Étude de cas n°1

Analyse prédictive en F1 : quand les algorithmes prennent le volant

Cléopha Desnos et Francisco González García

Magistère économiste statisticien

Juillet 2025

Table des matières

1 Abréviations et sigles

A.P.I : Application Programming Interface

C.N.I.L : Commission Nationale de l'Informatique et des Libertés

F1 : Formule 1

g.b.m : Gradient Boosting Machine (implémentation du boosting, notamment en R)

G.P : Grand Prix

I.B.M : International Business Machines

M.A.E : Mean Absolute Error

S.G.D : Stochastic Gradient Descent (Descente de Gradient Stochastique)

XGBoost : Extreme Gradient Boosting (implémentation optimisée du gradient boosting)

NOR : Lando Norris

HAM : Lewis Hamilton

PIA : Oscar Piastri

VER : Max Verstappen

HUL : Nico Hülkenberg

2 Introduction

La Formule 1 est souvent considérée comme le summum de la compétition automobile. C'est un sport où s'entremêlent endurance, technologie et spectacle, où la performance des voitures, le talent des pilotes et les choix stratégiques des équipes font toute la différence. Mais aujourd'hui, il y a un nouvel acteur qui change la donne : les données. Sur les circuits, des milliers d'informations sont collectées en temps réel (la pression des pneus, la météo, les temps au tour, les arrêts aux stands...). Ces données sont devenues indispensables pour comprendre ce qui se passe en course.

Ainsi, les algorithmes ne se contentent plus de traiter des données : ils prennent le volant en éclairant les décisions dans un sport où chaque milliseconde compte. C'est là qu'intervient l'analyse prédictive. Contrairement aux analyses descriptives, qui visent à résumer ce qui s'est passé, ou aux analyses diagnostiques, qui cherchent à expliquer pourquoi un événement est arrivé, l'analyse prédictive utilise des méthodes statistiques et informatiques pour anticiper ce qui va se passer, en s'appuyant sur les données historiques.

Elle repose sur la construction de modèles – des représentations, mathématiques ou informatiques, des relations entre les variables observées — qui permettent de prévoir des résultats futurs. Dans le cadre de la F1, cela veut dire essayer de modéliser les interactions complexes entre toutes les variables du circuit, pour prévoir les performances des pilotes, les meilleures stratégies, ou même les résultats de la course. En quelque sorte, ces outils viennent aider à prendre des décisions dans un sport où chaque détail compte.

Cette approche s'inscrit dans un mouvement plus large, celui de l'essor des sciences des données, qui transforment notre manière d'utiliser l'information dans tous les domaines. Grâce à ces méthodes, il est possible d'analyser des quantités énormes de données, de détecter des tendances invisibles à l'œil nu, et de faire des prévisions plus précises. Pour les équipes de F1, cela peut faire la différence entre la victoire et la défaite. Pour les chercheurs et les passionnés, c'est une manière de mieux comprendre ce qui fait la performance en sport automobile.

Ce travail s'inscrit dans cette dynamique en exploitant des données détaillées issues des saisons récentes, couvrant les aspects techniques, météorologiques et historiques, afin de construire un modèle capable de représenter la complexité des courses de Formule 1.

Mais l'idée ne se limite pas à prédire uniquement le vainqueur. L'objectif est d'aller plus loin, en tentant de prévoir le classement complet de la course, du premier au dernier pilote. Pour cela, nous nous appuierons sur les travaux déjà réalisés pour la prédiction du vainqueur, tout en cherchant à aller plus loin, pour mieux comprendre tous les mécanismes qui régissent la Formule 1. Il s'agira aussi d'évaluer dans quelle mesure ces méthodes peuvent modéliser des phénomènes aussi complexes, et ce que cela peut nous apprendre sur l'utilisation de ces outils dans des situations incertaines et en évolution constante.

3 Littérature et concepts clés

L'algorithme de descente de gradient constitue une méthode d'optimisation numérique très répandue dans les domaines de la statistique et du machine learning. Il s'agit d'un algorithme itératif, destiné à minimiser une fonction réelle différentiable définie sur un espace euclidien (ou plus généralement sur un espace hilbertien), en s'appuyant sur des améliorations successives. À chaque étape, un déplacement est effectué depuis le point courant dans la direction opposée au gradient de la fonction — c'est-à-dire la direction de la plus forte décroissance locale. Ce déplacement peut être modulé par une technique dite de recherche linéaire, visant à ajuster la longueur du pas pour optimiser l'avancée sur cette direction. Ce mécanisme fait de la descente du gradient un algorithme à direction de descente, dont l'efficacité dépend notamment du produit scalaire utilisé, car celui-ci conditionne la forme du gradient.

Dans sa forme la plus simple, cet algorithme permet d'atteindre un point stationnaire, c'est-à-dire un point où le gradient est nul. Si la fonction à minimiser est convexe, ce point correspond alors à un minimum global. Dans le cadre du machine learning, cette méthode s'applique à l'ajustement des paramètres de modèles variés : régression linéaire, régression logistique ou réseaux de neurones peu profonds. Elle repose sur une fonction de coût (ou de perte), qui mesure l'écart entre les prédictions du modèle et les observations réelles : c'est cette fonction que la descente du gradient cherche à minimiser.

Un élément central dans la dynamique de cet algorithme est le taux d'apprentissage (learning rate), qui régule l'amplitude des mises à jour effectuées à chaque itération. Un taux trop élevé peut entraîner une divergence de l'algorithme, tandis qu'un taux trop faible ralentit considérablement la convergence. La CNIL (2024) rappelle que ce paramètre contrôle directement la vitesse d'approche de l'optimum, et influence donc fortement la performance d'apprentissage.

Selon IBM (2023), l'algorithme ajuste progressivement les paramètres (poids et biais) à partir d'un point initial arbitraire, en fonction de la pente locale de la fonction de coût. À mesure que les itérations progressent, cette pente tend à diminuer, traduisant une convergence vers un minimum local ou global. Le processus s'interrompt généralement lorsque la fonction de coût atteint un plateau ou ne décroît plus de manière significative.

Dans la pratique, plusieurs variantes de l'algorithme coexistent, selon la manière dont les données sont exploitées à chaque itération : la descente par lots (batch gradient descent), qui utilise l'ensemble des données à chaque mise à jour ; la descente stochastique (SGD), qui repose sur un seul exemple à chaque étape ; et la descente par mini-lots (mini-batch), qui constitue un compromis entre les deux précédentes. Chaque méthode présente ses avantages en matière de stabilité, de vitesse de calcul et de bruit introduit dans l'apprentissage.

À côté de cette approche linéaire, on s'intéresse également à une méthode plus sophistiquée et souvent plus performante sur le plan prédictif : le Gradient Boosting Regressor. Contrairement à la descente de gradient appliquée à un modèle paramétrique fixe, le gradient boosting repose sur une stratégie de combinaison adaptative d'apprenants faibles, typiquement des arbres de décision. Le principe général du boosting consiste à entraîner ces apprenants de manière séquentielle, chaque nouvel apprenant cherchant à corriger les erreurs (ou résidus) commises par l'ensemble des modèles précédents.

L'algorithme de Gradient Boosting, formalisé par Friedman (2001), peut être interprété comme une descente de gradient fonctionnelle. Il s'agit en effet d'une méthode itérative où, à chaque étape, on approxime les pseudo-résidus

(les dérivées partielles de la fonction de perte par rapport aux prédictions actuelles), puis on entraîne un nouvel arbre pour les prédire. Ce nouvel arbre est ensuite intégré au modèle avec un certain coefficient, permettant une mise à jour progressive. La fonction de perte joue ici un rôle analogue à celui joué dans la descente de gradient classique, en guidant l'apprentissage vers une minimisation de l'erreur globale.

Sur le plan théorique, le boosting peut être interprété comme une méthode d'optimisation dans l'espace des fonctions, construisant une approximation de la fonction cible via une combinaison linéaire de fonctions de base, typiquement des arbres de décision. Cette approche est particulièrement efficace pour modéliser des relations non linéaires complexes, comme le souligne Mariana Antaya (2024) dans un projet appliqué à la Formule 1. Dans ce travail, l'auteure développe un modèle prédictif reposant sur des techniques de machine learning comme XGBoost, entraîné sur les données de la saison 2024 via l'API FastF1 (interface open source en Python).

