1981.97% |

We also extracted the movies which just released or are upcoming as our test dataset. The prediction results and their actual box office revenues so far are listed, and we have confidence that they are heading to our prediction. The whole test dataset is in appendix (6)

| Name_CN | Name_EN | Predicted | Exp | Current | Date_TW | Length | Director | Cast |
|---|---|---|---|---|---|---|---|---|
| 史努比 | A Peanuts Mo | 15.46 | 5,188,100 | 16,000,000 | 12/24/20 | 88 | 《冰原歷險記4：林 | |
| 紐約愛未 | Before We Go | 13.35 | 625,993 | 1,420,000 | 12/24/20 | 89 | 克里斯伊 | 《美國隊 |
| 真相急先 | Truth | 14.79 | 2,648,649 | 5,540,000 | 12/24/20 | 125 | 詹姆斯范 | 《藍色茉 |
| 翻轉幸福 | Joy | 17.40 | 36,085,718 | 1,830,000 | 12/31/20 | 124 | 《派特的 | 《飢餓遊 |
| 家有兩個 | DADDY'S HOM | 16.42 | 13,467,908 | 9,480,000 | 12/31/20 | 96 | 《老闆不 | 《官賤對 |
| 怪物遊戲 | Goosebumps | 15.74 | 6,848,908 | 11,000,000 | 12/31/20 | 103 | 《鯊魚黑 | 《格列佛 |
| 神鬼獵人 | The Revenant | 16.69 | 17,732,292 | 11,000,000 | 1/8/2016 | 151 | 阿利安卓 | 李奧納多 |
| 瞎趴姊妹 | Sisters | 16.21 | 10,917,346 | 780,000 | 1/8/2016 | 118 | 《歌喉讚 | 《愛在頭 |
| 女權之聲 | Suffragette | 15.01 | 3,315,551 | 340,000 | 1/8/2016 | 106 | 莎拉賈芙 | 《大亨小 |

## VII. Conclusions

1. To Our Client (theater managers)

Even though the overall accuracy is not high, we still get some insights for our client. First of all, when planning total released weeks and halls for a new movie, its budget, box office revenues, and released date difference between Taiwan and US should be considered. Second, some criteria about audience, like expectation rate, IMDB rate, and trailer page views are also seem to be important. Last, the movie type will also influence the box office revenues, especially action and crime movies.

2. To Future Studies

- Data size

The biggest weakness in our project is the data size. In order to train a more accurate model, future studies should focus more on dealing with missing values, and collecting more data.

- Number of rating people

For dimensions “Expectation” and “IMDB,” researchers should take the number of rating consumers or users into consideration. Otherwise we cannot rule out the bias situations that some high rating movies were actually only rated by very few people.

- Trailer page views

The page views of trailers on Youtube, especially of those trailers that were published years before, are accumulated over years and may not be precise as our predictor. Future research can find ways to eliminate this error.

- Environment changes

We think the changes of big environment should be considered, including floating exchange rate, increasing movie ticket price, and the change of consumer behavior (more and more people go to theaters to watch movies nowadays).

- More valuable predictors

When talking about box office revenues, released weeks of each movie should also be counted in. Besides this, we found other interesting and important predictors while reading papers and reports of similar topic. For example, some researchers give each movie a “star point,” indicating whether the director or cast of a movie is famous and has enough impact on audience.

- Classification method

If our client require only the level of revenues, not exact numbers, we suggest to change this project to a classification task, by coding the revenues to several classes and running classification methods.

