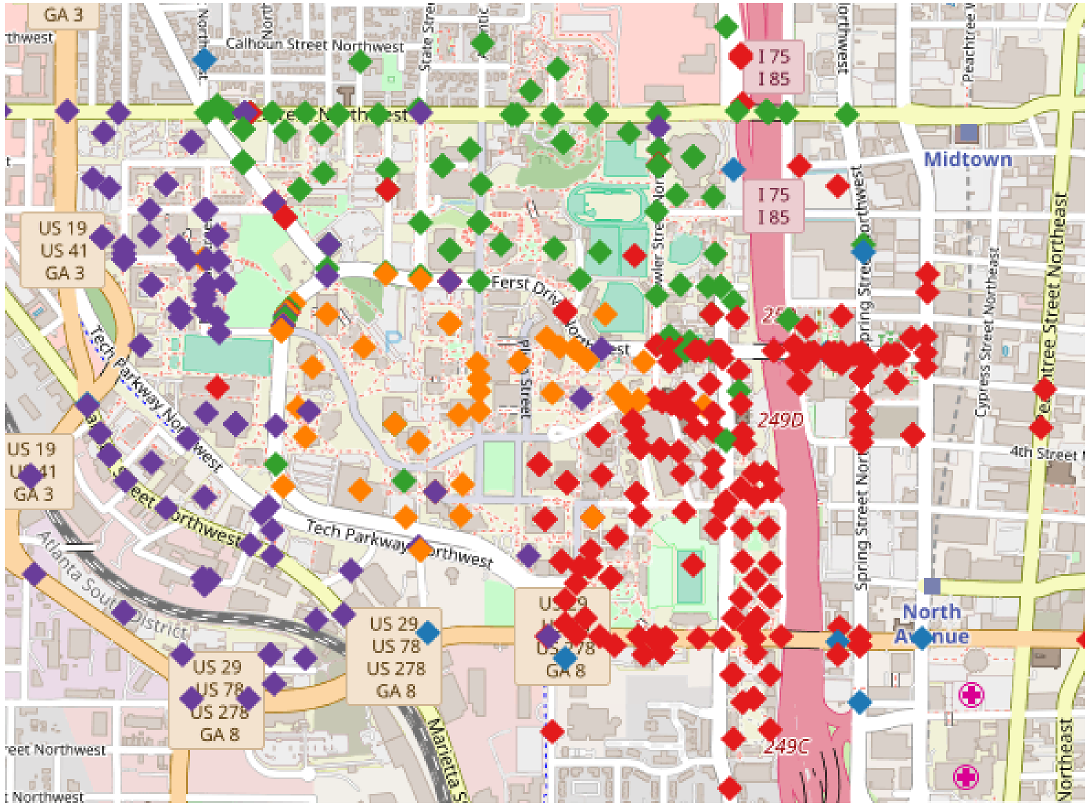
Forecasting Crime around the Georgia Tech Campus

Phillip Spratling

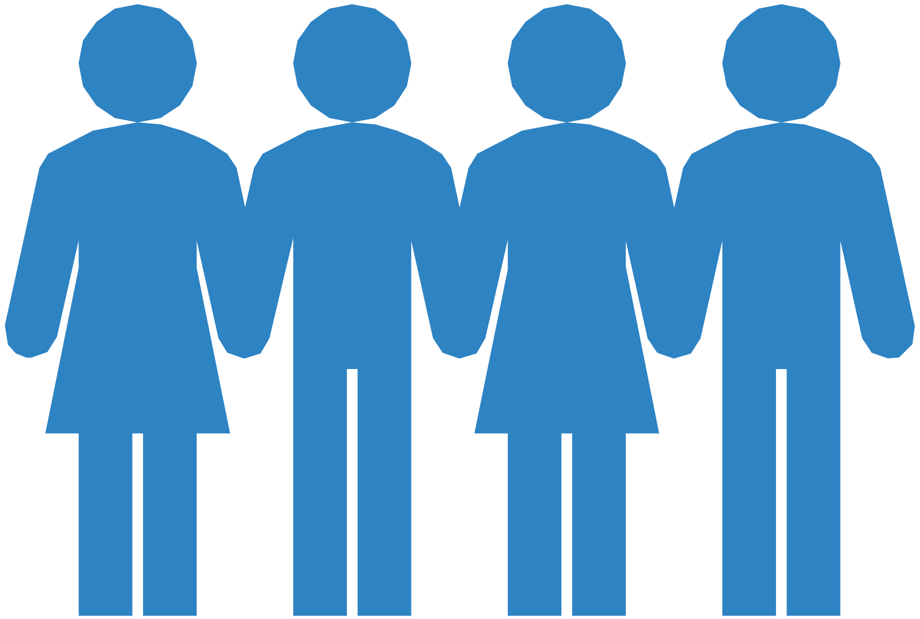
The Problem

- ~12,500 crimes on the Georgia Tech Campus since 2010
- ~3000 different locations where crime has occurred
- 6th most dangerous city in the USA (Forbes 2017)



Who does this affect?

- Students
- Georgia Tech Police Department
- Those living in residential areas surrounding campus



The Goal

Analyze	Analyze what factors affect crime levels so students know how to stay safe
Predict	Predict number of crimes per month to help GTPD keep crime on campus under control

