THE UNIVERSITY OF CHICAGO

Danil Meresenschi
Deepak Vanjani
Dharmi Gala
Prayut Jain
Christopher Marasco

Forecasting Travel Times for Uber Riders in Los Angeles

May 24th, 2023

Agenda

1: Problem Statement & Approach

2: Data Exploration

3: Modeling

4: Takeaways Recommendations, & Future Work

Problem Statement

Uber Movement shares anonymized data aggregated from over ten billion trips to help urban planning around the world.

One of these data sets is based on travel times between two points.

How can we build accurate and robust time series models to drive cost savings within Uber Operations while continuing to meet customer needs?

Our high-level analysis approach

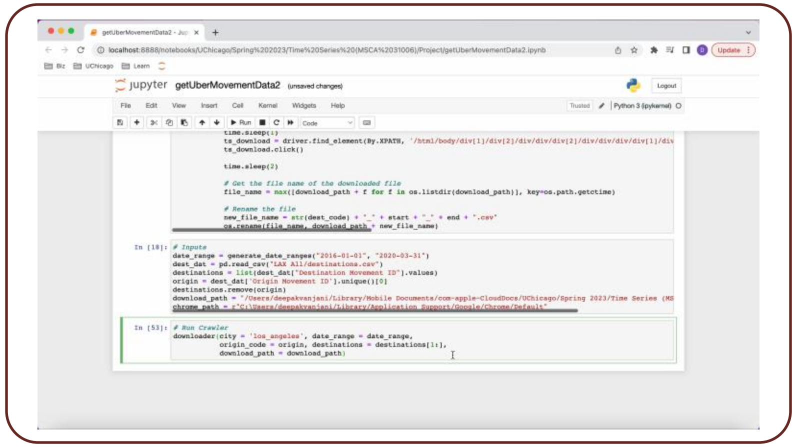
We leveraged the power of our automation-fueled pipeline to address the problem statement

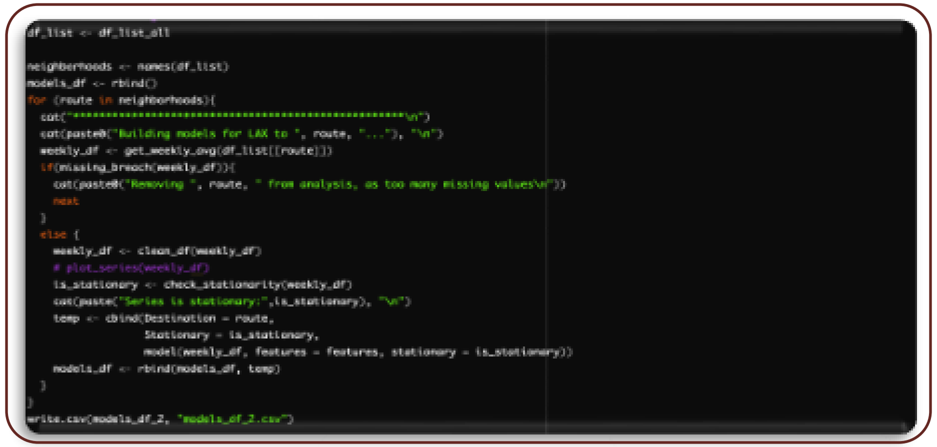
1. **Automated Data Download from Uber's Website**
   Built a custom Python bot to download publicly available time series data of Uber travel times for 60 routes with LAX as the starting point

2. **Data Cleansing & Preparation**
   Applied a series of steps to clean & prepare data including weekly averaging, correcting for stationarity, and addressing missing values

3. **Modeling & Analysis**
   Analyzed & evaluated 7 modeling methodologies:
   (1) Arima; (2) Regression;
   (3) Regression with Arima Errors;
   (4) Exponential Smoothing;
   (5) TBATS; (6) LSTM (Bonus);
   (7) XGBoost (Bonus)

