F-prediction
predição do ganhador da formula 1

Vitor H. O. Silva, Fabrício A. Silva

[1]Instituto de Ciências Exatas e Tecnológicas – Universidade Federal de Viçosa (UFV) – 35690-000 – Florestal – MG – Brasil

fabricio.asilva@ufv.br, vitor.h.oliveira@ufv.br

Abstract. *This article presents a project that employs machine learning to predict the probability of success of Formula 1 drivers in future races, both before and after the qualifying round. We explore relevant aspects of related work, data collection and feature engineering. We selected and validated the most appropriate algorithms, and presented the results obtained. This study provides significant insights for performance predictions in Formula 1, and strategies to maintain the model's relevance in the constant dynamics of the sport.*

Resumo. *Este artigo apresenta um projeto que emprega a aprendizagem de máquina para prever a probabilidade de sucesso de pilotos de Fórmula 1 em corridas futuras, tanto antes como após a etapa de qualificação. Exploramos os aspectos relevantes dos trabalhos correlatos, a coleta de dados e a engenharia de atributos. Selecionamos e validamos os algoritmos mais adequados, e apresentamos os resultados obtidos. Este estudo proporciona insights significativos para previsões de desempenho na Fórmula 1, e estratégias para manter a relevância do modelo na dinâmica constante do esporte.*

1. Introdução

O foco deste trabalho é desenvolver e apresentar um projeto que objetiva realizar a predição de probabilidade de sucesso para os pilotos na próxima corrida de Fórmula 1, tanto antes quanto após a etapa de qualificação.

Iniciaremos com uma discussão resumida sobre **pesquisas e estudos correlatos** no campo da predição de resultados em corridas de Fórmula 1. A seguir, abordaremos o **processo de busca e extração de dados**, detalhando as fontes utilizadas, os critérios de seleção de dados e como foram organizados para análise.

Na sequência, será explorado a etapa de **engenharia de atributos**, descrevendo as variáveis consideradas, os processos de transformação e a lógica subjacente à construção dos atributos utilizados no modelo de predição. Seguiremos para a **seleção e validação dos algoritmos**, explicando a metodologia aplicada na escolha do(s) algoritmo(s) e como foi realizado o processo de validação.

Após a discutirmos sobre da metodologia, será apresentado os **resultados obtidos**, comparando-os com as previsões realizadas e discutindo a eficácia e pontos fracos dos modelos. Finalmente, discutiremos a **disponibilidade pública** e as estratégias para a **atualização dos modelos**, a fim de garantir que continuem a ser úteis e relevantes no contexto em constante mudança das corridas de Fórmula 1.

2. Trabalhos relacionados

Nesta seção, discutiremos alguns trabalhos relacionados que usam tecnologia de dados e machine learning para analisar e prever resultados na Fórmula 1.

A AWS é um parceiro tecnológico da Fórmula 1, fornecendo soluções de nuvem e machine learning para gerenciar e analisar os dados do esporte. Uma das inovações resultantes dessa colaboração é o F1 Insights [Russo 2003], uma série de análises que ajudam os fãs a entender as complexidades do esporte. Estas análises fornecem informações baseadas em dados sobre a performance dos pilotos e das equipes, e destacam decisões que ocorrem em frações de segundo durante as corridas. A AWS também contribuiu para o gerenciamento de ativos de mídia da Fórmula 1, e suas simulações foram usadas para projetar os novos carros para a temporada de corridas de 2022/2023.

Um projeto semelhante buscou prever o curso da temporada de Fórmula 1 de 2020 usando machine learning [George 2021]. Este projeto coletou dados da Ergast API, que contém informações sobre corridas, resultados, pilotos, tempos de qualificação, pit stops, e posições de construtores e pilotos desde o início da Fórmula 1 em 1950. Os dados foram analisados para encontrar correlações e tendências úteis que poderiam informar o modelo de machine learning. Entre as descobertas, foi observado que a posição de qualificação tem uma correlação mais forte com a posição de chegada do que a posição de largada. O projeto também destacou a importância da idade dos pilotos e a imprevisibilidade inerente da Fórmula 1, com variáveis imprevistas como doenças, acidentes, falhas mecânicas e problemas nos boxes muitas vezes afetando os resultados previstos.

Um projeto similar, intitulado "F1 Champ: EDA + Classification (100% accuracy)" [Ganapathi 2023] foi publicado na plataforma Kaggle. O autor afirma ter alcançado 100% de precisão em suas previsões. No entanto, uma análise mais detalhada revela várias falhas na engenharia e na criação do modelo, particularmente no uso inadequado dos dados de treinamento como dados de validação.

