Analysis for predicting revenue of movie

Sin Wing Leong
Hong Kong University of Science and Technology

Abstract- In this study, predicting revenue of movie is performed using linear regression, LASSO, ridge regression and tree-based method, such as decision tree and bagging. The result is that tree-based out-perform the classical linear regression especially using boosting and bagging. During the study, R^2 is a criterion for evaluating the model.

I. Data Description

The data consists of 8 types of column including the response (revenue) and 7 features which are budget, genres, keywords, production companies, release date, cast and crew. There are 3376 movies as the sample but later we split into two types of data, i.e. training data and testing data. Continuous variable is budget, year and the others are categorical predictors.

<u>Clean data</u>

revenue		budget	
Min. :	5	Min. :	0
1st Qu.:	15352895	1st Qu.:	8500000
Median :	51751835	Median :	25000000
Mean :	117031353	Mean :	38884242
3rd Qu.:	140165096	3rd Qu.:	52000000
Max. :	2787965087	Max. :	380000000

Obviously, the data consists of some irregular condition such that the budget should not be 0.
The genres includes adventure, fantasy, animation, drama, horror, action, comedy, history, western, thriller, crime, documentary, science fiction, mystery, music, romance, family, war, foreign movie. (19 types of moives)

The keys words include total 8828 types of keywords, including 'japan'.

There are totally 3759 companies in data including Columbia Picture..

Regarding cast, there are 3 gender which is 0, 1, 2 level. There are 27608 gender = 0 casting, 19265 gender = 1, 39744 gender = 2. To simplify the model, we do not include the specific person who cast the movie.

The crews include 65517 gender = 0, 9965 gender = 1, 36583 gender = 2.

The release date of movie is ranging from 1916 to 2016. Before 1999, all data contains less than 100 data points each year. After 1999, most of data point contains more than 100 data points.
For example,

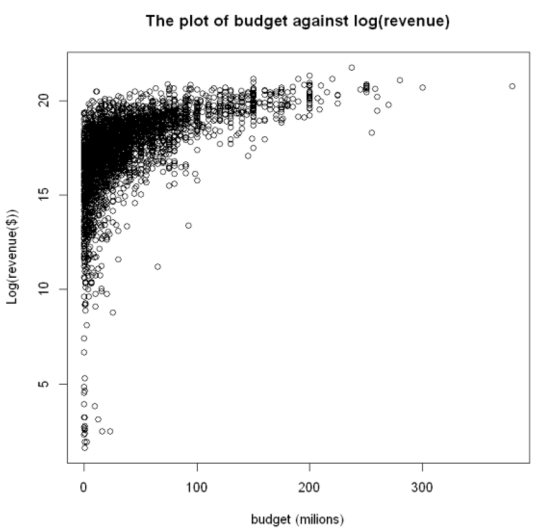
Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Count	79	91	111	109	125	135	113	143	151

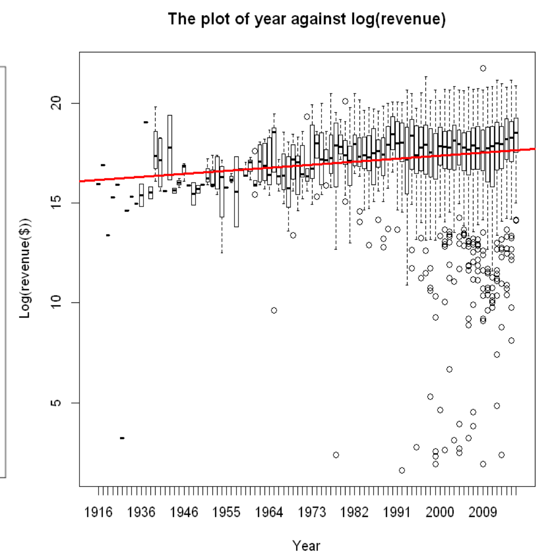
In this study, the date is split into year and seasons.