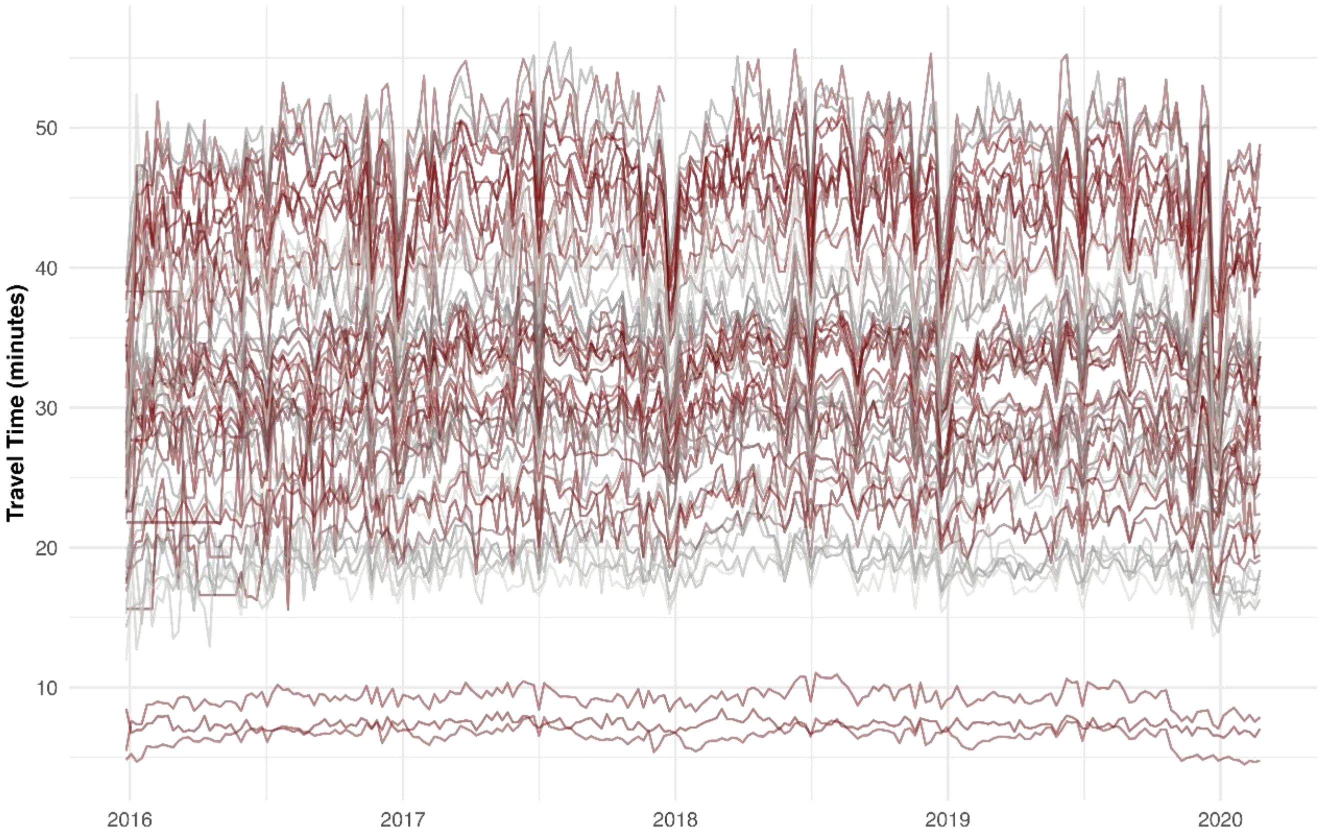
4. **Recommendations & the Future**
   Present recommendations & outline pathway for future work





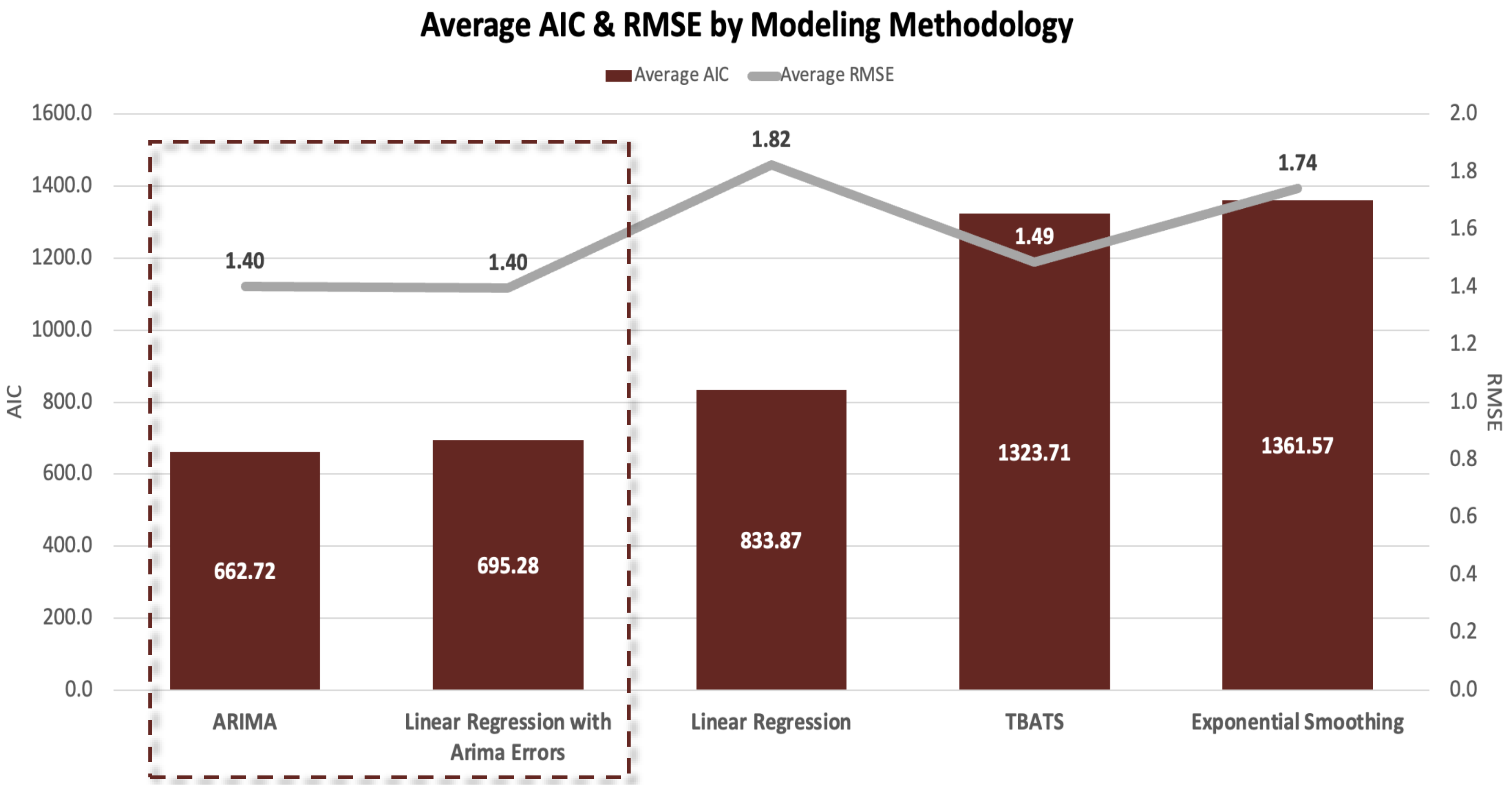
Key features of our data

Average Weekly Travel Time between LAX and 50 different locations in LA (Jan 2016 - Feb 2020)