## VIII. Appendix

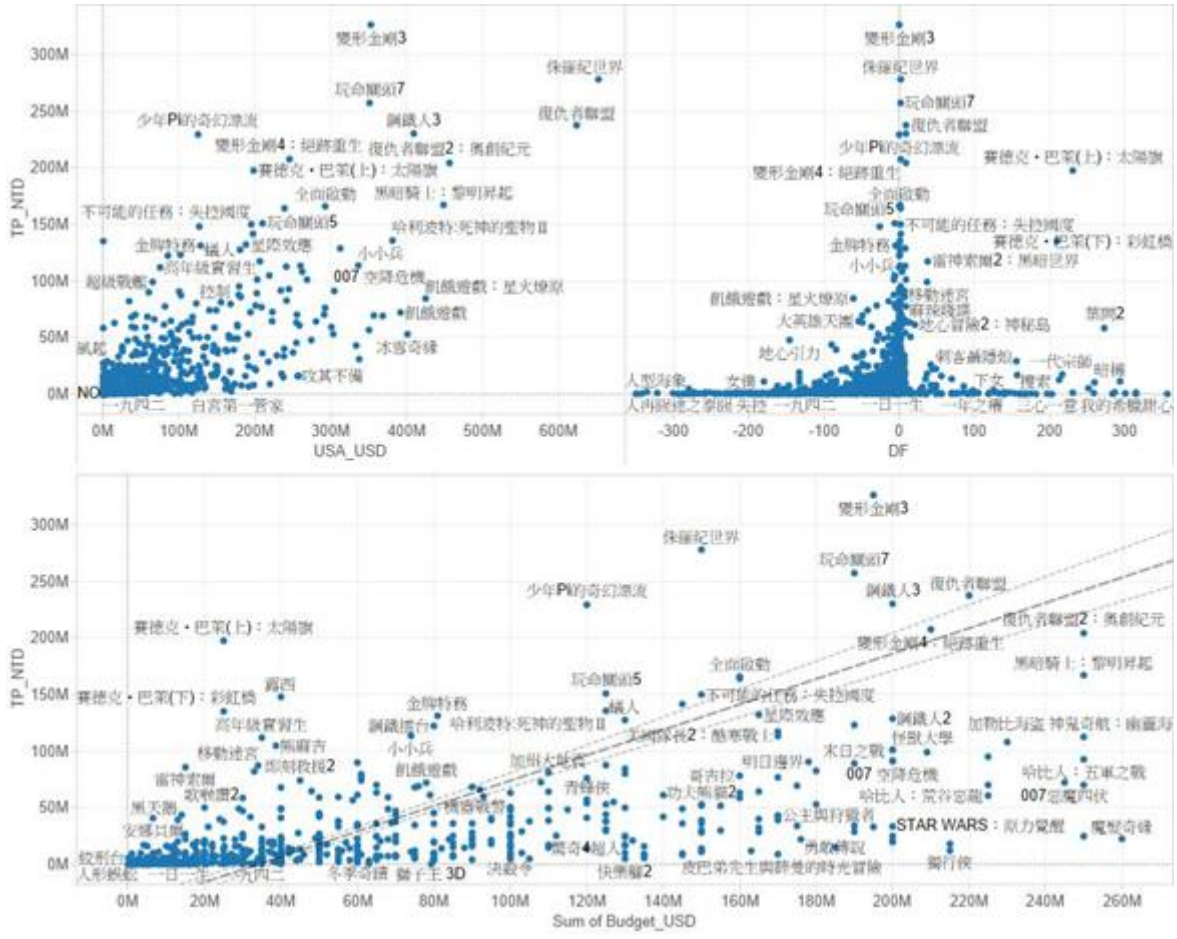
1. Sample data: We here show 10 rows of records.