Start	End	
March 1	May 31	Season = 1 (Spring)
June 1	August 31	Season = 2 (Summer)
September 1	November 30	Season = 3 (Autumn)
December 1	February 28 (29)	Season = 4 (Winter)

Obviously, there is positive relationship between budget and logscale of revenue. However, it may not linear relationship and not very clear when low budget level. Further investigation is needed.

From the plot between revenue and year, It was found that with year increases, the variance of revenue would increase as it was observed there are huge variance after 1990.

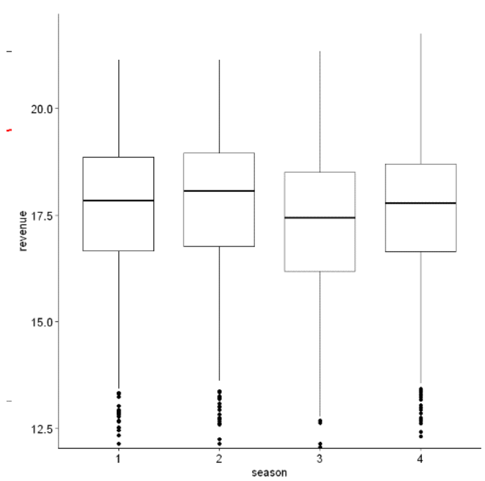




Now we have seen the relationship between budget, year and revenue.

<u>Season effect</u>

We will investigate the season effect. The class is 4 seasons and response is logscale of revenue.



In boxplot with x-axis = season and y-axis = revenue, it is hard to see whether the season effect is significant or not. Therefore, we use statistical model to test.

$$\begin{cases} H_0: \text{mean of each group is the same} \\ H_1: \text{two of groups are different} \end{cases}$$

	Df	F value	Pr(>F)
group	3	0.8758729	0.4527784

By Levene's test, There is no significant different variance between group(seasons). We do not reject equal variance between group.

However,

```
	Shapiro-Wilk normality test

data:  aov_residuals
W = 0.83353, p-value < 2.2e-16
```

By Shapiro-Wilk normality test, the distribution of revenue in each group is significant different from normal distribution. Therefore, we cannot use ANOVA model as the normality assumption is invalid.

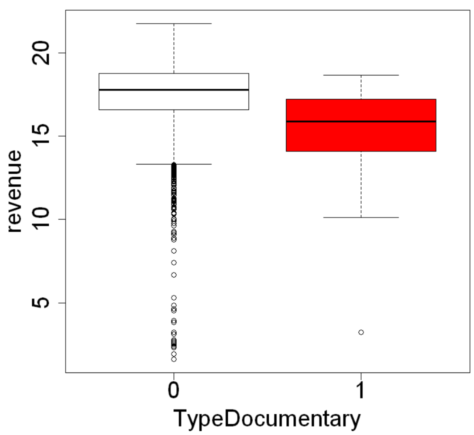
Therefore, Kruskal-Wallis test will be used.

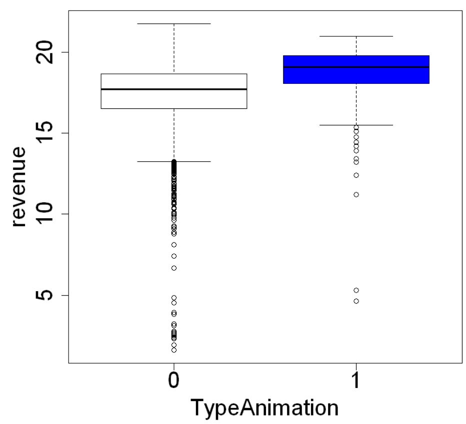
```
        Kruskal-Wallis rank sum test

data:  revenue by season
Kruskal-Wallis chi-squared = 53.119, df = 3, p-value = 1.729e-11
```

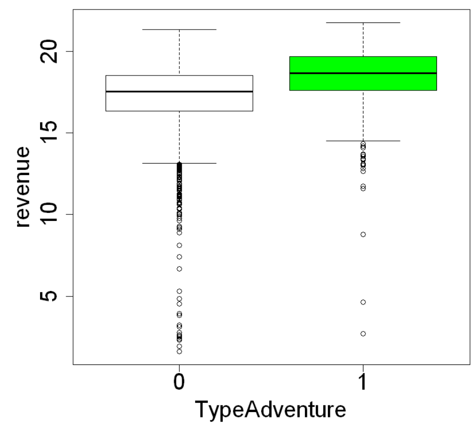
By Kruskal-Wallis test, p-value < 0.01 so that we reject the equality of mean for different group. Therefore, The season effect is significant. Therefore, we split into 4 seasons is statistically reasonable.