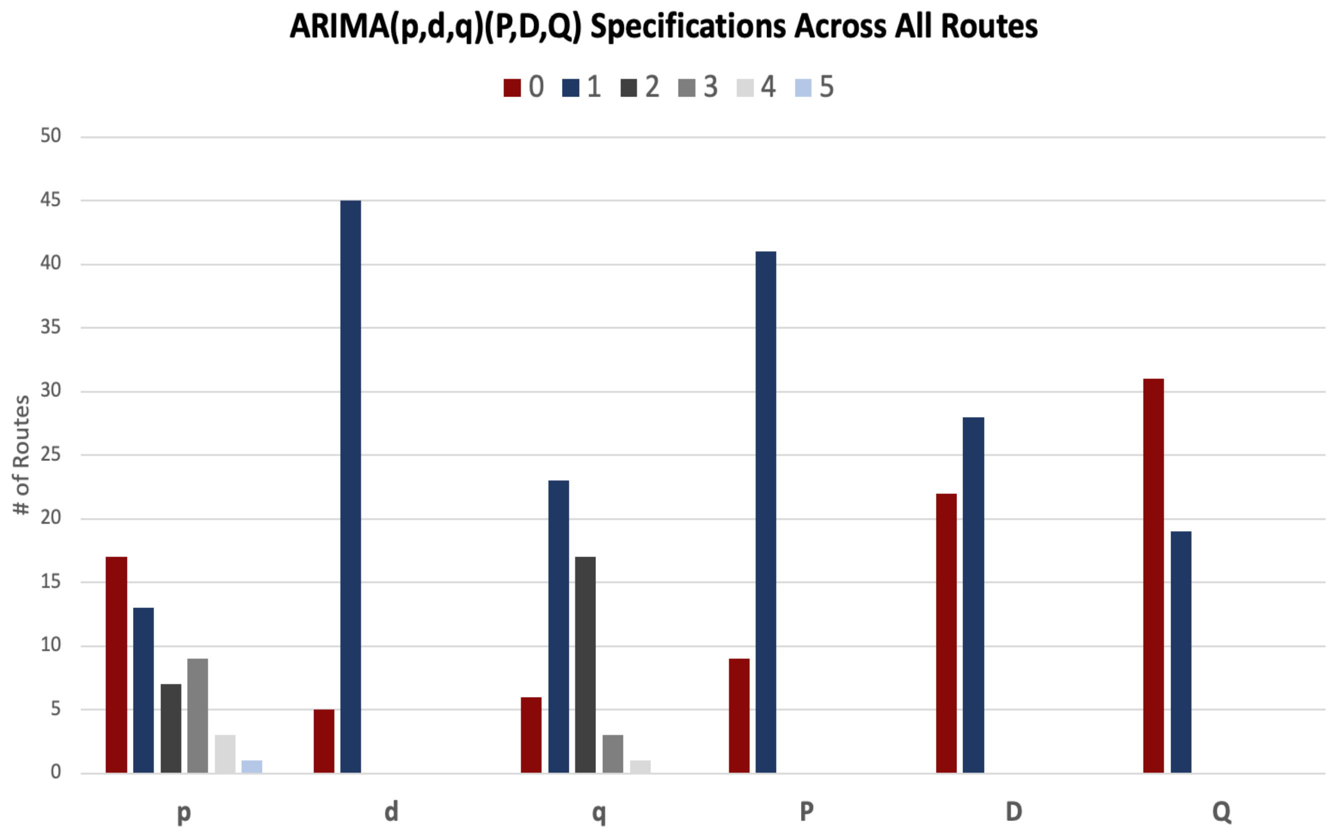
- **COVID Impact:** Data is available from January 2016 – March 2020, but we exclude March to avoid one-time COVID impact.
- **Missing Values:** 60 routes filtered down to 50 using a missing value threshold of 15%. Any remaining missing values were addressed using Last Observation Carried Forward imputation.
- **Seasonality & Trend:** Frequency of 52 weeks, with clear signs of annual seasonality. No clear trend present.
- **Stationarity:** All routes modeled were stationary using ADF test threshold of 0.05. Some routes required seasonal and/or non-seasonal differencing of order 1.
- **External Features:** Our analysis considered external variables including publicly available data for temperature in LA, relative humidity in LA, and a national holiday indicator.

Comparing model methodologies



ARIMA's show the lowest average training RMSE & AIC across all routes considered, followed by Linear Regression with ARIMA errors. As such, we focus our analysis on these two methods.