Outro trabalho que apresenta semelhanças com o nosso é a tese de bacharelado intitulada "Machine Learning Framework for Formula 1 Race Winner and Championship Standings Predictor" [SICOIE]. Este trabalho utiliza a mesma API Ergast que usamos em nosso estudo. Embora o trabalho de Sicoie não especifique claramente as características usadas para construir o modelo, nosso trabalho se diferencia ao incluir métricas de performance da temporada atual do piloto e ao criar um novo modelo com a característica da grid. Além disso, enquanto nosso projeto utiliza os algoritmos MLP e XGBoost, Sicoie aplica três algoritmos diferentes: Random Forest Regressor, Gradient Boosting Regressor e Support Vector Regressor. Os resultados obtidos por Sicoie serão usados como referência para nosso estudo.

Model	Spearman ρ	Pearson r	R squared	MSE	rMSE
RFR	0.902	0.880	0.616	4163.32	64.524
GBR	0.903	0.906	0.589	3166.80	56.274
SVR	0.883	0.917	0.630	2778.27	52.709

Figure 1. Resultados Sicoie

Esses trabalhos demonstram a aplicação e o potencial da tecnologia de dados e do machine learning para analisar e prever resultados na Fórmula 1. No entanto, também

destacam os desafios inerentes a este domínio, incluindo a complexidade e a imprevisibilidade dos dados da Fórmula 1. Estes são fatores importantes a serem levados em consideração ao desenvolver e avaliar modelos de previsão para a Fórmula 1.

3. Dados Utilizados

3.1. Formula 1 World Championship (1950 - 2023) - Kaggle

A fonte principal de dados foi o conjunto de dados "Formula 1 World Championship (1950 - 2023)" disponível no Kaggle [Vopani 2023]. Este conjunto de dados organiza as informações disponíveis na Ergast Developer API em tabelas CSV, facilitando a preparação e análise dos dados. O conjunto de dados inclui 14 tabelas diferentes: circuits.csv, constructor_results.csv, constructor_standings.csv, constructors.csv, driver_standings.csv, drivers.csv, lap_times.csv, pit_stops.csv, qualifying.csv, races.csv, results.csv, seasons.csv, sprint_results.csv e status.csv. As tabelas que terminam com "standings" representam as classificações da temporada, enquanto as tabelas que terminam com "results" representam os resultados de uma corrida específica.

A tabela "results.csv" foi a principal tabela utilizada, pois ela fornece os resultados de cada piloto para cada corrida desde 1950 até 2023. A partir desta tabela, foram realizados merges com as demais tabelas para agregar informações adicionais relevantes para cada corrida. No entanto, alguns dados de 2023 estavam ausentes no conjunto de dados do Kaggle, por isso foi criada uma rotina para atualizar os dados obtendo os resultados mais recentes diretamente da Ergast Developer API.

3.2. Ergast Developer API

A Ergast Developer API[**?**] é um serviço de web experimental que fornece um registro histórico de dados de corridas de motor para fins não comerciais. A API fornece dados para a série de Fórmula 1, desde o início dos campeonatos mundiais em 1950. A API é organizada de forma que as consultas são feitas através de um GET request, especificando a série, a temporada e a rodada que se deseja consultar. A API fornece informações sobre horários das temporadas, resultados de corridas, resultados de qualificação, classificações, informações de pilotos, informações de construtores, informações de circuitos, status de finalização, tempos de volta e pit stops. A API suporta formatos de resposta em XML, JSON e JSONP, e o número de resultados que são retornados pode ser controlado usando um parâmetro de consulta 'limit', até um valor máximo de 1000.

3.3. Open-Meteo

A API da Open-Meteo foi utilizada para coletar dados climáticos da localização de cada corrida, incluindo a umidade e as condições climáticas (seco, chuvoso, nevando, empoeirado). Esta API retorna dados hora a hora para cada localização geográfica especificada. Como uma corrida de Fórmula 1 dura em média 3 horas, os dados da hora de início da corrida e das três horas seguintes foram agregados.

Em resumo, as três fontes de dados foram integradas de maneira a obter uma visão ampla e detalhada de cada corrida, incluindo informações sobre os pilotos, as condições do circuito e o clima, o que tornou possível realizar análises estatísticas e preditivas abrangentes.

