DS 670: Fall 20255

Hamish Henderson

Predicting Lap Times in Formula One

This report presents a data driven analysis and modeling approach for predicting the fastest lap times in Formula 1 races for the 2025 season. It follows a structured methodology including data acquisition, preprocessing, model design, evaluation, and suggestion for future work.

1. Introduction

Predicting performance in Formula 1 has always been a challenge due to how many factors can influence a race; driver ability, car setup, tire choice, weather, and even track evolution throughout a session can all play a role. This project explores how well machine learning models can forecast the fastest lap times for a race by using real F1 data. With the help of the FastF1 API, historical lap times, sector splits, qualifying results, weather details, and team information were collected and combined into a dataset that captures both driver performance trends and circuit characteristics. The goal is to understand what impacts lap time the most and to see if a model can accurately predict how drivers will rank on different tracks.

To do this, several machine learning models were built and tested, including Gradient Boosting, XGBoost, Support Vector Machines, a neural network, and a stacked ensemble made from all four. These models were evaluated using common metrics like Mean Absolute Error to measure how close their predictions were to real lap times. By comparing results across a few different races, the project shows how each model handles different circuit styles and race conditions. Overall, this work not only predicts fastest laps but also provides insight into which features matter most and how data driven methods can support racing analysis and strategy.

2. Data

This project uses official Formula 1 timing and telemetry data obtained through the FastF1 API, which provides detailed session level information for each Grand Prix. The dataset includes lap by lap timing, driver and team identifiers, tire compounds, sector splits, and additional session metadata such as track conditions and session type. For each event, the analysis pulls data across multiple seasons to build a consistent and diverse historical dataset for model training.

To support prediction for a selected race in the 2025 season, qualifying and race session data were collected from previous years at the same circuits. Additional contextual information; such as team performance, weather conditions, and driver lineup change was also incorporated to capture broader factors that influence lap time. Data from the 2025 season, including prior race results, practice sessions, and earlier qualifying outcomes, was included to reflect current driver and car performance. All sources were merged into a single modeling dataset that links lap times with driver characteristics, car performance factors, circuit-specific variables, and environmental conditions.

Data Cleaning and Feature Engineering

To prepare the dataset for modeling, several feature engineering steps were applied to transform raw race data into meaningful predictive variables. These engineered features were designed to capture driver performance, tire behavior, session conditions, factors which could influence fastest lap times.

FastF1 provides lap times as timedelta objects, so all lap and sector times were converted into **milliseconds** to create a numerical target variable suitable for machine learning models.

Tire compounds (Soft, Medium, Hard, Inters, Wets) were transformed into categorical codes to reflect performance differences between compounds. Since compounds strongly influence lap time, this feature could play an important role in predictions. Normally a soft compound tire is the fastest, but it can depend on the conditions.

Each lap was linked to the driver's three-letter code, race number, and team name. These identifiers were encoded to help models learn consistent performance trends linked to specific drivers and cars.

Information about the session type (e.g., Practice, Qualifying, Race), lap number within the session, and remaining laps was added. These features could potentially allow the model to capture how performance changes as fuel loads drop or track conditions evolve.

Each record was tagged with the event name, circuit, and year to support cross event modeling and ensure that laps from different seasons remain distinguishable.

In laps, out laps, pit laps and laps with missing or incomplete data were removed. This ensures the model learns from laps that reflect competitive pace rather than pit entries, cool down laps, or aborted attempts. Essentially removing any outliers.

Together, these engineered features create a robust dataset that captures the key variables influencing lap time performance and enables the machine learning models to generalize across drivers, circuits, and seasons.

Models

To predict the fastest lap times for each Grand Prix, four individual machine learning models were trained and evaluated, along with a stacking ensemble model that combines their outputs. These models were chosen to represent a broad range of learning strategies, from tree based methods to kernel methods and neural networks. The models were selected to compare their strengths, weaknesses, and overall predictive performance. The models included a Gradient Boosting Regressor (GBR), an XGBoost Regressor (XGB), a Support Vector Regressor (SVM), and a Neural Network Regressor (NN).

A GridSearch process was used for each model to identify the best performing hyperparameters. After evaluating the individual models, a Stacking Regressor Ensemble was created to combine their predictions with the goal of further improving overall accuracy.

Once the models were trained, predictions were generated for each driver on the selected circuit. These predictions produced two key outcomes: the estimated fastest lap time for each driver and a predicted starting grid order. The grid position was determined by ranking the predicted lap times, with the fastest predicted time placing highest on the grid.

Model Results

When possible, the actual results from the 2025 races were loaded to compare against the model predictions. Since not every race had occurred at the time of writing, this comparison was only available for certain events. To evaluate the models, I created charts that display the predicted fastest lap time for each driver from each model and compared these values to the real fastest laps when available.

Several results from races are included at the end of this report. The first chart for each race shows all model predictions plotted against the actual 2025 results. At first glance, most drivers predicted times aligned reasonably well with the real outcomes, but a few drivers consistently appeared much slower. After investigating this pattern, I found that these spikes were caused by rookie or new drivers. Because they have limited historical data, the models had far less information to learn from, leading to noticeably less accurate predictions for these drivers.

To better evaluate model performance for the more established drivers, I created a second chart where rookies were removed. This made the comparison clearer and showed that most predictions were within one to two seconds of the actual fastest laps. In some cases, however, the real 2025 times were significantly slower, by as much as 20 seconds. Which likely reflects race day weather conditions. For example, wet sessions naturally produce slower lap times, and the model cannot predict this precisely without detailed weather forecasts.

Next, I created a bump chart. The chart provides a visual comparison of how each model ranked the drivers relative to the actual grid order for the selected Grand Prix. Each line represents a single driver, and the movement of the line across the chart shows how that driver's predicted qualifying position changed across the different models. Drivers whose lines remain relatively flat indicate consistent predictions across models, while larger jumps highlight disagreements or uncertainty in the models' estimations. As seen in the chart, most top drivers maintain similar positions across all models, while mid field and rookie drivers show more variation.

I then created a heatmap which illustrates the rank error for each model compared to the actual qualifying positions for the selected race, with positive values indicating that a model placed a driver lower (worse) than their real position and negative values showing where a driver was predicted higher than they actually qualified. This visualization highlights not only how far off each model was for each driver, but also where consistent patterns emerge across models. As shown in the chart, most top drivers exhibit relatively small errors, demonstrating strong predictive alignment near the front of the grid. In contrast, the largest errors appear for rookies, often out performing the models expectation for them.

To summarize model performance, I created a bar chart showing the Mean Absolute Rank Error (MARE) for each model. Across the races analyzed, most models produced a MARE between three and five positions, meaning the predicted grid positions were typically off by about three to five places from the real qualifying results. While higher than ideal, part of this error is likely influenced by rookie drivers, who often performed significantly differently than their limited historical data would suggest. The best performing model varied by race; in several cases the Neural Network delivered the lowest MARE, although this was not universal. For example, the British Grand Prix showed the NN performing worst. The stacked ensemble model tended to be the most consistent overall, generally providing solid mid-range predictions without extreme errors.

Future Work and Conclusion

Overall, this project provided a substantial amount of valuable insight into how machine learning can be used to analyze and predict performance in Formula 1. One of the biggest challenges that emerged was the lack of historical data for rookie drivers, which consistently impacted prediction accuracy. A potential solution for future iterations

would be to incorporate performance data from Formula 2, since many rookies have extensive F2 careers on the same circuits, even if in less powerful cars. Expanding the dataset in this way would likely lead to more reliable and meaningful predictions.

