STATS 415 Project: Video Game Sales

Wenfei Yan, Runhe Lu, Yunguo Cai

June 12, 2018

1 Introduction

1.1 Data

The dataset is found on Kaggle and contains 16375 games in total. There are 6 numeric variables for each observation: global sale, user score, critic score, user count, critic count, and year of release; and 5 categorical variables: platform, genre, publisher, developer and rating.

Out of all the observations, 9826 (59%) of the games have missing values in user score, critic score or year of release, most of which were released before 2000. This is reasonable as it's hard to find the user and critic reviews for these old games. Furthermore, the user and critic reviews could be important variables for our prediction goal (which will be described in section 1.2), we remove all the observations with any missing value.

After the removal, we have 6893 observations left. The rest of the dataset contains games released form 1996 to 2017, on 31 different platforms and of 12 different genres. The dataset is still comprehensive and is in a reasonable scale.

1.2 Project Goal

Observing the dataset, we notice the global sales would be a interesting and reasonable prediction response. In real world, people would likely to estimate how popular a game would be based on known information. We also have some meaningful variables for prediction on sales, such as the user and critic scores, genres and the year of release. The user and critic scores may reflect the game quality. Game popularity may differs among genres. Also, since the global sales variable records the accumulative sales of the game once it released, the year of release may also help.

There are mainly three part of the prediction question:

- What're the most useful predictors for predicting sales?
- What're the best prediction model?
- How accurate can we predict?

2 Exploratory Data Analysis

We first make some plots of the variables to better understand the dataset, and also get some basic but interesting findings.

2.1 Users and critics are more likely to give an average score

From the box plots below, it shows that most user scores and critic scores range from 60 to 80. Extreme scores (higher than 90 or lower than 40) are rare. This suggests that users and critics tend not to give an extreme score.

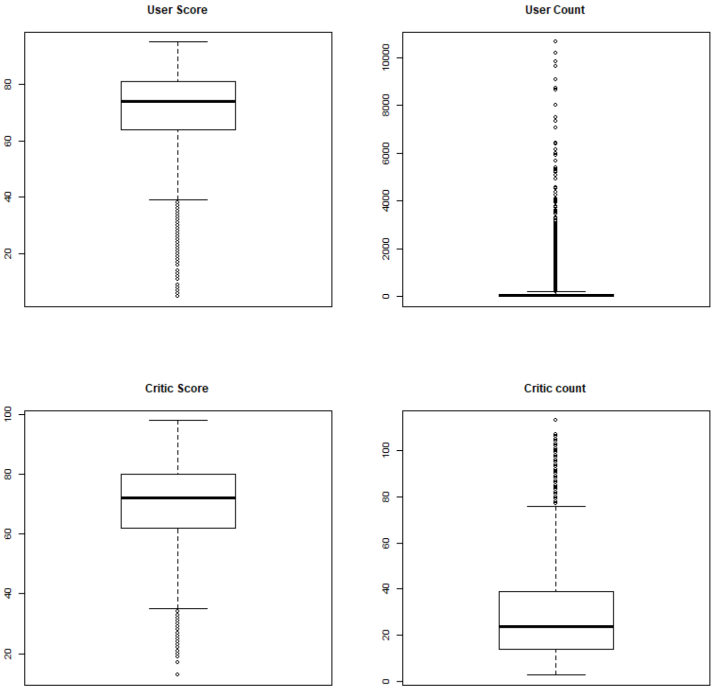

Figure 1: Box plots of User Score, User Count, Critic Score and Critic Count

2.2 Most users do not give a review

From the box plot of user count, we can observe that the majority of this variable is small. Actually, most games have a user count fewer than 200. Compared with the global sales in unit of million, the small user count indicates that most users do not give a review for the game they bought and played.

2.3 Most games have global sales lower than 1 million

min	Q1	median	mean	Q3	Max
0.01	0.11	0.29	0.77	0.75	82.53

From the summary statistics of the global sales shown in the table above, we know that sales of 75% of the games are lower than 0.79 million. In fact, only 19% of the games have global sales higher 1 million. Meanwhile, some extreme values in sales exist (max sale 82.53 million).

2.4 Some games are much more popular than others

The outliers in the user count box plot and the summary statistics of global sales both show that some games are much more popular than others. They have extremely high sales and more than 10,000 user reviews. These extreme outliers might have a bad influence on our prediction.

3 Regression

We first try to predict the global sales by regression. We randomly split 80% of original data as training set and the remaining as test set. We also replace year of release variable with the number of years that the game has been selling. There are 10 potential predictors in total, so we decide to use variable selection methods, shrinkage methods and dimension reduction methods to improve simple linear regression with OLS. Besides those parametric methods, we also conduct a non-parametric method: KNN regression.

