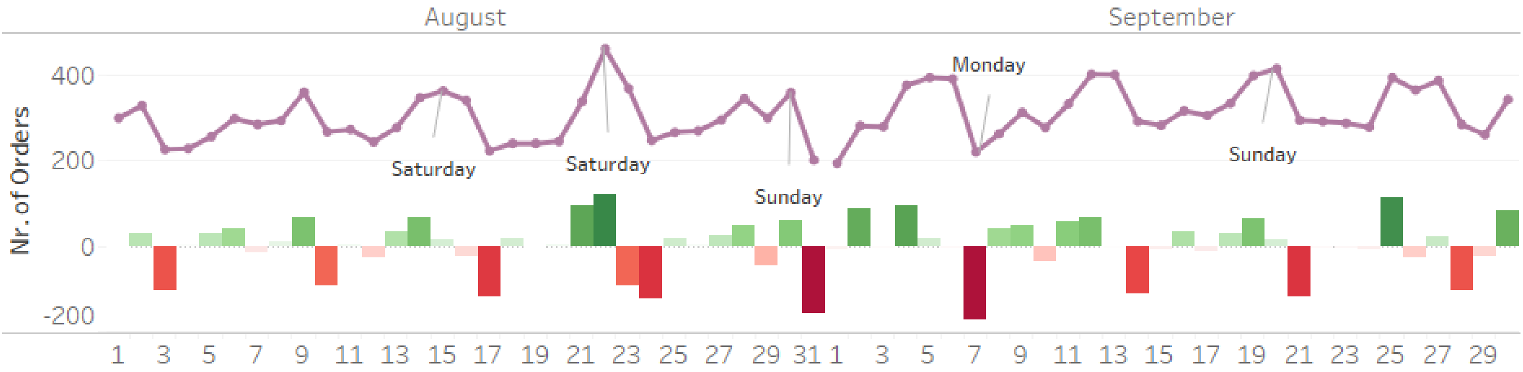
WOLT DATA SCIENTIST INTERN (2024) APPLICATION ASSIGNMENT

ENRICO SANTORO
ENRICO.SANTORO.N2@GMAIL.COM

Exploring Data and **Forecasting Delays** in Deliveries

Daily Orders

EDA PART 1

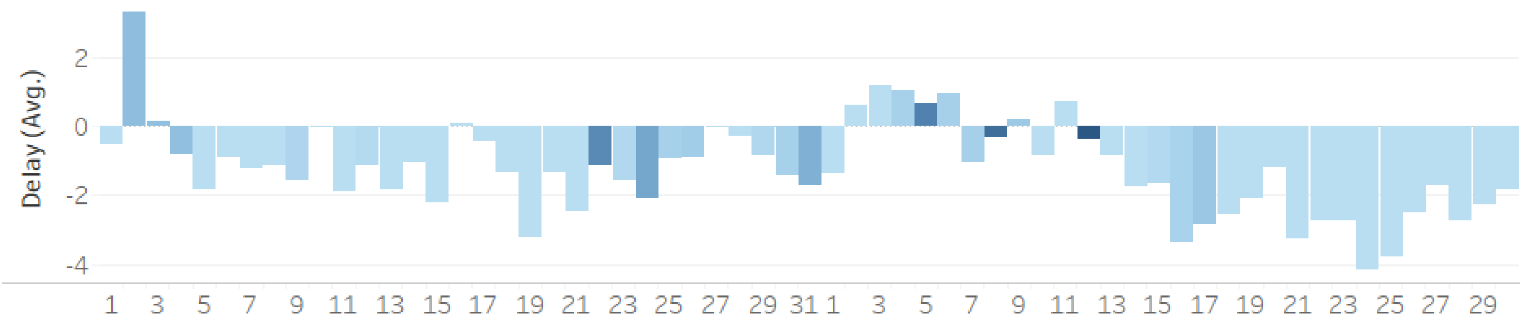


We can observe periodic patterns, due to the regular increase of orders during weekends.

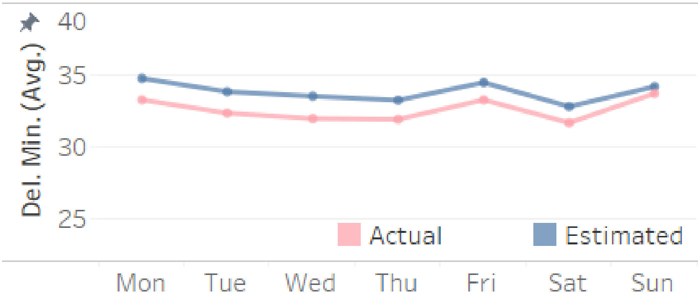
Are more orders linked with delays?
What other factors are delays affected from?

