Capstone Project 1: In-Depth Analysis

Machine Learning:

Predicting future 911 calls for violent and property crimes poses a particularly interesting data challenge because it has both geospatial and temporal dimensions and may be affected by many different types of features like weather, city infrastructure, population demographics, public events, government policy, etc.

Most of the models use extensive features than that were available here.Those include population demographics, employment rates, housing prices, race etc.

I used below mentioned features to predict the number of 911 calls for violent and property crimes from June 2018- December 2018.

Feature Engineering:

From EDA, a significant relevance between time of the call and number of calls was observed. This helped tremendously in manual feature selection process.

Year : Year of the call received will be extracted from call_dttm column and was converted to int64 data type
Month: Month of the call received will be extracted from call_dttm column and was converted to int64 data type
Date: Date of the call received will be extracted from call_dttm column and was converted to int64 data type
Hour: Hour of the call received will be extracted from call_dttm column and was converted to int64 data type
Day: Day of the week of the call received will be extracted from call_dttm column and was Encoded to 1-7 for the days from Monday-Sunday
Event: Number of the events per day

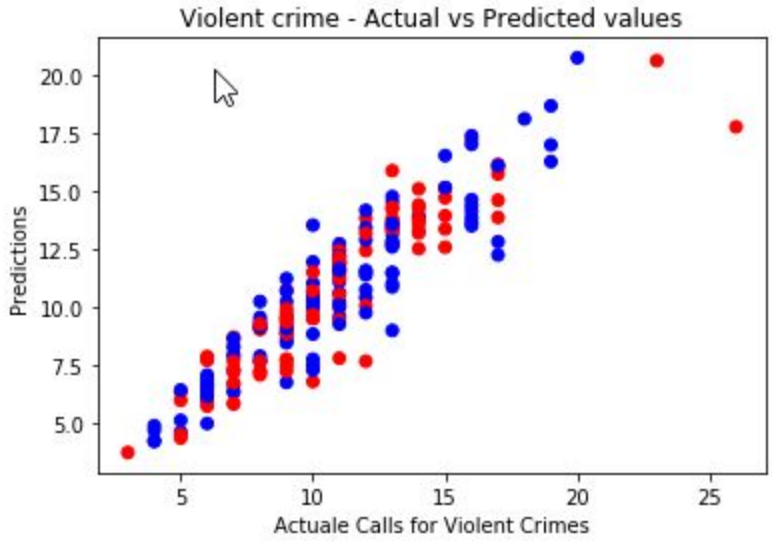
Number of events per day were calculated by grouping the year,month and date of the call and two dataframes (one for violent crimes and property crimes each) were being created with above features.

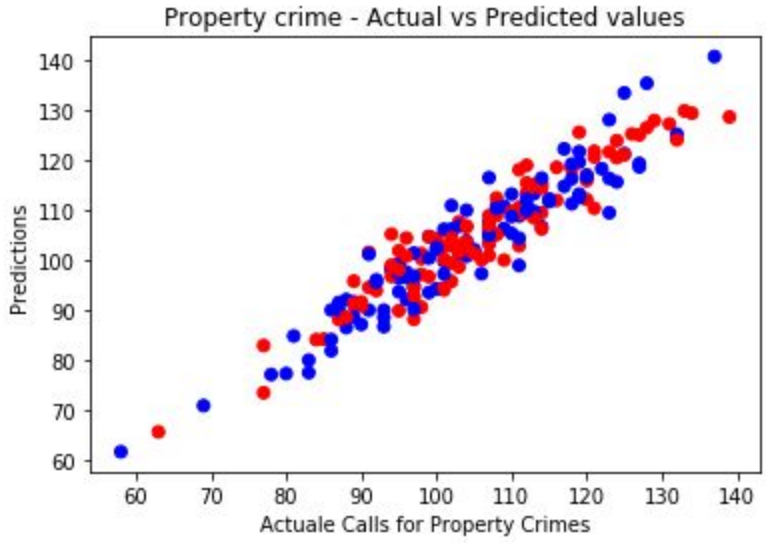
Predictions:

As the goal was to predict the number of calls for the period June 2018 - December 2018, I splitted the data accordingly. The data from March 2016 - May 2018 was planned to be used for training and data from June 2018- December 2018 was our testing dataset.

Predicting calls for violent crimes and property crimes:

To predict calls for violent crimes and property crimes, the base model was linear regression model. After fitting the model on training set , I tried to predict the events from testing dataset.The accuracy score of linear regression model was 81% for violent crimes and 89% for property crimes.





To know how good the models was, I used different evaluation metrics such as **Mean Squared Error** and **R-Square**. (Mean Squared Error is nothing but the difference between the observed value and predicted value, and R-square determines how much of the total variation dependent variable is explained by the variation in independent variable. In our case, event is the dependent variable and Year, Month, Date, Hour and Day are independent variables)

Later I implemented more models on our above problem and checked whether these models perform better than our linear regression model and below are the results.

Model comparisons:

	Linear Regression		Random Forest Regressor		Gradient Boosting Regressor	
	MSE	**r2_score**	**MSE**	**r2_score**	**MSE**	**r2_score**
Calls for Violent crimes	**2.60**	**0.811**	**2.54**	**0.815**	**2.16**	**0.843**
Calls for Property crimes	**20.46**	**0.890**	**23.38**	**0.874**	**21.71**	**0.883**

As we can see for violent crime call prediction, both the MSE and the value of R-square are less with Gradient Boosting Regressor model. Whereas for property crime prediction Linear regression model has small MSE and R-square value.