FINAL PROJECT SKILLCRAFT

Python for data analysis

Brief presentation

- SkillCraft is a dataset composed of a lot of features of over three thousand players playing at StarCraft 2 from bronze to professional gamers. In our study we will try to predict the league index of a player considering all his others features. Thus, it is a classification problem.
- The league index represent the level of a player between 1 and 8 (for example from Bronze to GrandMaster).

Data Exploration

- The first problem was the type of the features. Indeed, we have **Age**, **HoursPerWeek** and **TotalHours** which was detected as object, but they are numeric values. So, we have parse them to numerical values.
- Then, we check values quality. We discover that we have NaN values on the three previous columns. Indeed, all the Professional leagues players (8) misses their age, their total hours and their HoursPerWeek. Delete those observations was not relevant because knowing about professional players could be really helpful. So, we have just **changed those value to their respective mean value**.
- Finally, we had a look on the statistics of each features. We discover that someone has been playing 168 hours/week which means he/she plays 24h/24 and another playing 1 000 000 hours which means he/she played more than 100 years. Thus, we delete these observations.

Data Exploration

- Then, we had a look on the features distributions. We observe a **lack of data for the class 7 and 8.** Moreover it seemed important to **normalize and centralize the data.**
- By analyzing the correlation matrix we found that **ActionLatency, APM, NumberOfPACs** and **GapBetweenPACs** could be good predictors because they have hight absolute correlation. We confirm it by analyzing the box plot and the KDE plot of these features in function of LeagueIndex.
- Finally, we display a **scatter matrix** of these features and discover pretty linear relation which motive us next to build **logistic regression**.

Data preprocessing

- Here we start to try log transformation on different columns.
- Then, we split the dataset in train and test set.
- We normalize and centralize the data.
- We try to use polynomial features but we had a lower accuracy.
- We also try to apply a transformation to make the datas more normal but it completely destroy the information and we had a really really bad accuracy.

Modeling

- We start with a basic LDA model. Then we try other models like logistic regression, random forest or KNN.
- For each model we compute his accuracy, the R^2 and the MSE.
- For the logistic regression model we plot the coefficient of each features by class and discover something:

Age and HoursPerWeek see their importance boosted for class 8 while we set those features to the mean for this class. This is falsifying the predictions for class 8 as the model will tend to predict class 8 for all the players whose Age or HPW are near the mean. We shall drop Age and HPW.

For the rest MinimapRightClicks, CompexUnitsMade and UniqueUnitsMade are always in the least important for all classes meaning dropping them might optimize our model.

Modeling

- So we dropped unrelevant columns and we obtained a better score !
- Then we have compared all the metrics of our models and decide to choose the logistic regression which was the best.

Tuning

- When we have defined the models we also try to optimise the KNN model with a grid search on n_neighbors, weights and metric. It improve the score but was not good at all.
- We choose to try a grid search on the logistic regression to optimise the solver, the penalty and the parameter C.
- Finally, we build a pipeline with the scaler and the optimized logistic regression model and we saved it.

Remarks and reflections

- We note a significant lack of data, especially for the class 7 and 8 which leads to the model having a harder time making right prediction for those classes? Also we misses 3 parameters for all the class 8 observations, maybe those parameters would have been relevant.
- We noticed that we have only have mechanical skills (clicks, PAC informations…) which might just depends on the player strategy and/or computer quality. We should ask for more personal informations about the players (Married ? Type of work ? …)
- To conclude, we were surprised that the league index didn’t depend that much on total hours played and more on technical skill like Action Latency.