Forecasting Bixi rentals

By: Rima Bazerji

Introduction & Problem

Facts about BIXI

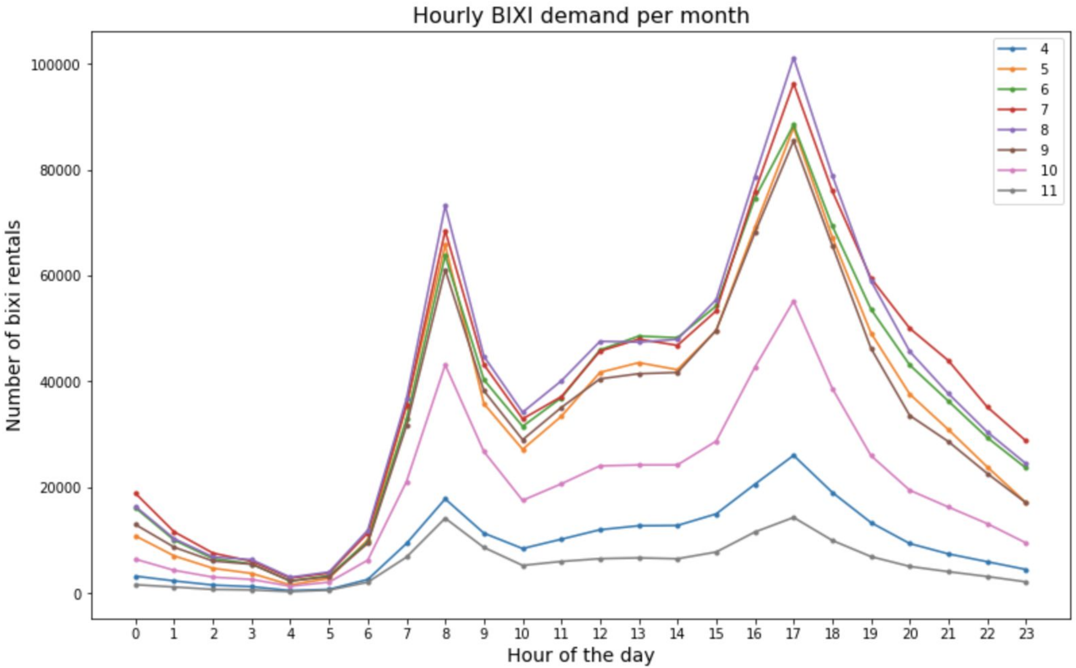
- 🚲 Bixi is a nonprofit organization created by the city of Montreal to manage its bike sharing system
- 🚲 7,250 bikes, 540 stations and 5,3M rides taken in 2018
- 🚲 Bike Sharing systems aim to reduce congestion, air and noise pollution in urban areas
- 🚲 One of the downside has been the lack of bike supplies for common routes and poor rebalancing of docks

Problem & Data

- 🚲 Forecasting hourly demand is a starting point to predicting at station level and help with dock rebalancing efforts as bikes tend to accumulate downtown
- 🚲 Data :
  - bixi.com
  - wunderground.com
  - www.tpsgc-pwgsc.gc.ca/