Genres Effect

In common sense, there are different revenue for different types of movies. Therefore before fitting regression or using tree-based method, investigating genres effect is essential. As for each movie, there are multiple in one movie so dummy variables are created with one rows possibly more than 1 type of movie = 1. Therefore, they are not mutually independent for each group(type). ANOVA and Kruskal-Wallis test are not available. Just use Boxplot now.





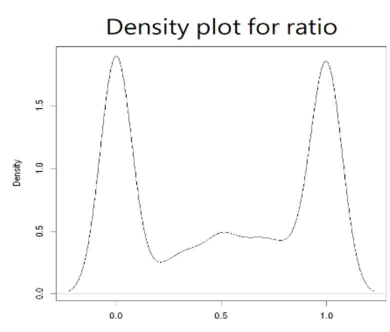


If we ignore interaction with other categorical variables, such as season, we can easily see that the mean of documentary is significantly lower than non-documentary and Animations are higher than non-animations. As the revenue is logscale, their difference may more than multiple 2. Therefore, genres effect is possible to add into the tree-based model or linear regression depending on the interaction effect with the other categorical variables.

Company

To categorize the company data, we analyze the mean of log revenue in each genres and calculate whether the company can earn more than the mean in the one genres. i.e. for each company,

name	id	above	below	ratio
Ingenious Film Partners	289	59	19	0.7564103
Twentieth Century Fox Film Corporation	306	436	154	0.7389831
Dune Entertainment	444	124	25	0.8322148
Lightstorm Entertainment	574	15	1	0.9375000
Walt Disney Pictures	2	264	56	0.8250000
Jerry Bruckheimer Films	130	56	5	0.9180328
Second Mate Productions	19936	6	0	1.0000000



'Above' is the amount of companies production that revenue is higher than mean of movie in that genre.

Ratio = simply(above/(above+below)) to estimate the probability that company produces one movie higher than mean in that genres in general. As density plot mentioned above, we may see huge difference. Therefore, we divided 4 groups in companies with [0,0.25], [0.25,0.5], [0.5,0.75], [0.75,1].

As there lots of categories for casts, crews and keywords. In linear regression (including shrinkage or not), we just consider three categorical variables- genres, seasons and companies in order to avoid overfitting.

Keywords labelling

It is hard to categories those keywords instead of using NLP method. However, It can be found by K-means clustering.

id	name	meanBudget	meanLogRevenue	Count
236	suicide	27951358	16.82973	37
392	england	33457209	18.20438	25
657	fire	52493324	17.80522	15
1655	country house	28575000	17.73732	4
1879	shower	24622455	16.74864	7

In 3000 training data, there are 8225 keywords. For example, the frequency of appearing 'suicide' is 37 and when it appears, mean of budget is 27951358 and log revenue = 16.82973.
(K =3) so we set 3 levels in the model. Those testing keywords did not appear in training data is not counted in the model. This rules also apply for company variables.
For example,

level1	level2	level3
0	0	0
0	16	0
0	4	0
0	2	0

There is no keywords data in first row and there are totally 16 cluster 2 keywords in movie.

Full training data (emit some of data in display for convenience)

	revenue	budget	TypeAction	TypeAdventure	TypeWestern	year	season	company1	company2	company3	company4	level1	level2	level3
2463	2.397895	11	1	0	0	1978	1	1	0	1	0	0	0	0
2511	17.977333	10000000	0	0	0	2016	4	1	0	2	2	0	16	0
2227	16.653429	15000000	1	0	0	1984	1	1	0	1	0	0	4	0

Company and keywords level are considered as continuous variable for simplified. Season and Type of movies are categorical variables. We set year as continuous variables as if we treat it is categorical variable, the data will be over-fitted as each group contains very few points if we split further to year-season-type. In linear regression, we only consider the interaction term between type and seasons.

