# Predict to Get Picked: Prediction Models for Formula 1 Constructors

Kent Grass-Valdivia (kfg29), Eric Sullivan (ebs226), Logan Abramowitz (la437), Maya Voloshin (mav99)
Department of Operations Research and Information Engineering, Cornell University

ORIE 3120: Practical Tools for Operations Research
Professor Peter Frazier
May 10th, 2025

# Table of Contents

# Abstract

In this report we seek to conduct an analysis of the Formula 1 World Championship dataset from Kaggle. This data set contains 14 tables and 120 fields. Our analysis was conducted on a subset of this dataset, starting from the year 2011 to 2024. This represents the era of Formula 1 where hybrid engines started to be used. This allows for the data to give better and more fair insights into the impact of certain metrics on driver performance. In terms of describing this dataset, there are many columns that are descriptive such as driver age, average lap time, number of points, number of pit stops, qualifying position, and many more. These descriptive columns are describing columns that provide for the identification of a constructor (Formula 1 team), a particular race, a specific circuit (race track), as well as the driver itself. In this report we investigate the significant features and their effect on race performance, particularly final position. After filtering our data to just include records for drivers that complete the race, we use statistical methods to create models that aid Formula 1 teams in developing their race strategies and selecting particular drivers.

*Key words: Formula 1, prediction, motorsports, sports analytics*

# Introduction

This dataset, because of its wide variety of fields, allowed us to execute data analysis methods to answer essential questions that would be of interest to Formula 1 teams (constructors).

Such questions include, *what is the final position of a driver, given certain descriptive variables of that driver including age and average lap time?* From this question we are able to pick up on the most important variables in predicting race performance.

An important variable that we analyze in our report is the timing of a pit stop, which is of great importance to Formula 1 constructors. Formula 1 is an interesting sport because team strategy is nearly as important as driver skill in the outcome. A large part of race strategy are pit stops, specifically the timing of one. Pit stops are important because they take time away from racing, but are essential to the performance of a car because of factors such as tire degradation. Drivers are required to take at least one pit stop, as two different types of tires are required to be used in a race (Motorsport Academy, 2021). There are two pit strategies that drivers use: Overcut and undercut. An overcut occurs when a driver will push on worn tires for a while and pitting closer to the end to switch their new tires and then speed ahead of opponents at the end. An undercut is when a driver will pit earlier in the race and attempt to gain leads on drivers with worn tires and get ahead during their later pit (Motorsport Academy, 2021).

Before a race, statistical analysis is usually performed to analyze things such as the weather and historical race data (Motorsport Academy, 2021) in order to prepare. In our analysis, we seek to provide models that could be used during these analyses conducted by Formula 1 teams. Most importantly, we hope to provide verification that such models provide accurate and relevant information.

Formula 1 teams each have two drivers per race. These constructors are interested in the outcomes of opposing constructors as this allows them to alter their strategies to place the best they can (e.g. when to come into the pitlane to replace tires or when to push the car). Placement is determined by how many points, which are determined by final placement of the two drivers and an audience survey for “Driver of the Day.” Constructor placement comes with great financial rewards (Mayne & Cox, 2025).

Essentially, our goal is to provide insight into what key features affect race outcomes for Formula 1 Teams. This will allow them to plan their race strategy and select particular drivers in order to gain the most number of points, and at the end of the day, to win.

# Data Analysis

Our team will utilize three statistical methods to conduct our data analysis. These three methods include linear regression, A/B testing, and random forest. We first conduct linear regression to narrow down the statistically significant features in the data set.

After figuring out what those important features are in predicting final position, we conduct an A/B test to analyze the causal effect of the important features on driver race placement. An important feature that was discovered was pit stop time. Using random forest as our third statistical method, we created a model for the goal of predicting when a driver will pit.

## Linear Regression Predicting Final Placement

The goal of this linear regression analysis is to predict where drivers will finish in a given race. Predicting where drivers finish the race is important as it allows the constructor to strategize who their driver should compete with. In the race, drivers cannot push their car to its maximum ability throughout the race due to tire degradation. Because of this, drivers must pick and choose when to push their car to overtake opposing drivers. Our linear regression model predicts drivers finishing position using starting position, driver age, number of stops, average pit time, and average lap time. Linear regression was used because of its ability to predict numeric outcomes and simplicity. Our model had an $R^2$ value of 0.541 (Appendix Figure 4) showing decent performance in predicting final position for Formula 1 drivers. This value means that the model explains 54% of variance in the dependent variable, the finishing position.

Figure 1.

