Orange Data Analysis

Overview

This project requires that you use the tools learned throughout this portion of the course to create a model for a real-world situation – creating a model to predict the success of NBA teams.

For the current 2021-2022 NBA season, use your advanced analysis skills to predict the number of wins for each team and what NBA teams will make the playoffs.

You will need to prepare two Excel files – the original Data Set and the Prediction File. Your goal is to use the past set of statistics to build the following two models for the current 2021 – 2022 season:

**Model 1:** Classification Model (Playoff Status – “yes” or “no”)

**Model 2:** Prediction (numeric) Model (Number of Wins)

IMPORTANT NOTE: Detailed instructions on how to prepare the Data Set and Prediction File are provided in the Appendix of this assignment.

Part 1: Collect and Prepare Data Set

* The Data Set will include 5 previous seasons of data (2020-2021) and prior. This will have all the statistics from the table (including wins and playoff status), but with the win-proxy statistics removed. (Refer to the Appendix for further information)
* You will start with the same Data Set for each of the Classification and Numeric prediction workflows in Orange.
* Classification Model (Playoff Status) - skip Wins column when you are developing the Orange algorithms (models)
* Prediction Model (Number of Wins) – skip Playoff Status when you are developing the Orange algorithms (models)

Part 2: Test and Score Data Various Models

Once your Data Set is prepared, you can begin using Orange to create and test your models. Use the same approach we used in class. (include all screenshots – see Deliverables section)

Part 3: Increase Accuracy of the Models

To increase the accuracy of the models, there are two approaches you can try:

1. **Change the Roles of the Fields in the Data Set** – Skip some Fields in the Data Set. You can also add Fields of your choice to the Data Set if you feel it improves the accuracy of the model. Track the changes made to the Role of the Fields in the original Data Set using the following table format. Be sure to note the original Role of all the fields in the table before you start making changes. Your actual table will be included in your paper. You can find the information to complete this table by double-clicking on your first file in Orange. (You will need to upload your Data Set first)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field Name | Type | Role (Original) | Role (Revised) | Why is this included (or skipped?) |
| EXAMPLE (not related to project): |  |  |  |  |
| AGE | Numeric | Feature | Feature | Age of the home is a major factor in the price of a home. |
| OR  AGE | Numeric | Feature | Skip | Age is not a factor in the price of a home. |

1. **Change the Settings on the Orange Models (Algorithms) Used** - You only need to complete this chart for the settings you changed in the Orange Models (Algorithms). Be sure to note the original setting in the table before you start making changes. Double click on the pink model widget to access the settings for that algorithm. Your actual table will be included in your paper.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model (Algorithm) Name | Setting Description | Original Setting | Revised Setting | Result |
| KNN | Number of Neighbors | 5 | 6 | Accuracy Improved |
| **OR** |  |  |  |  |
| KNN | Number of Neighbors | 5 | 6 | No impact on Accuracy |

As you are going through this process you must take note of what method you are using, what changes you make to the model data, and why you are making those decisions. You will need to present both your model, and the reasoning of why you built it as you did and why it is superior to the alternatives that proved to be less accurate. The process of developing your model is the most important part of this process, so ensure you are making logical improvements and documenting the reasoning and impact. Include supporting screenshots if applicable.

Part 4: Use the Prediction File to make Predictions for your model.

Prediction File:

• Use the current season (2021-2022). This will have the same set of statistics as the original data set, with the exception that the wins and playoff columns should be blank – this is what we are predicting. For the classification prediction in Orange, we will use the data to predict the playoff status (“yes” or “no”) and for the numeric prediction in Orange we will predict the number of wins. You can use the same prediction file for both the Classification and Prediction (numeric) Workflows. Include supporting screenshots.

**Overall Goal**

If you’re at all unclear, the main goal of all of this is to:

• Use past experiences – all the seasons that have been completed, along with all of the stats collected during those seasons to create two predictive models, one that guesses the number of wins, and another that guesses if a team makes the playoffs or not.

• Use those models to make predictions for the most recent season, where we (pretend we) don’t know the real number of wins or playoff status, but we do know all the stats.

Grading Deliverables

The deliverables for this project are:

• A formal paper that contains the results of your modeling exercise, the explanation of what you did, and the reasoning behind why your model is the best.

o The paper MUST have an executive summary at the front end that contains the results (for each of the wins prediction and the playoffs classification):

♣ Clear and labelled screenshots of the Orange Workflow, Data Sampler, Test and Score results, and Prediction Results for each of the Orange Algorithms (models) used.

♣ A table & supporting screenshots of your features included the original data set, along with a brief (< 1 sentence) statement of why that is included. (See Example above)

• In the table, if you’ve removed anything from the original feature set, the note should give a reason why.

• Reasons such as, “It is obviously irrelevant”, “\_\_\_\_\_\_\_ analysis indicated this feature should be removed”, “accuracy went up when this was added” are acceptable.

♣ A table & supporting screenshots stating what changes were made (to the Orange Algorithms (models), along with justification. (See Example Above)

o The remainder of the paper should be a brief discussion of your process and your findings. o The purpose of your paper is to present your process and results in a clear, understandable way.

• Your two Orange Workflow file(s) and two Excel files(s).

**Note:** As a very rough estimate, the written explanation, justification, and improvements will probably end up around 2 pages for most groups. The screenshots and tables will add to that, and those will make the overall length vary. Please do not aim for length, aim to provide the requested information.

\*You will only submit one Excel Data Set and one Excel Prediction File.\*

Appendix: Data Set & Prediction File Preparation

Your goal is to take the data on basketball statistics and clean it to the point that it may be analyzed using the data analysis techniques that we covered in class (Orange). The key goals of cleaning up this data are:

1. Assembling the data in a datasheet format:

a. One header row, with clear headers.

b. One instance (team’s season) per row.

c. Translation for making the playoffs (see more below).

d. A prediction file, containing the most recent season (2021-2022) on which the actual predictions will be made.

The source for our data is Basketball Reference (https://www.basketballreference.com/leagues/NBA\_2021.html There are links at the top of the page to go backwards/forwards for the year. The exact data used to build your model is something that you must determine – the starting point is to download the Advanced Stats table into Excel. A sample screenshot of the 2020-2021 Advanced Stats is shown below. You can download the Excel data by clicking on Share and Export, and then Get as Excel Workbook.

You may need to modify column names, or otherwise clean the data to prepare it for analysis.

**Playoff Status** – Teams that made the Playoffs have a \* beside their team’s name. You need to create a Playoff column that holds a variable denoting if the team made the playoffs (Yes = playoffs, No = no playoffs). Once you’ve assembled your sheet with the 2020-2021 data, repeat the process for at least 4 previous seasons, so you have 150 (30 Teams with 5 Years of information) total rows of data on which to build the model.

**Playoff Status Column**: =IF(ISNUMBER(FIND("\*",B2)),"Yes","No")

**Important Note**: When you’ve created the sheet, you must remove the Win-proxy stats. These are L, PW, PL, MOV, and SRS. There may be others depending if you’ve added items to your model. Ensure you double check this as inclusion of any win/loss statistics will ruin the accuracy of your model when we get to the project.

Prediction File To prepare your prediction file you must redo the same steps in setting up your original Data Set for the most recent full seasons – 2021-2022. Delete the Wins (W) column and the Playoff Status columns as those are our predicted values. Place this into another workbook.