Movie Revenue Prediction Using Regression and Clustering

Pakorn Walanaraya
Department of Industrial Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
pakornwalan@gmail.com

Weerapat Puengpipattrakul
Department of Industrial Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
weerapatp@gmail.com

Daricha Sutivong
Department of Industrial Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
daricha.s@chula.ac.th

***Abstract*—Among many movies that have been released, some generate high profit while the others do not. This paper studies the relationship between movie factors and its revenue and build prediction models. Besides analysis on aggregate data, we also divide data into groups using different methods and compare accuracy across these techniques as well as explore whether clustering techniques could help improve accuracy. Specifically, two major steps were employed. Initially, linear regression, polynomial regression and support vector regression (SVR) were applied on the entire movie data to predict the movie revenue. Then, clustering techniques, such as by genre, using Expectation Maximization (EM) and using K-means were applied to divide data into groups before regression analyses are executed. To compare accuracy among different techniques, R-square and the root-mean-square error (RMSE) were used as a performance indicator. Our study shows that generally linear regression without clustering offers the model with the highest R-square, while linear regression with EM clustering yields the lowest RMSE.**

Keywords-component: regression; predictive analytics; Clustering; Expectation-maximization; K-means; Movies

I. INTRODUCTION

The income of film industry comes from screening movie in the theater, which is called "Box-Office". Film industry is a highly competitive industry. Many new movies queue up to be released each week, so a theater owner has to decide on which movie to be shown, based mainly on revenue.

Regression analysis is a widely-used technique to predict revenue. This paper aims to compare accuracy among various types of regression analysis, with and without clustering techniques. Specifically, for regression analysis, we used three different types of regression to create prediction models, which are linear regression, polynomial regression, and support vector regression. As clustering into smaller groups of similar items may improve accuracy of the prediction models, we cluster the movie data and apply regression analysis on each cluster. Three clustering techniques are explored, namely clustering by movie genre, K-mean clustering and Expectation Maximization clustering. In order to compare performance across models, k-fold cross validation technique was employed to divide the data in two groups: training and testing data. The training data was used to create the prediction model, while the testing data was used to test accuracy of the regression models. R-square and root mean square error (RMSE) were used as performance indicators of the models.

II. RELATED WORKS

There exist many research works attempting to predict movie revenue using variations on the regression models, such as different techniques or factors, in order to investigate which approach could generate the highest accuracy. To perform linear regression, many used classic factors such as cast, producer, director and genre, while some employed other additional different factors. For example, in [1], classic factors were used along with social media data such as Facebook or Twitter to predict movie revenue. In addition, [2] used additional factors, such as Motion Picture Association of America (MPAA) rating, the number of screens and holiday-released date, while [3] proposed a model using MPAA rating and criticism from audiences to build a regression model for movie with over 50,000 dollars in revenue. Moreover, a research in China showed that directors had more influence on movie revenue than actors by using multiple linear regression [4]. To improve accuracy, [5] showed that doing regression after clustering based on budget and the number of theaters that show the movie could decrease error. However, [6] found that linear regression might not always work well because the dataset could be too small to generate an accurate model. Therefore, [7] used polynomial regression along with classic factors to predict movie revenue and discovered that a higher degree in equation might lead to more error. Support vector regression (SVR), one of non-linear regression methods, was shown to provide higher accuracy when compared with linear regression, ridge regression and logistic regression [8,9].

In [2], clustering by Expectation Maximization (EM) method was used to divide data into groups based on the number of theaters that showed a movie in the first week. The regression method was then applied to predict revenue of movie in each group. This study found that applying EM clustering onto data before doing regression decreased the prediction error. To consider influence of movie stars and directors on revenue, [4] used the number of movie star appearances as a parameter for each movie star called "star power" and averaged the star powers between leading actor and actress for each movie. In [6], the same approach was

978-1-5386-6107-9/18/$31.00 ©2018 IEEE

applied to directors of the movies. Furthermore, the Oscar Award was used to represent the influence of the movie stars in [10].

III. Methodology and proposed models

This paper aims to study and compare various methods for movie revenue prediction. We develop the prediction models using its related factors by applying different kinds of regression analysis. Moreover, this study also explores the concept of grouping data using different techniques before performing regression to study whether the approach can improve prediction accuracy.

