CS229 Final Project: Formula One Race Prediction with Deep Learning

Isaac Kleisle-Murphy (ikleisle@stanford.edu)

Project Overview

In this project, I explore how three popular machine learning techniques – multinomial logistic regression, feed-forward neural networks, and convolutional neural networks – can be leveraged to predict Formula One ("F1") race results. Using historical race and timing data, I find that the feed-forward neural networks perform best in predicting a driver's eventual finishing position at the outset of a race.

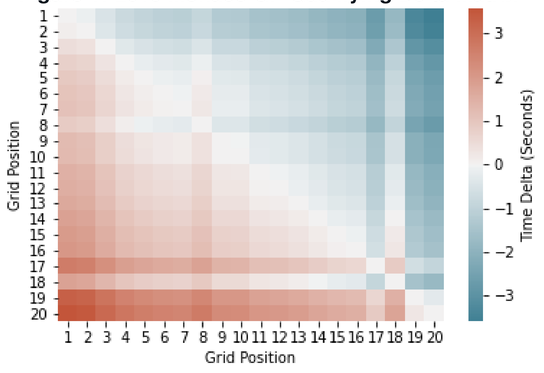
Goal: To predict a driver's race result – first, second, third, fourth, . . . , tenth, or eleventh+ – at the outset of a Grand Prix.

Data Ingest and Preprocessing

Using 2006-2020 race data obtained via the Ergast F1 API, I engineered three principal types of features likely to be predictive of race performance, namely:

1. **Historical Driver/Team Box Score Statistics:** *driver and team-specific win, podium, and points-scoring rates; driver-specific crash rates; team-specific mechanical failure rates; driver/team positions in the final standings in previous seasons; driver experience at circuit; driver career starts.*
2. **Circuit Features:** *circuit-specific indicator variables; circuit-specific crash/retirement rates; start-finish correlation in previous races (a proxy for overtaking difficulty); indicators for wet/dry track conditions.*
3. **Inter-Driver Performance Deltas ("Images"):** *race qualifying time deltas between starting grid positions; seasonal qualifying time deltas between grid positions; season top speed differentials between grid positions.*

Figure 1: 2019 Monaco GP Qualifying Delta Matrix



Above, each entry reflects an ordered pair of starting positions, as colored by the qualifying delta between that pair grid positions (delta = row time - column time). A larger delta (i.e. color intensity) reflects a greater gap in speed/performance between two drivers.

Models Considered

I considered three models and a baseline model, as set forth below. Data (spanning 2006-2020) was split into a training set, consisting of 246 races from 2006-2018, and a test set, consisting of 31 races from 2019-2020. With the exception of the naive classifier, these models were tuned over a grid of relevant hyperparameters using five-fold, single-repeat, randomized cross validation within the training set. For each model, the final hyperparameters were selected based on average cross entropy over the validation folds.

1: Naive (Baseline)

For training races $i = 1, \ldots, n$, starting grid positions $j = 1, \ldots, J$, and finishing position $k = 1, 2, 3, \ldots, 10, 11^+$, calculate sample probability vector:

$$\left\langle \sum_{i=1}^{i} \frac{\delta(\mathsf{S}=j, \mathsf{F}=1)}{n}, \ldots, \sum_{i=1}^{n} \frac{\delta(\mathsf{S}=j, \mathsf{F}=11^+)}{n} \right\rangle$$

This model, which returned simple historical frequencies, served as the baseline for performance.

2: Multinomial Logistic (L2)

For race i, starting grid position $j = 1, \ldots, J$, corresponding feature vector $x^{(i,j)}$, and corresponding finishing position $F_{i,j} \in \{1, \ldots, 10, 11^+\}$, model

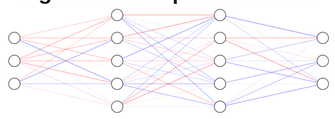
$$P(F_{i,j} = k) = \frac{\exp(\beta_k^T x^{(i,j)})}{\sum_{\ell=1}^{K} \exp(\beta_\ell^T x^{(i,j)})}.$$

with $\lambda \sum_k ||\beta_k||_2^2$ included in the MLLE maximisation.

