PREDICTING CHICAGO CRIMES

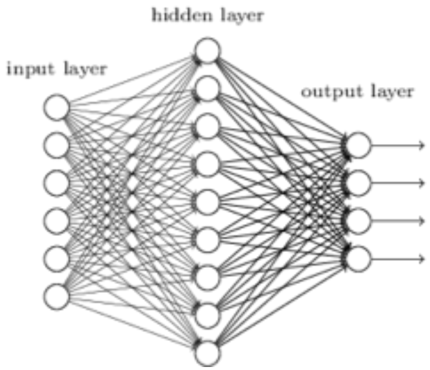
AMIT DINGARE

EXECUTIVE SUMMARY

- Chicago crimes were predicted using two different models;
  - a time series model based on Facebook's Prophet model and
  - a dense NN model with one hidden layer
- As expected, the **NN model outperformed** the time series model
- With **data augmented from external sources** such as median income by zip code and weather information such as temperature and humidity, the NN model can be enhanced
- If **explainability** of the model is more important than it's predictive power, the NN model can replaced by a simpler model or a technique such **LIME** (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations) can be applied

WHY I CHOSE THIS PROJECT?

- A fan of American police dramas
- The project allowed me to test two concepts that I am currently learning; Facebook's Prophet model for Forecasting & Neural Networks



- The project also allows me to demonstrate my comfort with leveraging APIs for data pull

BACKGROUND

A PREDICTIVE MODEL BUILT WITH THE CRIME DATA HAS A LARGE NUMBER OF USES FOR CPD

- Crime rate in Chicago has been substantially higher than the US average. To bring down the crime, CPD has been tracking various crimes diligently.
- A predictive model built on this data could help with:
  - **Staffing** the right number of Police Officers in various areas
  - **Equipping** Police Officers with appropriate weaponry and equipment depending on the crime rate in crime areas
  - Deciding routes of **patrol**
  - **Optimizing hours** for Police Officers to get the best outcome without causing personnel fatigue
  - And many more…

DATA PULL AND EDA

DATA EXTRACTION AND DATA SPARSITY WERE HANDLED PRIOR TO EDA

- The data was pulled through a **Socreta Open Data API**.
- In order to use the most **relevant data for the analysis**, data later than 2010 was pulled.
- Primary Crime Type indicates crime categories used while ascribing the crime. In order to reduce data sparsity problems, crime categories with less than or equal to 5% of total crimes were combined into a "Misc" crime category.
- From the datetime field, additional features such as day of week, hour and week are extracted

CRIME RATE IS DECREASING YOY WITH MID-YEAR SEASONALITY

- Overall crime rate has been **decreasing** year over year, but, has stayed about flat in last 3 years.

- Crimes seem to demonstrate a consistently **seasonality**; going up during the middle of the year and then down.

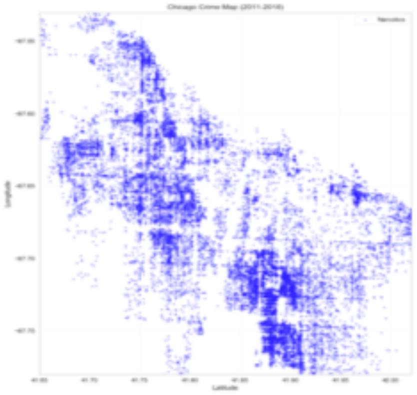
GEOGRAPHIC CONCENTRATION OF CERTAIN CRIMES HAS BEEN OBSERVED

Thef t

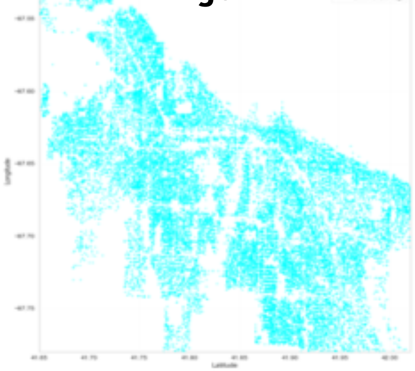
Misc. Crimes

Battery

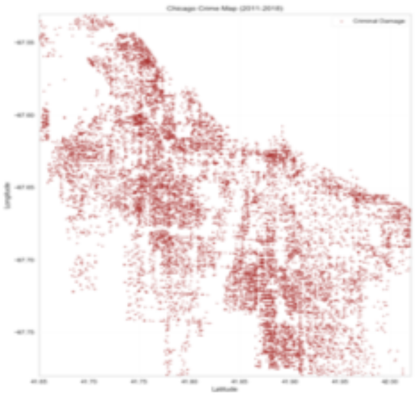
Narcotics



Criminal Damage

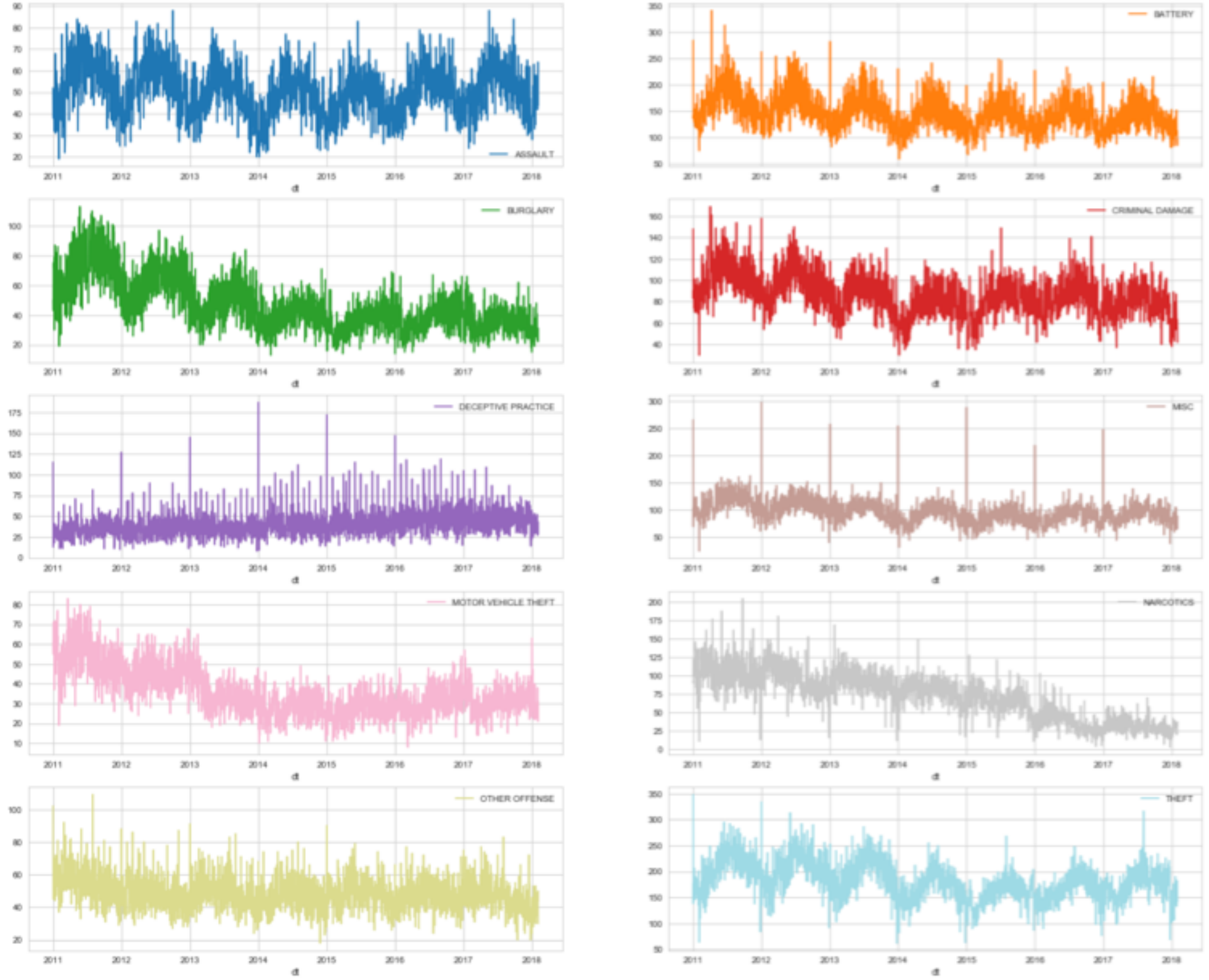


Assault



- Theft seems to concentrated at the lakeshore (probably rich areas)
- Criminal Damage and Misc. crimes seem to spread throughout the city.
- Narcotics, Battery and assault seem to be concentrated on the south side

PATTERNS EMERGE WHEN LOOKED ON A YoY BASIS FOR EACH OF CRIME CATEGORIES



- Most crime categories show a downward trend; most noticeably Narcotics, Motor Vehicle Trend and Burglary related crimes.
- Assault, Theft and Criminal Damage have stayed about the same or declined slightly.
- Surprisingly, Deceptive Practices have seen an upward trend.