Identifying Candidate Models (Arima)

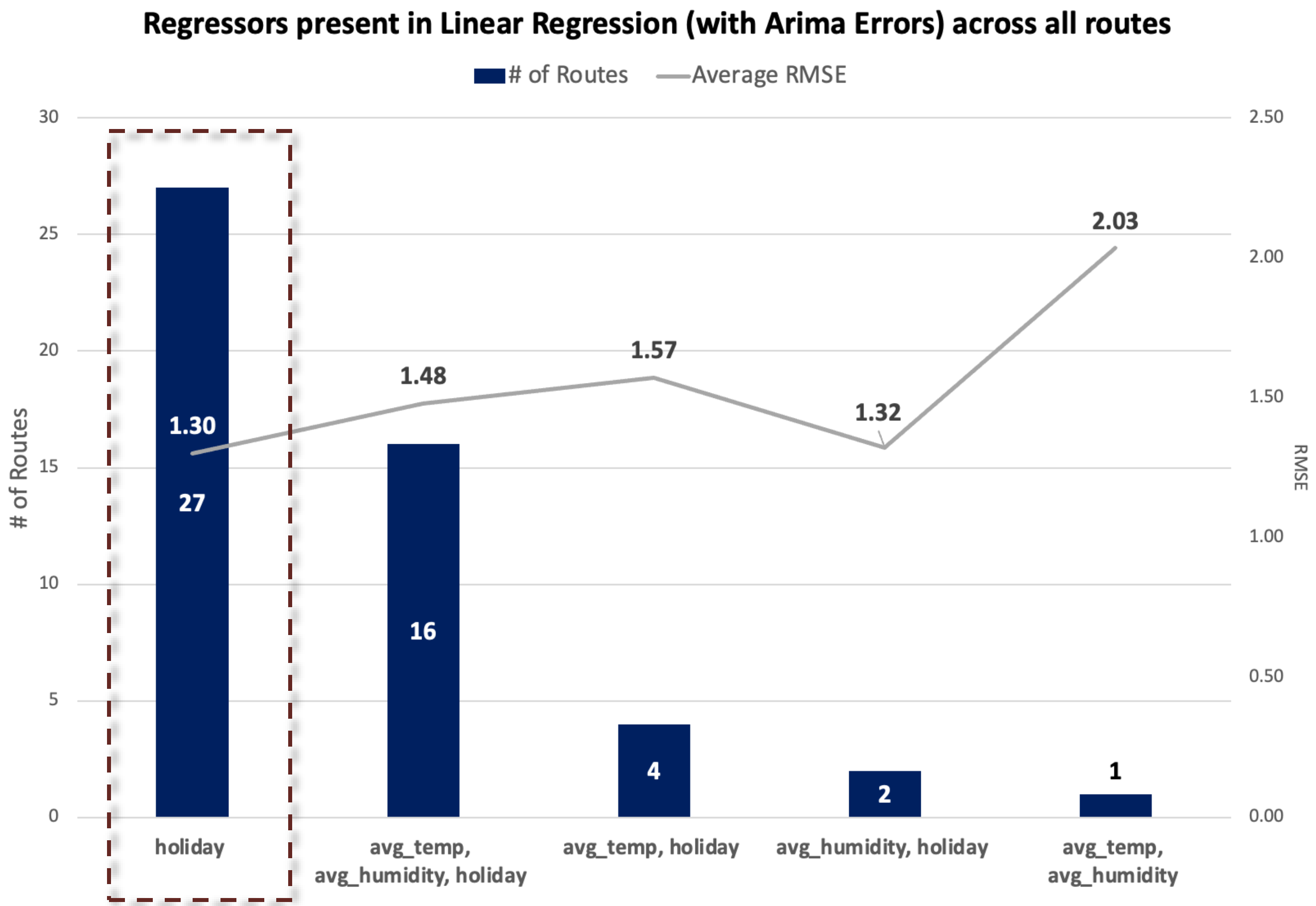


Candidate Models

#	Models	% of routes covered
1	**ARIMA(0,1,1)(1,1,1)[52]**	**10%**
2	**ARIMA(0,1,1)(1,1,0)[52]**	**8%**
3	ARIMA(0,1,2)(1,0,0)[52]	6%
4	ARIMA(1,1,2)(1,1,0)[52]	6%

Across all routes, optimal ARIMA models contain 0-1 non-seasonal (p) and seasonal (P) autoregressive terms, 1 non-seasonal (d) and seasonal difference (D), and 0-2 lagged seasonal (q) and non-seasonal errors (Q).

Identifying Candidate Models (Regression w/errors)