For future work, I would like to build a similar modeling pipeline that predicts finishing positions for an entire race, not just fastest laps. This was part of the original plan, but limitations with the API (the larger volume of data required) made it difficult to implement within the scope of this project. Additionally, I would like to develop a more user friendly interface for selecting races and visualizing the results, ideally through an interactive online dashboard. This would make the analysis more accessible and allow users to explore race predictions and model behavior in real time.

In conclusion, this project demonstrated that machine learning can effectively capture key patterns in Formula 1 performance and generate reasonably accurate predictions of fastest lap times. While challenges remain, especially regarding limited data for rookie drivers, the models showed consistent potential across multiple circuits. With expanded datasets and further refinement, this approach could support more reliable race forecasting and deeper performance analysis in future work.

Sources:

FastF1. (2024). FastF1 Documentation. Version 3.3. https://docs.fastf1.dev/index.html

Géron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

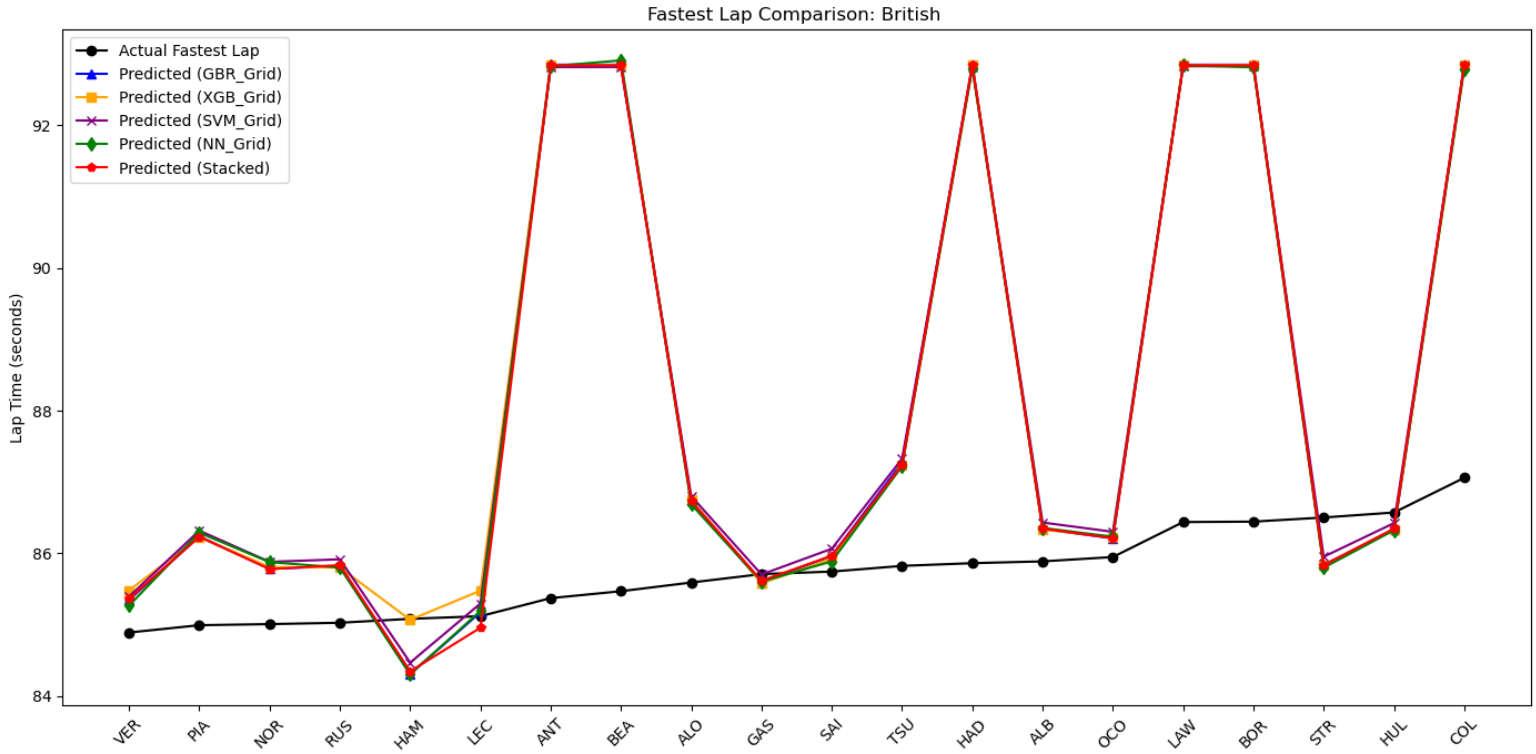
OpenAI. (2025, November 11). ChatGPT (GPT-5) [Large language model]. OpenAI. https://chat.openai.com

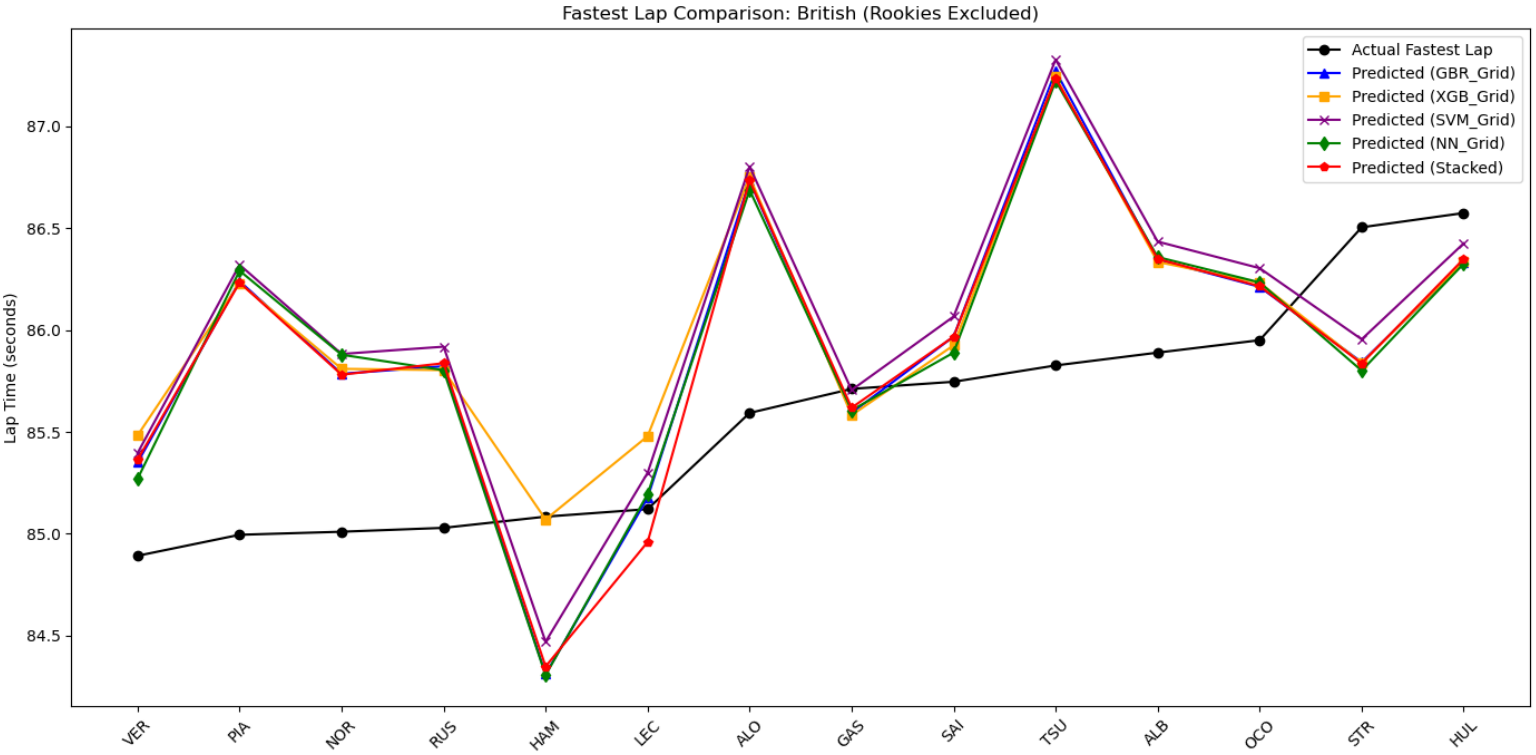
Saint Peter's University. (2025). DS 670: Capstone Department of Data Science, Saint Peter's University, Jersey City, NJ.

