World Cup Business LLM Forecast

Invest Atlanta | AI.Data Lab 2025

2025

- Mandy Chen, mandy.chen@emory.edu
- Sherelle Li, sherelle.li@emory.edu
- Madeleine Pelli, madeleine.pelli@emory.edu
- Daniel Cardenas, daniel.cardenas@emory.edu

More Info: https://github.com/sherelleli/aidatalab_g6

Problem Space

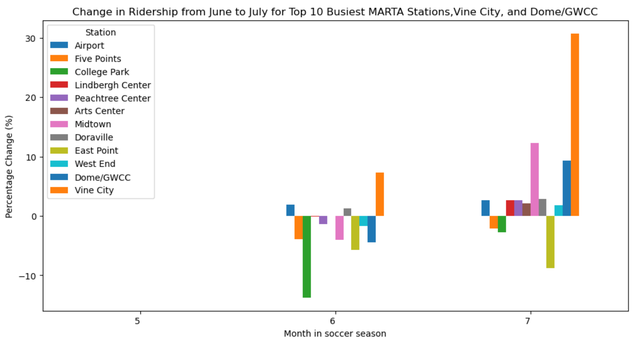
We want to help business owners in the Atlanta area understand the projected foot traffic and gauge its impact on their own business in order to accommodate the uptick. We wanted to know, based on ridership and business patterns from past major events, which Atlanta businesses are likely to see the largest increase in public transit use and potential customer traffic during the 2026 World Cup. We used our findings to create an LLM trained on our findings for business owners to query to provide specific impact data for their situation.

Methodology

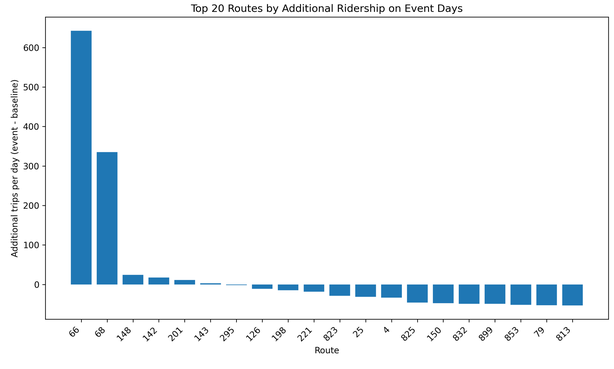
Major data sources:

- Business license data from the City of Atlanta
- MARTA bus and train ridership data from 1/1/23 to 8/31/25
- A meta-dataset of major sports events (Falcons, Atlanta United, SEC Championships, etc.) created using external sources

- Our goal is to link the event meta-dataset to corresponding changes in MARTA ridership and use these relationships to model the projected impact of the World Cup.

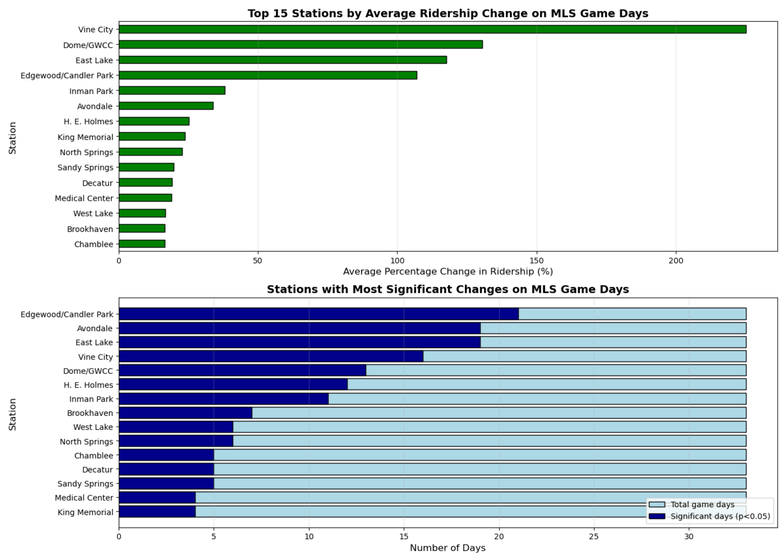


- The graph compares changes in ridership from June to July — the months when most major soccer events, including the 2026 World Cup, are hosted — across 10 MARTA stations and 2 stations near Mercedes-Benz Stadium.
- Dome/GWCC and Vine City have lower overall ridership than most stations, but both show substantial growth from June to July, with especially sharp increases at Vine City. A likely explanation is increased public transit use for attending soccer games.
- Consistent with this pattern, routes 66 and 68 also show notable ridership increases during similarly sized sporting events at Mercedes-Benz Stadium.