Variable Selection

We use C_p and BIC as the criterion in variable selection. The two methods both select user count, platform and genre as our predictors. We then used these three predictors to do linear regression on sales. The resulting test MSE is 6.98, adjusted R^2 is 0.2378, which shows that the true model is far from a linear one.

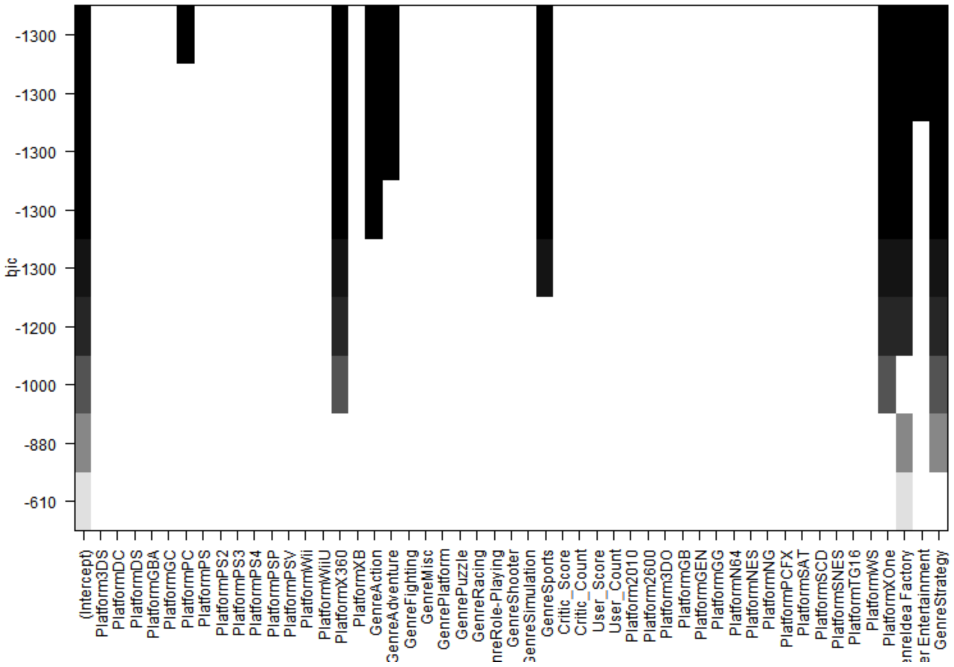

Figure 2: Variable Selection Result by BIC

Shrinkage methods

Variable selection is not satisfying. We then try the ridge and lasso regression to shrinkage the estimated coefficients. Before we conduct regression on our training data, we used cross validation to choose the tuning parameter(λ) which gives us the minimum cross validation MSE in ridge and lasso model. λ and corresponding test MSE are listed as follows:

	λ	Test MSE
Ridge	0.39	5.67
Lasso	0.01	5.54

In comparison with the only three predictors selected by BIC and C_p (critic score, platform, genre), lasso regression selects user score, user count, critic score, critic count, platform and genre as predictors. The better performance of Lasso implies that some important predictors were lost during BIC and Cp procedures.

Dimension reduction and KNN methods

We choose principal component regression to reduce the dimension and predict global sales. We also used the non-parametric method: KNN regression. Only the 5 numeric predictors are used in both methods, as it's hard to calculate the variance or distance when including categorical predictors. The number of principal components chosen in PCR by cross validation, tuning parameter K chosen in KNN by cross validation and the test MSE are listed:

	Parameter	Test MSE
PCR	#comp = 4	7.13
KNN	$K = 100$	6.80

Both methods give us a worse performance than lasso regression. The performance of PCR is the worst of all methods, which makes sense because it fails to reduce the dimension (changing dimension from 5 to 4) and meanwhile loses some information in predictors. The bad performance of KNN also indicates that some categorical variables could be pretty important for predicting sales.

Regression Conclusion

The performance of each regression method is concluded in the following table.

	Variable Selection	Ridge	Lasso	PCR	KNN
Parameter	p = 3	$\lambda = 0.39$	$\lambda = 0.01$	#comp = 4	K = 100
Training MSE	2.31	1.48	1.52	1.98	2.26
Test MSE	6.98	5.67	5.54	7.13	6.80

Lasso regression performs best on prediction, and it selects all the variables except for number of selling years, developers and platforms. We're surprised that the selling years is not included, as it could influence the accumulative global sales. Combining with the best subset selected in variable selection method, which is (critic score, platform, genre). These three predictors are also selected by Lasso, so we consider them to be the most important predictors in regression.

However, the best test error is 5.54, which is still too large compared to the scale of the global sales, as the global sales (in million) records are smaller than 1 for 80% of the games. Also, in all methods, the test MSEs are much larger than the training ones, which might imply that the regression models are all overfitting.

