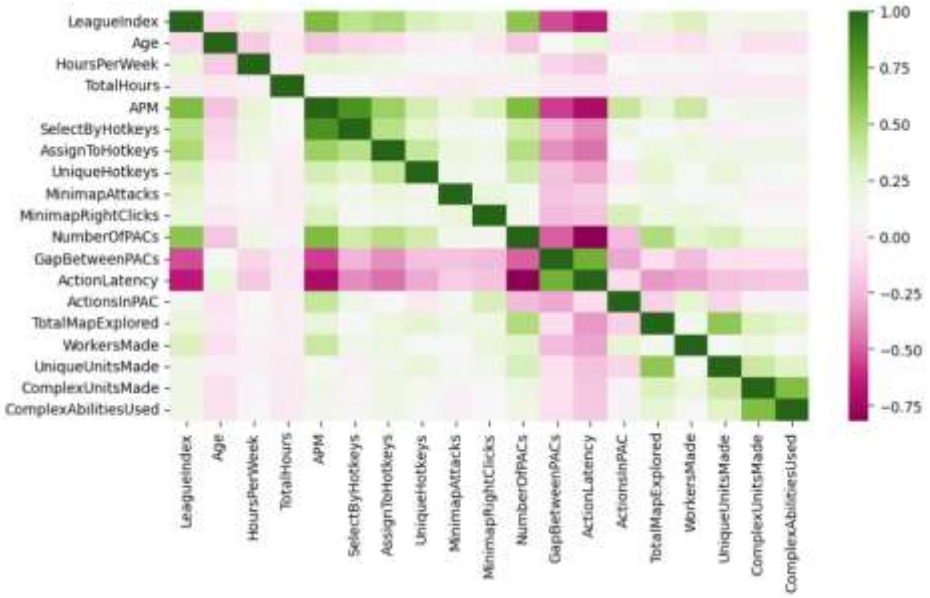
SUMMARY

- **Data characteristics:** The Dataset consists of 3395 rows and 20 columns. LeagueIndex is treated as the ranking of the player and is the target variable.

