Forecasting Outcomes in Formula 1 Racing

**Ali Arslanhan, Ryan Bausback,
Edward Voskanian**

Erdos Institute
Data Science Boot Camp, Summer 2024

Motivation

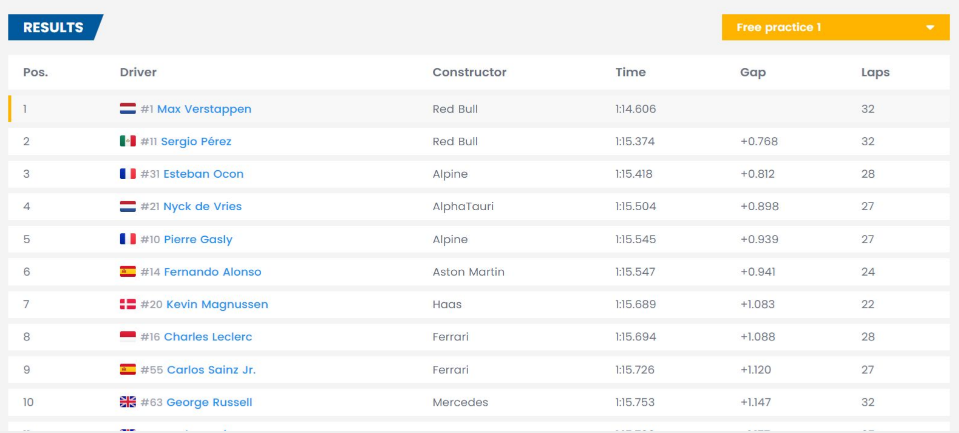
- A typical Formula 1 season calendar features a number of Grand Prix events, each with
  - three free practice sessions (Friday and Saturday),
  - three qualifying sessions (Saturday) to determine the starting grid on race day, and
  - the main race (Sunday)

- Research Questions
  - Explanatory Modeling
    - Which factors from practice sessions during a given Grand Prix have the most significant impact on the subsequent qualifying round results?
  - Predictive Modeling
    - By utilizing data from the practice sessions, what level of accuracy can be achieved in predicting the qualifying times and grid positions.

Dataset

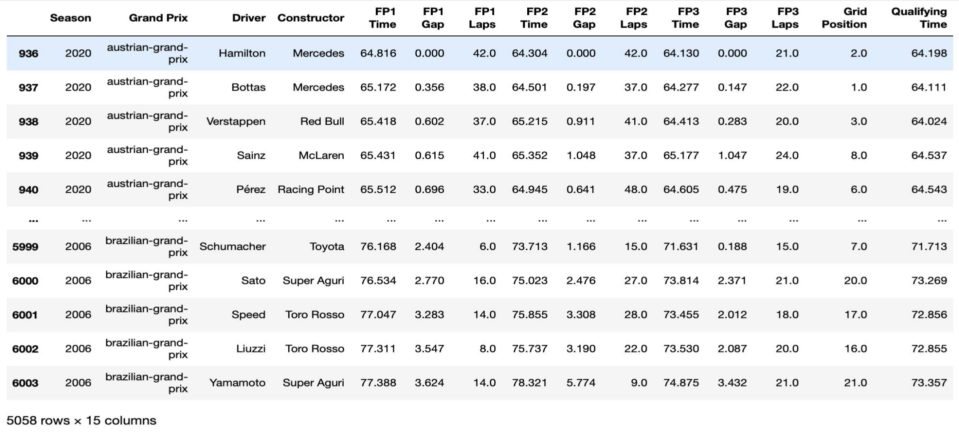
- Practice and qualifying race data for 2003-2023 seasons
  - Source: pitwall.app Formula 1 Database
- Train-Test-Split
  - Training: 2003-2020 seasons
  - Testing: 2021-2023 seasons
  - 85% training, 15% testing
- Variables/Feature
  - # of practice laps (x3)
  - Fastest lap time in practice (x3)
  - Gap to fastest lap time (x3)
- Outcomes
  - Grid position
  - Q1 time