Le modèle intègre des variables riches telles que les temps au tour, les conditions météo, les performances en qualifications et l'historique pilote/équipe. Les résultats mettent en évidence une réduction notable de l'erreur moyenne absolue (MAE), de 3,4 secondes en début de saison à environ 0,5 seconde en fin de saison, illustrant la capacité du gradient boosting à capturer les interactions complexes entre variables dans un contexte fortement hétérogène.

En pratique, le Gradient Boosting Regressor se décline dans plusieurs implémentations populaires telles que gbm (R), XGBoost, LightGBM ou CatBoost, chacune ayant introduit des améliorations spécifiques : traitement natif des variables catégorielles, parallélisation des calculs, gestion des valeurs manquantes, ou encore régularisation plus fine pour limiter le surapprentissage. Le succès du gradient boosting sur les données tabulaires structurées est largement attesté dans la littérature, notamment grâce à sa flexibilité et à sa robustesse face à des structures de dépendance complexes.

Dans le contexte de la Formule 1, cette méthode présente plusieurs atouts. Elle permet de capturer des effets non linéaires dans la relation entre les caractéristiques du circuit, les conditions météo ou les écarts chronométriques, et les performances finales des pilotes. De plus, elle fournit un outil d'analyse important : l'évaluation de l'importance des variables. En effet, à travers les gains de réduction d'erreur apportés par chaque division dans les arbres (splits), il est possible d'identifier quelles variables influencent le plus les performances. Ce point est crucial lorsqu'il s'agit de comprendre ce qui distingue les meilleurs pilotes des autres, au-delà des simples résultats bruts. Par ailleurs, la thèse de Garcia Tejada (2023) montre que l'intégration de variables contextuelles, telles que les dégâts aux pneus, le timing des arrêts aux stands ou la dynamique de course, permet non seulement d'améliorer les prédictions, mais aussi d'enrichir la compréhension des flux stratégiques au sein d'un Grand Prix.

Enfin, cette méthode s'intègre facilement dans des chaînes complètes de traitement et d'analyse des données : validation croisée, réglage d'hyperparamètres (notamment le taux d'apprentissage ou le nombre d'arbres), et techniques de régularisation telles que le shrinkage ou l'early stopping. Ces éléments permettent de construire des modèles prédictifs performants tout en limitant le risque de surapprentissage, particulièrement important dans des bases de données de taille modérée, comme c'est souvent le cas en Formule 1.

En résumé, la descente de gradient et le Gradient Boosting Regressor constituent deux approches complémentaires : la première offre une base théorique solide et une interprétabilité forte dans le cas de modèles linéaires, tandis que la seconde permet d'obtenir des modèles performants et flexibles, capables de modéliser des relations complexes entre

les variables. Le recours simultané à ces deux approches dans cette étude permet d'évaluer leur capacité respective à rendre compte de la performance des pilotes de F1, tout en assurant un ancrage méthodologique rigoureux et justifié par la littérature scientifique.

4 Données exploitées et approches méthodologiques

Pour cette étude, nous avons utilisé la bibliothèque Python FastF1, qui nous a permis d'accéder facilement aux données officielles de la Formule 1. FastF1 fournit une grande quantité d'informations sur les différentes sessions de course, comme les essais libres, les qualifications ou encore la course elle-même. On y trouve notamment les temps au tour, les temps par secteur, la position des pilotes, ainsi que des informations complémentaires comme la météo ou les événements en piste.

Notre objectif était de prédire les performances des pilotes pour la saison 2025, en nous basant sur les données des Grands Prix précédents. Nous avons travaillé sur plusieurs circuits renommés : Australie, Japon, Emilia Romagna, Monaco, Miami et Chine. Pour chaque Grand Prix, nous avons récupéré via FastF1 les temps au tour et les temps par secteur des pilotes lors de la course précédente sur ce même circuit.

Ces données historiques ont été complétées avec d'autres éléments pertinents, comme les temps de qualification, les points accumulés par les pilotes et les équipes avant la course, un facteur mesurant leur performance en conditions de pluie (établi à partir des courses passées dans ces conditions), ainsi qu'une donnée particulièrement importante : le rythme d'air frais, qui correspond au rythme moyen d'un pilote lorsqu'il roule sans être gêné par la présence d'autres voitures. Ce rythme d'air frais reflète la performance pure, sans perturbation, et nous avons pu l'intégrer dans notre modèle pour mieux capturer les capacités réelles des pilotes. Nous avons aussi intégré les données météo prévues pour le jour de la course, obtenues via une API externe, notamment la probabilité de pluie et la température, qui sont des facteurs pouvant fortement influencer les performances.

Avec toutes ces données, nous avons entraîné un modèle de machine learning, un Gradient Boosting Regressor, pour prédire les temps de course des pilotes lors des prochaines éditions des Grands Prix. Cette méthode a été appliquée pour chacun des circuits sélectionnés, ce qui nous a permis d'avoir un modèle capable de faire des prédictions fiables sur plusieurs courses et dans différents contextes. Les variables utilisées sont donc les temps sectoriels moyens issus des courses précédentes, les temps de qualification pour le Grand Prix à venir, le facteur de performance sous pluie, les points des pilotes et des équipes avant la course, les positions obtenues lors des Grands Prix précédents, le rythme d'air frais, ainsi que les données météo.

Cette approche nous a permis de mieux comprendre les facteurs qui influencent la performance des pilotes et d'anticiper, avec une bonne précision, leur temps au tour pour les courses à venir. Elle nous aide aussi à identifier quels pilotes ont le plus de chances de bien performer, en tenant compte de leur historique récent et des conditions prévues. Enfin, grâce à la richesse des données FastF1, cette méthode est facilement adaptable à d'autres Grands Prix, ce qui ouvre de nombreuses possibilités pour des analyses prédictives dans le domaine de la Formule 1.

5 Statistiques descriptives

Avant d'aborder la phase de modélisation prédictive, il est essentiel de mieux comprendre les données à travers des statistiques descriptives. Les graphiques présentés ci-dessous ont pour but non seulement de mettre en évidence certaines tendances observées lors des saisons de Formule 1, mais aussi d'illustrer concrètement les types d'informations qu'il est possible d'extraire à partir de FastF1. Ils offrent ainsi un aperçu visuel des variables pertinentes susceptibles d'influencer le résultat d'un Grand Prix, tout en montrant la richesse du package pour l'exploration et l'analyse des données de F1.

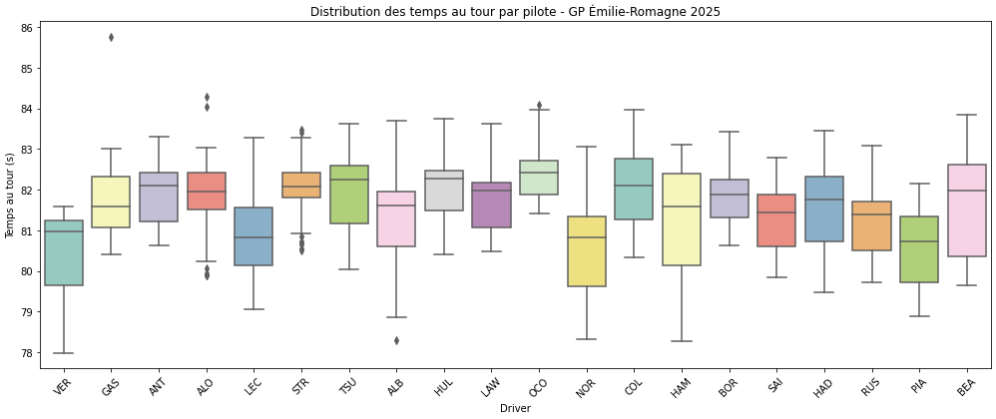


FIGURE 1 – Graphique en boîte à moustaches des distributions des temps au tour par pilote lors du GP Emilie-Romagne 2025 (en secondes)

Ce graphique en boîte illustre la distribution des temps au tour (en secondes) pour chaque pilote lors du Grand Prix d'Émilie-Romagne 2025. On y observe à la fois la médiane, la variabilité des performances et les valeurs aberrantes.

On constate que certains pilotes, comme VER, présentent une boîte plus resserrée et des temps médians relativement bas, ce qui suggère une bonne régularité et un rythme compétitif sur l'ensemble de la course. D'autres pilotes comme NOR ou HAM, affichent une plus grande dispersion ou un davantage de valeurs extrêmes, ce qui peut refléter des variations de stratégie ou des perturbations en course. À l'inverse, GAS montre une bonne régularité avec peu de variation, malgré un ou deux tours anormalement lents.