Saint Peter's University. (2025). DS 630: Machine Learning. Department of Data Science, Saint Peter's University, Jersey City, NJ.

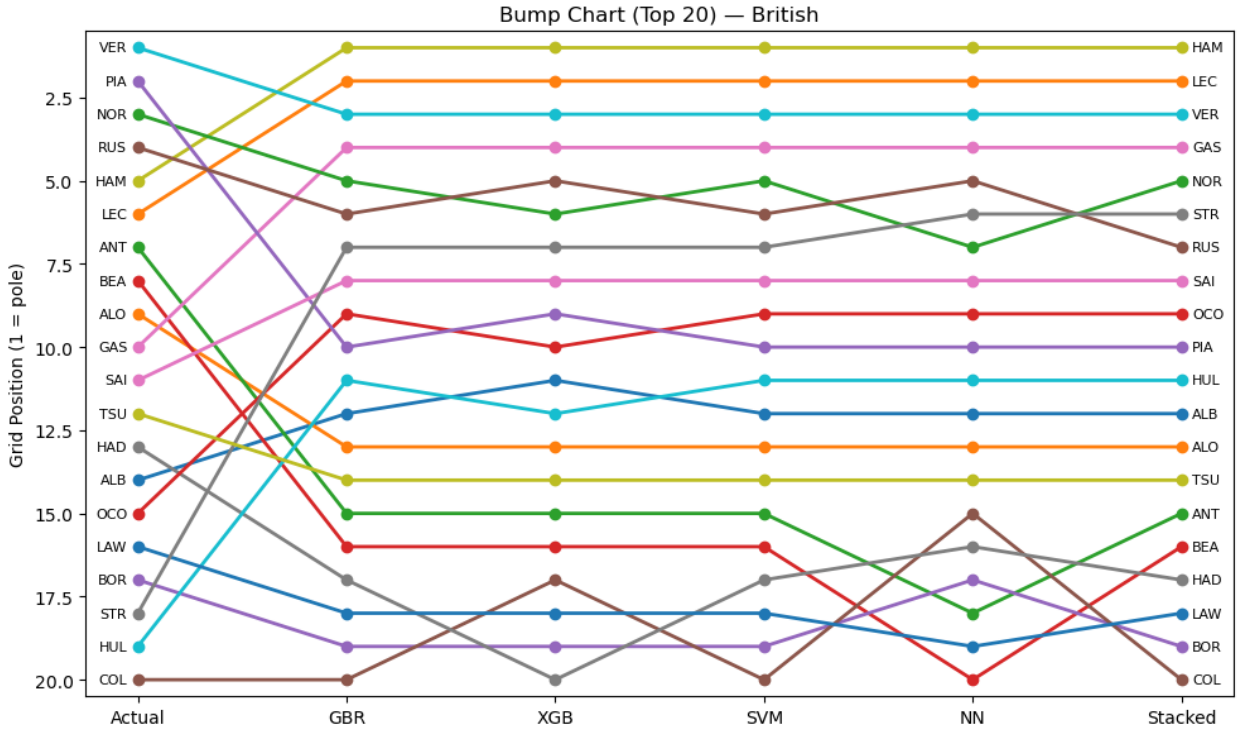
Scikit-learn Developers. Scikit-Learn API Reference. scikit-learn, https://scikit-learn.org/stable/api/sklearn.html. Accessed Nov. 2025.