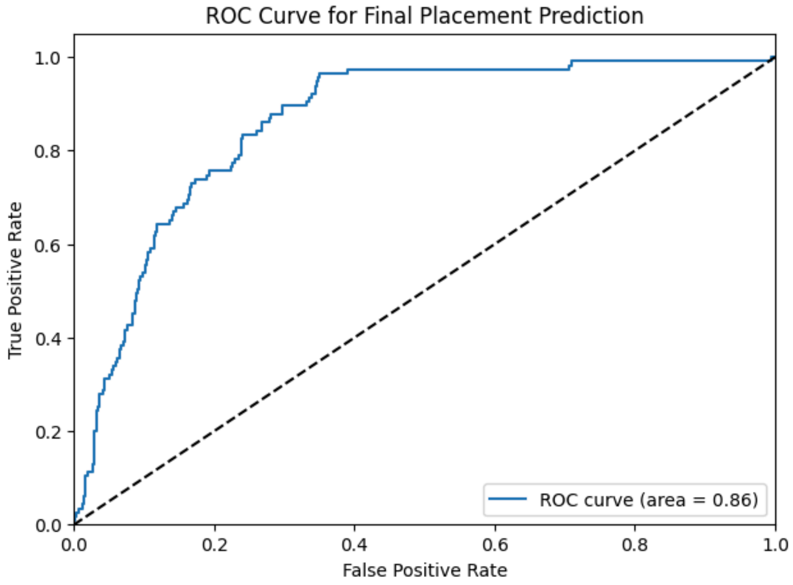


Figure 1 utilizes a ROC AUC graph showing the true and false positivity rates of our model. This ROC curve positively reinforces our analysis, as the graph shows that our model is powerful enough to predict with 86% accuracy if a driver placed in the top 10 drivers or the bottom 10.  The model has a high true positive rate and a low false positivity rate, showing its efficacy in predicting driver placement in top 10 and bottom 10 placement. This means the model doesn't make many mistakes in its prediction. If our regression model was  poor, we would see that the ROC curve is closer to or below the dotted line which means our model is based more on randomness than correctness. Our model can be used by constructors to predict which drivers their drivers are matched up against in the race. This is further supported by the fact that our assumptions for linear regression hold true (Appendix Figure 5). Based on the output, features that were most significant were starting position, driver age, and average pit time. Because the first two are difficult/or impossible for the constructors to control, later in this paper will perform an analysis on pitting.

A problem common with linear regression is collinearity in the dependent variables. This is when the dependent variables can be expressed in terms of each other, meaning they have their own dependence on each other. To check against this we produced a pairplot, graphing each dependent variable's raw data against all other variables. This figure is available in Appendix Figure 6. Each graph shows very little to no correlation among these variables, evidenced by the random spread of data points. This makes us confident there is no collinearity among our dependent variables.

To further support our model we take a look at the residual plot with respect to our dependent variable. The residuals can be thought of as the extra noise in our equation predicting the final placement of drivers, and they represent the difference between the actual final placement value from the data and the predicted value of our model. We graph a Locally Weighted Scatterplot Smoothing (LOWESS) line that shows the relationship between the residuals and the dependent variable. For an accurate model we should see a random distribution of data points and a horizontal LOWESS line running through 0 on the y axis. We see in Appendix Figure 5 that the LOWESS lines for each of the dependent variables are not strictly horizontal, but with the exception of some outliers the data points are randomly scattered (note that num_stops is the number of stops in a race and can only be an integer value, but the data is randomly spread among these integers). Additionally, there is no actual structure to the LOWESS line, so we can be fairly confident that there is no missed relationship among our dependent variables.

This means that our model is an actual linear function, but to perform further analysis on the accuracy of our model, we can perform an A/B test on the dataset to see if a random distribution of changes to our final placement can be accurately predicted by a linear regression model. This will give us further confidence in our results.

## A/B Testing for Placement Prediction

The purpose of A/B testing is to discern a difference between the groups A and B, where group A has a treatment or effect applied to it. In this experiment several of these tests were run with a treatment to the final placement of the drivers.

These tests can be important for either choosing between groups A and B, or analyzing the effect of a treatment on a dataset, and measuring a method's ability to detect this treatment. In this case the A/B test is used for the latter, to make a statement that a linear regression model is accurately able to interpret the dataset with statistical significance. This is especially useful for the F1 dataset with ~ 5000 data points that could confuse a regression model.

Figure 2.

