# Prediction of Road Accidents in the City of Austin: Time series Analysis and Regressive Machine Learning Models

Road accidents are one of the leading causes of death of young people in the world. Millions of people over the world lose their lives every year from auto accidents. Machine learning (ML) can help to avoid road accidents. We propose a simple machine learning model which can predict the number of timely accidents at various locations in Austin, Texas.

Unpredictable weather, hazardous road conditions and many other factors lead to the accidents. The number of vehicles in Austin city is increasing daily due to the rapid growth of residential population. Geographical ups and downs cause another problem on driving safety in Austin during the flood after heavy rain. Therefore, a detailed analysis of the dependencies of the factors on road accidents can help travelers and road safety authority to reduce the potential road accidents.

To understand the factors involved in causing road accidents, we analyze the problem by using multiple correspondence analysis. By using various statistical models, we can predict the potential road accidents. The accurate prediction will help travelers to avoid the accident risk zones by adopting an alternative way. This will also help the roadside safety management team to provide much attention to the predicted zone.

In this project, we apply a time *series regression* model to forecast the traffic accident in a particular area in the city. Using spatiotemporal analysis, we identify the traffic accident hotspots. Using machine learning models (Random Forest (RF) and eXtreme Gradient Boosting (XGB)), we predict the risk of traffic accidents in particular location and time. Python3 codes for data collection/cleaning, exploratory data analysis, statistical inferences and regression models can be accessed in the jupyter notebook Github.

## Data collection and cleaning:

Accident data reported every five second interval was collected as 'Real_time_traffic_incidence.csv' from https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x .

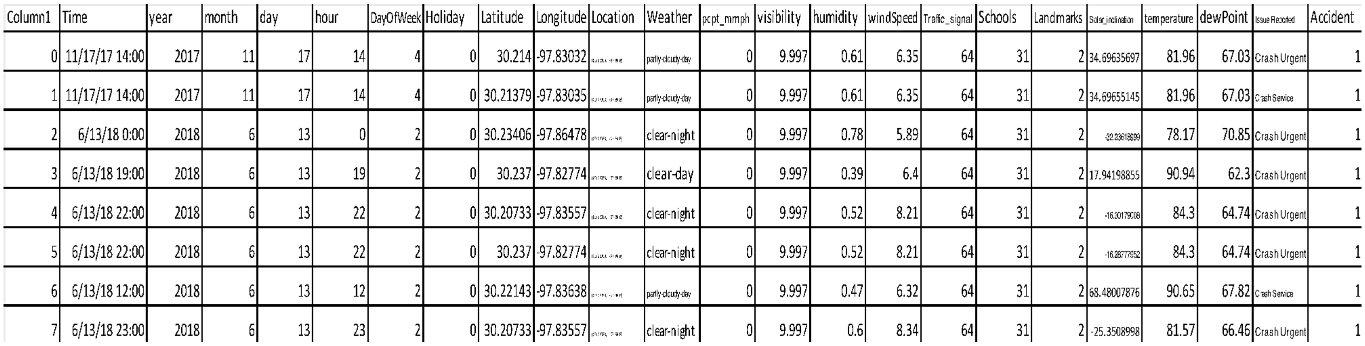
From the source https://data.austin.gov, we collected information about other spatial features, such as nearby traffic signals, road intersections, historical landmarks, school zones.

The dataset that contains a list of dates and holidays was downloaded from Kaggle.com and merged to our master datafile.

Illumination caused by sunlight may cause the imbalanced driving to the driver and result in an accident. Therefore, we included the solar inclination angle at the time of events. Using python module 'pysolar', we derived the solar inclination angle, at the particular location and time.

The weather-related data for specific locations and dates that we concerned were downloaded from https://api.darksky.net.