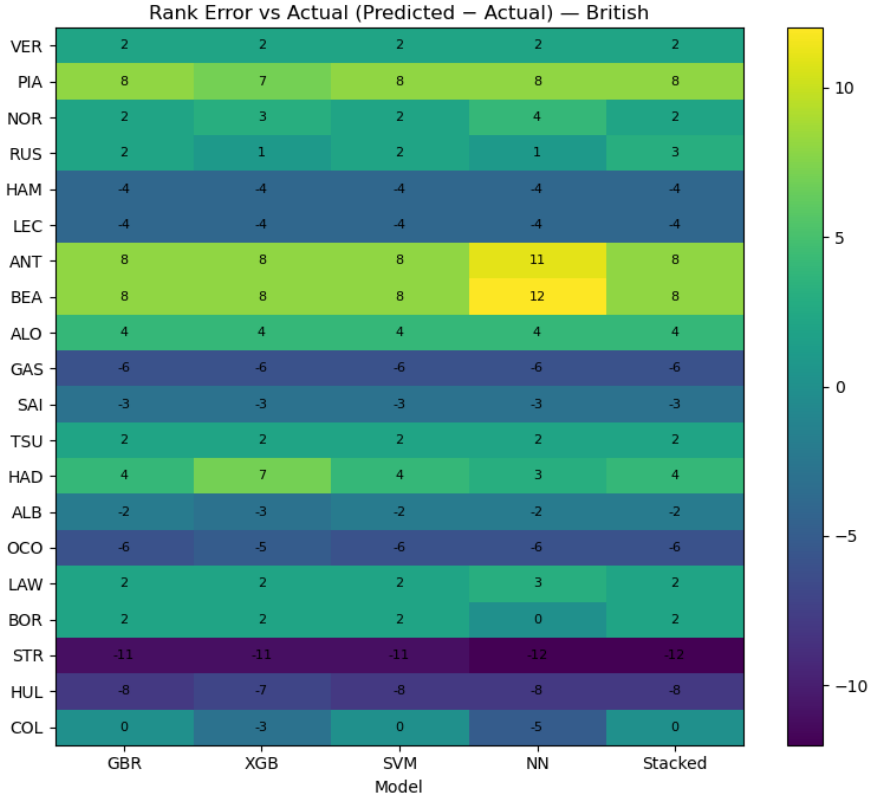
British Grand Prix Results

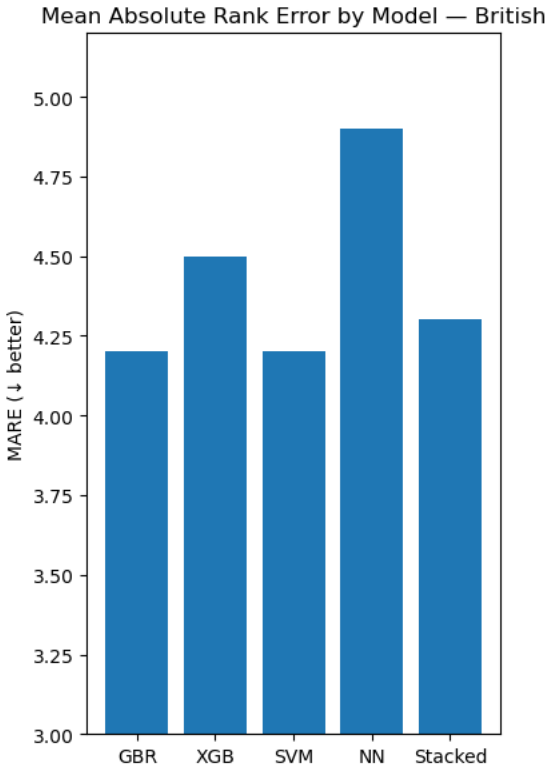




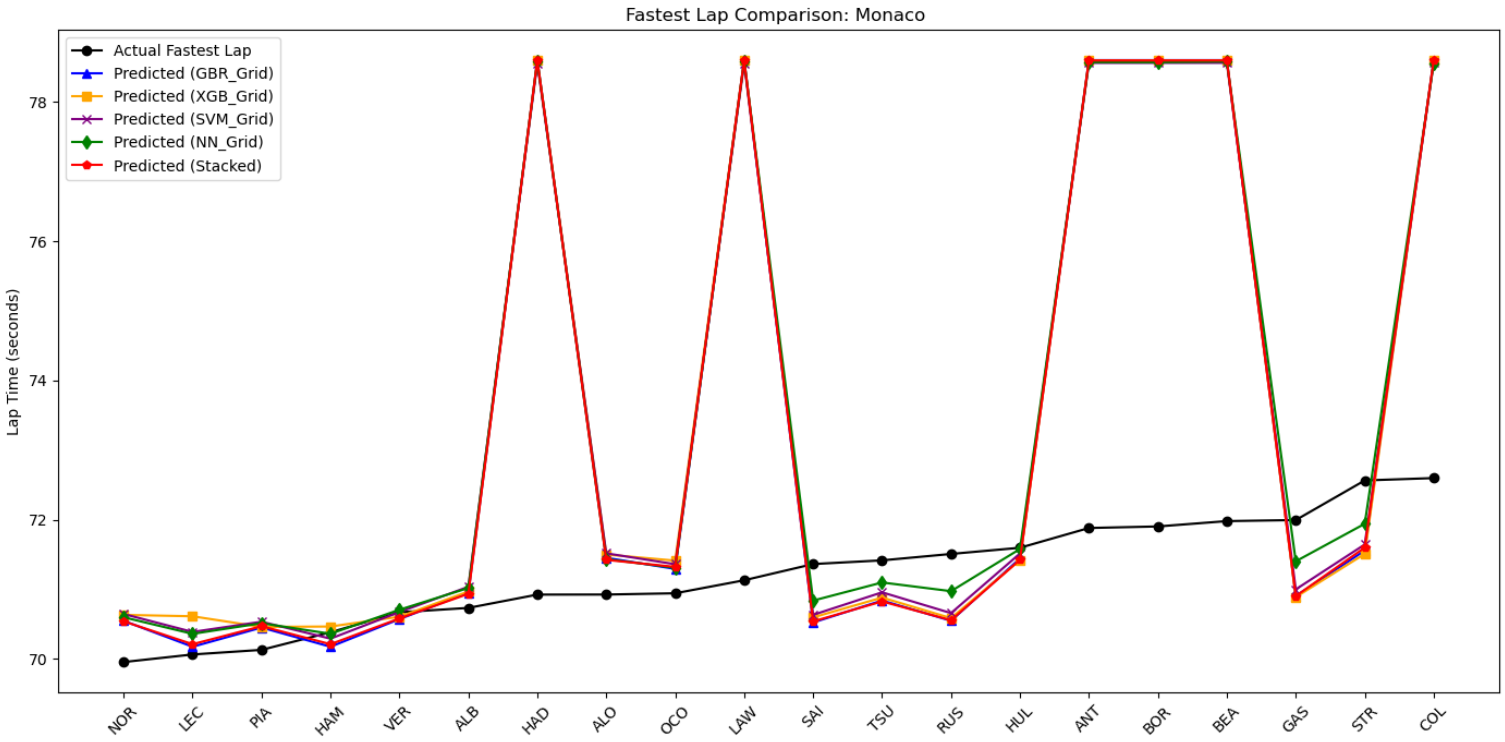
British GP Results Cont.