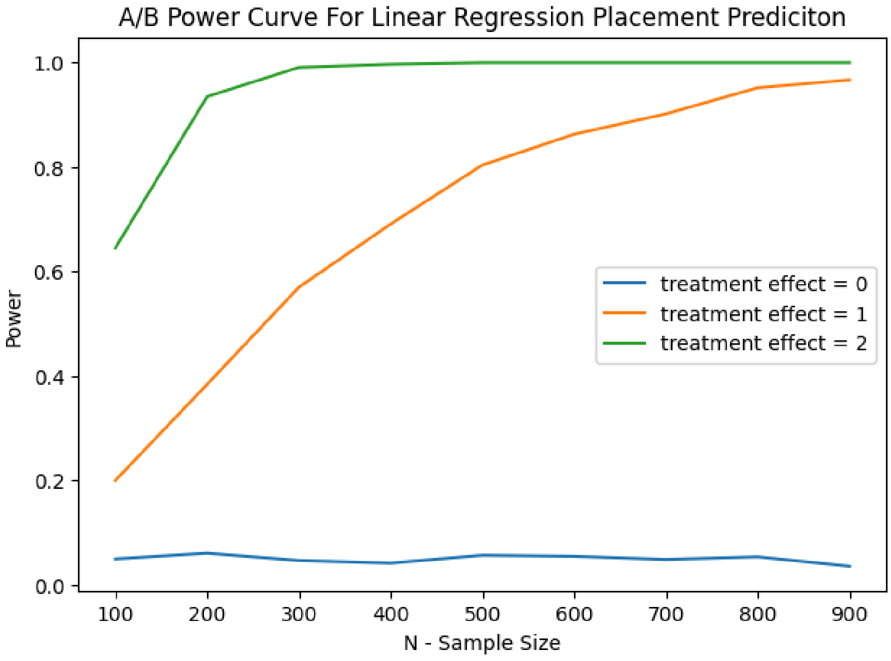


Figure 2: The power curve for the linear regression model designed to detect a treatment in the final placement of each driver. The power is a measure of the models ability to accurately state when there has been a statistically significant treatment to the outcome variable. In this case a treatment effect to the final placement of 0, 1, and 2 is applied, which moves each treated driver back that many places. The X-axis is the sample size of the larger dataset that was treated.

These tests check for statistical significance in the p-value of the treatment coefficient. What this means is that it is testing how much of an effect a feature actually has on the outcome. If this value is less than .05 then a success is recorded in binary output as a one, otherwise as a zero. We sum up these values over many trials and take the average as the power value, so a power value of one would mean the model is completely accurate 100% of the time in detecting a treatment. In Figure 2 we have the power curve for three different treatment levels. We see that at a treatment of zero the model consistently identifies no significance. For a treatment of one, that is the drivers in group A being moved back one in final placing, there is a low power that steadily increases as we increase our sample size. This is because the minute sample sizes do not carry enough information for the model to make an accurate decision. For a treatment effect of two there is considerable power throughout, showing that even at smaller sample sizes if the treatment is large the model will recognize it.

Standard practice suggests a power rule of .8 is optimal, and we achieve this at a sample size of 600 for a treatment effect of one. Since the previous linear regression model we analyzed is predicting placements of one or greater with a train/test split of .7/.3 (both larger than a sample size of 600), we can conclude that the model is accurately able to judge significance in a treatment effect.

This is important because we can now trust that the ROC curve in Fig 1 is an accurate representation of the model's performance on the test data. We can now also be more confident in the findings of the regression summary in Fig 8 of the appendix. Additionally, Fig 2 shows us that sample size has a strong effect on the power when treatment is small, and so we now know that more data will actually be beneficial and not a diminishing return. This gives us an avenue to explore further strengthening of the Final Placement Prediction model.

## Predicting Pitstop Timing

Predicting when opponent drivers will enter the pitlane to perform a pitstop is important as this is an opportunity to overtake and/or decide when you should pull your drivers into the pitlane. Therefore, we decided to try and understand when a driver might decide to take a pit stop as this is invaluable information for opposing teams when developing strategies.

Our first attempt at predicting pit stop timings came in the form of using logistic regression as it is widely used for predicting the probability of a binary event such as whether or not a driver will take a pit stop on a given lap. We found that this method would most likely not produce usable results due to the sparsity of the data. Most drivers pit one to three times a race, and spend the other fifty or more laps driving which resulted in the ratio of positive labels (pit stop on a given lap) to negative labels (no pit stop on that lap) to be about 1:66.

The results from the logistic regression (Appendix Figure 7) showed that even though it predicted whether or not a pit stop would occur with 97% correctness, it did not have any predictive value. This is because that correctness came from classifying every lap as a lap without a pit stop which is true approximately 97% of the time but we are interested in the 3% of the time where pit stops do occur and this model could not predict that. So, we pivoted to a more complex model that would hopefully produce better results.

The new model we used to try and better predict when a driver might take a pit stop was a random forest classifier. We thought that this model might do better because of its ability to classify much more complex patterns than a standard logistic regression. We were hoping that the Random Forest would be able to identify a specific trend that was missed before by the regression model.

The classification report (Appendix Figure 8) shows that it may have done a little better than the logistic regression as it had a precision of 0.67 for laps with pit stops. However, after closer inspection we found that there was only a recall of 0.12 which means that the model failed to identify a significant portion of the positive instances within the data set. On the other hand, it was near perfect in identifying the negative instances but this again could be due to the massive imbalance in the data set of positive to negative labels.

Despite the suboptimal results we constructed an ROC curve which plots the true positive rate vs the false positive rate of the model on the testing data. Similarly to our linear regression analysis, this ROC curve shows that our model correctly predicts when a driver will pit most of the time. It shows that the outputs of our model are not random, as the ROC curve is not close to the random chance line. The goal of this was to see another angle of our model and if its classification of data was at all usable. The graph of the ROC Curve can be seen in Figure 3.

Figure 3.

