Business Impact

The project shows the transformation of historic F1 race data into usable predictive systems which calculate the likelihood that a driver will finish Top 10 before the start of that race. This type of model serves a multitude of constituents from a business/strategic viewpoint. For example, teams and race analysts receive a data-based reference expectation regarding the likelihood of scoring points, enabling them to perform scenario analyses such as evaluating risk associated with qualifying setup decisions, tire selection trade-offs, or evaluating reliability management. TV/Media also use the probabilistic forecast to enhance the quality of their storytelling through determining competitive flow and providing quantitative evidence of trending in momentum. For fans participating in engagement platforms (including but not limited to fantasy sports, prediction markets, betting analytics), validating the precision of pre-event probability will provide a transparent and logical basis for making decisions. More generally, establishing a reproducible framework through which to assess longitudinal performance (over time) across drivers, constructors and eras resides at the core of this system. By bringing together short-term performance (via rolling periods), long-term career performance, constructor performance, track characteristics and competitive situations, this model provides actionable and interpretable probability estimates by modelling all historical interactions between these data elements.

Assumptions in the Analysis

This analysis is based on some key assumptions. The data used represents all engineered features of the race that were readily available prior to the start of the race. Rolling and expanding statistics, such as Top-10 rate over each of the last five races and constructor performance throughout the season, were constructed using historically shifted windows to prevent leakage from current race results. The binary variable “Top-10 Finish” is used as a proxy for race success across the eras, despite historical changes in points systems and grid size. In addition, qualifying information (if available) has been assumed to be reasonably approximated by grid position when it is not available, and the median value will provide a minimal level of bias after fallback has taken place. The assumption is made that tree-based models will be able to correctly identify entity differences and not create fictional ordinal relationships when using label encoding for categorical identifiers (drivers, constructors, circuits). Lastly, the temporal split between the data set (training: <2018, validation: 2019-2021, test: 2022-2024) assumes that seasons after 2018 will represent an appropriate modern day evaluation regime and data prior to 2018 will be adequate to provide a long term structure of the patterns.

Limitations

The analysis has a number of structural limitations even though there are strong indicators of prediction. An inherent instability characterizes Formula One, as the relative competitiveness and balance among teams have changed over many years due to factors such as regulations, developments in car design, reliability advances, cap on team budgets, and changes to aerodynamic rules. Therefore, correlations that were developed during earlier decades have been proven not to always apply to the current year or decades further into the future than they did previously. In addition, the analysis lacks critical contextual variables that have a significant impact on race results (e.g., weather, frequency of safety cars, info on pit strategy, penalty information, choice of tire compounds, use of engine components, and upgrades by teams). Another limitation is that due to missing qualifying data within the earlier years , the data had to be estimated creating the potential to decrease signal quality. Further, multiple variables, such as grid position, percentile rank, and derived binary flags, have exhibited considerable multicollinearity, leading to inflated importance measures, as well as making the interpretation of these variables difficult in linear models. Lastly, there is residual dependence (i.e. repeated driver by season) even though time-based splits reduce the potential for leaking, thus limiting the generalisability of the model unless observed closely.

Lessons Learned

A key takeaway from this study is the importance of controlling for leakage in sports prediction modelling. Features that are specifically included in finishing results or other race results such as finishing position, number of points awarded and number of laps completed could lead to falsely perfect predictive ability of the model if these features are not excluded from pre-race models. Temporal order by strict time or distance and careful shifting of features are critical for preserving the validity of the predictive models. Another important finding is that the predictive ability for formula 1 to have drivers on the grid is highly concentrated across only 3 structural dimensions (grid position and contractor ability, recent driver performance). Volatility between races at one track also plays an important role in determining pre-race predictions between grids(s) (grid position), as well as using other parts of data where there is no indicator to see that there may be different levels of volatility (DNFs). It was also found that combined short term rolling metrics with long term expanding career statistics produced a higher level of stability than using only static indices for pre-race predictions. Finally, this study demonstrates the use of a great deal of feature engineering in sports models helps make better use of available data for long term predictions. However, this increase in volume of data used requires an increased need for rigorous verification, calibration and interpretability protocols.

Next steps

Future projects should enhance three primary areas: robustness and interpretability of adopted techniques, and their application in real-world scenarios. First, the training of as a final

model should impose an unyielding 'pre-race only' feature set on the underlying data. Second, by implementing hyperparameter optimization as well as probability calibration techniques (e.g., Platt Scaling and Isotonic Regression), the reliability of forecasts will be increased. Third, future evaluation studies should include cross-era review of model performance degradation in response to regulatory changes; this will arguably promote the use of epoch specific models or time sensitive weighting methodologies. Fourth, external data (e.g., weather conditions, penalty record occurrences, safety car events and tire strategy) can add significant relevant content to improve predictive accuracy. Lastly, the model with the highest gross return on investment should be converted into an inference pipeline (either as a batch or an API), complete with mechanisms for detecting Parkinson's disease and automatic guideline based retraining at seasonally set intervals.