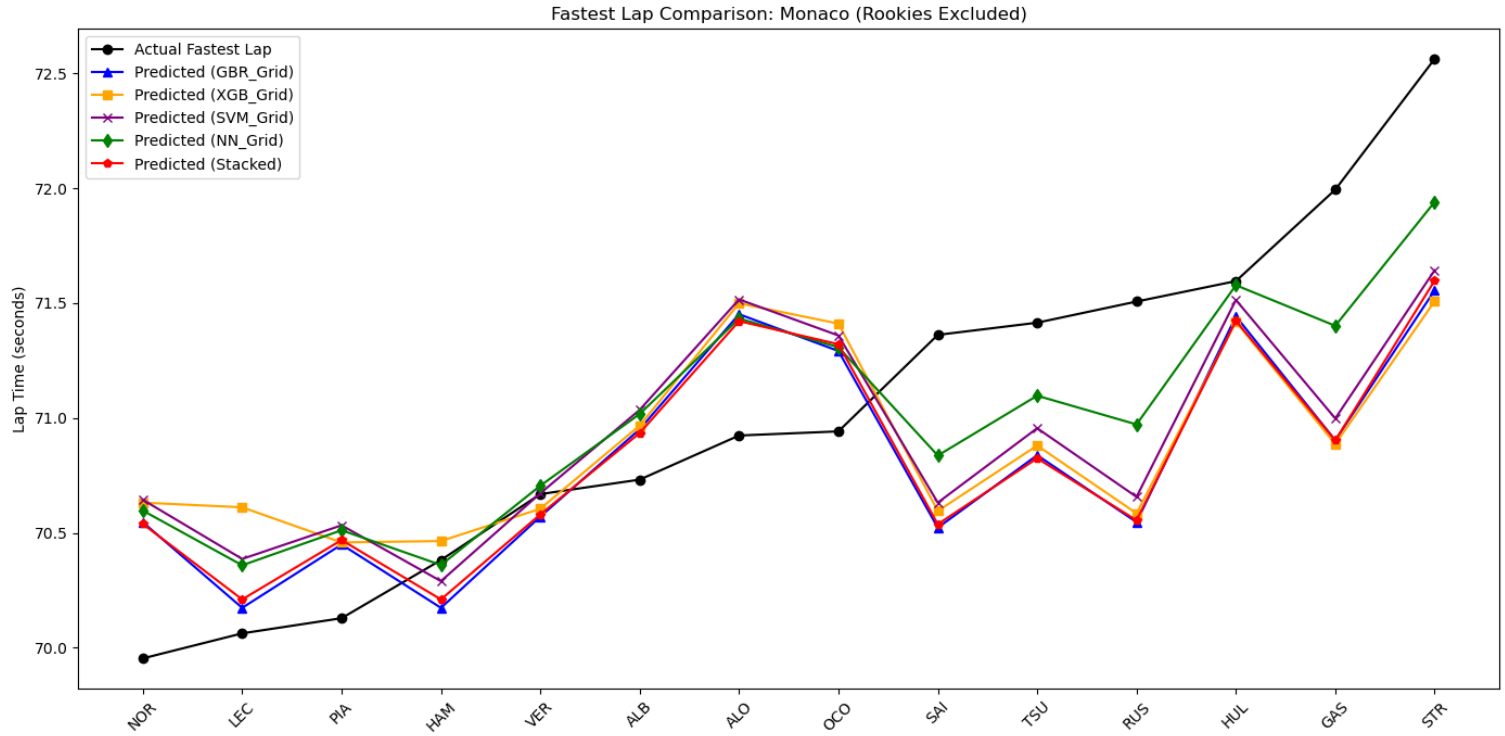




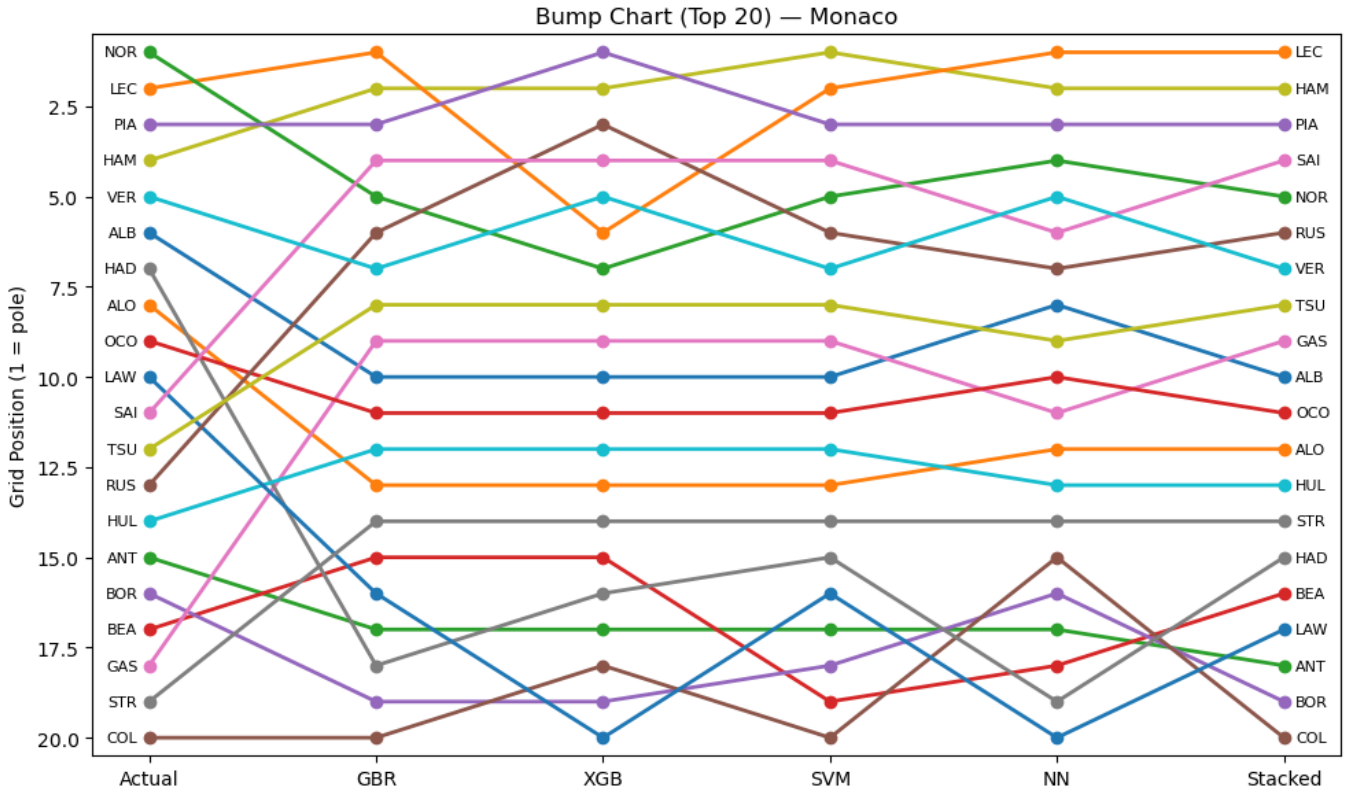


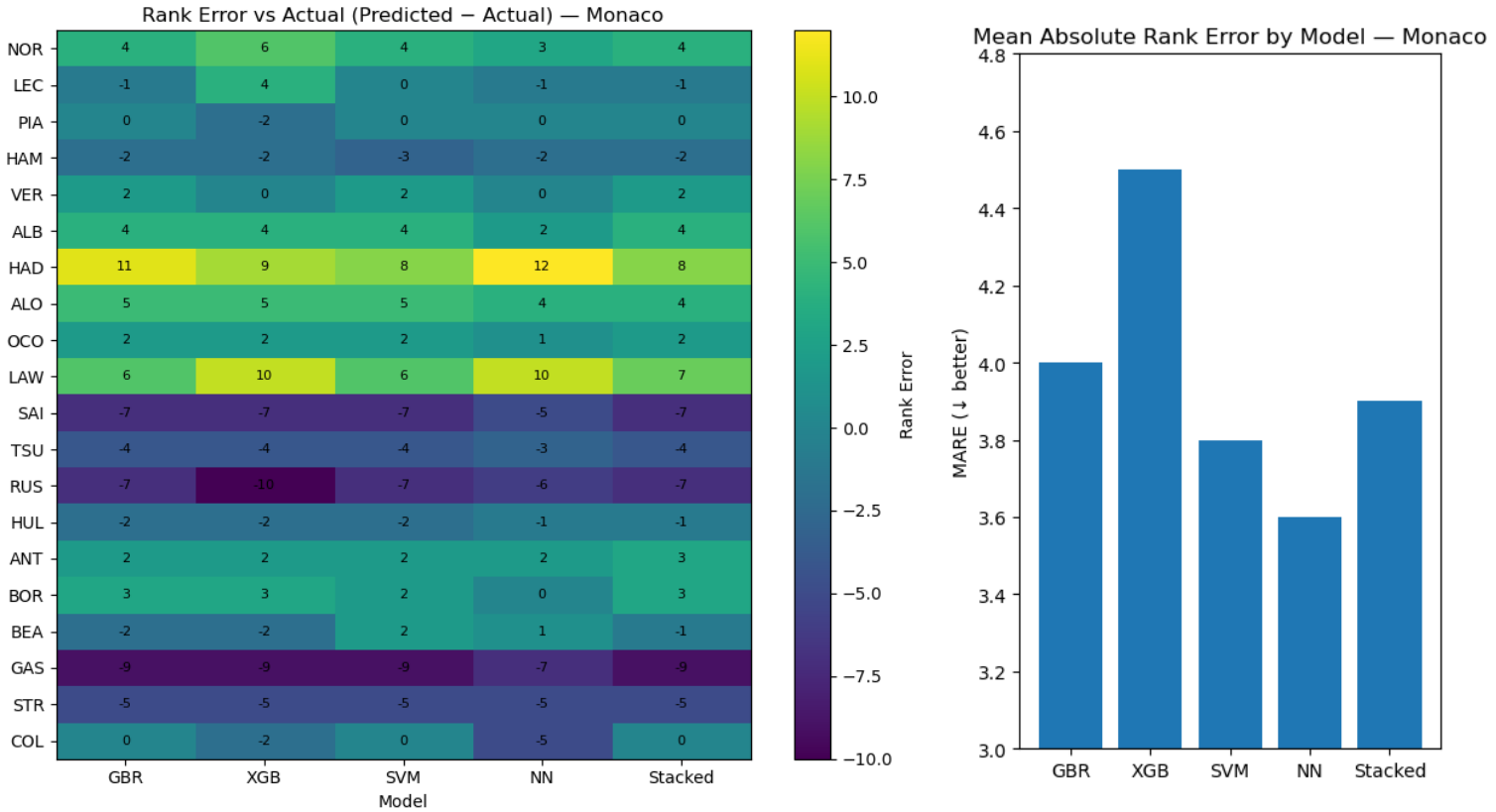
