Predict the winner of the F1 Grand Prix using Machine Learning

Jaspreet Singh ID:19150299

A4 CMP/DIG620 BCU 2021-2022

Introduction

Predicting winners in sports has always been of interest to many. One such sport is Formula One.

While it seems quite an easy motorsport to predict as the winner will be the driver in the fastest car, this is not necessarily the case as lots of other factors contribute to the final result and all these possibilities could lead to an unpredictable outcome but with the right techniques it’s possible to have an answer.

Problem definition

Every major sporting event can be bet on and used to win money, but is there a way to correctly predict the winner of a Formula One Grand Prix in order to gain an advantage in general betting that is more efficient and trustworthy than gut feelings and emotions?

Aims and Objectives

- Find publicly available historic F1 seasons datasets.
- identification of existing machine learning models for predicting race winners.
- Find the best set of variables inside the dataset.
- Utilize the best possible machine learning models.
- Evaluate the models to find the most accurate one.

Research Methodology

Regarding the different subjects that were studied during this review and subsequently, in order to complete the project, these different terms were searched and studied:

- Data analysis
- Data Pre–Processing (Data wrangling, feature engineering, and feature selection)
- Machine learning
- F1
- Betting Odds
- Performance Evaluation (Confusion Matrix and Classification Report)

Methodology

A waterfall methodology was used throughout the project, while test-driven development was executed while coding each part of the algorithm.

For the development of the artifact python language has been used to code, as well as some other non python libraries.

- Pandas
- Mathplotlib and seaborn
- NumPy
- Scikitlearn

Implementation (Data)

- Gather data from the Kaggle dataset ‘Formula 1 World Championship (1950-2022) (Vopani, Kaggle.com, 2022) as CSV files.
- Data wrangling : merge files, drop unused columns and rename useful ones.
- Feature engineering : create new usful columns (ex. age, overtakes)
- Data cleaning: fill null values (ex. fast_lap_rank,qualify_position)
- Label encoding: change categorical values to numeric. (ex. driver_nationality)
- Check skewness and drop columns too skewed

Implementation (ML)

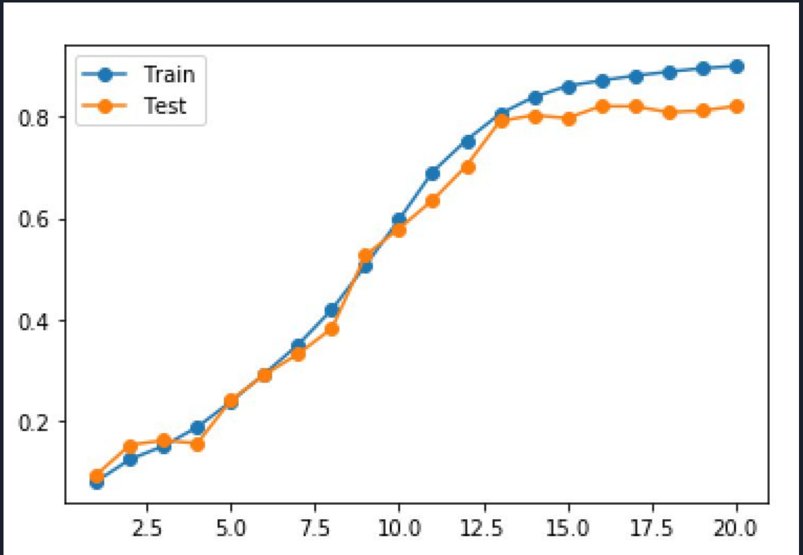
The problem can be seen as a classification or regression problem, but in the study it was only seen as a classification problem using Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. These next steps will be executed multiple times as a test-driven development.

- Split train (all season previus the test season) and test (chosen test season).
- Fit scalers (standard and minmax).
- Hyperparameter optimization with GridCV (for LR) and loops (for DTC).

Implementation (ML)

GridCV for LR gave the best possible parameters for the model after a long processing time.