| Name_CN | Name_EN | Date_TW | Length(raw | Length(min | Director | Cast | Agent | Expection | Type | Production | Country | Language | date_US | Budget_USD | USA_USD | IMDB | YouTube | DF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 破碎的擁 | Broken Er | 40186 | 0.09 | 128 | 《悄悄告 | 《情遇巴 | 山水 | 0.91 | 劇情/愛情 | Universal | Spain | Spanish | 40137 | 1.8E+07 | 4930000 | 7.2 | 1828 | -49 |
| 打不倒的 | INVICTUS | 40193 | 0.09 | 134 | 《登峰造 | 摩根費里 | 華納兄弟 | 0.94 | 劇情 | Warner B | USA | English | 40148 | 6E+07 | 3.7E+07 | 7.4 | 32940 | -45 |
| 鼠來寶2 | Alvin and | 40199 | 0.06 | 89 | 《怪醫杜 | 《全民公 | 福斯 | 0.95 | 喜劇 | Fox 2000 | USA | English | 40170 | 7.5E+07 | 2.2E+08 | 4.5 | 102997 | -29 |
| 歡吉桑卡 | Old Dogs | 40200 | 0.06 | 87 | 《荒野大 | 《荒野大 | 博偉 | 0.95 | 喜劇 | Walt Disr | USA | English | 40142 | 3.5E+07 | 4.9E+07 | 5.4 | 2257 | -58 |
| 帕納大師 | The Imag | 40200 | 0.08 | 122 | 《神鬼克 | 《黑暗騎 | 威視電影 | 0.95 | 奇幻/劇情 | Infinity Fe | France | English | 40172 | 3E+07 | 7670000 | 6.9 | 12153 | -28 |
| 食破天驚 | Cloudy W | 40207 | 0.06 | 90 | | (配音)《 | 博偉 | 0.9 | 動畫 | Columbia | USA | English | 40074 | 1E+08 | 1.2E+08 | 7 | 15190 | -133 |
| 墮落與救 | FILTH & V | 40207 | 0.06 | 84 | 瑪丹娜(M | 尤金・赫 | 聯影／聯 | 0.74 | 劇情 | Semtex F | UK | English | 39747 | 5300000 | 22406 | 5.6 | 1525 | -460 |
| 奪天書 | The Book | 40221 | 0.08 | 118 | 《關證手 | 《亡命快 | 博偉 | 0.91 | 動作/劇情 | Alcon Ent | USA | English | 40193 | 8E+07 | 9.5E+07 | 6.9 | 2599 | -28 |
| 沙灘上的 | The Beac | 40235 | 0.08 | 110 | 安妮華達(Agnès V | 聯影／聯 | 0.62 | 歷史/傳記 | Ciné Tam | France | French | 40034 | 4000000 | 239711 | 7.9 | 4452 | -201 |
| 攻其不備 | The Blind | 40235 | 0.09 | 128 | 《心蠶投 | 《愛情限 | 華納兄弟 | 0.96 | 劇情/溫馨 | Alcon Ent | USA | English | 40137 | 2.9E+07 | 2.6E+08 | 7.7 | 6922 | -98 |

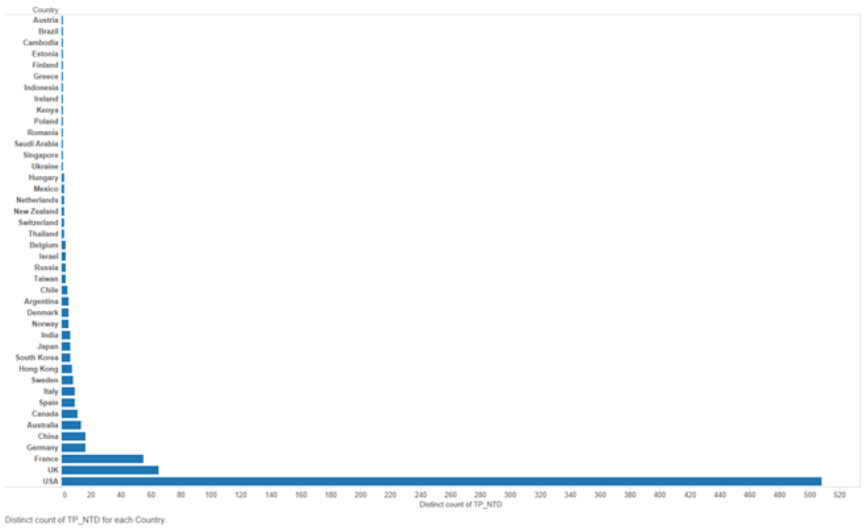
| Caring | War | Drama | Action | Horror | Suspense | Love | Fiction | Adventure | Inspiration | Comedy | Fantacy | Animation | Crime | Documentar | Musical | History | Rating_G | Rating_PG | Rating_PG-1 | Rating_R | TP_NTD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2370000 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 4510000 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 8840000 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1.3E+07 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1.3E+07 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1E+07 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 140000 |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 9510000 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 410000 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1.6E+07 |