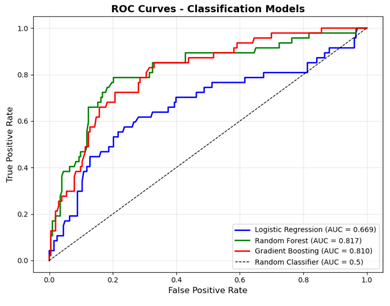
Metrics

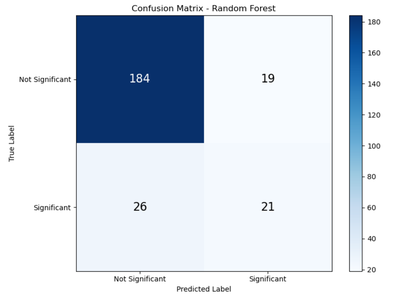
- "Success" meant that a model could effectively predict itself — accurately reproducing the historical data it was trained on, which increases confidence in its future forecasting ability.
- Significant ridership changes were measured relative to a baseline from the same day of the week and same month to account for normal seasonal and weekly patterns.

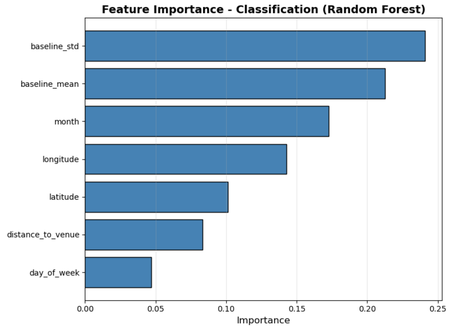


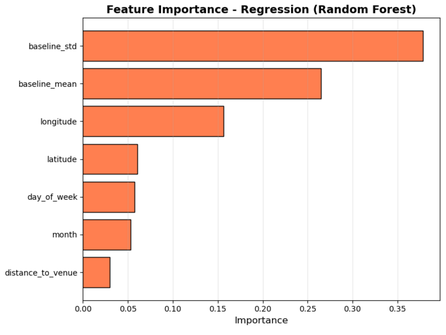
We then tested three **machine learning models** — logistic regression, random forest, and gradient boosting. Only the latter two yielded meaningful predictive performance, with AUC scores of 0.817 and 0.810, respectively. We cross validated with NBA game data to ensure gradient boosting was the best model.

Interestingly, models trained solely on 2023 data achieved a substantially higher R^2 score, reaching a precision of 88%. Even incorporating **time-series features into the random forest model did not outperform the single-year model.** The reason behind this discrepancy remains unclear, making it a valuable direction for future investigation.





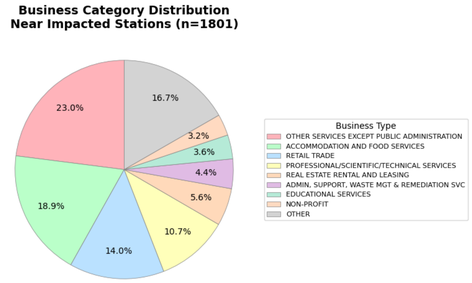




- Classification model — predicts whether ridership will experience an abnormal increase on event day
- Regression model — estimates the *amount* of additional ridership

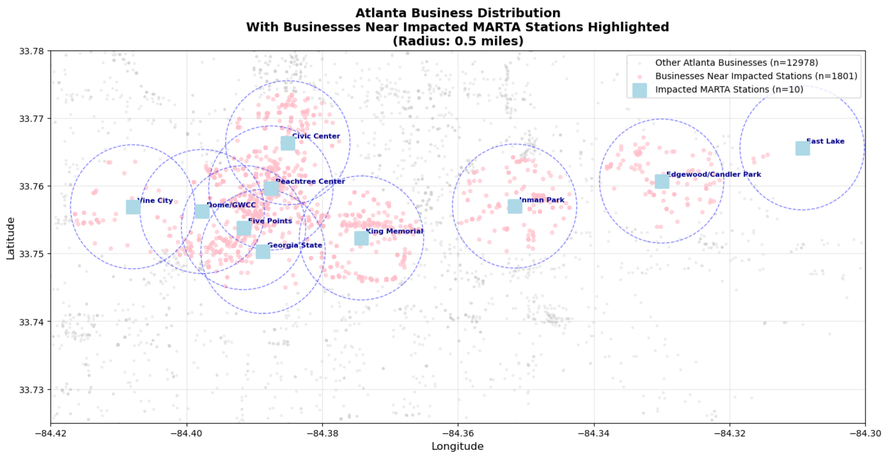
We found that the features that were the 'strongest drivers' of prediction were the variability in normal ridership and the ridership level, as well as temporal factors like the time of year (month).

Implications & Our Recommendations



Primary Recommendation: Deploy the classification model for identifying impacted stations and use the 2023-only regression model as an upper bound on predictive accuracy. Continue monitoring temporal patterns and consider year-specific models if distinct patterns persist.

We identified four key MARTA stations where businesses are likely to experience the highest increases in activity: Edgewood/Candler Park, Vine City, Dome/GW traffic during the World Cup, we acknowledge that MARTA ridership alone may underestimate tourist movement, since many visitors rely on alternative transportation. Even so, the model provides valuable guidance by highlighting which regions are most likely to be affected, helping CC and Doraville. As these areas should experience substantial foot traffic, businesses should prioritize preparation accordingly.



How this Helps Business Owners: Our Traffic-Predicting Chatbot

We used these findings to create a context-specific LLM that ingests all of our data to help any business owner prepare adequately and interpret these findings for their own business.