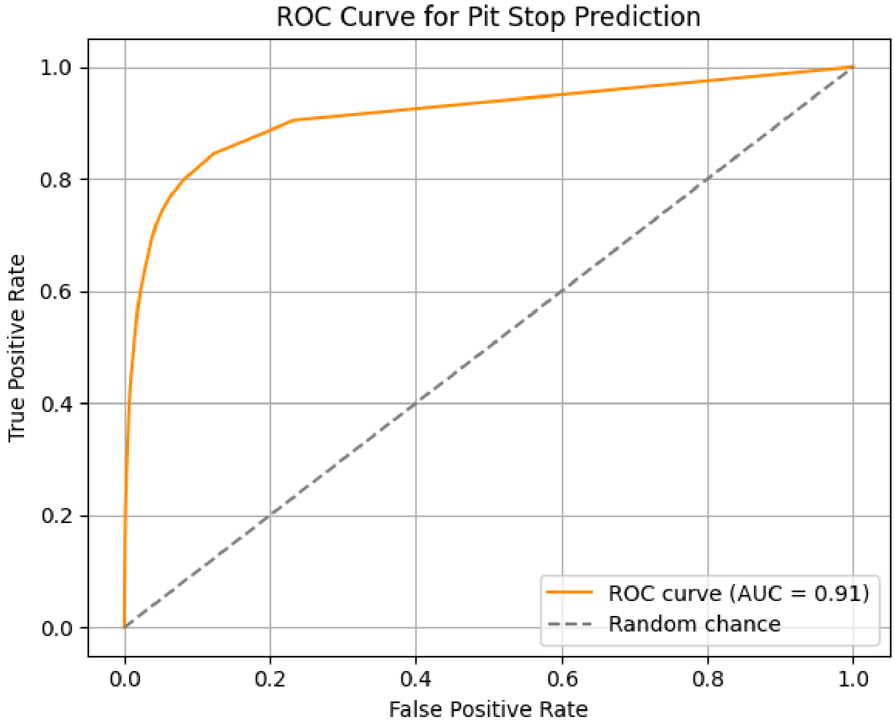


Figure 3: This is the ROC curve for the random forest model that predicts the lap where a pit stop will occur. This ROC curve is decently above the random chance line. The graph starts very steep and then flattens out indicating that when false positive rate is low, the true positive rate is high.

This ROC curve actually contradicts the suboptimal results from earlier as it is very close to the ideal curve which goes straight up along the y-axis and then straight across at y = 1. This curve is well above the random chance line of y = x and the area under the curve (AUC) is 0.91 which is very close to the ideal value of 1. An AUC of 0.91 means that the model correctly ranks a randomly chosen positive instance higher than a randomly chosen negative instance 91% of the time. This ROC curve shows that this model may be usable in predicting when a particular car may take a pit stop during a race.

One final take away from the random forest model is what features seem to be most important to predicting when a pit stop will occur. We found that the change in lap time between two consecutive laps is a very strong indicator of when a pit stop will occur with most other features being less important. These results can be seen in Appendix Figure 9.

To conclude this section, we attempted to accurately predict when pit stops would occur as to aid teams in their strategy making. We found that logistic regression was unsuccessful in predicting pit stops due to the sparsity of the labels. Upon pivoting we found that a random forest classifier did a better job and upon inspection of the ROC curve we found that it actually may have done a good job of classifying the data and could potentially be able to predict when a driver will take a pit stop.

# Further Analysis/Conclusion

From our three models, we gather insight on the ability to predict the final position of a racer, the specific causes for a final position, as well as the ability to predict the timing of a pit stop. From our first method, after running our linear regression model, we were able to figure out what features were the most important in predicting the final position. These features include average lap time and pit stop time.

After discovering these important features, we analyze the degree of the effect of these significant features by performing an A/B test. The results from this test showed the particular significance of the p values from our linear regression, highlighting the actual effect of our features on the final position. Our third model, the random forest model predicting pit stops, was able to provide for a decently usable model that would predict when drivers would enter the pitlane.

After creating these models, we gain more insight into what determines a driver's performance in a race. This insight provides potential value to the Formula 1 constructors. Because the main goal for these teams is to win championships, understanding what factors maximize their chances of winning is important and useful.

Further analysis that could be done in relation to team strategy and the outcome of a race includes taking into account weather conditions and tire type. Weather conditions have a significant impact on the race track. If the track is moist, tire degradation will be less than if it was dry, thus affecting the amount of pit stops required and the timing of pit stops.

# Recommendations

As a result of our data analysis, we recommend that the Formula 1 teams can utilize our models to inform their team building decisions and race strategies. When formulating a team, we propose that the significant features found in our regression analysis, such as average lap time will be taken into account when determining what the final position of their driver will be. In creating a race strategy, teams must carefully plan their timing of a pit stop. It is essential that they take into account the timing of a pit stop of their opponents in order to maximize their chance of gaining a lead. The decision of when to make a pit stop is heavily based on the actions of rival teams (Motorsport, 2021), so we hope these models aid in predicting those actions.

