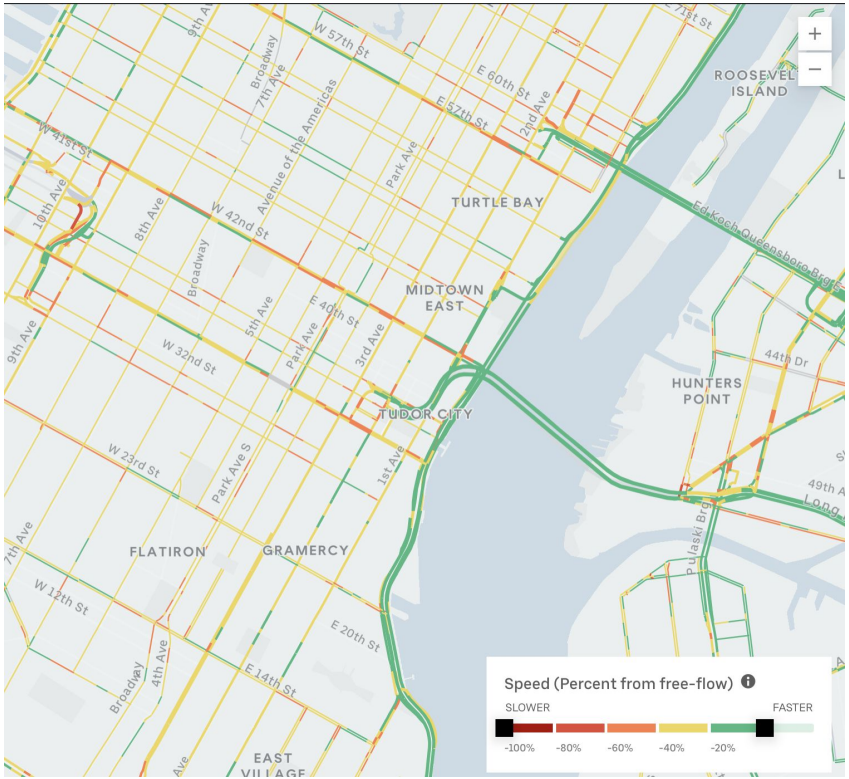
Forecasting NYC Street Speed using Weather Data

Levente Szabo, Yaowei Zong, Lingbo Ji

Introduction

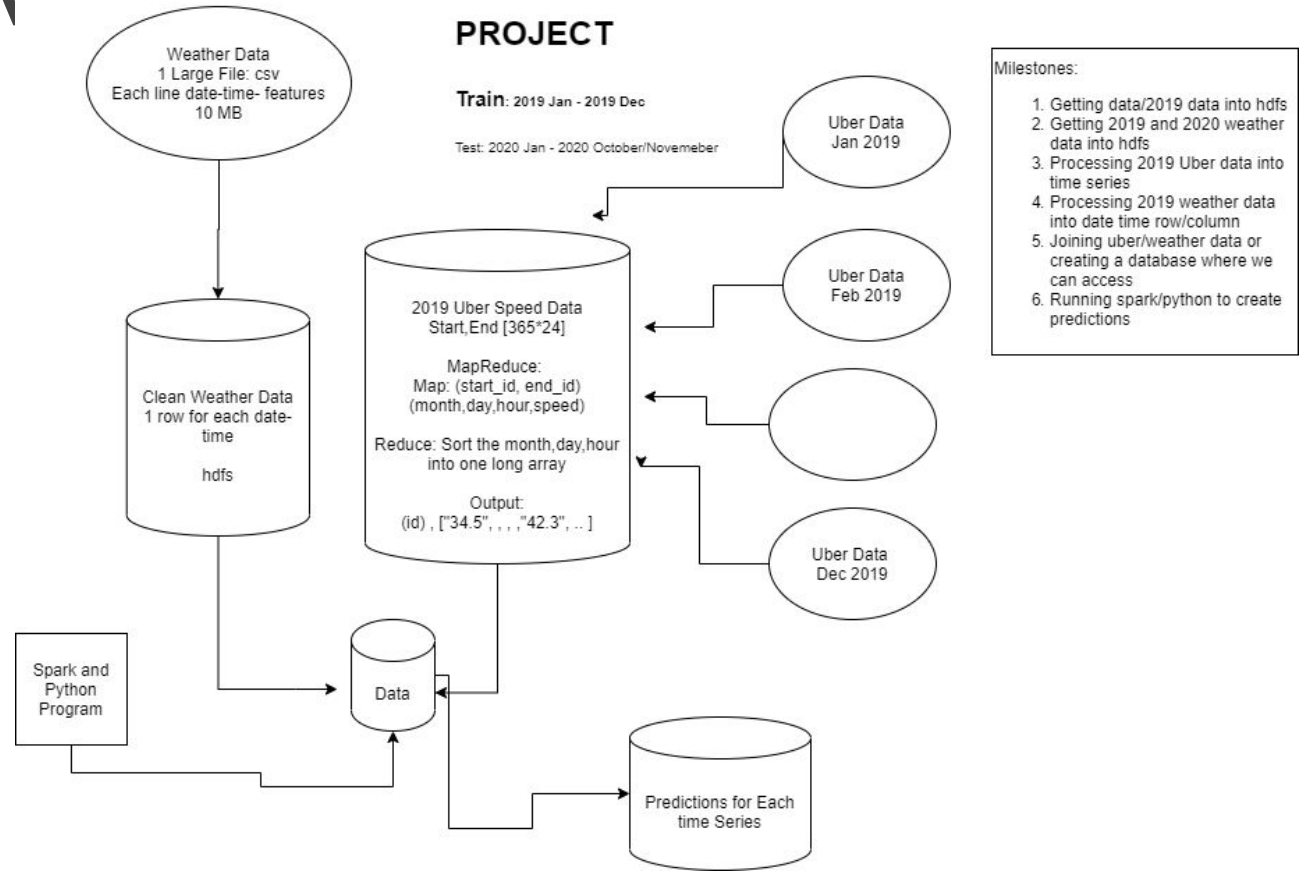
- Traffic prediction involves many features
- Predicting NYC Street Movement Speed based on previous data
- Understand the impact of weather conditions on street movement
- Help NYC Commuters and Uber Drivers