II. Predictive Modelling

In this study, we are investigating various method using R^2 .

Linear regression

$$Y = X^T\beta + error\ term$$

Where X is design matrix, β is the coefficient obtained by training data.
Linear regression is a classical method to predict response (revenue). Here, we used least squared method to obtain those coefficients.

Budget only	Variable	R^2 in training set	R^2 in testing set
	Budget	0.2619	0.3423604

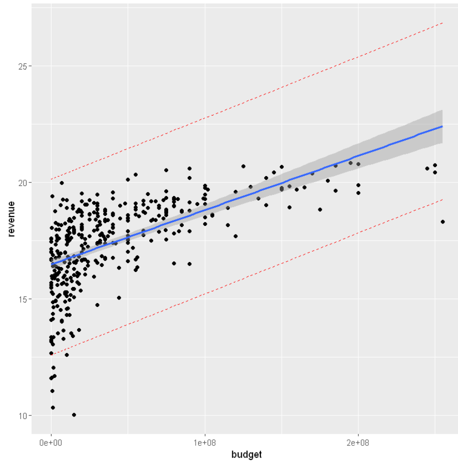


Figure 1 Regression line with 95% confidence interval (error term follow normal distribution)

, where the data point is from testing data set.
Assume There is no interaction term between genres of movie (too complicated and those effect may not be significant)

Full Model	Variable	R^2 in training set	R^2 in testing set
	Full variable with interaction (season: Type of movie)	0.3474	0.3582995

It reveals that even though the predictor becomes more, R^2 cannot be improved much. It may exist overfitting.
According to outlier test, there are 10 suspected outliers in the regression model. We kick it out first.
Using AIC as criteria, model selection (Stepwise regression) is performed.

$$\begin{aligned} revenue = \ & budget + TypeAnimation + TypeDocumentary + TypeDrama + TypeFamily + TypeForeign \\ & + TypeHistory + TypeMusic + TypeRomance + TypeScienceFiction + TypeWar + TypeWestern \\ & + year + season + company1 + company3 + company4 + level1 + level2 + level3 \\ & + TypeDocumentary{:}\,season \end{aligned}$$

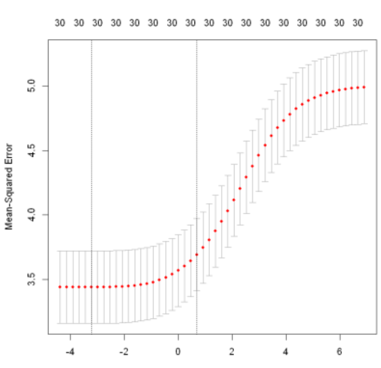
AIC selected model	Variable	R^2 in training set	R^2 in testing set
	Above mentioned	0.3323	0.3652548

$$\begin{aligned} revenue = \ & \log(budget) + TypeAnimation + TypeDocumentary + TypeDrama + TypeFamily + TypeForeign \\ & + TypeHistory + TypeMusic + TypeRomance + TypeScienceFiction + TypeWar + TypeWestern \\ & + year + season + company1 + company3 + company4 + level1 + level2 + level3 \\ & + TypeDocumentary{:}\,season \end{aligned}$$

By log transforming budget(Box-Tidwell transformation),

AIC selected model	Variable	R^2 in training set	R^2 in testing set
	Above mentioned but Log(budget)	0.3456	0.3232799

'As R^2 is larger when log(budget) in testing data, we do not consider this model.