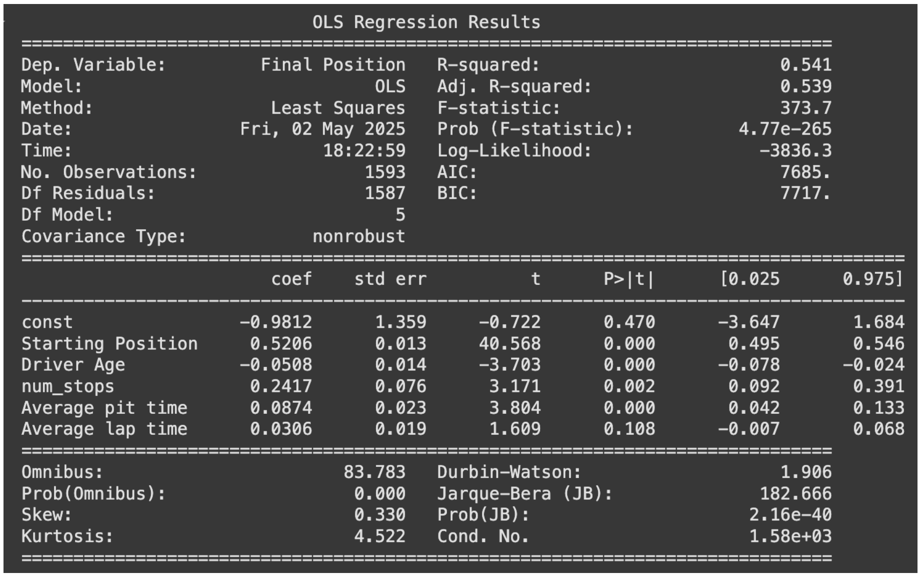
# References

Mayne, J., & Cox, B. (2025, March 11). *F1 prize money 2025: Payout breakdown, how much drivers and constructors earn*. Sporting News. https://www.sportingnews.com/us/formula-1/news/f1-prize-money-how-much-drivers-earn-constructors-championship/6e42726b49165fd1fec4ee7d

Motorsport Academy. (2021). *Defining Winning Strategies: Pit Stop Tactics.* Motorsport Engineer. https://motorsportengineer.net/how-race-strategy-works-in-formula-1/