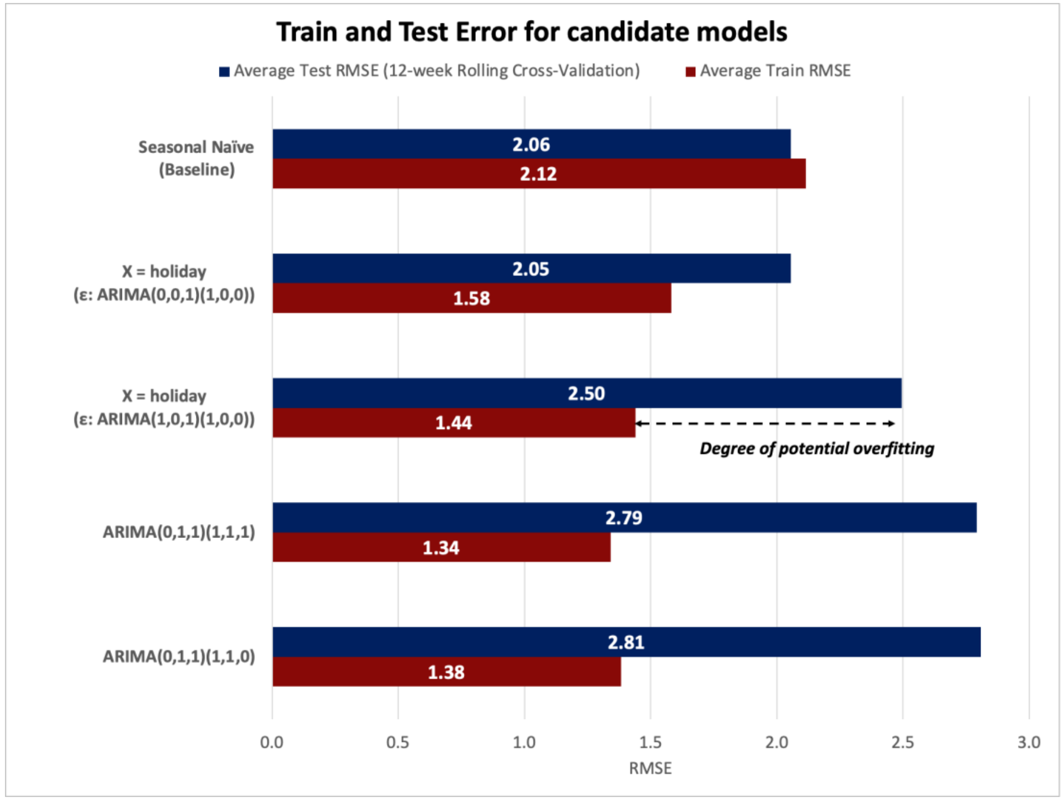
Candidate Models

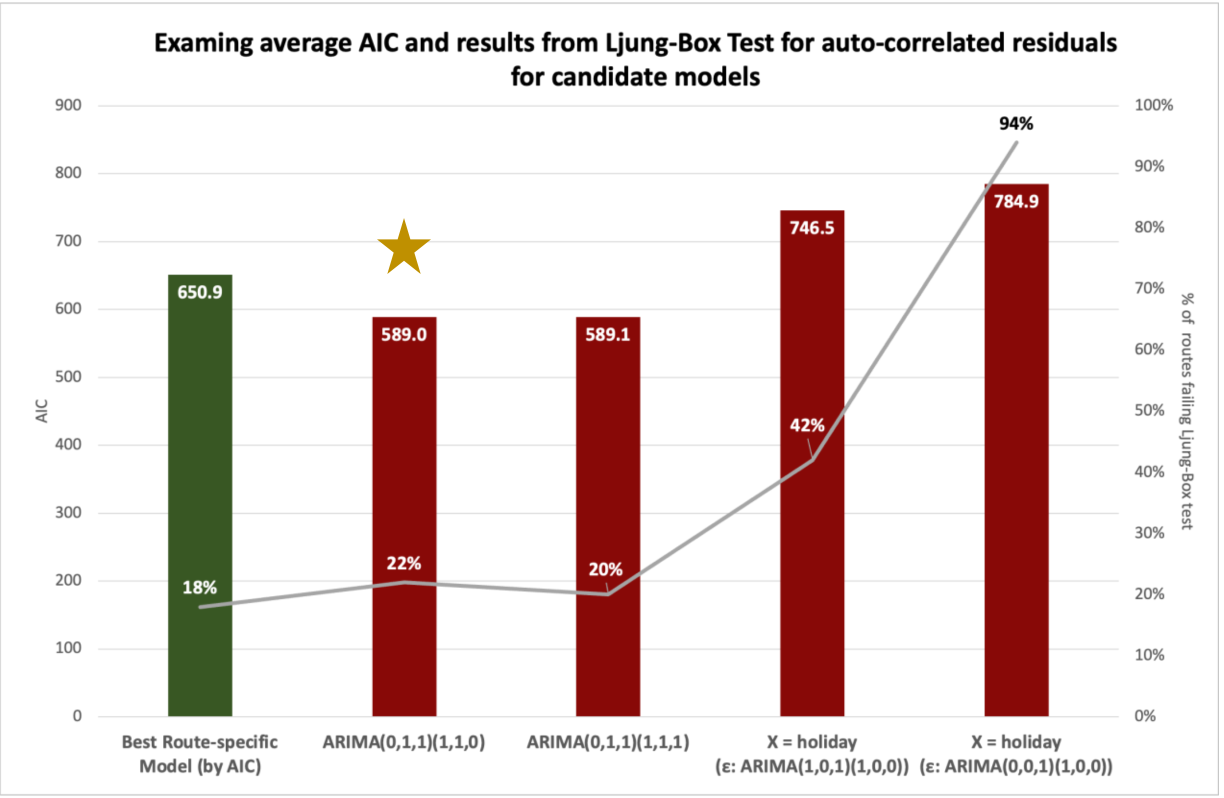
#	Models	% of routes covered
1	**X = holiday** *(Errors: ARIMA(0,0,1)(1,0,0)[52])*	**4%**
2	**X = holiday** *(Errors: ARIMA(1,0,1)(1,0,0)[52])*	**4%**
3	X = holiday *(Errors: ARIMA(0,0,1)(1,1,1)[52])*	4%
4	X = holiday (Errors: ARIMA(5,0,0)(1,0,0)[52])	4%

Linear regressions with a single holiday indicator variable outperforms models with additional variables.

Meeting our champion, Arima (0,1,1)(1,1,0)

This model shows the best balance across performance, simplicity, and robustness





Net Takeaways & future considerations

One Generalized Model to ”rule them all”

MODEL RECOMMENDATION SPECTRUM

“We’ve got compute power for days”

Arima (0,1,1)(1,1,0)

Arima & Regression with Errors with fixed parameters

Arima & Regression with Errors with varying parameters

Best Model, per route

Conclusions for Uber

We built a fully automated modeling suite empowering Uber Operations with the option to build route-specific models & more generalized models.