MODELING

TIME SERIES AND NEURAL NETWORKS ARE THE TWO TECHNIQUES TRIED

- Time Series Technique leveraging Facebook's Prophet model:

- Facebook's Prophet model is $y(t) = g(t) + s(t) + h(t) + \epsilon_t.$ Based on the **GAM** technique
  - where g(t) is the trend function
  - S(t) is the seasonality function for seasonality **(yearly and weekly)**
  - h(t) is function to account for effect of holidays **(can be provided from outside)**

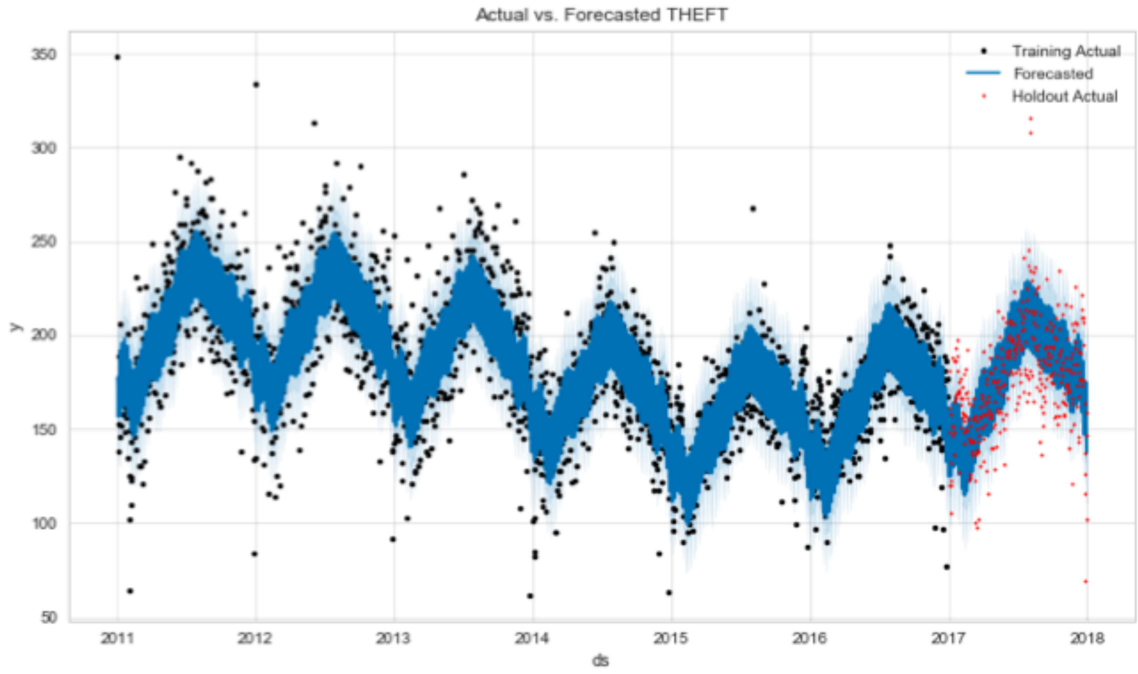
- Neural Network Technique:
  - A fully connected NN with one hidden layer
  - Relu activation function

[* https://peerj.com/preprints/3190.pdf](https://peerj.com/preprints/3190.pdf)

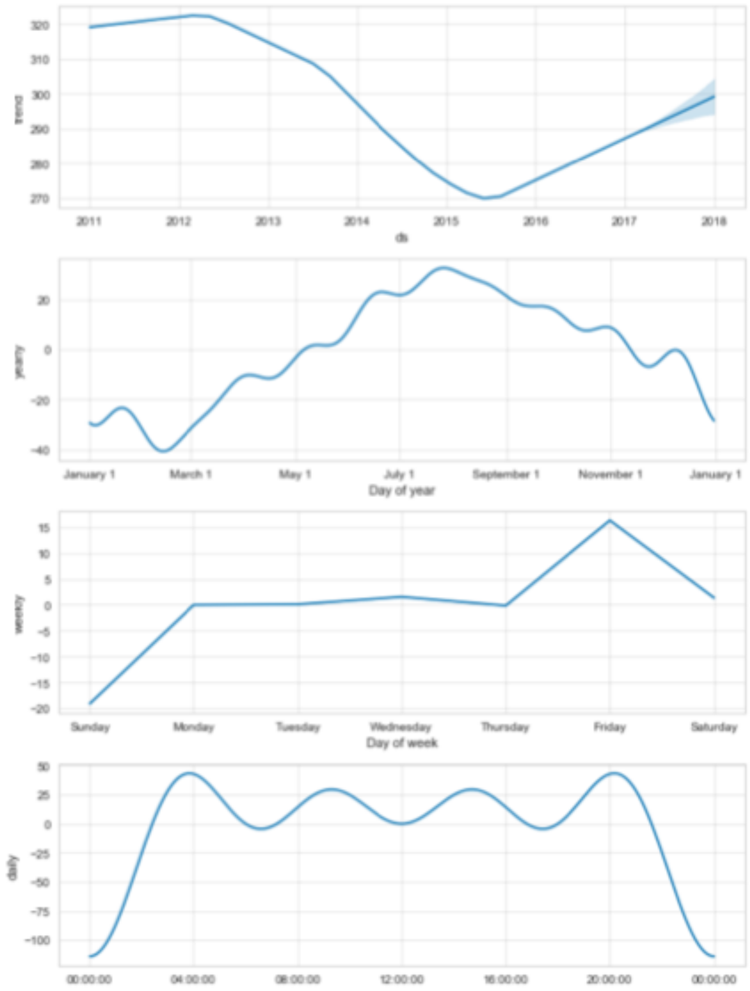
OPTION 1: TIME SERIES BASED PROPHET MODEL

- Data for 2011-2016 is used to build forecast model
- 2017 daily points are predicted using the forecast model
- Test set RMSE is calculated at daily and weekly levels and used as a comparative metric
- Forecast is plotted with respect to actual and forecast components are calculated