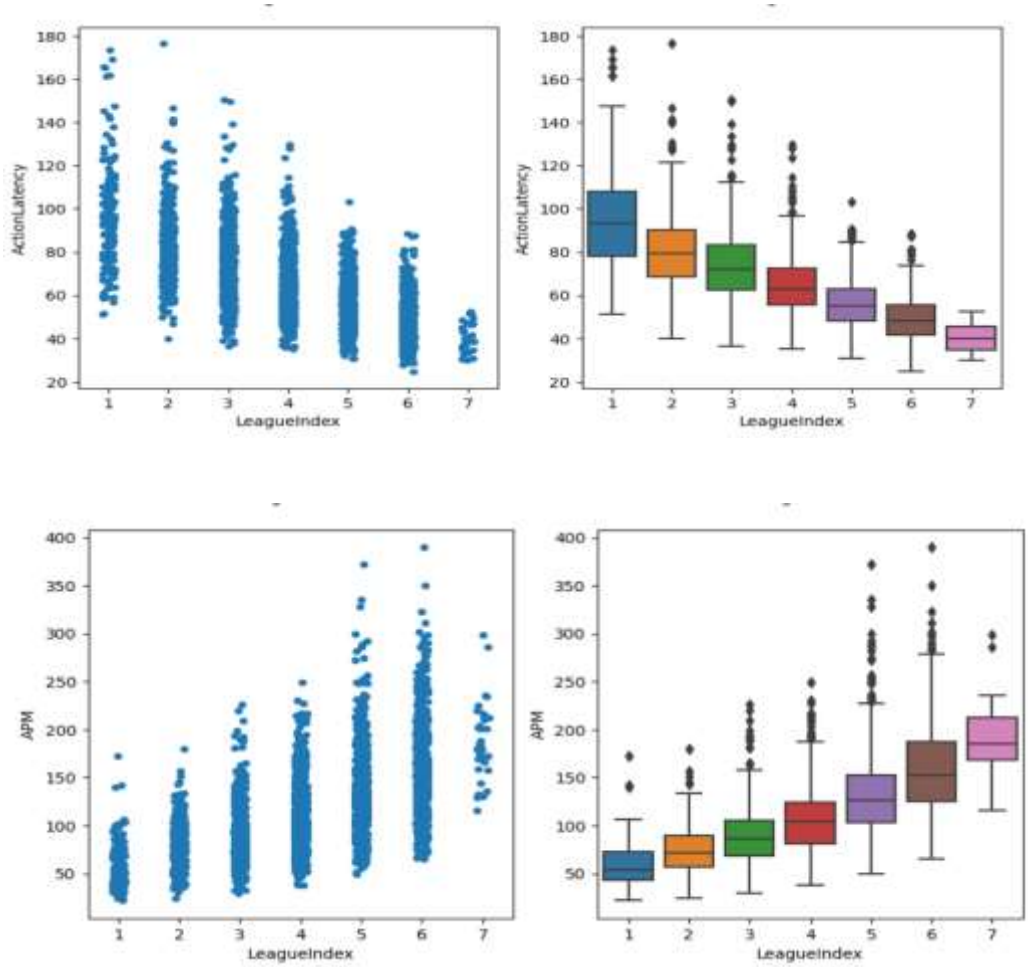
- **Key Takeaways:**

1. **Missing Value Analysis:** The Dataset had few columns (Age, HoursPerWeek, TotalHours) which had an object data type in spite of being quantitative. On further analysis, it was found that a few of the rows had the value "?". These values can be dropped or imputed. They have been imputed in this workflow using a Regression Model. The other features were used to predict the values for the three columns. This made more sense as compared to imputing it with the mean value or dropping the rows entirely as dropping all these rows with '?' would have meant that the all the rows for LeagueIndex = 8 would be removed from the data. Using a Regression Model imputation keeps the variation in other features in context while making predictions.

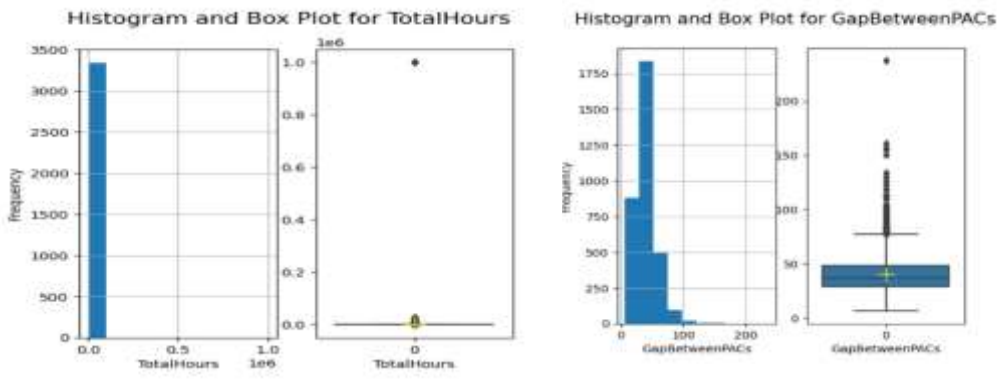
2. **Correlation:** Analyzing the correlation between variables is important as highly correlated variables can impact the performance of the model in the downstream workflow.



3. **Bivariate Analysis:** Seeing the relationship of all the independent variables with the dependent variable i.e., League Index. This was aimed at creating plots to gain to visualize the possible correlations found in the previous step. This helps in getting an idea about the features that would be helpful in the modelling process.



4. **Outlier Detection:** The Data had many columns which had very high values for some rows. Hence Outlier Detection was done using Interquartile Range to remove the outliers from the data.