The previous regression analysis has several limitations, which are mainly caused by the variables used (e.g. user score, user count, critic score, critic count, genre and platform) in regression techniques:

- **Missing Values between 1996 and 2000** may decrease the accuracy of prediction. Because of limited Internet access, it could be hard to collect reviews for the old games. All games before 1996 and 83% of the games during 1996 and 2000 are removed due to missing values, mostly in user score and critic score. Lacking data of old games, our model can perform badly when predicting the sales for the remaining old games.
- **Outliers with Large Global Sales** may also hurt the accuracy of prediction. Some games have sales larger than 10 million, while 80% of observations are smaller than 1 million. This could lead to large errors when predicting sales for these outliers. Meanwhile, the outliers in the training sample can also influence the model so that the prediction for those normal points may also be less accurate.
- **Interpretation of Categorical Variables** is hard. Platforms and Genres are chosen predictors, but they have too many levels, resulting in a lot of dummy variables. This makes it hard to interpret the result.

4 Classification

Since the results of predicting global sales with regression is not ideal, we reduce the regression problem to a classification one to see if we can predict better.

We classify the games to three classes according to their global sales. Since the median of global sales is 0.29 million, we consider the games with a global sale lower than 0.3 million as "low" sales games. Also, if a game has a global sale higher than 1 million, we label it as a "high" sales game. The rest of games are labeled as "soso". In short, the data set is comprised of 1310 "high" sales games, 2038 "soso" ones, and 3545 "low" sales ones. There's no imbalanced classes problem, so we can then start our classification directly.

LDA and QDA

We first try some simple methods, LDA and QDA, which don't need parameter tuning. Since categorical variables cannot be used in LDA and QDA, we only include the five numeric predictors.

	LDA	QDA
Training Error	0.4071	0.4173
Test Error	0.4039	0.4206

As shown in the table, neither LDA nor QDA performs well, and both training and test errors are large. This indicates that the normality assumption might not be reasonable. Therefore, we turn to some more flexible methods.

KNN and Decision Tree

Then, we try K-nearest neighborhood and choose the parameter, number of neighbors, according to cross validation. Surprisingly, the accuracy of KNN is even smaller than LDA and QDA, while using the exactly same predictors. It has a test error of 0.6027, which means that it's even worse than random guess. This might be resulted from the curse of dimensionality, so we try to use principal components analysis to reduce the dimensions.

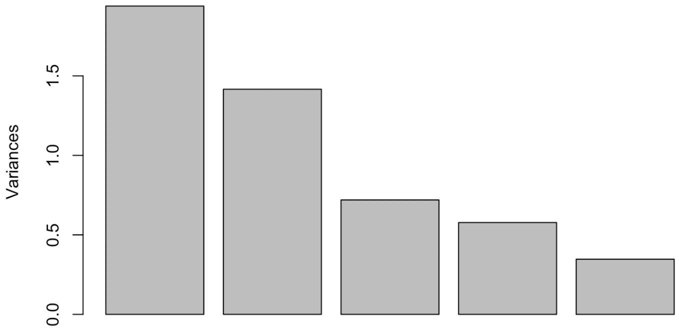

Figure 3: Scree plot of principal components

However, the scree plot of principal components suggests that PCA does not help. The first component only accounts for 38% of the variance, and we have to use 4 out of 5 components to represent 90% of the variance in x's. No matter we use 2, 3 or 4 principal components in KNN, the resulting error rates are still larger than 60%.

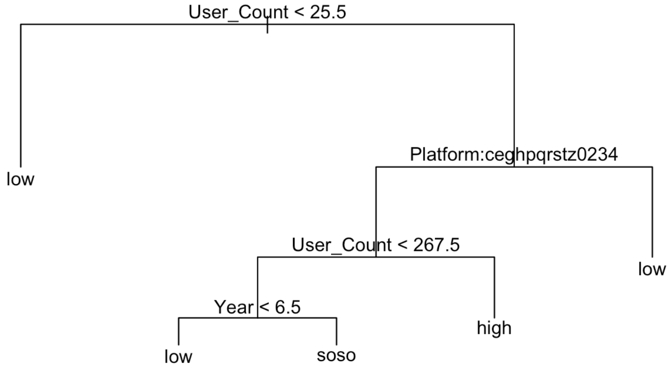

Figure 4: Decision tree

We also try the decision tree, choosing the number of splits by cross-validation. All predictors are used, except for Publishers and Developers, which have too many levels. The actually used variables in the tree are user count, platform and critic score, of which the user count is the most important one. The test error rate is 0.4423, still lower than the result of LDA and QDA. Nevertheless, since the prediction is better than random guess, we can use ensemble methods to improve it.