THEFTS

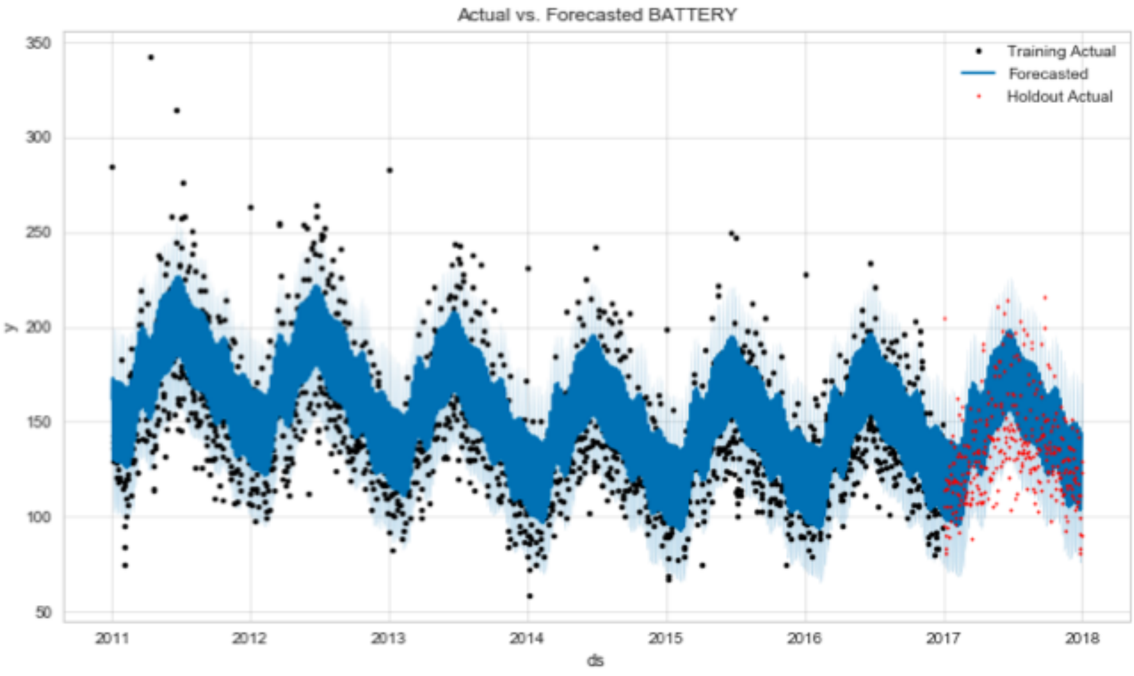


Weekly RMSE: 1214

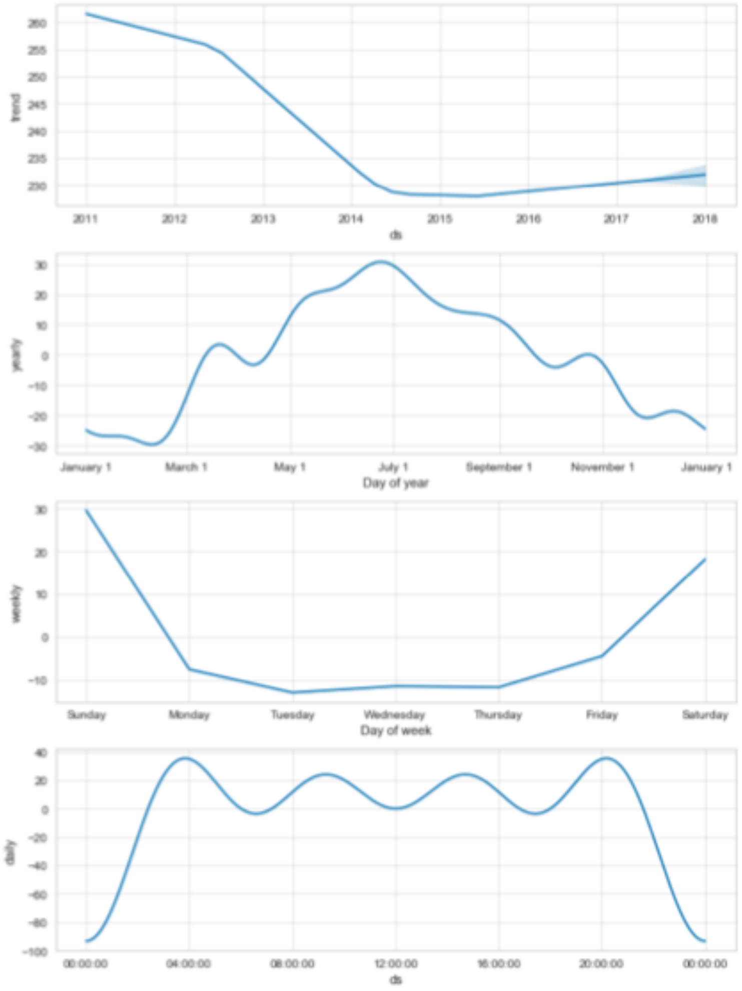


- YoY trend seems to be upward after going down until mid-2015.
- Thefts increase during summer.
- Fridays seem to be the most prone day for thefts.
- Daytimes are more prone periods for thefts.
- While, the RMSE seems to large, the 95% CI seems to capture a large percentage of thefts.

BATTERY

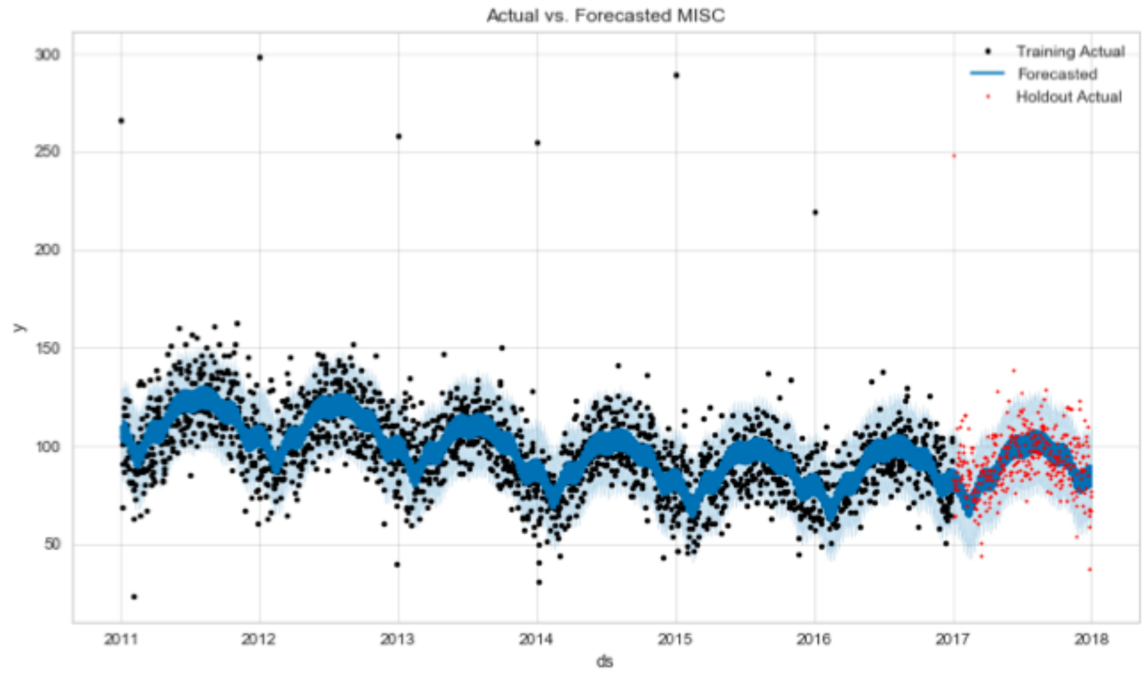


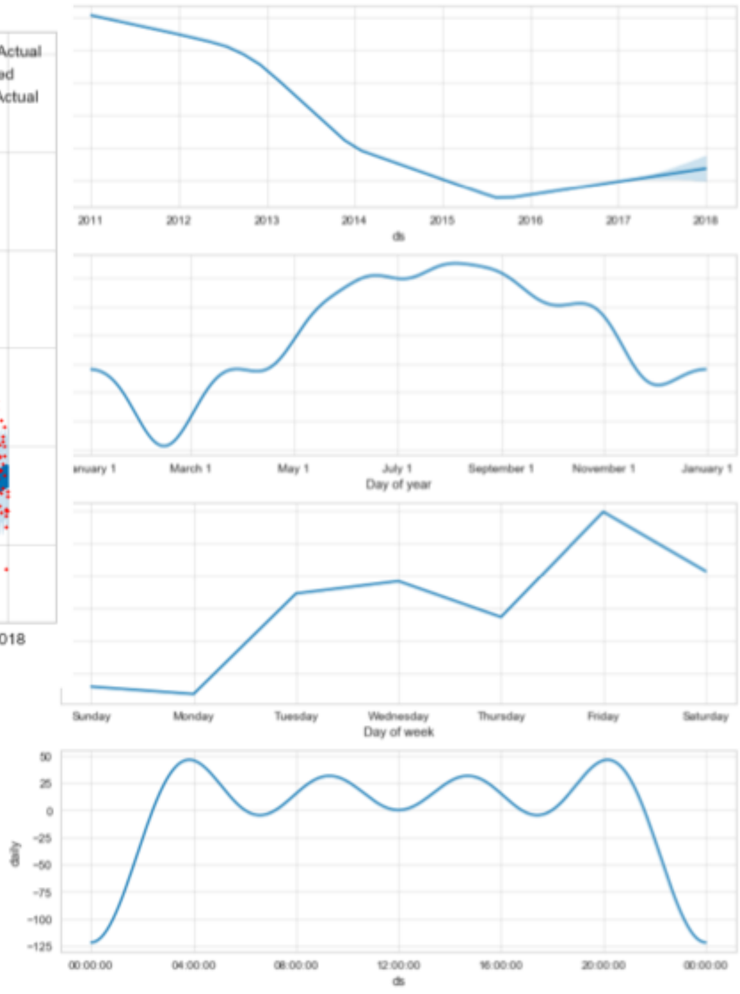
Weekly RMSE: 925



- YoY trend seems to be slight upward after going down until mid-2015.
- Battery increase during summer.
- Weekends seem to be the most prone day for battery.
- Daytimes are more prone periods for battery.
- The model seems to be biased a bit upward.