Pitwall Database



Pos.	Driver	Constructor	Time	Gap	Laps
1	#1 Max Verstappen	Red Bull	1:14.606		32
2	#11 Sergio Pérez	Red Bull	1:15.374	+0.768	32
3	#31 Esteban Ocon	Alpine	1:15.418	+0.812	28
4	#21 Nyck de Vries	AlphaTauri	1:15.504	+0.898	27
5	#10 Pierre Gasly	Alpine	1:15.545	+0.939	27
6	#14 Fernando Alonso	Aston Martin	1:15.547	+0.941	24
7	#20 Kevin Magnussen	Haas	1:15.689	+1.083	22
8	#16 Charles Leclerc	Ferrari	1:15.694	+1.088	28
9	#55 Carlos Sainz Jr.	Ferrari	1:15.726	+1.120	27
10	#63 George Russell	Mercedes	1:15.753	+1.147	32

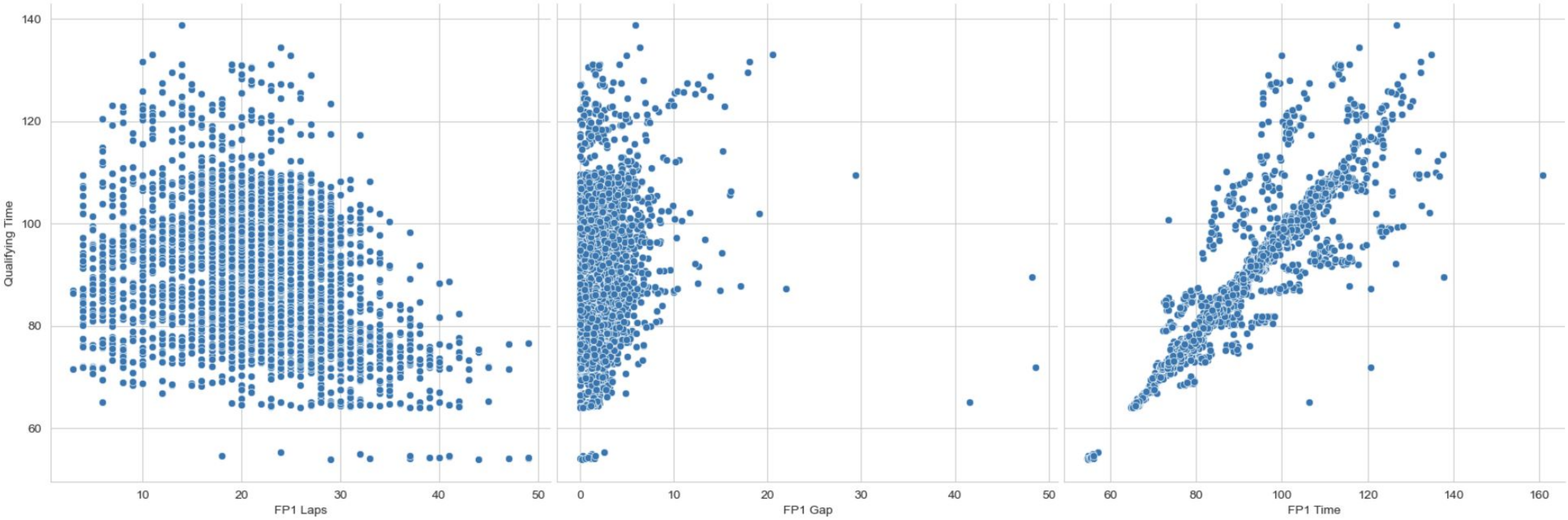
Pandas Dataframe



	Season	Grand Prix	Driver	Constructor	FP1 Time	FP1 Gap	FP1 Laps	FP2 Time	FP2 Gap	FP2 Laps	FP3 Time	FP3 Gap	FP3 Laps	Grid Position	Qualifying Time
936	2020	austrian-grand-prix	Hamilton	Mercedes	64.816	0.000	42.0	64.304	0.000	42.0	64.130	0.000	21.0	2.0	64.198
937	2020	austrian-grand-prix	Bottas	Mercedes	65.172	0.356	38.0	64.501	0.197	37.0	64.277	0.147	22.0	1.0	64.111
938	2020	austrian-grand-prix	Verstappen	Red Bull	65.418	0.602	37.0	65.215	0.911	41.0	64.413	0.283	20.0	3.0	64.024
