F1 Data Analysis: Exploring Performance Metrics to predict Race Outcomes

Name: Daniel Kim, Ryan Lee, Nikita Gohil

Background

Formula 1 has been a worldwide sport that dates back all the way to the 1940s. Over recent years, the popularity of the sport among fans worldwide has grown drastically and exponentially. As of 2024, the global fanbase of F1 has 750 million fans worldwide. From this, just over the past year the sport has garnered and generated a staggering revenue of $3.4 billion in just this past year. Considering the constantly growing fan base as well as the economic gains the sport has experienced, teams within F1 are now more than ever encouraged to perform well in the races across the year. Better performance by teams will lead to more fans and in turn, more profits for that team. Due to this, teams over recent years have hired and onboarded tons of engineers and analysts to look at which factors can maximize a driver's chance of winning a race and ultimately become the best team for that year. This data can include information on pitstop performance, engine performance to aerodynamics of the car build. For this, we want to analyze and predict future race outcomes based upon various metrics to determine which factors can increase a team's ability to win, particularly as the performance gap between teams has narrowed over the years.

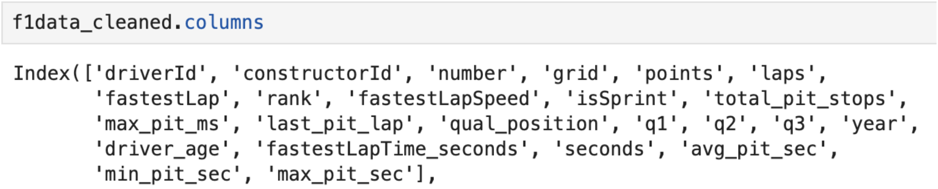
Question

Our research will investigate the development of different regression models to forecast the probability of certain race outcomes. Additionally, the research will gain insight into the features that are most important in predicting these outcomes. Thus, the primary questions were:

What factors can teams investigate to help predict their performance in races? What are the accuracies and performance of these different models to predict race results based on points earned?

Dataset

Our dataset consisted of multiple csv files regarding various race results, times, team information, and driver information. The aggregated dataset consists of data from races beginning from the inaugural season in 1950 until the 2024 season, resulting in roughly 27,000 data points with 38 features. After cleaning and preprocessing our data, the dataset used had a sample size of 2,020 rows over 24 features.



```
f1data_cleaned.columns
```

```
Index(['driverId', 'constructorId', 'number', 'grid', 'points', 'laps',
       'fastestLap', 'rank', 'fastestLapSpeed', 'isSprint', 'total_pit_stops',
       'max_pit_ms', 'last_pit_lap', 'qual_position', 'q1', 'q2', 'q3', 'year',
       'driver_age', 'fastestLapTime_seconds', 'seconds', 'avg_pit_sec',
       'min_pit_sec', 'max_pit_sec'],
```

Listed above are the features we used when creating our models. We chose to use 23 features to predict and analyze the dependent variable ‘points’, which represents the points a driver earned for a given race.

Dataset: F1 race data from 1950 to 2024

Methodology (*pre-processing or EDA*)

Although the aggregated dataset resulted in roughly 27,000 rows over 38 features, the final dataset was reduced to 2,020 rows over 24 features because of pre-processing and cleaning the original dataset. After converting the relevant columns to numeric values, where possible, the columns that measured time in milliseconds were converted to seconds as scale. The only feature that was converted from millisecond to seconds was the data on pitstop and time recorded for qualifying rounds. The fastest lap time was converted from minutes to total seconds. With these newly created time-scaled features, the originals were dropped as well as any redundant and irrelevant features such as any features with the function of only identifying races. The only identifying features were those used to encode the identification of the team and the driver. Considering the discovery of specific teams or drivers who have unique abilities to outperform the group was also targeted. Lastly, all rows with missing values for any features were dropped from the data set. This method was selected over dropping any features that had too many missing values within in as retaining the highest number of features was desired over retaining as many observed data points. As a result, all the data points in the final dataset were only those that were from recent years, dating back to more recent times than the inaugural season in 1950.