Monaco Grand Prix Results



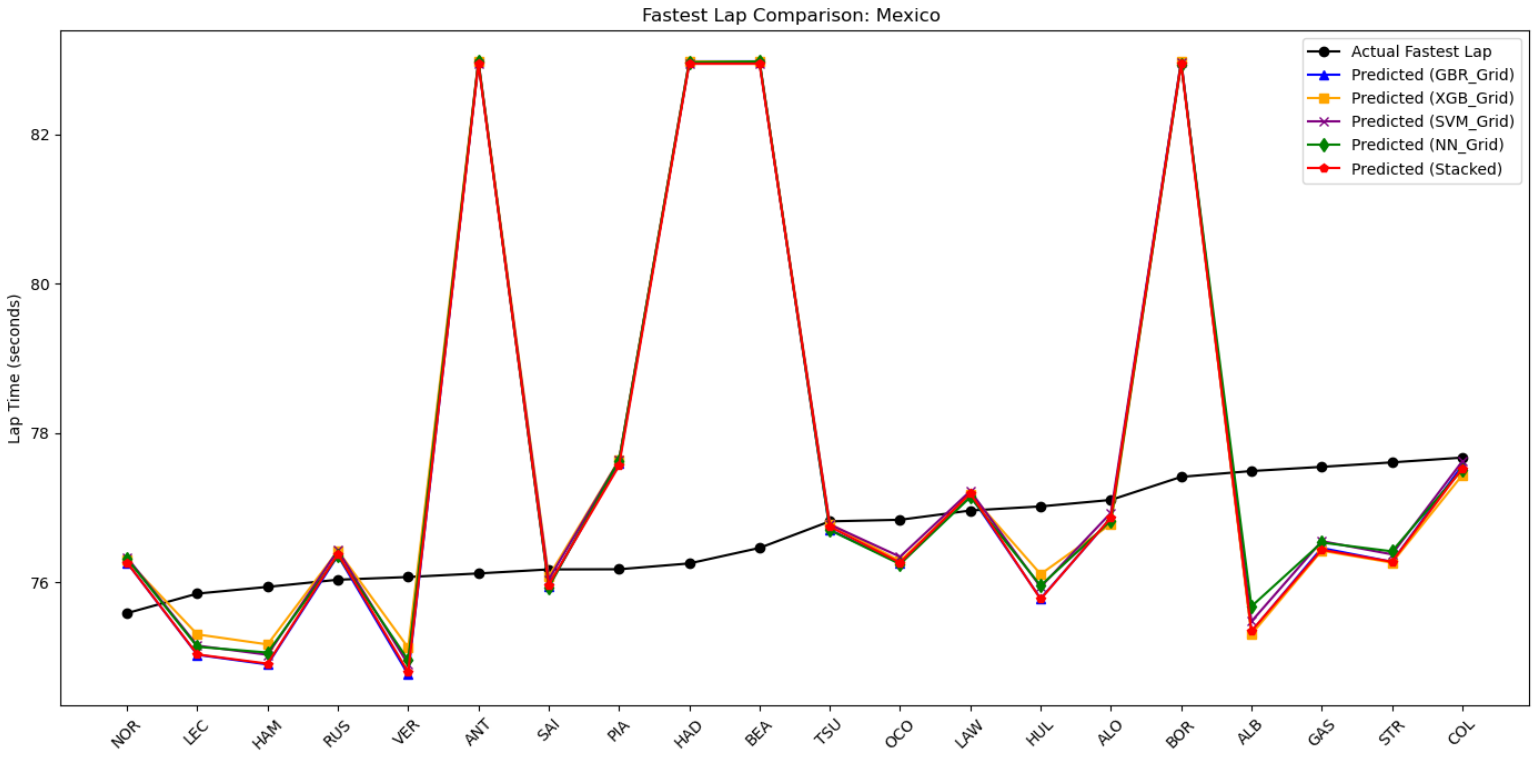


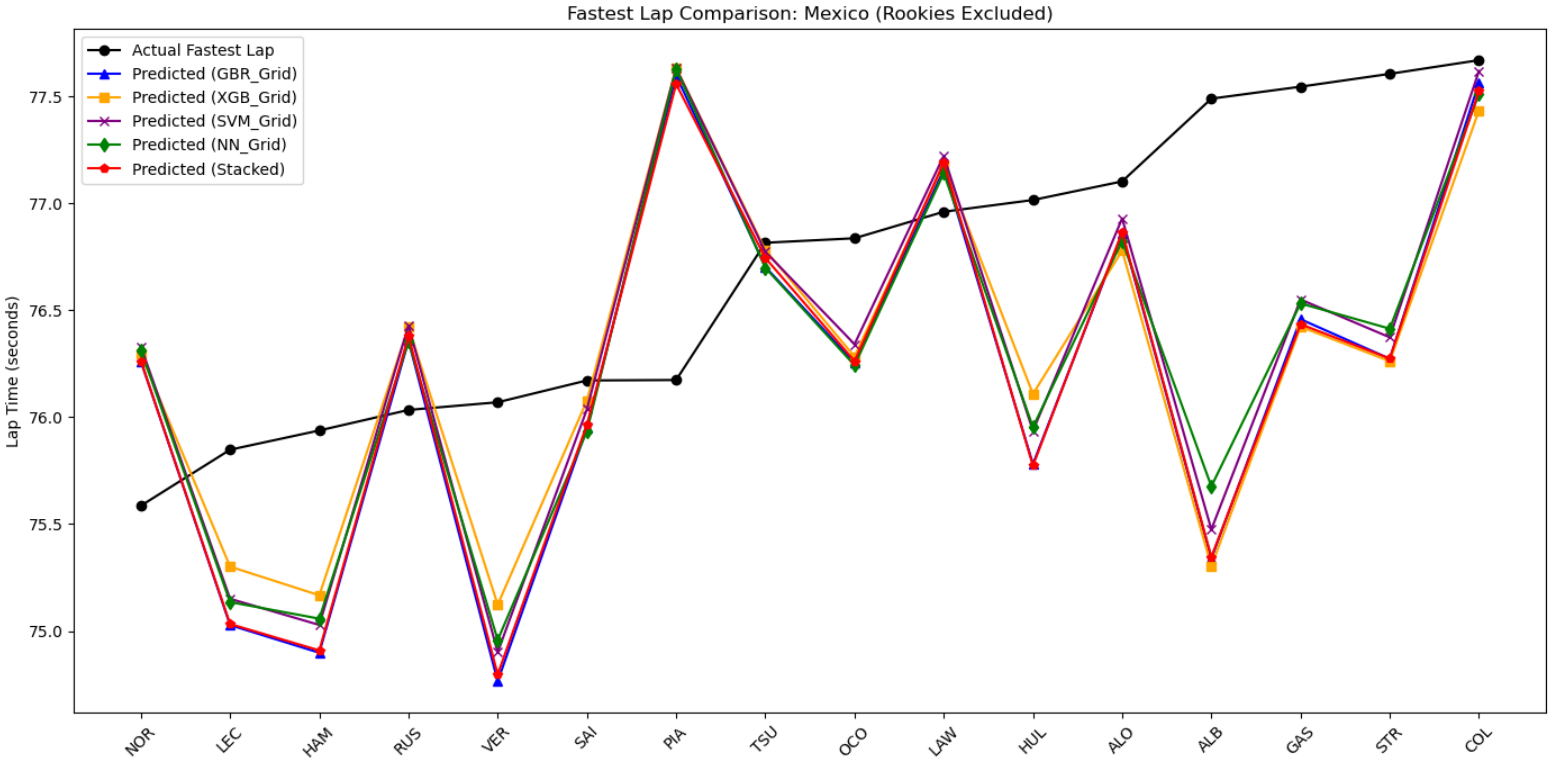
Monaco Grand Prix Results Cont.