We used seaborn heatmap and histogram to detect the unphysical and missed data. Unphysical data samples were removed, and missing data were filled by using its past immediate value at that location. After cleaning all files and merging to single file, the final data file was with 24 columns and 68420 rows.



| Column1 | Time | year | month | day | hour | DayOfWeek | Holiday | Latitude | Longitude | Location | Weather | pcpt_mmph | visibility | humidity | windSpeed | Traffic_signal | Schools | Landmarks | Solar_inclination | temperature | dewPoint | Issue Reoted | Accident |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 11/17/17 14:00 | 2017 | 11 | 17 | 14 | 4 | 0 | 30.214 | -97.83032 | [illegible] | partly-cloudy-day | 0 | 9.997 | 0.61 | 6.35 | 64 | 31 | 2 | 34.69635697 | 81.96 | 67.03 | Crash Urgent | 1 |
| 1 | 11/17/17 14:00 | 2017 | 11 | 17 | 14 | 4 | 0 | 30.21379 | -97.83035 | [illegible] | partly-cloudy-day | 0 | 9.997 | 0.61 | 6.35 | 64 | 31 | 2 | 34.69655145 | 81.96 | 67.03 | Crash Service | 1 |
| 2 | 6/13/18 0:00 | 2018 | 6 | 13 | 0 | 2 | 0 | 30.23406 | -97.86478 | [illegible] | clear-night | 0 | 9.997 | 0.78 | 5.89 | 64 | 31 | 2 | -32.23618999 | 78.17 | 70.85 | Crash Urgent | 1 |
| 3 | 6/13/18 19:00 | 2018 | 6 | 13 | 19 | 2 | 0 | 30.237 | -97.82774 | [illegible] | clear-day | 0 | 9.997 | 0.39 | 6.4 | 64 | 31 | 2 | 17.94198855 | 90.94 | 62.3 | Crash Urgent | 1 |
| 4 | 6/13/18 22:00 | 2018 | 6 | 13 | 22 | 2 | 0 | 30.20733 | -97.83557 | [illegible] | clear-night | 0 | 9.997 | 0.52 | 8.21 | 64 | 31 | 2 | -16.30179008 | 84.3 | 64.74 | Crash Urgent | 1 |
| 5 | 6/13/18 22:00 | 2018 | 6 | 13 | 22 | 2 | 0 | 30.237 | -97.82774 | [illegible] | clear-night | 0 | 9.997 | 0.52 | 8.21 | 64 | 31 | 2 | -16.28777952 | 84.3 | 64.74 | Crash Urgent | 1 |
| 6 | 6/13/18 12:00 | 2018 | 6 | 13 | 12 | 2 | 0 | 30.22143 | -97.83638 | [illegible] | partly-cloudy-day | 0 | 9.997 | 0.47 | 6.32 | 64 | 31 | 2 | 68.48007876 | 90.65 | 67.82 | Crash Service | 1 |
| 7 | 6/13/18 23:00 | 2018 | 6 | 13 | 23 | 2 | 0 | 30.20733 | -97.83557 | [illegible] | clear-night | 0 | 9.997 | 0.6 | 8.34 | 64 | 31 | 2 | -25.3508998 | 81.57 | 66.46 | Crash Urgent | 1 |

## Exploratory Data Analysis and Feature Engineering:

**Time Series** (daily accident counts, temperature and rain):

We explored the seasonal trend on the accident by plotting the number of accidents as a function of day. Observing the time series graph for the available dataset for two years, we found repeated peaks at a certain season of the year, around the first week of December as shown in green line of figure 1. To understand those systematic peaks, we plotted the time series of December each year and found peaks obtained on (6th of Dec 2017) and (7th of Dec 2018). We explored the effect of correlation between weather conditions and time to the peaks on that particular day of both of the years. We plotted the qualitative graph of precipitation intensity, temperature and number of accidents in the same graph and found that dips on the temperature and peaks on precipitation matched to the peaks on the number of accidents graph. In Austin, on rainy winter days when the temperature goes down to freezing points, the roads are covered by ice and a lot of accidents happen at that time.

