Starcraft Player Game Analysis

Evil Geniuses x Genius League: Data Scientist Internship

Author: Uditi Namdev, MS CS, UC San Diego

unamdev@ucsd.edu

Table of Contents

01

Problem Overview

Understanding the problem and data

Problem Description

- In this project, we have a dataset that contains information about how Starcraft players perform in ranked games.
- Our goal is to create a model that can predict a player's rank based on the data we have.
- By analyzing patterns and trends in the dataset, we aim to develop a tool that can estimate a player's skill level and rank them accordingly.

Data Description

Some of the important information which was in the dataset is mentioned below:

Feature Name	Explanation
Age	Denotes the Age of the player
TotalHours	The total hours spent playing the game
UniqueHotkeys	Number of unique hot keys* used per timestamp
APM	Action per minute

*Hot keys: A hotkey is a key or set of keys which perform a specific function with regards to time efficiency.

02

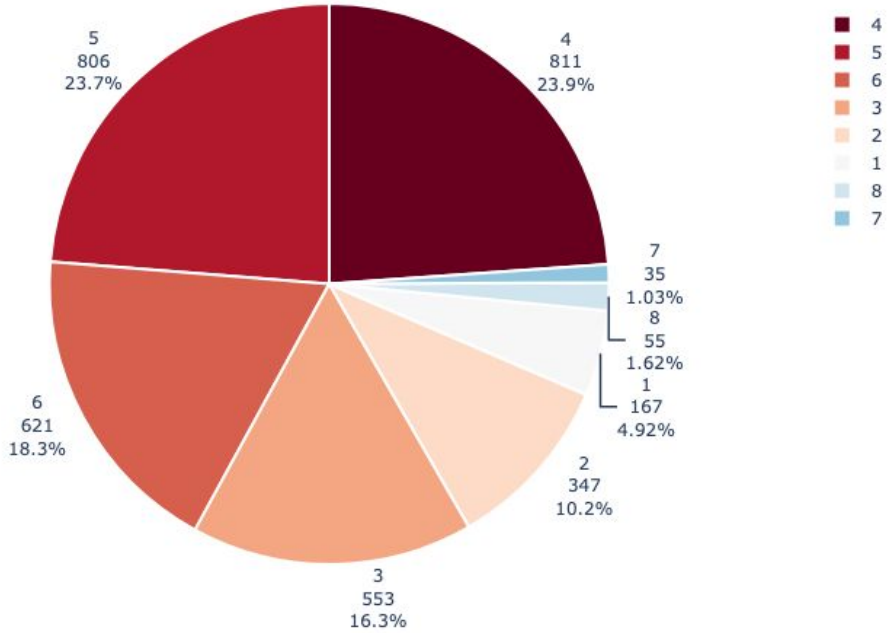
Exploring the data

Analyzing the distributions of the data

Visualizing the distribution of the data

This pie chart shows the distribution of the data according to different ‘LeagueIndex’.
In other words, it tells us how many examples of each ‘LeagueIndex’ is present in the dataset.

It can be observed that LeagueIndex = 4 has the highest number of examples and LeagueIndex = 7 has the lowest number of examples.

