Predict box office revenue in China to maximize profits of movie theaters in China

Business Analytics using Data Mining

Group 1

李童宇 Lee, Tung-Yu

陳毅寰 Chen, Yi-Huan

陳鈺方 Chen, Yu-Fang

澤凡 Jevon Mckenzie

蓋亞德 Dimitri CAYARD

Summary

Business Problem:

This research is in an effort to increase the profits of movie theaters in China. It is with no doubt that the movies which a theater decide to show differ in the revenue that they contribute. With that in mind, this research aims to identify movies that are more likely to generate higher revenue because:

1. If the demand for a particular movie is too low in comparison to the supplied amount of shows and venues, the movie theater will incur a loss.
2. If the demand for a particular movie is very high in comparison to the supplied amount of shows and venues, the movie theater will not be able to maximize their revenues.

Data:

We retrieved data about information of the movies both in China and Taiwan (i.e Movie Title, type, directors, actors, released date) , the responses towards the movies from Taiwanese customers (i.e # of total comments, # of actual rating, # of expected rating, actual rate, expected rate) , and the box office revenue of both countries.

Analytics solution:

We used the information about movies in Taiwan to predict the box office revenue of movies which will be released in China. Data visualization allowed us to get a quick vision of the effects and correlation among the predictors and the outcome variable. Various data mining strategies and models were then applied from which the most successful model was chosen. Only the visualizations that deemed meaningful to us were added to the appendix.

Recommendations:

According to the box office revenue we predicted, movie theaters in China can better arrange their shows and venues. Furthermore, they will also know which factors have significant influence on their revenue.

Detailed Report

Problem description:

Our business goal is to maximize the profits of China movie theaters by meeting the demands of their customers, and our data mining goal is to use the information about movies in Taiwan to predict the box office revenue of the movie released in China.

Data description:

Our data has 187 records and 67 columns. Each record is a movie. The output variable is "China box office revenue" and the input variables are showing in the bottom pictures.

- **sample of 5 rows:**

total_comments	taiwan_release_date	taiwan_day	taiwan_month	TotalMinutes	expected_rating	num_exp_votes	actual_rating	num_actual_votes	boxoffice_taiwan(10thousand)
531	41089	Friday	June	136	82	2302	4.5	2760	10800
623	41010	Wednesday	April	131	98	1997	4.5	3061	9900
1050	41024	Wednesday	April	142	99	6506	4.8	7353	23700
259	41124	Friday	August	118	97	828	4.4	1268	4400
259	41052	Wednesday	May	106	99	1740	4.6	1652	9500

china_release_date	china_day	china_month	type_action	type_adventure	type_animation	type_comedy	type_crime	type_drama	type_fantasy	type_horror	type_other	type_romance
41148	Monday	August	1	0	0	0	0	0	0	0	0	0
41017	Wednesday	April	1	0	0	0	0	1	0	0	0	0
41034	Saturday	May	1	0	0	0	0	0	0	0	0	0
41202	Saturday	October	1	0	0	0	0	0	0	0	0	0
41054	Friday	May	0	0	0	0	0	0	0	0	0	0

type_sci-fi	type_thriller	taiwan_day_Friday	taiwan_day_Thursday	taiwan_day_Tuesday	taiwan_day_Wednesday	taiwan_month_April	taiwan_month_August	taiwan_month_December	taiwan_month_February
0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0
0	0	0	0	0	1	1	0	0	0
1	1	1	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0

taiwan_month_January	taiwan_month_July	taiwan_month_June	taiwan_month_March	taiwan_month_May	taiwan_month_November	taiwan_month_October	taiwan_month_September	china_day_Friday
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	1

china_day_Monday	china_day_Saturday	china_day_Sunday	china_day_Thursday	china_day_Tuesday	china_day_Wednesday	china_month_April	china_month_August	china_month_December	china_month_February
1	0	0	0	0	0	0	1	0	0
0	0	0	0	0	1	1	0	0	0
0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

china_month_January	china_month_July	china_month_June	china_month_March	china_month_May	china_month_November	china_month_October	china_month_September	boxoffice_china(100million)
0	0	0	0	0	0	0	0	3.16
0	0	0	0	0	0	0	0	3.1
0	0	0	0	1	0	0	0	5.68
0	0	0	0	0	0	1	0	1.16
0	0	0	0	1	0	0	0	5.03