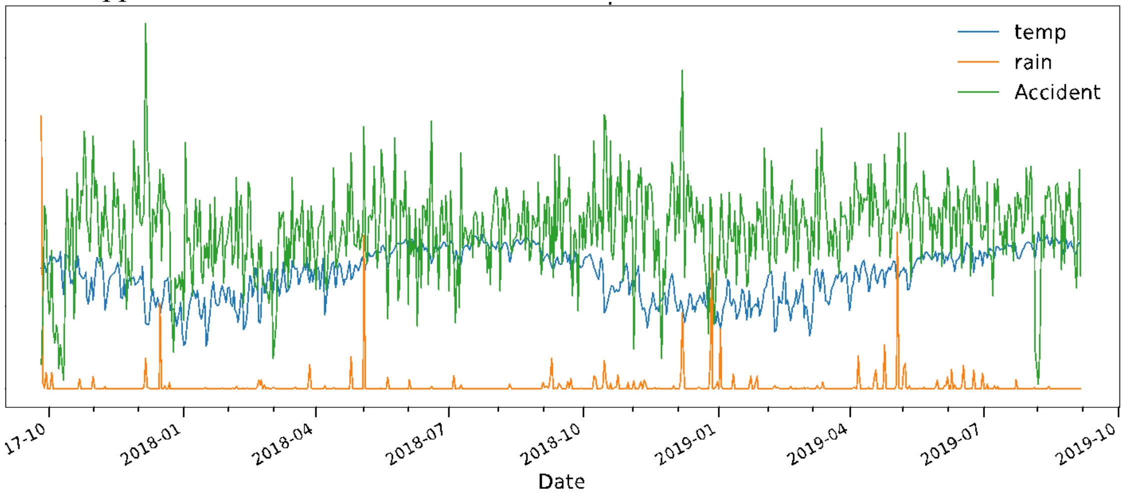


Figure 1 Time Series graphs for Number of accidents, rain in 1000mm/hour, and temperature in degree F.

**Accidents on Days and Months:**

The total counts of accidents in the city of Austin were found to be independent of days and months, days of the week and holidays. In Fig. 2 (iv), the small number of accidents during September is because of sampling issues. Our time series data starts on 26th Sep of 2017 and ends on 9th Sep of 2019. We found a smaller number of accidents on Thanksgiving and Christmas days, which was not surprising because on those days people celebrate those festivals inside their home.

