TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Machine Learning for Basketball Memorabilia Investments

In this assignment you will use real data on past basketball players to use ML to inform memorabilia investment decisions.

You are provided with data on past players that includes 27 different predictors/attributes on each player. Note that the attribute names are often abbreviated and the complete list is included at the end of this document.

You will use the historical data set, Hall-of-fame-train.arff (495 players), to build a classification model predicting whether or not a given player will be inducted (and the probability thereof). You will then apply your model to select players to invest in. The goal of this assignment is to develop hands-on experience in developing and evaluating machine learning models to inform interventions (investment decisions in this case) and to maximize profitability.

The data set includes a target variable (i.e., the class) which indicates whether or not the player has been inducted to the hall of fame. Note that in this data set, we assume that a player who has not been inducted 45.4 years after retirement, will not be inducted. The target variable is binary and has value 1 if the player has been inducted or 0 otherwise. Note that the data includes the actual target variable value for each player. This information is available for this assignment in order to allow you to evaluate the efficacy of investment decisions and to compare amongst investments selected by different models to identify the best model to use in the future.

Step 1: Launching Weka

In the Weka GUI Chooser, select **Explorer.**

Step 2: Exploring the data

For this step, use the Hall-of-fame-train.arff as the training data (open the file from the **preprocess** tab).

2.1 Use the Preprocess or the Visualization tab to examine the distribution of different variables with respect to the target variables. Suggest two attributes that are likely to be predictive of a player's induction to the hall of fame. Describe the pattern you observe to support your choice. (5 points)

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Building, Evaluation and Comparing Different Machine Learning Models

Select the **Classify** tab**.** In the **Classifier** box you can **Choose** a classifier/model. Compare the results of the following models using 10-fold cross-validation on your training set (select **Cross-validation** in **Test options**):

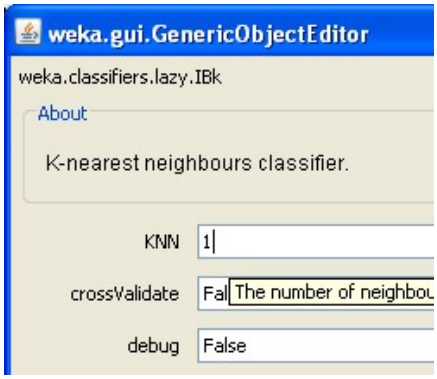
1. Classification **trees** (J48): Click on the **Choose** button. Then, select classifiers→ trees→ j48.

2. Bagging : begin with 30 base models, and you may then explore the performance of bagging with a larger number of models. For the base models, use classification trees (J48). To do so, click on Choose. Select: classifiers→ meta → bagging. To specify the number of models to build and the base modeling technique to use for these models, click on the command line on the right to the Choose button. For **classifier**, click on the Choose button and select : classifiers→ trees→ j48. The numIterations filed specifies the number of models (try 30 or more). You may try a different number of base models and examine whether there is any improvement in performance.

3. Random Forest: classifier → Tree→ RandomForest. Click on the command line to the right of the Choose button. In the form specify the following:
   a. In numFeatures: specify the number of randomly selected attributes to be used to build each model in the ensemble.
   b. In numIterations: specify the number of tree base-models to be induced

   You may explore different numbers for either parameters (there is no need to make an exhaustive search). You may start with 6 attributes and 60 models

4. Naïve **Bayes:** classifiers→ Bayes → NaiveBayes
5. K-Nearest Neighbor: classifiers→ lazy → IBk. By default the number of nearest neighbor used is 1. To change the number of nearest neighbors you would like to use click on command line to the right of the Choose button. The following dialog box will appear:



The KNN field specifies the number of neighbors used for prediction.

Answer the following questions using <u>10-fold cross-validation:</u>

2.2 Using classification accuracy rate as the evaluation measure, which model best predicts whether or not a player will be inducted? (5 points)

2.3 Using the Area under the ROC curve (ROC area), which model yields the largest area under the ROC? (10 poinst)

2.4 Given your goal is to identify and invest in a set of players likely to be inducted to the Hall of Fame, which of the two measures above (classification accuracy rate or ROC area) is more appropriate for selecting the model to inform your investments? Explain your answer. (Note that while classification accurate rate and ROC area may often agree, in some cases the classification accuracy of two models may be comparable, while the ROC area of one may be larger than the other.)

