Brian Johns - Capstone Project README File
April, 2022
BrainStation

File List

Text File For Environment Set-Up:
requirements.txt

Jupyter Notebooks
Brian Johns Capstone #1 - Data Cleaning.ipynb
Brian Johns Capstone #2 - Regression Modeling.ipynb
Brian Johns Capstone #3 - Clustering Models.ipynb
Brian Johns Capstone #4 - Findings.ipynb
Brian Johns Capstone - Appendix NB#2 Copy with Grid Searches

CSV Files Needed For Notebooks
CLEAN_CAPSTONE_032522.csv
CLEAN_CLUSTERING_032522.csv
VISUALIZATION_DATA.csv

CSV Files of Data - To be kept in *data* folder
These files were downloaded from evolving-hockey.com and requires a subscription to download from the site:

EH_gar_ratings.csv
EH_gar_totals.csv
EH_standard_boxscore_ratings.csv
EH_standard_boxscore_totals.csv
EH_standard_onice_ratings.csv
EH_standard_onice_totals.csv
EH_xgar_ratings.csv
EH_xgar_totals.csv

PDF Files
Brian Johns Capstone - Glossary_Hockey Statistics
Brian Johns Capstone - README
Brian Johns Capstone - Report

Tableau
Capstone_Viz.twb

Other Notes:
CapFriendly website was used to scrape the financial data for the NHL players

Project Data

There are two sources of data for this project.

1. CapFriendly

I webscraped the information off of this page in order to acquire it for my data set. The code that I used to scrape the website is in the ‘Brian Johns Capstone #1 - Data Cleaning’ notebook.

On March 25th, I ran this notebook and output a csv file: *CLEAN_CAPSTONE_250322.csv.* This data is what I use for the rest of my modeling. This file has been included in the submission and should be kept in the same folder as the notebooks.

If running through the notebook, the end of it will output a new csv file which will overwrite the csv that is used for modeling. I have commented out the cell to ensure the data that is used for modeling is not overwritten prematurely.

2. Evolving Hockey

I downloaded this data directly from the website in .csv format. It is taken from several pages, so there are 8 csv files that are uploaded for this project. These files have been included in the submission.

The directory for this data should be ‘data/eh/’. The data folder should be in the same location as where the notebooks are being run. There should be no other files kept in this folder.

Environment

This project will run on the base environment for python, with the exception of using lxml, XGBoost and SHAP. A requirements.txt file has been provided to run this environment. The following commands can also be input in the terminal to set-up the environment.

```
conda create -n bj_capstone python=3.8 numpy pandas matplotlib seaborn statsmodels scikit-learn=0.24.1 jupyter jupyterlab
```

```
conda activate bj_capstone
```

```
conda install -c conda-forge shap
```

```
conda install -c conda-forge xgboost=1.1.1
```

```
conda install -c conda-forge lxml
```

```
ipython kernel install --name "bj_capstone" --user
```

Project Walkthrough

Notebook #1: ‘Brian Johns Capstone #1 - Data Cleaning’

- Running the notebook in full will output a .csv file titled ‘CLEAN_CAPSTONE.csv’ in your data folder. Given the nature of live-scraping information from CapFriendly, this data is NOT be used for the Regression or Cluster modeling
- A .csv file has been provided that the modeling is run on: ‘CLEAN_CAPSTONE_250322.csv’. Ensure this file is in the same folder as where you will be running the notebooks.
- Ensure that all Evolving Hockey .csv’s are located in a ‘data/eh’ directory with the data folder in the same location as the notebook
- The Webscraping for this notebook does take some time, around 10-15 minutes on my computer. Otherwise the notebook should run quickly.

Notebook #2: ‘Brian Johns Capstone #2 - Regression Modeling’

- Prior to this notebook I suggest going through the pdf titled ‘Brian Johns Capstone - Glossary Hockey Statistics’. This will provide a basic understanding of the stats that are included in the data and analysis and to better understand my explanations and processes.
- Ensure that the ‘CLEAN_CAPSTONE_250322.csv’ is located in the same folder as where this notebook is running
- Tuning Hyperparameters through GridSearch
    - I ran numerous grid searches (approximately 50) in order to tune my hyperparameters, I was given tips on how to shorten the run-time of my notebook after I had gone through everything and I wanted to keep my work and thought process.
    - I copied my full Regression Modeling notebook and turned it into ‘Brian Johns Capstone - Appendix NB#2 Copy with Grid Searches’ and have removed the GridSearches from this finalized notebook
    - The notebook with the full Grid Searches will take several hours to run. However, my full notation and process is documented in the Appendix on how I came to my optimized Hyperparameters are outlined in detail in that notebook.
- Running the notebook in full will output a .csv file titled ‘CLEAN_CLUSTERING.csv’ in your data folder. Again, to ensure there are no changes in the data from what I ran the modeling on, this file will NOT be used for the next notebook
- With the Grid Searches removed from this Notebook, this notebook should run fairly quickly, less than 10 minutes.

Notebook #3: Brian Johns Capstone #3 - Clustering Models

- Ensure that the ‘CLEAN_CLUSTERING_032522.csv’ is in the same folder as the one you are running this notebook.
- Since this is a cluster model, if you run the model then the results in the notebook will change. Please read through my notes prior to running the notebook.
- This notebook exports a ‘VISUALIZATION_DATA.csv’ into the data folder.
- The ‘VISUALIZATION_DATA.csv’ file that I have provided should be kept in the same folder as where the notebooks are run. No modeling is run on this data and it is only used to supply the data for my Tableau visualizations. This csv is not needed for any of my notebooks
- There is some hyperparameter tuning to visualize TSNE in this notebook. Run time was approximately 45-50 minutes.

Notebook #4: Brian Johns Capstone #4 - Findings

- There are no special instructions for this notebook
- This notebook is purely a written summary of the findings I had for my project