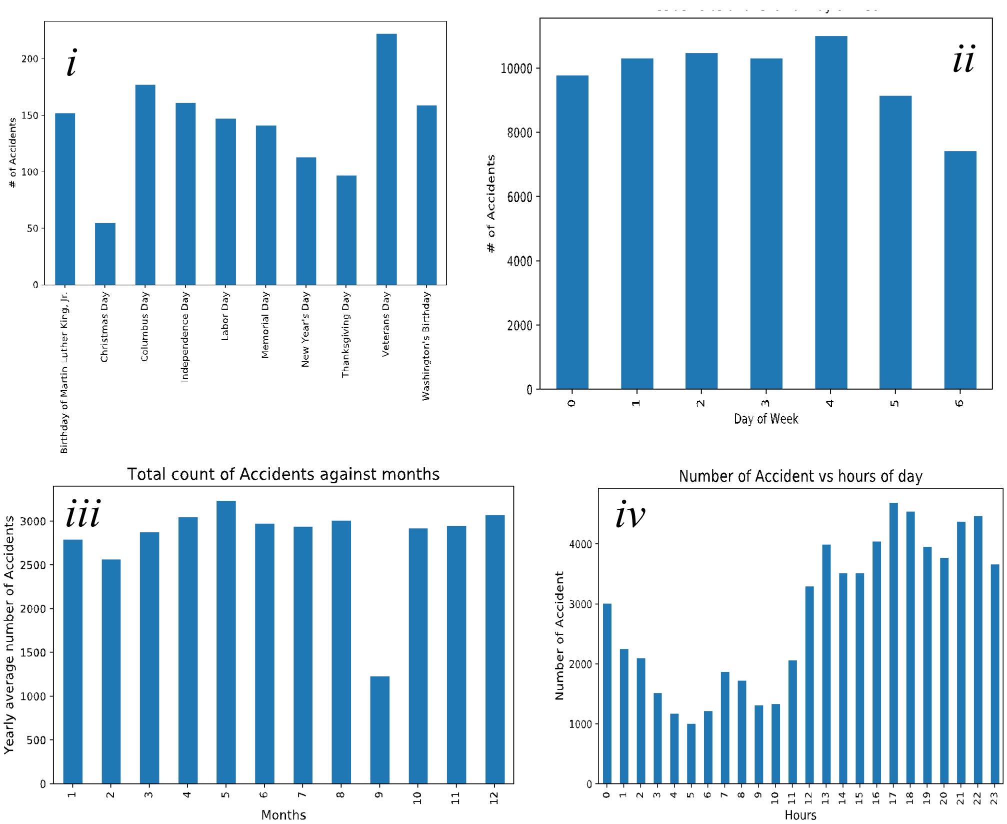


Figure 2: Number of Accidents as a function of (i) US holidays (upper left), (ii) Day of Week (upper right), (iii) Months (lower left) and (iv) Hours of day (lower right)

Solar inclination was calculated using hour of the day and location of accidents. From the analysis of hour of day dependence, we observed that most of the accidents occurred in the afternoon to midnight with maximum in the evening. Based on the available data for two years, the solar illumination was not the major factor for occurring the accidents. Solar illumination at the driver's eyes should be maximum at sunrise or sunset (when the sun is near horizon i.e. around solar inclination of -80$^{o}$ or 80$^{o}$). But, the number of accidents at that solar inclination were found to be less than at other angles.

**Road accidents at different locations:**

We analyzed the accidents data on the basis of geographical location i.e. on the basis of latitude and longitude. Firstly, we plotted the accidents on the geographical map of Austin in Fig. 3. The large number of accidents occurred near the downtown area and alongside the interstate highway. We defined different locations equally spaced obtained as cross points of 20 grids on longitudes and 20 grids on latitudes. We assigned the accidents to the particular closest cross point in the mesh by calculating the distance from the actual location of accidents to the nearby four cross points of grids. Bar plot (Figure 4) shows that the majority of the accidents occurred nearby locations and most of the locations were with negligible number of accidents.

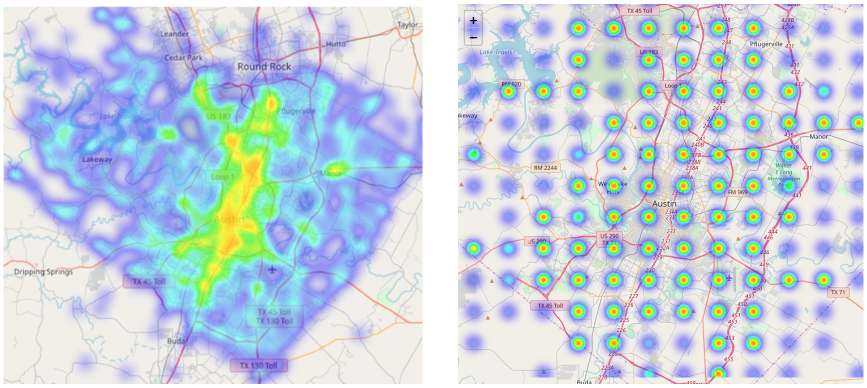


Figure 3: The heatmap of accidents as a function of geographical map of Austin (left). The right heatmap shows the number of accidents nearby specific locations (cross points created by drawing vertical and horizontal grids in between the whole range of latitudes and longitudes). Areal resolution of every location ~ 6 square miles.

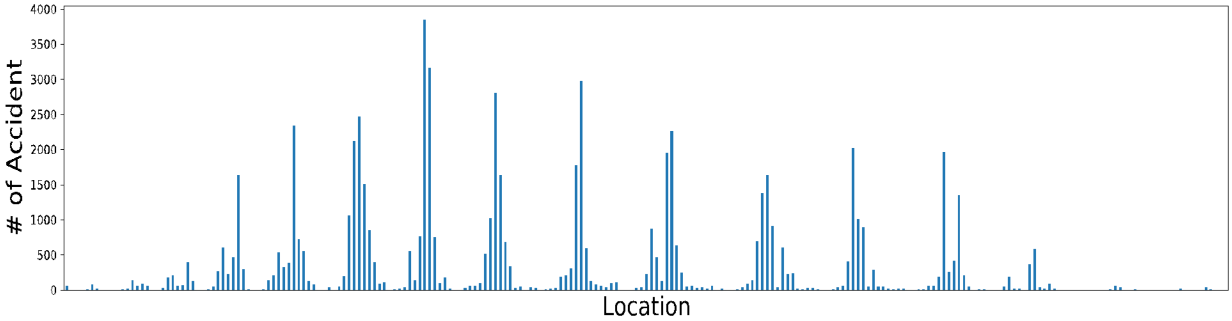


Figure 4: Histograms as a function of specific locations (cross points of the latitude-longitude grids)