The Data

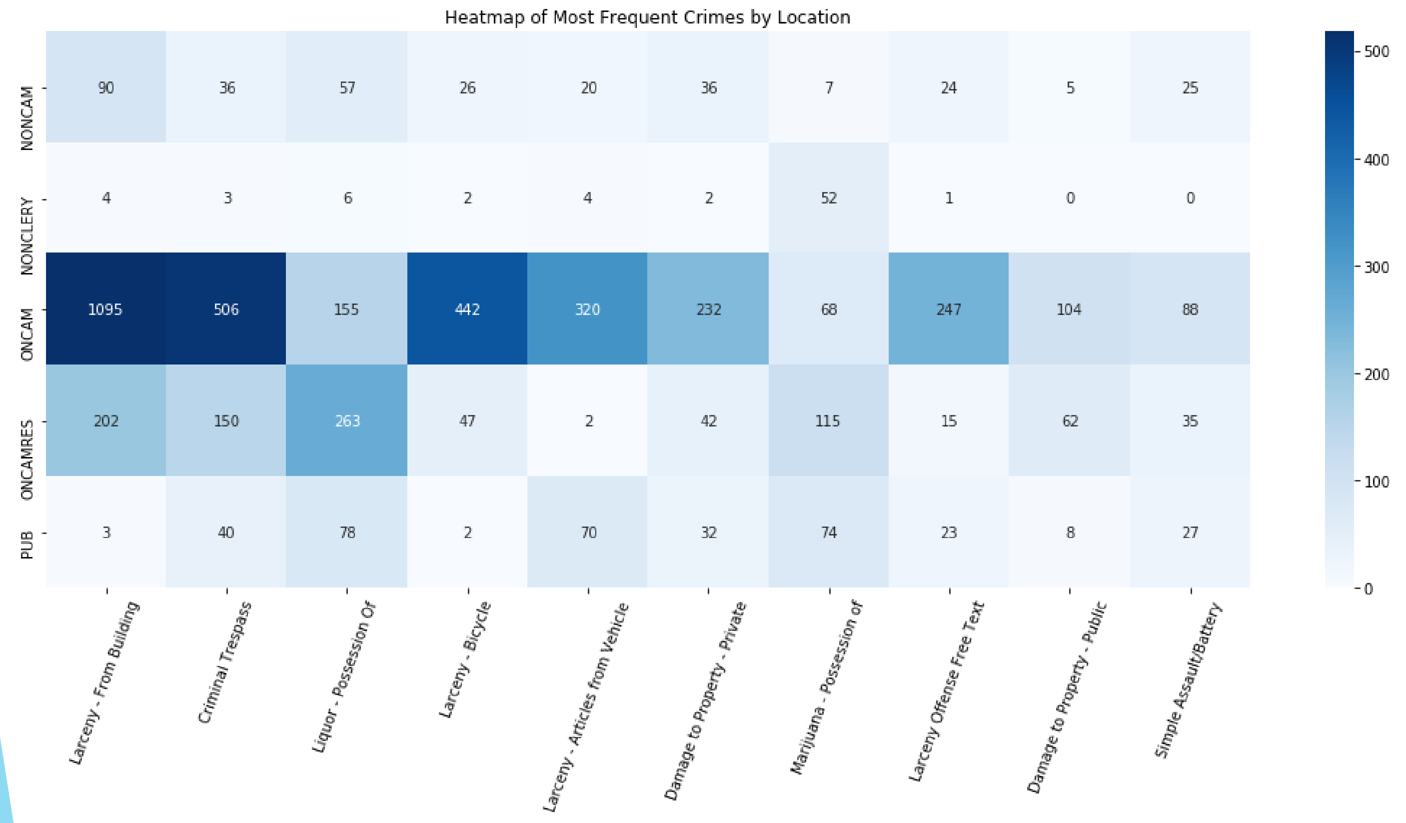
- Georgia Tech public crime logs from 2010-today
- Atlanta census demographic data
- Georgia Tech enrollment data

time	code	description	disposition	location	patrol_zone	landmark	lat	long	year	month	...	inc_15_25	inc_25_35	inc_35_50	inc_50
2018-06-29 14:31:00	2901	Damage to Property - Business	NaN	ONCAM	Z2	NAA Dining Hall	33.779500	-84.402337	2018	6	...	21826	16939	23089	31
2018-06-29 21:30:00	2317	Larceny - Bicycle	NaN	ONCAMRES	Z1	Graduate Living Center	33.781507	-84.397034	2018	6	...	21826	16939	23089	31
2018-06-30 00:58:00	5499	Traffic Offense (describe offense)	Cleared by Arrest	NONCLERY	OFFCAM	North Avenue NW @ Spring Street NW	33.770969	-84.389392	2018	6	...	21826	16939	23089	31
2018-06-30 22:50:00	5311	Disorderly Conduct	Cleared by Arrest	NONCAM	Z1	Phi Gamma Delta Fraternity	33.777970	-84.393620	2018	6	...	21826	16939	23089	31
2018-06-30 23:55:00	5499	Traffic Offense (describe offense)	Cleared by Arrest	PUB	Z2	North Avenue NW @ Fowler Street NW	33.770974	-84.393940	2018	6	...	21826	16939	23089	31