Exploratory data analysis

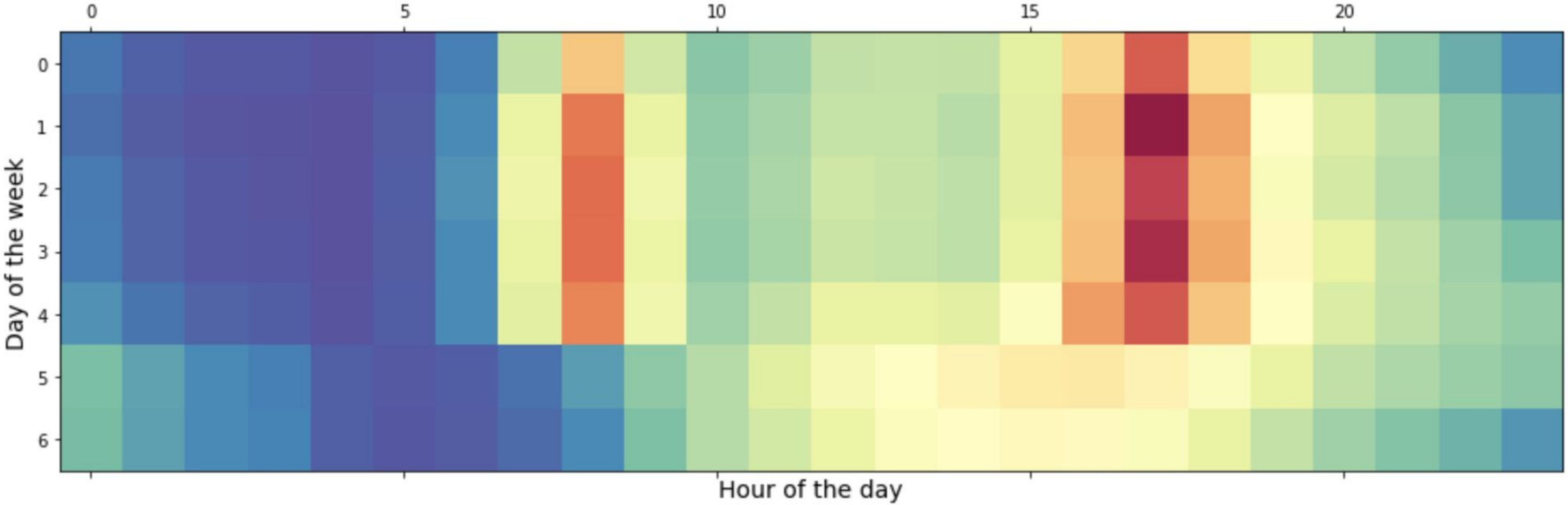
- Monthly pattern & Hour of the day usage



Exploratory data analysis

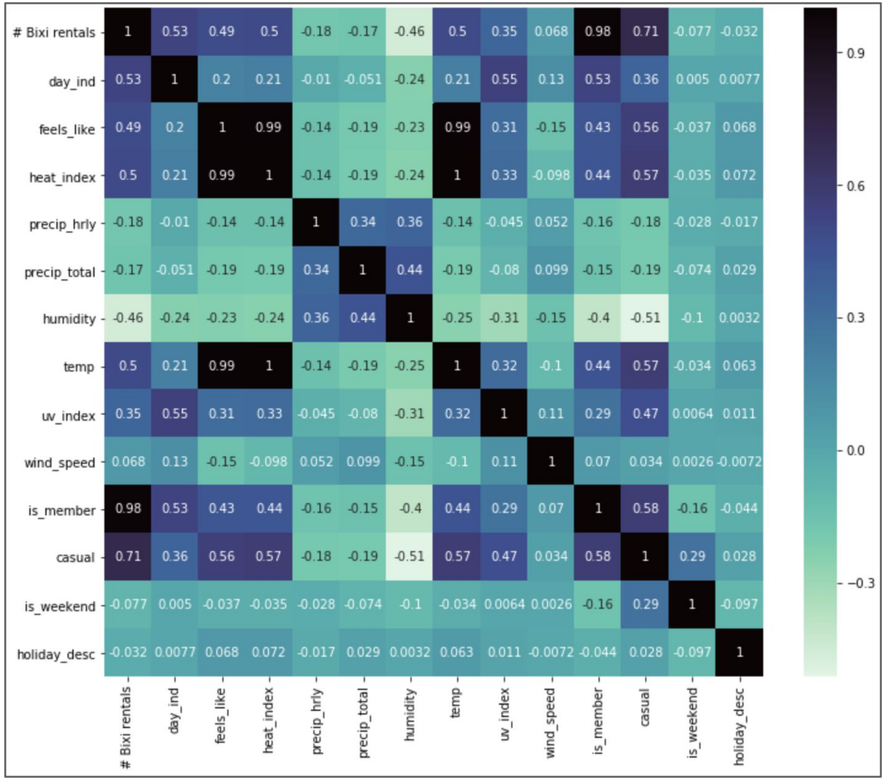
- Weekdays vs Weekends

- Rush Hour vs Not



Exploratory data analysis

- Casual members' behaviour is 10% more correlated with weather features
- Feels likes, Relative humidity, Uv index are amongst the most correlated with rentals
- Missing values (Backward fill)

