Analysis Report

Shen Zhang

4/23/2019

1. Introduction

People play board games in different cultures and societies for a long time. There are many different types of board games, like Worker Placement, Deck Builders, Area Control and so on. The popularity of board games is different. It's very interesting to see which board game is more popular, and which attribute affects more about the ranking of a board game.

2. Data Description

2.1 Collection Procedure

The data set is collected from www.boardgamegeek.com. Basically, several variables are parsed by the python program. After the data cleaning procedure, there are 7 variables in the dataset. They are GameName, AvgRating, GameRank, GeekRating, ListPrice, NumVoters and RankCategory. The total number of observations is 8,005.

I create the variable RankCategory by myself based on GameRank. GameRank in the first quartile will be classified in group 1. Then GameRank in the second quartile in group 2, and so on. There are four categories for this RankCategory variable for analysis purposes.

Table 1 shows the summary statistics of continuous variables.

Table 1

	AvgRating	GeekRating	ListPrice	NumVoters
Min	1.050	3.466	3.99	30
Median	6.710	5.711	39.84	372
Mean	6.678	5.888	39.84	1530
Max	9.590	8.610	299.99	84548

2.2 Missing Values

Missing values come from the variable List Price. The other variables do not have missing values. In other words, not all board games have a list price. Only 2,239 observations have a list price out of 8,005.

By dealing with Missing values, I simply use the mean value for the NAs of Listprice variable.

Basically, I have two datasets. One data set with the missing values filling with mean values. The other dataset is the original one which has the missing values in the dataset.

3. Process of Analyzing

By given a new board game, we want to see the potential rank group based on rating, list price, and the other variables. I will use two supervised machine learning method to predict the RankCategory. RankCategory is our dependent variable, and the other 6 variables are the independent variables. K-nearest-neighbor and randomForest will be used to do the analysis.

I use those two datasets to do the analysis and compare the results to see if there is a difference affected by missing values. Also, I will check the importance of the variable Listprice.

Cross-Validation will be used to compare the accuracy of the prediction for the two methods. Also, we will see which independent variable is more important.

4. Analysis Results

4.1 Data set with Missing Values filled

4.1.1 KNN

Table 2

	Predict 1	Predict 2	Predict 3	Predict 4
True 1	350	46	1	0
True 2	94	269	41	1
True 3	12	84	299	16
True 4	15	33	91	249

Table 2 shows the confusion matrix by using KNN method. The accuracy score is about .7289 which is the sum of diagonal divide by the total number of prediction in table 2. We can see KNN method is not very good.

4.1.2 Random Forest

Table 3

	Predict 1	Predict 2	Predict 3	Predict 4
True 1	397	0	0	0
True 2	1	404	0	0
True 3	1	1	61	0
True 4	0	0	26	69

Table 3 shows the confusion matrix by using the RandomForest method. The accuracy score is about 0.9981. We see the RandomForest has a good prediction. Importance probability is like following, AvgRating 0.1308, GeekRating 0.6232, ListPrice 0.00011, NumVoters 0.2448.

We can see after we fill the missing value, variable ListPrice does not important. Actually, the Greeking rate and the number of voters will affect more about the rank of a board game.

4.2 Data set with Missing Values

4.2.1 KNN

Table 4

	Predict 1	Predict 2	Predict 3	Predict 4
True 1	91	12	0	0
True 2	31	69	23	1
True 3	6	31	61	20
True 4	1	7	26	69

4.2.2 Random Forest

Table 5

	Predict 1	Predict 2	Predict 3	Predict 4
True 1	102	1	0	0
True 2	0	122	2	0
True 3	0	0	118	0
True 4	0	0	1	102

Table 4 shows the confusion matrix by using KNN method. The accuracy score is about 0.647. Table 5 shows the confusion matrix by using the RandomForest method. The accuracy score is about 0.9911.

Compare by both cases, we can see the prediction of KNN method is worse than RandomForest.

Importance probability of RandomForest is like following, AvgRating 0.114, GeekRating 0.6988, ListPrice 0.0029, NumVoters 0.1838. Variable ListPrice still not very important. Actually, the Greeking rate and the number of voters will affect the rank of a board game. The missing values do not affect the results too much.

5. Conclusion

From the analysis results, we can see Greek rating and the number of voters are important to the rank of a board game.

On the contrary, the price of the board game is not that important. Maybe in general, the price for the board game is not very high, people care more about the interesting of board game instead of price.