PYTHON FOR DATA ANALYSIS

ESILV – devoir 2021

TRAN-HONG Nicolas

Summary

- Introduction of the dataset
- Objective with this dataset
- Variables importance for the prediction
- Exploratory Data Analysis
- Modelisation
- Model Comparison
- API Flask of our Machine Learning model
- Conclusion

Introduction

- The dataset : Skillcraft1 Master Table Dataset https://archive.ics.uci.edu/ml/datasets/SkillCraft1+Master+Table+Dataset#
- This dataset was created to study human expertise in a specific field.
- Here, the domain of expertise is in a video game called « StarCraft II ».
- This game is a Real-Time Strategy (RTS) game in which players positions structures and maneuvers multiple units to destroy their opponents' headquarter.
- It requires especially skills and knowledge to play well.

Ingame image

- Because differences in skill can be important in this game : there is a matchmaking system that enables to play against players of their level.
- Players are spread over several distinct levels of skills called leagues (Bronze, Silver, Gold, Platinum, Diamond, Masters, GrandMaster).
- Because the dataset contained very few players from the League Index 7 (GrandMaster, which corresponds to the Top 200 players from each region). Data from 55 professional players were added. They were attributed the League Index 8.

-> we add these professional players to the league 7 to have enough data for the GrandMaster

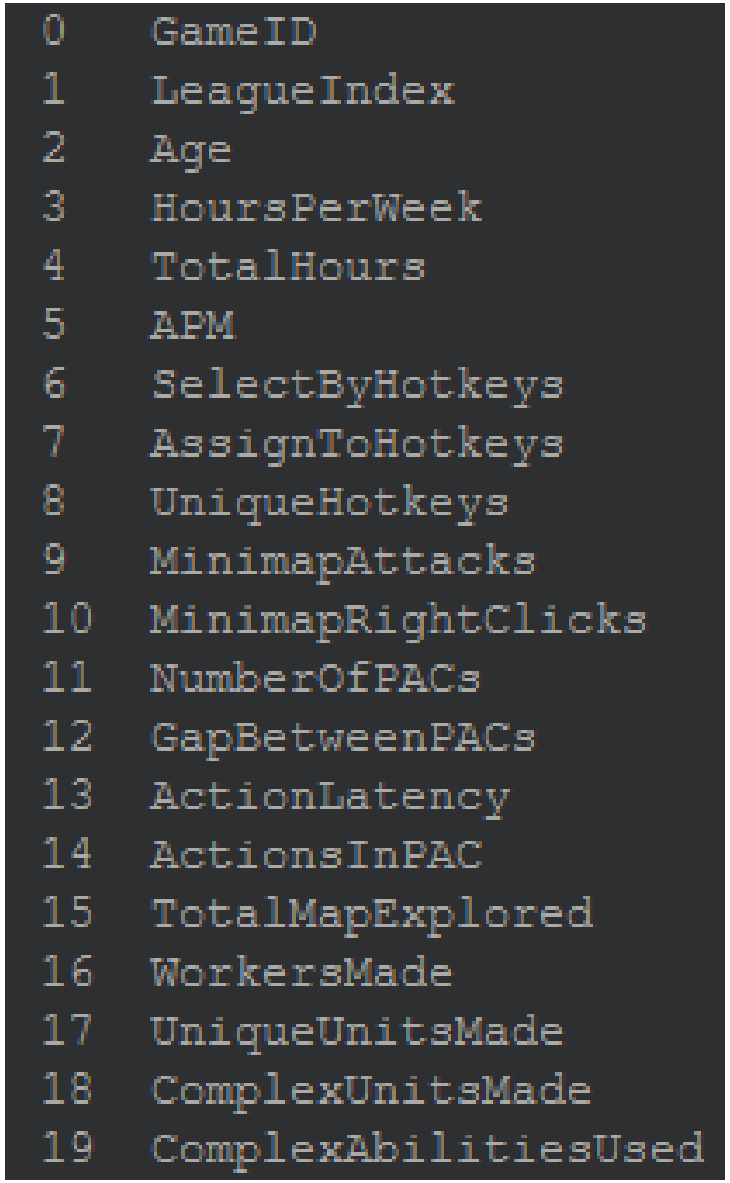
- This dataset contains telemetric data obtained from over 3000 game replay of players from all these leagues.
- The selected variables are related to cognitive-motor abilities.
- Some variables can be classified into these categories :