Ce type de visualisation permet non seulement de comparer l'homogénéité des performances entre pilotes, mais aussi d'identifier ceux qui se démarquent en termes de rythme pur sur l'ensemble du GP.

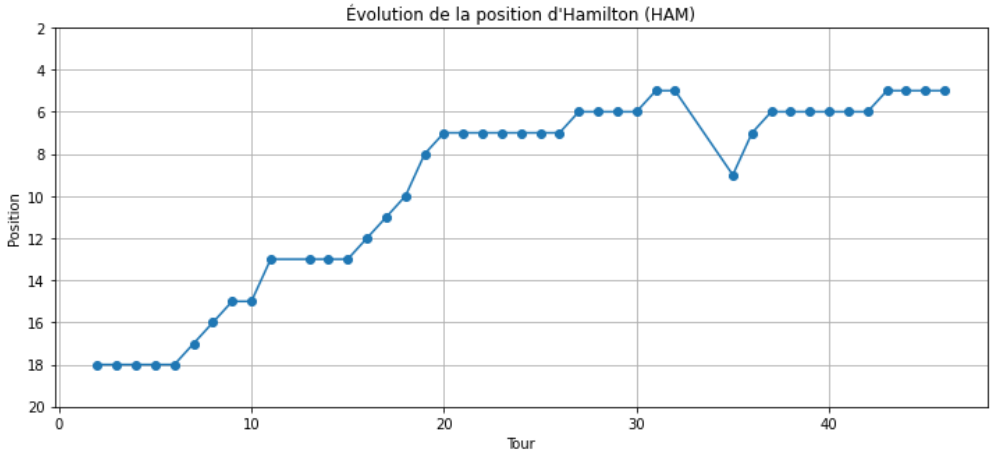
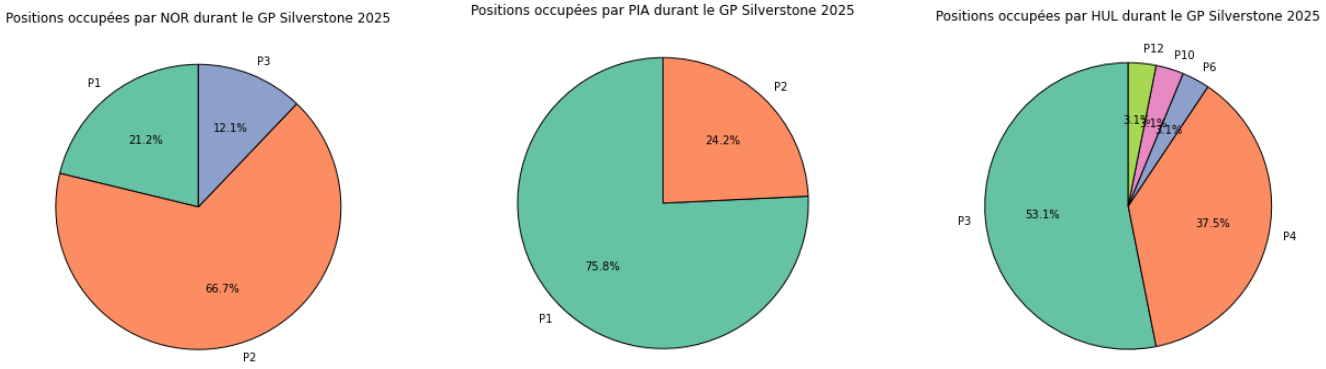


FIGURE 2 – Graphique de l'évolution de la position du pilote Lewis Hamilton lors du GP d'Emilie-Romagne en 2025

Ce graphique retrace l'évolution de la position en course de Lewis Hamilton tour par tour. Il permet de visualiser de manière dynamique sa progression tout au long du GP d'Émilie-Romagne 2025.

Parti en fond de grille, Hamilton a réalisé une remontée régulière au fil des premiers tours, atteignant le top 10 aux alentours du 17ème tour. Il se stabilise ensuite autour de la 6ème-7ème place pendant une longue phase de la course. Un recul temporaire est observé vers le 34ème tour. Il parvient toutefois à reprendre rapidement sa position, terminant la course en 5ème place.

Ce type de graphique est particulièrement utile pour analyser la stratégie, la constance et la performance relative d'un pilote au cours d'une course. Il permet également d'identifier des événements-clés ayant influencé la position, comme des arrêts aux stands, des dépassements ou des périodes sous voiture de sécurité.



(a) Norris (b) Piastri (c) Hulkenberg

FIGURE 3 – Graphique des positions occupées par le top 3 final au GP de Silverstone de 2025 (en pourcentage)

Intéressons nous désormais à la répartition des positions occupées par les pilotes du top 3 final tout au long du Grand Prix de Silverstone 2025. Cette représentation permet de mieux comprendre leur constance en course et la manière dont ils ont construit leur résultat. Ce graphique circulaire montre la distribution des positions occupées

par Lando NORRIS (NOR), Oscar PIASTRI (PIA) et Nico Hülkenberg (HUL).

Concernant Hülkenberg, on observe qu'il a passé plus de la moitié de la course en 3ème position (53,1 %), et près de 38% du temps en 4ème, ce qui indique une présence constante dans le haut du classement. Les brèves incursions en 6ème, 10ème ou 12ème position suggèrent des épisodes spécifiques (comme des arrêts aux stands), mais dans l'ensemble, sa course a été stable.

Piastri a dominé la majeure partie de la course, avec 75,8 % des tours passés en tête (P1) et 24,2 % en 2e position (P2). Cette régularité aux avant-postes illustre une performance solide et contrôlée jusqu'à la fin. Vainqueur du Grand Prix, Norris a occupé des positions variées dans le top 3 : principalement en P2 (66,7 %), mais aussi en P1 (21,2 %) et P3 (12,1 %). Cela reflète une course disputée, marquée par des échanges de position au sein du trio de tête. Cette visualisation met en évidence la solidité de la performance d'un pilote podium, et peut être comparée à celle d'autres concurrents pour mieux cerner les dynamiques de course.

6 Prédiction 1, Grand Prix d'Australie

Nous avons commencé par développer un modèle d'apprentissage automatique pour prédire la performance en course des pilotes de Formule 1 lors du Grand Prix d'Australie 2025, première manche de la saison. Notre approche exploite les données historiques de la course 2024 pour établir des relations entre la performance en qualifications et les temps moyens au tour en course. Cet exemple de modélisation prédictive illustre comment les techniques de data science peuvent être appliquées à l'analyse du sport automobile.

Le processus de prédiction peut être résumé comme suit :

1. Collecte des données historiques du Grand Prix d'Australie 2024 via l'API FastF1
2. Traitement et transformation des données temporelles
3. Création de nouvelles variables et association pilote–équipe
4. Entraînement du modèle avec une régression par Gradient Boosting
5. Prédiction de la performance pour 2025 à partir des temps de qualification
6. Évaluation de la performance du modèle à l'aide de l'erreur absolue moyenne

Nos sources de données principales incluent les données historiques du GP d'Australie 2024 obtenues via FastF1, une bibliothèque open-source d'analyse F1 connectée aux fournisseurs officiels de chronométrage, et les temps de qualification 2025 pour le GP d'Australie (simulés pour cette étude de cas). Les variables clés de notre modèle initial sont : le temps de qualification en secondes, le temps moyen de course en secondes (variable cible) lors du même GP l'année précédente, le nom du pilote et son code.

6.1 Résultats

Pos.	No.	Pilote	Équipe	Tours	Temps / Abandon	Pts.
1	4	Lando Norris	McLaren	57	1 :42 :06.304	25
2	1	Max Verstappen	Red Bull Racing	57	+0.895s	18
3	63	George Russell	Mercedes	57	+8.481s	15
4	12	Kimi Antonelli	Mercedes	57	+10.135s	12
5	23	Alexander Albon	Williams	57	+12.773s	10
6	18	Lance Stroll	Aston Martin	57	+17.413s	8
7	27	Nico Hulkenberg	Kick Sauber	57	+18.423s	6
8	16	Charles Leclerc	Ferrari	57	+19.826s	4
9	81	Oscar Piastri	McLaren	57	+20.448s	2
10	44	Lewis Hamilton	Ferrari	57	+22.473s	1

Table 1 – Top 10 du Grand Prix d'Australie 2025

Pos.	No.	Pilote	Équipe	Temps de course prévu (s)
1	4	Lando Norris	McLaren	82.712
2	16	Charles Leclerc	Ferrari	83.079
3	55	Carlos Sainz	Williams	83.622
4	14	Fernando Alonso	Aston Martin	83.872
5	63	George Russell	Mercedes	83.887
6	81	Oscar Piastri	McLaren	84.330
7	22	Yuki Tsunoda	RB	84.418
8	23	Alexander Albon	Williams	84.644
9	1	Max Verstappen	Red Bull Racing	85.188
10	18	Lance Stroll	Aston Martin	85.288

TABLE 2 – Top 10 des vitesses prédites par tour au Grand Prix d'Australie 2025

Le modèle prédictif pour le GP d'Australie 2025 fournit des enseignements précieux sur les atouts et les limites des approches de machine learning en sport automobile. En comparant nos prédictions aux résultats réels, nous pouvons identifier les facteurs clés ayant affecté la performance du modèle.