## Feature correlations:

We tested the strength of correlation by calculating the correlation coefficients between the possible pair of variables. We calculated the Pearson's correlation coefficient as

$$\rho_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

where, $cov(X,Y) = \frac{1}{n}\sum_i \quad x_i y_i - \underline{x}\,\underline{y}$ and $\sigma_x = \frac{1}{n}\sqrt{x_i - \underline{x}}$ , $\sigma_y = \frac{1}{n}\sqrt{y_i - \underline{y}}$ are the standard deviations of features X and Y, respectively.

We plotted the correlation matrix in a heatmap as shown in Fig. 5, where the lower triangular matrix shows the color codes as indicated in the color bar and the upper triangular part shows the actual values of the Pearson's correlation coefficient between the corresponding pair of variables. The levels of accidents were assigned as: 1 (issue reported with word 'crash'), 2 (issue reported with word 'collision') and 3 (issue reported with word 'injury or fatal'). The issue reported was found to be independent of all features except the latitude, longitude, humidity, number of traffic signals, number of schools and number of historical landmarks.

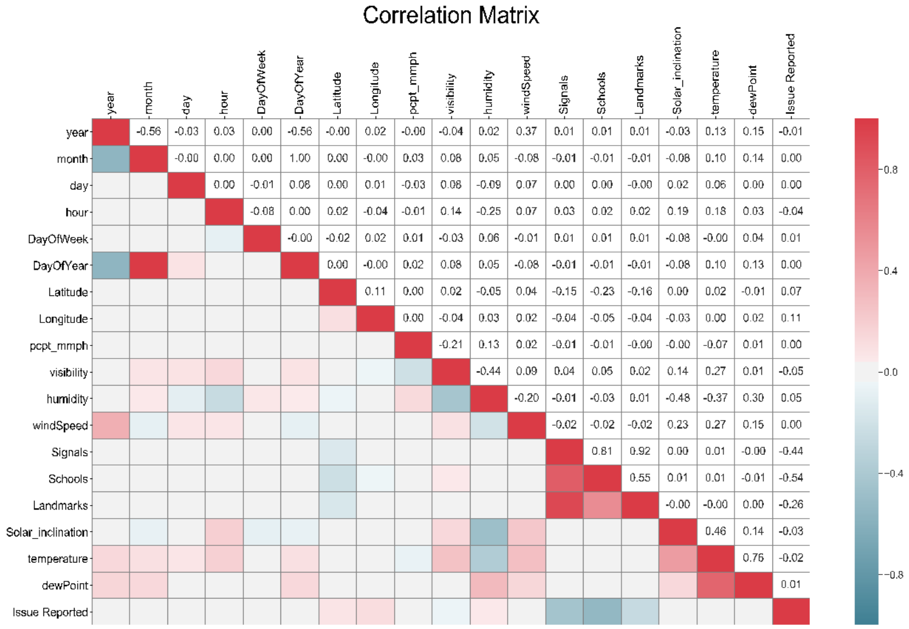


Figure 5: Heat map for the Pearson's correlation coefficient. The lower triangular part represents the color codes corresponding values of the coefficient are shown in the upper triangular region. The values of the color codes are shown in the color bar.

**Daily accidents analysis:**

We analyzed the distribution of the daily sum of the accidents in the city of Austin. The histogram of the daily accidents looks normal. Mean and standard deviation of the daily accidents in the city of Austin were confirmed to be 96 and 26, respectively.

