***Introduction to Data Mining***

***DATS 6103***

***Final Project***

*Milestone -5-*

**Data Mining Model Evaluation and Results communication**

## Author

Provide your Full Name here and your GWID.

## Michael Wegan G45192209

## Zheying Chen G43061235

## Weiwei Li G27494462

## Xueli Zhou G29955935

## Introduction

Milestones 5 is stepping stones to complete your semester Final Project. The goals of these millstone deliverables it to help you progressively achieve your end goal while you understand, practice and master each building block of your final product.

## Explore and Prepare Data:

Please answer the questions below and provide the following:

1. Test your results and discuss quality of implementation

The results were very mixed. Overfitting was a problem with many of the modeling techniques. Most of the models performed fairly well against the training data, but many were not very useful against the testing data. Our best models were the support vector machine, random forest, and neural network. These models were superior to the others in their ability to correctly classify, even if they all had issues with false negative rates. So in milestone 5 we updated our model: Ensemble 4 different models(SVM, ANN, Random Forest and KNN) to reduce the false negative rate. We also enlarged the size of the training data by duplicating the true MVPs in order to solve the overfitting problem. The new model effectively reduced the false negative rate and still kept a high accuracy. So the quality of implementation is good. We finally got a better result.

1. Provide explanation of results

The new model (after ensembling 4 models and enlarging the size of the train data) performed well. Result showed that the false negative rate is 13.33%, which also meant that the model had a relatively high sensitivity (true positive rate). Thus the result showed that 86.67% (sensitivity)of the true MVP winners can be recognized when we adopted the new model. In the meanwhile, the accuracy of the new model is still very high (97.19%), although this index is a little lower than the other models we used.

1. Evaluate your findings against another model

Comparing the other models we used, the new model effectively reduced the false negative rate and still kept a high accuracy. So to some extent, we can conclude that the new model is better than SVM, ANN, Random Forest and KNN. The details of the contrast among these different models as benchmarked against the test data partition is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | False Positive Rate | False Negative Rate | Sensitivity | Specificity | Precision |
| SVM | 98.62% | 1.21% | 36.07% | 63.93% | 98.79% | 20.64% |
| ANN | 98.75% | 1.11% | 29.51% | 70.49% | 98.90% | 23.76% |
| Random Forest | 97.37 | 2.6% | 8.2% | 91.8% | 97.4% | 14.74% |
| KNN | 99.5% | 0% | 100% | 0% | 99.98% | 0% |
| Ensembled Model | 98.28% | 1.7% | 24.49% | 80.33% | 98.30% | 18.85% |

More details can be seen in the attached .R files.

1. Discuss Results impact on initial hypothesis.

Our initial hypothesis is that we can successfully predict who are mvp winners in a relatively small scope, because mvp winners should have better performance statistics. In the initial analysis we can easily find out the margin between the mvp and non-mvp winners based on the analysis of primitive dataset through some boxplots and scatter plots. But the later results showed that too many players have similar statistical profiles and there is not a clear margin between the mvp and non-mvp ones. So it is hard to distinguish who would be the true mvp winners. Therefore, we should take more factors (such as media exposure rates) into consideration, so that the model could provide more accurate prediction, but those data are difficult to find.

1. Discuss Results potential application and future research

Many sports networks and media outlets prognosticate the MVP winners every year, but none have models to actually predict them. All the predictions are based on the knowledge of the person making the prediction. The research can be used to guide people in deciding which lotteries to buy regarding guessing the final mvp winners or champion teams. Or the research can be used to help lottery companies to decide the odds of each team or player. Moreover, some sports media companies will probably be interested in a product like this since if they can publish prediction results correctly before the final election, they can surely attract more readers and make more profits. Our concentration of the future research should be put on how to perfect the prediction models. Given that our dataset is too large and timescale is too long, some variables we want to put into our model are difficult to collect complete data, so we can cut time range(for example, we only choose date in the last 10 years as our sample). If so, it will be easy to put the “likeability” of the candidate winners as an important variable into our model. We can also take the media exposure rate of the candidates into consideration in future research.