  - Perception-Action-Cycle variables

  - Hotkeys usage variables

  - Complex unit production and use variables

  - Minimap variables



```
0    GameID
1    LeagueIndex
2    Age
3    HoursPerWeek
4    TotalHours
5    APM
6    SelectByHotkeys
7    AssignToHotkeys
8    UniqueHotkeys
9    MinimapAttacks
10   MinimapRightClicks
11   NumberOfPACs
12   GapBetweenPACs
13   ActionLatency
14   ActionsInPAC
15   TotalMapExplored
16   WorkersMade
17   UniqueUnitsMade
18   ComplexUnitsMade
19   ComplexAbilitiesUsed
```

Objective

- Predict the league placement of a Starcraft II player by using the data that can be obtained from a game replay
- Target feature : League index from 1 – 7 (Bronze – GrandMasters)
- Multiple labels classification problem

Variables importance

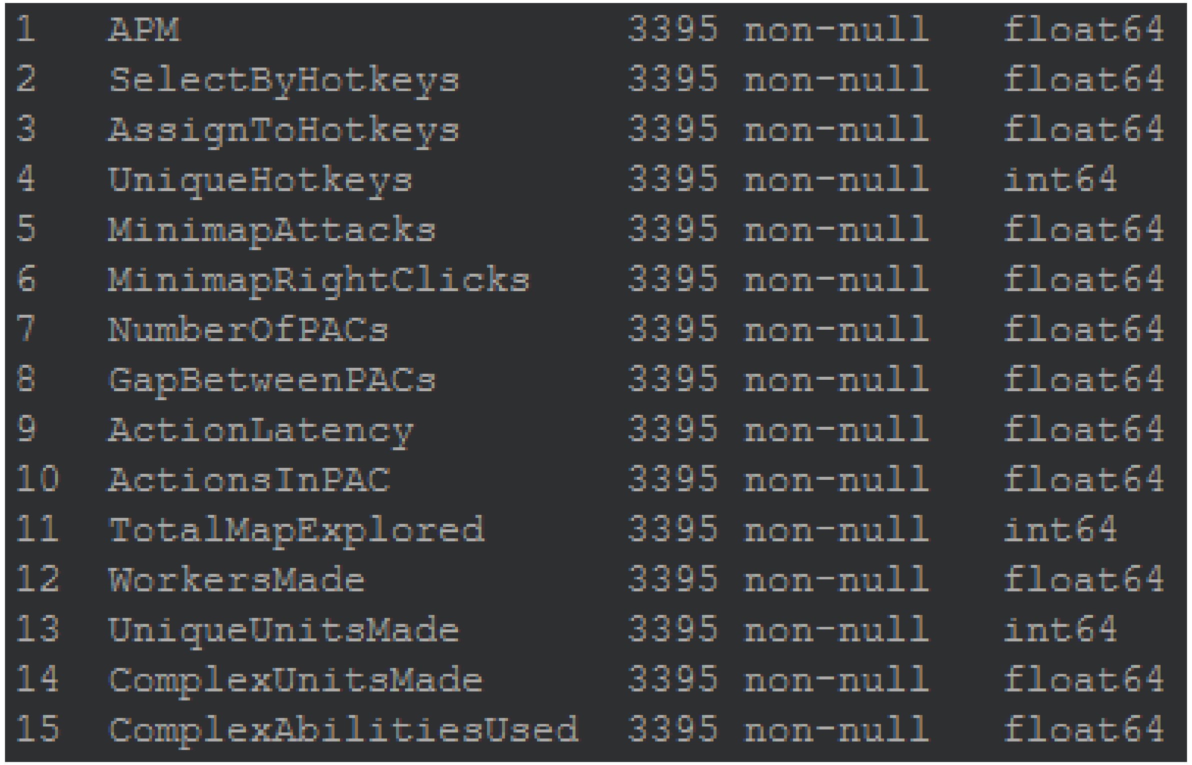
- Before doing some Exploratory Data Analysis (EDA), a question must be asked.
- Are all variables important for the prediction ?

-> Because the expertise is the center of the study, we doesn’t need variables like “Age”, “HoursPerWeek” and “TotalHours”(they can still be useful for the EDA).

-> “GameID” is useless for the prediction and the EDA

We will then drop these variables.