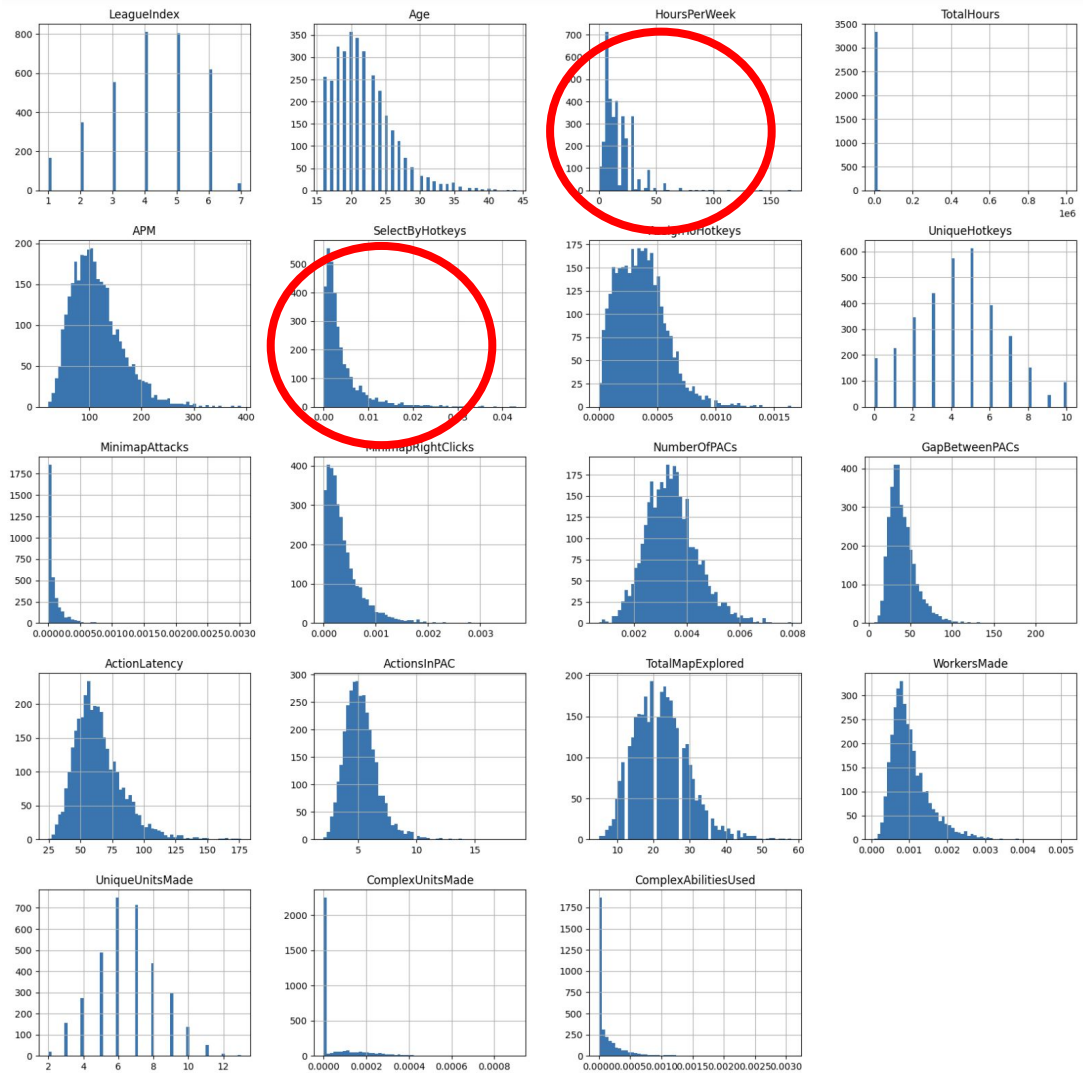


Histogram Plots

Plotting these histograms helps us see how our data is distributed with respect to every single feature.

How does this help?

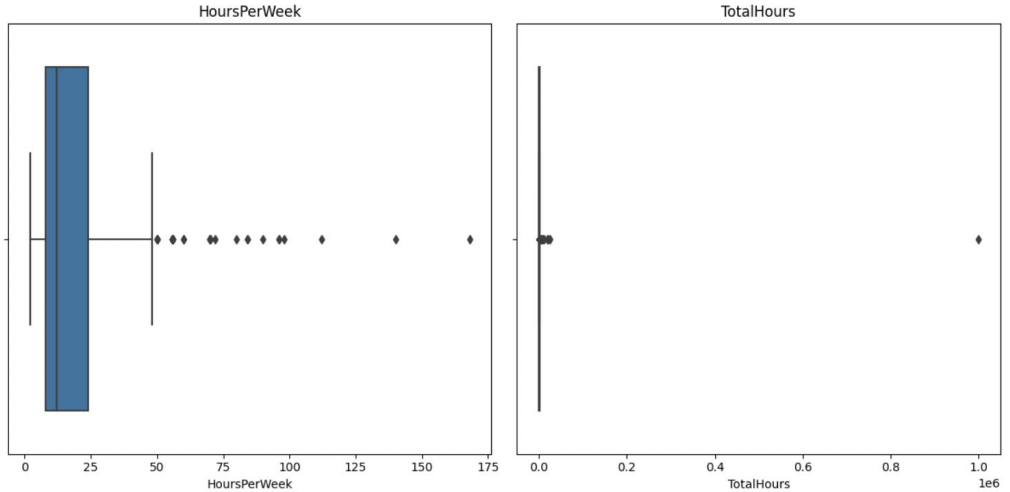
From the red circles, we can see that there are some high peaks on one side of the plot. This denotes extreme values on the right. These have to be fixed to get accurate results.