Datasets

- **Uber Movement:** NYC Street Speed (speed)
- **NCEI (formerly NCDC) :** Local Climatological Dataset (weather)
- **NYC Open Data:** Motor Vehicle Collisions (crashes)

Overview


PROJECT
Train: 2019 Jan - 2019 Dec
Test: 2020 Jan - 2020 October/Novemeber
Weather Data
1 Large File: csv
Each line date-time- features
10 MB
Clean Weather Data
1 row for each date-time
hdfs
Uber Data
Jan 2019
Uber Data
Feb 2019
Uber Data
Dec 2019
2019 Uber Speed Data
Start,End [365*24]
MapReduce:
Map: (start_id, end_id)
(month,day,hour,speed)
Reduce: Sort the month,day,hour into one long array
Output:
(id) , ["34.5", , , , "42.3", ..]
Spark and Python Program
Data
Predictions for Each time Series
Milestones:
1. Getting data/2019 data into hdfs
2. Getting 2019 and 2020 weather data into hdfs
3. Processing 2019 Uber data into time series
4. Processing 2019 weather data into date time row/column
5. Joining uber/weather data or creating a database where we can access
6. Running spark/python to create predictions

Dataset(Weather)

- Range: 2019-01-01 ~ 2020-03
- Size: 12119 + 2980 = 15099 Rows

  124 Columns, comma divided

"72505394728","2019-12-30T16:04:00","FM-16","7",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,"29.78","39","40","0.05","RA:02 BR:1 |RA |RA",,,"97",,"OVC:08 7","29.61","0.75","40","040","18","13",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,"MET10812/30/19 16:04:02 SPECI KNYC 302104Z 04011G16KT 3/4SM RA BR OVC007 04/04 A2978 RMK AO2 P0005 T00440039 $ RTX","FM-16","7",,,,,,,,,,,,,,,,,,,,,,,,,,,,

[0] STATION	"72505394728"
[1] DATE	"2019-12-30T16:04:00"
[2] REPORT_TYPE	"FM-16"
[3] SOURCE	"7"
...	
[41] HourlyAltimeterSetting	"29.78"
[42] HourlyDewPointTemperature	"39"
[43] HourlyDryBulbTemperature	"40"
[44] HourlyPrecipitation	"0.05"
[45] HourlyPresentWeatherType	"RA:02 BR:1 |RA |RA"
[46] HourlyPressureChange	
[47] HourlyPressureTendency	
[48] HourlyRelativeHumidity	"97"
[49] HourlySeaLevelPressure	
[50] HourlySkyConditions	"OVC:08 7"
[51] HourlyStationPressure	"29.61"
[52] HourlyVisibility	"0.75"
[53] HourlyWetBulbTemperature	"40"
[54] HourlyWindDirection	"040"
[55] HourlyWindGustSpeed	"18"
[56] HourlyWindSpeed	"13"
...	
[93] REM	"MET10812/30/19 16:04:02 SPECI KNYC 302104Z 04011G16KT 3/4SM RA BR OVC007 04/04 A2978 RMK AO2 P0005 T00440039 $ RTX"
[94] REPORT_TYPE	"FM-16"
[95] SOURCE	"7"
...	
[120] Sunrise	
[121] Sunset	
[122] TStorms	
[123] WindEquipmentChangeDate	