15 variables for the prediction



```
1    APM                   3395 non-null   float64
2    SelectByHotkeys       3395 non-null   float64
3    AssignToHotkeys       3395 non-null   float64
4    UniqueHotkeys         3395 non-null   int64
5    MinimapAttacks        3395 non-null   float64
6    MinimapRightClicks    3395 non-null   float64
7    NumberOfPACs          3395 non-null   float64
8    GapBetweenPACs        3395 non-null   float64
9    ActionLatency         3395 non-null   float64
10   ActionsInPAC          3395 non-null   float64
11   TotalMapExplored      3395 non-null   int64
12   WorkersMade           3395 non-null   float64
13   UniqueUnitsMade       3395 non-null   int64
14   ComplexUnitsMade      3395 non-null   float64
15   ComplexAbilitiesUsed  3395 non-null   float64
```

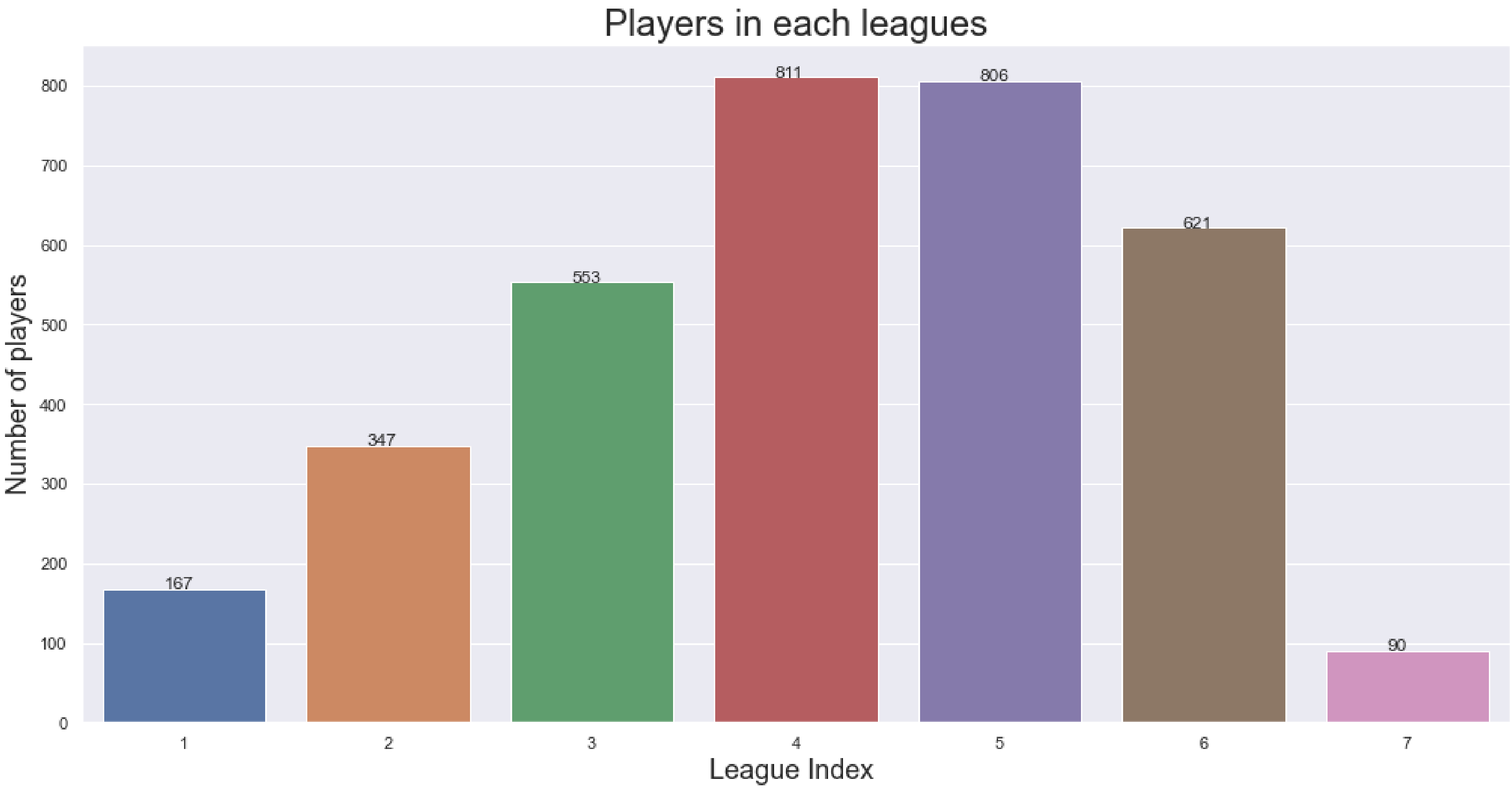
Exploratory Data Analysis

Distribution of the target feature :

LeagueIndex

Most players of the dataset are in high leagues.