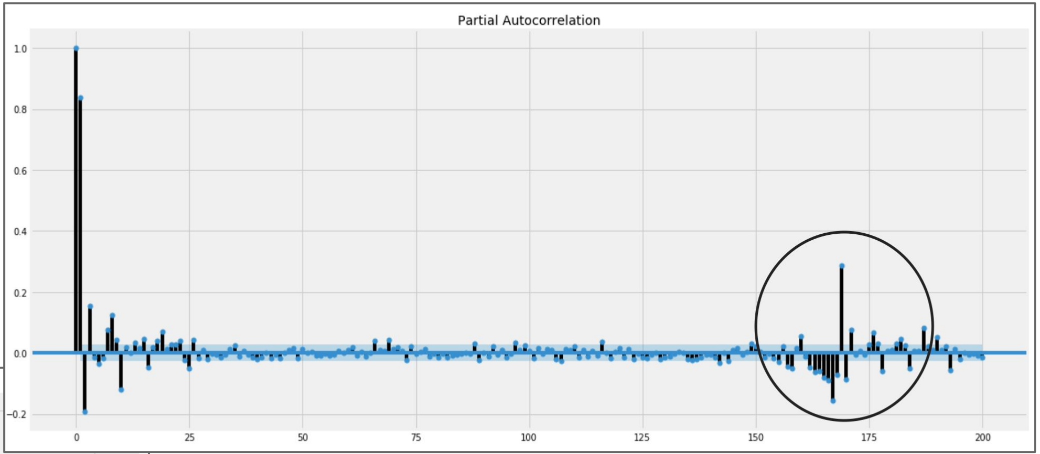


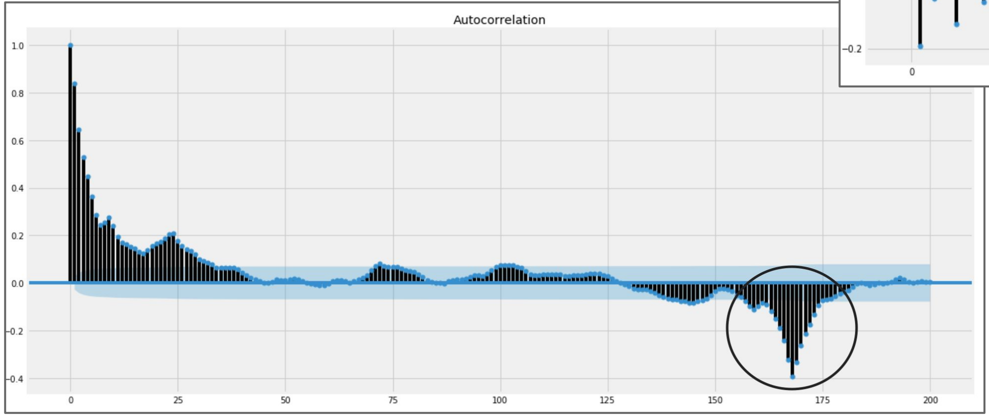
Modeling Process

- Modeling
  - Time dependent
    - Time Series - SARIMAX
    - FB Prophet
  - Time independent
    - Random Forest Regressor
    - XGBoost Regressor

Base Model 1A: SARIMAX (1,0,2)(1,0,1)24

- AdFuller test for Stationarity ✅
- Daily + Weekly seasonality
- Shifted distribution by 1 week (168 hrs)
- m = 24
- PACF to set AR parameters p and P





- ACF to set MA parameters q and Q
- Train test split ~ 20%:
  - Test set starts September 21, 2018