Ensemble Methods

We apply both random forest and adaboost. The best result that we can get is shown in the table. We first choose the parameters by a sparse grid search. We also find that the random forest is not sensitive to the parameter tunning, so we finally just use the default parameter values. For AdaBoost, we mainly tune the number of iterations.

The test error rates are 0.3067 and 0.3445 for random forest and adaBoost, respectively. There's a large improvement in the test error rate compared to all methods used before. However, it's still not a good prediction, with accuracy lower than 70%.

According to the variables importance plot from random forest, user count is always the most important predictor. Platform, critic score and critic count are also relatively more important than others.

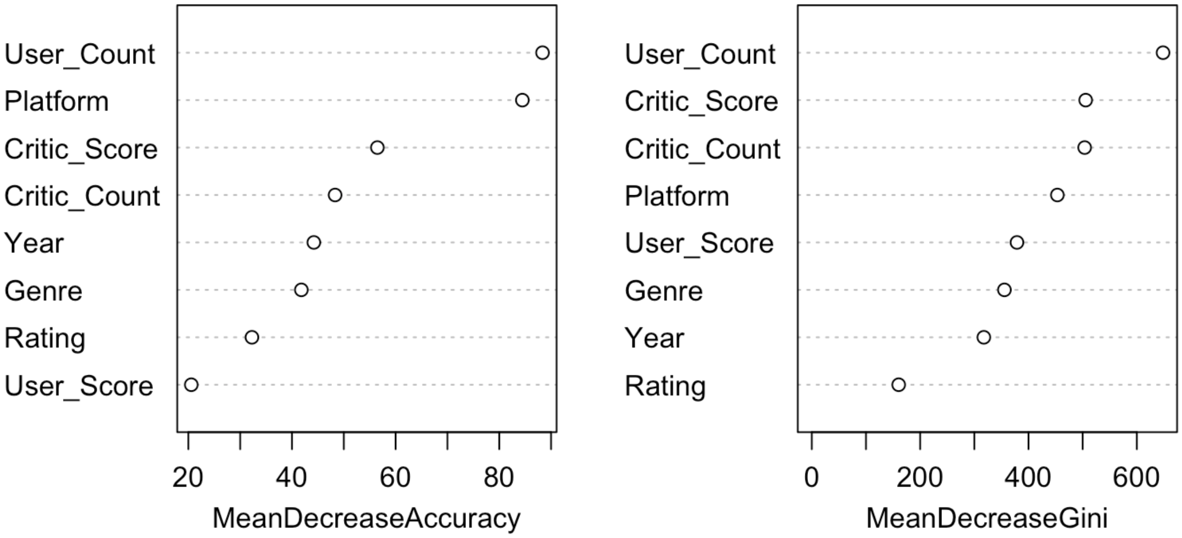

Figure 5: Variables importance plot from random forest

SVM

We also try support vector machine classifier, as its popularity indicates it can have a good performance on many problems. We use both ”linear” and ”radial” kernals. The parameters are determined by a sparse grid search. As shown in the table below, the performances of SMV using both kernals are similar. The radial one is a little bit better and its accuracy is close to the random forest.

	test error	cost	gamma
linear	0.3147	0.5	/
radial	0.3075	2	0.5

Classification Conclusion

The performance of each classification method is concluded in the following table.

	LDA	QDA	KNN	Tree	RF	AdaBoost	SVM, radial
Parameter	/	/	$K = 33$	split=5	ntree=500	#tree=761	$C = 2, \gamma = 0.5$
Test error	0.4039	0.4206	0.6027	0.4423	0.3067	0.3445	0.3075

The random forest and the SVM perform equally well, and are the best two among these classifiers. According to the classification tree and random forest, the user count is the most important predictor in classification. Also, the platform, critic score are more important than the rest.

The three most important predictors for classification are different from those for regression. Platform and critic score are important for both classification and regression. However, user count is the most important for classification, while it's even not in the three most important predictors for regression.

However, the best classification error is 0.3067, which is still not a good performance in prediction. We investigate the confusion table from random forest and "radial" SVM prediction to see more details of the prediction error. Both classifiers predict the "high" sales games best, and it seems hardest to tell difference between a "soso" game and a "low" sale one.

	actual		
random forest	high	soso	low
high	159	49	11
soso	73	183	84
low	30	176	614

	actual		
svm	high	soso	low
high	138	33	5
soso	69	191	78
low	55	184	626

The possible reasons for the bad performance in classifying non-"high" sale games could be similar to the regression part. The outliers with high global -ales influence the model a lot. One possible way to solve this problem is to build prediction models for "high"-sales games and non-"high"-sales games separately.

Finally, as both the results of the regression and classification problem are not ideal, we suspect whether we have information to make good prediction for the sales. For example, the game quality could be quite different from the game sales, so the user and critic scores are not that useful as we expected.