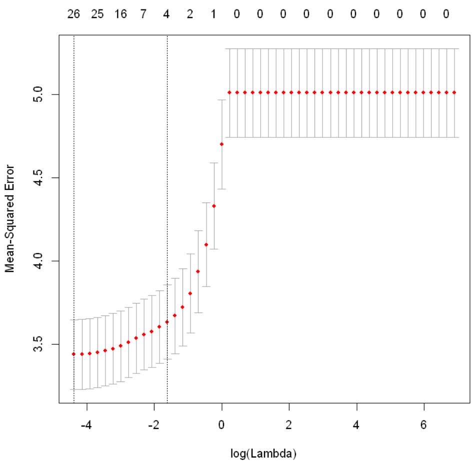
Ridge regression

Using ridge regression, there are one more hyperparameter than linear regression which is λ. Then we use Cross Validation to select this hyperparameter. (interaction term does not include)

Here, MSE is shown according to log(lambda). (by the figure on the left, log(λ) vs MSE)
Therefore, λ =0.03162is selected by CV
In testing data set, ridge regression' R^2 is 0.3854971

Lasso regression

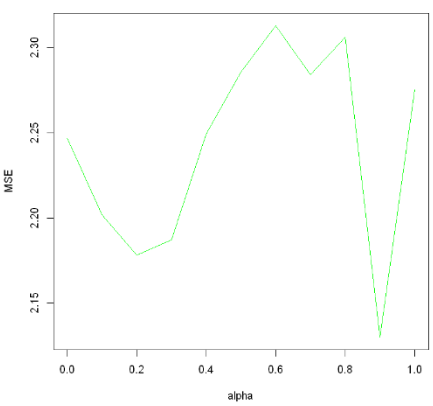
Lasso regression is similar to ridge regression but the shrinkage is using L_1 penalty, i.e $|\beta| < t$. Using full model (ignore interaction also first.)



Here, MSE is shown according to log(lambda). (by the figure on the left, log(λ) vs MSE)

Therefore, λ = 0.0125892541179417 is selected by CV.
In testing data set, ridge regression' R^2 is 0.3959779, higher than ridge regression a little.

Elastic Net



$$\lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2+\alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

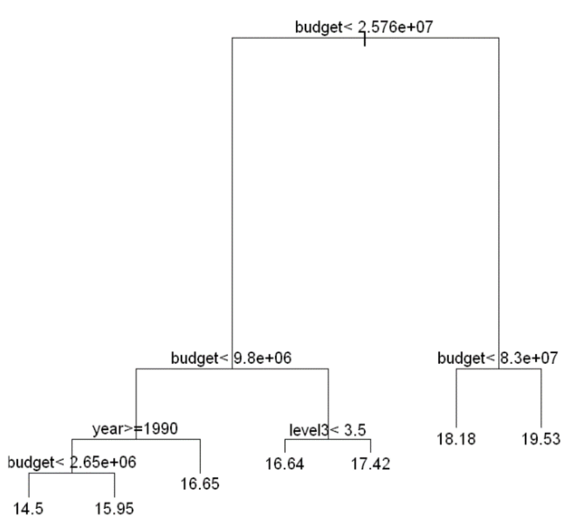
Elastic net combined Lasso and Ridge regression, where alpha is considered as the weighing we used LASSO or ridge regression Here we use some alpha to elaborate the model.

From the graph on the left, which x-axis is α and y-axis is Mean Square Error in testing set respectively, it is found that there is not specific pattern for finding optimal α.

In the aspect of R square, it is 0.40037 if α = 0.9 (near LASSO regression)

Tree based Method

After removing the weird budget data which is near 0, we obtained decision tree.



$$\hat{f}(X)=\sum_{m=1}^{3}c_m I(X_1,X_2)\in R_m$$

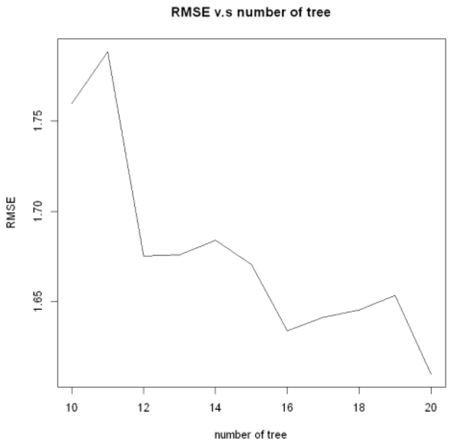
Regression tree is used.