Dataset(Weather)

-SN:03 FZ:8 FG:2 |FG SN |

‘- ‘: Light
SN: Snow
FG: Fog
FZ: Freezing

```
"72505394728","2019-12-30T16:04:00","FM-16","7",...,"29.78","39","40","0.05","RA:02 BR:1 |RA |RA",,,"97",...
"72505394728","2019-12-30T16:11:00","FM-16","7",...,"29.80","39","40","0.06","RA:02 BR:1 |RA |RA",,,"97",...
"72505394728","2019-12-30T16:38:00","FM-16","7",...,"29.81","38","39","0.16","+RA:02 BR:1 |RA |RA",,,"96",...
"72505394728","2019-12-30T16:45:00","FM-16","7",...,"29.81","37","39","0.18","RA:02 BR:1 |RA |RA",,,"93",...
"72505394728","2019-12-30T16:49:00","FM-16","6",...,"29.82","37","39",,"RA:02 BR:1 |RA |RA",,,"93",
"72505394728","2019-12-30T16:51:00","FM-15","7",...,"29.81","38","39","0.17","-RA:02 BR:1 |RA |RA",,,"96","29.79",...
```

	time	temp	precipitation	visibility	wind_gust	wind_speed	rain	snow	mist	fog	haze	freezing
8724	2019-12-30 15:00:00	39.75	0.03	2.19	22.00	10.50	2	0	1	0	0	0
8725	2019-12-30 16:00:00	39.33	0.12	1.75	25.67	14.67	3	0	1	0	0	0
8726	2019-12-30 17:00:00	39.00	0.03	7.00	39.00	23.00	1	0	0	0	0	0
8727	2019-12-30 18:00:00	37.50	0.01	7.00	32.50	14.50	0	1	0	0	0	0

External Data Sources

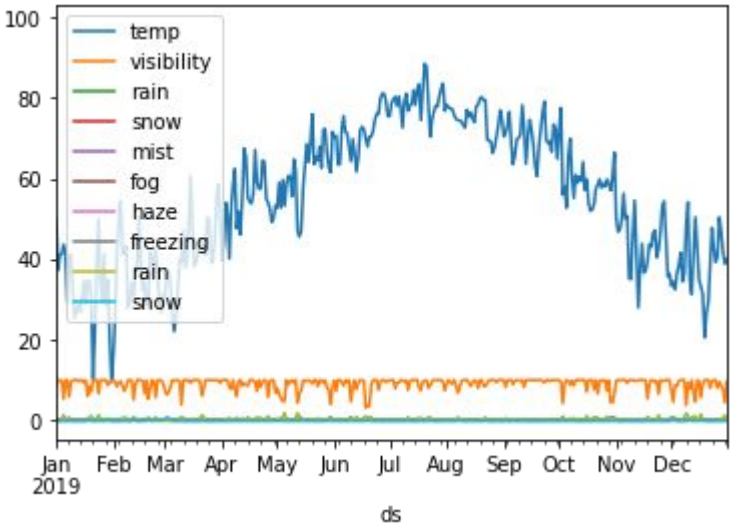


Fig 1. NYC Weather 2019

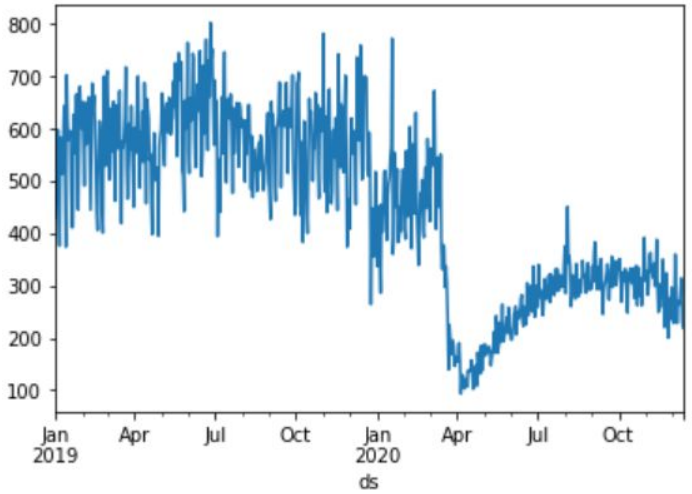


Fig 2. NYC Daily Crash 2019-2020

Dataset(speed)