Model 1B: SARIMAX (1,0,2)(2,0,3)24

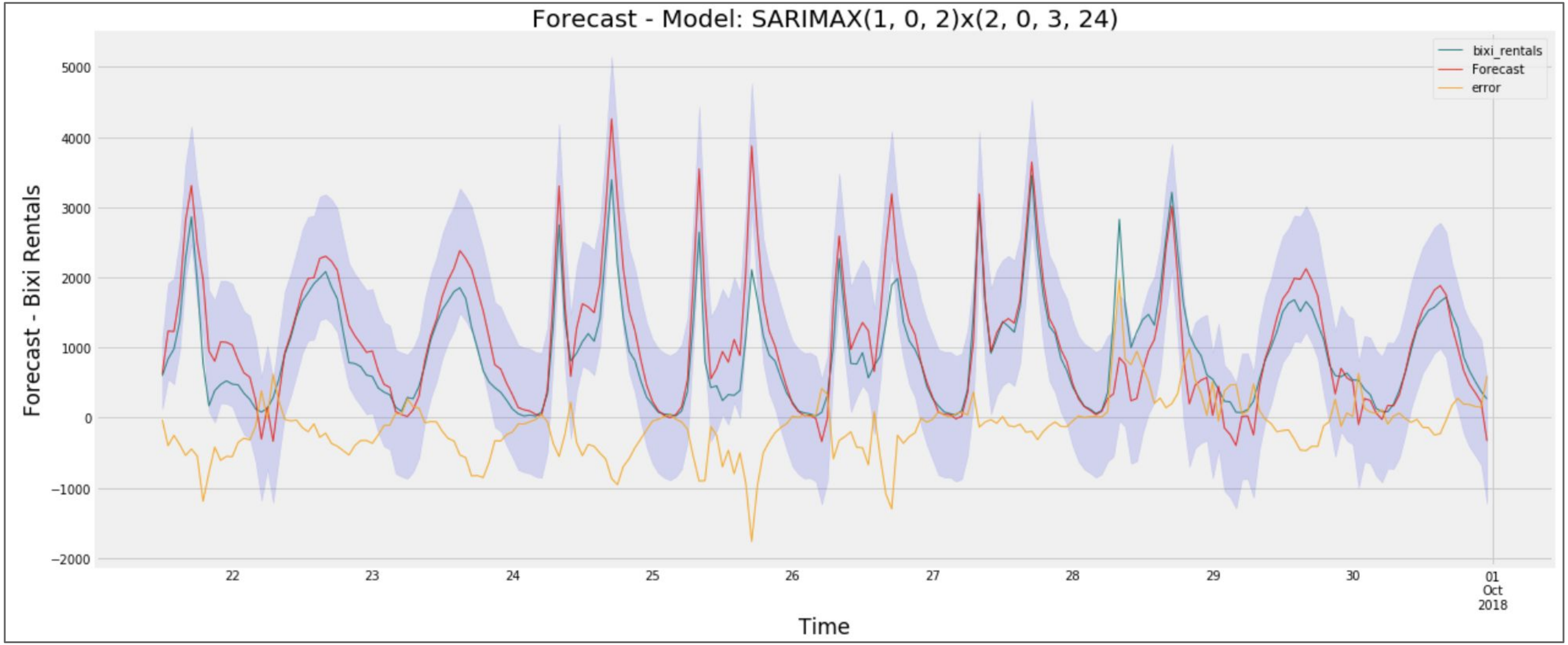
- Optimized SARIMAX parameters (GridSearch and minimize AIC) – on the right
- Base Model – below

	coef	std err	z	P>|z|
humidity	-3.0901	0.470	-6.580	0.000
feels_like	3.3516	0.426	7.874	0.000
precip_hrly	-28.7542	15.054	-1.910	0.056
ar.L1	0.7428	0.008	89.143	0.000
ma.L1	0.3150	0.014	21.848	0.000
ar.S.L24	0.9850	0.008	117.691	0.000
ma.S.L24	-0.9999	0.605	-1.654	0.098
sigma2	5.855e+04	3.5e+04	1.672	0.094

	coef	std err	z	P>|z|
humidity	-4.3268	0.501	-8.639	0.000
feels_like	5.7809	0.568	10.186	0.000
precip_hrly	-425.5591	11.893	-35.783	0.000
ar.L1	0.8194	0.014	57.286	0.000
ma.L1	0.1047	0.017	6.210	0.000
ma.L2	-0.1271	0.018	-6.877	0.000
ar.S.L24	-0.7948	0.035	-22.670	0.000
ar.S.L48	-0.7276	0.037	-19.920	0.000
ma.S.L24	0.9042	0.037	24.188	0.000
ma.S.L48	0.9143	0.033	27.421	0.000
ma.S.L72	0.1918	0.015	12.464	0.000
sigma2	6.546e+04	754.544	86.758	0.000