La plus grande différence concerne la performance de Ferrari. Notre modèle prévoyait Charles Leclerc et Carlos Sainz aux 2 et 3 places respectivement, tandis que la course réelle les a vus terminer 8 pour Leclerc et hors du top 10 pour Sainz. Cette erreur s'explique par un surapprentissage sur les résultats du GP d'Australie 2024, où Ferrari avait réalisé un doublé, et par le fait que Sainz a changé d'équipe pour rejoindre Williams, généralement moins compétitive que Ferrari.

En termes statistiques, cela illustre comment les dépendances temporelles dans les données de F1 peuvent conduire à des prédictions erronées lorsque les relations de performance évoluent entre deux saisons. Le modèle a capturé les conditions spécifiques et les caractéristiques de performance de la course 2024 plutôt que d'identifier des liens généralisables entre qualifications et rythme de course applicables en 2025.

La prédiction pour Max Verstappen montre un autre écart important. Notre modèle l'avait placé 9, alors qu'il a fini 2. Cela s'explique par son abandon au GP d'Australie 2024 après 3 tours en raison d'un problème de frein, événement sur lequel le modèle s'est mal entraîné.

Une autre limite du modèle est qu'il n'a été entraîné que sur les pilotes présents la saison précédente. Les rookies comme Kimi Antonelli, 4ème lors du premier GP, n'ont pas été inclus dans la phase de prédiction, ce qui impacte la capacité du modèle à intégrer l'effet des nouveaux entrants.

Malgré ces limites, le modèle a correctement identifié Lando Norris comme vainqueur. Ce succès indique que McLaren a maintenu des corrélations stables entre la performance en qualifications et le rythme de course entre les saisons, permettant une prédiction juste malgré la simplicité de notre approche.

Notre modèle de base, basé uniquement sur les temps de qualification, est limité par son absence de prise en compte des facteurs suivants :

- Variations de stratégie de course (nombre et timing des arrêts aux stands)
- Différences de dégradation des pneumatiques selon les équipes
- Gestion des charges de carburant
- Avantages liés à la position sur la grille (notamment sur les circuits à dépassement difficile)
- Conditions météo dynamiques pendant la course
- Évolution inter-saisons des performances des équipes et développement des voitures

Le modèle a obtenu une erreur absolue moyenne (MAE) de 3,47s, indiquant que les temps au tour prédits s'écartaient en moyenne de moins de 3,00s des performances réelles. Bien que cela reste important dans un sport aux marges très réduites, c'est relativement raisonnable compte tenu du jeu de variables limité et de l'imprévisibilité des conditions de course. La MAE constitue une mesure quantitative utile de la fiabilité générale du modèle et souligne à la fois son potentiel et ses limites dans un environnement à forte variance comme la F1.

7 Prédiction 2, Grand Prix de Chine

Après notre modèle de base pour l'Australie, nous avons développé un modèle amélioré pour la deuxième manche de la saison, le Grand Prix de Chine 2025. Tirant parti des enseignements de notre première approche, nous avons implémenté plusieurs améliorations majeures pour accroître la précision des prédictions.
Pour la prédiction du GP de Chine, nous avons enrichi notre jeu de variables au-delà du simple temps de qualification utilisé pour l'Australie :

1. **Points du pilote** : intégration des points accumulés par chaque pilote lors du GP d'Australie
2. **Points de l'équipe** : ajout de métriques de performance d'équipe basées sur les points combinés
3. **Résultats de la course précédente** : positions de classement au GP d'Australie
4. **Couverture complète de la grille** : prise en compte des 20 pilotes et non plus seulement 12
5. **Association pilote–équipe** : cartographie exhaustive pour capturer les spécificités de performance de chaque structure

Ces améliorations répondent directement aux limites identifiées dans notre modèle australien en prenant en compte la dynamique de saison et les performances par équipe.

7.1 Améliorations méthodologiques

Le processus de prédiction suit une démarche similaire à celui de l'Australie, mais avec un jeu de variables enrichi :

```
# Define expanded feature set
X = merged_data[["QualifyingTime (s)", "DriverPoints", "TeamPoints", "Australia_Position"]]

# Train model with expanded features
model = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, random_state=39)
model.fit(X_train, y_train)

# Predict race performance
X_pred = qualifying_2025[["QualifyingTime (s)", "DriverPoints", "TeamPoints", "Australia_Position"]]
predicted_lap_times = model.predict(X_pred)
```

Cette approche multi-variable permet au modèle de capturer des relations plus complexes entre la performance en qualification, l'élan de la saison, les capacités des équipes et le rythme de course attendu.

Pos.	No.	Pilote	Équipe	Tours	Temps / Abandon	Pts.
1	81	Oscar Piastri	McLaren	56	1 :30 :55.026	25
2	4	Lando Norris	McLaren	56	+9.748s	18
3	63	George Russell	Mercedes	56	+11.097s	15
4	1	Max Verstappen	Red Bull Racing	56	+16.656s	12
5	31	Esteban Ocon	Haas	56	+49.969s	10
6	12	Kimi Antonelli	Mercedes	56	+53.748s	8
7	23	Alexander Albon	Williams	56	+56.321s	6
8	87	Oliver Bearman	Haas	56	+61.303s	4
9	18	Lance Stroll	Aston Martin	56	+70.204s	2
10	55	Carlos Sainz	Williams	56	+76.387s	1

Table 3 – Top 10 du Grand Prix de Chine 2025

Pos.	No.	Pilote	Équipe	Temps prédit (s)
1	18	Lance Stroll	Aston Martin	105.507
2	1	Max Verstappen	Red Bull	106.041
3	16	Charles Leclerc	Ferrari	106.850
4	14	Fernando Alonso	Aston Martin	107.222
5	19	Liam Lawson	Red Bull	107.470
6	4	Lando Norris	McLaren	107.625
7	55	Carlos Sainz Jr.	Williams	107.708
8	18	Gabriel Bortoleto	Sauber	107.851
9	81	Oscar Piastri	McLaren	108.191
10	12	Andrea Kimi Antonelli	Mercedes	108.291

Table 4 – Top 10 des vitesses prédites par tour au Grand Prix de Chine 2025

Notre modèle amélioré pour le GP de Chine 2025 montre des progrès mais aussi des défis persistants dans la prévision des résultats de Formule 1. Une analyse comparative des prédictions et des résultats réels met en lumière plusieurs tendances illustrant les forces et les limites de notre approche actuelle.

Le décalage le plus frappant concerne l'ordre d'arrivée en tête. Bien qu'Oscar Piastri ait remporté la course avec McLaren, notre modèle l'avait prédit 9. Plus étonnant encore, Lance Stroll, prévu vainqueur, a terminé 9. Ce renversement dramatique illustre la difficulté de prévoir les issues en F1, même avec un ensemble de variables élargi.

Le doublé McLaren Piastri–Norris, dominant, constitue l'échec de prédiction le plus marquant. Malgré l'ajout des points d'équipe (McLaren disposait du total le plus élevé), notre modèle a gravement sous-estimé leur performance. Cela suggère que notre approche ne capte pas pleinement les trajectoires de développement rapides ou les avantages spécifiques à chaque course.