We have model recommendations across it all. Ultimately it boils down to a choice of simplicity & operational efficiency or large-scale improvement & better fit.

Future considerations

Investigation into what’s important for Uber. Did our problem statement address a pressing need?

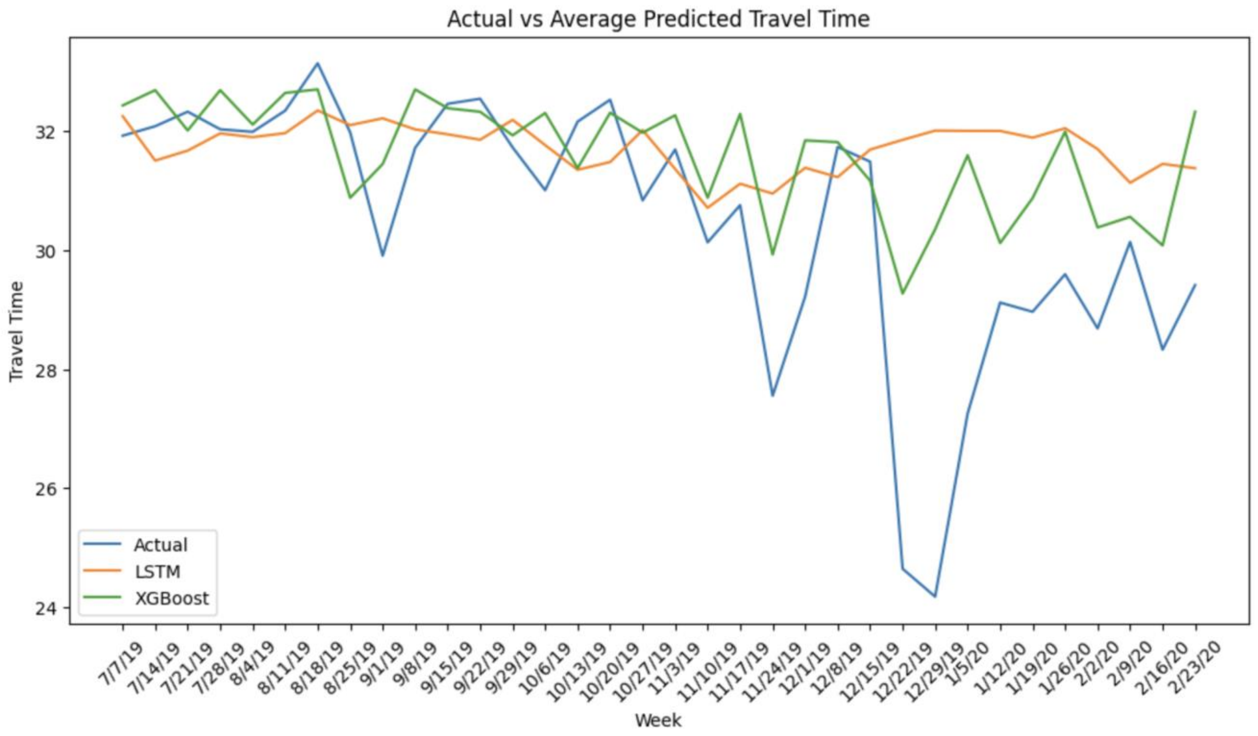
More routes, more locations, more test & learn. We need to test our modeling assumptions more broadly;

- New starting points, end points, across major metros, beyond urban, and extended globally
- Including new externalities. We leveraged weather & Holidays in our modeling. But what are Uber’s most important additional features?

BONUS: LSTM & XGBoosted Regression

Methodology:

- Computational Efficiency ➞ used average travel times of all locations, instead of average output of every model iteration through the locations.
- Experiment with more sophisticated models to assess their prowess to yield more performant results.
- Strong outcomes can be achieved with LSTM and XGBoost models.
- Results are close to the evaluation scores of the core 5 route-specific models ➞ a model that incorporates all locations might also be viable.
- Details on each model's methodologies are in the appendix section.



Model Evaluation RMSE:

XGBoost= 1.9496

LSTM = 1.0841

THE UNIVERSITY OF CHICAGO

THANK YOU

Contributions

The entire team touched on nearly every aspect of this project. We met weekly (9 hours in total), coordinated deliverables, and paced, as a group, to our finished presentation. **Below we have only highlighted the components each member over-indexed in.**

Deepak Vanjani = R & Python Scripts (Web Scraper, Automation, Aggregation, Reporting), Modeling, Project Brief