This section explains our proposed approach and models in four major steps. First, data source and preparation are described. Second, we start developing models by applying three types of regression, i.e. linear regression, polynomial regression, and SVR, onto the aggregate cleaned data. Third, we take the same dataset and divide data into groups using three methods, i.e. by movie genres, using EM clustering, and K-means clustering. In addition to EM clustering, which was found to produce good results [2], we have added grouping by genre and K-means clustering, a commonly used clustering technique, for further technique comparison. Finally, we examine the model performance using R-square and RMSE.

A. Data source and preparation

In this study, the data used for analysis comes from a public database at https://www.kaggle.com/tmdb/tmdb-movie-metadata/data.

We first examined the data and eliminated attributes which are not useful for our analysis, such as the movie's homepage, the movie's id, etc. We separated 10 remaining attributes into two major types: numerical attributes and non-numerical attributes, as shown in Table I.

TABLE I. Movie data attributes

Numerical attributes (5)
Budget, Revenue, Vote Average, Vote Count, Runtime
Non-numerical attributes (5)
Genres, Spoken Language, Production Companies, Release Date, Cast

Our original dataset contains 4804 movies. We then removed the movie data with missing values because accuracy may decrease if we use incomplete data to perform regression. We also removed the movie data with total revenue lower than 100,000 dollars as they were hard to predict [3] and were insignificant in our case. After data cleansing, 3121 movies were remained for analysis.

Next, we considered both major types of attributes. Numerical attributes were ready for the analysis in the next step, but non-numerical attributes were not. Therefore, we converted these non-numerical attributes to numerical attributes to be used for regression.

For Genres, Spoken Languages, and Production Companies, we used binary variables for each value of these attributes. For example, Aladdin's genre is Adventure and Romance, its Spoken Language is English, and its Production Company is Walt Disney Pictures, the data values were converted as shown in an example in Table II.

TABLE II. Example Of conversion from Non-numerical To Numerical ATTRIBUTES

	Aladdin	**Now You See Me 2**	**Johnny English Reborn**
Budget ($M)	28	90	45
Revenue ($M)	504	335	160
Vote Average	7.4	6.7	6
Vote Count	3416	3235	1007
Runtime (min)	90	129	101
Genres			
Adventure	1	1	1
Action	0	1	1
Romance	1	0	0
Spoken Language			
English	1	1	1
Chinese	0	1	1
Production Companies			
Walt Disney Pictures	1	0	0
Summit Entertainment	0	1	0
Universal Pictures	0	0	1

From the Released Date attribute, we compute its day of the week, call it Day of Week attribute and use binary variables for each day of week in order to explore whether there exists a relationship between a movie's day of week of the release date and its revenue. For example, we convert 14 September 1995 to Thursday.

TABLE III. Example Of Prepared Data

	Aladdin	**Now You See Me 2**	**Johnny English Reborn**
Budget($M)	28	90	45
Revenue ($M)	504	335	160
Vote Average	7.4	6.7	6
Vote Count	3416	3235	1007
Runtime(min)	90	129	101
Star Power	21	16	6
Genres			
Adventure	1	1	1
Action	0	1	1
Romance	1	0	0
Spoken Language			
English	1	1	1
Chinese	0	1	1
Production Companies			
Walt Disney Pictures	1	0	0
Summit Entertainment	0	1	0
Universal Pictures	0	0	1
Day of week			
Wed	1	0	0
Thu	0	1	1

For the Cast attribute, we convert it to numerical attribute using star power. Our star power value of a movie is a sum of star power values of two leading cast members. Star power value of a cast member is calculated by counting the number of occurrences of the star in the dataset. For example, Aladdin's leading casts are Robin Williams and Scott Weinger. Robin Williams has star power value of 20 because there are 20 Robin Williams movies in this dataset, and Scott Weinger has star power of 1 by the same analogy. Therefore, the star power value of Aladdin is 21. An example of completely prepared data is shown in Table III.

B. Aggregate data analysis

After the dataset was fully prepared, it is ready for regression analysis. In an aggregate data analysis, different kinds of regression were applied to the entire movie data. Our study explores three types of regression analysis:

- **Linear regression**. The objective of linear regression is to find the response value by generating a linear equation as a function of predictor variables, as shown in equation (1)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \quad (1)$$
where y denotes the response variable, which is Revenue in our study. β_0 denotes a constant; β_i denote co-efficients of predictor variables x_i, such as Budget, Vote Average, etc.
- **Polynomial regression.** This non-linear regression models relationship between dependent variable and independent variables that may be non-linear, as shown in equation (2)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \beta_i x_i^n \quad (2)$$
where y denotes the response variable, which is Revenue in our study. β_0 denotes a constant; β_i denote co-efficients of predictor variables x_i, such as Budget, Vote Average, etc.
- **Support vector regression (SVR).** This regression method helps alleviate an overfitting problem of the regression method and may increase accuracy of the model. Support vector regression selects the widest gap between data points and create a hyperplane to separate data into two groups. Hyperplane equation is shown in equation (3)
$$\vec{w} \cdot \vec{x} - b = 0 \quad (3)$$
where $\vec{w}$ denotes a normal vector of the hyperplane; $\vec{x}$ denotes a data point of any dimension; b denotes a constant.