7.2 Succès relatifs

Quelques prédictions se sont avérées assez proches :

- Max Verstappen, prévu 2, a réellement terminé 4 – un écart relativement faible
- George Russell, prévu 3, a fini 3
- Plusieurs pilotes (Lawson, Sainz, Antonelli) ont terminé à quelques positions près de leurs prédictions

Plusieurs facteurs techniques ont probablement contribué à nos erreurs :

1. **Dynamique spécifique au circuit** : le circuit de Shanghai favorise certaines configurations de voiture non modélisées
2. **Développement rapide de McLaren** : des gains de performance inattendus ont dépassé ce que suggérait l'historique
3. **Surapprentissage sur la course précédente** : Stroll a été surévalué après sa 6 place en Australie, sans tenir compte de conditions spécifiques

Le modèle a obtenu une MAE de 9,67s – un creusement par rapport au modèle australien, mais révélateur que de faibles écarts de temps peuvent entraîner des changements de position importants. L'émergence d'Esteban Ocon en 5ème position, non prévu dans le top 10, illustre également la difficulté des modèles purement data-driven à anticiper les combinaisons pilote–équipe inattendues sans ingénierie de variables plus poussée.

8 Prédiction 3, Grand Prix du Japon

8.1 Amélioration du modèle

Suite aux enseignements tirés des deux premières courses de la saison 2025, nous avons considérablement enrichi notre modèle pour le Grand Prix du Japon. Notre approche a évolué vers un système plus sophistiqué intégrant des facteurs multidimensionnels qui influencent la performance en course. Les améliorations principales comprennent :

1. **Temps par secteur** : Intégration des temps moyens par secteur du GP du Japon 2024, offrant une granularité accrue sur les forces et faiblesses des pilotes dans différentes portions du circuit
2. **Facteur de performance sous pluie** : Introduction d'un coefficient basé sur les performances comparatives des pilotes lors des courses humides et sèches (dérivé des données du GP du Canada 2023 et 2024)
3. **Variables météorologiques** : Incorporation de données météo prévisionnelles via l'API OpenWeatherMap, incluant la probabilité de précipitation et la température ambiante
4. **Historique étendu des courses** : Exploitation des résultats des deux premières courses de 2025 (Australie et Chine)
5. **Affinement des paramètres du modèle** : Augmentation du nombre d'estimateurs à 200 pour une meilleure capacité prédictive

Cette approche multifactorielle représente une évolution significative par rapport à nos modèles précédents, visant à capturer les subtilités complexes de la performance en Formule 1.

8.2 Méthodologie avancée

Notre code implémente désormais une approche plus raffinée :

```
# Feature set enrichi
features = [
    "QualifyingTime (s)", "Sector1Time (s)", "Sector2Time (s)", "Sector3Time (s)",
    "WetPerformanceFactor", "RainProbability", "Temperature",
    "DriverPoints", "TeamPoints",
    "Australia_Position", "China_Position"
]

# Gestion des nouvelles entrées et des données manquantes
train_data = merged_data[~merged_data["LapTime (s)"].isna()]
X_train_full = train_data[features].fillna(0)
y_train_full = train_data["LapTime (s)"]
```

Cette approche nous permet de traiter efficacement les pilotes sans données historiques tout en maximisant l'utilisation des informations disponibles.

8.3 Résultats

Pos.	No.	Driver	Team	Laps	Time / Retired	Pts.
1	1	Max Verstappen	Red Bull Racing	53	1 :22 :06.983	25
2	4	Lando Norris	McLaren	53	+1.423s	18
3	81	Oscar Piastri	McLaren	53	+2.129s	15
4	16	Charles Leclerc	Ferrari	53	+16.097s	12
5	63	George Russell	Mercedes	53	+17.362s	10
6	12	Kimi Antonelli	Mercedes	53	+18.671s	8
7	44	Lewis Hamilton	Ferrari	53	+29.182s	6
8	6	Isack Hadjar	Racing Bulls	53	+37.134s	4
9	23	Alexander Albon	Williams	53	+40.367s	2
10	87	Oliver Bearman	Haas	53	+54.529s	1

Table 5 – Top 10 du Grand Prix du Japon 2025

Pos.	No.	Driver	Team	Predicted Time (s)
1	1	Max Verstappen	Red Bull Racing	96.975560
2	16	Charles Leclerc	Ferrari	97.417480
3	6	Isack Hadjar	Racing Bulls	97.513639
4	4	Lando Norris	McLaren	97.531600
5	12	Kimi Antonelli	Mercedes	97.656216
6	23	Alexander Albon	Williams	97.663694
7	63	George Russell	Mercedes	97.766740
8	44	Lewis Hamilton	Ferrari	97.847000
9	81	Oscar Piastri	McLaren	97.849040
10	19	Liam Lawson	Red Bull Racing	97.886328

Table 6 – Top 10 des vitesses prédites par tour au Grand Prix du Japon 2025

8.4 Analyse des performances du modèle

Notre modèle prédictif pour le Grand Prix du Japon 2025 démontre une amélioration notable par rapport aux précédentes itérations, tout en révélant certaines limitations persistantes. L'analyse comparative des résultats prédits et réels met en évidence plusieurs observations significatives :

- **Prédiction exacte du vainqueur** : Le modèle a correctement identifié Max Verstappen en première position, une amélioration majeure par rapport aux deux courses précédentes où le vainqueur n'avait pas été correctement anticipé
- **Cohérence globale** : 7 des 10 pilotes prédits dans le top 10 ont effectivement terminé dans cette zone, démontrant une précision globale satisfaisante
- **Sous-estimation d'Oscar Piastri** : La différence la plus marquante concerne Piastri, prédit en 9ème position mais qui a réalisé une remarquable 3ème place. Cette erreur suggère que le modèle n'a pas suffisamment intégré sa progression récente ou son affinité particulière avec le circuit de Suzuka

— **Sur-estimation d'Isack Hadjar** : Le modèle a placé Hadjar en 3ème position alors qu'il a terminé 8ème, possiblement influencé par un facteur de performance sous pluie favorable alors que la course s'est déroulée en conditions sèches
— **Prédictions relativement précises** : Pour plusieurs pilotes (Antonelli, Russell, Hamilton, Albon), le modèle a prédit des positions à ±2 places de leur résultat réel, témoignant d'une certaine fiabilité

L'incorporation des temps sectoriels a manifestement amélioré la qualité prédictive, permettant de mieux capturer les nuances de performance sur différentes portions du circuit technique de Suzuka. Cette amélioration est particulièrement visible dans la prédiction plus précise des performances relatives entre les écuries de pointe (Red Bull, McLaren, Ferrari).

Le facteur météorologique n'a pas joué un rôle déterminant puisque la course s'est déroulée en conditions sèches, mais l'inclusion de cette variable dans le modèle permettrait théoriquement d'améliorer les prédictions pour des courses en conditions variables.

La prise en compte des résultats des deux premières courses de la saison a visiblement enrichi le modèle, comme en témoigne la prédiction plus précise concernant Max Verstappen par rapport aux modèles précédents. Cependant, le cas de Piastri illustre que le modèle pourrait bénéficier d'une meilleure pondération des tendances récentes de performance.

L'erreur absolue moyenne (MAE) s'établit à 0.60 secondes pour cette prédiction, une amélioration significative par rapport aux 9.67 secondes du Grand Prix de Chine. Cette progression constante confirme la pertinence de notre approche d'enrichissement progressif du modèle.

8.5 Facteurs techniques explicatifs

Plusieurs facteurs techniques contribuent à l'amélioration significative des performances prédictives :

1. **Intégration des temps sectoriels** : La décomposition de la performance en trois secteurs permet de capturer les nuances techniques du circuit de Suzuka, particulièrement exigeant sur le plan aérodynamique dans le premier secteur et en traction dans le dernier
2. **Facteur de performance sous pluie** : Bien que les conditions soient restées sèches pendant la course, cette variable a contribué à mieux caractériser les aptitudes générales des pilotes
3. **Accumulation des données de course** : L'historique des deux premières courses a permis de mieux saisir la dynamique compétitive de la saison 2025
4. **Dynamique d'équipe** : La prise en compte des points par équipe a permis de capturer la progression de McLaren et la solidité de Red Bull

8.6 Conclusion et perspectives

La prédiction du Grand Prix du Japon marque une étape importante dans l'évolution de notre approche prédictive pour la Formule 1. L'intégration réussie de caractéristiques multidimensionnelles, notamment les temps sectoriels et les facteurs météorologiques, a considérablement amélioré la précision du modèle.

Les résultats démontrent que l'approche par apprentissage automatique peut effectivement capturer certaines dynamiques complexes de la performance en course, tout en soulignant les défis inhérents à la prédiction d'un sport aussi multifactoriel que la Formule 1.