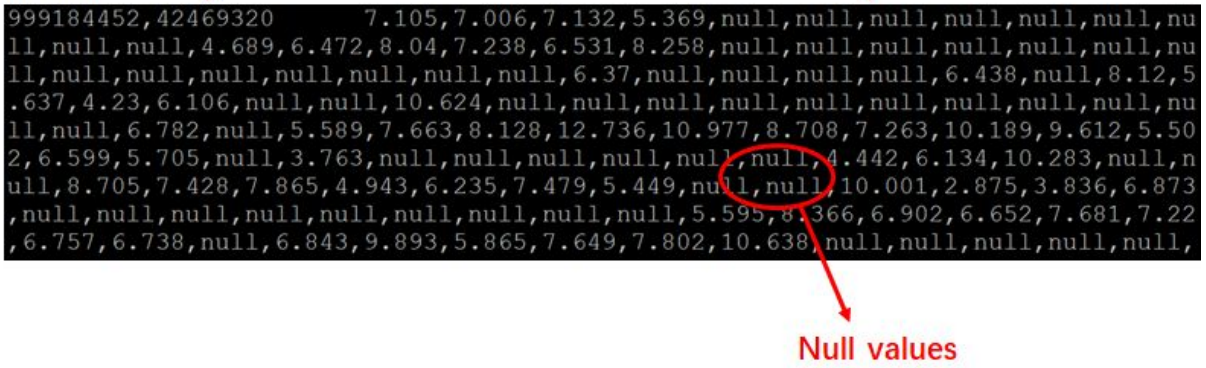
- Almost 72 GB in total with a lot of null values.

Dataset(speed)

Step1

- Selecting useful columns from the original dataset and convert them into a specified format:
- **(Key: start_node_id, end_node_id Value: speeds in time series format)**

Result:

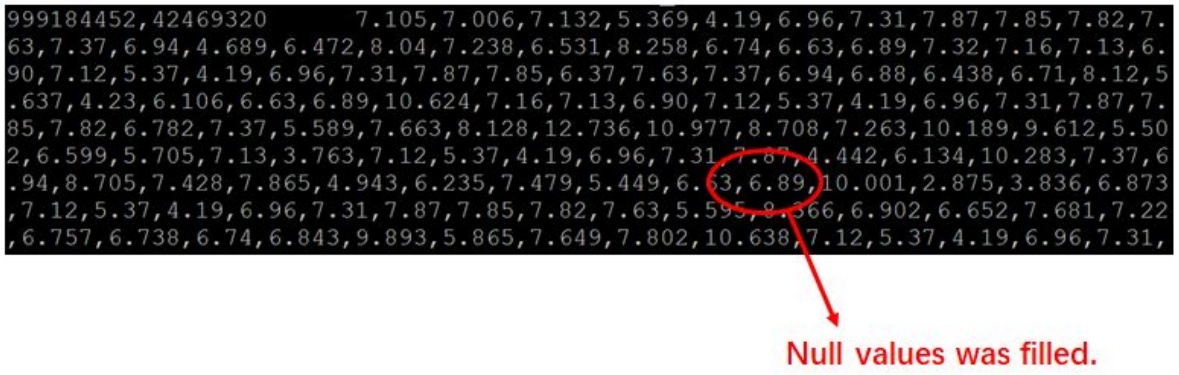


Dataset(speed)

Step2

- For each road, calculating the average speed for each hour of a day(24 hours).
- Filling in the null values according to the corresponding hour.

Result:



Dataset(speed)

Step3

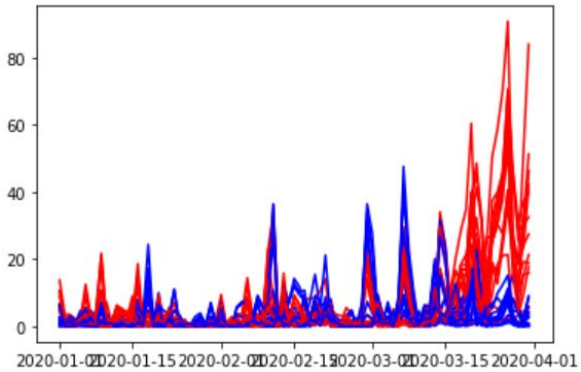
- If a road still has some null values, calculating the average speed and using the result to replace null.