5. **Scaling the Data:** The Data was scaled using a Standard Scaler to ensure it can be used for different model. Algorithms like XgBoost don't require data to be scaled however, there might be distance-based algorithms like KNN where performance might be impacted due to the scale of data. Hence, it is usually recommended to scale the data.

6. **Modelling:**

Regression

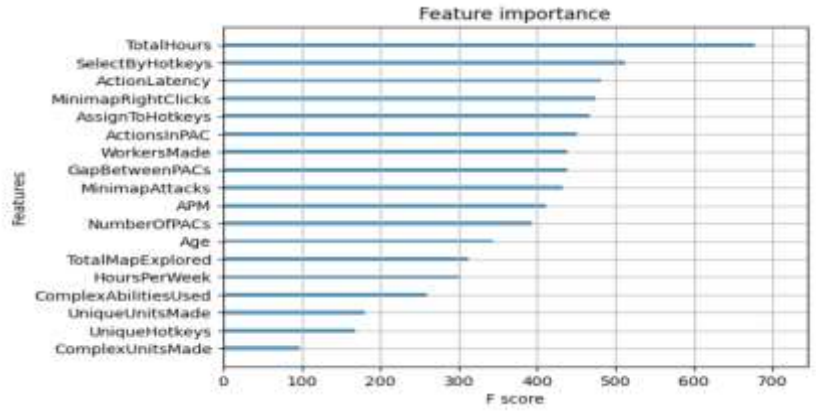
In the first experiment, this problem is considered as a regression problem. This makes more sense intuitively as we are trying to predict the rank of a player independently. Even if the rank needs to be assigned for a 8 player multiplayer session, the output of the regression model can be sorted to assign ranks and the updated ranks can be used as ground truth for further training.

- **Algorithms:** Ridge Regression, Random Forest, XGBoost Regressor were tried out. Ridge Regression serves as a baseline model against which we can evaluate the performance of other algorithms.
- **Hyperparameter Tuning and Performance Metric** – Mean Squared error was used as the scoring metric to tune the hyperparameters. R Squared has been used to evaluate the performance on test data as it helps in evaluating how much variance in dependent variable is being explained by our model. The best model which was XGBoost Regressor has an R Squared of 0.64. This is decent as this is just the first iteration of the model building process.
- **Feature Engineering** can be further explored to improve the performance of the model further. Highly correlated features can be removed which would help eliminate features that are not contributing in the model. Adjusted R Squared can be used as the performance metric as it penalized on the addition of extra variables which R Squared does not.
- **Classification –** The problem has also been explored as a classification problem where Logistic Regression, Random Forest and XGBoost have been tried.

7. **Model Interpretation:** Visualized Feature Importance and Shapley values to interpret the model.

FEATURE IMPORTANCE

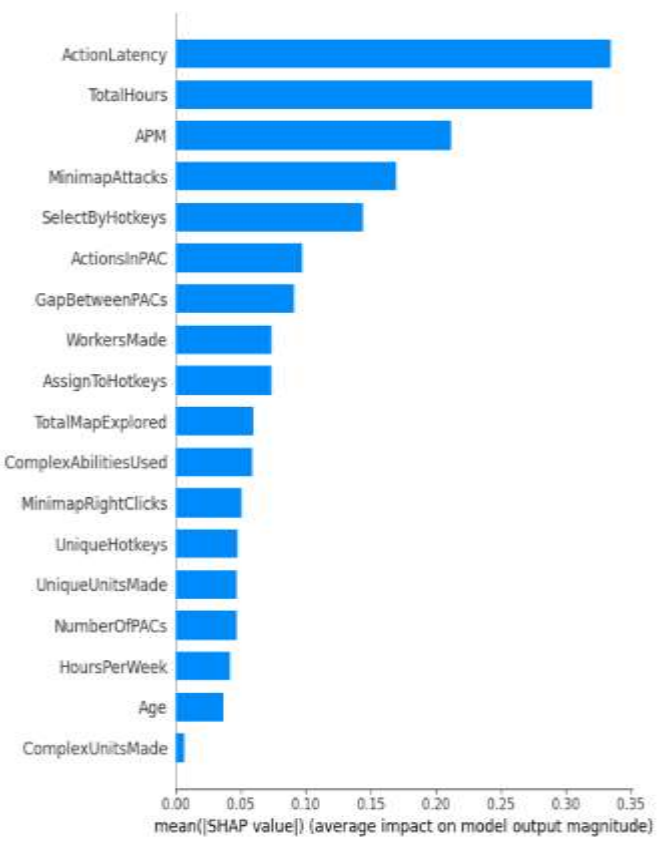
- This provides an insight on which variables played a major role in the construction of trees in XGBoost while training the model