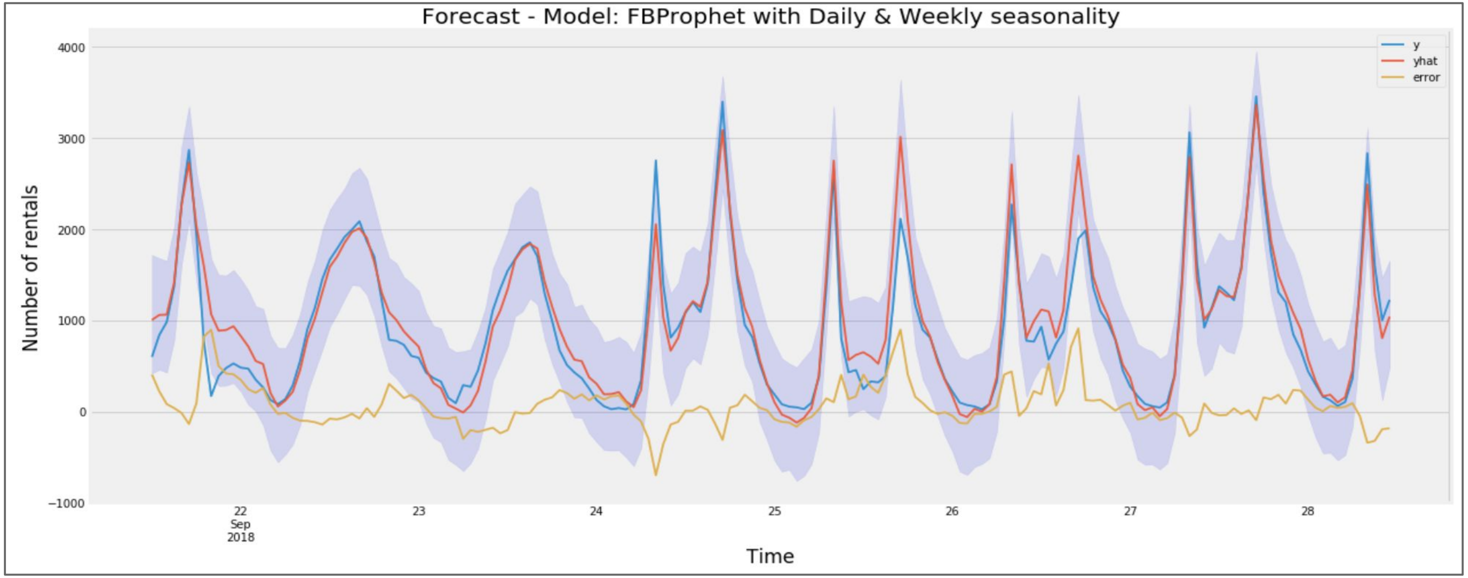
Model 1B: SARIMAX (1,0,2)(2,0,3)24

- Forecast



Model 2: FB Prophet

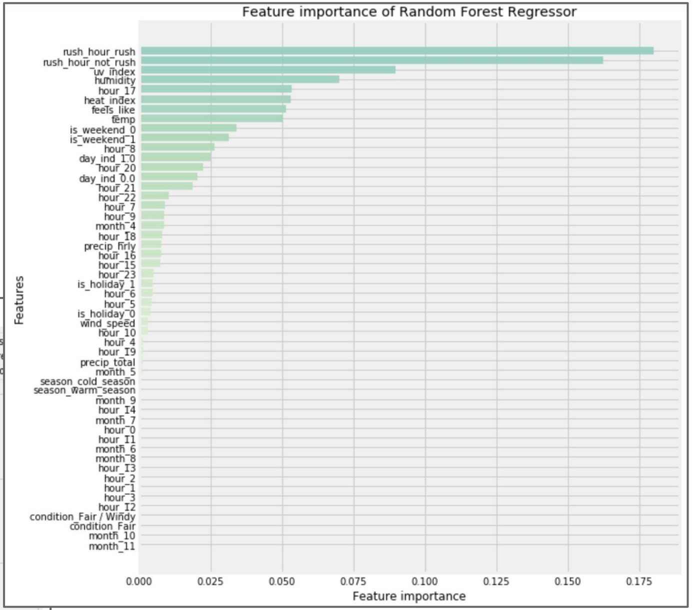
- Curve fitting problem where time is a regressor
- GridSearch hyperparameters:
- Daily/Weekly Fourier Order (Increased -> Model can fit at higher frequency)
- Change point scale (trend)

