Final Project Report

Predicting Formula 1 World Champions Using Logistic Regression

Introduction

I am looking to see if a logistic regression model can predict which Formula 1 drivers become World Champions using career statistics and demographic attributes. Formula 1 has a history, with hundreds of drivers who raced at different levels. The dataset, for this project holds 868 drivers representing the group of competitors. Only 34 of the 868 drivers have ever won a World Championship title. Champions are a slice of the whole population so the prediction task is hard. The test results for 174 drivers come from the 80/20 train-test split. The 80/20 train-test split uses 20 percent of the dataset, which's one fifth of all drivers as the test set. The test set of 174 drivers reflects the population. The test set of 174 drivers keeps the model evaluation fair on data.

Dataset Description

The dataset contains variables, for each driver. The dataset includes career starts, race entries, points scored podium finishes, wins, team affiliations, seasons raced and more. When I examine the dataset I see observations before any modeling. Out of the 868 drivers only 346 have ever scored a Formula 1 point. That means 60 percent of the drivers finish their careers without scoring any points. I see that modern scoring systems give points to, about half of the grid in each race. I see that modern scoring systems show that point-scoring opportunities were very different, across eras.

A single race win, under the scoring system gives a driver career points than all but 140 drivers in the whole dataset. The statistic shows that the points distribution has been unbalanced historically. The statistic also shows that career length differs a lot among drivers. The dataset has 489 drivers with than ten race entries. The dataset has than five hundred drivers with, than ten race starts. I notice that the majority of Formula 1 drivers do not have career racing to build performance stats that're as good, as the long-term competitors. Champions usually race for seasons. Champions collect many points many podiums and many wins. These facts strongly affect the modeling and the results.

Exploratory Data Analysis

I notice that the performance-related metrics, like points, wins, podiums and starts are very right-skewed. The performance-related metrics have a group of drivers, with high values. Most drivers have zero values. The champions sit at the end of the performance-related metrics. The champions are a few compared with the rest of the data. Because of that the differences do not always show up clearly in visualizations or basic statistical comparisons.

Other patterns also appear. I see a link, between the number of starts and the total points scored. The link is expected because the drivers who race often have chances to earn points.. Even the drivers with starts only a few ever reach championship-level performance. Many drivers who stay in the sport for a time collect statistics and do not get close to championship-level performance. The result adds difficulty, to modeling. Predictive modeling becomes more complex.

Preprocessing

Before I built the model I cleaned the data. I dropped columns that were not needed or that duplicated information. I turned variables into one- encoded columns. I scaled the features with a scaler. The target variable, for the regression model is a flag that shows if a driver is a World Champion. After I finished preprocessing I split the dataset into training and testing sets. I used an 80/20 split. I stratified the split so that the share of champions stays the same in both sets. There are 868 drivers, in total. I used about 694 of the drivers for the training and about 174 of the drivers, for the testing. The numbers fit the size of the dataset. Make sure the final model is evaluated correctly.

Logistic Regression Model

I chose the regression model. The logistic regression model is simple and easy to understand. In my view the logistic regression model works well for classification. The logistic regression model tries to tell champions from non-champions using the features we have. The dataset is imbalanced—34 champions and 834 non-champions—so the logistic regression model does better on the majority class.. The logistic regression model still gives clues, about which drivers are more likely to be champions based on their career statistics.

Evaluation Metrics

I saw the classification report that the model gave for the test set of 174 drivers.

	Precision	Recall	F1-Score	Support
Class 0:	0.97	1.00	0.99	167
Class 1:	1.00	0.29	0.44	7

I looked at the results. Saw that the model got an accuracy of 0.97 while the macro-average F1 score was 0.71 and the weighted-average F1 score was 0.96. The model did well on the non-champion class. The model struggled on the champion class. Out of the seven champions, in the test set the model correctly identified two. The precision for champions was perfect. When the model predicted champion the model was always right. The recall, for champions was 0.29. The model missed most of the champions. The model failed.

Interpretation of Results

I notice that the model performance depends heavily on the dataset structure. In the data ninety six percent of all drivers are not champions. The model learns that saying a driver is not a champion always gives an answer. That is why the model shows an accuracy and a perfect recall, for the non-champion class. The small number of champions in the dataset makes it hard for the model to find patterns that separate champions, from the rest of the drivers. Also most drivers have a few Formula 1 races. More, than half of the drivers have than ten entries. The data on the drivers is very small. The model cannot detect patterns in the data, for drivers. The model finds the task almost impossible because the data is limited.

Championship outcomes, in Formula 1 depend on more than driver statistics. The car quality, the team resources, the era-specific competitiveness, the changes, the political and financial dynamics all play a role in championship success. Career longevity is heavily influenced by team decisions than driver performance, for drivers. The car quality, the team resources, the era-specific competitiveness, the changes the political and financial dynamics are not represented in the dataset. The missing car quality missing team resources missing era- competitiveness missing changes, missing political and financial dynamics limit the prediction ability of the model.

Findings and Conclusion

I found that the logistic regression does not work well for predicting World Champions when I only use driver statistics. The rarity of champions the spread of performance metrics and the short career lengths of drivers all make the prediction task hard. The model does on non-champions. The model still fails to pick out champions. The recall, for champions stays low.

I think that statistical career data alone is not enough to model or predict championship outcomes in Formula 1. I see that the sport's structure makes sure that only a small group of drivers ever gets the chance to race at the level needed for a championship. I suggest that future work could look at models. I suggest that future work could add team-level performance data. I suggest that future work could use resampling methods to fix class imbalance. The findings of this project match the world of Formula 1. I see that the findings of this project show the limits of modeling when the data are unbalanced and when the context changes a lot.