Formula 1 Predictor

Presentation by: Vignesh, Vinay, Shah, Rama and Sunanda

Why Formula 1

- Formula 1 represents the pinnacle of automotive technology with cars being fitted with the latest sensors to detect gather and transmit gigabytes of data to further push the boundaries of automotive technology
- Formula 1 has faced many criticisms from top analyst as well as the the racers that participate that the sport is too predictable and that there is not enough competition and we want to test this theory
- One other topic is that the same racers and teams tend to always win so we want to determine if their strategy is better than other teams or if it is a matter of having the best car and driver

Description of Source Data

- Data provided by Kaggle
- Abundance of data which is highly labelled and has data stretching back 40 years
- Scrapped the formula 1 website to retrieve weather data for each event for the last 40 years

Question we Aim to Answer

1. Given the amount of historical data gathered from past races and events we seek to answer the question, Are formula 1 races too predictable?
2. Given certain locations of races require different strategies and approaches to maximize results we seek to answer the question, Can we create segmented race strategies for F1 constructors depending on circuit level historical performance?

Exploratory Analysis

- First determined whether the same team has been winning most of the events or if the winning is distributed amongst the teams. Noticed that from 2000-2009 Ferrari has been dominating the F1 year and from 2010-2019 Mercedes has been dominating
- When analyzing the points scored by each team we notice that Ferrari scored the most points from 2000-2009 while Mercedes scored the most points from 2010-2019
- Michael Schumacher for Ferrari dominated from 2000-2009 while Lewis Hamilton dominated from 2010-2019.
- One other trend noted irrespective of year is that racers that start in the initial grid position have won the event
- The leader of the race at the fifth lap and the first lap has overwhelming success in finishing first as fifth lap leader wins 57.9% while first lap leader wins 55.2% of the time

Analysis and Machine Learning- Race Prediction

- Initial approach to analyze data started with joining the data from the results and races table. This was done as these tables contained base information for all the races. We further joined in the constructor table and circuits table as these were used to ID the event and the team
- The full table was than trained on using a logistic regressor. Initial accuracy scores were poor as data contained many 0-values
- In second attempt we joined weather data into the existing table. Weather was added as it affects the circuit and drivers have to adjust for the weather
- We retrained the data using the following models: Random Forests, Support Vector Classifier and a Neural Network

Analysis and Machine Learning- Race Prediction

The Following are the results of the 2 iteration of Machine-Learning:

	Machine Learning Models	Accuracy Score	Number of times predicted driver won the race	Number of times predicted driver finished in the top two	Number of times predicted driver finished in the top three
Win Predictor	Logistic Regression	93%	8 out of 21 races	11 out of 21 races	11 out of 21 races
Race Predictor	SVM	95%	12 out of 21 races	15 out of 21 races	16 of 21 races
	Random Forest	94.52%	10 out of 21 races	18 out of 21 races	21 of 21 races

Machine Learning Models	Predictions	Predicted %	Additional Predictions
SVM	Predicting 20 Race Positions With a Spread of 2	38.09%	
	Predicting Driver Finish Bins for 2019	55.95%	Model Predictor correctly for all Podiums - 58.73% Model Predicted correctly for positions three to six - 33.33% Model Predicted correctly for positions seven to ten - 40.47% Model Predicted correctly for Bottom Ten - 68.09%
Random Forest	Predicting 20 Race Positions With a Spread of 2	72.14%	
	Predicting Driver Finish Bins for 2019	77.14%	Model Predictor correctly for all Podiums - 68.25% Model Predicted correctly for positions three to six - 47.61% Model Predicted correctly for positions seven to ten - 67.85% Model Predicted correctly for Bottom Ten - 92.38%

Analysis and Machine Learning- Race Strategy

- During exploration, analysis showed that there are different factors on which a race can be segmented using a clustering algorithm
- Retrieved finishing status of each driver and constructor for each circuit over the past 2 decades
- Used K-means and Hierarchical clustering models to cluster historical performance data of each constructors based on the circuit.
- Moved forward with K-means algorithm as it is easier to determine number of clusters through an elbow curve and model interia.
- Used K-means based on average of fastest lap times to determine fast, medium and slow circuits