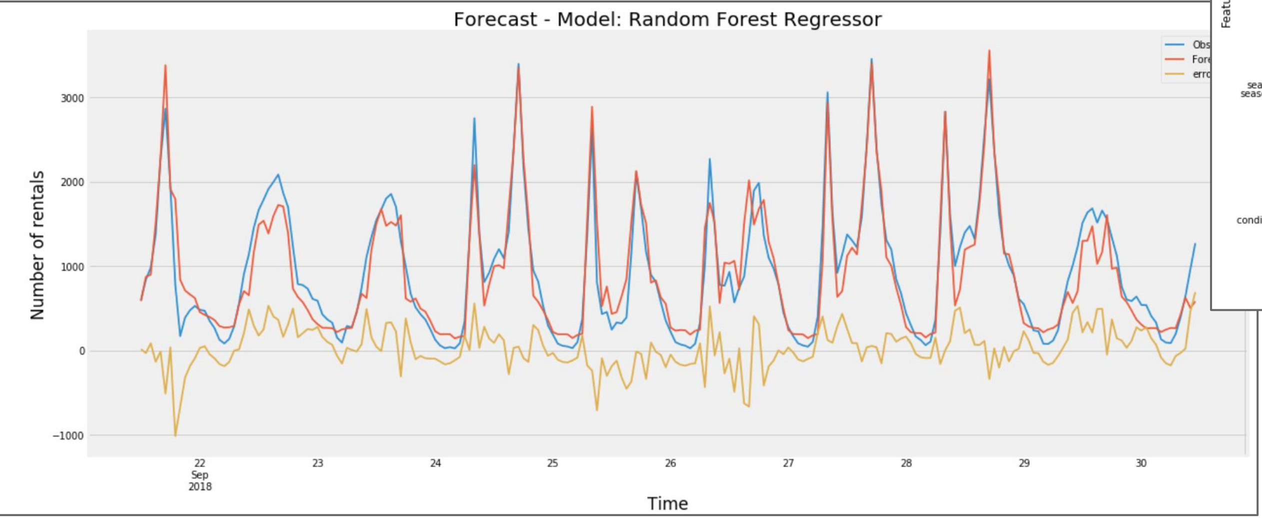


Model 3: Random Forest Regressor

- Decompose Time into Descriptive Features

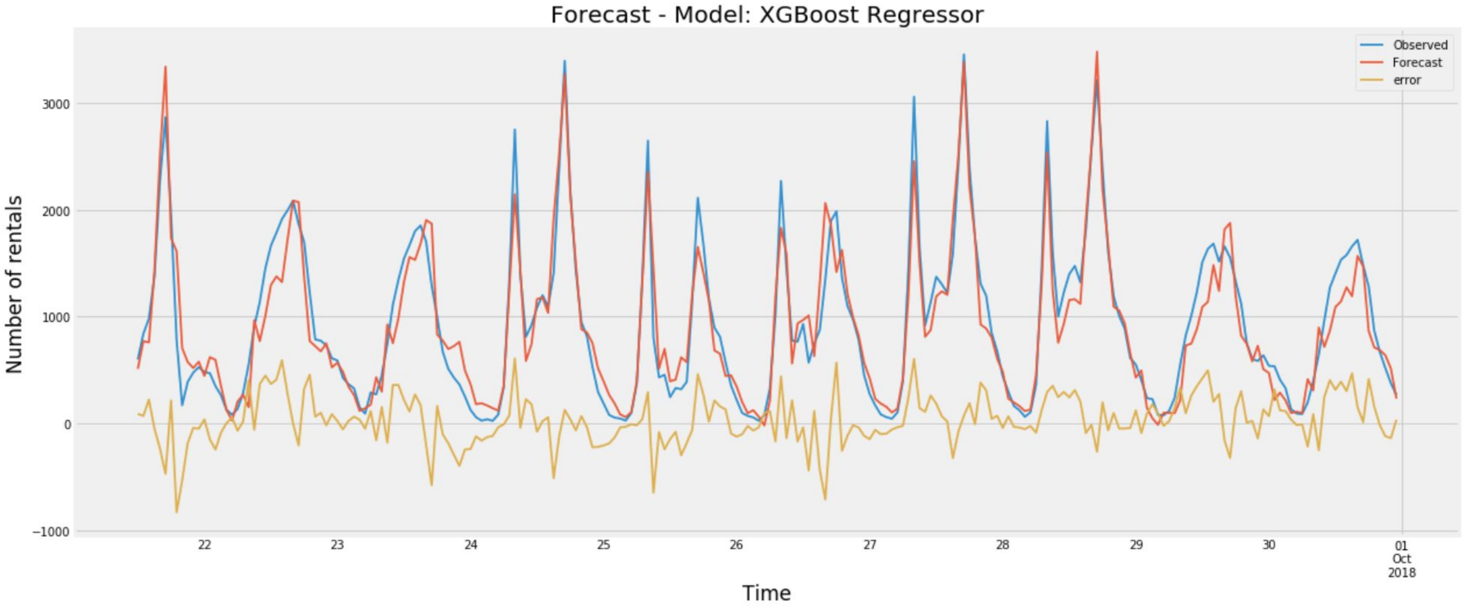
  – Hour, month, rush hour, off-peak season etc.
- Feature engineering based on EDA
- GridSearch