Boxplot Plots Pt. 1

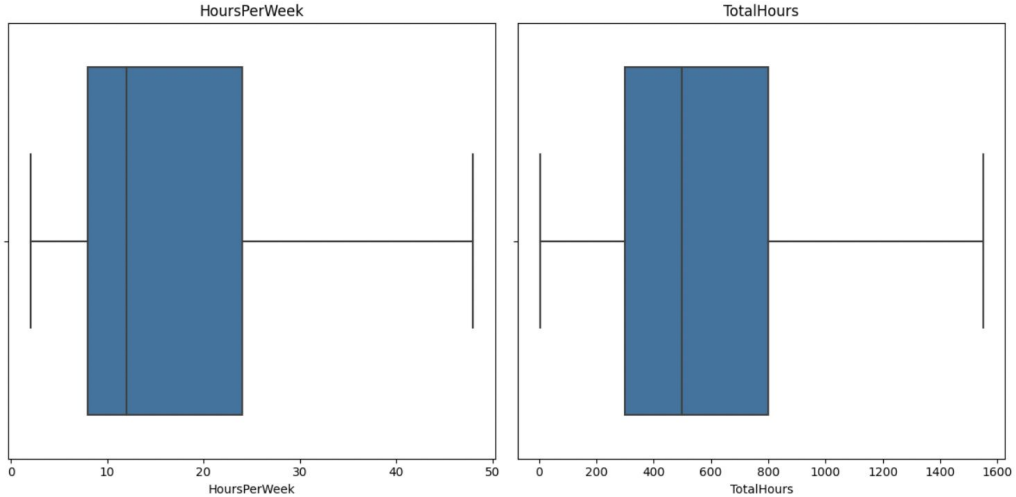
These boxplots are concentrated towards one side and not uniform. This indicated that there are extreme values.

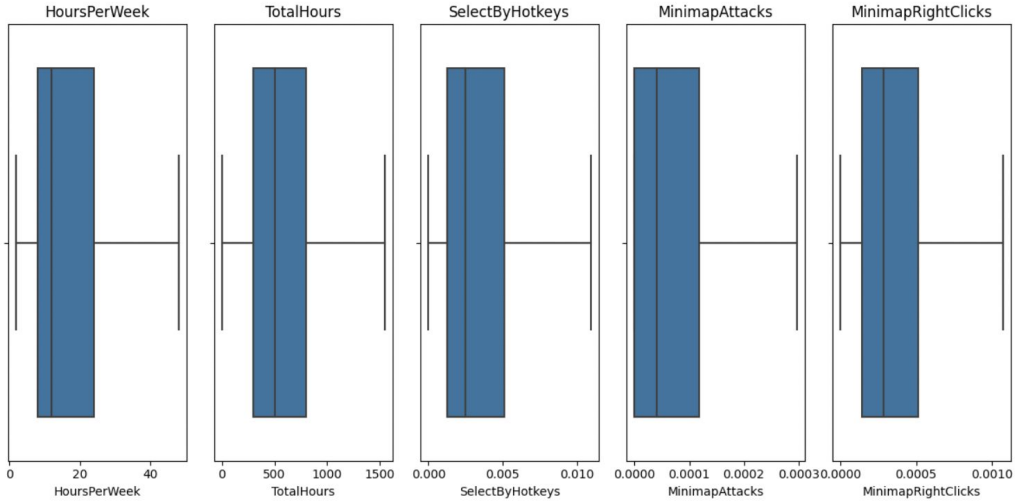
When we have such extreme values, it is often difficult for an algorithm to predict something accurately.