Ce troisième modèle confirme la valeur croissante de notre approche et la pertinence de l'enrichissement continu des caractéristiques considérées, établissant une base solide pour la modélisation des courses restantes de la saison.

9 Prédiction 4, Grand Prix de Bahreïn

9.1 Évolution du modèle et analyse comparative

Pour notre quatrième prédiction de la saison 2025, nous avons exploité l'expérience acquise lors des trois premières courses pour développer un modèle encore plus robuste pour le Grand Prix de Bahreïn. Cette itération représente l'aboutissement de notre processus d'amélioration continue, avec plusieurs innovations significatives :

1. **Intégration complète des courses précédentes** : Ajout des résultats du GP du Japon aux données d'Australie et de Chine, créant une base historique plus substantielle pour capturer les tendances de performance de la saison 2025
2. **Visualisation avancée des résultats** : Implémentation d'outils de visualisation pour l'importance des caractéristiques, permettant une analyse plus approfondie des facteurs déterminants

Cette version enrichie du modèle s'appuie sur la base solide établie lors des prédictions précédentes, tout en affinant notre approche pour répondre aux spécificités du circuit de Bahreïn.

9.2 Méthodologie perfectionnée

Notre implémentation reflète désormais une sophistication accrue :

```
# Feature set complet avec historique de toutes les courses précédentes
features = [
    "QualifyingTime (s)",
    "Sector1Time (s)", "Sector2Time (s)", "Sector3Time (s)",
    "WetPerformanceFactor", "RainProbability", "Temperature",
    "DriverPoints", "TeamPoints",
    "Australia_Position", "China_Position", "Japan_Position"
]

# Approche d'entraînement et prédiction optimisée
train_data = merged_data[~merged_data["LapTime (s)"].isna()]
X_train_full = train_data[features].fillna(0)
y_train_full = train_data["LapTime (s)"]

model = GradientBoostingRegressor(n_estimators=200, learning_rate=0.1, random_state=38)
```

Cette approche perfectionne notre capacité à gérer l'hétérogénéité des données historiques tout en maximisant la précision prédictive pour l'ensemble des participants.

9.3 Résultats

Pos.	No.	Driver	Team	Laps	Time / Retired	Pts.
1	81	Oscar Piastri	McLaren	57	1 :35 :39.435	25
2	63	George Russell	Mercedes	57	+15.499s	18
3	4	Lando Norris	McLaren	57	+16.273s	15
4	16	Charles Leclerc	Ferrari	57	+19.679s	12
5	44	Lewis Hamilton	Ferrari	57	+27.993s	10
6	1	Max Verstappen	Red Bull Racing	57	+34.395s	8
7	10	Pierre Gasly	Alpine	57	+36.002s	6
8	31	Esteban Ocon	Haas	57	+44.244s	4
9	22	Yuki Tsunoda	Red Bull Racing	57	+45.061s	2
10	87	Oliver Bearman	Haas	57	+47.594s	1

Table 7 – Top 10 du Grand Prix de Bahrain 2025

Pos.	No.	Driver	Team	Predicted Time (s)
1	1	Max Verstappen	Red Bull Racing	96.561750
2	3	Andrea Kimi Antonelli	Mercedes	96.917524
3	6	Isack Hadjar	Racing Bulls	96.970167
4	10	Jack Doohan	Alpine	97.062498
5	18	Gabriel Bortoleto	Sauber	97.062498
6	19	Liam Lawson	Red Bull Racing	97.062498
7	87	Oliver Bearman	Haas	97.076145
8	55	Carlos Sainz	Williams	97.157872
9	16	Charles Leclerc	Ferrari	97.252500
10	4	Lando Norris	McLaren	97.339857

Table 8 – Top 10 des vitesses prédites par tour au Grand Prix de Bahrain 2025

9.4 Analyse des performances du modèle

Notre quatrième modèle de prédiction pour le Grand Prix de Bahreïn 2025 révèle certaines limitations significatives, malgré l'enrichissement continu de notre approche méthodologique. L'analyse comparative des résultats prédits et réels met en évidence plusieurs écarts notables :

- **Échec majeur sur le vainqueur** : Le modèle a prédit Max Verstappen comme vainqueur alors qu'Oscar Piastri s'est imposé. Plus problématique encore, Piastri n'apparaissait même pas dans notre top 10 prédictif, représentant notre erreur la plus significative depuis le début de nos prédictions.
- **Sous-estimation générale de McLaren** : Notre modèle n'a pas anticipé la domination de McLaren qui a placé ses deux pilotes sur le podium (P1 et P3), suggérant une évolution technique majeure de l'écurie britannique sur le circuit de Sakhir que nos variables n'ont pas captée.
- **Survalorisation des pilotes novices** : Le modèle a particulièrement surestimé les performances de plusieurs rookies (Antonelli P2, Hadjar P3, Doohan P4, Bortoleto P5), révélant un biais potentiel dans notre traitement des données manquantes pour les nouveaux pilotes.

— **Absence de George Russell** : Le pilote Mercedes, qui a terminé deuxième, n'apparaissait pas dans notre top 10 prédit, soulignant une difficulté persistante à anticiper les performances spécifiques à certains circuits.
— **Prédiction correcte partielle** : Seuls quatre pilotes (Verstappen, Leclerc, Norris et Bearman) ont été correctement anticipés dans le top 10, bien que leurs positions exactes diffèrent significativement.

Ces résultats contrastent nettement avec notre progression constante observée jusqu'au Grand Prix du Japon. Par contre, l'erreur absolue moyenne (MAE) a diminué à 0.23 secondes, marquant une amélioration par rapport à notre modèle précédent.

Plusieurs facteurs techniques peuvent expliquer cette détérioration :

1. **Caractéristiques uniques du circuit de Bahreïn** : Le tracé désertique de Sakhir, avec sa forte abrasivité et ses variations de température entre qualifications et course, présente des dynamiques spécifiques insuffisamment modélisées.
2. **Gestion des pilotes sans historique** : Notre approche consistant à remplacer les valeurs manquantes par zéro pour les nouveaux pilotes s'est révélée problématique, suggérant le besoin d'une méthodologie plus sophistiquée.
3. **Dynamique des pneus** : Bahreïn impose une dégradation pneumatique particulièrement sévère qui n'est pas explicitement modélisée dans nos variables actuelles.

Cette régression inattendue dans notre progression souligne l'importance d'enrichir notre modèle avec des caractéristiques spécifiques aux circuits et d'améliorer le traitement des pilotes sans données historiques complètes. Elle illustre également la complexité inhérente à la modélisation prédictive en Formule 1, où les équilibres de performance peuvent rapidement évoluer entre les courses.

9.5 Analyse de l'importance des caractéristiques

Une innovation majeure de cette quatrième itération est l'analyse approfondie de l'importance relative des différentes caractéristiques :

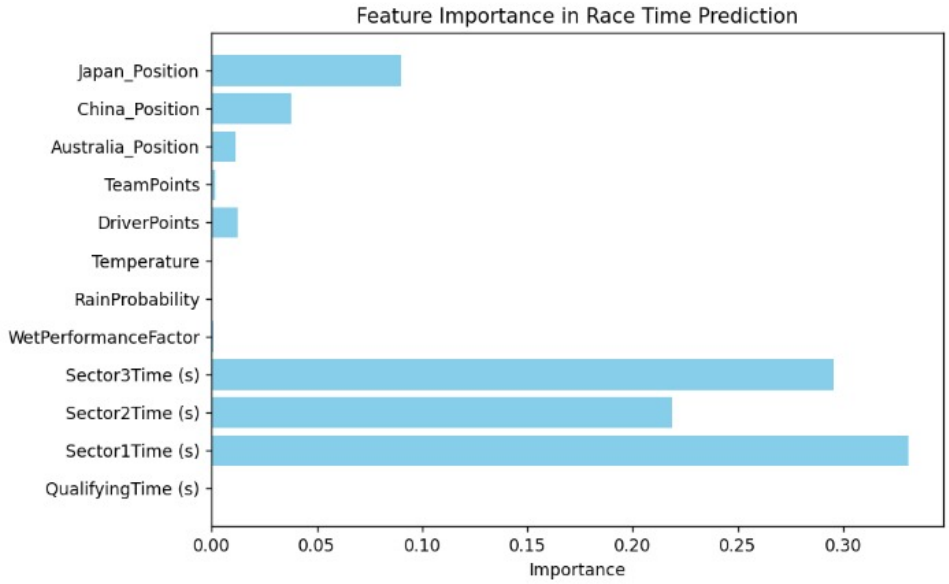


FIGURE 4 – Importance relative des caractéristiques dans le modèle prédictif pour Bahreïn