Data preparation details:

- **Data Collection:**

We collected movie data from Yahoo Movie Taiwan. The data collected included variables such as movie title, type, directors, expected rating, actual rating. We used Ruby to scrape the data and got 5,000 records. Then, we used three other websites to manually collect box office revenue in Taiwan and China. Those websites are: 台灣偶像劇場, douban, and CBO.

- **Data Manipulation:**

To preprocess the data, we deleted records which didn't have Taiwan box office revenue, China released date or China box office revenue. Second, we deleted the data for which China release dates were earlier than the Taiwan release dates as well as the records for which China release dates were one week later than Taiwan release dates. In an effort to deal with categorical variables, we created dummy variables of movie types as well as Taiwan and China release date. Subsequent to that, we changed the format of both Taiwan and China release date. Finally, we converted the units for both Taiwan box office revenue and China box office revenue in such a way that they are on the same scale. As a result, the converted box unit for Taiwan box office revenue was converted to 10 thousand and the unit for China box office revenue was converted to 100 million. In an effort to do classification, we needed to bin box office revenue into different classes. In doing so, we chose 5 and 3 classes.

Data mining solution:

- **Methods applied:**

We have two different types of model, one is prediction, the other one is classification. Furthermore, we used two different tools (XLMiner and R Studio) to accomplish the data mining task.

Prediction

Tool	Data Mining Method	Variable Selection
XLMiner	Linear regression	All
		From PCA
		From LR Variable selection
	KNN	All
		From LR Variable selection
R	Linear regression	All
		From PCA
		From LR Variable selection

We used XLMiner and R to do linear regression. We used all variables to train the model as our benchmark. In addition to that, we selected some variables using two different methods: (1) PCA and (2) variable selection, which resulted from the linear regression in XLMiner.

One way we applied KNN was by using all variables and setting k to 20. Another way we applied it was by using variables selected from LR variable selection with k set to 18.

Classification

Tool	Data Mining Method	Number of class	Variable Selection
XLMiner	Classification tree	3	All
			From LR Variable selection
		5	All
			From LR Variable selection
	KNN	3	All
			From LR Variable selection
		5	All
			From LR Variable selection

For the Classification Method we used XLMiner to do classification tree and KNN. Because we are doing classification, so we need to transform the numerical variable into bin variable. We bin the box office in China into 3 and 5 classes. These classes will be our output of classification tree and KNN.

- **performance evaluation:**

For prediction methods, we used SSE and RMSE. We looked at which model could give us the smallest results. The best model is Linear Regression using Variable Selection (3 variables).
For classification methods, we looked at overall error. The best model is Classification Tree using Variable Selection (3 variables).

Conclusions:

- **advantages:**

The model described is not costly to implement, and it does not introduce any treath to the existing performance of the movie theaters. In Reference to the quality of variables, we currently have more data that we are preparing to add to the model.

- **limitations:**

The limitations we faced in accomplishing our objectives are related to the fact that we could only use the movies that were released both in Taiwan and China. Furthermore, China movie theaters preferred to broadcast the movies produced in their own country. In addition to that the initial dataset contained important variables that we had to remove from the model. The removal of those variables was because we could not identify a method for handling such categorical variables. Those variables include: Movie Actors, Movie Producers and Comments.

- **how to choose variables:**

Although we used both PCA and variable selection from linear regression, the results from PCA didn't perform well. The issue we encountered may be linked to the fact that PCA is an unsupervised method, so the important variables which showed up didn't depend on our output variable. So if we want to run linear regression, we could just use variable selection which is already included within the linear regression method.