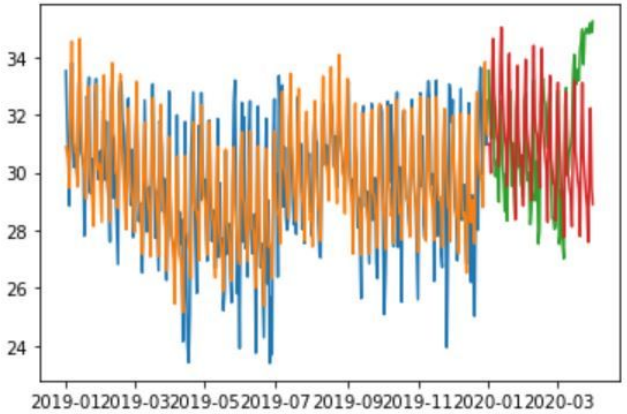
Methods

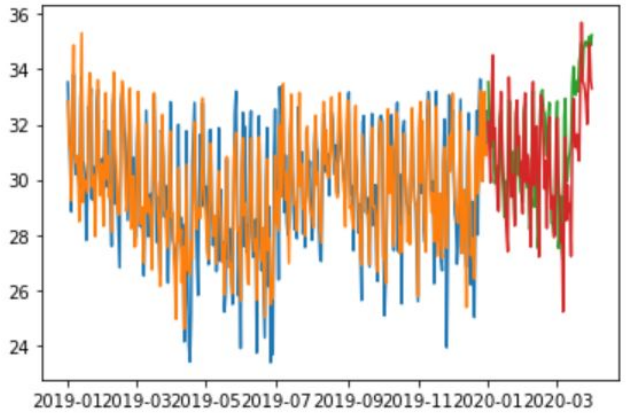
- To test the efficacy of using weather features and traffic collisions
- Fbprophet- Out of the box time series forecasting model
  - y(t) = trend(t) + seasonality(t) + holiday_effects(t) + e
  - Can add additional regressors to account for weather conditions and collisions
- PySpark : We need a way to interface between data stored in Spark and Fbprophet forecasting model.
- Next, we run a baseline model against our proposed solution and calculate the RMSE over the testing period
- Hyperparameters:
  - Trend: Linear
  - Seasonality: Additive, Multiplicative
  - Added Regressors: Additive, Multiplicative

Results

- Minimal effect of weather features pushed us to test a range of hyperparameters and include the daily crash data.
- Lockdown effects during march cause high error for the baseline model.
- Top: Model 1: Baseline Bottom: Model 2: Added Regressors
- Squared Residuals over Validation Set

