***Introduction to Data Mining***

***DATS 6103***

***Final Project***

*Milestone -4-*

**Data Mining Model Build**

## Author

## Michael Wegan G45192209

## Zheying Chen G43061235

## Weiwei Li G27494462

## Xueli Zhou G29955935

## Introduction

Milestones 4 is a stepping stone to complete your semester Final Project. The goals of this millstone deliverable is 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 Data mining Analysis:

1. Describe you Data mining model in details

We mined our data with several different techniques including k-nearest neighbor, neural network, support vector machine, naïve Bayesian, random forest, and logistic regression models. The three best mining techniques for this classification process were the support vector machine, random forest, and neural network.

As the data had previously been cleaned and scaled, there were very few data manipulations required. Most of our energy went into deciding which techniques to attempt, implementing and evaluating them.

We did suffer from overfitting of the data in all the techniques – they worked very well against the training data but only moderately well against the testing data. The implementation of all of the models was relatively straightforward given the available packages in R. Our next steps are to work on parameter tuning and implementing a cross validation process across the three best models to generate a classification probability for all our testing data.

1. List you Data Mining Results

See the attached .html file for the results from each of our modeling attempts.

1. Our goal was to successfully predict the most valuable player (MVP) from the list of all eligible players in a given year. Major League Baseball is divided into two leagues, the National League (NL) and the American League (AL). One MVP is selected from each league every year. Non-pitchers are selected as MVPs at a rate of 87% and are generally evaluated on their offensive statistics. When pitchers are selected as MVP, they are judged on their pitching performance throughout the year. In order for us to predict the MVP in the majority of years, we chose to omit the pitchers

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

1. What models did you use initially?

k-nearest neighbor, neural network, support vector machine, linear regression, naïve Bayesian, random forest

1. Why did you choose those models?

They are commonly used data mining techniques for classification problems. Most of the time, we may not use as many model types but we felt it was important to use as many as we felt comfortable to try so that we could have experience with them. We also implemented some models that we used in class to show that we understood how to use them. Those that were not taught in this class, we have learned elsewhere and thought they could be useful.

1. How did you do this in R?

See the attached .R files.

1. What were the results?

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 positive rates.

We think the failure of these models to appropriately classify some of the testing data, especially those that were false positives, is dual in nature. First, there is a huge class imbalance in our data set. The true positives numbered 166 and the true negatives 41,498. Also, there were 3-5 players that had MVP-like statistics in a given year, only one of which actually won the MVP. Taken across more than 100 years, there are many players that had statistics similar to the true MVPs and did not win. This makes it difficult for any model to discern which individuals at the upper end of the statistical spectrum actually won because there are many other players at a similar level.

For milestone 5, we are going to attempt to tune the model parameters to increase accuracy within each model. Then we will build a probability-based cross-validation framework which is informed by all of the models. When each player is evaluated in every model, we can see if and how many times they were classified as MVP. From this we can assign a probability of them having actually won it. We are hoping that this can help to overcome our problems of class imbalance and clustering of players.

1. How did evaluate different models and their results for your project?

We looked at their rates of correctly classifying the testing data. Ultimately, this is the point of the exercise and correctly identifying the true MVPs is the criteria upon which the model should be evaluated.

1. How do the models or model help you to evaluate your original question?

Our original question was framed around whether we could mine the data to predict the MVP in any given year. These models show that we can, but we also select too many as winners. We are getting too many false positives. At the heart of this issue is that too many players have similar statistical profiles as MVPs. In any year, there are about 300 players that play enough to be legitimate MVP candidates. If we assume that the MVP will be in the top 3% of those players in offensive output, there are 9 players from which 1 is chose as MVP. That makes it very difficult for a model to discern which player is truly the MVP and which is very similar to an MVP – so it chooses both. If we are willing to have false positives, then our models can perform well. I think that once we implement the cross-validation process to use the output from the three best models, we will have fewer false positives.

1. Now that built your model, what data would improve the performance of that model if you could acquire it? (if you had the time)

I don’t think time is the ultimate problem here. Of the players that are the most productive every year, it is really difficult for a model to choose which one will be the MVP. There may not be additional data that can give us a better estimate. The one thing that comes to mind would be a survey of “likeability”. A vote is held by the Baseball Writers Association of America to “elect” the MVP. There are some intangibles that give certain players an advantage. One of these is how the player is perceived by the media and the general public. If the person appears to be friendly and has a good attitude, they may win the MVP even if they didn’t have the best year statistically. This information could help us to further discriminate the MVPs from non-MVPs and reduce the false positive rates. These data would be nearly impossible to gather though. Many of the players and past voters are dead.