Methodology (*Models)*

Linear Regression

To establish a baseline for predicting race points in Formula 1, we implemented a multiple linear regression model. This approach was selected due to its interpretability and ability to reveal linear relationships between predictor variables and the target outcome—driver points.

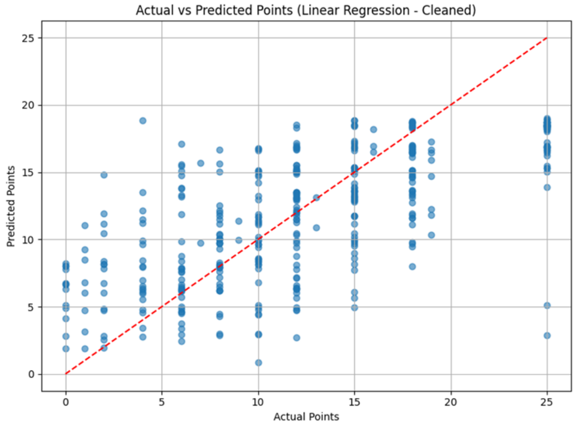


Figure 1: Linear Regression Model

We started with a dataset containing over 27,000 records across 38 features, which was reduced to 2,020 observations and 24 features after comprehensive preprocessing. This process involved removing irrelevant or redundant identifiers (e.g., race IDs), dropping rows with missing values, and retaining only numeric and essential encoded features related to race performance. From this cleaned dataset, we engineered and selected the following key features based on domain relevance and correlation analysis: driver_age, fastestLapSpeed, grid, avg_pit_sec, position_gain, and lap_speed_ratio. All numerical variables were standardized using z-score normalization to ensure comparability across scales. The cleaned dataset was split into 70% training and 30% testing sets using stratified sampling to ensure representative target variable distribution. The linear regression model was trained using scikit-learn's LinearRegression function, without regularization, to maintain transparency in coefficient interpretation. This allowed us to identify key variables with strong linear relationships to the number of points earned.

Ridge and Lasso Regression

The further develop the work done on the linear regression analysis, the lasso and ridge regularized regression analysis was conducted to penalize any coefficients that added noise to the model, only causing concern for overfitting. The dataset was split into 70%/30% training and testing sets to develop this model. The features, or independent variables, were also scaled to ensure not one feature would dominate its impact on the dependent variable. Prior to training the lasso and ridge regressions, the optimal alpha values were determined by iterating through different appropriate alpha levels that would optimize the penalty parameter to reduce coefficients responsible for overfitting. In Figures 3 and 4 respectively, the optimal alpha value was 0.1 for the Lasso regularized regression while the optimal alpha value was 100 for the Ridge regularized regression as these values produced the highest accuracies for their respective models.