939	2020	austrian-grand-prix	Sainz	McLaren	65.431	0.615	41.0	65.352	1.048	37.0	65.177	1.047	24.0	8.0	64.537
940	2020	austrian-grand-prix	Pérez	Racing Point	65.512	0.696	33.0	64.945	0.641	48.0	64.605	0.475	19.0	6.0	64.543
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5999	2006	brazilian-grand-prix	Schumacher	Toyota	76.168	2.404	6.0	73.713	1.166	15.0	71.631	0.188	15.0	7.0	71.713
6000	2006	brazilian-grand-prix	Sato	Super Aguri	76.534	2.770	16.0	75.023	2.476	27.0	73.814	2.371	21.0	20.0	73.269
6001	2006	brazilian-grand-prix	Speed	Toro Rosso	77.047	3.283	14.0	75.855	3.308	28.0	73.455	2.012	18.0	17.0	72.856
6002	2006	brazilian-grand-prix	Liuzzi	Toro Rosso	77.311	3.547	8.0	75.737	3.190	22.0	73.530	2.087	20.0	16.0	72.855
6003	2006	brazilian-grand-prix	Yamamoto	Super Aguri	77.388	3.624	14.0	78.321	5.774	9.0	74.875	3.432	21.0	21.0	73.357

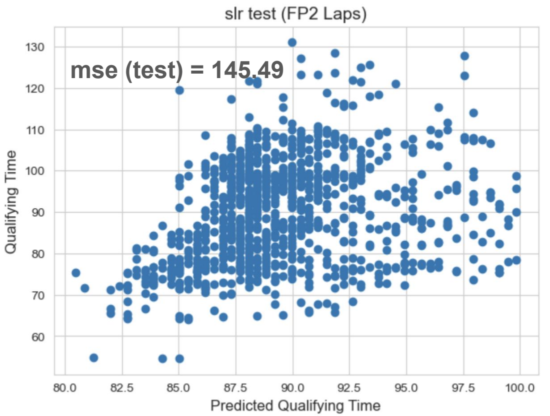
5058 rows × 15 columns

Data Visualization

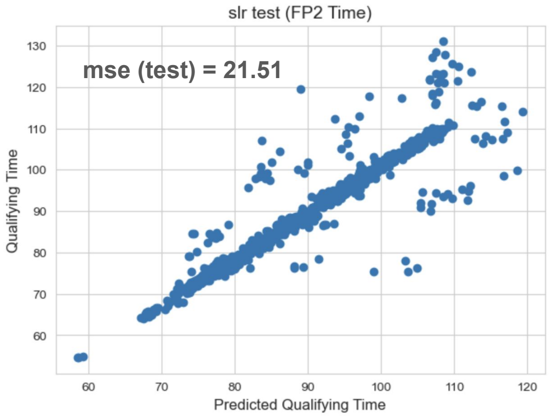


Linear Regression and KNN (Predicting Q1 Times)

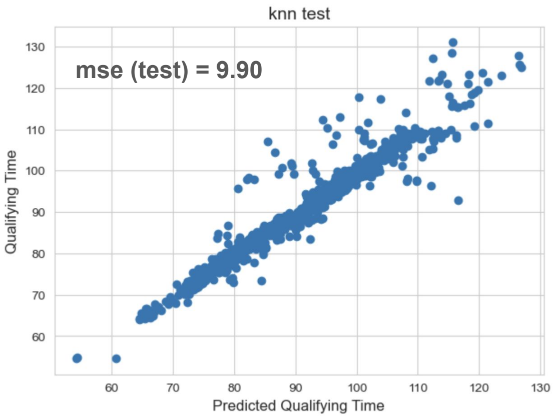
Five-fold cross-validation (train_test_split 20% testing) using mean squared error


slr test (FP2 Laps)

