MEEN 423

PROJECT 2

REPORT

PREDICTING A FORMULA 1 RACE

By: Akhil Jayadeep, 533002964
Section 500

1. Introduction:

Formula One (F1) is the highest class of international racing for open-wheel single-seater formula racing cars sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA Formula One World Championship has been one of the premier forms of racing around the world since its inaugural season in 1950. This sport has seen long periods of absolute domination especially in the recent years with Mercedes dominating most of the hybrid engine era which started in 2014. This domination period proved to be one of the longest in the sport's history and it has produced the of the most successful racers of our time, Lewis Hamilton. 2021, marked the end of this domination, with Red Bull Racing's Max Verstappen clinching the Championship. However, the dominant periods of the sport have caused critics of the sport to call it predictable. This project will explore whether racing results can be predicted on the basis of historical race data available online.

2. Problem Definition:

This project will attempt to use historical data to predict the results of Formula 1 races. The data will be used to train a machine learning model that can predict whether a particular driver will finish in the top ten during a race. The model will be trained on data from previous races, such as race results, driver age, cumulative wins, constructor wins and weather conditions. The model will then be tested on a held-out set of data to see how well it can predict the results of new races.

The goal of this project is to develop a model that can accurately predict the results of Formula 1 races. This information could be used by teams and drivers to make strategic decisions about how to race. Furthermore, because of the nature of the sport, knowing whether a driver will be in the top ten could also be used by fans to make more informed bets that have higher winnings.

3. Description of Dataset:

The dataset was obtained from Kaggle (linked here). The dataset comprises comprehensive information on Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, and championships spanning from the inaugural 1950 season to the latest 2023 season. The dataset was created using the Ergast Developer API (linked here). Upon initial analysis, it was observed that the dataset was fairly clean and had relatively few missing values. The separate datasets for race results, circuit data, driver standings, constructor standings, qualifying times, driver information, race finish status, and racing season data were all merged to create one main dataset.

Track weather is a very important factor that affects race outcomes, but this data was not readily available anywhere. Python web scraping tools, Selenium and Beautiful Soup were used to scrape race Wikipedia pages and extract the weather information from the first Info box table that is usually

seen on the right-hand side of Wiki pages. The weather information was then sorted into warm, cold, dry, wet, and cloudy, based on a custom keyword dictionary. When the weather information was not available, warm and dry weather was assumed.

Best qualifying times were calculated for each driver for all the races. However, the qualifying information was not complete, and upon further analysis the dataset had far too many empty values, for it to be a viable feature. As such best qualifying time was dropped as a feature. This was done with the understanding that best qualifying time was directly correlated to the starting grid position, and thus was not a necessary feature.

Additional features were added to the dataset based on existing columns. Nationality of the driver and the track location were compared to determine whether the driver was participating in a race in his home country. This feature was added, due to the widely known knowledge that drivers tend to try harder to win races in their home country.

Driver age at the time of race was calculated along with driver points before the start of the race. The finish status of the of each driver for every race was used to determine their finishing position as well. Finally, the dataset was trimmed down to only include race information starting from the year 2001 to 2021. This was done because prior to this the regulations of the sport were vastly different, which would not help with accurate prediction.

4. Supervised Learning Strategy:

The variable chosen to be predicted was a binary variable called in_top_ten. As such classification methods were chosen to be used to create the models.

The supervised learning methods used in this analysis were Random Forest classifiers, Logistic Regression, Support Vector Machines (SVM), and Neural Network classifiers. Multiple methods were used to increase the chances of finding a good fit. These methods were chosen in particular because of their abilities to handle both linear and non-linear relationships between the features and the target, as well as high-dimensional data. Data from races ranging from 2001 to 2020 were used for training and the 2021 championship was used as the test set for all the models.

The features used for the models were 'year', 'round', 'weather_warm', 'weather_cold', 'weather_dry', 'weather_wet', 'weather_cloudy', 'cum_wins', 'grid', 'driver_age', 'home_race' and 'points_before_race'.

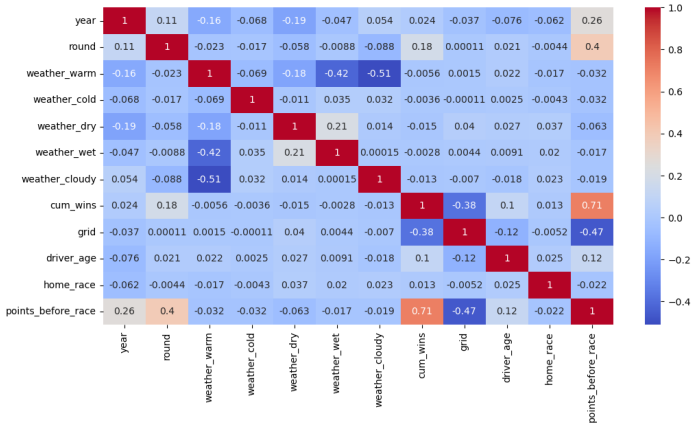

Fig. Feature Correlation Heatmap

The feature correlation matrix showed that none of them were highly correlated to each other, which meant that principal component analysis could not be performed for this particular dataset.