Danil Meresenschi = Model Recommendations, Data Strategy, Parallel Python work

Dharmi Gala = Model Recommendations, Data Collection, Outcome Strategy

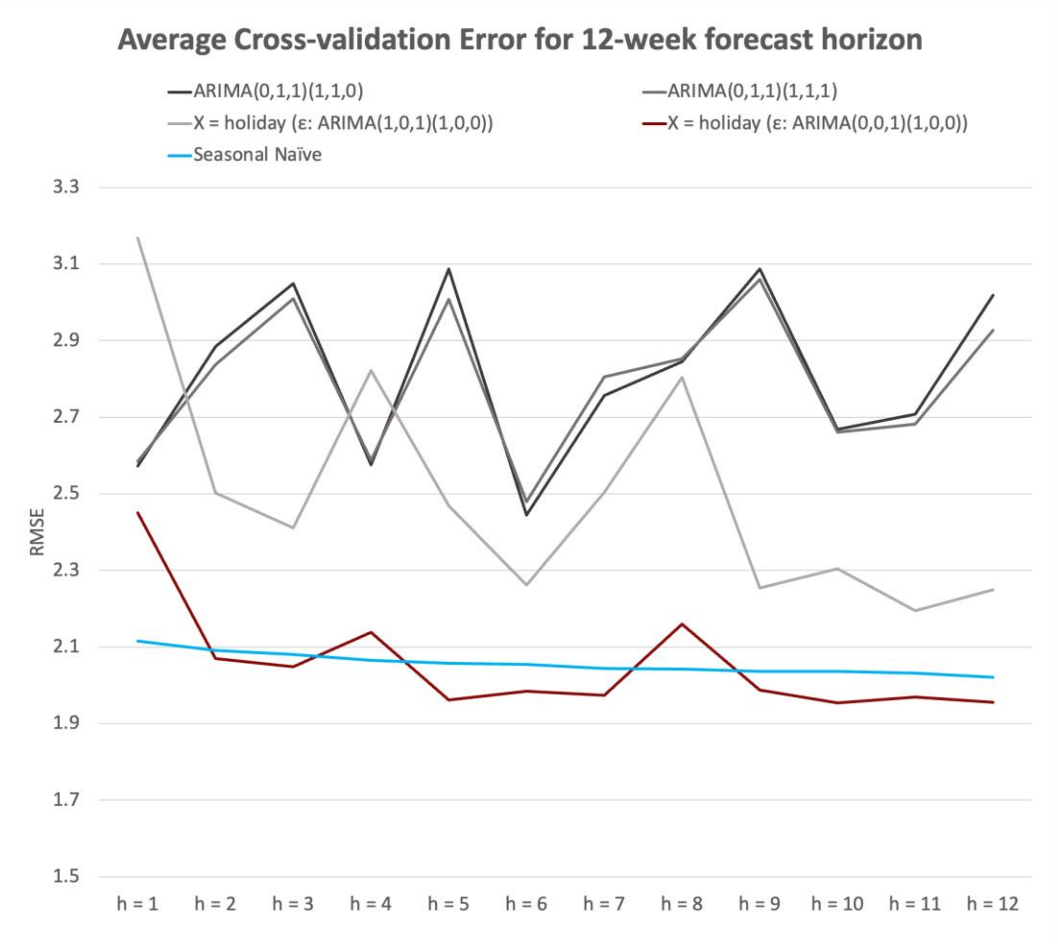
Prayut Jain = Model Recommendations, Data Collection, Outcome Strategy

Christopher Marasco = Project Management, Data Collection, Narrative & Presentation

THE UNIVERSITY OF CHICAGO

APPENDIX

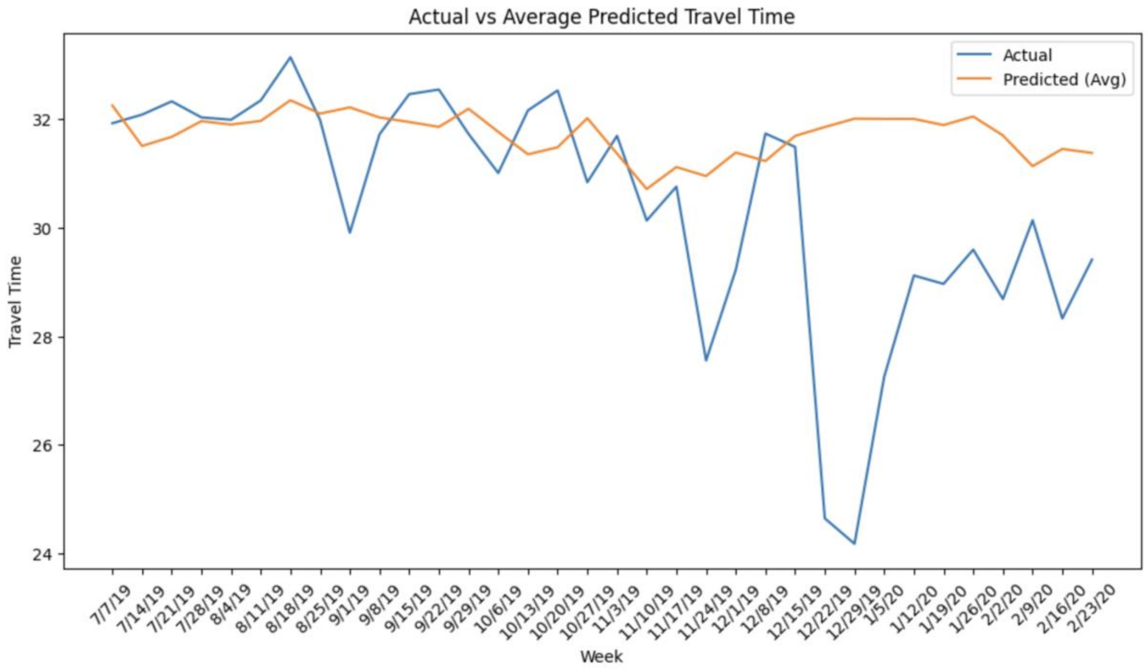
Appendix 1. Cross-validation results for candidate models



Methodology:

- Applying cross-validation with rolling 12-week windows shows that X = holiday (Errors: ARIMA(0,0,1)(1,0,0) performs the best in terms of average RMSE at each forecasting horizon