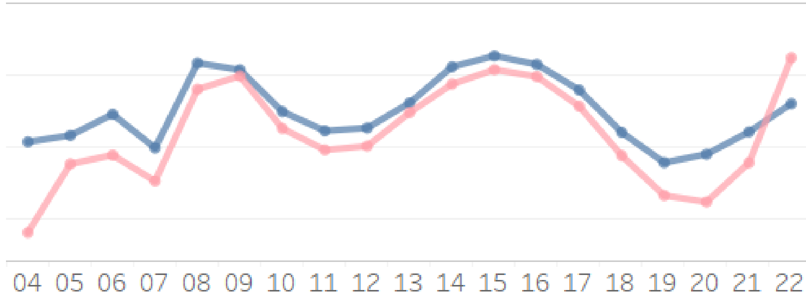
Daily Delays (Actual - Estimated Delivery Minutes)

and how they're affected by weather, weekday, hour.



Weather Condition
Rain (mm) 0 3.288





On average, **Estimated Delivery time is lower than the Actual one.** The first one is correctly adjusted during busy hours.

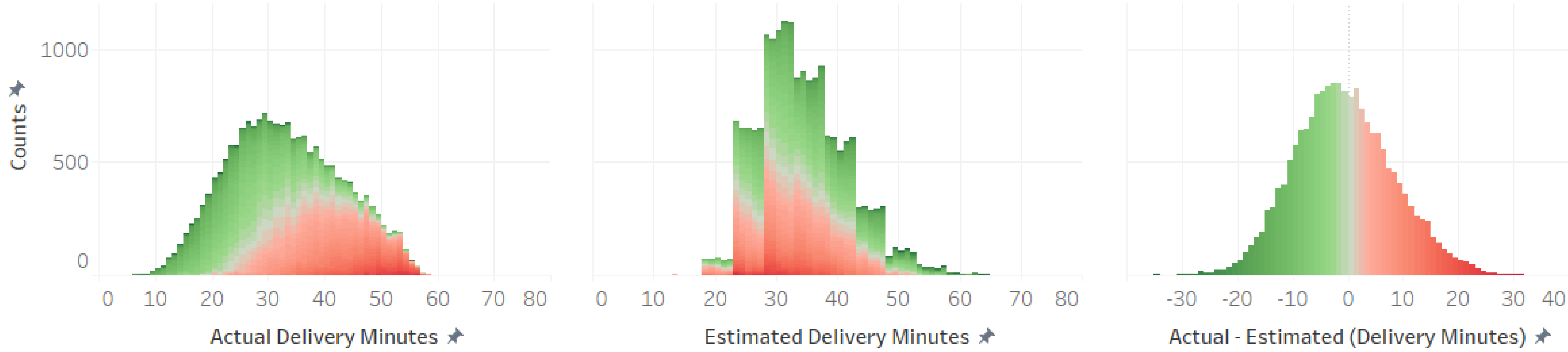
Delivery time doesn't seem to be strongly influenced by weather or weekday.

Interactive Dashboard at this [link]

