F1 Predictive Analytics

Brainstation Capstone - Sprint 2

Manny Brar

Agenda

PROJECT SUMMARY

PROJECT OBJECTIVE

SWOT ANALYSIS

DATA SNAPSHOT

DATA PIPELINE

NEURAL NETWORK MODEL

As Christian Horner, CEO of Oracle Red Bull Racing, says, “Data is in the team’s lifeblood. Every element of performance – how we run a race, how we develop a car, how we select and analyze drivers – it’s all driven by data.”

Project Summary

Back to Agenda Page

Predictive Analytics in Formula 1

This project aims to explore and develop predictive analytics models that utilizes historical race data, weather conditions, and car performance metrics to forecast race outcomes in Formula 1. By analyzing patterns and correlations within the data, the model seeks to provide actionable insights that can help F1 teams make informed decisions regarding race strategies, car setups, and driver management to maximize their competitive edge.

Big Business, Big Risks, Big Data Driven Decissions

With Formula 1 being one of the most valuable sports empires in the world at a staggering $18 billion dollars... This is a big business, where every decision has a major dollar figure attached.

Throughout the history of F1, teams have utilized a combination of machine learning techniques and statistical analysis to process and analyze a comprehensive dataset comprising lap times, tire usage, pit stop strategy, weather conditions, and other relevant variables from past F1 races.

Project Objective

Race Placement & Lap Time Predictive Modelling

This project will utilize machine learning algorithms to predict the AVG lap times for the race based on qualification lap data and explore an extensive dataset that includes lap times from past races and qualification lap results, tire usage, pit stop strategies, weather conditions, and additional pertinent variables from historical Formula 1 races. The predictive model developed from this analysis will undergo thorough training, testing, and validation phases using these datasets, to confirm its precision and dependability in predicting outcomes.

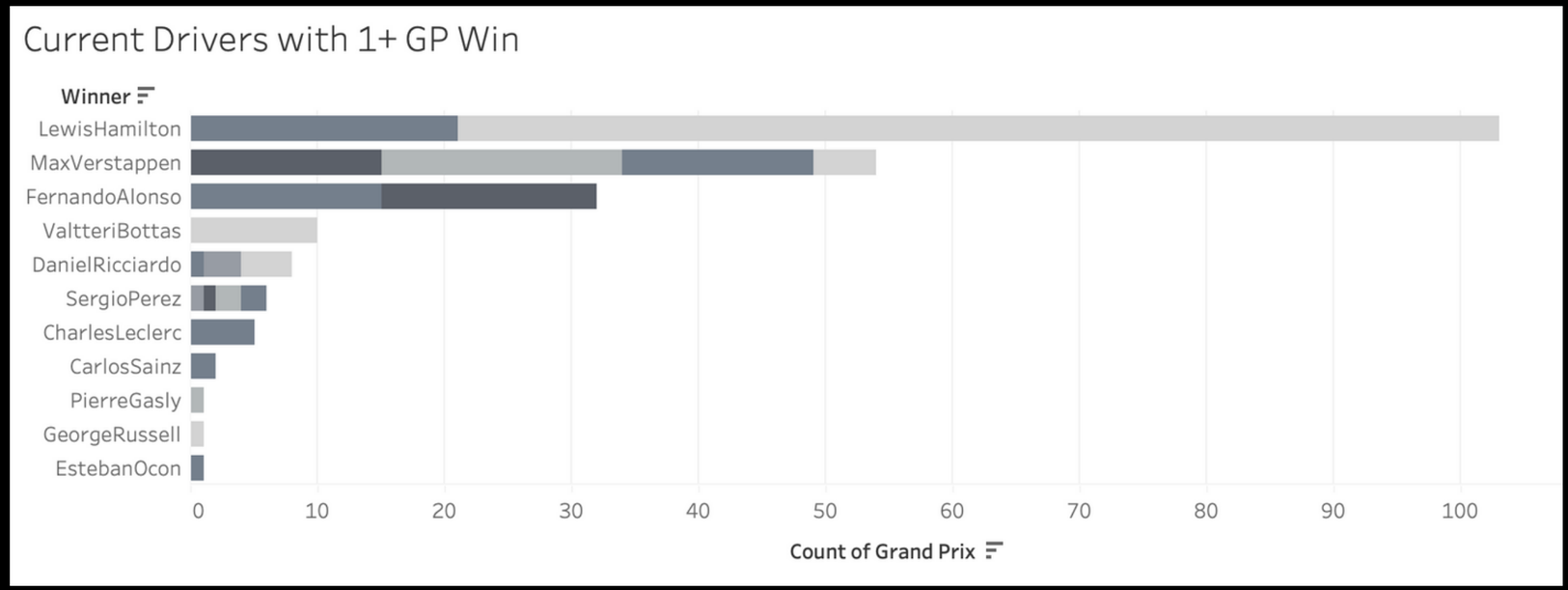
S	W	O	T
Strengths	Weaknesses	Opportunities	Threats
• I was able to extract and clean all the data I wanted to gather from the various sources. I was able to create relationships and create new features that would be beneficial for modeling.	• Tuning the model has been a challenge, but I have made a lot of progress. However I have a long way to go with this process and gaining further understanding of various parameter tuning options when and where to utilize these options and how best to optimize the overall data for modeling.	• I see a lot of potential in next steps and additional models for subsequent predictions but that would after I get this model dialed in.	• Important factors to keep in mind for this project would be, the amount of time required to get everything completed, any further issues with data collection and data integrity could cause further delays. So I will need to be mindful and very calculated as I progress.