- **operational recommendations:**

China movie theaters can use our models to identify which factors have more influence on the revenue from their box office movies. They can also use the models to predict each movie's revenue. We recommend that you consider the application of our model. The implementation of our model is not costly; furthermore, it doesn't introduce any threat to the existing performance of the movie theaters. Adopting our model is considered a 'win-win' situation, as various factors suggest that our model will allow for the identification of those 'high' and 'low' revenue movies. Ultimately, the movie theaters will have an additional resource to assist with the identification of quantity supply needed for their movie shows and venues.

Appendix

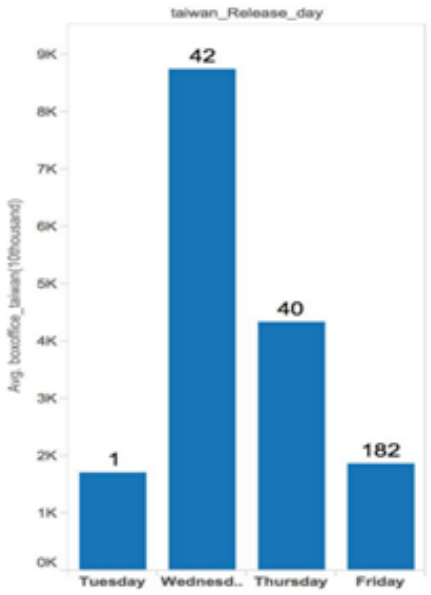
Dummies for movie type:

We chose the top 12 popular movie type to create dummies and put other type to “other”. Top 12 popular type includes:

Crime	Animation	Action	Adventure	Sci-fi	Fantasy
Comedy	Thriller	War	Romance	Drama	Horror

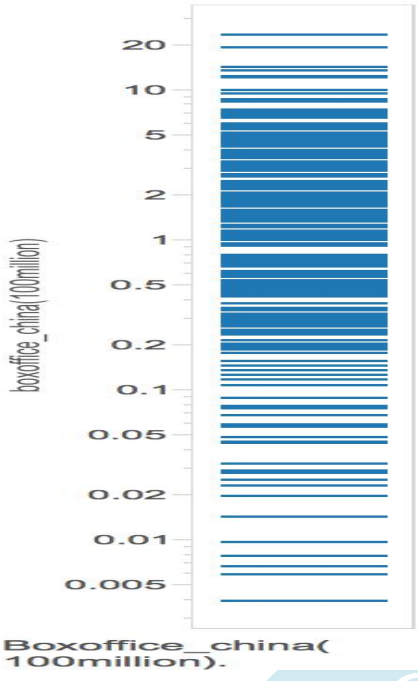
Data Visualization:

We found that although the number of movies released on Friday was the highest, the average box office revenue was the highest on Wednesday. The reason caused this situation may be that the movies released on Friday including the good movies and bad movies.



Distribution of China box office:

This is the distribution of China box office. We bin the China box office based on the distribution.



Output from Models:

Method	Variables		SSE	RMSE
XLMiner - **Linear Regression**	ALL	Training	184.77	1.402
		Validation	370.14	2.571
		Testing	358.16	3.111
	PCA	Training	292.51	1.764
		Validation	317.08	2.379
		Testing	274.50	2.723
	Variable Selection (3 variables)	Training	386.88	2.028
		Validation	259.41	2.152
		Testing	274.83	2.725

Method	Variables	cross validation (80/20)	SSE	RMSE
R - **Linear Regression**	ALL	Testing	66	1.328
	PCA	Testing	60.98	1.277
	Variable Selection (3 variables)	Testing	73.58	1.403

Method	Variables		SSE	RMSE
XLMiner - **KNN**	ALL (k=20)	Training	≈0	≈0
		Validation	617.944	3.321
		Testing	325.406	2.965
	Variable Selection (3 variables) (k=18)	Training	≈0	≈0
		Validation	709.089	3.558
		Testing	336.026	3.013