2. Data visualization:

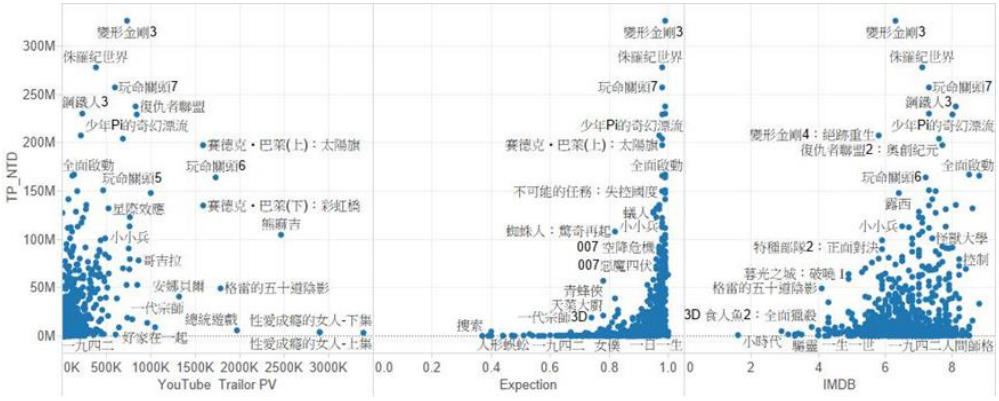
First, we found small positive correlation between box office revenues in Taipei and in US. In addition, the visualization of DF and TP_NTD shows that, when DF is getting close to 0, the TP_NTD is increasing, which means that a movie will has better box office revenues if its released days in US and Taipei are near or even the same.



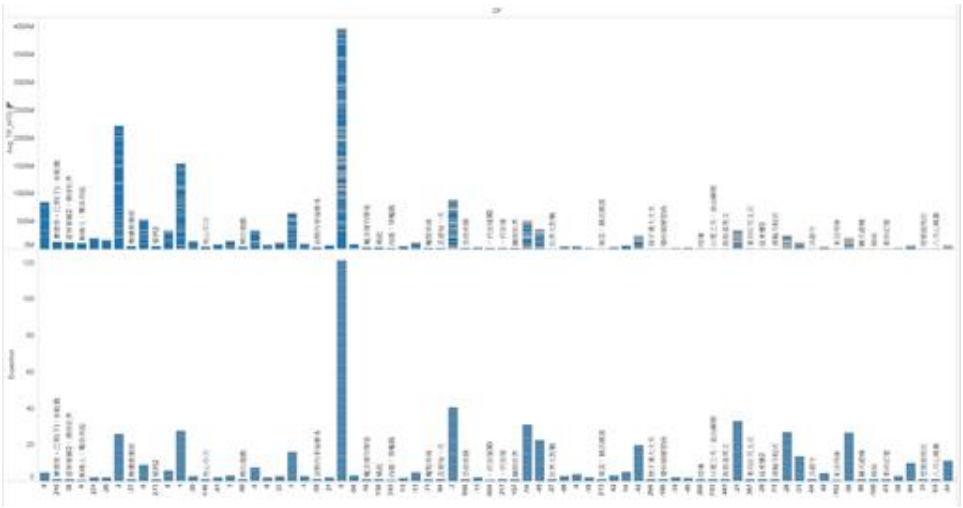
Second, we found that over 60% of the movie country in our dataset are US. Hence we decided not to use Country in our prediction.




Third, there's almost no correlation among TP_NTD with IMDB, Youtube PV and Expectation.