Exploratory Analysis

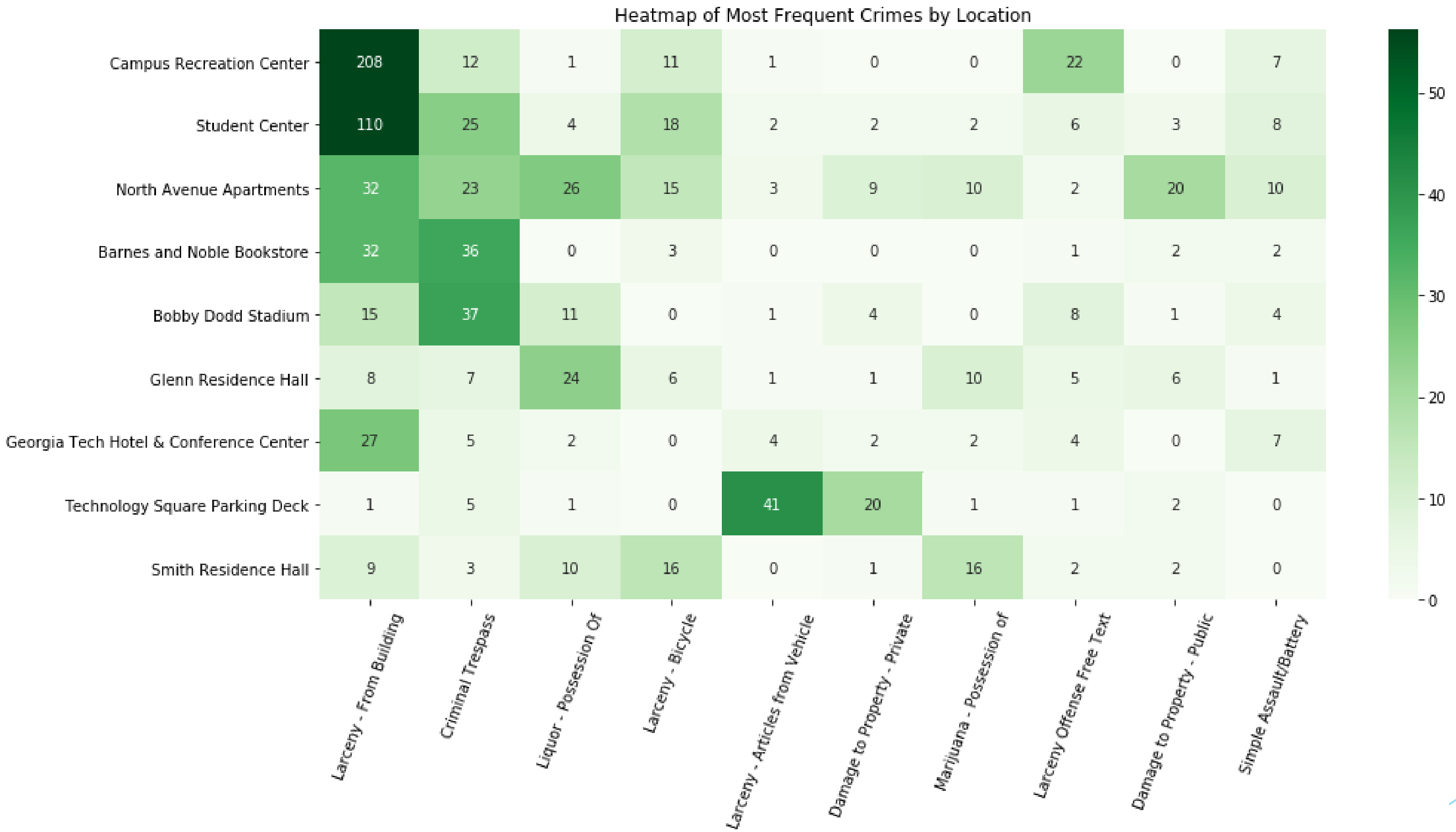
What factors can we use to predict crime?

Location-Based Crime Trends



- The majority of crimes reported happen on campus
- Possession of drugs and alcohol most likely in residential areas
- Larceny most prevalent type of crime across campus

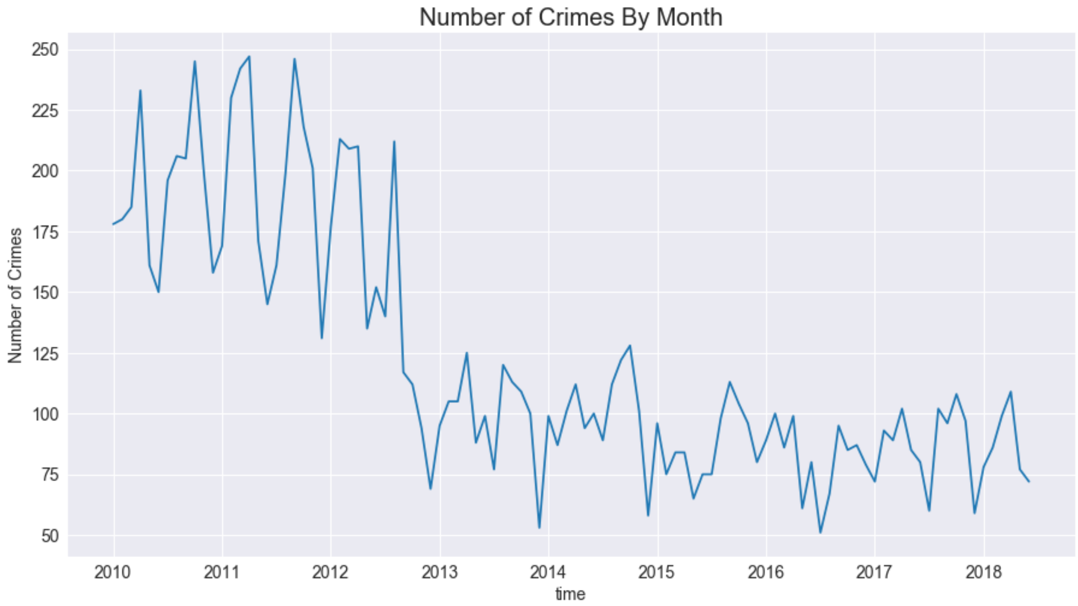
Location-Based Crime Trends



- Larceny common in CRC locker rooms
- Crime levels in CULC very low – not even in top 10 most frequent locations
- Tech Square parking deck most frequent location of theft from vehicle