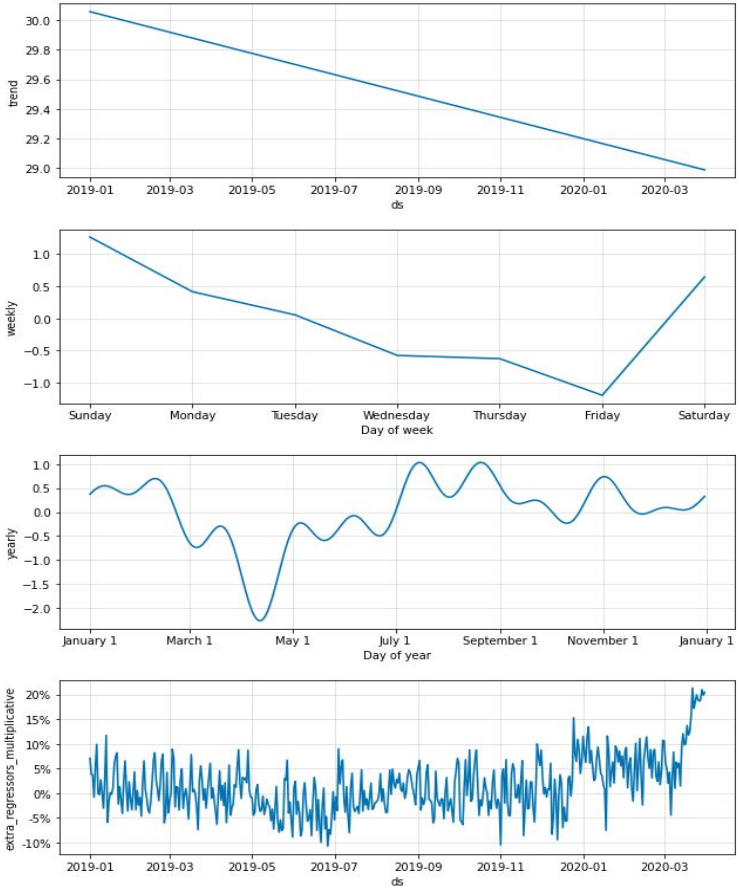






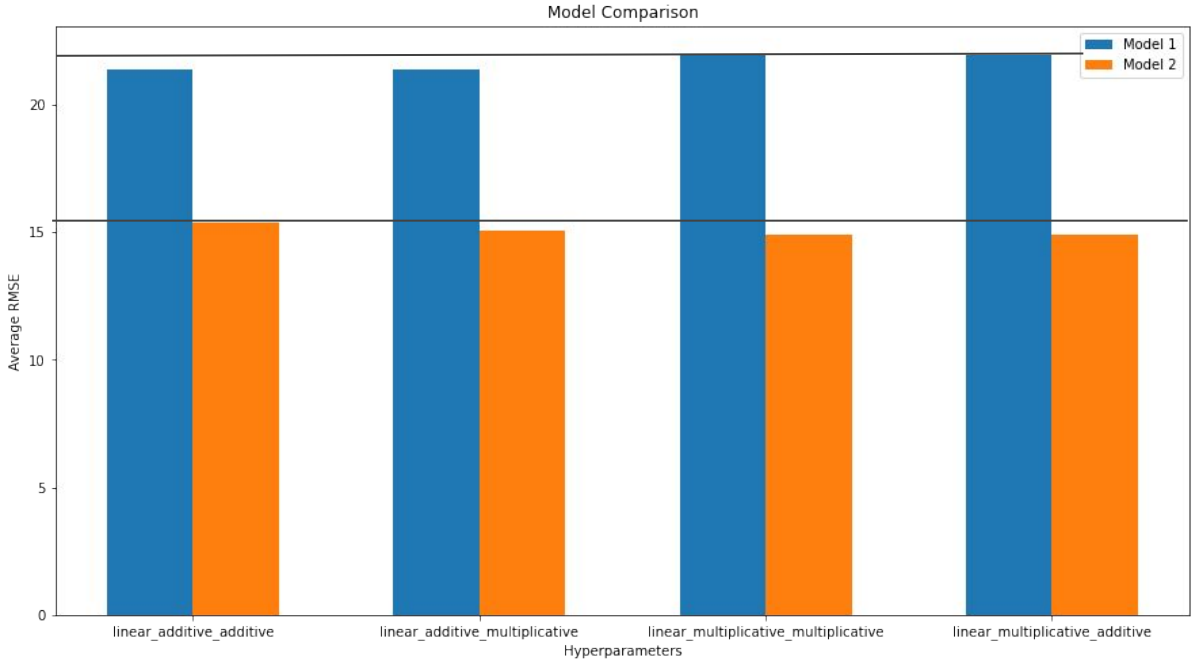
Model Components

- Plotting model components allow us to see learned parameters for specific time series and hyperparameters.
- Trend: In this case we learn a linear data trend over the period
- Seasonality: We can see the weekly and yearly fluctuations
- Additional Regressors: Multiplicative regressors produce a dilation of the speed.



Model Comparisons

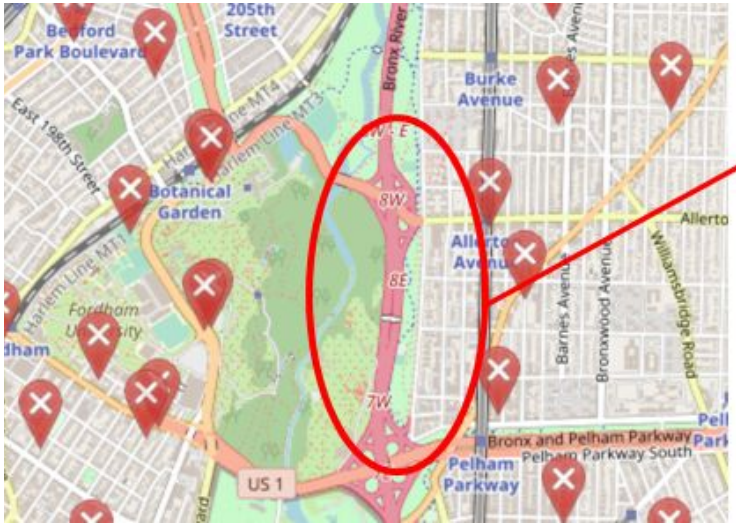
- Added regressors outperformed baseline model in all cases.
- Best model was using a linear trend with multiplicative seasonality and added regressors.