MISC.

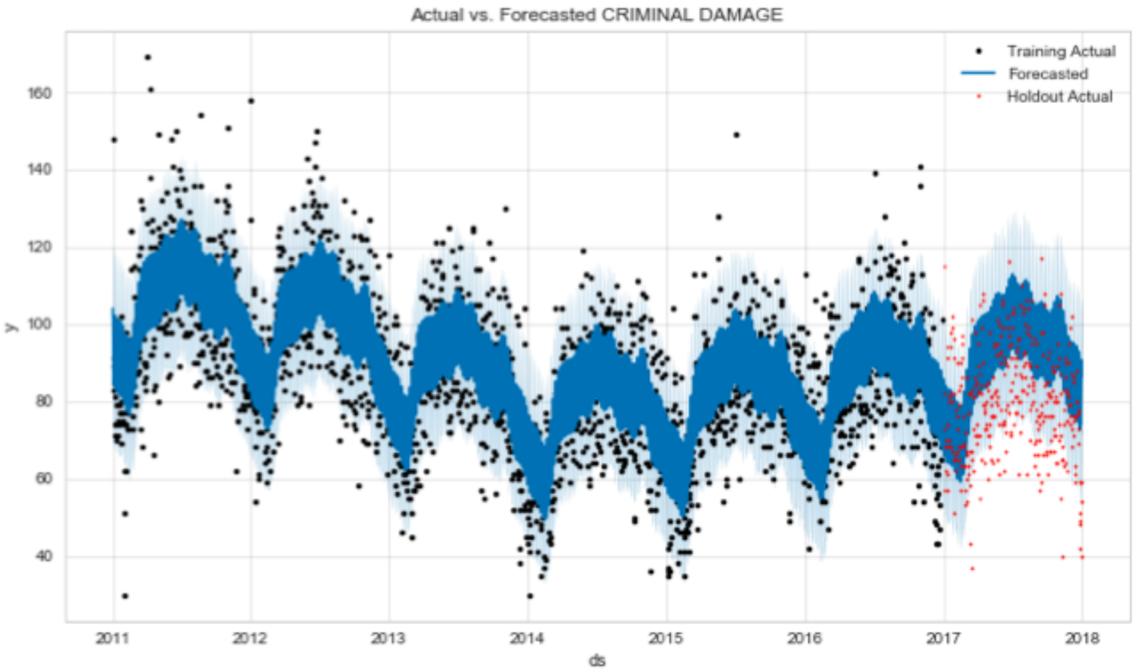




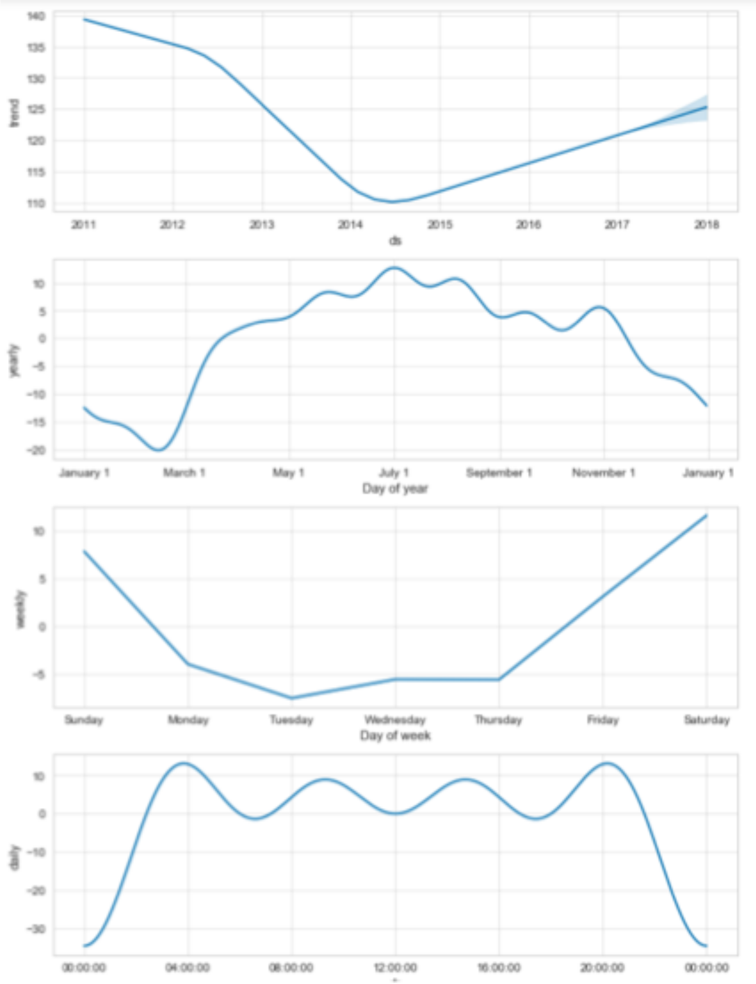
Weekly RMSE: 623

- YoY trend seems to be slightly downward
- Misc. crimes seem to increase during summer/fall.
- Friday & Saturdays seem to be the most prone day for Misc. crimes.
- Daytimes are more prone periods for misc. crimes.
- 95% CI of the model seems to capture Misc. crimes pretty well.