C. Group-based analysis

Our study also investigates whether grouping data before applying the regression onto each group will yield a higher accuracy than applying the regression directly to all data together. To achieve this goal, we proposed three models to group data. Specifically, we proposed segmenting movie data by its genre, as revenue of movie type may be driven by different factors. Moreover, we proposed grouping data using two different clustering techniques, namely EM clustering and K-means clustering. EM clustering has been shown to produce a good prediction [2], while K-means clustering is a popular clustering tool in data mining.

Before applying the two clustering techniques, the data was normalized in order to adjust values of different scales into a common scale. For example, Budget value could be more than a million whereas Vote Average value would be less than 10. Therefore, the data were normalized so that the clustering techniques can work across various dimensions or attributes. The normalization computation is shown in equation (4)
$$z = \frac{x - \mu}{\sigma} \quad (4)$$
where z denotes a normalized value of an attribute; x denotes an actual value of an attribute; µ denotes an average value of an attribute; σ denotes a standard deviation of the attribute value.

The three proposed grouping techniques are as follows:

- ***Grouping movie by their genres***. There are 18 movie genres, including Action, Adventure, etc., in this dataset, so there are a total of 18 groups. A movie may belonged to several genres; therefore, a movie and its data may appeared in several groups.
- ***Using Expectation Maximization clustering (EM).*** This technique is known to cluster non-linear data more suitably and leads to a higher prediction accuracy. The technique assigns data to clusters probabilistically. Initially, the value of k was specified by trial and error, such as 2, 3, etc. Then the mathematical function for each cluster was generated to calculate the logarithm of the probability that any data belongs to this cluster (log-likelihood function). This step is called the "expectation" step. The data is then assigned to the cluster so that a log-likelihood function of every cluster was maximized. This step is called the "maximization" step. These two steps were done iteratively until the log-probability values were maximized. In our study, all numerical attributes, namely Budget, Revenue, Vote Average, Vote Count, and Runtime, were used together as the factors for clustering.
- ***Using K-means clustering.*** This technique is one of the most popular and commonly used technique for clustering. The algorithm is quite simple. The value of k was specified by trial and error, such as 2, 3, etc. Then, k observations were randomly selected as centroid points for each cluster. Each data point was assigned to the cluster whose centroid is nearest to that data point (minimum Euclidean distance). The new centroid was then calculated from the current members of each cluster. These processes were done iteratively until the total of Euclidean distance between data points and their clusters' centroid was minimized, and the clusters' centroids converge. This technique assigns data to clusters deterministically, contrary to the EM technique. Again, all numerical attributes were used as the factors for clustering.

After all data points were assigned into groups, linear regression, polynomial regression, and support vector regression were then applied to each group to generate a prediction.

D. Performance Comparison

We compared the model accuracy among various types of regression and among various grouping techniques. The performance indicators used were R-square and RMSE. R-square is a value which indicates how well the model explains variation. The R-square computation is shown in equation (5).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (5)$$

where f_i denote predicted values (predicted revenue); y_i denote actual values (actual revenue); $\bar{y}$ denotes an average of the actual revenue.

RMSE represents the difference between the predicted values and the actual values. The RMSE computation is shown in equation (6)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (f_i - y_i)^2}{n}} \qquad (6)$$

where f_i denote predicted values (predicted revenue); y_i denote actual values (actual revenue); n denotes the number of data points considered.

IV. RESULTS AND PERFORMANCE ANALYSIS

In order to reliably compare performance among models, 5-fold cross validation technique was applied to the dataset in order to reduce the overfitting of the prediction model. To perform 5-fold cross validation technique, we randomly splitted data into five equal-size groups. Four groups serve as training data and the other group serves as testing data. Once the regression was performed with corresponding R-square and RMSE, training and testing data were changed. The procedure was repeated until every group has been used as a testing data set. We used an average of five R-square values and five RMSE as performance indicators of each model.