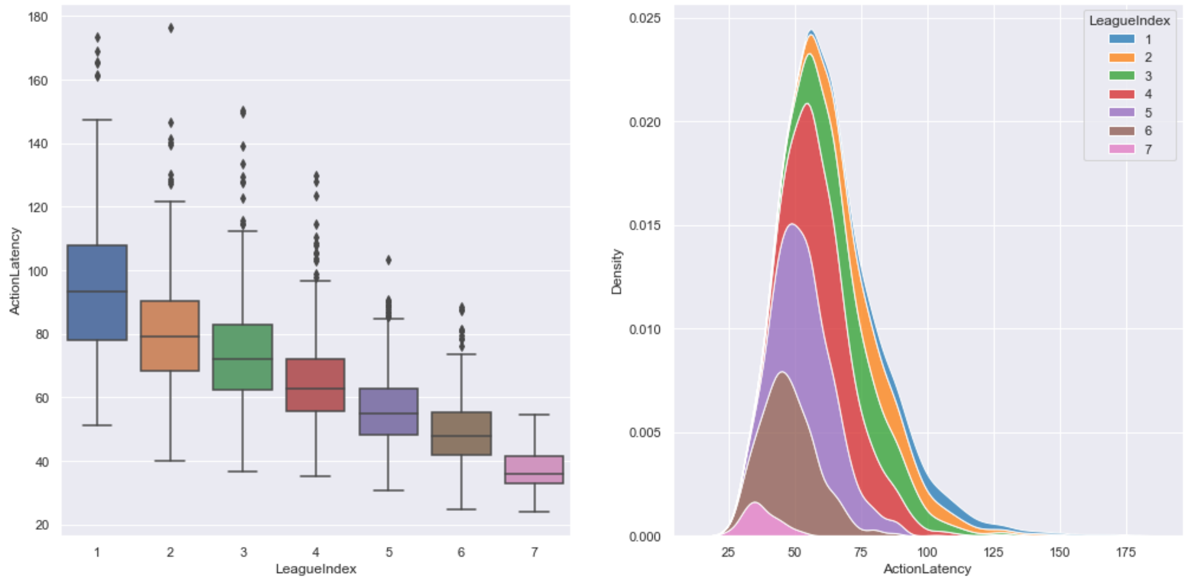
It is not surprising because data were obtained by asking game replays on online gaming communities and social media, where most determined players spend time on.



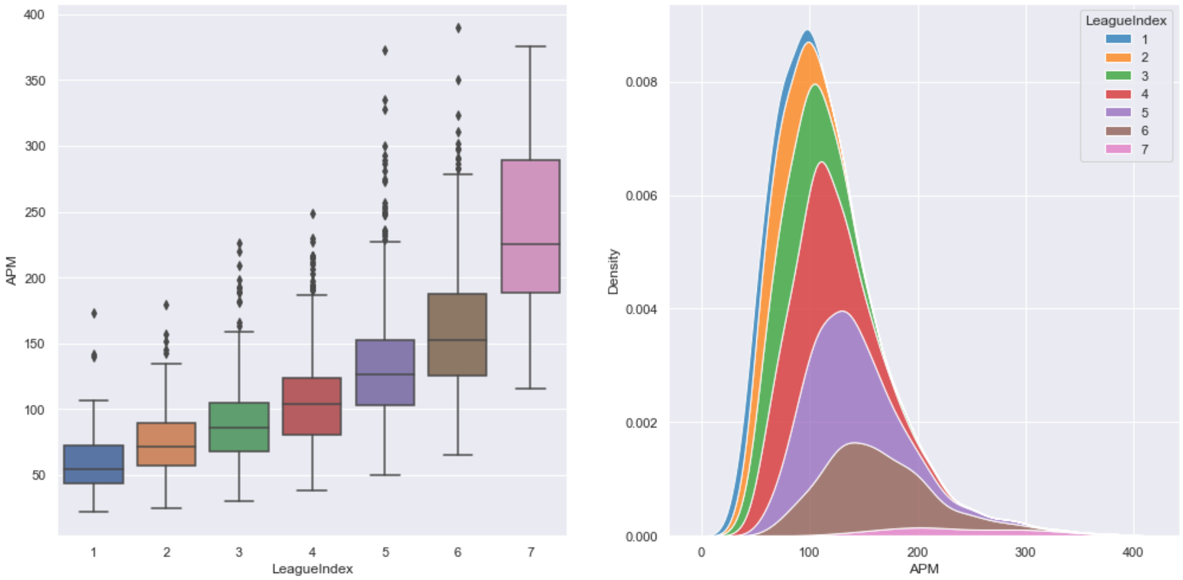
Thanks to an EDA, we can know that some variables will be very important for prediction.