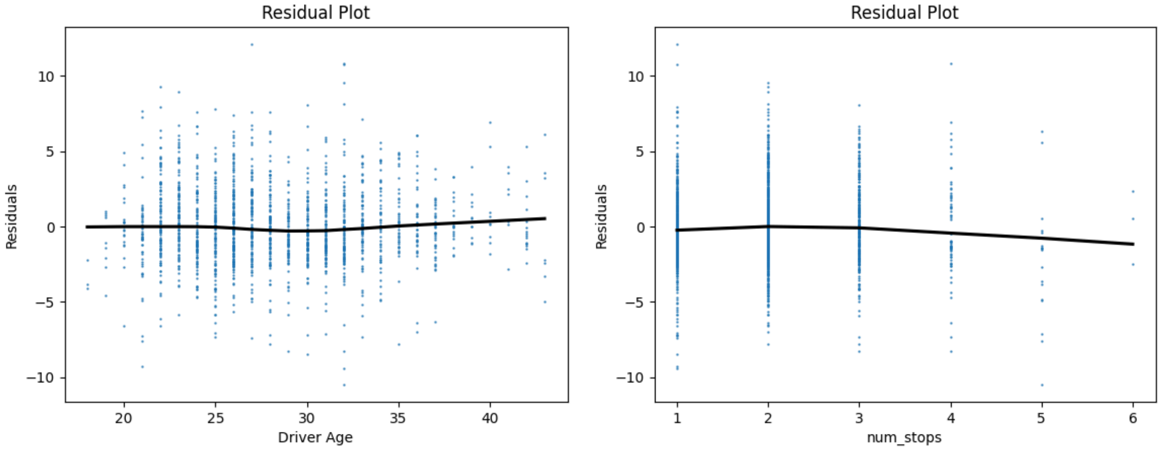
# Appendix

Figure 4. OLS Regression Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Final Position   R-squared:                       0.541
Model:                            OLS   Adj. R-squared:                  0.539
Method:                 Least Squares   F-statistic:                     373.7
Date:                Fri, 02 May 2025   Prob (F-statistic):          4.77e-265
Time:                        18:22:59   Log-Likelihood:                -3836.3
No. Observations:                1593   AIC:                             7685.
Df Residuals:                    1587   BIC:                             7717.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.9812      1.359     -0.722      0.470      -3.647       1.684
Starting Position     0.5206      0.013     40.568      0.000       0.495       0.546
Driver Age           -0.0508      0.014     -3.703      0.000      -0.078      -0.024
num_stops             0.2417      0.076      3.171      0.002       0.092       0.391
Average pit time      0.0874      0.023      3.804      0.000       0.042       0.133
Average lap time      0.0306      0.019      1.609      0.108      -0.007       0.068
==============================================================================
Omnibus:                       83.783   Durbin-Watson:                   1.906
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              182.666
Skew:                           0.330   Prob(JB):                     2.16e-40
Kurtosis:                       4.522   Cond. No.                     1.58e+03
==============================================================================
```

Figure 5. Residual Plots for checking linear regression assumptions



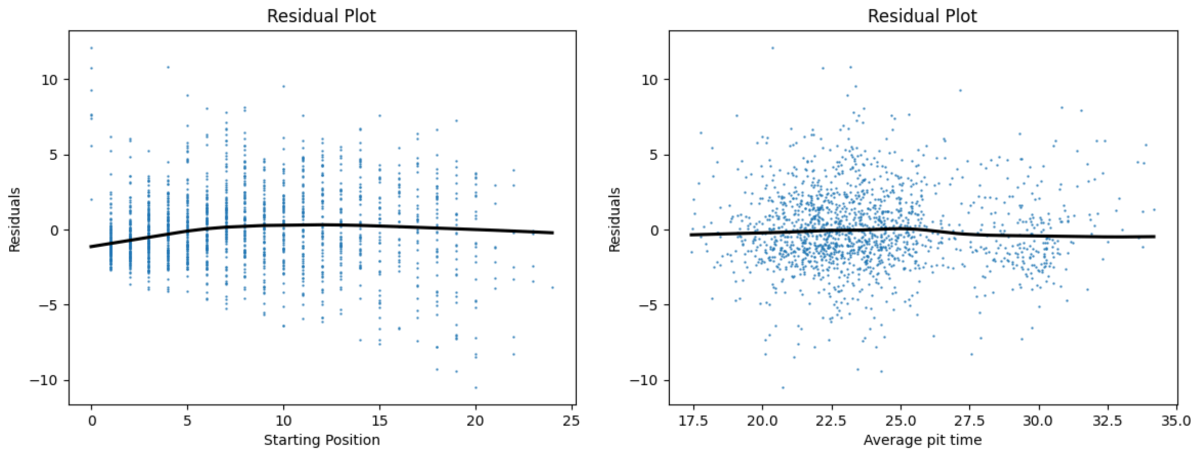