A. Aggregate data analysis

For the aggregate data analysis, the R-square and RMSE of linear regression, polynomial regression, and SVR are shown in Table IV. Noted that for the aggregate data analysis, where all the movie data were used to generate a model for each type of regression, the best technique is the linear regression. Its R-square is the highest at 69.7%, and its RMSE is the lowest at 101.43 million dollars. In terms of R-square, polynomial regression performs the worst, while SVR's RMSE is the largest.

TABLE IV. R-SQUARE AND RMSE VALUES OF EACH REGRESSION TECHNIQUE ON AGGREGATE DATA

Regression	**R-square (%)**	**RMSE (M$)**
Linear regression	69.7	101.43
Polynomial regression	53.2 (maximum degree = 7)	127.41
SVR	60.5	194.16

B. Group-based analysis by movie genre

In this section, we divided movies into groups by its genre. With 18 genres of movies, the R-square and RMSE results were shown in Table V and Table VI, respectively. The results of each regression type is the weighted average according to the number of data points in each genre.

TABLE V. R-SQUARE VALUES (%) OF GENRE-BASED ANALYSIS

Genres	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
Action	902	**70.9**	48.9 (2)	61.3
Adventure	653	**72.6**	50.6 (8)	66.4
Animation	185	**63.2**	48.3 (10)	55.7
Comedy	1068	**69.3**	51.7 (7)	60
Crime	510	**67.1**	55.6 (2)	55.7
Documentary	31	**85.4**	58.6 (9)	65.1
Drama	1382	**64.8**	50.6 (9)	57
Family	360	**68.5**	50.4 (9)	62.8
Fantasy	336	**72.2**	49.1 (8)	63.4
History	143	**61.2**	51.6 (7)	53.8
Horror	315	**51.2**	41.3 (10)	45.5
Music	109	53.3	**60.2 (10)**	50.5
Mystery	258	**71.5**	57.8 (3)	60.9
Romance	555	**65.6**	40.4 (2)	55.3
Sci-Fi	422	**67.9**	46.6 (9)	58.9
Thriller	912	**70.4**	58.3 (4)	62.1
War	116	**71.5**	62.5 (2)	59.8
Western	56	**41.7**	35.1 (2)	38.9
Average	-	**67.54**	50.81	57.39

TABLE VI. RMSE (MILLION $) OF GENRE-BASED ANALYSIS

Genres	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
Action	902	**127.22**	172.18 (2)	236.87
Adventure	653	**126.60**	171.32 (8)	244.00
Animation	185	**105.71**	188.90 (10)	188.90
Comedy	1068	**92.02**	115.58 (7)	167.91
Crime	510	**108.35**	126.03 (2)	200.89
Documentary	31	**41.36**	58511.24 (9)	113.17
Drama	1382	**92.20**	109.82 (9)	165.13
Family	360	**104.89**	134.41 (9)	190.75
Fantasy	336	**128.09**	171.40 (8)	247.07
History	143	**88.98**	111.42 (7)	140.48
Horror	315	**96.82**	146.57 (10)	146.57
Music	109	**86.31**	138.94 (10)	138.94
Mystery	258	**109.07**	136.66 (3)	204.44
Romance	555	**66.92**	109271.243 (2)	112.59
Sci-Fi	422	**113.97**	146.714 (9)	197.39
Thriller	912	**111.57**	203.39 (4)	210.68
War	116	**96.87**	117.00 (2)	179.72
Western	56	**101.44**	128.15 (2)	148.58
Average	-	**104.02**	7649.18	189.39

From Table V, the linear regression yields the highest average R-square of 67.54%. Noted that this average number is still lower than 69.7% of the linear regression on the aggregate data in the previous section. However, there are various genres that outperform the aggregate data model, such as Action, Adventure, Documentary, Fantasy, Mystery, Thriller and War.

From Table VI, the linear regression has the lowest RMSE for all movie genres. Its average is 104.02, which is slightly higher than 101.43 of the aggregate data model. Once again, certain genres yield lower RMSE than the aggregate data model, such as Comedy, Documentary, Drama, History, Horror, Music, Romance and War.

To summarize, for group-based analysis by genre, the linear regression technique provides the best accuracy, when compared with the polynomial regression and SVR. Overall, the accuracy of group-based analysis is slightly worse than the aggregate data analysis. However, dividing movies into groups based on its genres may help improve the revenue prediction accuracy in some genres. In particular, Documentary and War movies are better predicted using data of its own genre, probably due to its own nature and appeal.