Data Snapshot

Back to Agenda Page

Race_Winner	Race_AVG_Lap_Time_sec	DriverNumber	Q_LapTime(s)	LapNumber	Stint	PitOutTime	PitInTime	Sector1Time	Sector2Time	Sector3Time	SpeedI1
VET	92.637931	27	112.720	1.0	NaN	262.434	NaN	42.085	28.028	42.607	226.0
VET	92.637931	27	112.720	1.0	NaN	262.434	NaN	42.085	28.028	42.607	226.0
VET	92.637931	27	85.348	2.0	1.0	NaN	NaN	28.492	23.123	33.733	276.0
VET	92.637931	27	85.348	2.0	1.0	NaN	NaN	28.492	23.123	33.733	276.0
VET	92.637931	27	116.609	3.0	1.0	NaN	NaN	40.615	30.899	45.095	141.0
...	...	...	...	...	...	...	...	...	...	...	...
VER	90.034483	63	NaN	13.0	5.0	3877.412	NaN	NaN	46.782	40.992	242.0
VER	90.034483	63	84.152	14.0	5.0	NaN	NaN	17.256	36.420	30.476	291.0
VER	90.034483	63	101.723	15.0	5.0	NaN	4181.281	20.542	42.105	39.076	242.0
VER	90.034483	63	NaN	16.0	6.0	4353.278	NaN	NaN	44.465	41.951	255.0
VER	90.034483	63	83.788	17.0	6.0	NaN	NaN	17.097	36.326	30.365	291.0

37323 rows × 152 columns



DATA PIPELINE

EXTRACTION:

- Web scraping Formula1.com and utilized FASTF1 API to collect data for historical data collection.
- Mage pipeline will handle upcoming race event for data collection.

TRANSFORM & LOAD:

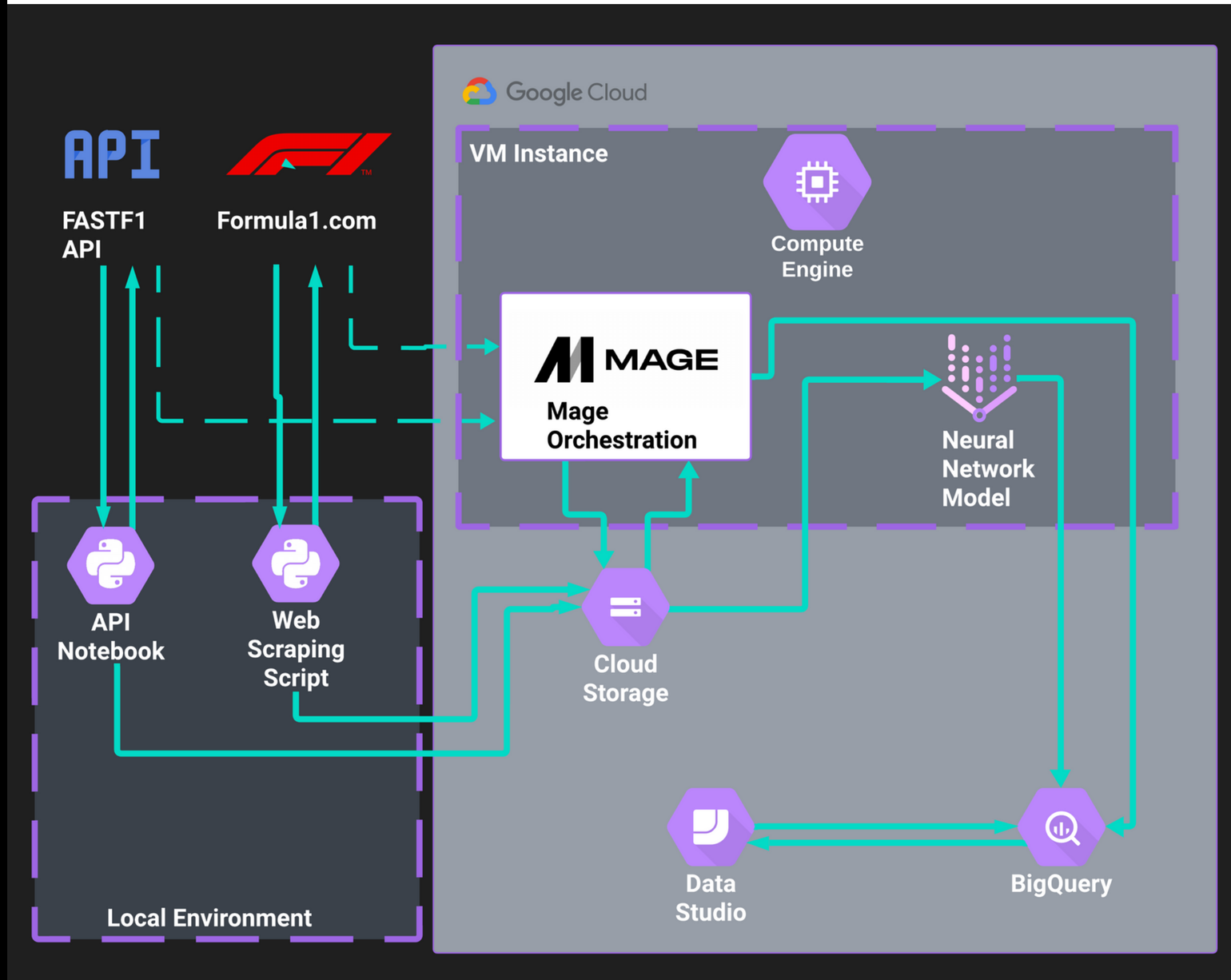
- Scripts will load raw dataframes into GCS bucket and process data for modeling and load that table to a GCS bucket as well.

MODELING:

- Run neural network modeling in Compute Engine VM.

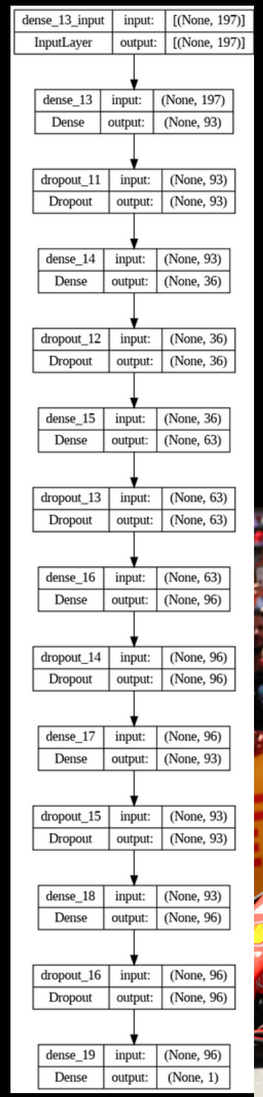
DATA VISUALIZATION:

- After modeling, write prediction data to BigQuery data warehouse and connect a Data Studio dashboard for visualizations.



NEURAL NETWORK MODEL

	Event_id	Driver	Predicted Avg Lap Time (sec)	Race_Winner	Race_AVG_Lap_Time_sec
6	979	HAR	90.631149	VET	92.637931
2	979	ERI	90.989670	VET	92.637931
15	979	SIR	91.392822	VET	92.637931
7	979	HUL	91.529671	VET	92.637931
4	979	GRO	91.622879	VET	92.637931
3	979	GAS	91.700531	VET	92.637931
9	979	MAG	91.761978	VET	92.637931
17	979	VAN	91.801529	VET	92.637931
19	979	VET	92.021461	VET	92.637931
14	979	SAI	92.101997	VET	92.637931
5	979	HAM	92.149628	VET	92.637931
1	979	BOT	92.159927	VET	92.637931
16	979	STR	92.229454	VET	92.637931
12	979	RAI	92.234894	VET	92.637931
18	979	VER	92.248245	VET	92.637931
13	979	RIC	92.268265	VET	92.637931
0	979	ALO	92.317322	VET	92.637931
11	979	PER	92.346344	VET	92.637931
8	979	LEC	92.359962	VET	92.637931
10	979	OCO	92.367645	VET	92.637931



Model Predictions

The model does a good job of predicting AVG Lap time for the Race but some improvements still need to be worked on, next steps are to get better at predicting overall driver placement results.

Thank You