A Survey in Delays

Looking at Actual Delivery Minutes and Estimated Delivery Minutes distributions

EDA PART 2

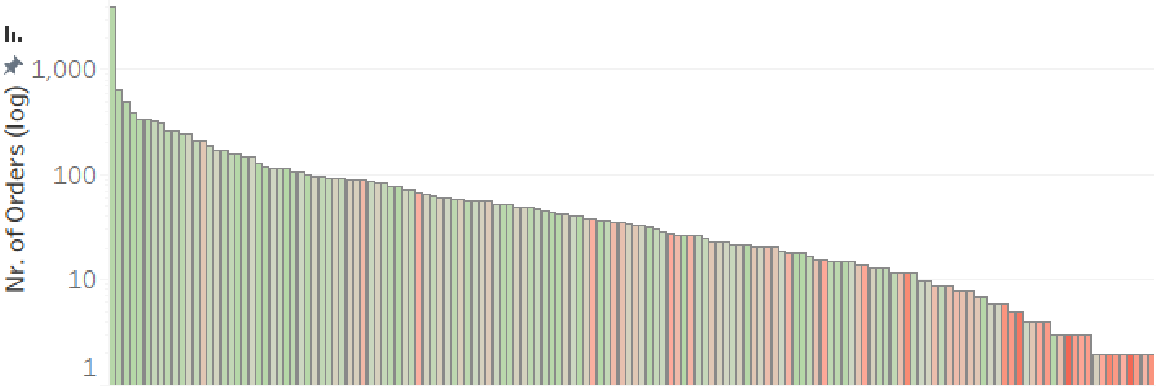


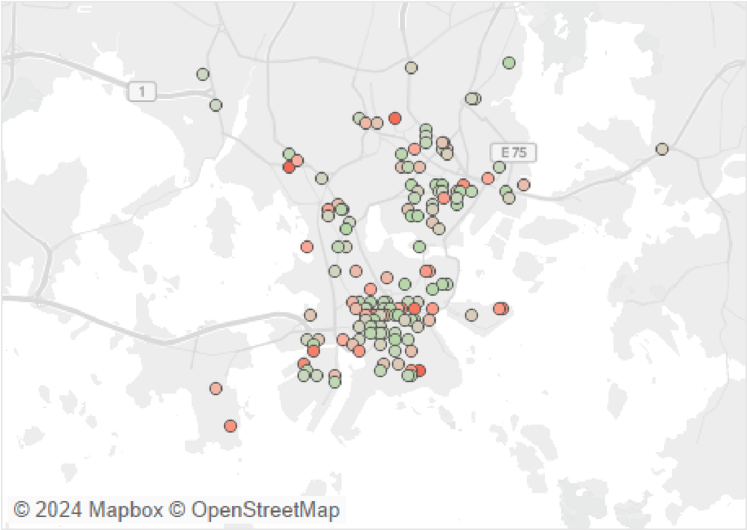
Distribution (number of occurrences) for delivery minutes (Actual, Estimated and their difference).

Ranking Venues

According to the **Avg. Difference** between Actual and Estimated Delivery Time.

Are delays linked to Position/Popularity of Venues?





Most **popular venues (higher number of orders)** show **shorter average delays**.
No obvious correlation between delays and Venue Position.

Top N Venues (Actual - Estimated Del. Min.)

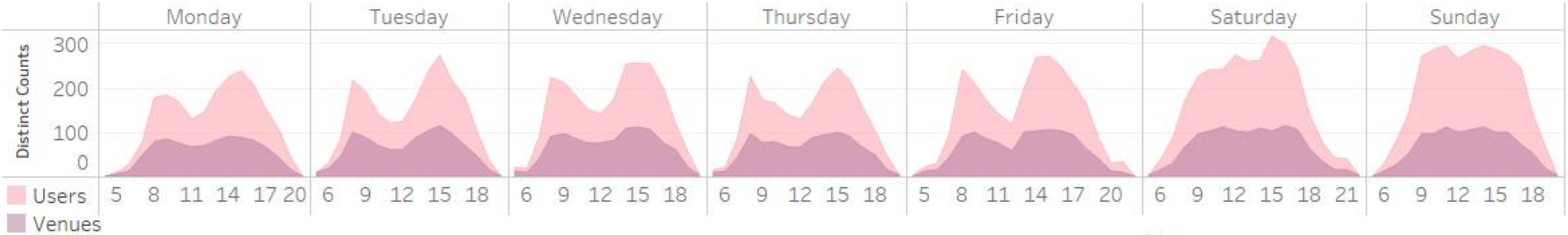
150

Interactive Dashboard at this [link]

HeatMap

Positions of Users and Venues

EDA PART 3



Is the **location** of ordering users influenced by hour or weekday?

Is there a **shift** of *hot* regions?

How do the numbers of users and venues **change over time**?