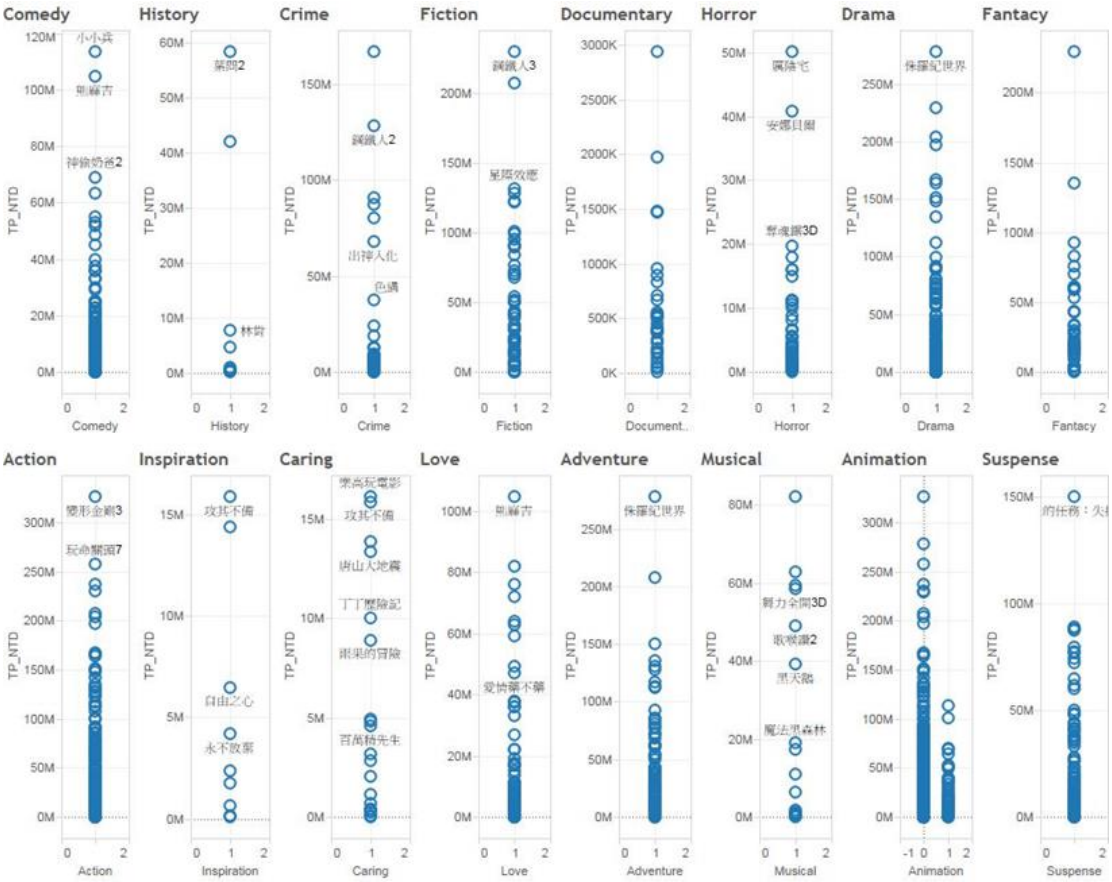



To understand more, we do the further exploration on Expectation, TP_NTD and DF, found that the better expectation comes with better box office revenue.



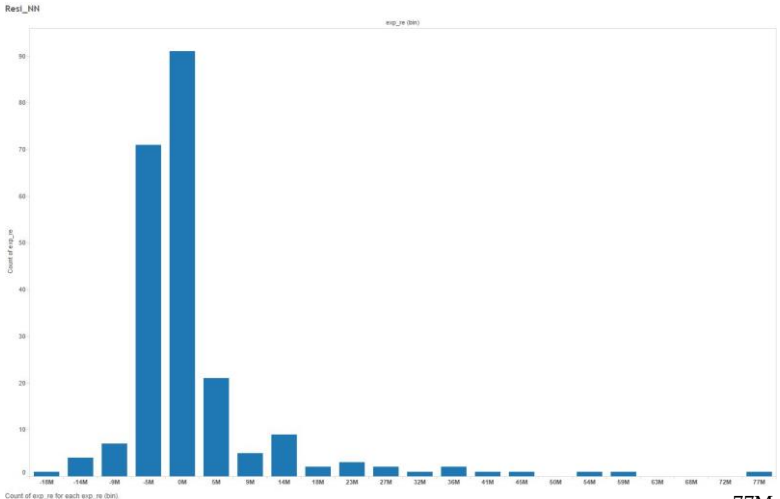
In addition, we found July is the month which has the greatest box office revenues in Taipei, however, in the mean time, it's also the month that release least movies.


On the other hand, we also visualized the relationship between TP_NTD and movie types to see whether there's correlation within these variables. We found out Action, Adventure, Animation movies tend to have better performance in box office revenues, which may implies the movie preference of people in Taipei.