CRIMINAL DAMAGE

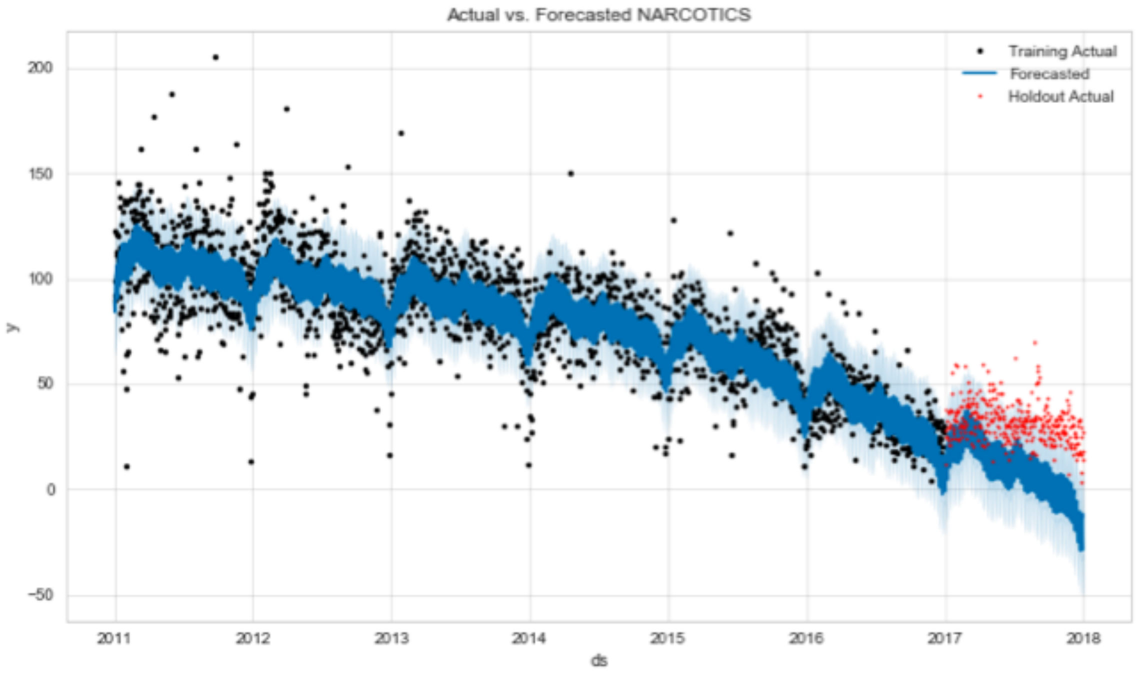


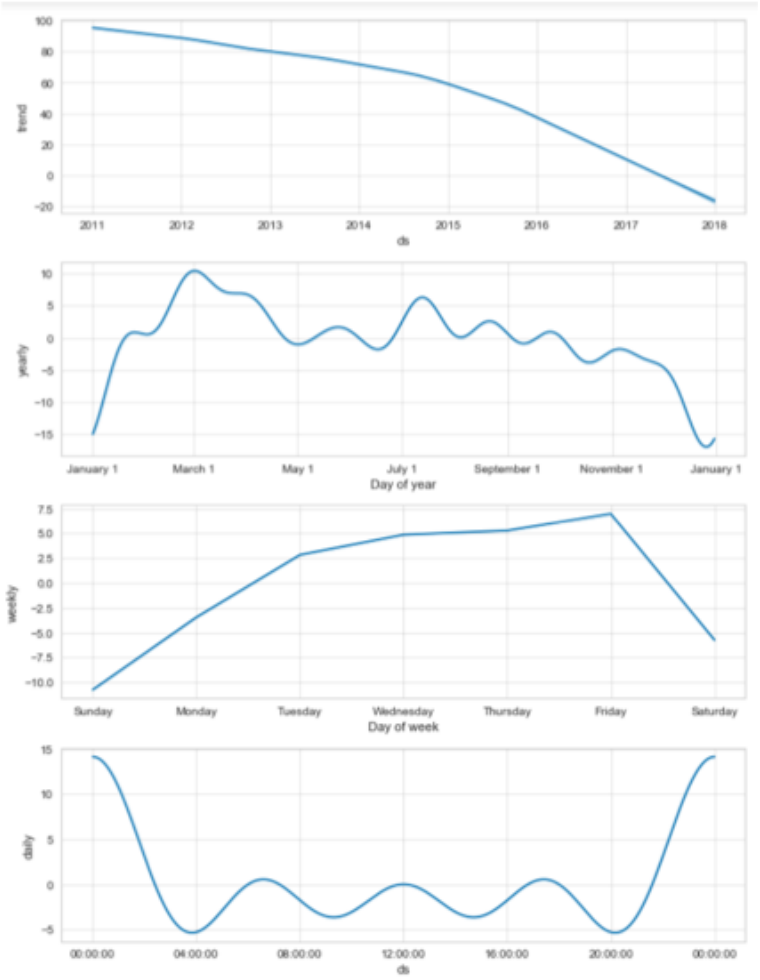
Weekly RMSE: 534



- YoY trend seems to be trending upward since mid-2014.
- Misc. crimes seem to increase during summer/fall.
- Friday, Saturday and Sundays seem to be the most prone day for criminal damages.
- Daytimes are more prone periods for criminal damages.
- The model is biased upwards.

NARCOTICS

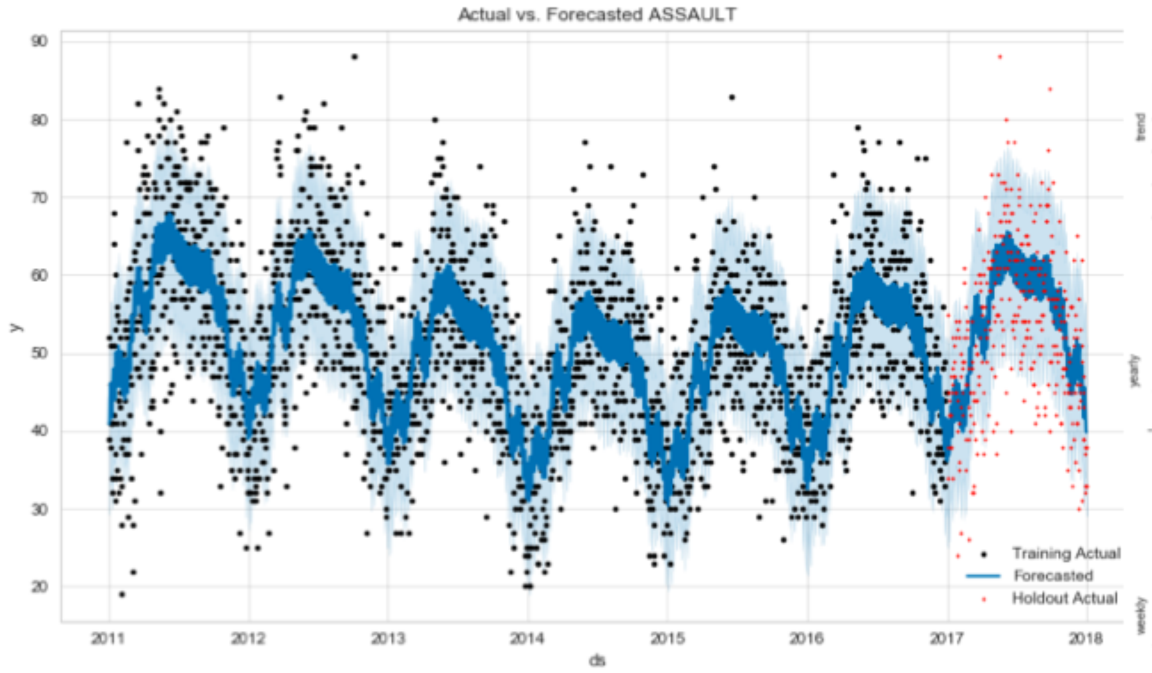


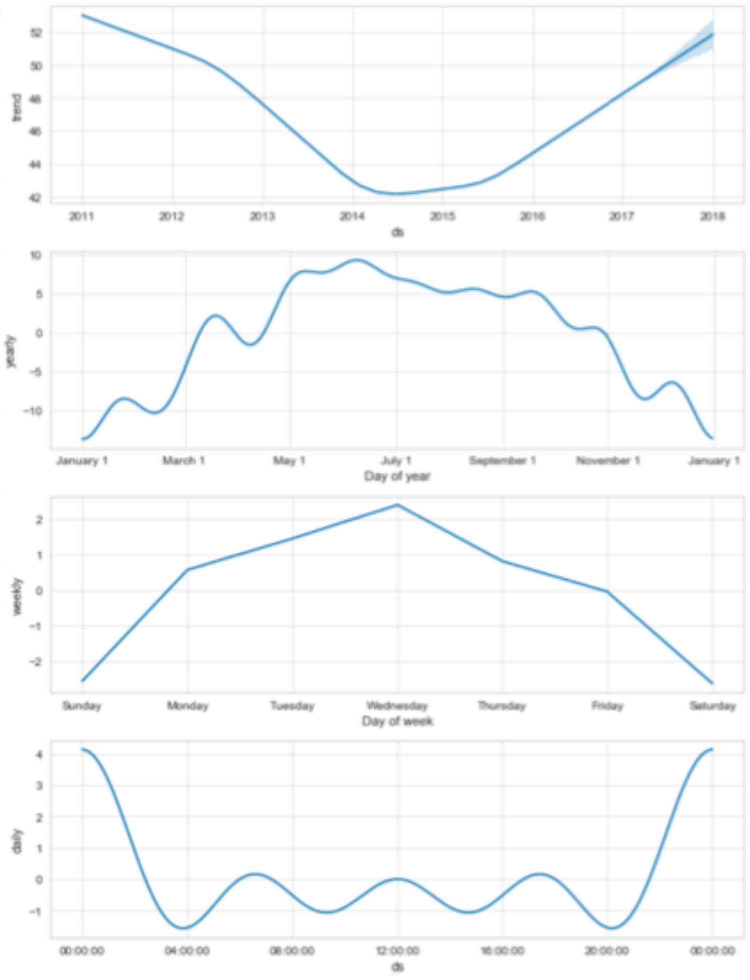


Weekly RMSE: 200

- YoY trend seems to be trending downward since 2011.
- Narcotics crimes seem to be around throughout the year.
- Narcotics crimes seem to increase during the week and then fall on the weekend.
- Nights (8 PM – 4 AM) are more prone periods for Narcotics.
- The model is biased downwards.