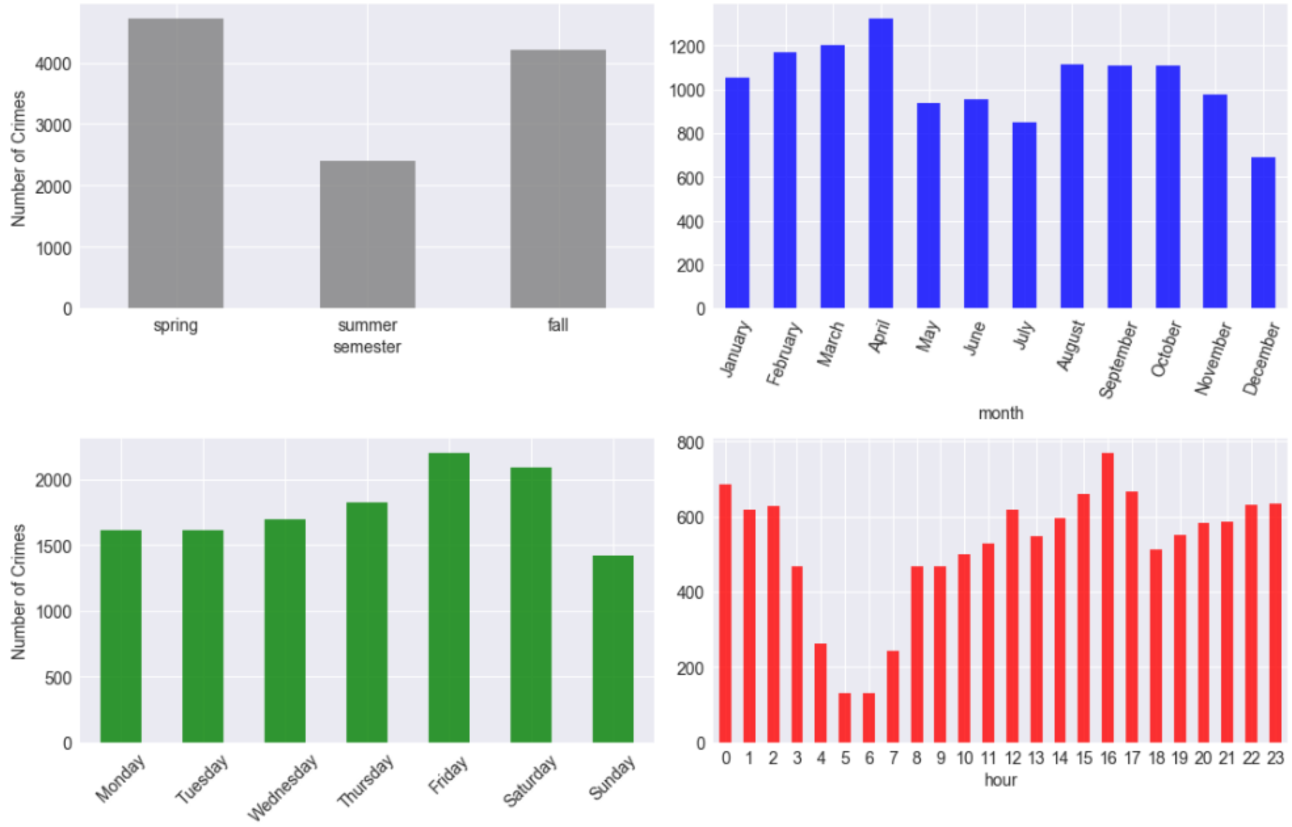
Temporal-Based Crime Trends

- Crimes have decreased overall since 2010
- Periodic – spiking twice per year
- Large sudden decrease starting fall semester 2012



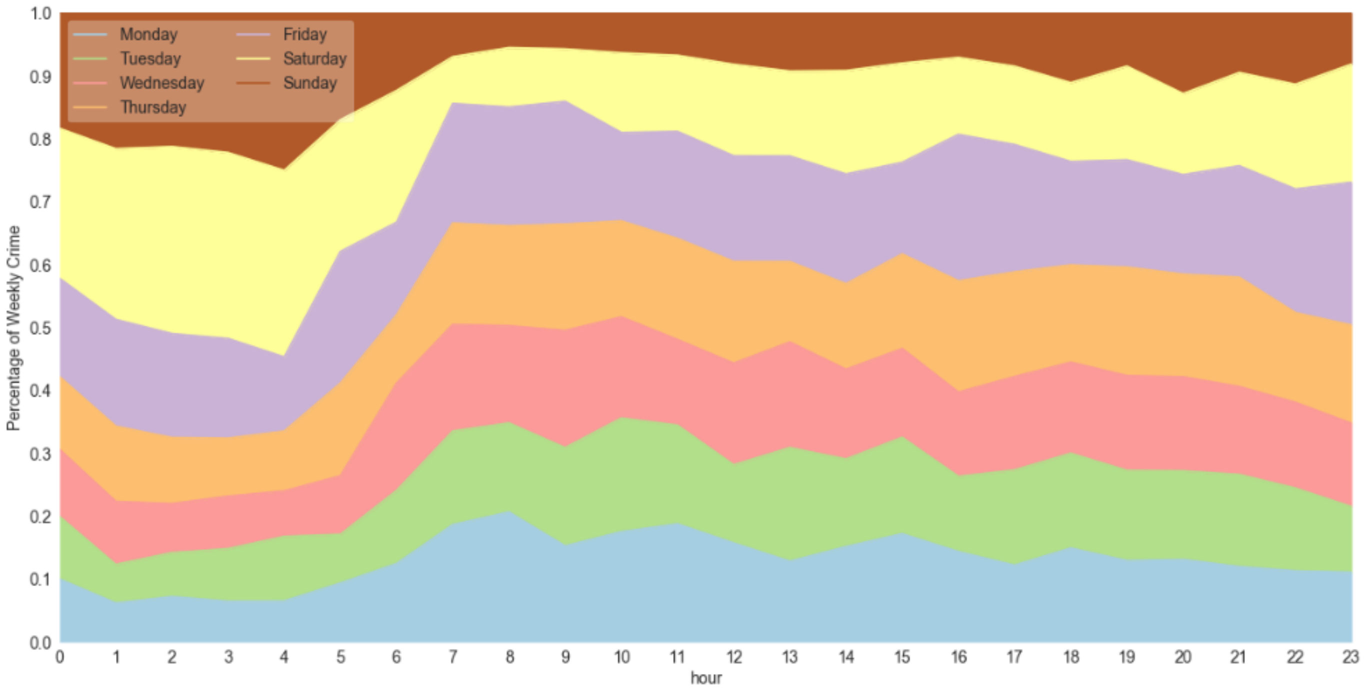
Temporal-Based Crime Trends

- Less crimes in summer months – due to lower enrollment in summer semester?
- Peak in crimes on Friday and Saturday. Sunday has lowest number
- Crimes highest in afternoon and late night, lowest around 5-6 AM