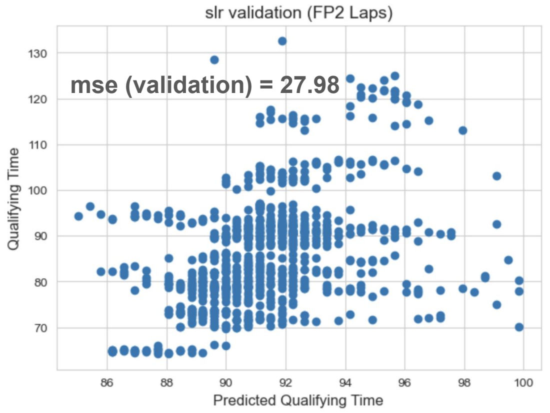
mse (test) = 145.49


slr test (FP2 Time)

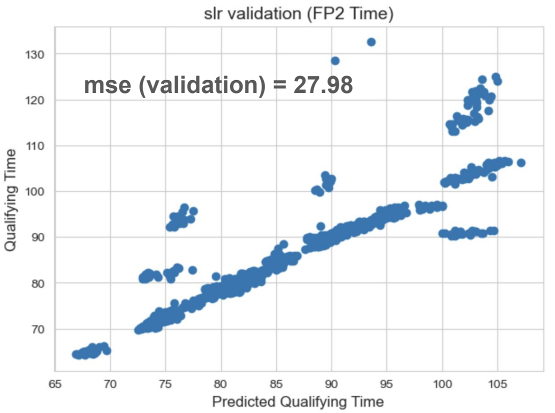
mse (test) = 21.51


knn test

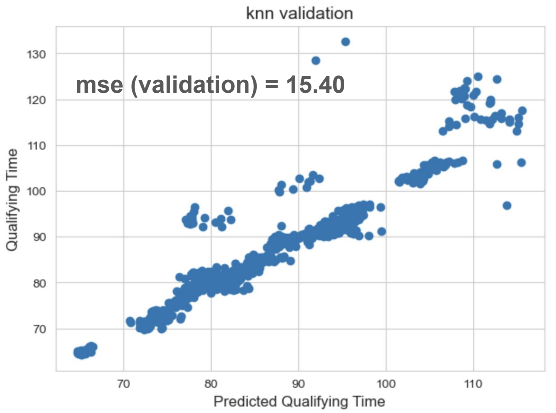
mse (test) = 9.90


slr validation (FP2 Laps)

mse (validation) = 27.98


slr validation (FP2 Time)