Figure 6. Linear Regression Pairplot

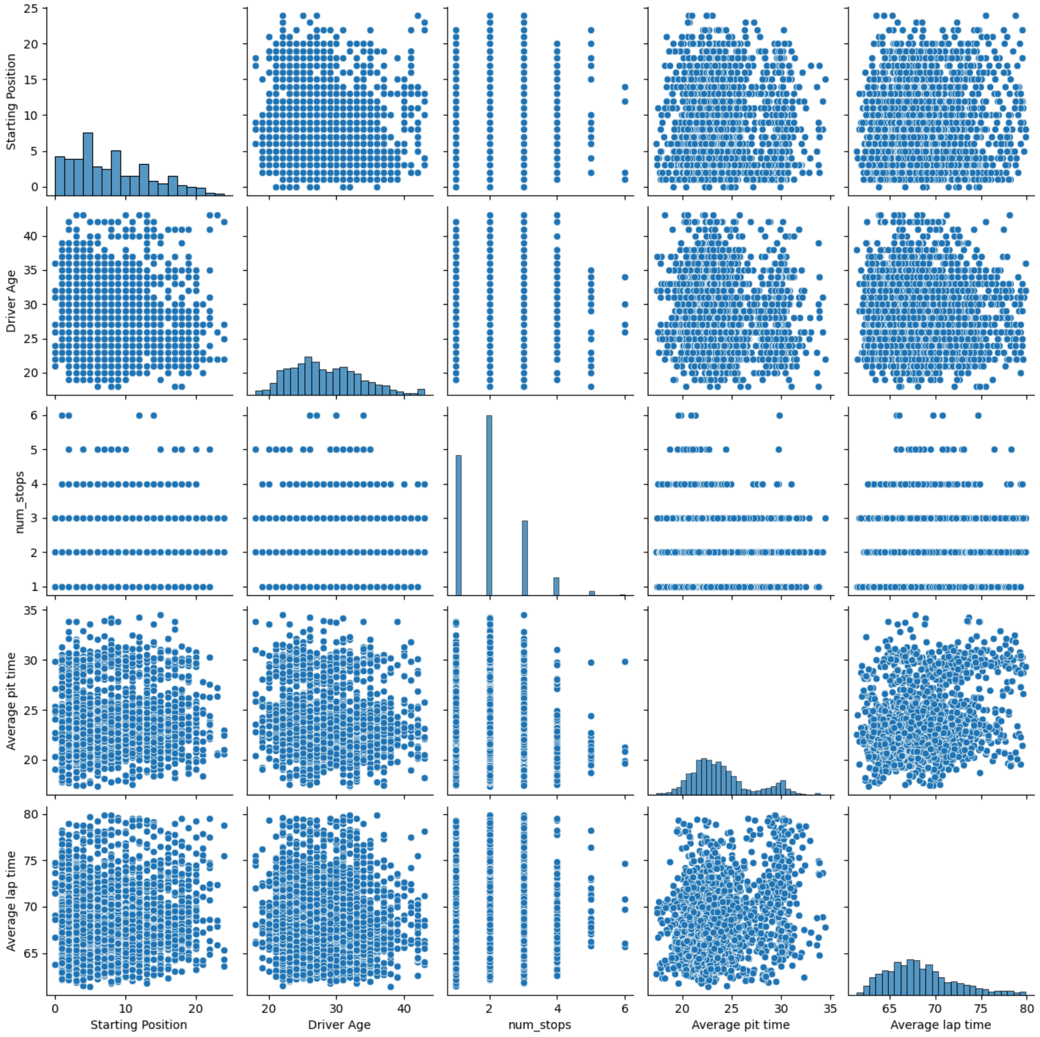
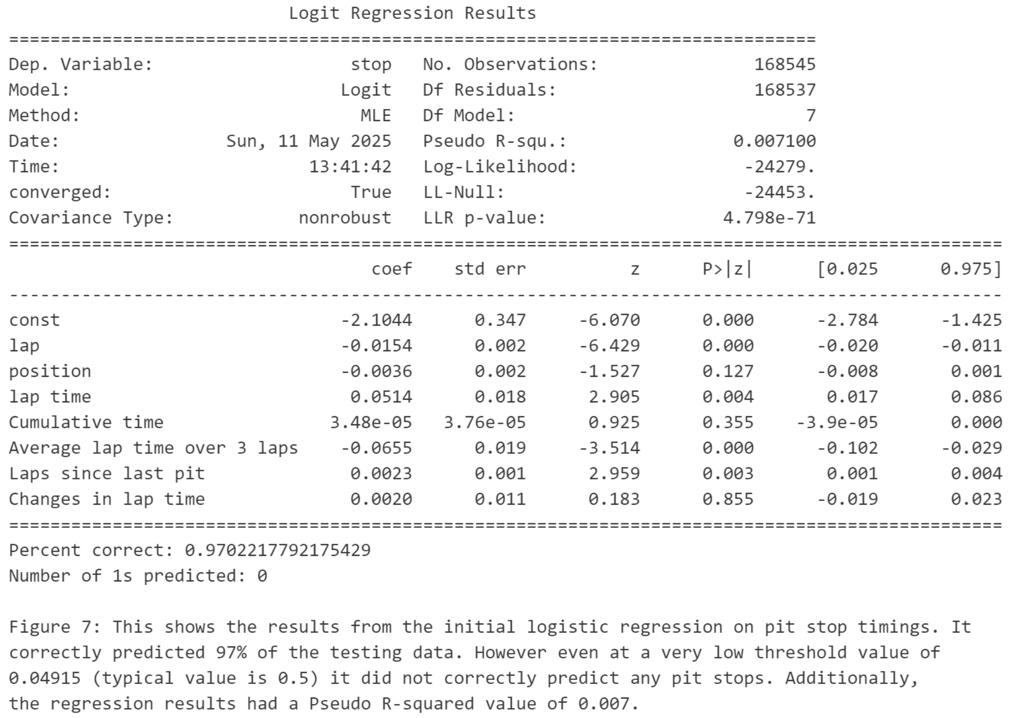


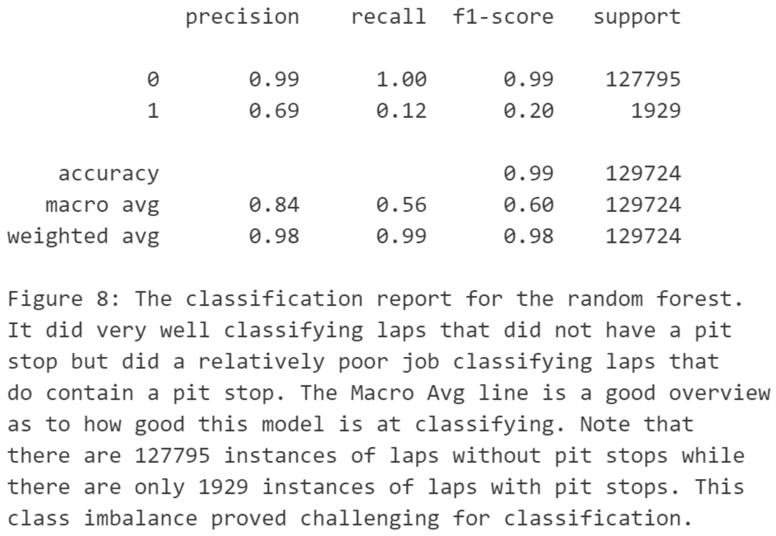
Figure 7. Logistic regression results



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                   stop   No. Observations:               168545
Model:                          Logit   Df Residuals:                   168537
Method:                           MLE   Df Model:                            7
Date:                Sun, 11 May 2025   Pseudo R-squ.:                0.007100
Time:                        13:41:42   Log-Likelihood:                -24279.
converged:                       True   LL-Null:                       -24453.
Covariance Type:            nonrobust   LLR p-value:                 4.798e-71
================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -2.1044      0.347     -6.070      0.000      -2.784      -1.425
lap                           -0.0154      0.002     -6.429      0.000      -0.020      -0.011
position                      -0.0036      0.002     -1.527      0.127      -0.008       0.001
lap time                       0.0514      0.018      2.905      0.004       0.017       0.086
Cumulative time              3.48e-05   3.76e-05      0.925      0.355    -3.9e-05       0.000
Average lap time over 3 laps  -0.0655      0.019     -3.514      0.000      -0.102      -0.029
Laps since last pit            0.0023      0.001      2.959      0.003       0.001       0.004
Changes in lap time            0.0020      0.011      0.183      0.855      -0.019       0.023
================================================================================================
Percent correct: 0.9702217792175429
Number of 1s predicted: 0
```