1. what are the assumptions of your model?

k-nearest neighbors: data lie in a feature space; data have binary classifier

support vector machine: independent and identically distributed

logistic regression: binary classifier, independent observations, linear relationship between independent and log-transformed dependent variables

random forest: sample is representative

neural network: sample is representative

naïve Bayesian: conditional independence, positional independence

1. What are the constraints?

The support vector machine can only support two dependent classes and cannot handle discrete independent variables. The random forest minimized the issues inherent in the decision tree, but is still greedy – focusing on information gain at the decision point and not considering future information gain for alternate decisions. Also, the trees in the forest can only segregate on axis-aligned splits. The naïve Bayesian classifier is unable to handle correlated continuous data and repetitive categorical data.

1. Limitations?

All of the classifying models are prone to overfitting. The real issue that we have with our data set is a lack of significant segregation of the two classes in multi-dimensional space. The optimal hyperplane in the SVM cannot classify without including many false positives, unless we overfit to a very high degree. At this point, the model would then perform poorly on the test data. This is the crux of the issue with all our models – there is so little separation between the MVPs and high-performing non-MVPs that many of the non-MVPs will always be selected as positives by the models. Maybe through parameter optimization we can balance the overfitting and false positive issues.

1. What are your recommendations to improve model effectiveness?

We are going to try to optimize the models through tuning of the parameters and implement the probabilistic cross-validation framework. This should improve the individual models’ performances and the overall results of the analysis. We may also try to subset the data to reduce the magnitude of the class imbalance and possibly allow for some separation between the best players. This may reduce the number of false positives but will not eliminate them. There is just too much intermingling of non-MVP and MVP winners to develop a perfect hyperplane for separation of classes.

We will also explore other options to reduce the overfitting problems we have encountered. Our assumption is that if we reduce the level of overfitting, we will see fewer false negatives in our models because the selection space will be larger and the boundaries smoother.

1. What are real world implications or possible costs for implementing your model in the real world?

There are no real significant costs for implementing the findings of this analysis in the real world. If we can get reasonable results after amending our models in milestone 5, there could be positive implications, however. Many sports networks and media outlets prognosticate the MVP winners every year. None have models to actually predict them, though. All the predictions are based on the knowledge of the person making the prediction. Comparing their predictions to ours prior and the actual MVP being named would make for interesting media piece, especially if done prior to the MVP being named. Once the MVP is selected the model and media predictions can be revisited for additional conversation. Some sports media companies will probably be interested in a product like this, if only for the novelty and manufacturing of a conversation around the MVP selection. I wouldn’t expect it be extremely valuable as a product, but it would probably get some interest.

1. What factors of data availability or computer processing costs( for example ec2 instances or cloud computing) ?

Some of these models are very computationally expensive, especially when trying to tune the parameters through cross-validation. Running many iterations of the models can tie a machine up for a significant amount of time. Cloud computing or farming the job to a server machine over a secure shell would be valuable if we were doing this as part of our career. It is not reasonable to expect to build these types of models at the speed necessary in the business world using one machine. Optimizing code is also paramount when attempting to build these models. It is easy to build poorly constructed code that runs properly but wastes valuable time. Finally, one should be very comfortable with the types of models they are using so that they can optimize the parameters in an efficient way. A failure to understand how the models operate will lead to wasted time and computational power.

## Evaluation and Summary

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

We have had some limited success mining these data to predict the MVP winners, as evidenced in the confusion matrices in the attached .html file. We believe that we can improve on these results in two ways: model refinement and cross-validation of testing data results across the three models. As we move forward to milestone 5, we are planning to focus on the SVM, neural network, and random forest as they have provided the best preliminary results. However, we can probably get better results through parameter tuning. Once we have fine-tuned the parameters of these three models, we can then compare results across them from the testing data. This will allow us to generate a probabilistic estimate of each player winning the MVP. The cross validation of the final models should reduce the impact of the limitations of each model type and allow for more robust estimates of class for each player.

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 are excited to present our final results in milestone 5.