4. Engenharia de Características e Seleção de Algoritmos

A engenharia de características e a seleção de algoritmos são partes fundamentais no desenvolvimento de um modelo de aprendizado de máquina.

4.1. Engenharia de Características

A engenharia de características iniciou com a organização dos dados. A primeira transformação realizada foi um *shift* nos resultados da temporada, proporcionando uma visão do estado atual do piloto na temporada. Esta perspectiva se baseou em três características: `points_season`, `position_season` e `wins_season`.

Em seguida, a data de nascimento do piloto foi transformada em anos, e foram coletados dados climáticos para cada corrida, resultando nas características `humidity`, `temperature` e `weather_condition`. Foi realizada também uma soma cumulativa da média da posição de Grid e da posição final do piloto, considerando todas as corridas já realizadas por cada piloto, gerando as características `AvgGrid`, `AvgFn` e `wins_acc`.

Após essa etapa, foi realizado um *parse* de todas as nacionalidades, países dos construtores e países dos circuitos, para um mesmo *encoder*, com a intenção de encontrar possíveis correlações. Ao final, as colunas nulas foram removidas, resultando em um total de 18 colunas, ou 19 quando o grid está presente. Estas características são:

- `constructorId`: Identificador do construtor do veículo.
- `position`: Posição final do piloto na corrida.
- `laps`: Número de voltas necessárias para terminar a corrida.
- `nationality`: Nacionalidade do piloto.
- `round`: Etapa da temporada da corrida.
- `circuitId`: Identificador do circuito onde a corrida ocorre.
- `country_circuit`: País onde o Grand Prix acontece.
- `height`: Altura acima do nível do mar do circuito.
- `nationality_constructors`: Nacionalidade do construtor do carro.
- `points_season`: Pontos acumulados pelo piloto durante a temporada.
- `position_season`: Posição atual do piloto na temporada.
- `wins_season`: Vitórias do piloto durante a temporada.
- `age`: Idade do piloto.
- `weather_condition`: Condição climática durante a corrida.
- `humidity`: Umidade durante a corrida.
- `temperature`: Temperatura durante a corrida.
- `AvgGrid`: Média de posições de grid do piloto.
- `AvgFn`: Média das posições finais do piloto.
- `wins_acc`: Vitórias acumuladas pelo piloto ao longo de sua carreira.

Para a feature target, fizemos uma transformação, no qual as posição de 1 a 10 recebiam uma porcentagem de 100 a 10, de 10 em 10, representando a probabilidade de ganho de cada piloto.

Essas características representam uma combinação de informações sobre o piloto, o carro, o circuito e as condições durante a corrida. Acreditamos que este conjunto de características fornece uma visão abrangente dos fatores que podem influenciar o desempenho de um piloto em uma corrida de Fórmula 1.

4.2. Seleção de Algoritmos

A seleção de algoritmos focou em cinco modelos: Random Forest, Gradient Boosting, XGBoost, SVR e MLP. Esses algoritmos foram escolhidos devido aos melhores resultados obtidos em testes preliminares, exceto pelo SVR, que foi selecionado para comparação com a literatura existente.

Para o Random Forest e SVR, foram utilizados os hiperparâmetros padrões. Já para o Gradient Boosting, foi aplicado um `RandomizedSearchCV` do scikit-learn para encontrar os melhores parâmetros, resultando em: {`'validation_fraction': 0.1, 'tol': 0.0001, 'subsample': 1.0, 'random_state': 42, 'n_iter_no_change': 20, 'n_estimators': 500, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'log2', 'max_depth': 5, 'loss': 'huber', 'learning_rate': 0.1, 'criterion': 'friedman_mse', 'ccp_alpha': 0.1, 'alpha': 0.9`}.

Para o XGBoost, também foi aplicado o `RandomizedSearchCV`, resultando nos seguintes melhores parâmetros: {`'subsample': 0.7, 'reg_lambda': 0.1, 'reg_alpha': 0.2, 'n_jobs': -1, 'n_estimators': 1500, 'min_child_weight': 7, 'max_depth': 8, 'learning_rate': 0.01, 'gamma': 0.2, 'colsample_bytree': 0.5`}.

Finalmente, para o MLP, foi usado o `HalvingRandomSearch`, que busca os parâmetros que resultam em melhorias. Este método resultou em: {`'solver': 'adam', 'momentum': 0.95, 'max_iter': 500, 'learning_rate_init': 0.001, 'learning_rate': 'adaptive', 'hidden_layer_sizes': (100, 50), 'epsilon': 1e-06, 'beta_2': 0.9, 'beta_1': 0.9, 'batch_size': 32, 'alpha': 0.1, 'activation': 'tanh'`}.