Decision tree is a easy interpreting model. From the tree on the left, it easily classified that budget and year are mainly used for predicting revenue.

R^2 in testing dataset is 0.4647382. It is much higher than the linear regression model, even Elastic Net model with α = 0.9

Analysis of decision tree using RMS

Even though decision tree is very easy and higher R^2. However, the variance of one decision tree is very high as if the training data set is changed, the testing error will be varied a lot.



Using Bagging method to estimate bagging root mean squared error, we found that root mean squared error is decreasing respect to the number of the tree.

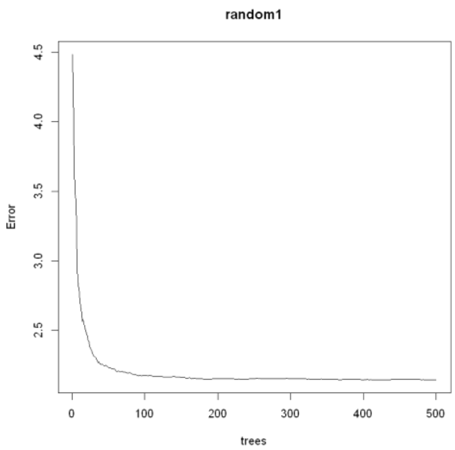
As it is out of bagged (OOB) estimation in training data set, it should be unbiased estimator of the testing error. Therefore, we expected number of trees increases to achieve better prediction.

Using 10-fold cross validation, we obtained optimal N-tree model. The R^2 result is 0.49835.

However, bagging relied on low correlation between variables. Therefore, random forest is also implement.

Random Forest

The variable is selected randomly. In our cases, 500 trees are chosen and each trial we split to 10 variables each time.



Random Forest is one of bagging methods.
Error in training data set is decreasing according to the number of trees.
The result is consistent with bagging method in decision tree.

In testing data set, the R^2 result is 0.5678786, which is also higher than standard bagging method and linear regression method..

Gradient Boosting

Gradient boosting is determined by the number of trees and essentially the loss function is important.
Here, Gaussian distribution for selecting loss function is used. Again, K-fold cross validation is used in the model for selection of trees. K = 10 is used.
A gradient boosted model with gaussian loss function. 1000 iterations (maximum)were performed. The best cross-validation iteration was 401.
R^2 in test data is 0.55095 which is near random forest performance.

To evaluate the K in cross validation,

K	4	6	8	10	12	14	16	18	20
Optimal tree	354	343	265	276	351	326	264	289	296
R^2	0.5565154	0.5652	0.5601	0.5509	0.5581	0.5629	0.5621	0.5637	0.5655

There is no significant difference from different K value in Cross Validation.

III. Discussion

Predicting the revenue using linear regression generally is worse than tree based method. It may related to how we define the group for companies and type of key words. Clustering key words and companies are significantly impact our final model. Therefore, if the method of clustering is changed using other methods, such as LDA, Logistic regression, QDA. In addition, the predictor for clustering may not be easily budget and revenue itself. The interaction between companies and keywords is possibly exist. For example, when Disneyland launch new movies, the keywords may not be violent. Therefore, the effect should be investigated. To improve the model, data engineering is worth to spend more time.

In aspect of modelling, the example in gradient boosting showed no matter how we use K-value for cross validation, the R-squared did not differ too much. Therefore, it is suggested that using K = 10 for lowering computing time. Bagging is one of powerful method but it still worse than gradient boosting. However random forest performed as same as gradient boosting with gaussian distribution.

In the example of model selection, most of genres is significant in the process of stepwise regression. It may indicated that genres is worth to add into predictive model. However, year (p-value is 0.08) is comparatively insignificant with budget. It may due to the positive correlation with budget which means when year is increasing, budget is also increasing. Most of the interaction term is emitted, which means in linear regression the interaction term is not very significant in modelling in aspect of AIC as a criterion. LASSO performed better than ridge and ordinary regression because LASSO tends to selection the variable which is not significant by the penalty terms. It may show that overfitting problem exist in linear regression so LASSO is better than that. Ridge regression does not contain those effects.