Method	Output Class	Variables		Overall Error
XLMiner – **KNN**	5 classes	ALL **(k=1)**	Training	0%
			Validation	69.64%
			Testing	83.78%
		Variable Selection (3 variables) **(k=1)**	Training	0%
			Validation	67.86%
			Testing	67.57%
	3 classes	ALL **(k=4)**	Training	45.74%
			Validation	37.5%
			Testing	56.76%
		Variable Selection (3 variables) **(k=2)**	Training	24.47%
			Validation	46.43%
			Testing	59.46%

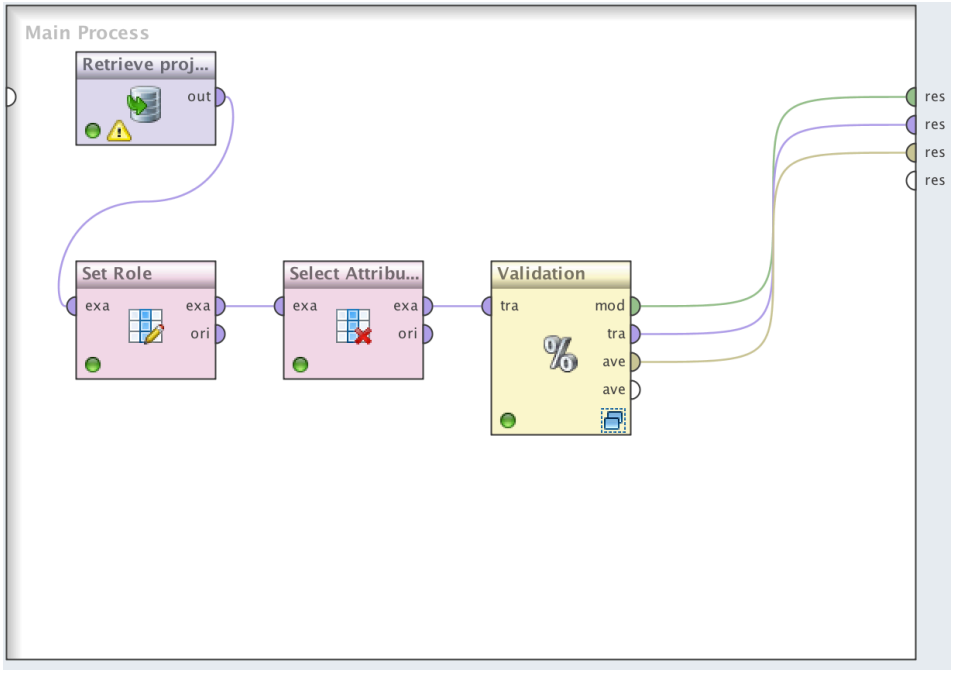
Method	Output Class	Variables		Overall Error
XLMiner – **Classification Tree**	5 classes	ALL	Training	13.83%
			Validation	66.07%
			Testing	70.27%
		Variable Selection (3 variables)	Training	29.79%
			Validation	57.14%
			Testing	56.76%
	3 classes	ALL	Training	11.70%
			Validation	44.64%
			Testing	45.95%
		Variable Selection (3 variables)	Training	13.83%
			Validation	39.28%
			Testing	27.03%