Figure 7: This shows the results from the initial logistic regression on pit stop timings. It correctly predicted 97% of the testing data. However even at a very low threshold value of 0.04915 (typical value is 0.5) it did not correctly predict any pit stops. Additionally, the regression results had a Pseudo R-squared value of 0.007.

Figure 8. Random Forest classification report



```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    127795
           1       0.69      0.12      0.20      1929

    accuracy                           0.99    129724
   macro avg       0.84      0.56      0.60    129724
weighted avg       0.98      0.99      0.98    129724
```

Figure 8: The classification report for the random forest. It did very well classifying laps that did not have a pit stop but did a relatively poor job classifying laps that do contain a pit stop. The Macro Avg line is a good overview as to how good this model is at classifying. Note that there are 127795 instances of laps without pit stops while there are only 1929 instances of laps with pit stops. This class imbalance proved challenging for classification.

Figure 9. Features of importance from the random forest classifier

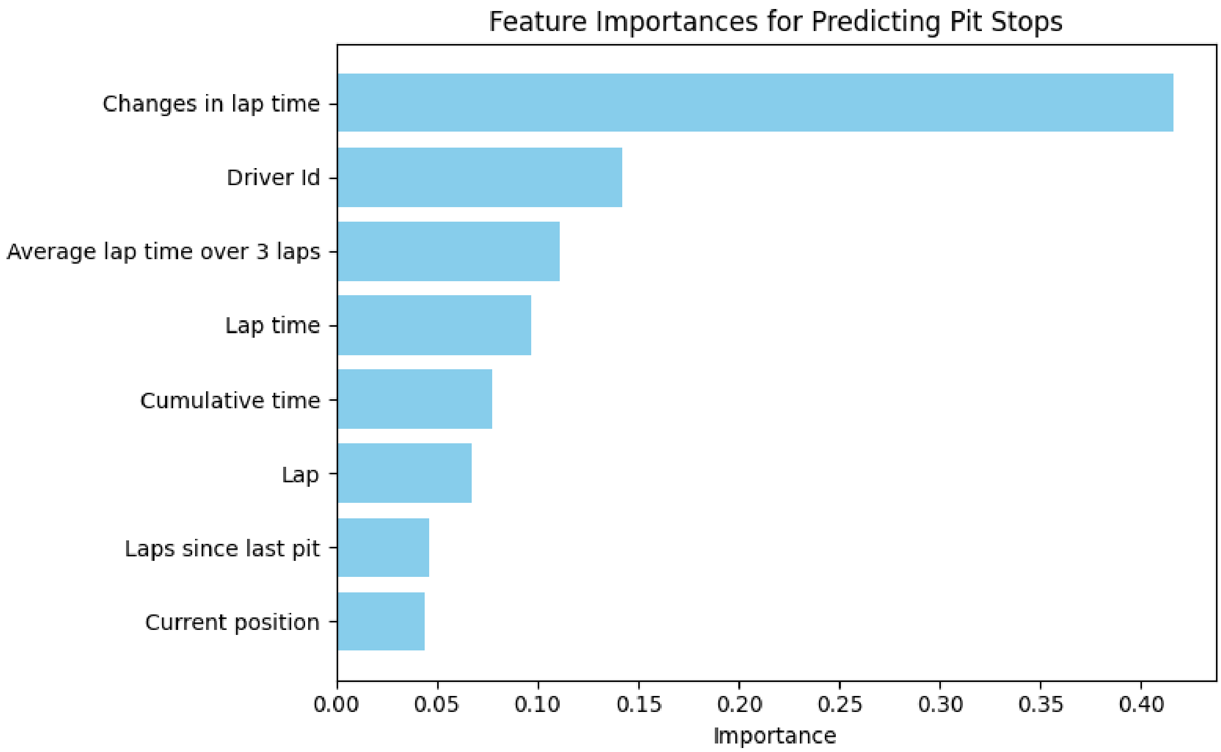


Figure 9: Important features based on Random Forest Classification. These results are from using sklearn's Random Forest Classsifier on our data. From this graph, we can see that changes in lap times was by far the most important predictor for when a pit stop will occur with an importance level over 0.40. After that, Driver Id and Average lap time over 3 laps were the next mostimportant but at importance levels of less than 0.15. Lap time, cumulative time, lap number, number of laps since the last pit and current position were relative unimportant predictors according to this model.