1. Provide References for your work.

Reference:

[1]. Leung C K, Joseph K W. Sports data mining: predicting results for the college football games[J]. Procedia Computer Science, 2014, 35: 710-719.

[2]. Schumaker R P, Solieman O K, Chen H. Predictive modeling for sports and gaming[M]//Sports Data Mining. Springer US, 2010: 55-63.

[3]. Cao C. Sports data mining technology used in basketball outcome prediction[J]. 2012.

[4]. Solieman O K. Data mining in sports: A research overview[J]. Dept. of Management Information Systems, 2006.

[5]. Miljković D, Gajić L, Kovačević A, et al. The use of data mining for basketball matches outcomes prediction[C]//IEEE 8th International Symposium on Intelligent Systems and Informatics. IEEE, 2010: 309-312.

[6].When can we use support vector machine basic over other classification algorithms?(2016, November 24). Retrieved November 25, 2016, from Quora：<https://www.quora.com/When-can-we-use-support-vector-machine-basic-over-other-classification-algorithms>

[7]. Why use SVM?(2016, November 24). Retrieved November 25, 2016, from The Yaksis: <http://www.yaksis.com/posts/why-use-svm.html>

[8].Classification (machine learning): When should I use a K-NN classifier over a Naive Bayes classifier?(2016, November 24). Retrieved November 25, 2016, from Quora: <https://www.quora.com/Classification-machine-learning-When-should-I-use-a-K-NN-classifier-over-a-Naive-Bayes-classifier>

Below are questions that you should investigate, answer and report back in as many details as possible (written summaries, pictures, code, snapshots, ….)

1. After using different modeling techniques on your data, how did the different models perform?

We initially used 6 kinds of models, including K-Nearest Neighbor(KNN), neural network, support vector machine, logistic regression, naïve Bayesian, random forest. Support vector machine, random forest, and neural network performed better due to the fact that they can correctly classify our data with fewer false negatives, even though they also had issues with false negative rates. KNN, logistic regression and Naive Bayesian were not good enough because they have too many outliers and false negatives which cannot help us decide the range of the mvp candidates.

1. How did you use R to evaluate your models?

We used table() and confusionMatrix() functions in R to looked at the model’s rates of correctly classifying the testing data. We got the accuracy , false positive rate, false negative rate, sensitivity, specificity and precision of different models. Through comparing all the rates of these models, we found that the ensembled model is the best one.

More details can be seen in the attached .R files.

1. For your specific project, which model performed best and why?

In milestone 4, after trying 6 models, we found that support vector machine, random forest, and neural network performed best due to the fact that they can correctly classify our data, even though they also had issues with false negative rates, but when cross-validation is used, we can get better results.

So in milestone 5, we adopted a new approach to improve our model. We mainly used ensembles package in R to integrate support vector machine, random forest, neural network and KNN into one model. Using this new model, we got a better result, which means lower false negative rate and higher accuracy. The new model performed best because it succeeded reducing the impact of the limitations of each model type and allowing for more robust estimates of class for each player.

1. Explain how people in different circumstances might choose models differently?

All of the 6 models we used had their own constraints. So one should choose different model according to the specific circumstances including data characteristics, data quality, the specific problem needed to fix, etc. Specifically, the support vector machine can be chosen when you have a huge number of features. For example for text classification in a bag of words model. This algorithm is also particularly useful when you want ordinal classification (ranking), it is widely used in "learning to rank" algorithms. And support vector machine with nonlinear kernels perform quite well in most cases and are usually head to head with random forests, sometimes random forest works slightly better and sometimes support vector machine wins. The naïve Bayesian classifier is a simple and easy to implement algorithm. It is unable to handle correlated continuous data and repetitive categorical data. Due to its simplicity, naïve Bayesian classifier might outperform more complex models when the data set is not large enough. You can try it whenever you don’t need the actual probabilities for each class, but only to know which is the most likely.

1. If you were to continue this project after the class, how ideally like to communicate the results?

We can communicate with each other about the results of the latest research via emails or we can post these results (including new hypothesis and implemented codes of project) on GitHub repositories for us to exchange information on time. What’s more, we can also post blogs on famous websites to publish results so that everyone, but not limited to ourselves can see these results, and they can offer useful recommendations for us to improve our research.