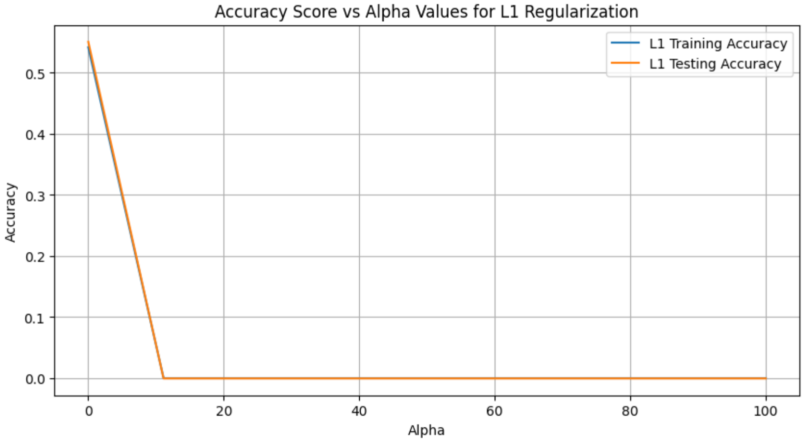


Figure 2: Optimal Alpha Value for L1 (Lasso) Regularization

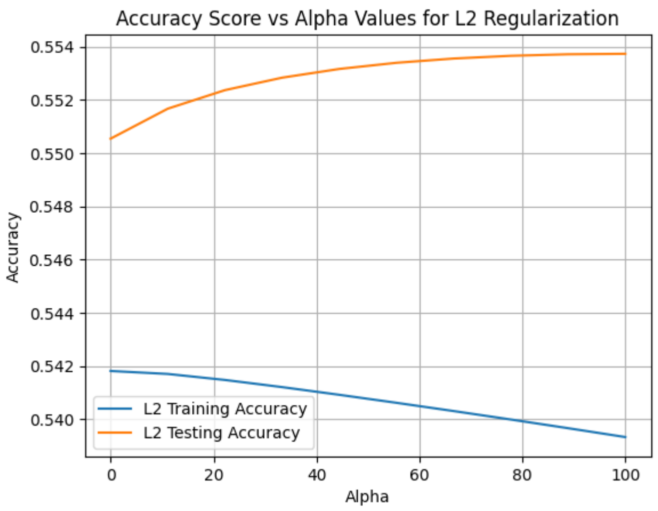


Figure 3: Optimal Alpha Value for L2 (Ridge) Regularization

Gradient Boosting

Our last model was created utilizing a gradient boosting algorithm in order to predict points of score per race. Gradient boosting was chosen due to its ability to iteratively correct prediction errors and capture complex non-linear interactions, particularly useful for sports performance datasets where feature effects may not be additive or independent. Like the previous models, a train-test split of 70/30% was used. With the gradient boosting model, a grid search cross-validation was used to check all potential combinations and determine the optimal hyperparameters to utilize. The hyperparameters that were tuned include the number of estimators, the learning rate of the gradient boost, the max depth, the min sample split, and the min sample leaves. From this, optimal parameters were determined in accordance with our training data.

Optimal Parameters

n_estimators	200
learning_rate	0.05
max_depth	3
min_samples_split	2
min_samples_leaf	1

Table 1: Determined Parameters for Gradient Boosting

Above are the optimal parameters of our grid search with cross validation determined for our Gradient Boosting model. Uses these parameters the training data was fit onto the model and used to predict the test data point outcomes. From this R2, MAE, MSE, and RMSE scores were calculated against the actual point outcomes to determine the accuracy of this regression model.

In addition, feature importances were also looked at to see which metrics attributed the most to predict point outcomes for races.

Results

Linear Regression

The linear regression model produced moderate but meaningful predictive accuracy. Evaluation on the test data yielded the following performance metrics:

Metric	Value	Interpretation
R² Score	0.4474	Model explains ~44.74% of the variance in points scored
Mean Absolute Error	4.00	On average, predictions are off by 4 points
Mean Squared Error	25.77	Average squared prediction error; more sensitive to large errors
Root Mean Squared Error	5.08	Standard deviation of prediction errors; indicates typical prediction deviation

Table 2: Evaluation Results for Linear Regression Model

These results indicate a moderate fit to the data, which is expected from a linear model without regularization or interaction terms. The R² score of 0.4474 suggests that while the model captures some key patterns, over half of the variance in driver points remains unexplained. The scatter plot (Figure 3) of actual versus predicted values shows a general alignment along the diagonal line (y = x), but with increasing spread at higher point values. This implies that the model is less accurate when predicting top-performers or edge cases—likely due to non-linear relationships that linear regression cannot model effectively. In summary, the model performs reasonably well as a baseline and successfully identifies meaningful predictors of F1 driver performance. However, its limitations in capturing complex interactions justify the use of regularized and non-linear models, as explored in the following sections.

Lasso and Ridge Regression

Given the alpha values that optimized each model, the model performance for the respective models is shown in the table below. As presented, the performance of both models was comparable, showing that one model was not better than the other.

Metric	L1 (Lasso)	L2 (Ridge)
R2	0.5556972190746714	0.5537311838338472
RMSE	4.817681337490508	4.828328667121476
Training Accuracy	0.5393374187683425	0.5393374187683425
Testing Accuracy	0.5537311838338472	0.5537311838338472

Table 3: Performance Metric for Lasso and Ridge Regressions

With the respective alpha values, the Lasso and Ridge coefficients were modified to account for the penalty parameters. As seen in the table below, the Lasso regression zeroed out a few features, essentially eliminating these features from the model, while the Ridge regression only reduced all the features to near zero, retaining all the features and basically minimizing the impact of these features on the model. As both models performed similarly, the coefficients from the Lasso regression can be utilized, eliminating unnecessary features. In both, however, rank, qualifying positions, and grid positions had a negative relationship with the dependent variable which meant that the lower values indicated an improved likelihood of scoring more points in each race.