Step 3: Analysis and Model Improvement

3.1 Using an appropriate measure list the three best predictors of a player's future induction to the hall of fame. Here, do not use visualization, as above, but use a relevant measure to quantify and rank the attributes by their predictive value. Explain what measure you used to select these predictors and why. 10 points

3.2 Most predictors reflect career statistics such as points, assists, and rebounds represent a cumulative value, over a player's career. Do you believe this an informative predictor for both young and older players' likelihood of being inducted to the Hall of Fame? Explain your answer. If your answer is no, please suggest how these attributes may be revised such that they useful information to predict a player's merit. If your answer is yes, comment on why the cumulative values may contain predictive information on a player's likelihood of being inducted. 10 points

3.3 Improving the model's performance (20 points)

As discussed in class, an important step in developing predictive models pertains to (a) removing predictors (features or variables) that do not improve or that undermine out-of-sample prediction (through overfitting), (b) adding new information that is not included in the current data set (birth place of a player), or (c) constructing new variables (i.e., Feature Construction). New variables can be constructed by applying transformations on existing variables, such as by calculating the average career points per game from the cumulative number of pointe), or by producing a function using multiple variables (e.g., computing the product or ratio of the number of points and the number of rebounds).

Start with your best model produced so far and aim to improve this model through either removing existing attributes, or the creation of new attributes. See guidelines below on how to produce new variables.

To evaluate whether the changes you have made yield an improvement, examine the difference in performance before and after the change (namely before and after the removal of an existing variable or after the inclusion of a new variable you created). Towards the performance evaluation use the 10-fold-cross-validation procedure and a relevant measure of performance for our problem. In particular, use the measure of performance that you suggested in question 2.4 to be more appropriate for this task.

If you found that the removal of an existing variable improves your model specify which variables were removed. In addition, describe all your attempts to produce new variables (even if the change did not improve the model) and outline what the new variables where and what was the effect on your model's performance.

Using 10-fold-cross-validation procedure report your final (best) model's classification accuracy rate and AUC area.

Guidelines on how to produce new predictors from existing ones:

a) In WEKA, save the file you would like to make changes to as CSV file. This CSV file can be then easily opened in Excel.
b) Open the CSV file in Excel to either make changes to the attribute values (e.g., divide a value by the number of games, etc.), You may also create entirely new attributes derived from existing attributes to produce new and informative predictors that may improve the prediction of whether a player's is likely to be inducted to the HOF.
c) Be sure to provide a name for each attribute you add, and avoid using characters other than English letters in the name.
d) Be sure the file remains as a .csv (do not save as Excel spreadsheet, etc.)
e) Read the file into WEKA and save as ARFF file.

Notice that the class variable is read as numeric (this is because the values for our class variables are 0 or 1). To change it back to nominal do the following:
In the Preprocess tab, click on Filters→Unsupervised→ Attributes →NumericToNominal
Then click on the command line. Set attributeindices to: last
Click OK, and then Apply. This will set the type of the last attribute (class) from Numeric to Nominal. You may want to save the revised file on your hard drive before continuing your work.

Guidelines on how to remove an existing variable:

As done in class, before considering attributes for removal, it is useful to rank the attributes in order of their predictive value and then remove the attributes one at a time in ascending order of their predictive value (starting with the least predictive attribute).

a) In the preprocess tab, check (mark with a √) the attribute you would like to remove.
b) Click on the 'remove' button at the bottom of the screen.
c) To assess the effect, build and evaluate the new model without the variable removed and compare the model's performance after the removal with the performance before the attribute was removed.
d) If your model's performance did not improve, return to the pre-process tab and click on **Undo** (at the top of the screen) to return the attribute which was removed.