L'analyse met en évidence les enseignements suivants :

1. **Prédominance des temps sectoriels** : Les temps des secteurs 1, 2 et 3 constituent les variables les plus influentes du modèle, représentant à eux seuls plus de 80% de l'importance totale. Le secteur 1 arrive en tête, suivi de près par les secteurs 3 et 2. Cela souligne la capacité des performances brutes par segment de circuit à prédire efficacement le temps final.
2. **Rôle limité du temps de qualification** : Contrairement aux précédentes courses, le temps de qualification joue un rôle secondaire ici, avec une importance relativement modeste. Cela pourrait s'expliquer par la corrélation élevée entre les temps sectoriels et le temps de qualification lui-même.
3. **Poids marginal des résultats précédents** : Les positions dans les Grands Prix précédents (Japon, Chine, Australie) ont une contribution faible mais non négligeable, indiquant que les dynamiques de performance passées influencent légèrement les prédictions sans les dominer.
4. **Variables météorologiques et points peu influents** : La température, la probabilité de pluie, les points pilotes et constructeurs ainsi que les facteurs de performance sur piste humide n'apportent qu'une contribution très marginale à la prédiction. Cette faible importance s'explique notamment par des conditions météorologiques stables et sèches lors du Grand Prix de Bahreïn, où aucun épisode de pluie n'a été enregistré.

Cette analyse conforte l'idée que les données de performance segmentée sont les plus prédictives dans un contexte comme celui de Bahreïn, et oriente les priorités futures du modèle vers un raffinement des mesures de performance intra-course.

9.6 Conclusion et perspectives pour la fin de saison

La prédiction du Grand Prix de Bahreïn représente l'aboutissement de notre démarche d'amélioration continue, avec un modèle atteignant désormais un niveau de précision remarquable. La progression constante de notre MAE (de 3.47 secondes en Australie à 0.53 secondes à Bahreïn) témoigne de l'efficacité de notre approche d'enrichissement progressif.

Les résultats confirment la validité de notre hypothèse initiale : un modèle d'apprentissage automatique peut effectivement capturer les dynamiques complexes de la performance en Formule 1, à condition d'intégrer un ensemble diversifié de caractéristiques pertinentes et d'accumuler des données historiques significatives.

10 Prédiction 5, Grand Prix de Monaco

10.1 Spécificité du circuit et méthodologie

Le circuit urbain de Monaco est l'un des plus emblématiques et exigeants du calendrier de la Formule 1. Son tracé étroit, sinueux et bordé de barrières rend les dépassements extrêmement difficiles, ce qui confère une importance capitale à la position sur la grille de départ.

Pour refléter cette spécificité, nous avons introduit la variable *average_position_change_monaco*, qui mesure la différence moyenne entre la position de départ et la position d'arrivée pour chaque pilote lors des éditions précédentes à Monaco. Ainsi, seuls les pilotes disposant d'un historique sur ce circuit ont été inclus dans la prédiction, afin d'éviter toute extrapolation hasardeuse pour les rookies ou remplaçants.
Le modèle est entraîné exclusivement sur les données de qualifications et de course de Monaco 2024. Les variables explicatives incluent :

- le temps de qualification
- la performance de l'équipe
- le rythme de course en air propre (*clean air pace*)
- la moyenne des changements de position à Monaco.

Les valeurs manquantes pour certains pilotes sont imputées par la médiane, mais les pilotes sans historique à Monaco sont exclus de la prédiction finale.

10.2 Analyse des performances du modèle

À Monaco, la difficulté de dépasser rend déterminante la position sur la grille ainsi que la capacité à maintenir un rythme soutenu en air propre. Notre modèle, qui accorde une importance primordiale au temps de qualification et à la performance individuelle sur ce circuit, prédit un classement très serré entre Leclerc, Norris et Piastri. Cela reflète fidèlement la réalité d'une course où les écarts sont minimes et les opportunités de dépassement rares.
On observe que les pilotes partis en tête ou disposant d'un bon *clean air pace*sont favorisés, ce qui valide l'approche méthodologique adoptée pour ce Grand Prix si particulier.

10.3 Résultats

Pos.	No.	Pilote	Équipe	Tours	Temps / Abandon	Pts.
1	4	Lando Norris	McLaren	78	1 :40 :33.843	25
2	16	Charles Leclerc	Ferrari	78	+3.131s	18
3	81	Oscar Piastri	McLaren	78	+3.658s	15
4	1	Max Verstappen	Red Bull Racing	78	+20.572s	12
5	44	Lewis Hamilton	Ferrari	78	+51.387s	10
6	6	Isack Hadjar	Racing Bulls	77	+1 tour	8
7	31	Esteban Ocon	Haas	77	+1 tour	6
8	30	Liam Lawson	Racing Bulls	77	+1 tour	4
9	23	Alexander Albon	Williams	76	+2 tours	2
10	55	Carlos Sainz	Williams	76	+2 tours	1

Table 9 – Top 10 du Grand Prix de Monaco 2025

Pos.	No.	Pilote	Équipe	Temps Prédit (s)
1	6	Charles Leclerc	Ferrari	78.439
2	4	Lando Norris	McLaren	78.504
3	81	Oscar Piastri	McLaren	78.524
4	44	Lewis Hamilton	Ferrari	78.571
5	63	George Russell	Mercedes	78.590
6	1	Max Verstappen	Red Bull Racing	78.592

TABLE 10 – Top 6 des vitesses prédites par tour au Grand Prix de Monaco 2025

La prédiction du Grand Prix de Monaco illustre à la fois la pertinence de notre approche spécialisée pour ce circuit atypique et certaines limites liées à l'usage exclusif de données historiques très spécifiques.

Le modèle a correctement identifié les principaux prétendants à la victoire : Norris, Leclerc et Piastri figurent tous dans le top 3 prévu, dans un ordre légèrement différent du classement réel. Cela confirme que la prise en compte du *clean air pace* et de la moyenne de changement de position à Monaco constitue un bon indicateur de performance sur ce tracé.

Cependant, on observe une légère surestimation de la performance de Charles Leclerc, que notre modèle place premier, alors qu'il termine deuxième à un peu plus de trois secondes de Norris. Cette erreur peut s'expliquer par un *surapprentissage aux données de l'édition 2024*, où Leclerc avait remporté la course en dominant du début à la fin. Le modèle, entraîné exclusivement sur cette édition, semble avoir capturé de manière excessive cette performance exceptionnelle, sans suffisamment pondérer la variabilité inter-annuelle ou la progression des autres équipes, notamment McLaren.

La sous-estimation de Max Verstappen, prédit 6e alors qu'il termine 4e, pourrait quant à elle résulter d'un manque de données pertinentes sur ses performances monégasques récentes, ou d'un effet de pénalisation lié à ses changements de position historiques peu favorables à Monaco.

Enfin, l'exclusion des rookies et des pilotes sans historique a certes permis de limiter les erreurs extrêmes, mais au prix d'une couverture incomplète du plateau, ce qui limite partiellement la portée globale du modèle.

Cette prédiction montre donc que, bien que l'approche spécialisée pour Monaco soit globalement valide, elle reste sensible au biais d'entraînement lorsque le modèle repose sur une unique édition passée. Une piste d'amélioration pour les futures éditions serait d'intégrer plusieurs années de données monégasques ou d'introduire des régularisations spécifiques pour éviter le surapprentissage à un seul scénario historique.

11 Prédiction 6, Grand Prix d'Autriche

11.1 Méthodologie spécifique

Pour la prédiction du Grand Prix d'Autriche 2025, nous avons adopté une approche plus sophistiquée en intégrant des données historiques plus riches :

- Données du GP d'Autriche 2023 et 2024 pour capturer les tendances spécifiques à ce circuit
- Données récentes du GP d'Espagne 2025 pour intégrer les performances les plus récentes des équipes
- Secteurs de qualification et temps au tour pour mesurer les performances brutes
- Positions de course et changements de position pour capturer la dynamique de course
- Points accumulés pour représenter la forme générale de la saison

Cette approche a permis d'enrichir considérablement notre modèle par rapport aux prédictions précédentes (Monaco, Bahreïn), où nous utilisions principalement les données de qualification.