To find the best hyperparameter for each of the models, GridSearchCV was used. This helped prevent overfitting to the data. Once the best hyperparameters are found, a new model is trained using them, and predictions are made on the test set. The predictions are rounded to 0 or 1 based on a threshold of 0.5, and a new column is added to the test data frame to store the predicted values. Finally, the model is evaluated using the accuracy score, confusion matrix, and classification report.

5. Results:

Model accuracy here is defined as the ability to accurately predict whether a driver will score points and finish in the top ten of the race.

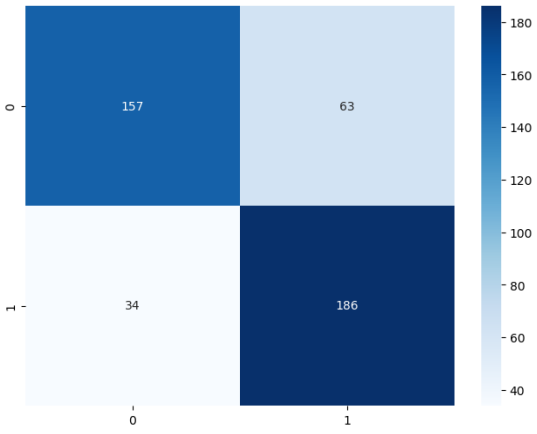

Fig. Confusion Matrix (Random Forest Classifier), Model Accuracy: 0.75

Best Hyperparameters: {'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}

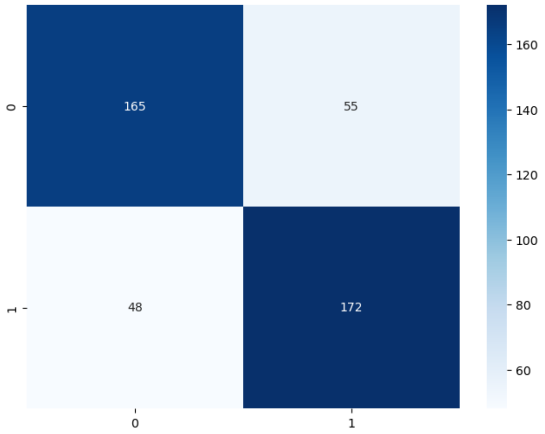

Fig. Confusion Matrix (Logistic Regression),

Model Accuracy: 0.76

Best Hyperparameters: {'C': 0.001, 'penalty': 'none'}

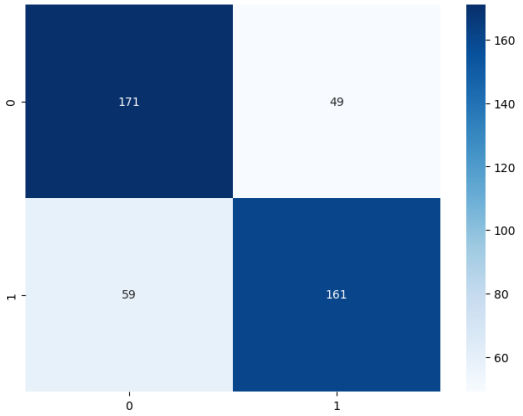

Fig. Confusion Matrix (SVM Classifier),

Model Accuracy: 0.77

Best parameters: {'C': 10, 'degree': 2, 'kernel': 'linear'}

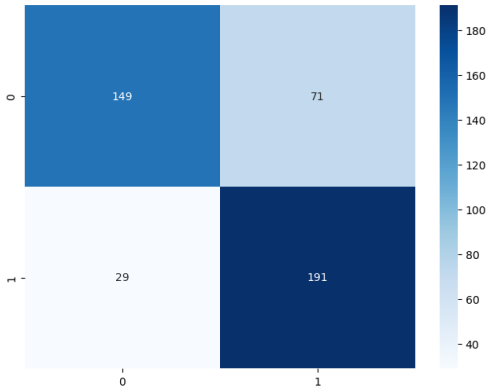

Fig. Confusion Matrix (Neural Network Classifier), Model Accuracy : 0.73

Best parameters: {'activation': 'linear', 'dropout_rate': 0.2, 'neurons': 64}

Overall, the SVM Classifier model produced predictions with the highest accuracy of 77%.

6. Conclusions

In conclusion, this project helped create a fairly accurate model when it comes to predicting the drivers finishing in the top ten and scoring points. This SVM model produces an acceptable level of accuracy for prediction because, F1 races have a certain degree of unpredictability as well when it comes to random factors such as spontaneous driver error and opening lap collisions which affect the race to produce outlier results for 4 or 5 races in a season. However, it is important to note that this model will not be as accurate for race seasons past 2021 due to the major regulation changes that were introduced in 2022 to change the design of the cars to make us of the ground-effect that enables easier

overtaking. This model can be improved in the future by exploring more comprehensive datasets that include vehicle telemetry as well as measurable changes that have been caused by regulation overhauls.

7. References:

i. Ergast API. (n.d.). Retrieved May 6, 2023, from https://ergast.com/mrd/

ii. Wikipedia contributors. (2023, May 4). Formula One. In Wikipedia, The Free Encyclopedia. Retrieved 03:37, May 7, 2023, from https://en.wikipedia.org/w/index.php?title=Formula_One&oldid=1153118871