Step 4: Investment Decisions

Problem description:

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Assume you have $4,000,000 at your disposal to buy players' memorabilia. Each individual player's memorabilia costs $40,000 (you can buy a maximum of 100 players' memorabilia) when the player is not a hall of famer. At this time, it is unknown whether any of the players in the test set will be inducted. Assume that the NBA association will release the list of hall of famers some time in the future, before which you must decide whether to buy the memorabilia of any given player.

Also assume that if you invest in a player who will later be inducted to the hall of fame, your estimated revenue from selling the corresponding player's memorabilia after the induction is $120,000 (the memorabilia will appreciate by $80K). Also assume that if you decide to buy a given player's memorabilia and to your disappointment the player is not inducted, you will lose 50% of your initial investment in the corresponding player's memorabilia (i.e., a loss of $20,000).

Your goal now is to develop, evaluate, and then apply a memorabilia investment strategy. A strategy here refers to the model(s) you will use and how the predictions produced by the model(s) should be used to select the set of players to invest in (e.g., above what threshold to invest). In practice, the actual outcome for a player is unknown at the time an investment decision is made. Hence, in addition to your choice of predictive model to estimate the probability of induction to the hall of fame, your strategy must also the threshold of the predicted probability, above which it would be profitable to invest in a player (i.e., invest if a player's estimated probability of induction by the model is above T).

Splitting your data set into training and test set

In what follows, your will first split the data into a train and test data sets. Consider the training data as historical data that you were provided with to develop an investment strategy. The test set includes players, the memorabilia of which you are considering to invest in.
You will initially use only the training (historical) data to develop and evaluate different approaches. Once you identify a winning strategy, you will apply it to make predictions for the players in the test set, and select the players that, per your best strategy, it would be optimal to invest in.

Split your data into training and test sets and save each set as a separate ARFF file:
Because the minority class in your data is rather rare, it is recommended to use a stratified sampling approach to splitting your data. This ensures that the proportion of hall of famers in both the training and test sets remains similar to that in the original data set. To do so, follow the steps below:

To do so, follow the steps below:

- Click on Filters
- Select: supervised →instance → StratifiedRemoveFolds
- Click on the command line: StratifiedRemoveFolds and select the following parameters:
- numFolds: 3
- Click OK, and then Apply.
- The result would be your test data. Confirm that the proportion of the minority class remained similar to that of the original data set. Save the sample as your test set (make sure to include the word "test" in the file name).
- Click Undo (button is at the top of your screen). Click on the command line again, and select "True" for Inverse selection.
- Click OK, and then Apply.
- Save the resulting sample as your training set.

-

Open the training data set in the preprocess tab so that it is used to build the model(s). The model you will induce from the training set and select to deploy in practice, will be used to make predictions for players in the test set you created.

Produce models, Evaluate and Deploy a Machine Learning-Based Investment Strategy

Your next step is to select one of the four machine learning approaches considered above to select players to invest in. The approaches are: classification tree, Naïve Bayes, Bagging (choose the parameters that worked best in Step 2) and Random Forest (choose the parameters that worked best in Step 2).

Note that previously, you evaluated models using different measures. however, our focus now is how profitable it would be to use the model to rank and select players for your investment.

4.1 What evaluation measure/plot (i.e., classification accuracy rate, ROC area, lift chart, or profit chart) would be most appropriate in order to select the predictive model based on which to select players to invest in? Why? (5 points)

4.2 Use the training data with 10-fold cross-validation, as well as the evaluation measure/plot you selected in (4.1), to compare between models and to select the model that performs best. Present your results to support your choice of the best model to use. Explain in words how the evidence you present supports your choice of model. (20 points)

4.3 Using the model selected in 4.2, formulate clearly the investment strategy you will use to decide which players to invest in. The strategy should include the following:

a. Which model to use (along with corresponding parameters, such as number and type of models in bagging ensemble, number of K nearest neighbors, etc.)

b. Given your model's estimated probability of induction, above what probability threshold to invest in a player's memorabilia.