Actions per minute (APM) and Action Latency (which is related to the reaction time in game) are some of these.

Both variables describe well the skill of a players and ease the comparison between low and high-skill players.



Action Latency By Leagues

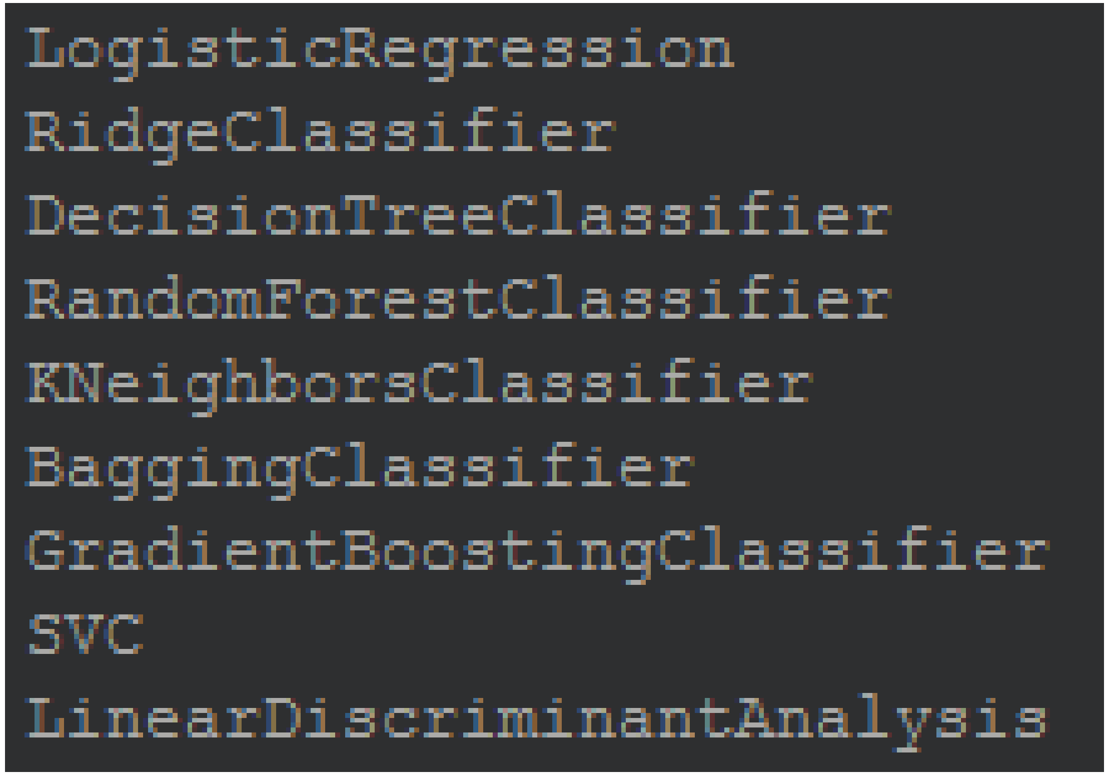


APM By Leagues

Modelisation

- Here we have a multi classes classification problem
- We will use these different algorithms
- For each algorithm : thanks to a grid search, we try to obtain the best hyper parameters possible.
- But before that we have to split dataset into a training set and a test set to validate whether our models will generalize well to new data.