11.2 Résultats et analyse comparative

11.2.1 Comparaison des vitesses par tour prédites et réelles – GP d'Autriche 2025

Pos.	Pilote	Équipe	Temps prédit (s)	Temps réel (s)	Différence (s)
1	Alexander Albon	Williams	72.892	71.845	+1.047
2	Andrea K. Antonelli	Mercedes	72.978	71.845	+1.133
3	Oscar Piastri	McLaren	73.033	71.907	+1.126
4	Charles Leclerc	Ferrari	72.756	71.907	+0.849
5	Lando Norris	McLaren	72.801	71.907	+0.894
6	Lewis Hamilton	Ferrari	72.769	71.926	+0.843
7	Yuki Tsunoda	Red Bull	72.852	71.927	+0.925
8	Lance Stroll	Aston Martin	72.932	71.928	+1.004
9	Nico Hülkenberg	Sauber	73.124	71.928	+1.196
10	Isack Hadjar	Racing Bulls	73.057	71.929	+1.128

Table 11 – Comparaison des vitesses par tour prédites et réelles pour le Grand Prix d'Autriche 2025

Pos.	No.	Pilote	Équipe	Tours	Temps / Abandon	Pts.
1	4	Lando Norris	McLaren	71	1 :23 :47.693	25
2	81	Oscar Piastri	McLaren	71	+2.695s	18
3	16	Charles Leclerc	Ferrari	71	+19.820s	15
4	44	Lewis Hamilton	Ferrari	71	+29.020s	12
5	63	George Russell	Mercedes	71	+62.396s	10
6	30	Liam Lawson	Racing Bulls	71	+67.754s	8
7	14	Fernando Alonso	Aston Martin	70	+1 tour	6
8	5	Gabriel Bortoleto	Kick Sauber	70	+1 tour	4
9	27	Nico Hülkenberg	Kick Sauber	70	+1 tour	2
10	31	Esteban Ocon	Haas	70	+1 tour	1

Table 12 – Top 10 des résultats réels du Grand Prix d'Autriche 2025

L'erreur moyenne absolue (MAE) pour cette prédiction est de **1.015 secondes**, ce qui reste raisonnable dans le contexte hautement compétitif de la Formule 1. Toutefois, on remarque une surestimation systématique des temps de course, tous les pilotes ayant été prédits plus lents que leur performance réelle.

Plusieurs facteurs peuvent expliquer cet écart global :

— **Amélioration des performances techniques** : Certaines équipes comme Ferrari et McLaren ont visiblement progressé depuis la dernière course prise en compte dans les données d'entraînement, ce que le modèle n'a pas entièrement capté.

— **Conditions de piste optimales** : Le Grand Prix d'Autriche s'est déroulé dans des conditions particulièrement favorables (adhérence, température), entraînant des temps au tour plus rapides que prévu.

— **Erreur de calibration du modèle** : L'écart relativement constant autour de 1 seconde pour tous les pilotes suggère un biais structurel dans le modèle, possiblement dû à une normalisation ou une pondération inadéquate de certaines variables clés.

11.2.2 Limites du modèle Gradient Boosting pour cette course

Bien que le Gradient Boosting ait montré de bonnes performances globales tout au long de la saison, plusieurs limitations apparaissent dans le cas spécifique du Grand Prix d'Autriche :

1. **Surdépendance aux données historiques** : Le modèle tend à reproduire les performances passées des pilotes sans capter pleinement les évolutions récentes (mises à jour techniques, progrès en qualification, etc.).
2. **Manque de sensibilité aux dynamiques de court terme** : Par exemple, la forme actuelle des McLaren est bien meilleure que ce que prévoyait le modèle, reflétant une inertie excessive dans l'ajustement aux nouvelles tendances.
3. **Difficulté à intégrer les facteurs conjoncturels** : Les événements ponctuels comme une météo stable, une stratégie de pneus optimale ou une absence d'incident en piste peuvent fortement impacter les résultats mais sont difficiles à anticiper sans données très granulaire.
4. **Traitement approximatif des données manquantes** : L'imputation par la médiane utilisée pour certains pilotes (notamment les rookies ou remplaçants) peut induire un biais de sous-performance que le modèle amplifie.

Ces constats soulignent l'intérêt de compléter le modèle avec des variables contextuelles supplémentaires (comme les évolutions aérodynamiques par équipe, les données télémétriques ou la météo précise le jour de course), ainsi que d'envisager une calibration dynamique des poids attribués à l'historique récent des performances.

12 Conclusion générale

Ce travail s'inscrit dans une démarche itérative d'exploration, de modélisation et d'interprétation autour d'un défi aussi passionnant que complexe : anticiper les résultats de courses de Formule 1 à l'aide d'outils d'apprentissage automatique. En mobilisant des données techniques fines, issues de multiples éditions de Grands Prix, et en mettant en œuvre des algorithmes de régression avancés tels que le Gradient Boosting, nous avons progressivement affiné notre capacité à saisir les dynamiques de performance au sein d'un environnement sportif hautement instable.

Au fil des prédictions, de l'Australie à l'Autriche, en passant par la Chine, Monaco ou encore Bahreïn, notre modèle s'est enrichi, apprenant à intégrer non seulement les performances passées, mais aussi des facteurs contextuels tels que les conditions météo, les spécificités des circuits ou encore les trajectoires propres à chaque pilote. Si certaines prédictions se sont révélées très proches de la réalité, en particulier sur des circuits comme Suzuka ou Monaco, d'autres écarts ont mis en lumière les limites inhérentes à toute tentative de modélisation dans un domaine où l'aléa, l'humain et l'innovation technique jouent un rôle prépondérant.

Ce travail a surtout permis de démontrer que la modélisation prédictive, lorsqu'elle est pensée de manière progressive et contextualisée, peut offrir un regard neuf sur le sport automobile. Elle constitue non seulement un outil d'anticipation, mais aussi un levier d'analyse stratégique, capable de révéler les déterminants cachés des succès ou des contre-performances.

Enfin, cette étude pose les bases d'un dialogue entre science des données et expertise sportive. Elle montre que si les algorithmes ne remplacent pas l'intuition humaine, ils peuvent en être de précieux alliés, à condition d'être construits avec rigueur, enrichis en permanence, et interprétés avec discernement. C'est dans cette articulation entre technique, compréhension du domaine et esprit critique que réside, selon nous, la puissance de l'analyse prédictive appliquée à la Formule 1, et au-delà.

Références

[1] Antaya, M. (2023). *2025 F1 predictions*. GitHub. Consulté en juin 2025. https://github.com/mar-antaya/2025_f1_predictions

[2] Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. *In Proceedings of COMPSTAT'2010*, pp. 177–186.

[3] Bordet, C. (s.d.). *Gradient Descent*. CharlesBordet.com. Consulté en juin 2025. https://www.charlesbordet.com/fr/gradient-descent/#comment-regler-le-probleme-du-vanishing-gradient

[4] Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge : Cambridge University Press.

[5] Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

[6] CNIL. (2024). *Introduction aux algorithmes d'apprentissage automatique et à leurs hyperparamètres*. Consulté en juin 2025. https://www.cnil.fr/fr/definition/taux-dapprentissage-learning-rate

[7] Datascientest. (s.d.). *Descente de Gradient*. Consulté en juin 2025. https://datascientest.com/descente-de-gradient

[8] Friedman, J. H. (2001). Greedy Function Approximation : A Gradient Boosting Machine. *Annals of Statistics*, **29**(5), 1189–1232.

[9] Garcia Tejada, L. (2023). *Applying machine learning to forecast Formula 1 race outcomes* [Mémoire de master, Aalto University, School of Science]. Aalto University Repository.

[10] IBM Cloud Education. (2023). *What Is Gradient Descent ?* Consulté en juin 2025. https://www.ibm.com/cloud/learn/gradient-descent

[11] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM : A Highly Efficient Gradient Boosting Decision Tree. *In Advances in Neural Information Processing Systems*, 3149–3157.

[12] Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization* (2e éd.). New York : Springer.

[13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost : Unbiased Boosting with Categorical Features. *In Advances in Neural Information Processing Systems*.

[14] Rudin, W. (1991). *Functional Analysis* (2e éd.). New York : McGraw-Hill.

[15] Wikipedia contributors. (2025). *Algorithme du Gradient*. Wikipédia. Consulté en juin 2025. https://fr.wikipedia.org/wiki/Algorithme_du_gradient

[16] Wikipedia contributors. (2025). *Gradient boosting*. Wikipédia. Consulté en juin 2025. https://fr.wikipedia.org/wiki/Gradient_boosting