Loops for DTC were relatively faster on used machine and a subsequent overfitting test was done with the plot of the precision of test and train model's scores with the tuning.



Evaluation

The two possible problems that have been analysed during the development:

- Multiclass classification: for the positions of each driver on the grid .
- Binary classification: predict the race winner with purely 1s and 0s.

The evaluation metrics used for both of these classifications were:

- Confusion matrix
- Classification report
- Precision score (Correct winner predictions for each race of a season)

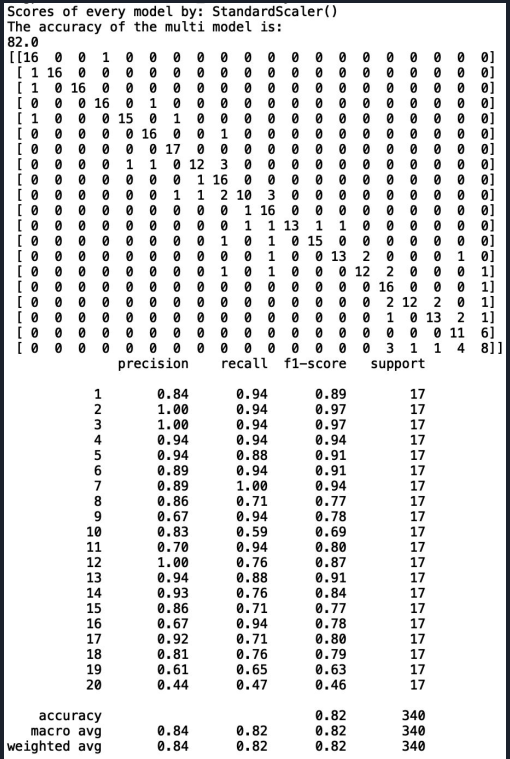
Last evaluation has been done by confrontation between model’s predictions and betting odds.

Evaluation (Multiclass)

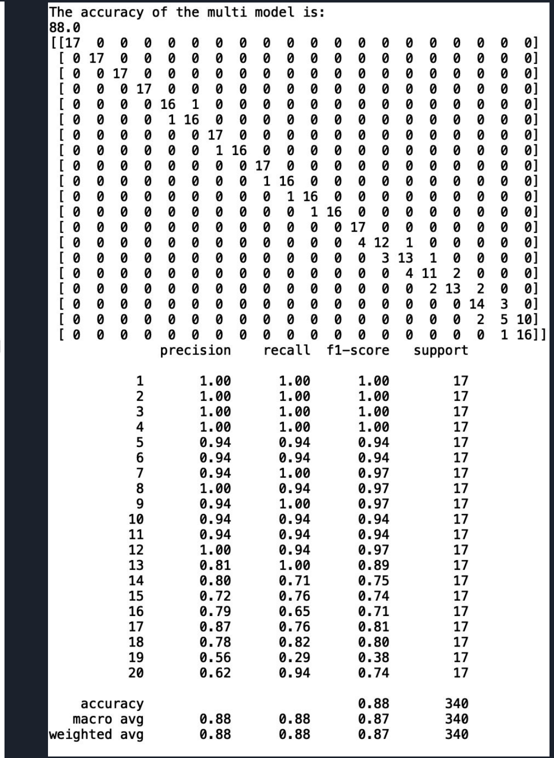
Only LR and DTC will be fully considered as final models as both reach a decent score for the multiclass classification:

- DTC : 82
- RFC : 56
- LR : 88

Models are still not smart enough to predict every position on the grid.



```
Scores of every model by: StandardScaler()
The accuracy of the multi model is:
82.0
[[16  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 1 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 16  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0 15  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 16  0  0  1  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  1  1  0 12  3  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  1  1  2 10  3  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1  1 13  1  1  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  0  1  0 15  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  0 13  2  0  0  0  1  0]
 [ 0  0  0  0  0  0  0  0  1  0  1  0  0  0 12  2  0  0  0  1]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 16  0  0  0  1]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2 12  2  0  1]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0 13  2  1]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 11  6]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3  1  1  4  8]]
              precision    recall  f1-score   support

           1       0.84      0.94      0.89        17
           2       1.00      0.94      0.97        17
           3       1.00      0.94      0.97        17
           4       0.94      0.94      0.94        17
           5       0.94      0.88      0.91        17
           6       0.89      0.94      0.91        17
           7       0.89      1.00      0.94        17
           8       0.86      0.71      0.77        17
           9       0.67      0.94      0.78        17
          10       0.83      0.59      0.69        17
          11       0.70      0.94      0.80        17
          12       1.00      0.76      0.87        17
          13       0.94      0.88      0.91        17
          14       0.93      0.76      0.84        17
          15       0.86      0.71      0.77        17
          16       0.67      0.94      0.78        17
          17       0.92      0.71      0.80        17
          18       0.81      0.76      0.79        17
          19       0.61      0.65      0.63        17
          20       0.44      0.47      0.46        17

    accuracy                           0.82       340
   macro avg       0.84      0.82      0.82       340
weighted avg       0.84      0.82      0.82       340
```



```
The accuracy of the multi model is:
88.0
[[17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 16  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  1 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1 16  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 17  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  4 12  1  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  3 13  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  4 11  2  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2 13  2  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 14  3  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  5 10]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1 16]]
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        17
           2       1.00      1.00      1.00        17
           3       1.00      1.00      1.00        17
           4       1.00      1.00      1.00        17
           5       0.94      0.94      0.94        17
           6       0.94      0.94      0.94        17
           7       0.94      1.00      0.97        17
           8       1.00      0.94      0.97        17
           9       0.94      1.00      0.97        17
          10       0.94      0.94      0.94        17
          11       0.94      0.94      0.94        17
          12       1.00      0.94      0.97        17
          13       0.81      1.00      0.89        17
          14       0.80      0.71      0.75        17
          15       0.72      0.76      0.74        17
          16       0.79      0.65      0.71        17
          17       0.87      0.76      0.81        17
          18       0.78      0.82      0.80        17
          19       0.56      0.29      0.38        17
          20       0.62      0.94      0.74        17

    accuracy                           0.88       340
   macro avg       0.88      0.88      0.87       340
weighted avg       0.88      0.88      0.87       340
```

Evaluation (Binary)

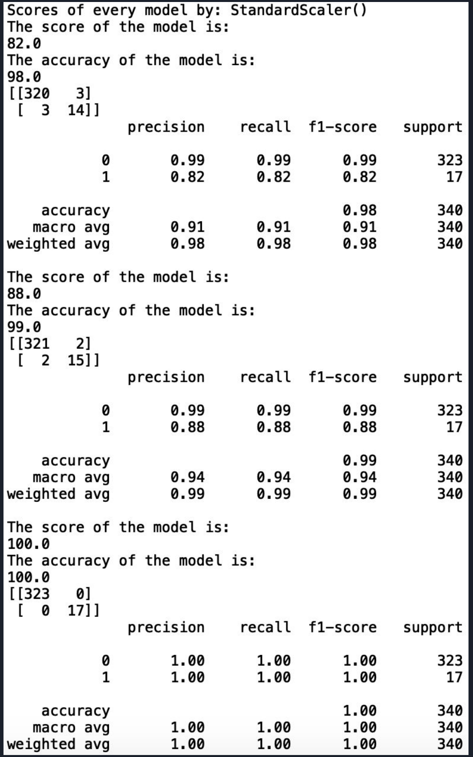
Only the standard scaler will be evaluated as its precision and score are high enough to be considered a successful attempt at prediction.