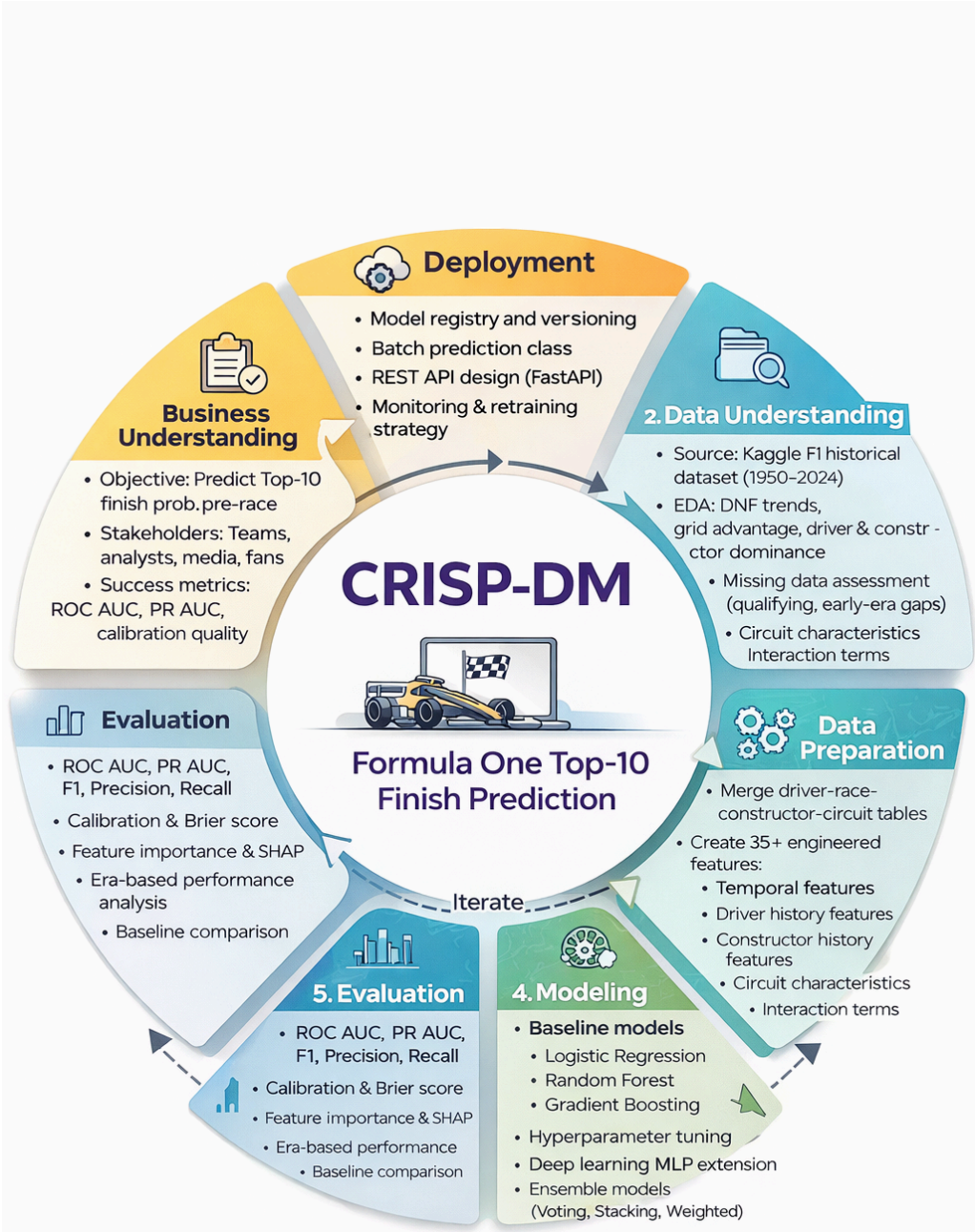
CRISP-DM Diagram

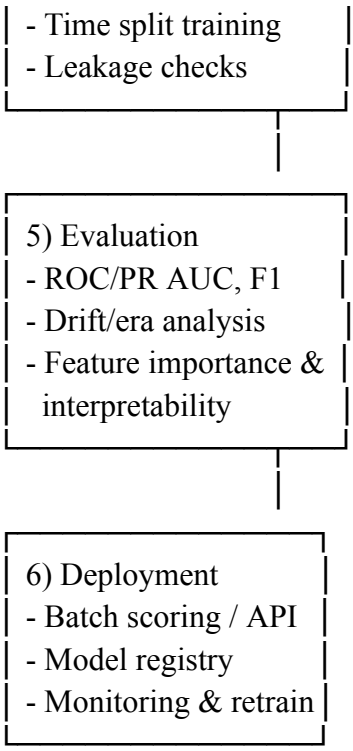
1) Business Understanding
- Predict Top-10 finish pre-race

2) Data Understanding
- Kaggle F1 tables
- EDA: trends, DNF, grid, constructors

3) Data Preparation
- Merge to driver-race
- Handle missing qual
- Engineer pre-race history/context feats
- Encode categoricals

4) Modeling
- Baselines (LR/RF/GB)




- Time split training
- Leakage checks
5) Evaluation
- ROC/PR AUC, F1
- Drift/era analysis
- Feature importance &
interpretability
6) Deployment
- Batch scoring / API
- Model registry
- Monitoring & retrain