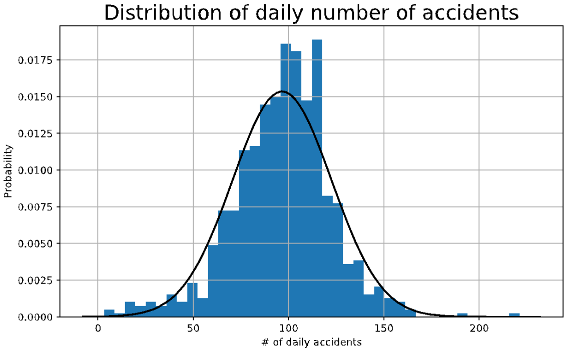


Figure 6: The histogram of the distribution of the daily accidents in Austin. The black solid line represents the corresponding normal probability distribution function

## Time Series Forecasting Models:

### Baseline Model (Persistence forecasting model):

Persistence forecast algorithm is a simple forecasting model. In this model, we assumed that one of the past values persists as the data in the future, i.e. we expected the periodic repetition of the event with some period. The root mean squared errors (RMSE) for each value of the assumed period were plotted and the value of time corresponding to the repeated minimum error will be selected as the period and apply as the predicted value for the future. In our data series, minimum RMSE was repeated every interval of 7. Therefore, our time series data was repeated weekly and we made forecasts accordingly.

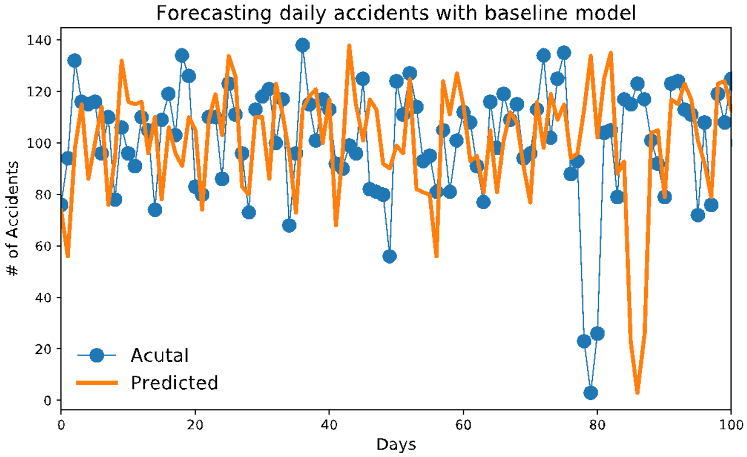


Figure 7. Daily road accidents forecast by using baseline model (persistence forecasting). Blue line with circles represents the actual test data and the orange line represents the corresponding predicted data. Days along the x-axis were ordered counts from the starting date of the test set.

**Autoregressive integrated moving average (ARIMA) model:**

From the above baseline models, we came to know that the forecasting is not efficient and does not perform well. Therefore, we used a regressive model which learns from the past data and predicts the future. One popular model is autoregressive moving average which uses the linear regression method to learn from the past data points. A condition for application of this method is that the past data should be uncorrelated, i.e. the time series data should be stationary. If the time series data has some trend, then the ARIMA algorithm uses the differentials until all the trends are removed and the data settled stationary. ARIMA has three steps: make series stationary (differentiation with degree 'd'), learn from the past data points (autoregression with number of past data 'p') and moving average of the past values (controlled by parameter 'q'). Therefore, ARIMA is a model controlled by three parameters (p, d, q).

To check stationarity of data, we applied ad-fuller-test. The null hypothesis: the data is not stationary therefore requires differentiation to make the data stationary. Based on the p-value, we decided whether the null hypothesis to be accepted or rejected. From the test, we found the ADF Statistic: -5.180979 and p-value: 0.000010. The data was stationary, no differentiation was required. Therefore, we chose d = 0.

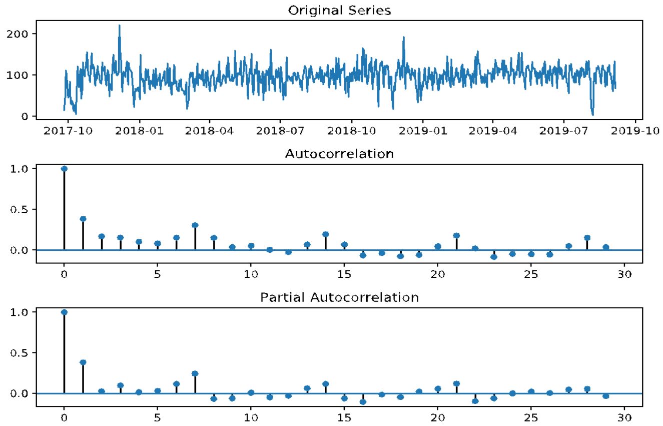


Figure 8. Original series, time autocorrelation and partial autocorrelation with respect to time from top panel to the bottom respectively.