All models produce incredible results for the binary classification:

DTC : Score = 82 / Accuracy = 98

RFC : Score = 88 / Accuracy = 99

LR : Score = 100 / Accuracy = 100

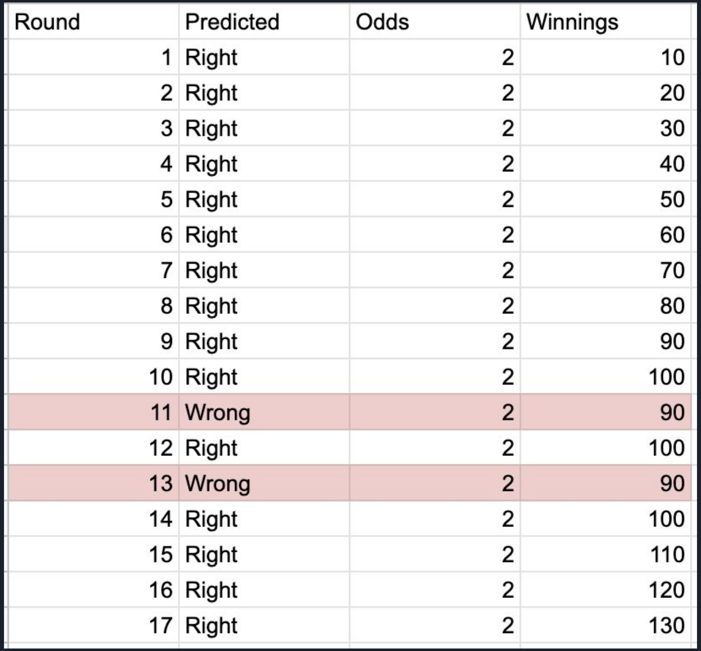


```
Scores of every model by: StandardScaler()
The score of the model is:
82.0
The accuracy of the model is:
98.0
[[320   3]
 [  3  14]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       323
           1       0.82      0.82      0.82        17

    accuracy                           0.98       340
   macro avg       0.91      0.91      0.91       340
weighted avg       0.98      0.98      0.98       340

The score of the model is:
88.0
The accuracy of the model is:
99.0
[[321   2]
 [  2  15]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       323
           1       0.88      0.88      0.88        17

    accuracy                           0.99       340
   macro avg       0.94      0.94      0.94       340
weighted avg       0.99      0.99      0.99       340

The score of the model is:
100.0
The accuracy of the model is:
100.0
[[323   0]
 [  0  17]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       323
           1       1.00      1.00      1.00        17

    accuracy                           1.00       340
   macro avg       1.00      1.00      1.00       340
weighted avg       1.00      1.00      1.00       340
```

Evaluation (Odds)

When confronting predictions with betting odds, the models' performances are similar and both allow an opportunity to achieve a profit.

Model confronted with same odds for all the season granted a profit of 130£ when betting 10£ for each race. In this case, the majority of the bets were won.



Round	Predicted	Odds	Winnings
1	Right	2	10
2	Right	2	20
3	Right	2	30
4	Right	2	40
5	Right	2	50
6	Right	2	60
7	Right	2	70
8	Right	2	80
9	Right	2	90
10	Right	2	100
11	Wrong	2	90
12	Right	2	100
13	Wrong	2	90
14	Right	2	100
15	Right	2	110
16	Right	2	120
17	Right	2	130

Conclusions

This study presents an empirical analysis of the predition of the F1 GP winner. To address the challenge,

- Historical racing data was gathered and used to create a big dataset, and various data processing and machine learning steps were carried out.
- LR, DTC, and RFC with standard scaler, minmax scaler and fine-tuned parameters were analysed.

LR and DTC models' performances were enhanced confortably to 80%–100% depending on the season and randomness. These results were similar to the study conducted by Anandaram Ganapathi using the same dataset.

Future Work

Different new experiments using various models can be operated in the future, or modifying existing models and comparing them can also provide meaningful data to work on.

- Training models for longer times, if not creating a new models altogether.
- Increase in the amount of data available and relevant could help to improve the model’s accuracy and relevance to the new regulations and changes.
- Increase the utility of the project by creating a webapp to permit public access to the model.

Thank you for your attention.

Jaspreet Singh   ID:19150299

A4 CMP/DIG620   BCU 2021-2022