Hour

11

Weekday

All

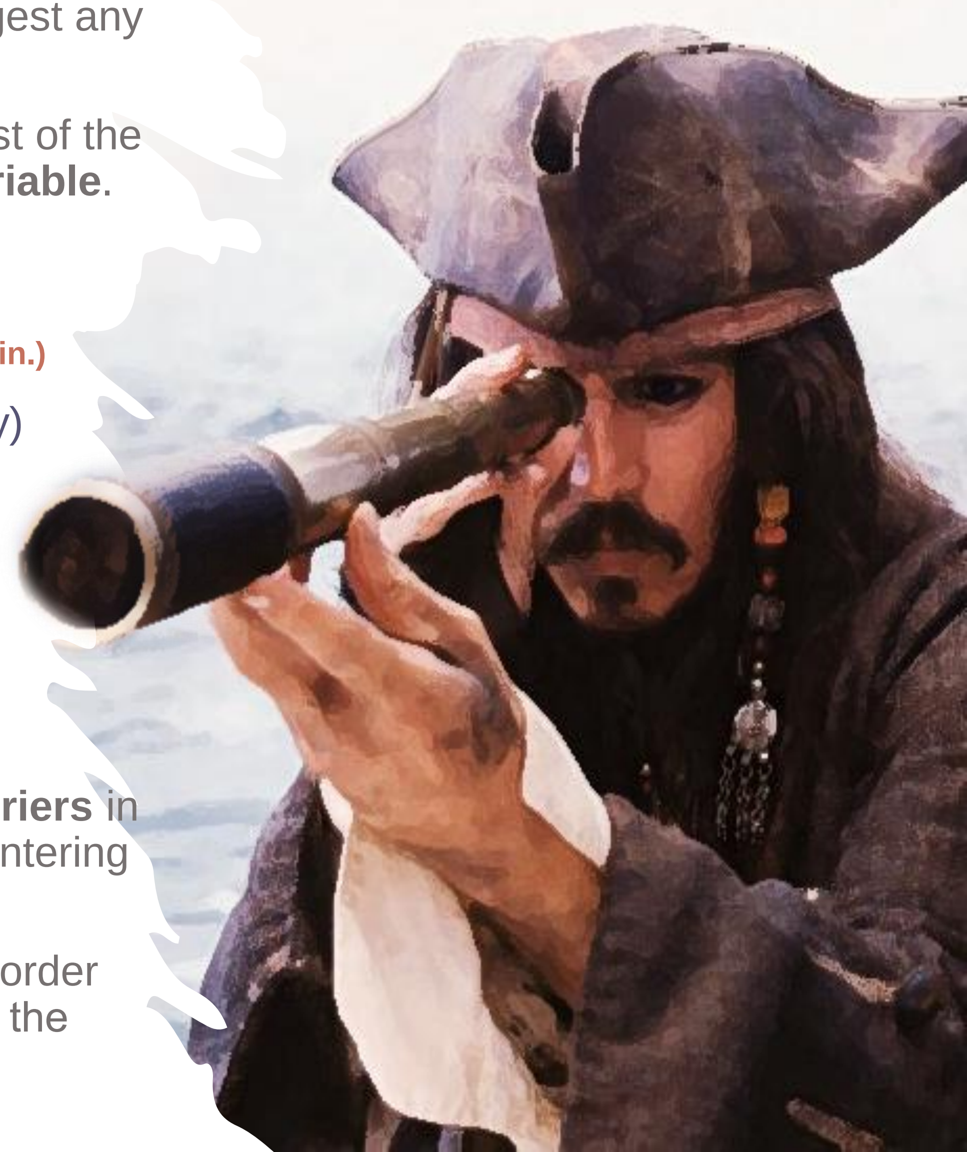
Interactive Dashboard at this [link]

MAKING PREDICTIONS
BASED ON DATA

- The EDA showed interesting trends but didn’t suggest any surprising phenomenon.
- To yield **quantitative predicitons**, making the most of the data, we stick to a **well defined, interpretable variable**.

We **forecast**:
- **Delays (Act. – Est. Del. Min.)**

considering their dependence from:
- time (hourly avg. delay)
- weekday
- hour of the day
- rain