ASSAULT

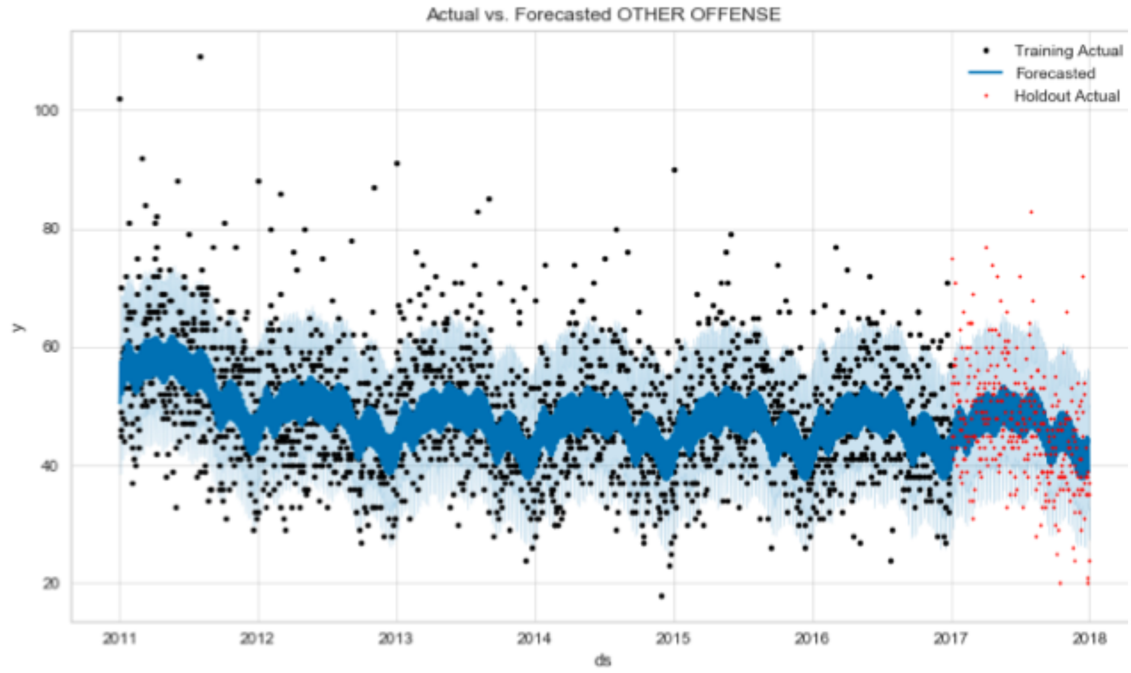


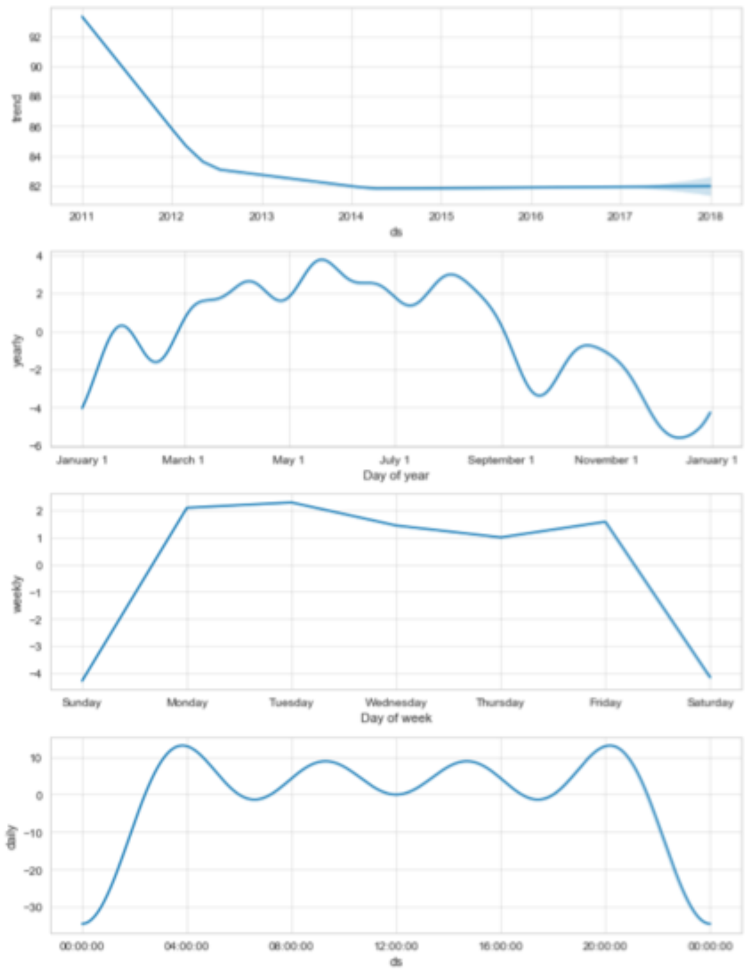


Weekly RMSE: 348

- YoY trend seems to be trending upwards since mid-2014 after going down.
- Assault crimes seem to be go up till summer and then go down.
- Assault crimes seem to increase during the week and then fall on the weekend.
- Nights (8 PM – 4 AM) are more prone periods for Narcotics.
- 95% CI seem to capture assaults correctly.

OTHER OFFENSE

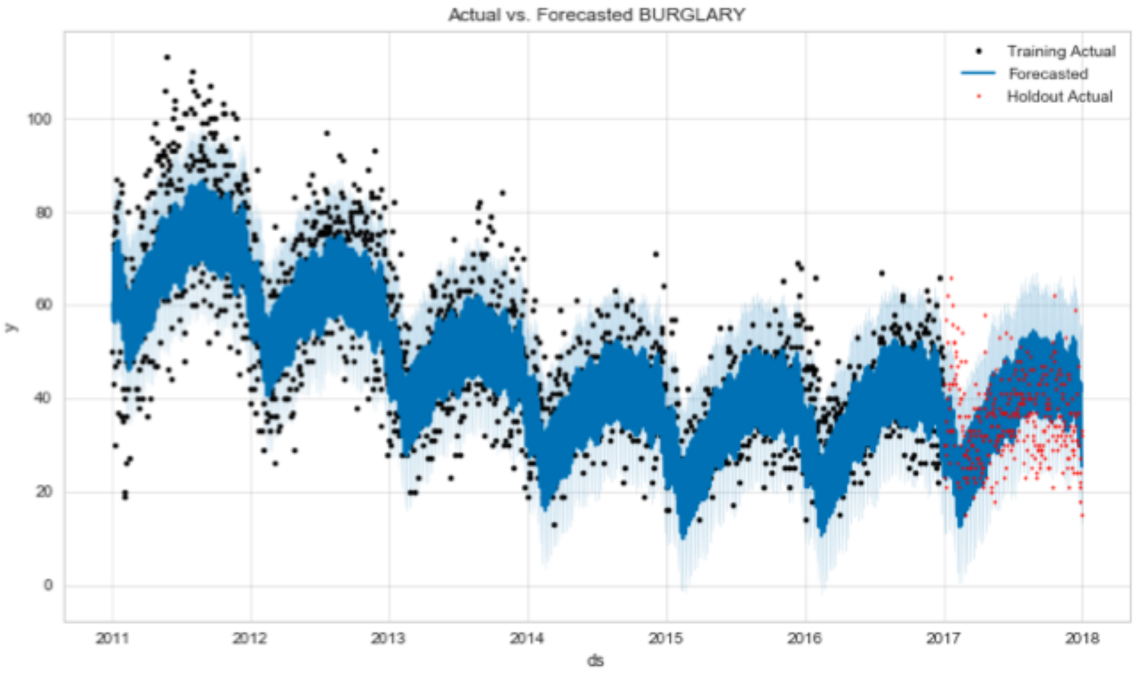


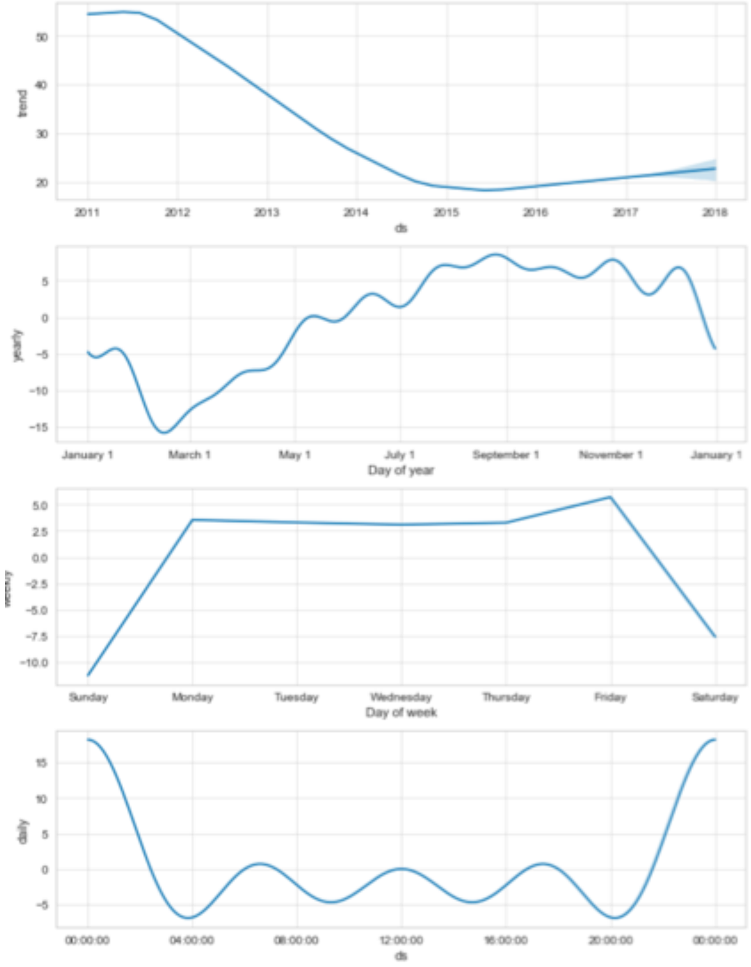


Weekly RMSE: 305

- YoY trend trended downwards until 2014 and then stayed flat.
- Crimes seem to be around throughout the year.
- Narcotics crimes seem to increase during the week and then fall on the weekend.
- Daytimes are more prone periods for Other Offenses.
- The model seems to be less accurate as a result of the large variance in the data.