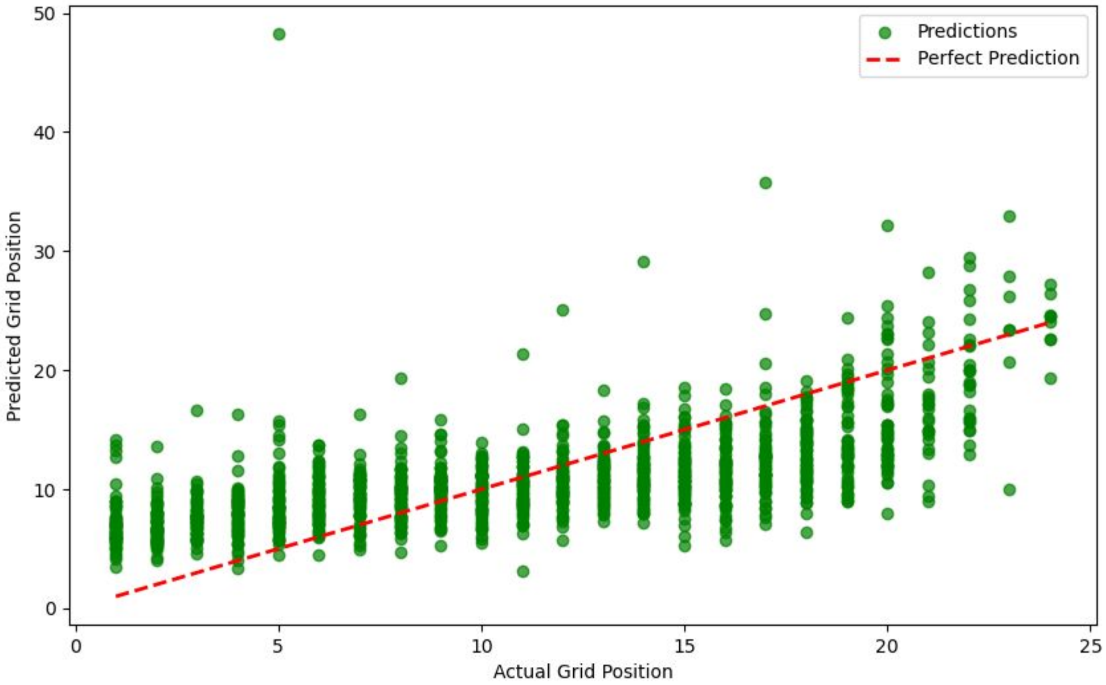
mse (validation) = 27.98


knn validation

mse (validation) = 15.40

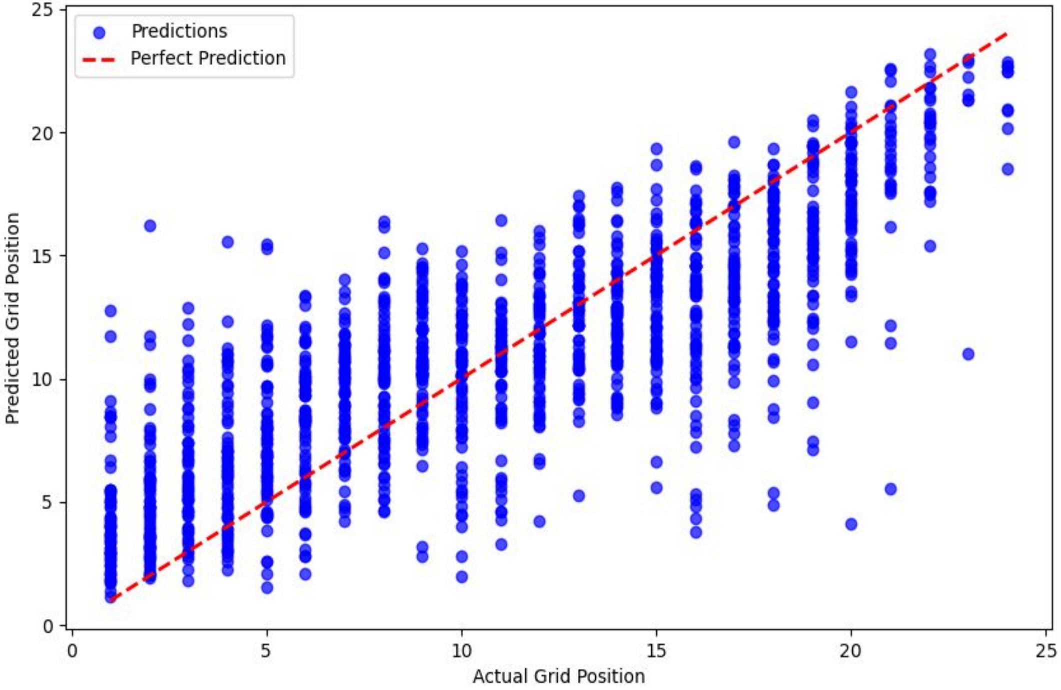
Linear Regression (Predicting Grid Position)