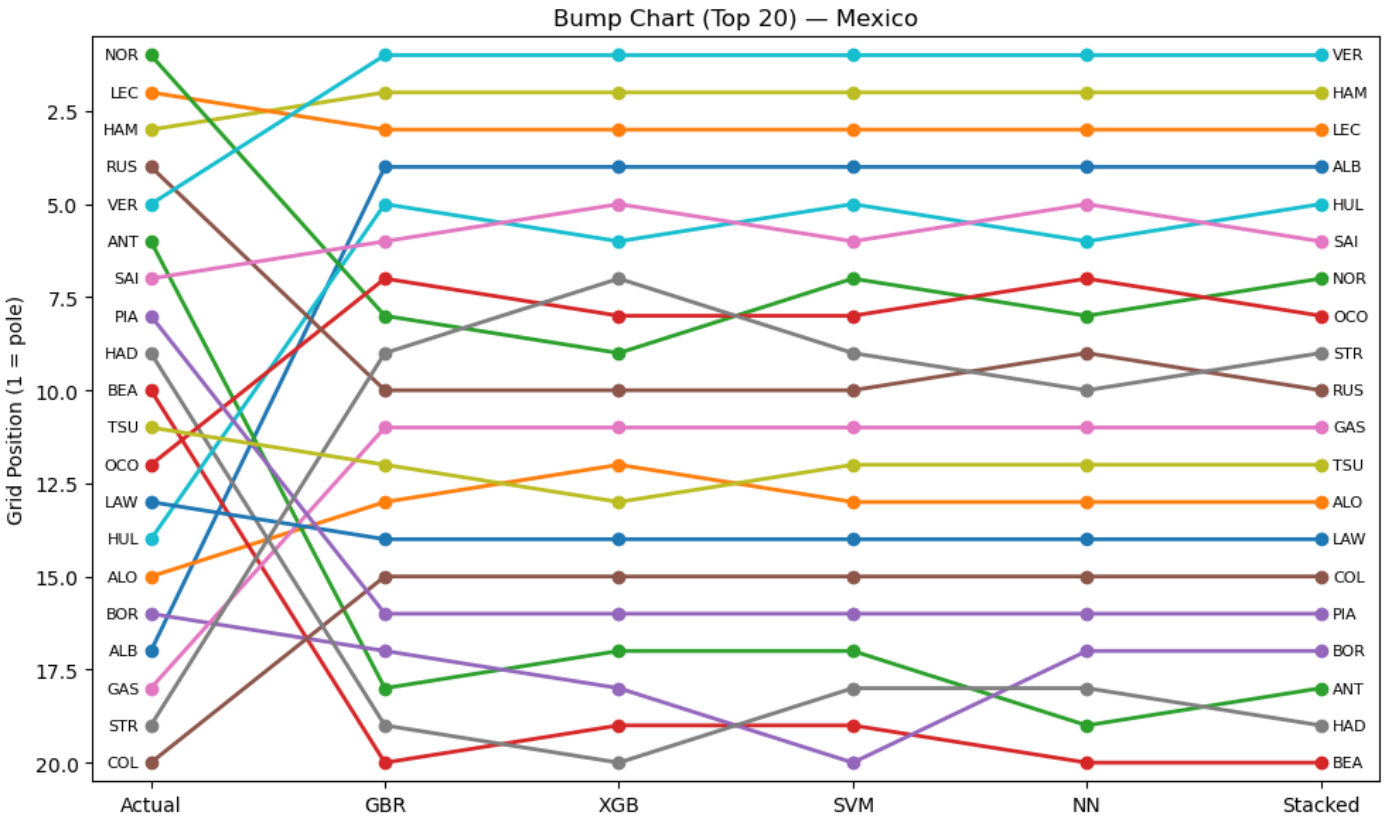


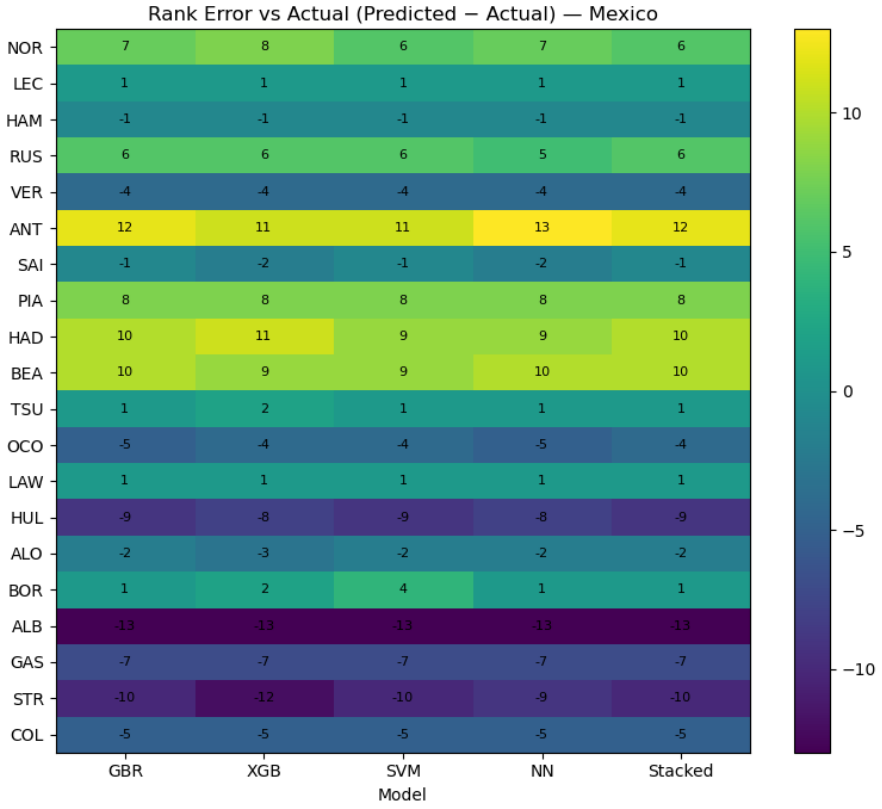
Mexico Grand Prix Results

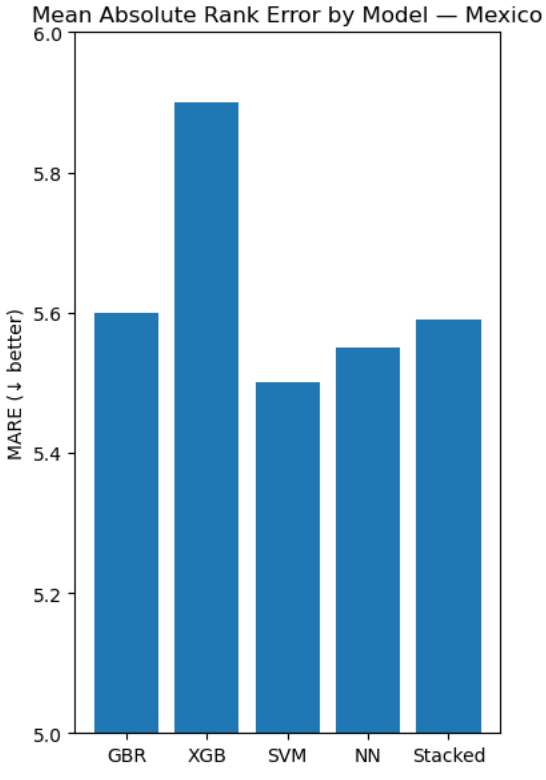




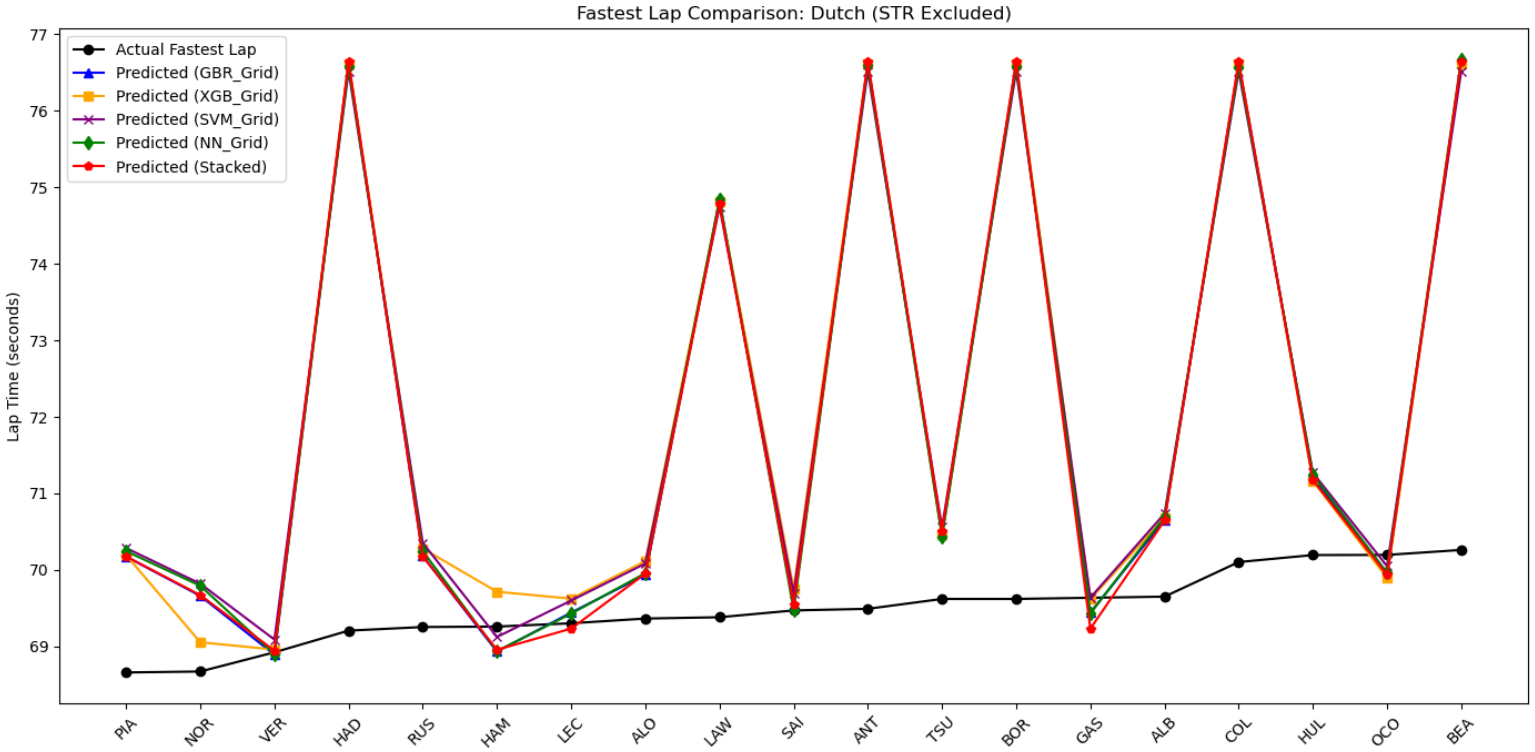