Boxplot Plots Pt. 2

So we replaced all such extreme values and got these plots which made the predictions more accurate.

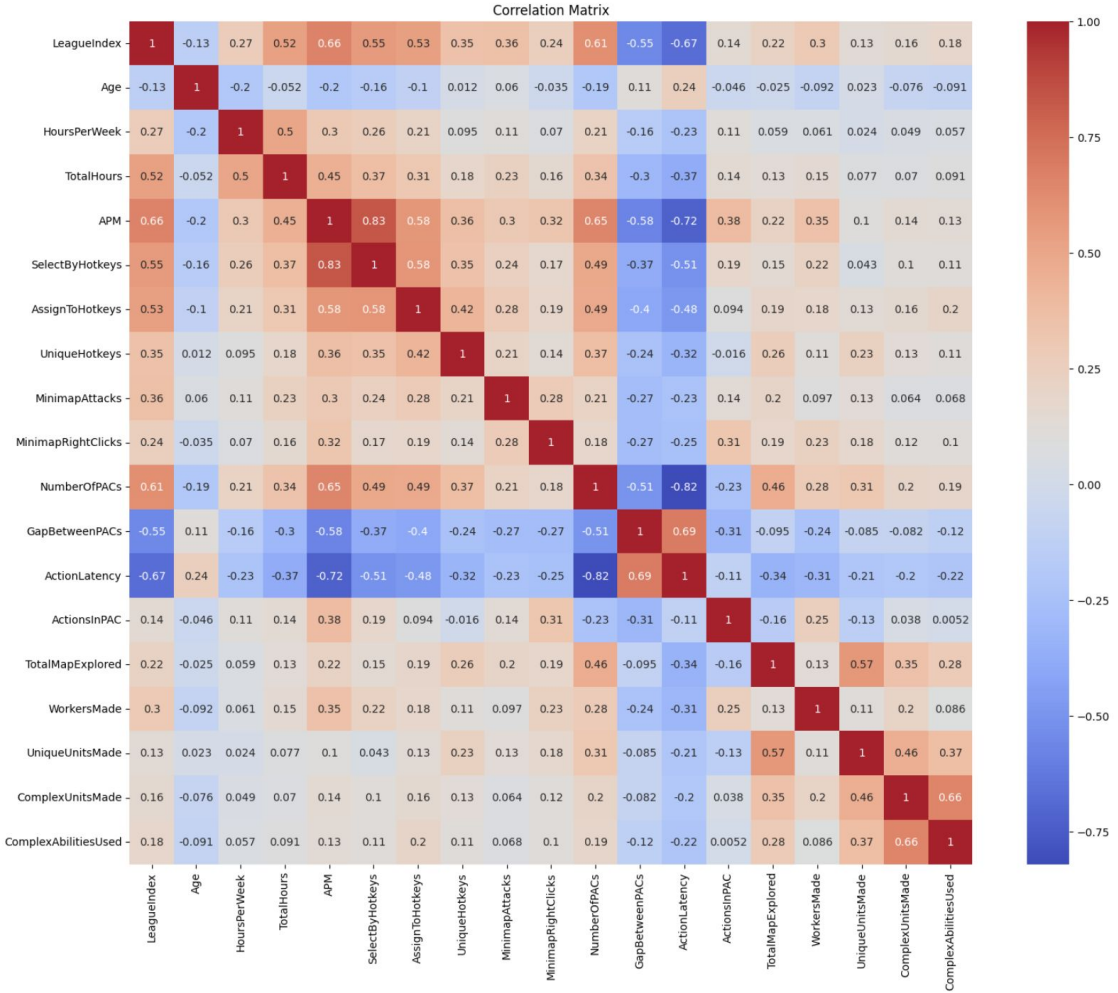




Correlation Matrix

Correlation Matrix helps us understand how one particular feature is correlated with another feature. Seeing the correlation of all the features with ‘LeagueIndex’ helps us understand which features play a prominent role in predicting ‘LeagueIndex’.

Darker shades of blue and red denotes higher correlation.