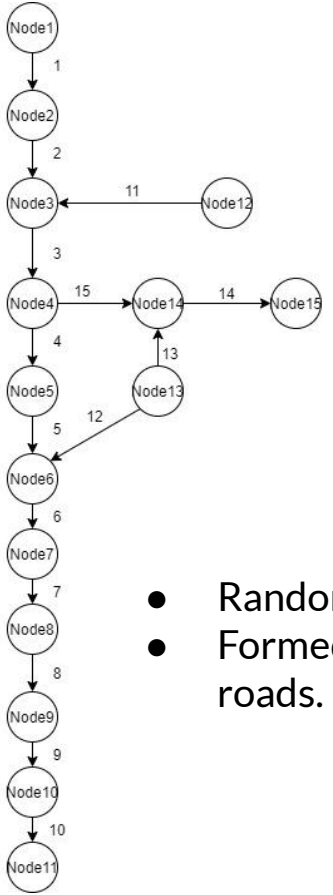
Results

- Testing with N = 295
  - Model 1 Baseline : Average RMSE = 19.55
  - Model 2 Added Regressors : Average RMSE = 16.55
- Formatted PySpark function to produce predictions or error calculations over specified subset of the data and desired forecasting period.

Real road example

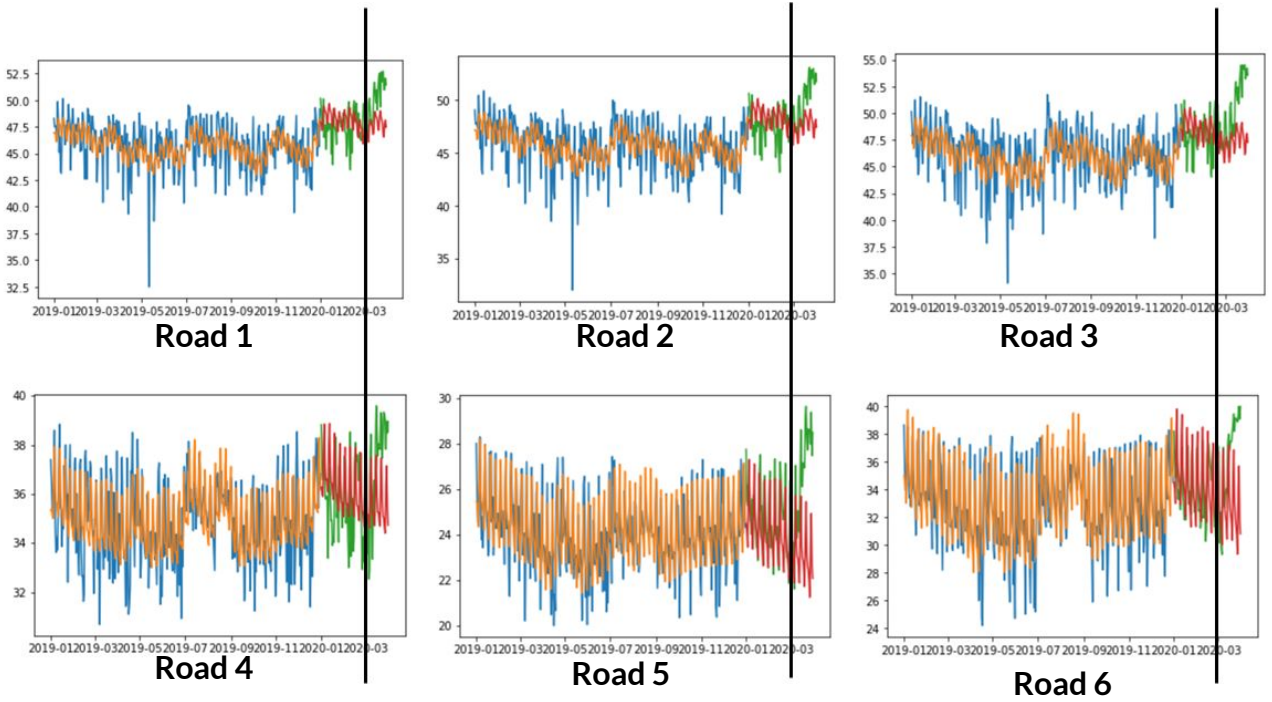


https://www.openstreetmap.org/



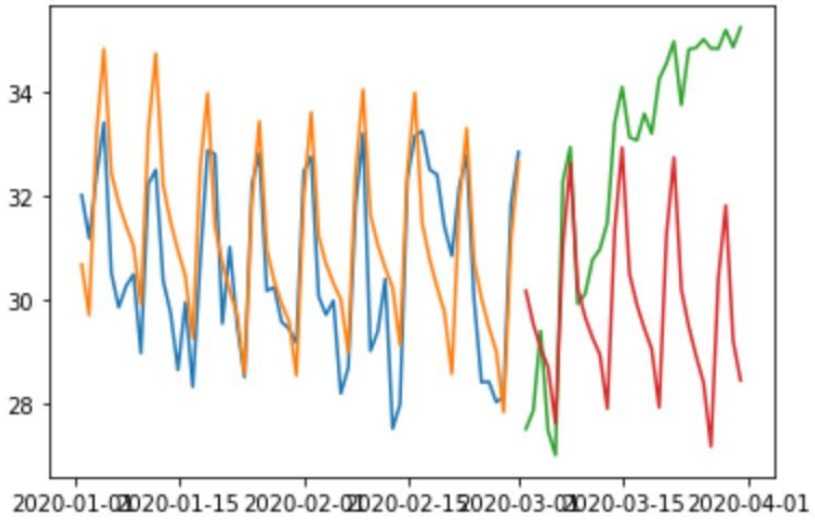
- Randomly selected 15 nodes.
- Formed 15 interconnected roads.

Real road example



Real road example

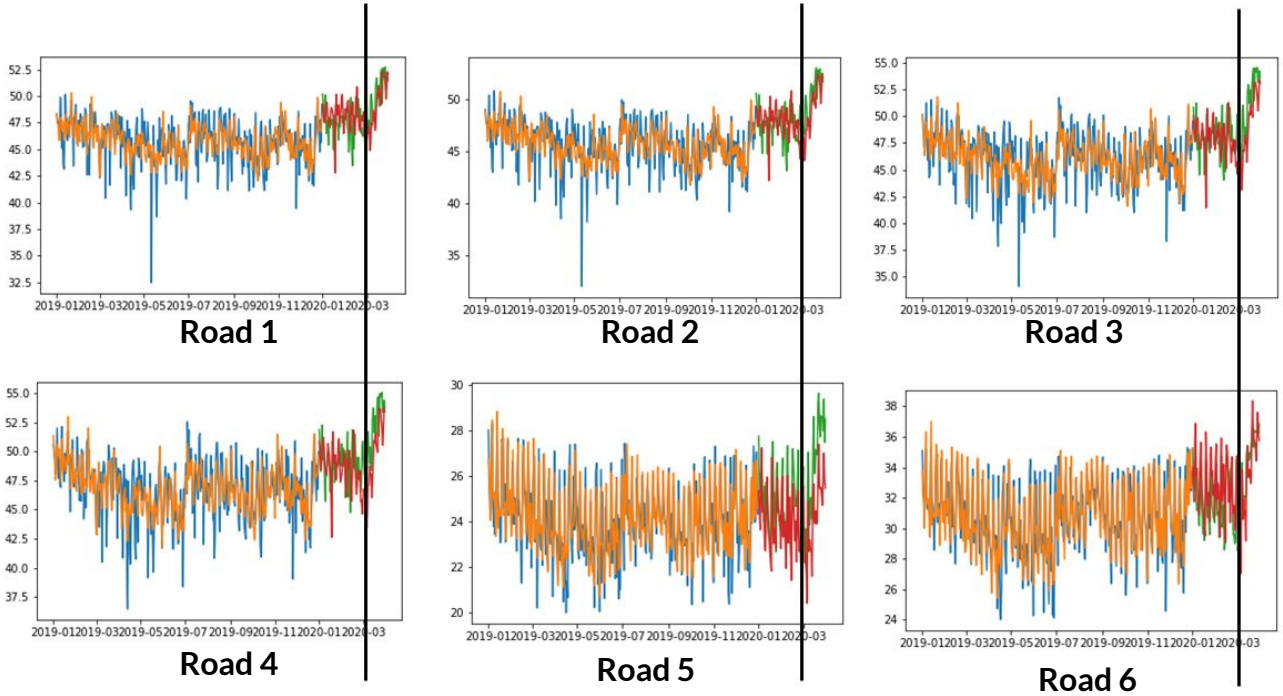
Road 1:



- **Threshold: March 1st.**
- **First part of testing data:**
  - **Blue line**: Original data
  - **Orange line**: Prediction
  - **Error: 10.20**
- **Second part of testing data:**
  - **Green line**: Original data
  - **Red line**: Prediction
  - **Error: 19.63**

Real road example

After add new feature(crash data):



Conclusion and Further Steps

- **Conclusion:**
  - Speed prediction model
  - Weather features we selected has no obvious improvement on result of prediction model while traffic collisions did

- **Further steps:**
  - Try more features in the weather dataset and use the weather forecasting data.
  - Try to find out some other feature that may impact the speed prediction result.(Pandemic Impact, etc.)
  - User input queries and application

Q&A

Thank you for listening to our presentation.

Are there are any questions?