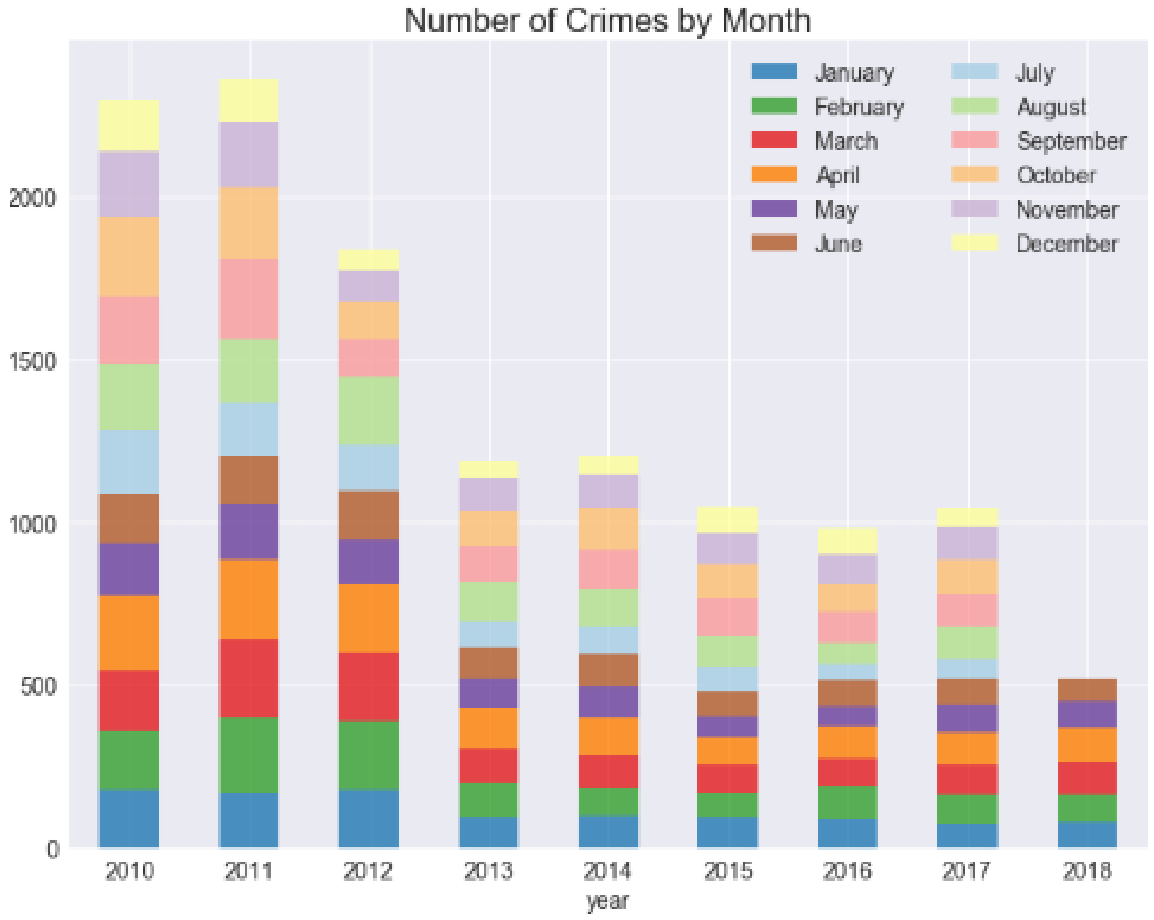


Temporal-Based Crime Trends

- Crimes on weekends happen later at night than on weekdays
- Majority of Sunday's crimes happen 1-4 AM (i.e. Saturday night after midnight)