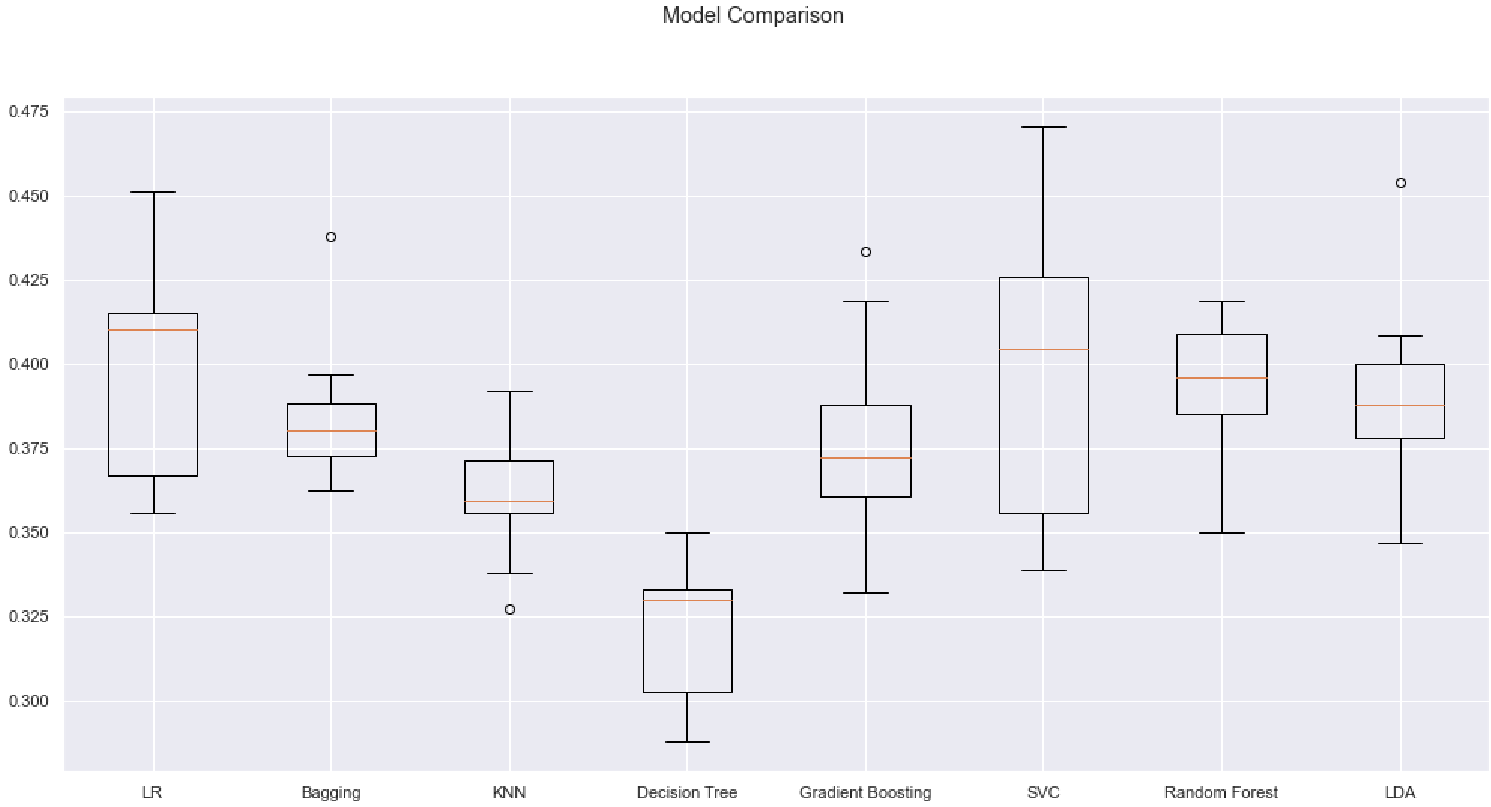
Algorithms used :