Appendix 2. LSTM

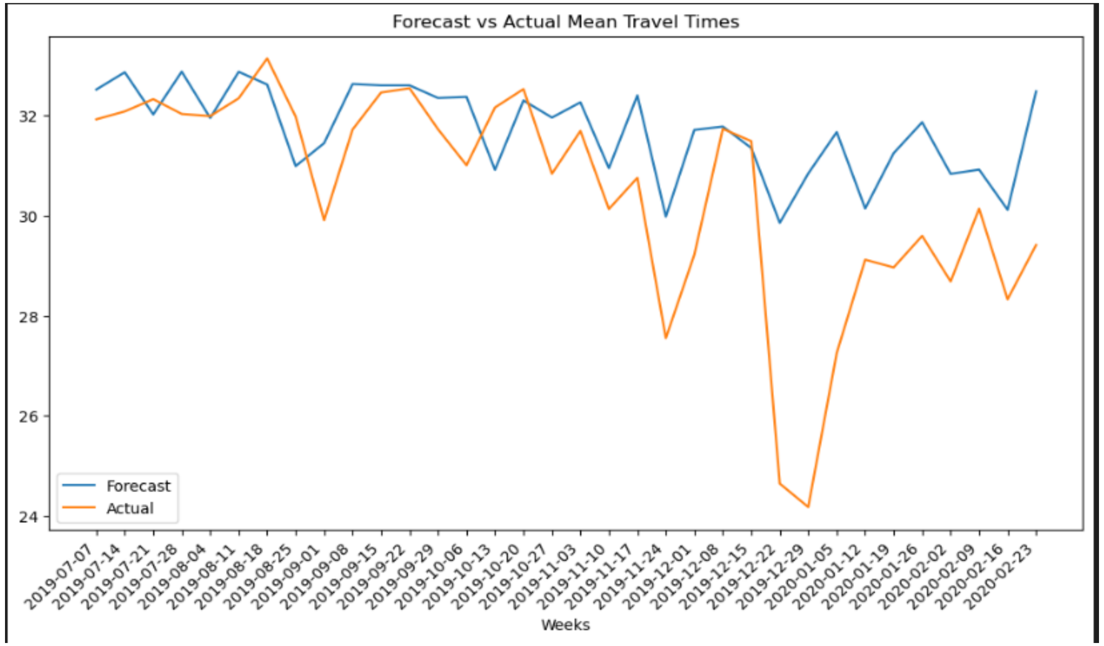


Model Evaluation:
RMSE = 1.0841

Methodology:

- Used normalized variables.
- Cross-validated using TimeSeriesSplit with 5 folds
- Implemented with 64 units and ReLU activation function.
- Compiled with the Adam optimizer and mean squared error loss function
- Trained using the training data for 100 epochs and a batch size of 32

Appendix 3. XGBoost Regression with Randomized Search.



Order of important features:

1. Holiday (True/False)
2. Average Temperature
3. Average Humidity

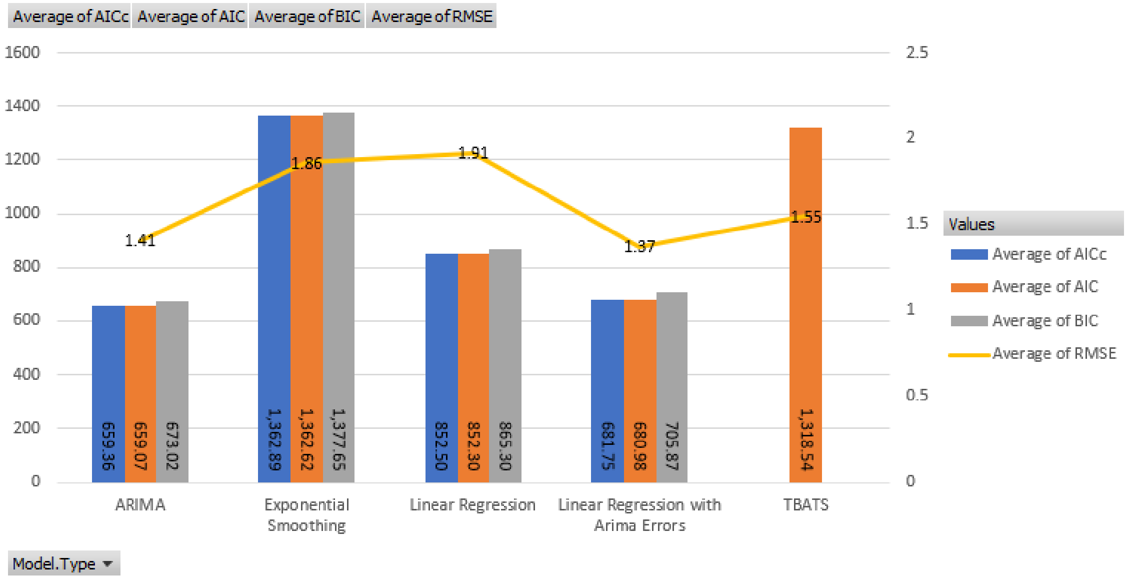
Model Evaluation:

RMSE = 1.9496

Methodology:

- Utilized Randomized Search to determine the best parameters.
- Implemented 5-Fold Cross-Validation in Randomized Search.
- The best XGBoost Regression model had the following parameters:
  - Learning Rate = 0.1
  - Max Depth = 3
  - Gamma = 1
  - Number of Estimators = 100

Appendix 4. Tourist Locations



Methodology:

- We included tourist-centric locations in our analysis to provide an opportunity to isolate the seasonality of travel times to potentially episodic locations.
- But from what you can see, the best options remain ARIMA and Linear Regression with ARIMA errors!