- Objective: Predict starting grid positions based on practice session data.
- Features: FP1, FP2, FP3 times, gaps to fastest lap, number of laps.
- Performance: MSE: 21.46, R²: 0.427



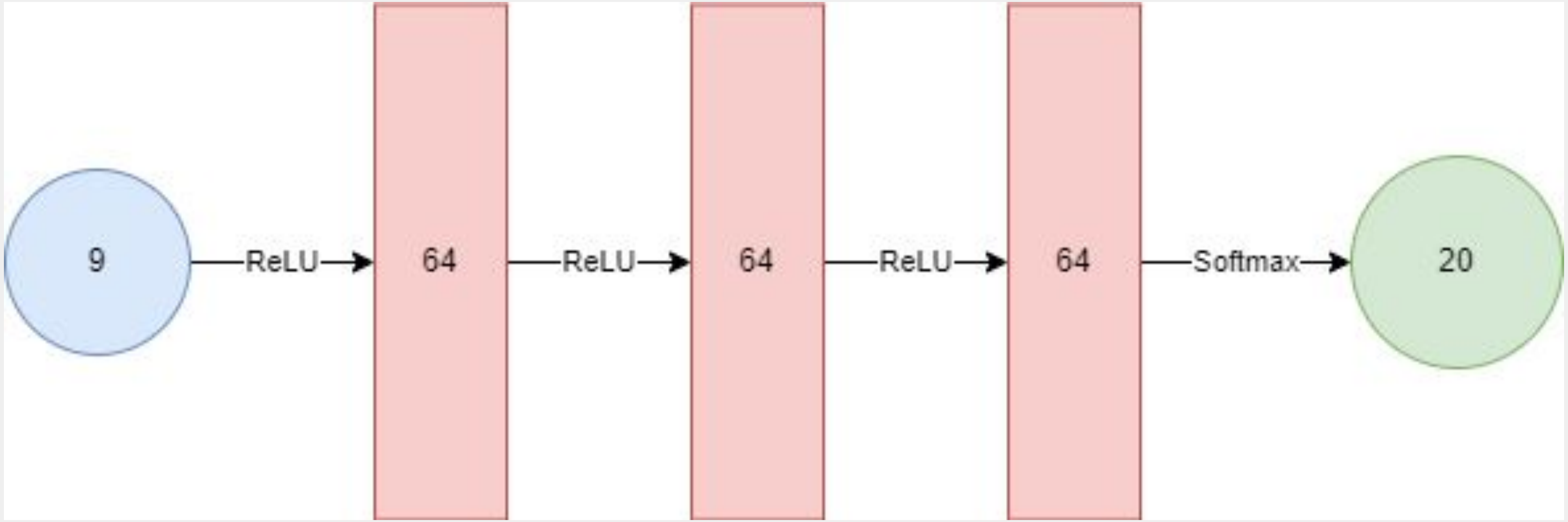
Random Forest (Predicting Grid Position)

- Objective: Predict starting grid positions based on practice session data.
- Features: FP1, FP2, FP3 times, gaps to fastest lap, number of laps.
- Performance: MSE:13.45, R²: 0.64