C. Group-based analysis by EM clustering

In this section, we used EM to group movies into clusters and then apply regression analysis on each group. The R-square and RMSE results are shown in Table VII - XII for k = 2, 3, 4. The average results are weighted average over all clusters.

TABLE VII. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 2 CLUSTERS (K=2).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	2079	**45**	41.6 (2)	44.6
2nd	1042	56.3	**57.1 (2)**	56.2
Average	3121	**48.77**	46.77	48.47

TABLE VIII. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 2 CLUSTERS (K=2).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	2079	**33.44**	41.66 (2)	46.92
2nd	1042	**167.21**	208.81 (2)	254.94
Average	3121	**78.10**	97.47	116.37

TABLE IX. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 3 CLUSTERS (K=3).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	997	**27.6**	12.7 (3)	27
2nd	646	44.6	**45 (2)**	43.8
3rd	1478	**26**	9.7 (9)	24.5
Average	3121	**30.36**	17.96	29.29

TABLE X. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 3 CLUSTERS (K=3).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	997	**14.13**	16.52 (3)	16.76
2nd	646	**202.09**	252.21 (2)	280.03
3rd	1478	**5.31**	9.35×10^{14} (9)	61.77
Average	3121	**48.86**	4.43×10^{7}	92.57

TABLE XI. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 4 CLUSTERS (K=4).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	308	**30.8**	29.6(5)	28.9
2nd	882	**24.8**	7.2(7)	24.5
3rd	611	**69.7**	42.9(3)	41.7
4th	1320	**28.5**	13.4(2)	27.8
Average	3121	**35.74**	19.02	29.7

TABLE XII. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING EM CLUSTERING WITH 4 CLUSTERS (K=4).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	308	**29.80**	3230.86 (5)	35.47
2nd	882	**11.24**	16.32 (7)	13.12
3rd	611	**205.84**	253.49(3)	277.90
4th	1320	**51.85**	63.07 (2)	61.11
Average	3121	**68.34**	399.75	87.46

Noted that from Table VII – XII, the linear regression outperforms the polynomial regression and SVR both in terms of R-square and RMSE. From Table VII, Table IX and Table XI, noted that the average R-square values of the linear regression are 48.77%, 30.36% and 35.74% for 2, 3 and 4 clusters, respectively. All R-square values are lower than the R-square of the aggregate data model. However, from Table VIII, Table X and Table XII, RMSE are 78.10, 48.86 and 68.34 for 2, 3 and 4 clusters, respectively. RMSE are all lower than the RMSE of the aggregate data model. Therefore, from R-square perspective the aggregate data model outperforms the group-based model using EM clustering, while from RMSE perspective the group-based model using EM clustering produces lower errors in prediction.

Our results do not show a clear relationship between accuracy performance and the number of clusters. However, we have observed from the detailed cluster characteristics that the ones with high budget and high vote count tend to produce high R-square.

D. Group-based analysis by K-means clustering

In this section, we used K-means clustering technique to group movies into clusters, and then apply regression analysis on each group. The R-square and RMSE results are shown in Table XIII – XVIII for k = 2, 3, 4. The average results were weighted average over all clusters.

TABLE XIII. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 2 CLUSTERS (K=2).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	2674	**36.7**	24.2 (4)	36.6
2nd	447	**40.2**	39.6 (10)	39.5
Average	3121	**37.2**	24.41	37.02

TABLE XIV. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 2 CLUSTERS (K=2).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	2674	**57.51**	67.02 (4)	76.09
2nd	447	**216.06**	260.83 (10)	286.61
Average	3121	**80.22**	94.78	106.24

TABLE XV. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 3 CLUSTERS (K=3).

Cluster	**No. of data**	**Linear**	**Polynomial (max degree)**	**SVR**
1st	1711	**47**	40.4 (6)	38.1
2nd	261	23.7	**27.5 (7)**	22.1
3rd	1149	**31.8**	29.6 (9)	30.3
Average	3121	**39.45**	35.35	33.89

TABLE XVI. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 3 CLUSTERS (K=3).

Cluster	No. of data	Linear	Polynomial (max degree)	SVR
1st	1711	**47.44**	54.43 (6)	68.15
2nd	261	**250.25**	303.11 (7)	288.15
3rd	1149	**86.05**	7.29×10^{24}(9)	106.62
Average	3121	**78.61**	2.68×10^{24}	100.71

TABLE XVII. R-SQUARE VALUES (%) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 4 CLUSTERS (K=4).