Feature	Feature Description	Lasso_Coefficient	Ridge_Coefficient
6	rank	-2.671264	-2.629244
12	qual_position	-2.586789	-2.345589
3	grid	-0.909122	-1.103787
0	driverId	-0.503575	-0.816568
9	total_pit_stops	-0.613870	-0.705988
11	last_pit_lap	-0.509167	-0.549576
13	q1	-0.062125	-0.394993
4	laps	0.120196	0.334346
17	driver_age	-0.054630	-0.307168
10	max_pit_ms	0.051506	0.266933

Table 4: Top 10 Feature Coefficients for Each Model

Gradient Boosting

Gradient Boosting Score Metrics

Training R2 Score	0.7190813657669826
Test R2 Score	0.5950811796501595
Test Mean Absolute Error (MAE)	3.5646155072701595
Test Mean Squared Error (MSE)	21.152664072000047
Test Root Mean Squared Error (RMSE)	4.599202547398847

Table 5: Performance Metric for Gradient Boosting

After performing our Gradient Boosting model, the model reported a training data R2 score of 0.71908 and a test data R2 score of 0.59508. The Mean Absolute Error on the test data was reported to be 3.5646, the Mean Squared Error on test data was reported as 21.15266, and the Root Mean Squared Error was reported to be 4.59920. These metrics show improvements compared to the accuracy scores reported in the linear, lasso, and ridge regression models.

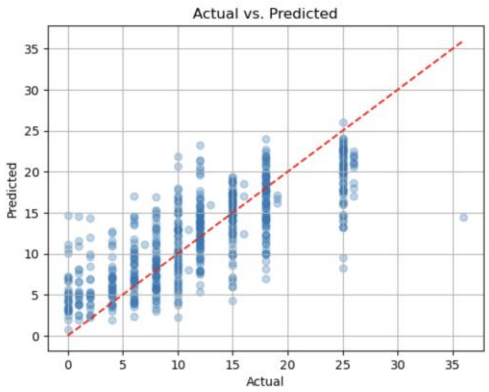


Figure 4: Actual vs. Predicted Points from Gradient Boosting

To further analyze the performance of the Gradient Boosting model, a plot of predicted point values for the test data compared to the actual point values of the test data was created. Looking at the plot, we can see that while the model was able to correctly predict some point outcomes, there is still a decent amount of error for certain point outcomes in our test data.

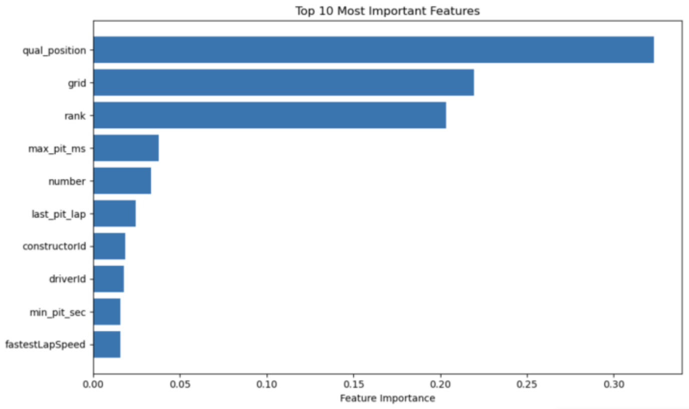


Figure 5: Feature Importance of Gradient Boosting

After our model, the gradient boosting also determines feature importances for each predictor used which were the same features determined to be of importance in the Lasso regression. Qualifying position grid and rank made up the top three respectively demonstrating that point outcomes from a race aren’t necessarily affected based upon race metrics but also how well the driver performed in events leading up to the race. With our models, our results