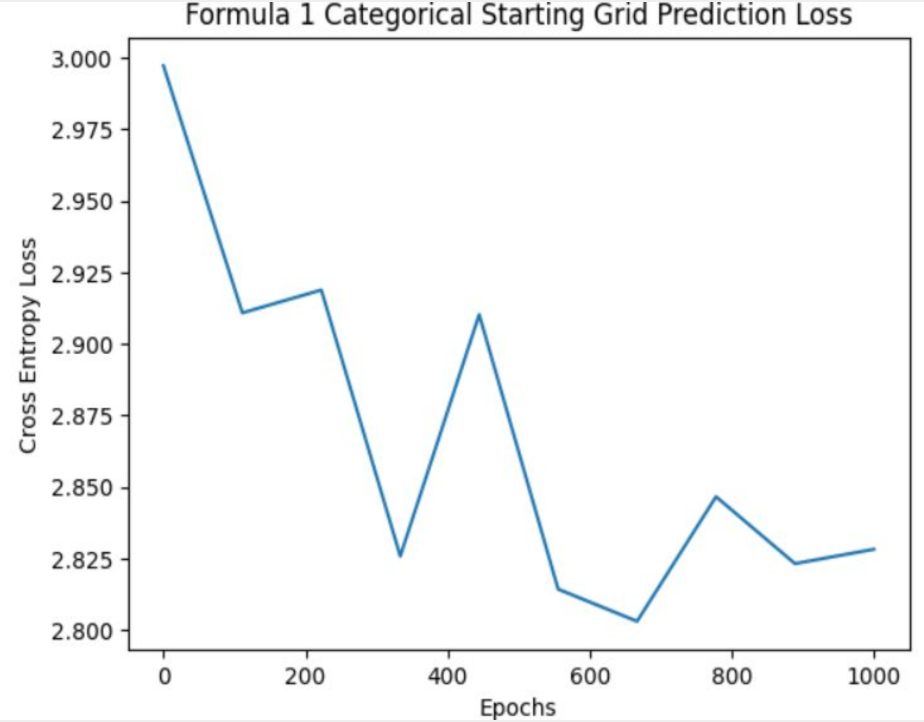


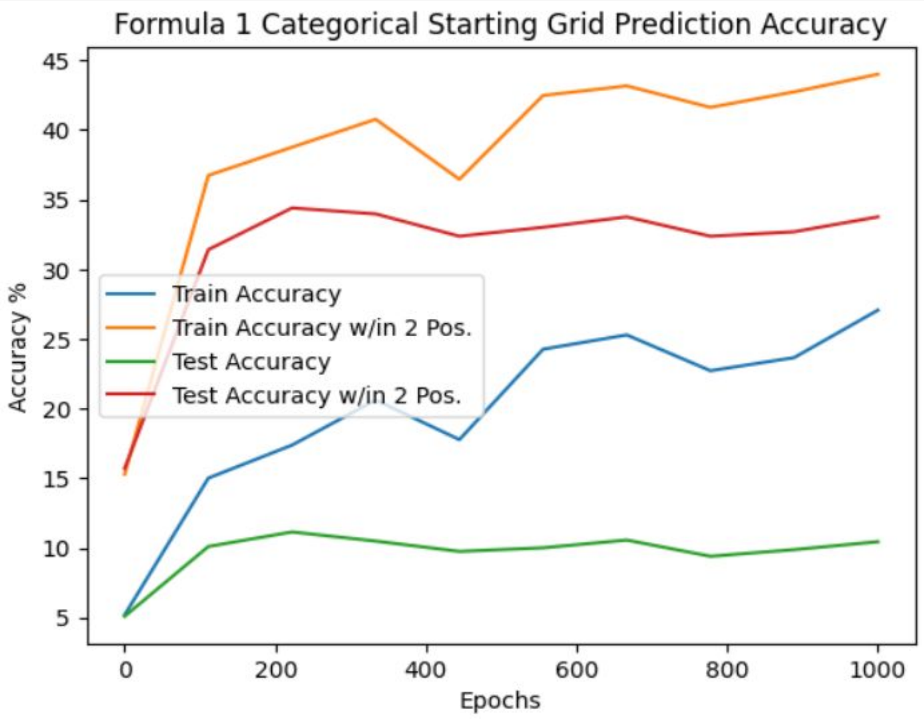
Dense Neural Network (Predicting Grid Position)

- Objective: Predict starting grid position as categorical variable
- Learning Rate = 0.001
- Loss: Cross Entropy



Dense Neural Network Performance





Conclusion

- Predicting Formula 1 starting grids is difficult.
- Potential Improvements:
  - Additional data (ex. Weather conditions: practice v. qualifying, tires, etc.)
  - Construct additional relative features (ex. Grid position in previous race)
  - Predict either race grid as single example?