Cluster	No. of data	Linear	Polynomial (max degree)	SVR
1st	809	**37.3**	33.6 (4)	36.6
2nd	1072	**52.7**	44.6 (2)	51.7
3rd	226	17.1	**17.9 (2)**	14.3
4th	1014	**33**	30.9 (9)	31.9
Average	3121	**39.73**	35.36	38.64

TABLE XVIII. RMSE (MILLION $) OF GROUP-BASED ANALYSIS USING K-MEANS CLUSTERING WITH 4 CLUSTERS (K=4).

Cluster	No. of data	Linear	Polynomial (max degree)	SVR
1st	809	**8.9**	3.93×10^{6}(4)	112.83
2nd	1072	**49.33**	57.42 (2)	73.62
3rd	226	**261.29**	327.87 (2)	295.60
4th	1014	**61.04**	72.89 (9)	77.61
Average	3121	**58.00**	1018769.49	101.15

Noted that from Table XIII – XVIII, the linear regression outperforms the polynomial regression and SVR, both in terms of R-square and RMSE. From Table XIII, XV and XVII, noted that the average R-square values of the linear regression are 37.2%, 39.45%, and 39.73% for 2, 3, and 4 clusters, respectively. All R-square values are lower than the R-square of the aggregate data model. However, from Table XIV, XVI and XVIII, RMSE are 80.22, 78.61, and 58.00 for 2, 3, and 4 clusters, respectively. These RMSE are all lower than the RMSE of the aggregate data model. Therefore, from R-square perspective, the aggregate data model outperforms the group-based model using K-means clustering, while from RMSE perspective the group-based model using K-means clustering produces lower errors in prediction.

Noted that from Table XIII, XV, and XVII, as the number of clusters increases, R-square value tends to increase as well. From Table XIV, XVI, and XVIII, as the number of clusters increases, RMSE tends to decrease. Therefore, the higher number of clusters tends to lead to a more accurate prediction.

From these results, we conclude that from R-square perspective the most accurate method to predict movies' revenue is to apply linear regression onto all movie data without clustering. It is shown, however, that grouping movies by its genres may improve accuracy in some genres. However, from RMSE perspective, group-based analysis using EM clustering with k = 3 produces the lowest prediction error.

V. CONCLUSION

Predicting movie revenue can be done by using various data analysis techniques. Regression analysis is one of the techniques that uses numerical data to predict the related response. In this paper, we apply various regression techniques onto movie data, with and without grouping, in order to predict the movie revenue. We found that the linear regression without applying clustering technique is the most accurate method in terms of R-square. Moreover, dividing movie data by its genre before doing regression can improve revenue prediction accuracy in some cases. However, applying clustering techniques, such as EM and K-means, can improve revenue prediction accuracy in terms of RMSE. Future work may include the integration of social factor data, such as reviews on website, into the analysis to further improve the accuracy of the prediction.

REFERENCES

[1] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, "Role of different factors in predicting movie success," *2015 International Conference on Pervasive Computing (ICPC)*, pp. 1-4, June 2015.

[2] G. He and S. Lee, "Multi-model or Single Model? A Study of Movie Box-Office Revenue Prediction," *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, Liverpool, pp. 321-325, 2015.

[3] D. Im and M. T. Nguyen, "PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET," CS229, Stanford University, Fall 2011

[4] Y. Yongbin and O. Rongzhao, "A study on the relationship among the leading actors, directors, and the box office income of a film — Based on multiple linear regression model," *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Xi'an, pp. 469-471, 2013.

[5] S. Shim and M. Pourhomayoun, "Predicting Movie Market Revenue Using Social Media Data," *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, San Diego, CA, pp. 478-484, 2017.

[6] S. Yoo, R. Kanter, D. Cummings, and A. Maas, "Predicting Movie Revenue from IMDb Data," 2011.

[7] N. Apte, M. Forssell, and A. Sidhwa, "Predicting Movie Revenue," CS229, Stanford University, December 16, 2011.

[8] C. Lee and M. Jung, "PREDICTING MOVIE SUCCESS FROM SEARCH QUERY USING SUPPORT VECTOR REGRESSION METHOD, " *International Journal of Artificial Intelligence & Applications (IJAIA)*., Vol. 7, No. 1, January 2016.

[9] B. Flora, T. Lampo, and L. Yang, "Predicting Movie Revenue from Pre-Release Data," CS229, Stanford University, December 12, 2015

[10] K. Taewan, J. S. Uk, and D. Son, "Influence of Star Power on Movie Revenue. Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology," 2016