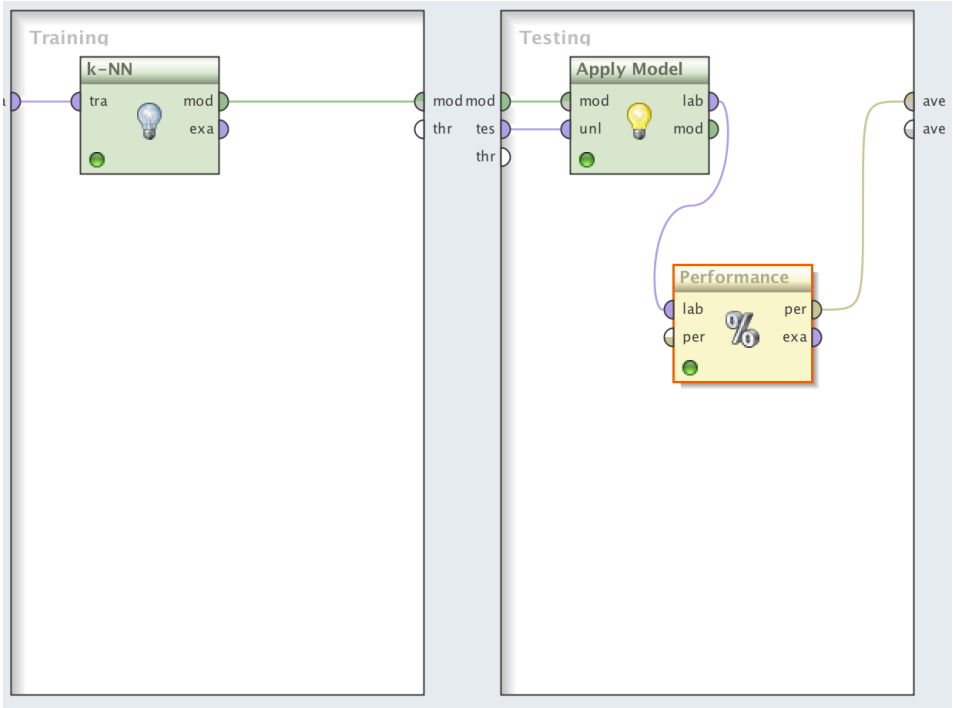
RapidMiner:

When we were going to build models, we also tried to use RapidMiner.

- **KNN (for prediction)**

We used cross-validation and set to 10-folds to run KNN models.