4.4 Deploy your strategy to select NBA players from the test set:

First, generate predictions for players in the test set and then use these predictions to identify an advantageous probability threshold to use towards your future investments

To do so:

In the **Classify** tab, first choose the model

a. Then, for your **Test Option**, select "**Supplied test set**" and click on the **Set** button to the right. Click **Open File** and select test set file . Click on **Open**, and then click on **Close**.

b. Because you want to see the probability predictions produced by the model so as to decide which of players to invest in, you would need to select the option to output the model's individual predictions for each customer. To do so, click on the "**More options**…" button in the **Test Options**. Then click on the "**Choose**" button to the right of the text "**Output prediction**" . Select **CSV**, and then click **OK**.

c. You are now ready to generate your model's predictions for the new data set. Click **Run**.
The model's prediction for each players in the test set are listed in the result buffer to the right. Scroll up to see the table of predictions (following the text : === Predictions on test set ===). Each player's prediction is listed in the order in which the players appears in the test data set.

The classifier's output includes the following columns:

- Example number
- Actual class
- Predicted class
- Error (+ is the model was incorrect)
- The model's estimated probability of the predicted (most likely) class (i.e., the class predicted by the model and that is listed in the Predicted column). Note that, as in assignment 2, to review the probability of induction for all players, you would need to calculate the complementary probability for players predicted not to be inducted to the hall of fame (for these players, the probability displayed is the probability of not being inducted).

e. From the result buffer on the right, copy the table with your model's predictions into a **new** (blank) text editor file (e.g., Microsoft Word file).

f. Copy the table with the test examples' predictions to a text editor. Remove the following text from the document:" *" , "+" , "1:" , "2:" (it's easy to use "find and replace" and leaving the "replace" text empty) .

g. Save the file a **simple text file**. In addition, if need to, change your text file extension to .txt

h. From Excel, open the text file. Specify the file is space delimited (first specify that the file is delimited, then check space).

The table now will include the following columns from left to right (you may need to fix the column names after opening the text file from Excel)

- Example number
- Actual class
- Predicted class
- The model's estimated probability for the most likely class

4.5 Print the list of players (along with their Example number) you will invest in, and report your profit. (15 points)

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Descriptions of Attributes

Field	Description
League	N = National Basketball Association (NBA); A = American Basketball Association (ABA)
Games	Games Played
minutes	Minutes Played
pts	Points
offReb	Offensive Rebounds
defReb	Defensive Rebounds
reb	Rebounds
asts	Assists
stl	Steals** SEE NOTE BELOW
blk	Blocks** SEE NOTE BELOW
turnover	Turnovers** SEE NOTE BELOW
pf	Personal Fouls
fga	Field Goals Attempted
fgm	Field Goals Made
fta	Free Throws Attempted
ftm	Free Throws Made
tpa	Three Pointers Attempted* SEE NOTE BELOW
tpm	Three Pointers Made* SEE NOTE BELOW
totalSeasons	Total number of seasons played. This value is calculated as follows: "lastSeason - firstSeason + 1"
Position	C = Center; F = Forward; G = Guard
firstSeason	First season played. The year corresponds to the first year of the season (i.e. a value of 2000 represents the 2000-2001 season).
lastSeason	Last season played. The year corresponds to the first year of the season (i.e. a value of 2000 represents the 2000-2001 season). Note that 2004 (2004-2005 season) is the last year for which there is data.
careerEnded?	Boolean field for if the player's career has ended (1 if career has ended, 0 otherwise). This field was calculated as follows: if the "lastSeason" field is earlier than 2004, the value is 1, otherwise 0. Note that this calculation naively assumes that no players retired at the end of the 2004 season.
yrsRetired@2004	The number of seasons that a player has been retired as of the 2004-2005 season.
Class (HallofFamer)	A Boolean field showing whether or not a player was inducted to the Basketball Hall of Fame (HoF) as a player. This field has value 1 if the player has been inducted, and 0 otherwise.

* The NBA did not have three point shots until 1979

** Steals, Blocks and Turnovers were not recorded in the NBA until 1973