3: Feed-Forward Neural Net

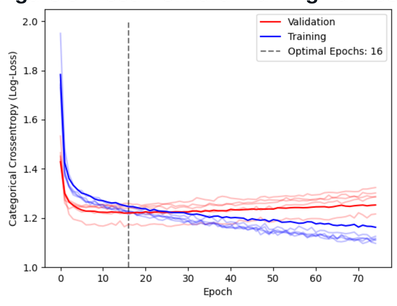
Identical feature and target structure to Model 2; however, model is now a directed/layered structuring of nodes, weights, biases, and activation functions.

Figure 2: Example Neural Net



Tuned hyperparameters included: dropout rate (.1), hidden layers (4), nodes/layer (15), activation (σ), learning rate (.001), and epochs (16). Batch normalization between layers.

Figure 3: Feed-Forward Training/Val Loss



4: Convolutional Neural Net

This model has two input streams. First, the delta matrices ("images") for race i are fed through two max-pooled, batch-normalized convolutional layers, before being flattened and merged with the second input (consisting of the remaining features for race i). The model then outputs J probability vectors (one for each driver in race i) for the $K = 11$ categories of finishing positions.

Figure 4: Example CNN

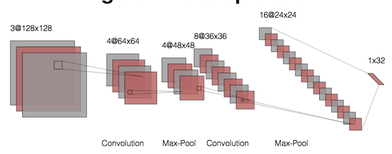
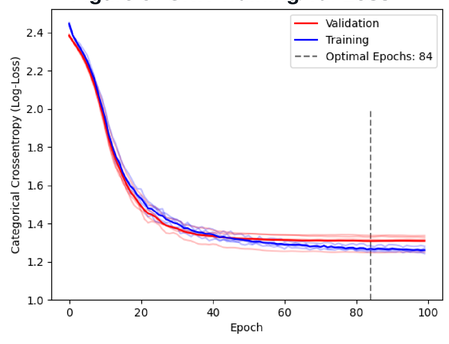


Figure 5: CNN Training/Val Loss



Findings

Cross entropy was my principal scoring measure, with predicted vs. observed Spearman correlation and categorical accuracy as secondary measures:

Table I: Test Results (2019-20 Races)

Model	Avg. Val. Cross Entropy	Cross Entropy	Semi-Spearman	Acc.
Naive	N/A	1.2182	0.6000	0.5450
MLR	1.2536	1.1954	0.6178	0.5550
FFNN	1.2201	1.1909	0.6042	0.5417
CNN	1.3084	1.2439	0.5935	0.5433

As evidenced by the cross entropy and Spearman correlations, the multinomial logistic regression and feed-forward neural network performed best. In my view, this makes sense: during the testing period, one team and driver (Mercedes/Lewis Hamilton) increasingly dominated the sport, with three distinct tiers of teams following. The simpler models could readily capture this. Further, with only 246 training examples, the dataset may have been too small for the CNN to generalize well.

Figure 6: L. Hamilton Win Probabilities (2019-20)

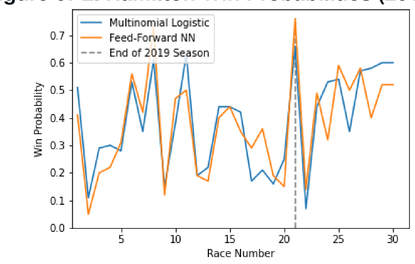
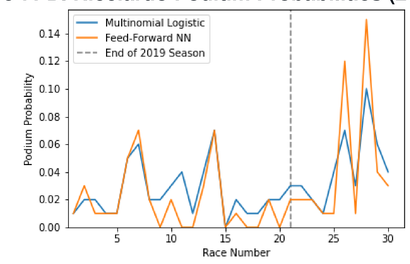


Figure 7: D. Ricciardo Podium Probabilities (2019-20)



Future research would likely entail: additional feature engineering, examination of time-dependent algorithms (i.e. LSTM), and/or ensembling.

Select References/Works Consulted

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)*. Antalya: IEEE, Aug. 2017, pp. 1–6. ISBN: 9781538619490. DOI: 10.1109/ICEngTechnol.2017.8308186. URL: https://ieeexplore.ieee.org/document/8308186/ (visited on 11/16/2020).

[2] *CS231: Convolutional Neural Networks for Visual Recognition*. publisher: Stanford University. URL: https://cs231n.github.io/convolutional-networks/#conv (visited on 11/16/2020).

[3] *Ergast Developer API*. en-US. URL: https://ergast.com/mrd/ (visited on 11/16/2020).