Mexico Grand Prix Results Cont.

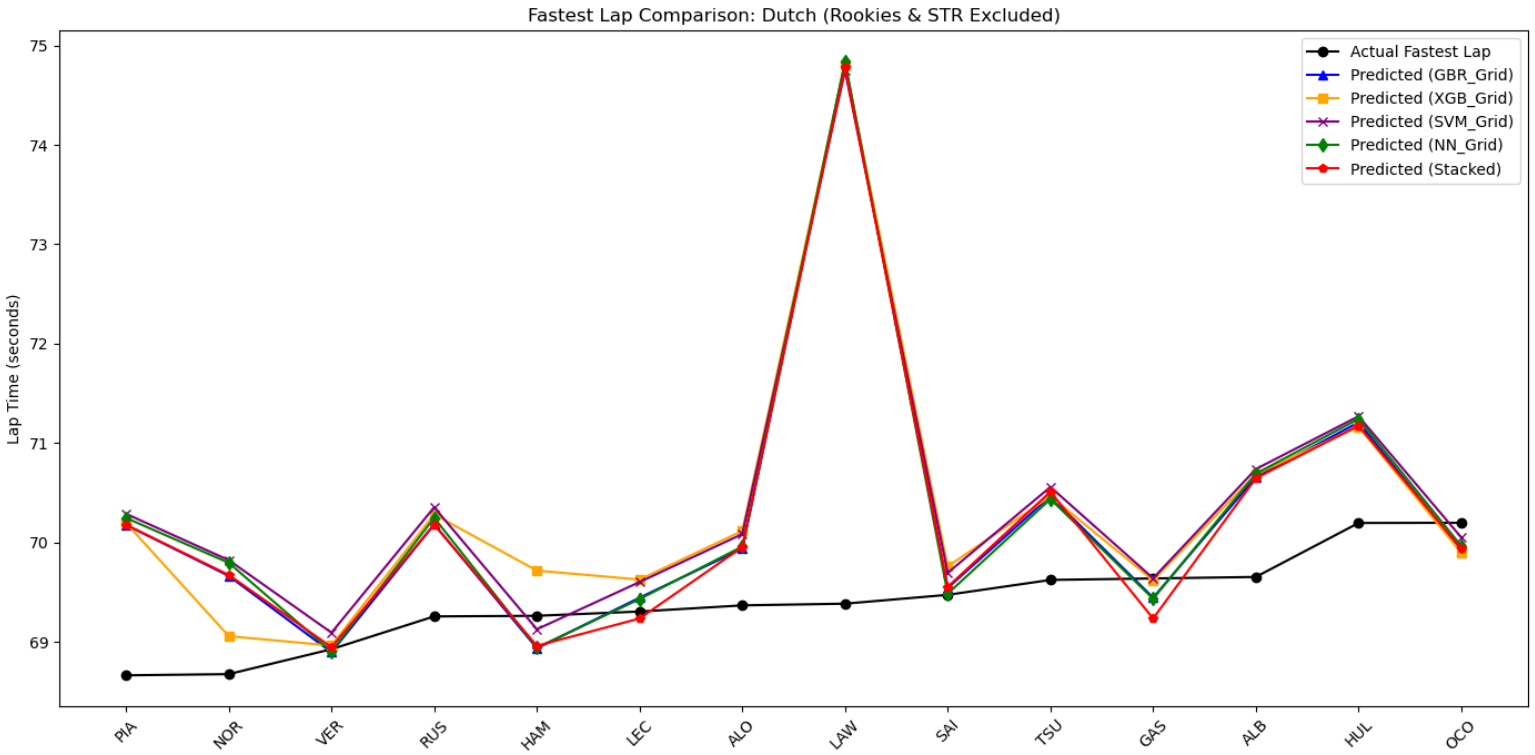






Dutch Grand Prix Results (STR was removed due to crash)





Dutch Grand Prix Results Cont.