Todos os modelos, foram validados através de validação cruzada com 5 folds.

Depois de realizar a engenharia de características e avaliar o melhor algoritmo para os modelos com e sem Grid, os modelos foram salvos usando a biblioteca pickle para uso futuro.

4.3. Resultados

Os resultados obtidos através da da engenharia de atributos e modelos de aprendizado de máquina são exibidos na Figura 2 e Figura 3, que representam os cenários sem e com a utilização da característica `Grid`, respectivamente.

Ao analisar os resultados sem a utilização do `Grid` (Figura 2), observa-se que os desempenhos dos modelos não são tão promissores. Isso pode ser atribuído à complexidade intrínseca desse tipo de previsão, que precisa levar em consideração uma grande quantidade de fatores externos.

No entanto, ao incorporar a característica `Grid` nos modelos (Figura 3), notamos uma melhoria significativa no desempenho. Embora os resultados ainda não sejam ideais, eles se assemelham àqueles encontrados na literatura existente, validando de certa forma

Sem GRID			
	R2	MAE	MSE
RF	0.552 (+/- 0.116)	15.220 (+/- 2.454)	511.629 (+/- 167.864)
GB	0.559 (+/- 0.130)	15.378 (+/- 2.203)	503.815 (+/- 188.130)
XGB	**0.592 (+/- 0.137)**	**14.716 (+/- 2.749)**	466.096 (+/- 190.817)
SVR	0.178 (+/- 0.293)	21.436 (+/- 3.835)	940.823 (+/- 414.893)
MLP	0.477 (+/- 0.153)	16.423 (+/- 2.929)	**16.423 (+/- 2.929)**

Figure 2. Resultados sem a utilização da característica `Grid`.

nossa abordagem. É importante notar, no entanto, que a literatura sugere um desempenho superior do modelo SVR, um aspecto que difere dos nossos resultados.

Com GRID			
	R2	MAE	MSE
RF	0.599 (+/- 0.122)	14.166 (+/- 2.691)	458.226 (+/- 172.413)
GB	0.610 (+/- 0.136)	14.217 (+/- 2.381)	445.944 (+/- 190.628)
XGB	**0.633 (+/- 0.143)**	**13.799 (+/- 2.837)**	**420.075 (+/- 195.011)**
SVR	0.203 (+/- 0.274)	21.120 (+/- 3.749)	911.605 (+/- 390.605)
MLP	0.559 (+/- 0.128)	14.971 (+/- 3.110)	502.929 (+/- 175.641)

Figure 3. Resultados com a utilização da característica `Grid`.

Em resumo, embora os resultados mostram algum potencial, porém ainda há espaço para melhorias. A complexidade da previsão do sucesso em corridas de Fórmula 1, que depende de uma ampla gama de fatores, torna esse um desafio significativo.

5. Abertura

Após a conclusão e validação dos modelos de previsão, foi disponibilizado para acesso publico por meio de uma instância na plataforma Azure (20.213.156.162), a predição para os pilotos de formula 1. Essa instância hospeda uma API, a qual inclui a rota /predict. Essa rota é capaz de retornar uma classificação dos pilotos com maior probabilidade de vitória na próxima corrida de Fórmula 1.

Logo após a etapa de qualificação, e capturado os resultados e atualizado os modelos com esses novos dados. Isso permite gerar uma previsão atualizada para a corrida de domingo. Além disso, na segunda-feira, após o término de um Grand Prix, atualizamos os dados dos modelos usando a API do Ergast. Com esses dados atualizados, rodamos o nosso pipeline de criação e validação de algoritmos.

Essa abordagem nos permite manter nossas previsões atualizadas com os dados mais recentes, aumentando a relevância e a precisão das previsões. Através do uso eficaz de dados em tempo real, buscamos fornecer previsões confiáveis para cada corrida de Fórmula 1.

References

Ganapathi, A. (2023). F1 champ: Eda: Classification 100% accuracy.

George, W. (2021). Formula 1 championship predictor: A machine learning solution.

Russo, A. J. (2003). A aws desenvolve o f1 insights.

SICOIE, H. *MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR*. PhD thesis, tilburg university.

Vopani, R. (2023). Formula 1 world championship (1950 - 2023).