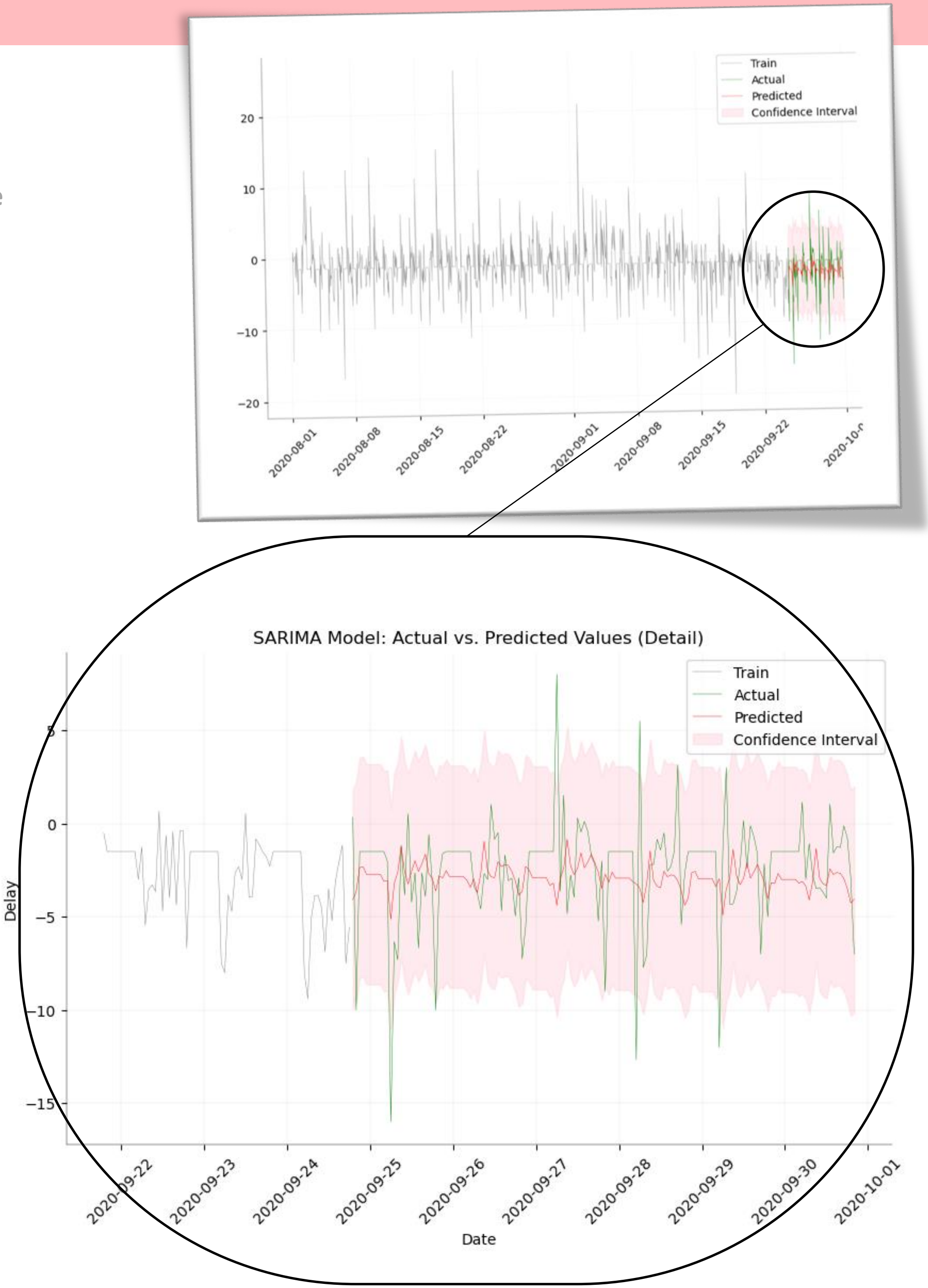
WHY?

- The company could favor a **higher number of couriers** in those hours where **more delays** are expected, countering the latter by properly distributing deliveries.
- The app could **estimate higher delivery times**, in order not to disappoint users, making them 'conscious’ of the longer waiting times.

FORECASTING WITH SARIMAX

Details in the notebook

- **S**easonal **A**uto**R**egressive **I**ntegrated **M**oving **A**verage with He**X**ogenous features
- Consider the series **global trends** (MA) and the **autocorrelation** (AR), for stationary and seasonal component.
- We used **correlograms** to identify the order parameters **(p, d, q)** for the ARIMA part and **(P, D, Q, s)** for the SARIMA part



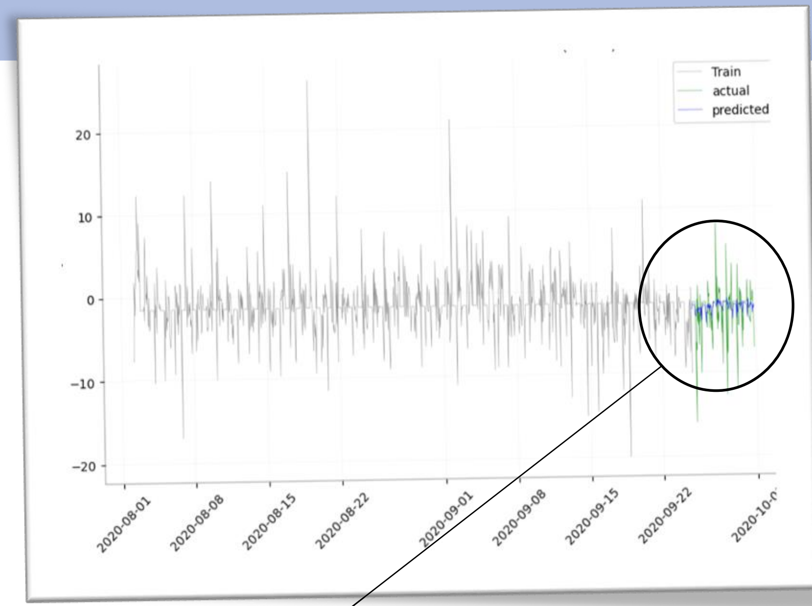
Results:

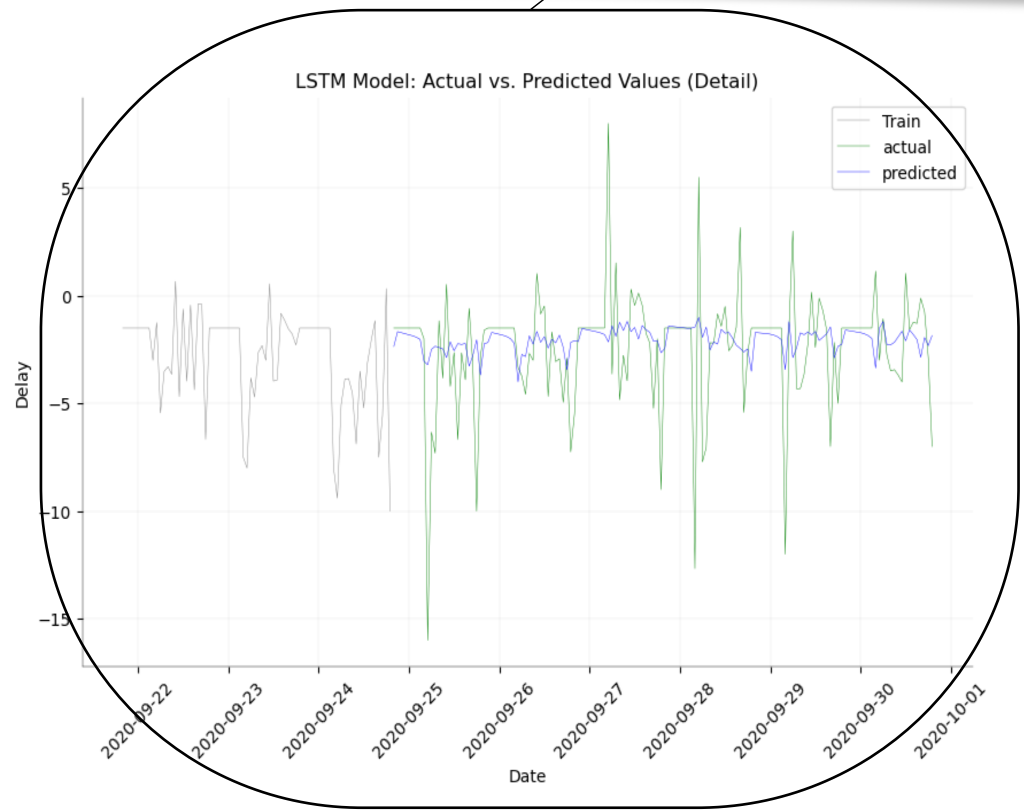
- The **model captures** global behaviours and patterns
- **High values for hourly delay fail to be predicted**, sometimes even trespassing outside of the 95% confidence band (pink).
- **RMSE** is about **2.9** (minutes)

FORECASTING WITH LSTMs

Details in the notebook

- **L**ong **S**hort-**T**erm **M**emory Recurrent Neural Networks.
- Ability to **remember long-term dependencies** and handle variable-length sequences.
- We build two basic architectures and set a 24 hours lookback window.





Results:

- The **model captures** global behaviours and **patterns**, but not better than SARIMAX.
- **High values for hourly delay fail to be predicted**. The predicted values tend to lay around the mean.
- **RMSE** is about **2.8** (minutes)

CONCLUSIONS AND FURTHER DEVELOPMENTS

- Forecasting **daily delays** rather than hourly ones would be **reasonable**, and likely yield **better results**. **More data is needed** in order to do that.
- **SARIMAX and LSTM** models should be **optimized** by **tuning parameters** and modifying **architectures.**
- **Anomaly detection** could be employed to predict spikes in the delivery time.
- **More advanced models** should be considered. LSTMs could be replaced by **Transformers**.