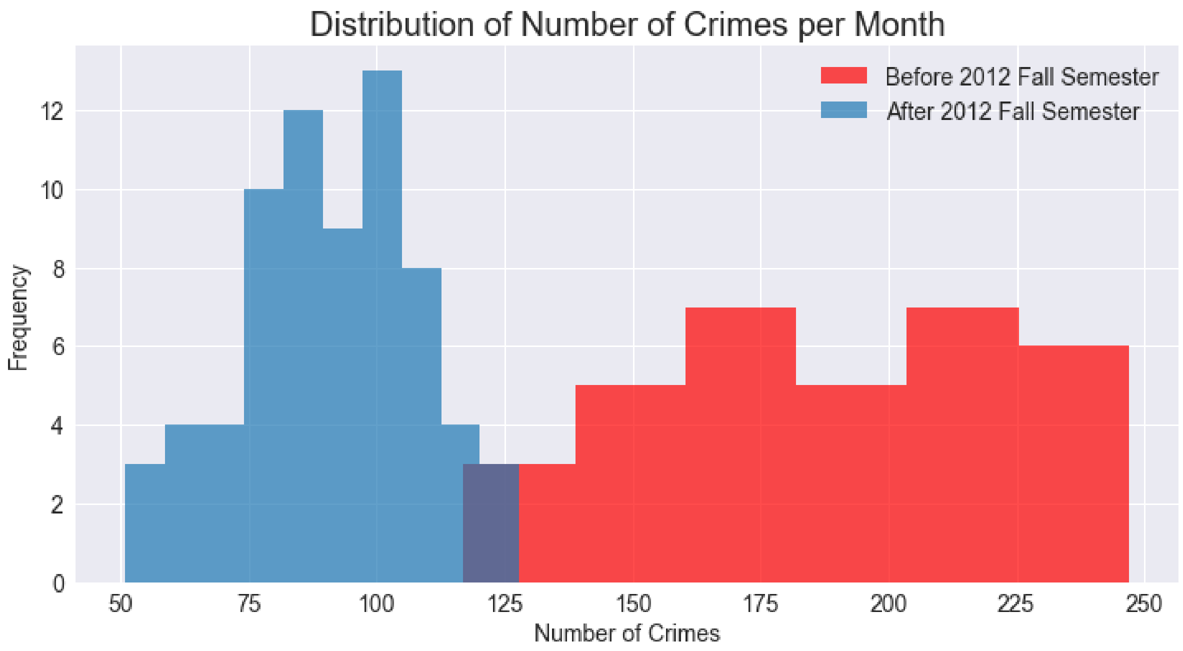


Crimes by Month – Further Analysis



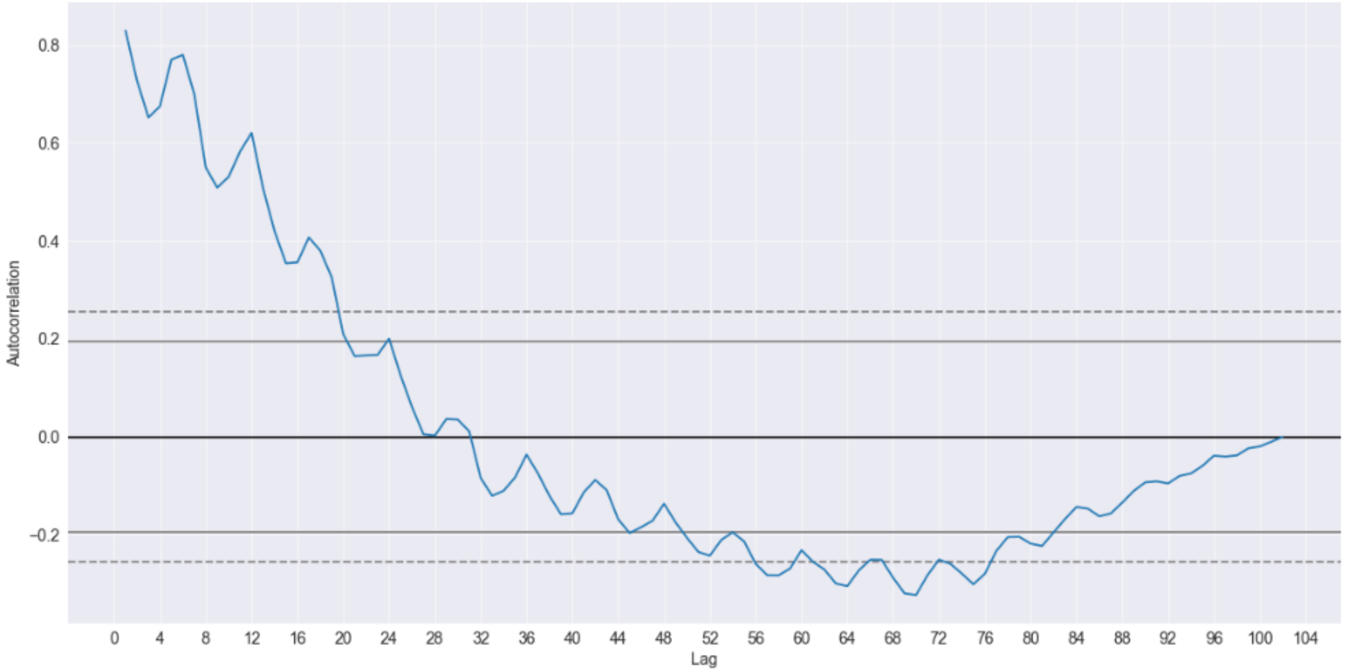
- Distribution of crimes per month stays relatively constant throughout the years
- October and April have the most crimes, June has the least

Crimes by Month – Further Analysis



- Bimodal distribution split at 2012 fall semester
- Crimes before centered around 175-200 per month
- Crimes now centered around 75-100 per month

Crimes by Month – Further Analysis



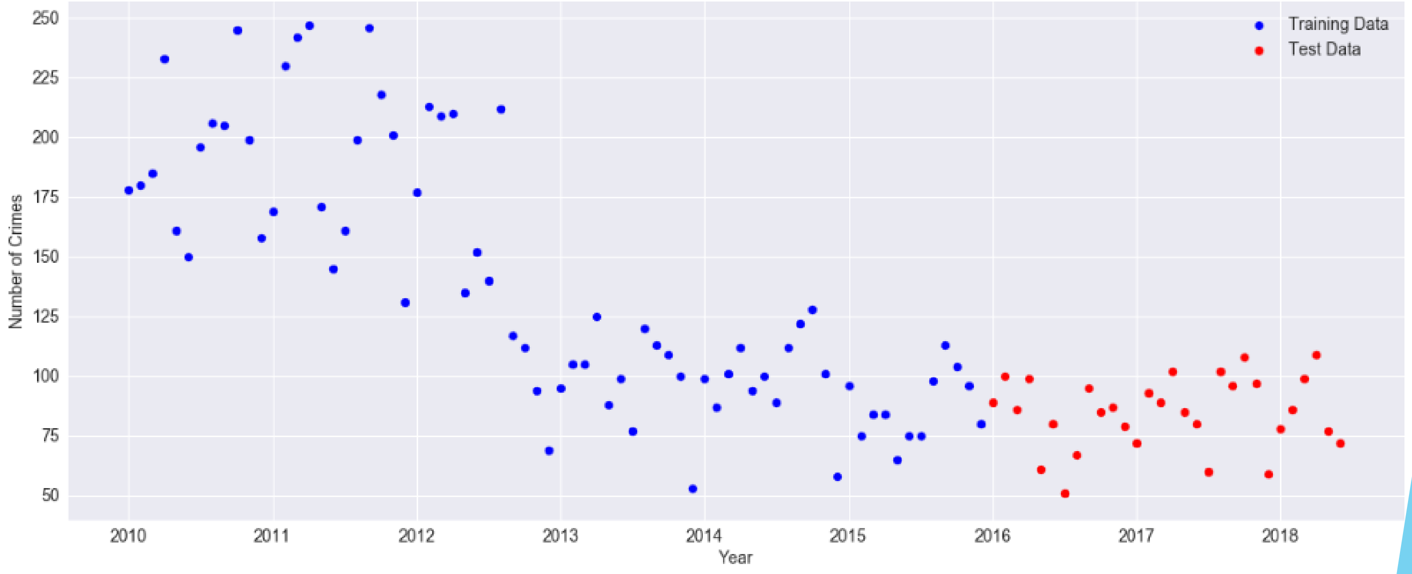
- Significant autocorrelation of crimes per month with lags less than 20
- Spikes in autocorrelation every 6 months

