Data Scientist @La Javaness
Hiring test

TENNIS MATCH PREDICTION

THE DATA

The website http://tennis-data.co.uk/alldata.php gathers outcomes of both WTA (Women Tennis Association) and ATP (Association of Tennis Professionals - men only) tennis games over several years. A brief description of each variable can be found here:

http://www.tennis-data.co.uk/notes.txt

This test comes with a Jupyter Notebook called "starter-kit.ipynb". Executing all the cells should provide you with:

- one pandas dataframe for all ATP matches (df_atp)
- one pandas dataframe for all WTA matches (df_wta)

WHAT IS EXPECTED OF YOU

First, we would like you to answer the questions on page 2.

Then, we want you to predict the outcome for each ATP tennis game during the year 2017, using one or two models of your choice.

To do so, we would like you to send us a self-contained Python notebook that handles:

- the data preparation process
- the modeling phase
- a discussion of the performances of your model(s)
- any observation or explanation you deem interesting

INDICATIONS

Your notebook will be evaluated on the cleanliness of your code, the clarity of your explanations, the rigor of your reasoning and your answer to the questions – not on how sophisticated your model is!

Extra points for:

- notebooks that can be understood without having to read every line of code
- one or two graphics that are both interesting and pretty

QUESTIONS

Please answer the following questions about the dataset with the appropriate line(s) of code.

Example

Question ⇒ How many ATP matches are there in the dataset?
Answer ⇒ len(df_atp)

1. Who are the three ATP players with the most wins?
2. How many sets did the player “Federer R.” win in total?
3. How many sets did the player “Federer R.” win during the years 2016 and 2017?
4. For each match, what is the percentage of victories of the winner in the past?