3. Neural Network Results

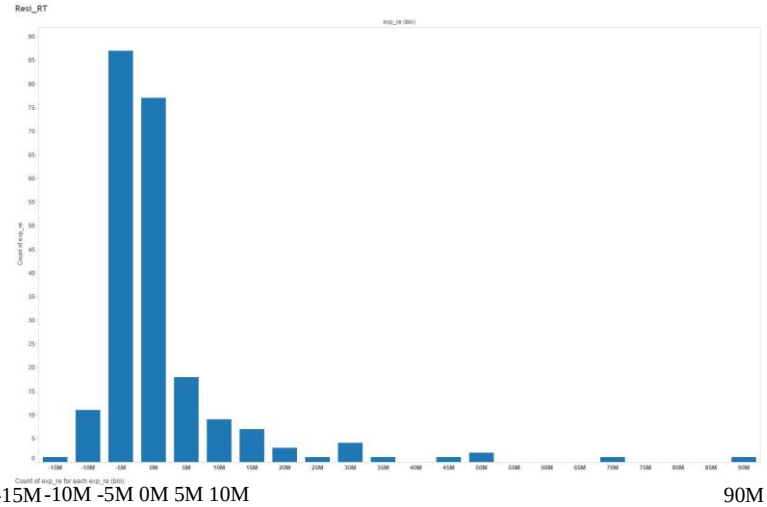
| Total sum of squared errors | RMS Error | Average Error |
|---|---|---|
| 3.10888E+16 | 11780879.4 | 3698755.059 |




-18M -9M -5M 0M 5M 9M 77M

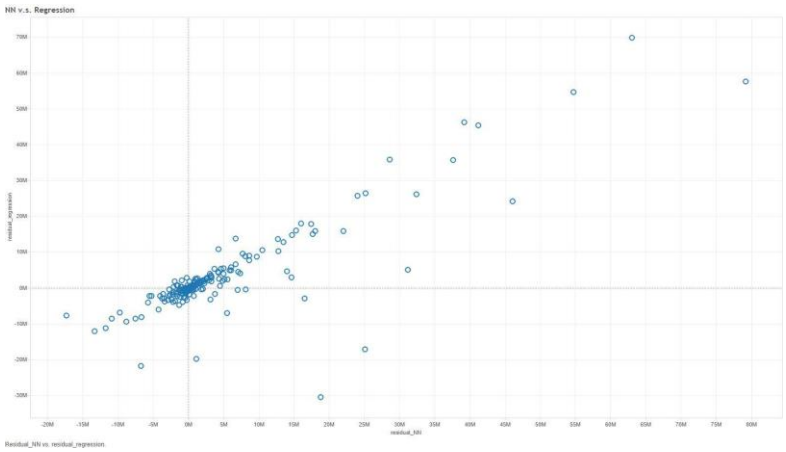
4. Regression Trees Results

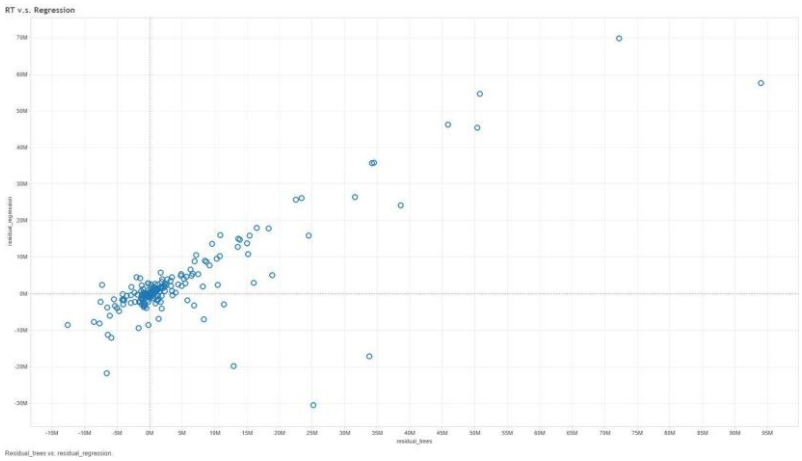
| Total sum of squared errors | RMS Error | Average Error |
|---|---|---|
| 3.48641E+16 | 12475699.7 | 3834408.328 |