Model 4: XGBoost

- GridSearch with pruned trees to avoid overfitting (max depth & Gamma)
- Best model based on RMSE



Stacked Model with Linear Regression

- Weighted average with Linear Regression, where:

  X = Predictions from all models &

  y = Observed

- Weights:

```
RandomForestRegressor: 0.2955404138782251
XGBoost: 0.6192844978845874
SARIMAX: 0.0857284784492l101
FB_Prophet: 0.12688710988303
```

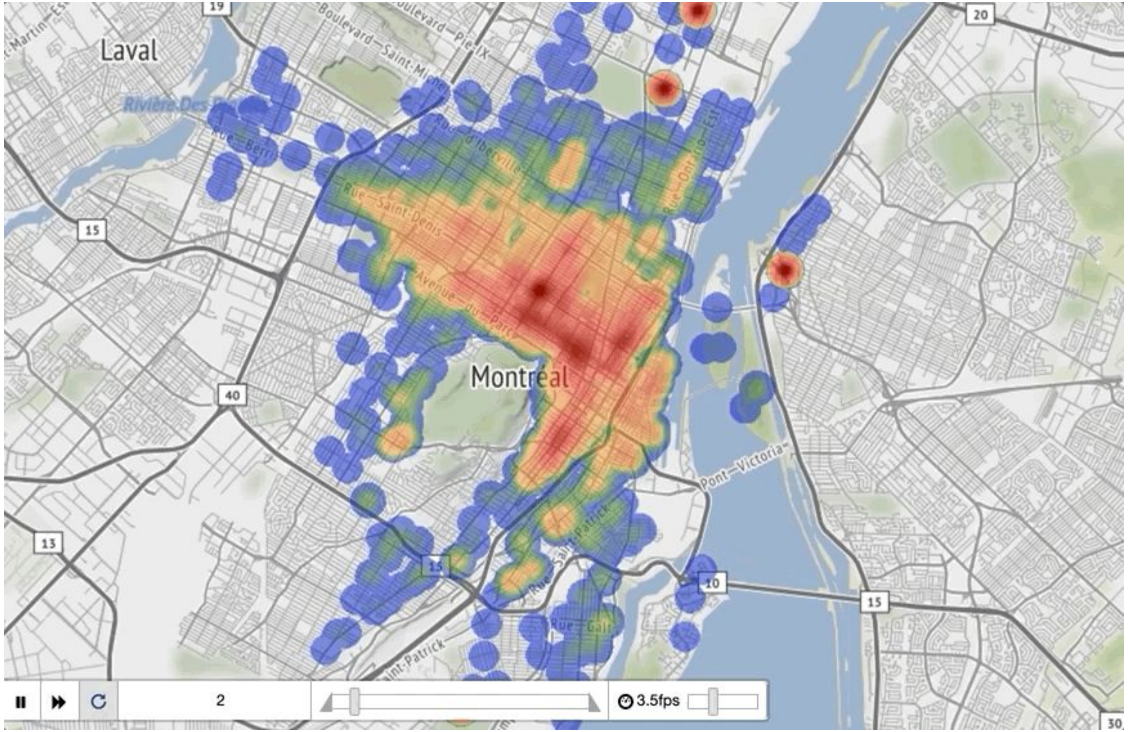
- Final Model: 0.3* RF + 0.62*XGBoost + 0.09*SARIMAX + 0.13*FB

Summary and Stacked model

Model	RMSE	R2
SARIMAX(1,0,2)(2,0,3)24	372.53	0.75
FB Prophet	405.55	0.72
Random Forest Regressor	266.48	0.80
XG Boost Regressor	255.65	0.85
Stacked Model	**251.33**	**0.86**

Takeaways & Next Steps

- General model can be applied to other cities
- Station Level demand
- Rebalancing efforts
- Typical Pickup & Delivery problem



Thank you