1. Who is your audience for this project?

Considering it is the final project in data mining class, our teacher Professor Bari and assistant teacher are the most important audiences for this project because the knowledge we learned in this class can be shown in this project, through which our data mining ability can be proved adequately. In addition, all of the classmates are also our audiences because they provided us helpful advice, just as our teachers. Outside of the classroom, anyone interested in sports analytics may find the results compelling

We hope to get some useful suggestions which can improve our project through the final presentation of this project.

1. How would your project inform decision-making?

We made decisions mainly by majority and by consensus which implied open mindedness and a willingness to let go of attachment to one's own idea and accept the ideas of others with objectivity. For example, in the beginning of this project(milestones 1&2), one of our team-members, Michael Wegan, proposed an interesting topic—predicting the future most valuable player (MVP) of the league baseball games through mining the performances of past players. All other members in our team thought it’s a great idea because we had sufficient raw data (a longitudinal data set of individual player statistics of all players from 1871 until present), and sports data mining is indeed an exciting topic to explore. Majority and consensus decisions also be used often in the following steps when we continue our project.

We also used group discussion and brainstorming to advance our work. For instance, during the model building and evaluation process (milestone 4), each member in our team did some research through reviewing related literature and searching for related cases by internet. Then we carried out group discussion and decided to try several models through using different classify methods. We believe that the best model for our project can be found through this process. And this does work. We finally found out that support vector machine, random forest, and neural network are best models for our project.

1. What lessons learned

The first lesson we learned is data mining is not omnipotent. Thus we should not overestimate the effect of data mining models, although they are indeed very useful. We implement sufficient and reasonable data cleaning before building models, and we tried all models we knew, but the result was not so good as we expected. Overfitting was a problem with many of the modeling techniques. Most of the models performed fairly well against the training data, but many were not very useful against the testing data. Even the best ones still had issues with false negative rates.

In addition, we lack of experience to tune suitable parameter in sports data mining case. It took too much time to try different parameters before we got the relatively satisfactory result.

1. If you were to go back and redo this project, what advice would you give to future students?

Students should first choose popular topics which interest most group members so that they will be inspired to explore it and then collect relatively complete dataset to deal with. The way how to tell a good dataset is whether the variables are more associated with each other, and whether there exist tractable links between them and easy to correlate them. When it comes to adding new variables beyond the original ones, students should choose to create ones whose data can more easily be collected from database. Moreover, students should refer to more other people’s projects on how to collect data, clean data, model, test models, and choose the final model, and read more practical related books. Reading and typing more codes will help students get more familiar with various parameters so that they can actually save time on tuning parameters.

1. How has your project transformed data into information? How does your project help people make decisions? How does it add value to the data?

We used R to clean data, build model and get the analysis result. Through the use of R, we successfully transformed boring data into easily understood conclusions through which people can easily tell who are the most probable candidates for the mvp winners and then they can use the knowledge to earn profits for them. In the last step of this project, we will present all the process and result by a 15-20 minutes’ presentation. We believe that presenting the result is as important as the data analysis process. And the presentation can make people easily learn what we did and help them make decisions.

Through this project, we indeed learned how to transfer data into information, even useful knowledge which can help people make decisions. It’s an exciting process. We hope to do more and improve our research in the future.

## Evaluation and Summary

Make your milestone summary main points here. Evaluate your findings so far. Draw some early potential conclusions.

Through milestone 5, we improved our model and succeeded predicting the MVP winners by using the ensemble of 4 different models, as evidenced in the confusion matrices in the attached files. To some extent, the result is good because it reduced the impact of the limitations of each model and allowed for more robust estimates of classification for each player.However, we can probably get better results through adding new variables such as the “likeability” of the candidate winners into our model. It’s difficult to gather data of the potential new variable “likeability” because many of the players and past voters are dead. A possible way to solve this problem is narrowing the time range of the sample. For example, we can only use the last 10 years data as our sample.

This project has given us a good basis for the process of data mining from the acquisition of data to cleaning it to implementing and refining models. We did learned a lot as well as some lessons which needed to improve in our future research. We are excited to present our final results.