Regression

Predicting Number of Crimes per Month

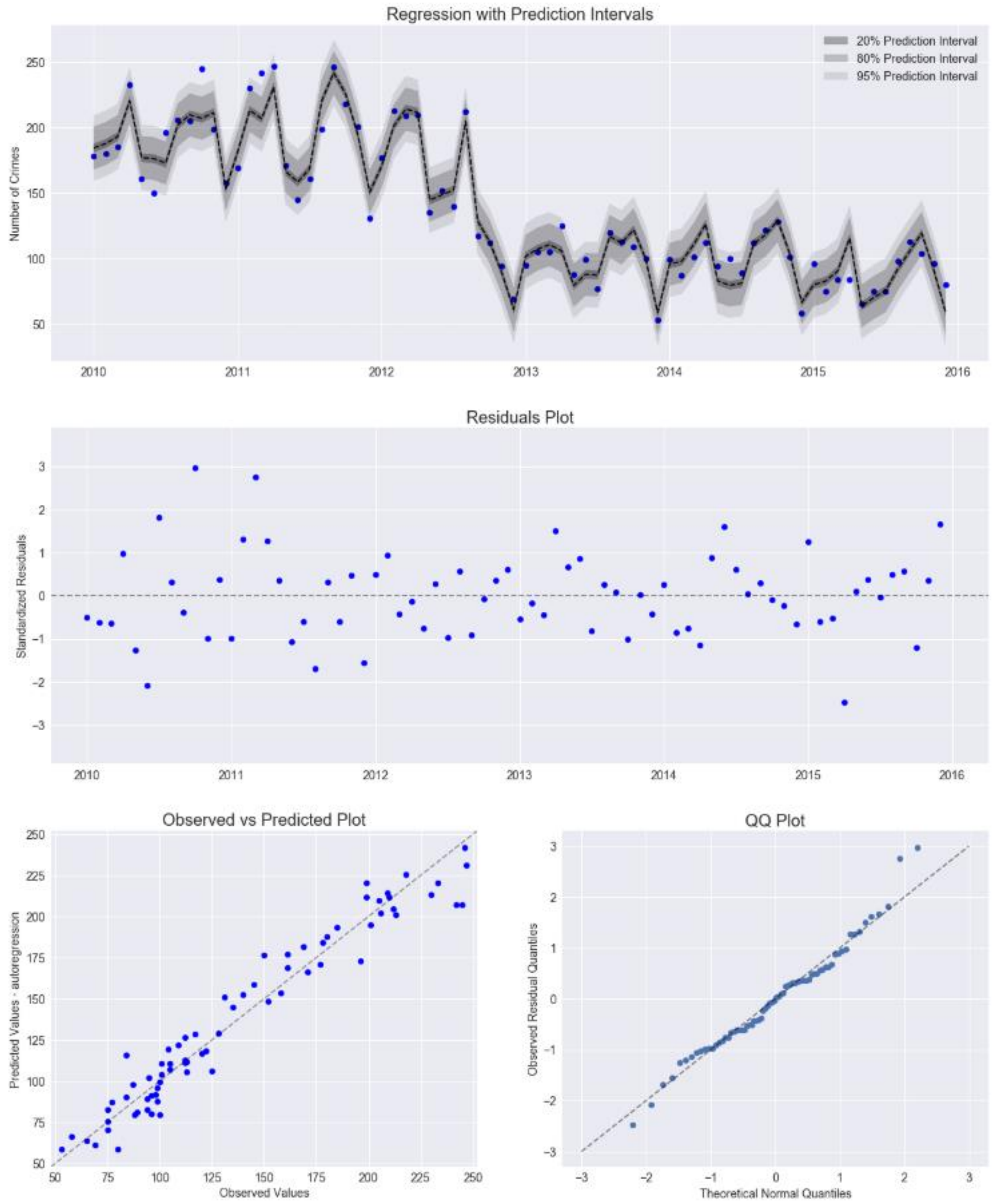
Train-Test Split

- 2010-2015 used as training data
- 2016-18 used as testing data



Testing Model Assumptions

- Each model fit was tested to ensure model assumptions were held
- Residuals plot should show constant variance
- QQ Plot should show normality of residuals
- Observed vs Predicted shows how the model performs with high vs low numbers of crime



Evaluation and Feature Selection

- Models evaluated by MSE, R^2, adjusted R^2, and trends accuracy (if crime will increase or decrease next month)
- With algorithms sensitive to overfitting with too many features, weakest predictors were removed
- Used recursive feature evaluation to maximize adjusted R^2

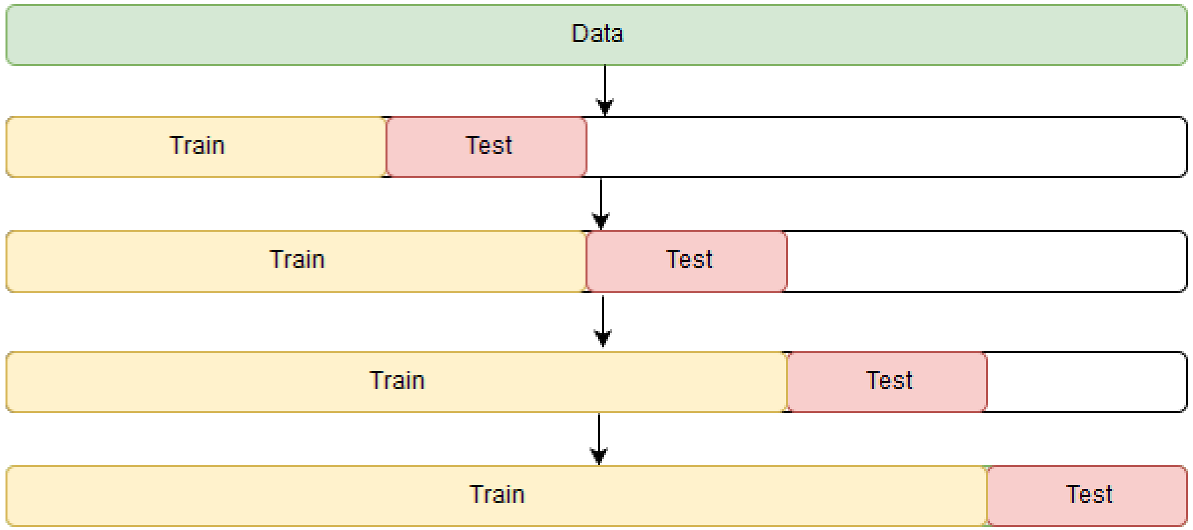
```
MSE: 167.6584
RMSE: 12.9483
R-squared: 0.9426
Adjusted R-squared: 0.9052
Trends accuracy: 0.7465 or 53/71
```

```
Strongest predictors:   abs(coefficent)
is_before_2012_fall     42.962057
enrollment              19.581002
past_num_crimes_8       16.051123
is_month_7              13.404389
is_month_6              12.769827
dtype: float64
```

```
Weakest predictors:   abs(coefficent)
pop_45_54      0.064846
pop_female     0.055369
pop_65_74      0.028566
pop_total      0.019091
pop_25_34      0.002308
dtype: float64
```

Time Series Cross-Validation

- Time series grid search cross-validation used on training set to find best model hyperparameters
- Model hyperparameters with lowest MSE on validation set chosen to use on test set