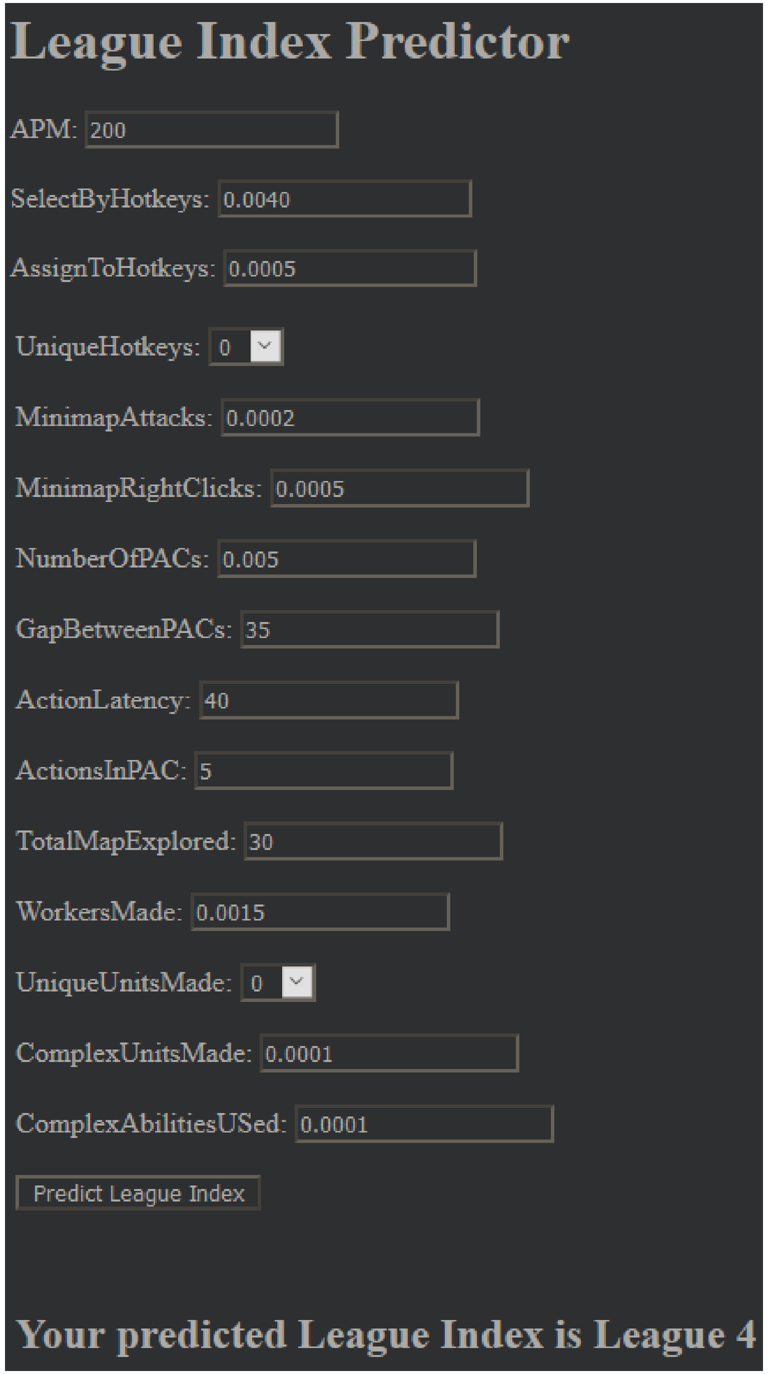
```
LogisticRegression
RidgeClassifier
DecisionTreeClassifier
RandomForestClassifier
KNeighborsClassifier
BaggingClassifier
GradientBoostingClassifier
SVC
LinearDiscriminantAnalysis
```

Model Comparison

- The following boxplot shows the spread of accuracy scores across each cross-validation fold for each model
- It seems that Logistic Regression and SVC are the most interesting models and worthy of further study for this problem.



API Flask of our ML model



League Index Predictor

APM: 200

SelectByHotkeys: 0.0040

AssignToHotkeys: 0.0005

UniqueHotkeys: 0

MinimapAttacks: 0.0002

MinimapRightClicks: 0.0005

NumberOfPACs: 0.005

GapBetweenPACs: 35

ActionLatency: 40

ActionsInPAC: 5

TotalMapExplored: 30

WorkersMade: 0.0015

UniqueUnitsMade: 0

ComplexUnitsMade: 0.0001

ComplexAbilitiesUSed: 0.0001

Your predicted League Index is League 4

- League Index Predictor, which uses the model I think will predict the best.
- Following the model comparison, I used the Logistic Regression model.
- So enter your data to predict your league index !

Conclusion

- Thanks to these RTS game replays, we can obtain interesting data to study expertise and cognitive science in general.
- We can have a better idea on the important variables like APM and Action Latency to predict League Index.
- However, the dataset is a very small proportion of the existing player base. It is mostly composed of active players. We lacked players in some leagues.
- Our model is far from being accurate. To improve it, we can do some serious variable selection and also deepen the grid search of the hyper parameters.