Important Correlated Features

These were the important correlated features that were present in the dataset. We included only these features in the next steps of our analysis and removed all the uncorrelated features.

```
['LeagueIndex',
 'TotalHours',
 'APM',
 'SelectByHotkeys',
 'AssignToHotkeys',
 'UniqueHotkeys',
 'MinimapAttacks',
 'NumberOfPACs',
 'GapBetweenPACs',
 'ActionLatency']
```

03

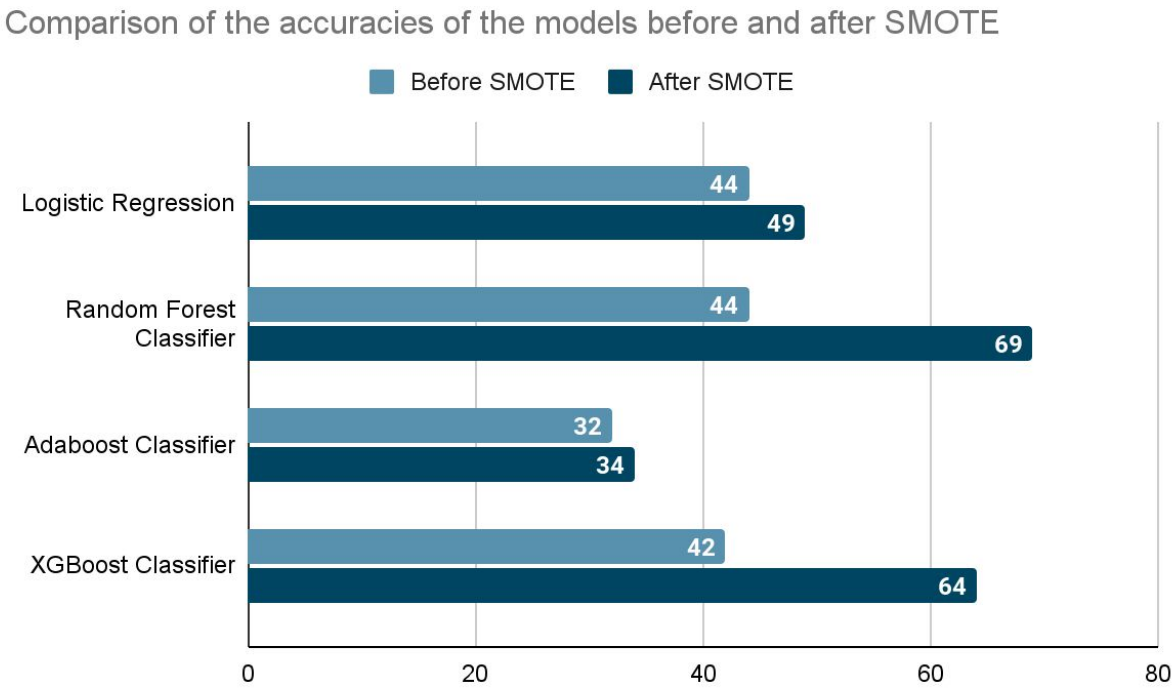
Uncovering Insights

Using data science methods to find interesting insights

Power of having more data

The bar chart on the right denotes how accurately different models could predict the ‘LeagueIndex’ after applying a technique called SMOTE.

So, the question is what is SMOTE and what is the significance of SMOTE?



Synthetic Minority Over-sampling TEchnique (SMOTE)

- SMOTE, which stands for Synthetic Minority Over-sampling Technique, is a method used to address imbalanced datasets.
- It looks at information about some things that happen less often and makes up new examples that are similar to them.
- By using SMOTE, we can enhance the accuracy and reliability of our models when dealing with imbalanced data, leading to more robust and fair predictions.

04

Future steps

Reviewing the results and formulating the next steps

Conclusion and Future Steps

- We observe that by artificially generating the data, we could improve the accuracy of the model by a large margin.
- It is possible that the artificially generated data is sometimes not a true representation of the real-life data. However, we cannot overlook the power of having a balanced dataset.
- This shows that it is essential to gather more data of those players whose LeagueIndex currently does not contain sufficient data.
- It is possible to develop robust models that can make more accurate predictions when we have sufficient data.

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**