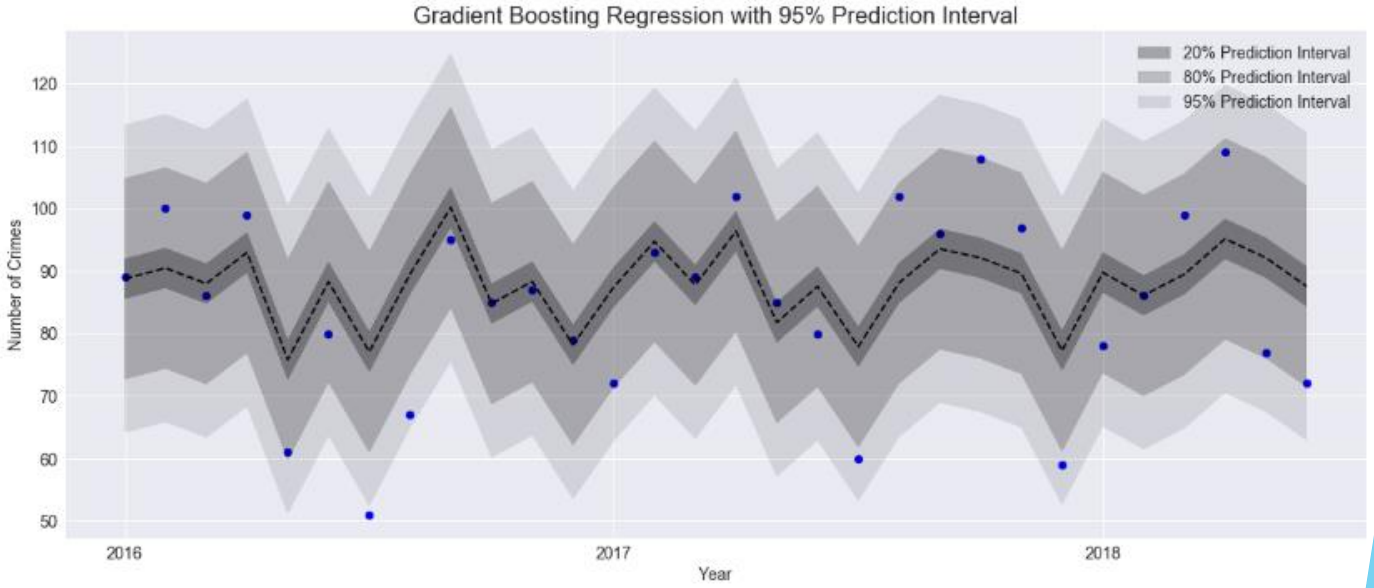



```
Best MSE: 158.8485812418725
Best RMSE: 12.603514638459881
Best Parameters: {'regression__C': 0.25, 'regression__epsilon': 4.0}
```

Evaluation

- Gradient boosting performed the best out of all models on test set
- Predicted crime trend correctly 24/29 times
- 9/30 in 20% prediction interval
- 29/30 in 95% prediction interval

```
MSE: 134.2860
RMSE: 11.5882
R-squared: 0.3951
Trends accuracy: 0.8276 or 24/29
```

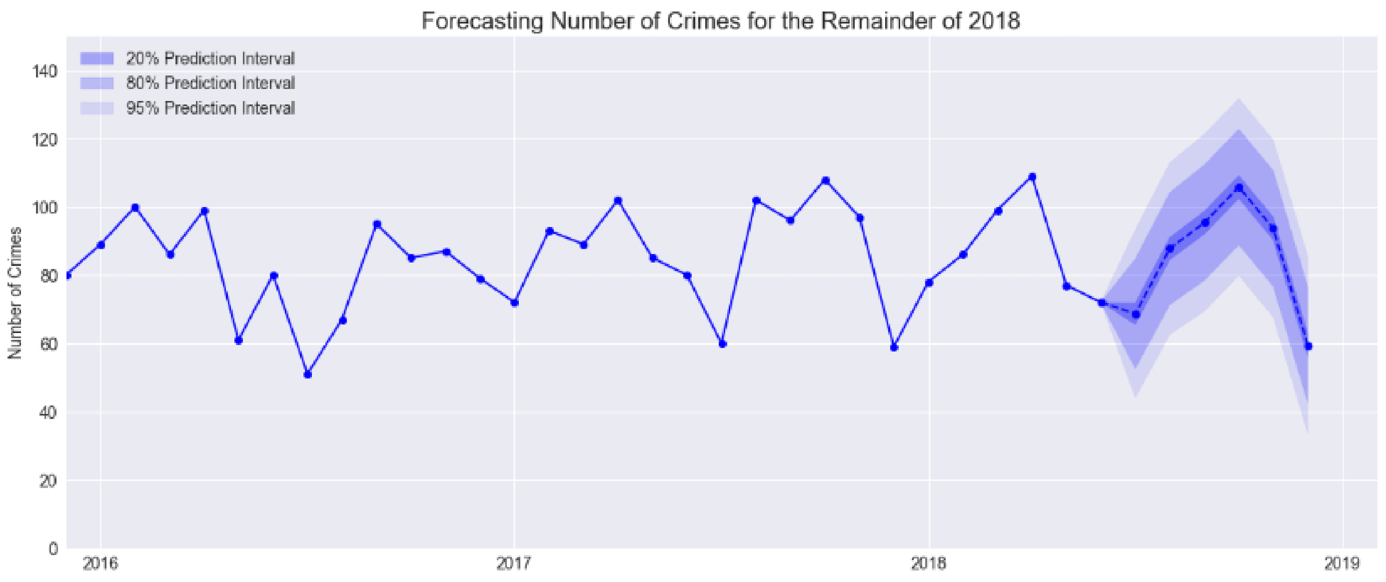


Conclusion

Forecasting and Next Steps

Forecasting the Remainder of 2018

- Gradient boosting model used to forecast for the remainder of 2018
    - July, 2018: 68
    - August, 2018: 87
    - September, 2018: 95
    - October, 2018: 105
    - November, 2018: 93
    - December, 2018: 59



Next Steps and Extensions

01

Build a model to predict the next crime to occur

02

Extend analysis to other locations

03

Use analysis to create new optimal patrol zones