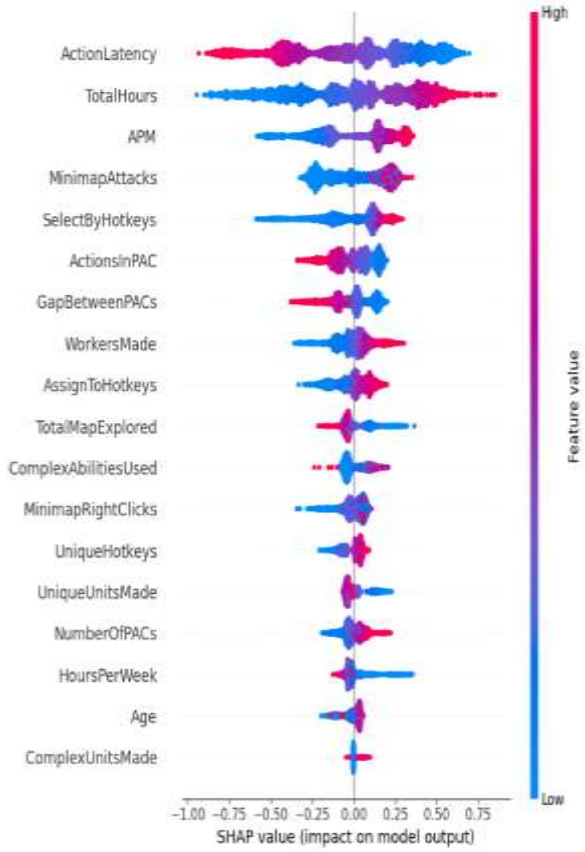


SHAPLEY

Shapley values help understand the features which have the largest contribution to the model prediction. This can be done on a global level as well as on a local level where we can drill down on a particular row to see which variables helped in making the prediction

- **Global Level:** The SHAP value plot can show the positive and negative relationships of the predictors with LeagueIndex. The variables are ranked in descending order of importance. It can be interpreted from the below plots that high value of ActionLatency has a high and negative impact on the ranking. On the contrary, high value of TotalHours has a high and positive impact on the ranking. Further importance of other features can be observed from the plot.





- **Local Level:** This can help in deep diving into individual predictions to look at every test prediction to analyze which features have contributed to the output. This will also help in root cause analysis for predictions which are way off from the expected output and identifying features that are contributing to the wrong prediction. In the below plot, it can be seen that for this row, the baseline prediction when the model did not use any of the independent features would have been 4.073 which is the mean prediction. However, the model predicted a score of 5.29 which can be rounded off to a ranking of 5. Action Latency of 0.40 played a major role in pushing the rank to 5.

Hypothetical: after seeing your work, your stakeholders come to you and say that they can collect more data, but want your guidance before starting. How would you advise them based on your EDA and model results?

1. My recommendation would be getting more data on features related to the top contributing features to the model such as ActionLatency, Hotkeys and APM. This might involve getting more aggregated level data for these features which would give further help in identifying patterns. It might also be helpful to get more session level details, that is how player are performing in the 8-player multiplayer game.
2. We can also delve into the domain of the game and accumulate more categorical data such as race number out of the 3 races, number of missions completed which can give more context about the ranking of the players.