BURGLARY

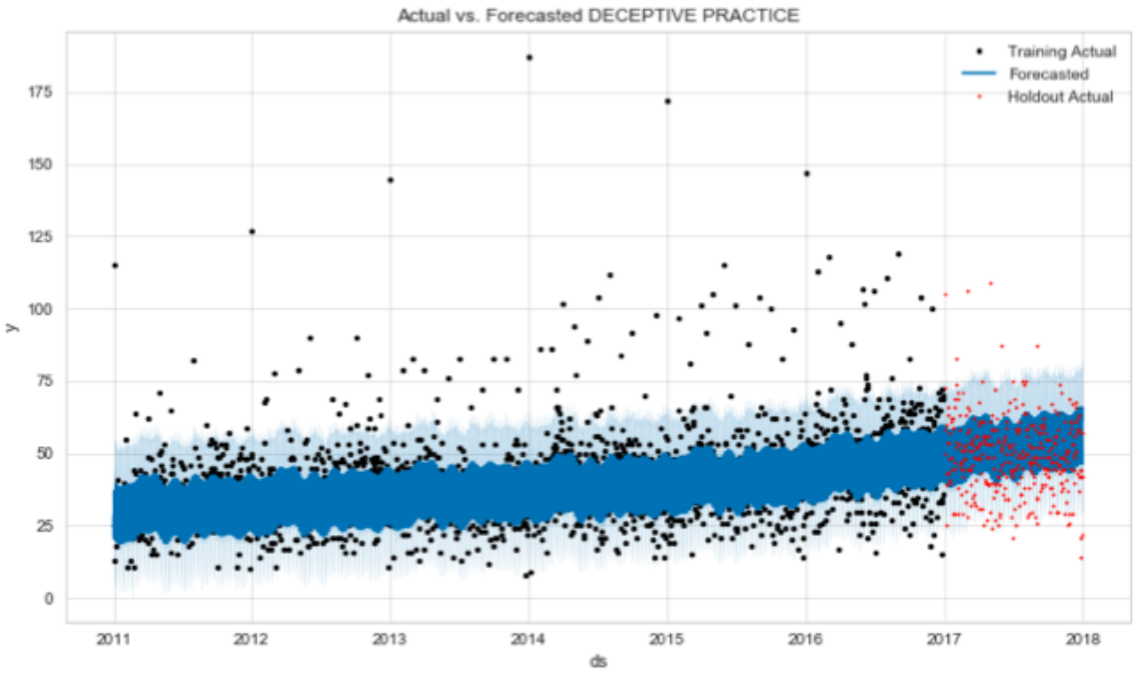




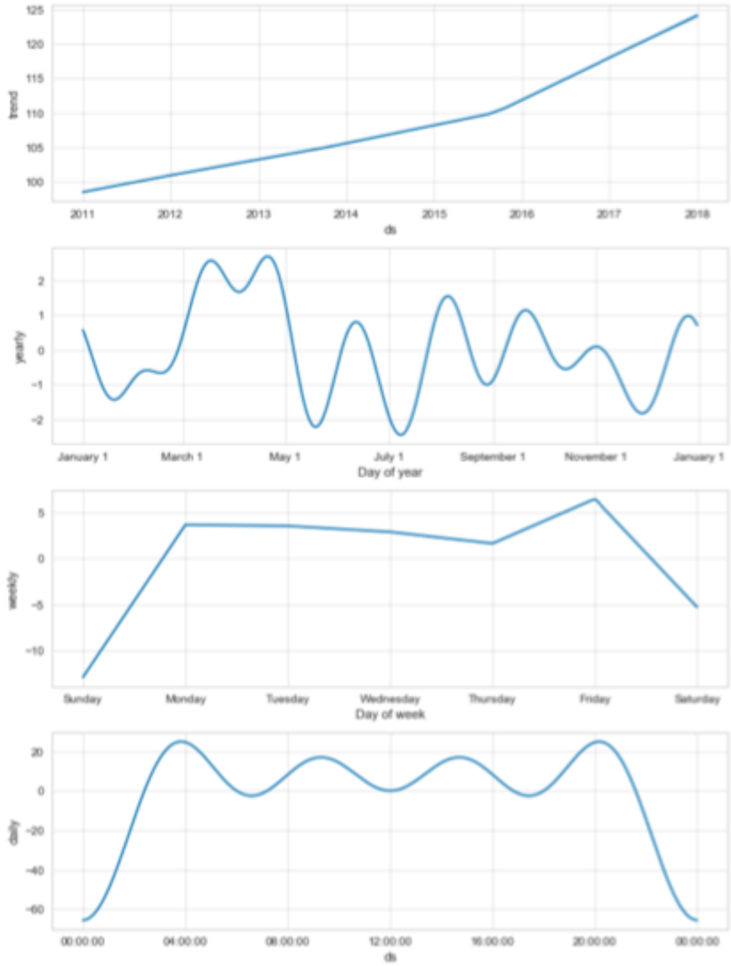
- YoY trend trended downwards until 2014 and then increasing slightly.
- Crimes seem to be going up starting Spring.
- Narcotics crimes seem to increase during the week and then fall on the weekend.
- Night times are more prone periods for burglaries.
- The model seems to be biased upwards.