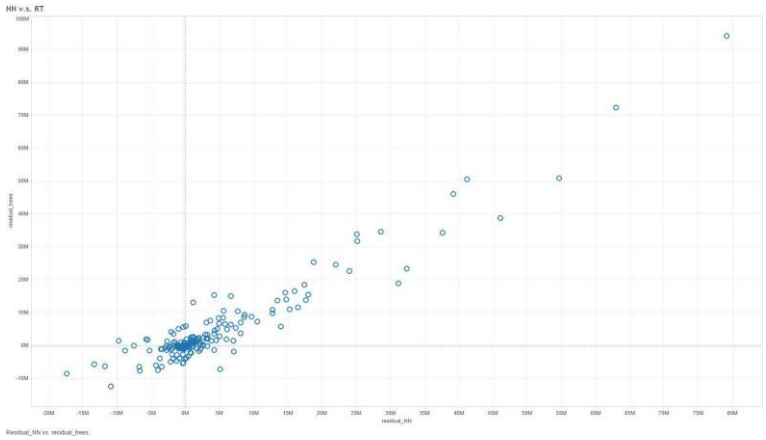



-15M -10M -5M 0M 5M 10M 90M

5. Correlation Check for Ensemble






## 6. Test Data

| Name_CN | Name_EN | Predicted | Exp | Current | Date_TW | Length | Director | Cast |
|---|---|---|---|---|---|---|---|---|
| 史努比 | A Peanuts Mo | 15.46 | 5,188,100 | 16,000,000 | 12/24/20 | 88 | 《冰原歷險記4：棒 |
| 紐約愛未 | Before We Go | 13.35 | 625,993 | 1,420,000 | 12/24/20 | 89 | 克里斯伊 | 《美國隊 |
| 真相急先 | Truth | 14.79 | 2,648,649 | 5,540,000 | 12/24/20 | 125 | 詹姆斯范 | 《藍色茉 |
| 翻轉幸福 | Joy | 17.40 | 36,085,718 | 1,830,000 | 12/31/20 | 124 | 《派特的 | 《飢餓遊 |
| 家有兩個 | DADDY'S HOM | 16.42 | 13,467,908 | 9,480,000 | 12/31/20 | 96 | 《老闆不 | 《官賤對 |
| 怪物遊獸 | Goosebumps | 15.74 | 6,848,908 | 11,000,000 | 12/31/20 | 103 | 《鯊魚黑 | 《格列佛 |
| 神鬼獵人 | The Revenant | 16.69 | 17,732,292 | 11,000,000 | 1/8/2016 | 151 | 阿利安卓 | 李奧納多 |
| 瞎趴姊妹 | Sisters | 16.21 | 10,917,346 | 780,000 | 1/8/2016 | 118 | 《歌喉讚 | 《愛在頭 |
| 女權之聲 | Suffragette | 15.01 | 3,315,551 | 340,000 | 1/8/2016 | 106 | 莎拉賈芙 | 《大亨小 |
| 45年 | 45 Years | 12.79 | 358,801 | | 1/15/201 | 95 | 《愛在週 | 《里斯本 |
| 大賣空 | The Big Short | 15.87 | 7,783,085 | | 1/15/201 | 130 | 《銀幕大 | 《黑暗騎 |
| 史帝夫賈 | Steve Jobs | 15.25 | 4,212,680 | | 1/22/201 | 122 | 《貧民百 | 《X戰警 |
| 鼠來寶：D | Alvin and The | 15.43 | 5,016,163 | | 1/22/201 | 86 | 《荒野大 | 傑森李/賈 |
| 恐龍當家 | The Good Din | 15.04 | 3,413,257 | | 2/5/2016 | 93 | Peter Soh | (配音)茱 |
| 扣押幸福 | Freeheld | 12.41 | 244,745 | | 2/19/201 | 104 | 《愛情無 | 《我想念 |
| 驚爆焦點 | Spotlight | 14.29 | 1,602,435 | | 2/19/201 | 128 | 《幸福來 | 《鳥人》 |
| 八惡人 | The Hateful E | 15.18 | 3,908,780 | | 2/19/201 | 182 | 昆汀塔倫 | 山繆傑克 |
| 丹麥女孩 | The Danish Gi | 15.28 | 4,332,077 | | 3/4/2016 | 120 | 《王者之 | 《愛的萬 |