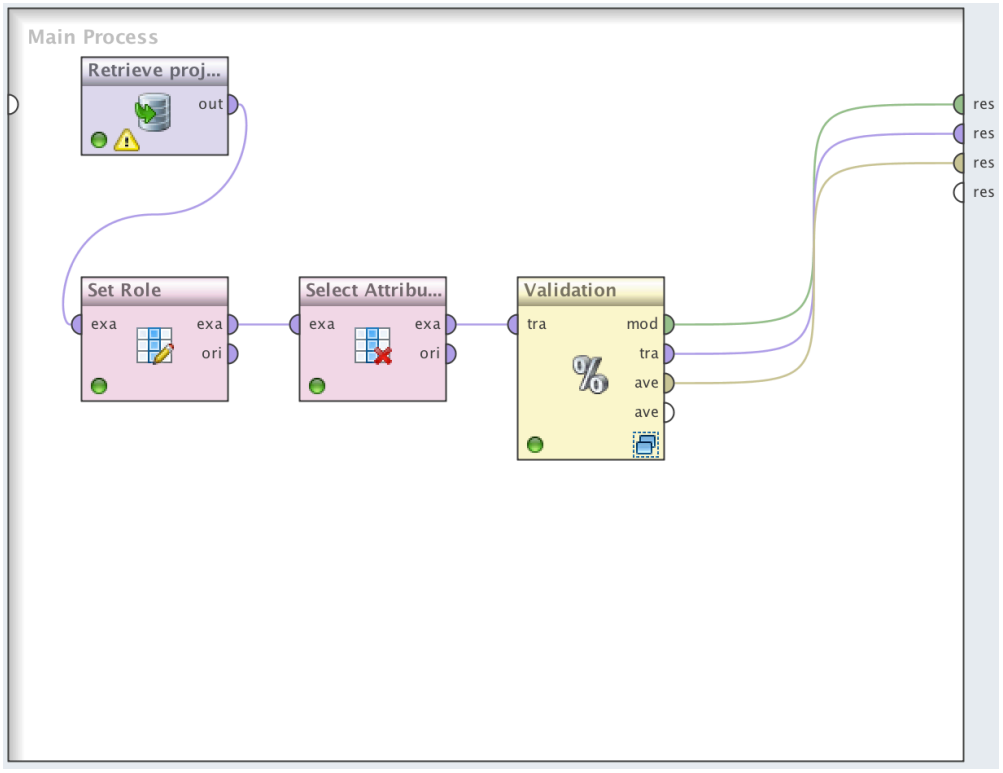


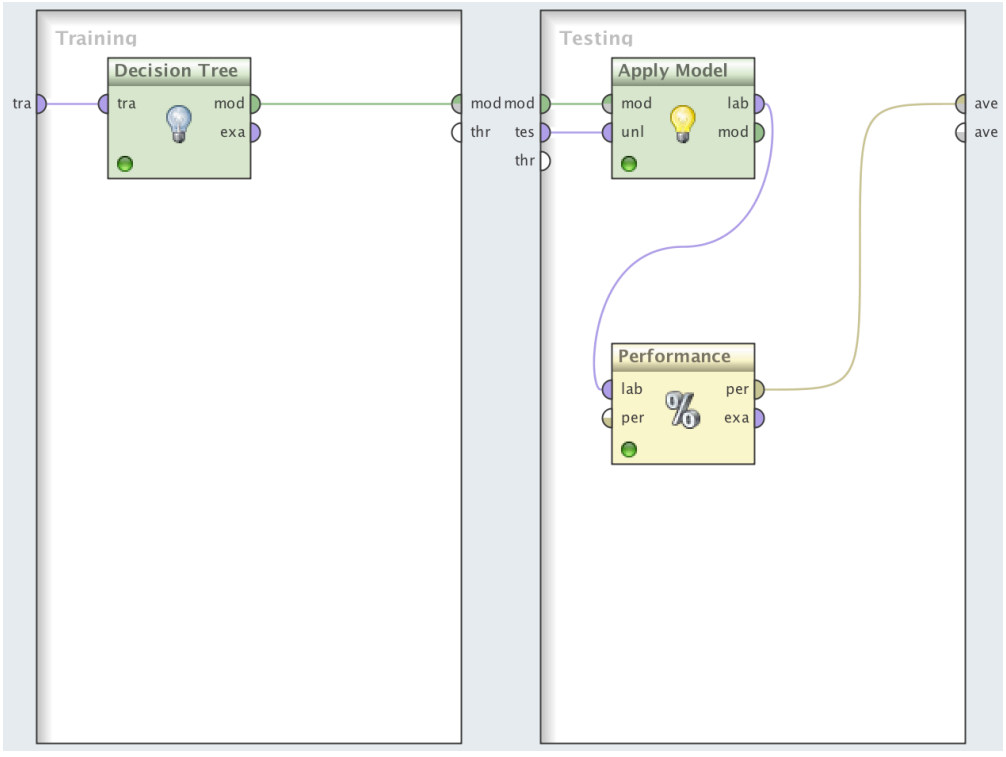
root_mean_squared_error

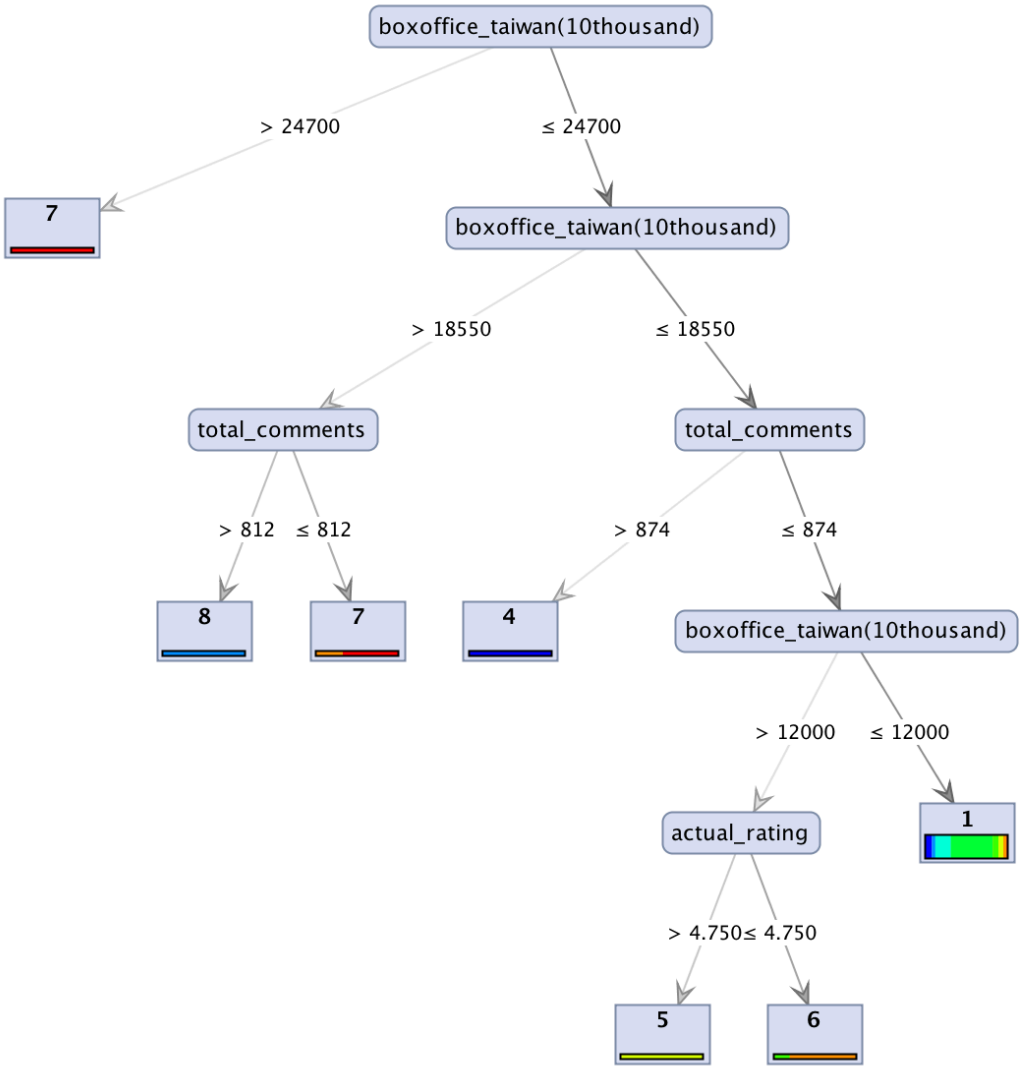
```
root_mean_squared_error: 2.317 +/- 1.024 (mikro: 2.518 +/- 0.000)
```

- **Classification Tree (Decision Tree)**

We used cross-validation and set to 10-folds to run Classification Tree models. The bin classes here are not same as what we used in XLMiner. This is the middle process of our project to look for better models.







accuracy: 50.35% +/- 5.01% (mikro: 50.27%)

	true 4	true 8	true 2	true 1	true 3	true 5	true 6	true 7	class precisi
pred. 4	2	1	1	1	2	1	0	0	25.00%
pred. 8	1	1	0	0	0	0	0	2	25.00%
pred. 2	1	0	3	3	1	1	0	0	33.33%
pred. 1	11	5	27	83	9	8	8	1	54.61%
pred. 3	1	0	0	0	0	1	0	0	0.00%
pred. 5	0	0	0	0	0	0	0	0	0.00%
pred. 6	1	1	1	0	1	1	3	0	37.50%
pred. 7	0	1	0	0	0	0	1	2	50.00%
class recall	11.76%	11.11%	9.38%	95.40%	0.00%	0.00%	25.00%	40.00%	