Weekly RMSE: 225

DECEPTIVE PRACTICE

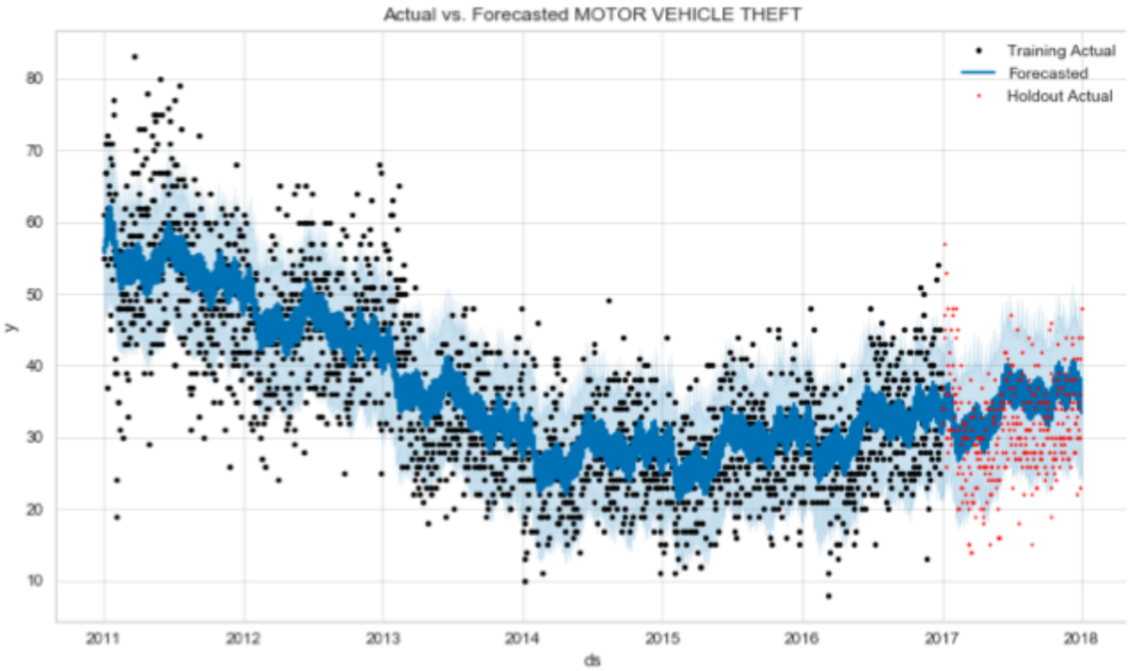


Weekly RMSE: 321

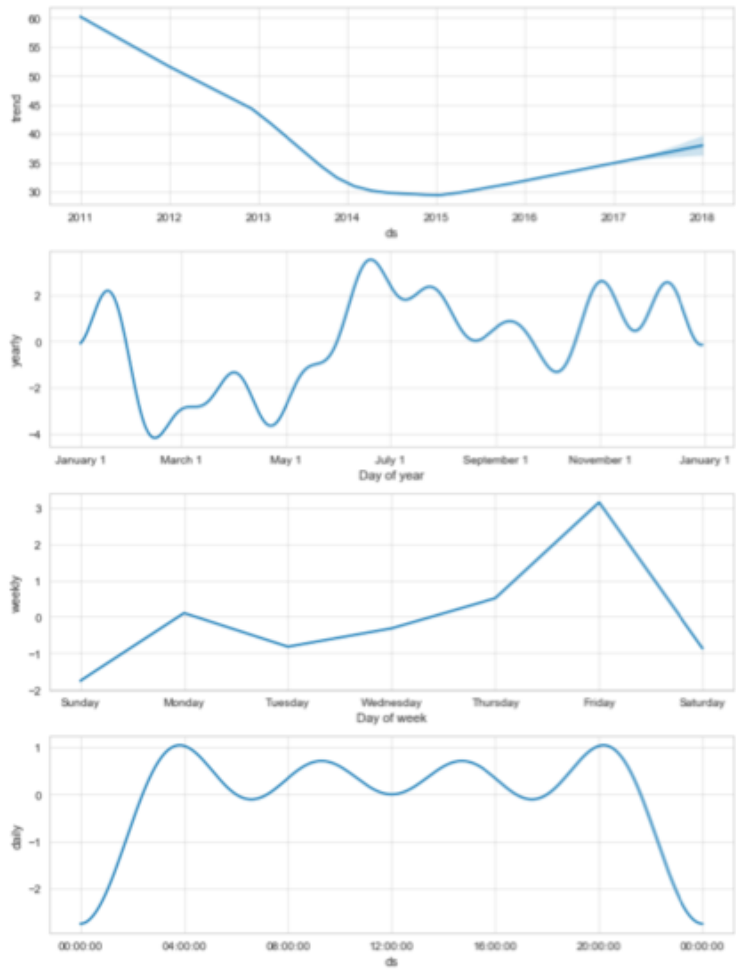


- YoY trend trended upwards.
- Crimes seem to be up and down throughout the year.
- Deceptive Practices seem to up during the week and then fall on the weekend.
- Daytimes are more prone periods for deceptive practices.
- The model seems to be biased slightly upwards.

MOTOR VEHICLE THEFT



Weekly RMSE: 196



- YoY trend trended upwards starting 2014.
- Crimes seem to be up and down throughout the year.
- Motor vehicle thefts are the highest on Fridays.
- Daytimes are more prone periods for motor vehicle thefts.
- The model seems to be biased upwards.

OPTION 2: NEURAL NETWORK BASED MODEL

- Used a dense network with 4 input dimensions; **hour, day of week, week of the year and beat** (representing location)
- The network has **one input layer of 4 inputs**, **one hidden layer with with 4 neurons** connected to the input layer with a '**relu**' activation function and the **output layer with one output value**.
- The model optimizes for mean squared error.
- **training data: 2011-2016, test data: 2017**
- Weekly RMSE are calculated for comparison

	RMSE	RMSE_Week
BATTERY	0.160268	5.577753
THEFT	0.282039	30.562912
NARCOTICS	0.259559	4.452092
ASSAULT	0.100309	2.378946
MISC	0.183790	8.048224
DECEPTIVE PRACTICE	0.131703	2.609011
OTHER OFFENSE	0.112722	2.542436
CRIMINAL DAMAGE	0.211461	4.339582
BURGLARY	0.132168	2.170177
MOTOR VEHICLE THEFT	0.107537	1.657921

Simple NN model outperforms the time series model

SUMMARY & NEXT STEPS

SUMMARY

- A simple NN model leveraging time and location based features outperforms a time series model.
- CPD can use this predictive output in such decisions as staffing, training, equipment procurement, patrols and many more.
- Using external data sources such as median income, property prices and weather conditions, the NN model can be enhanced
- Time series model can also be enhanced by incorporating holidays

NEXT STEPS (A.K.A IF I HAD MORE TIME...)

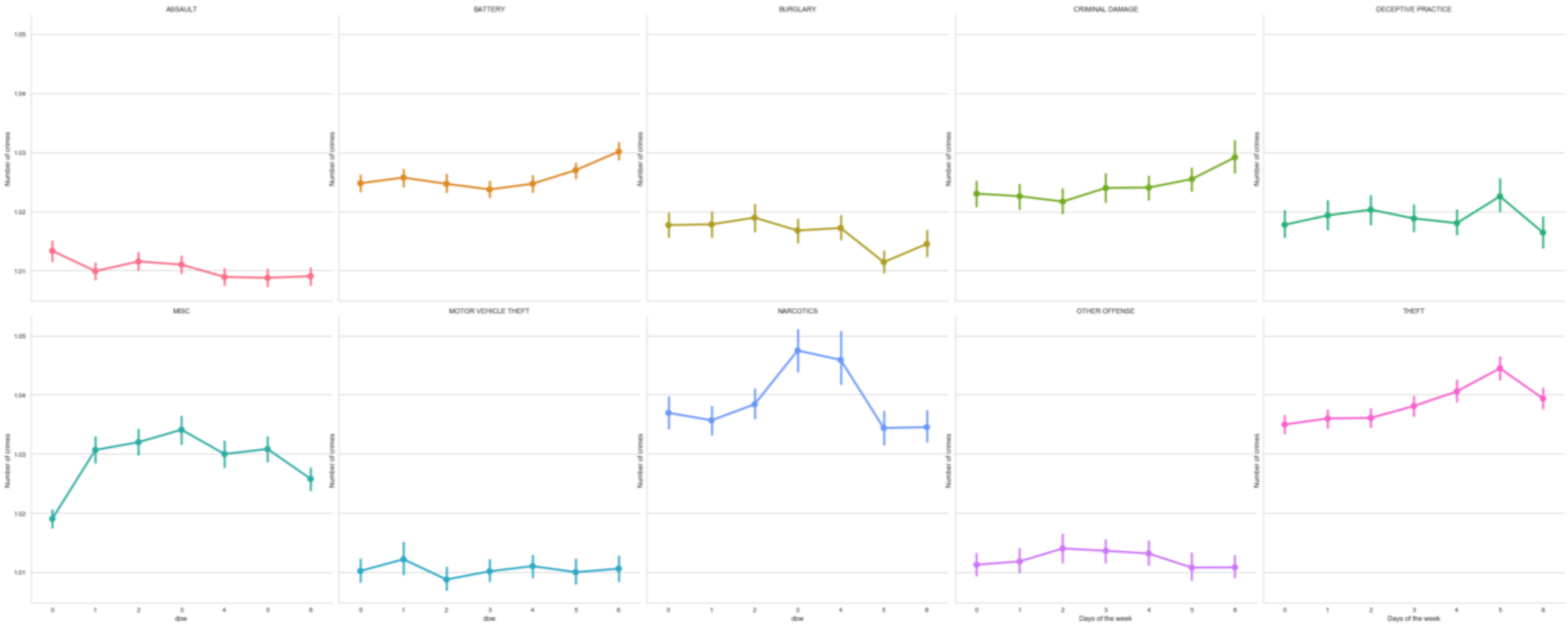
- **Feature Engineering**
  - Economic features by location such as median income, median house prices etc.
  - Weather features such as temperature, humidity, precipitation etc.
  - List of holidays passed on to the Prophet model

- **Model Selection**
  - Trying out various algorithms

- **Model Tuning**
  - with grid search for hyperparameter tuning
  - Cross validation

APPENDIX

PLOT BY DAY OF WEEK



PLOT BY HOUR OF THE DAY

PLOT BY WEEK OF THE YEAR