To know the values of autoregression steps (p) and moving average steps (q), we plotted the autocorrelation and partial auto-correlation functions. After observing the autocorrelation and partial autocorrelation plots, we fixed the values of p and q as 3 and 2, respectively

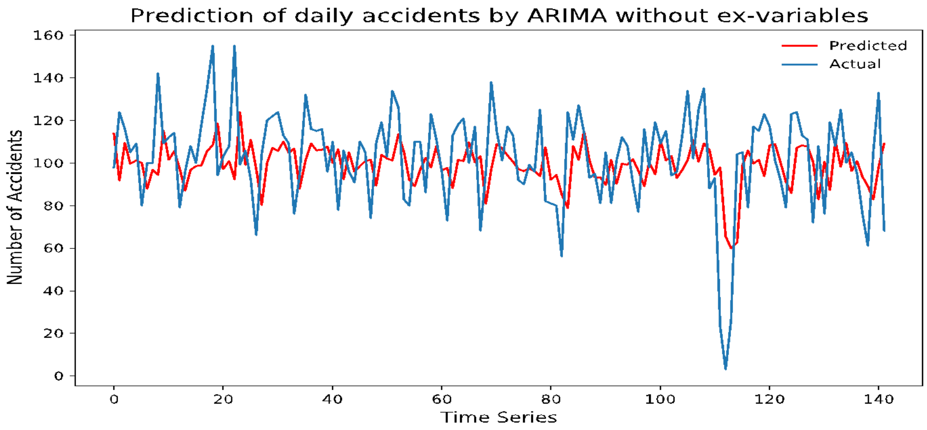


Figure 9. Predicted (red line) and actual (blue line) number of daily accidents for the test set of data. 'Time series' along x-axis were ordered counts from the starting point of the date in the test set.

Although ARIMA for our time series data was found to perform better than the baseline model, there was a systematic lag in predicted and actual values as plotted for the test set of data as in Fig. 9. Therefore, we applied the ARIMA with exogenous variables (external variables that influence the road accidents). We found a slight improvement of performance after including the exogenous variables.

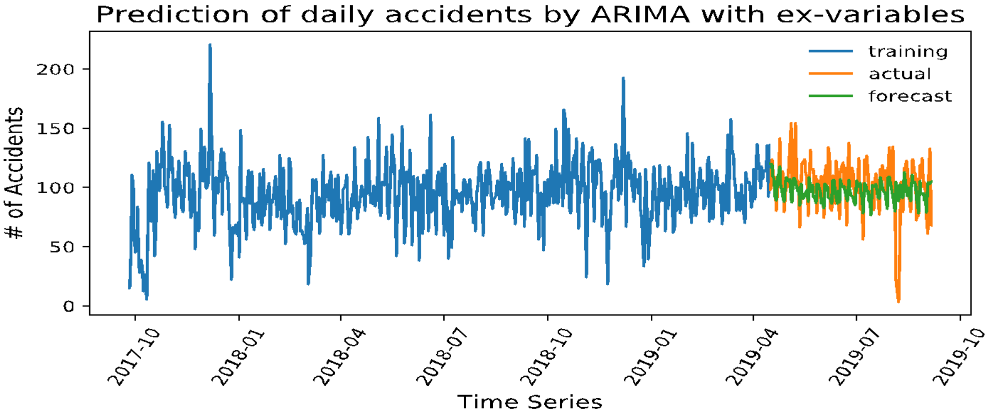


Figure 10. Predicted (green line), actual (orange line) daily counts of accidents in the city of Austin against date for the test set. Blue line represents the training set.

## Regression Models:

From the time series data analysis, we found that our datasets didn't follow any trend and seasonality. Therefore, we applied some predictive models to forecast the number of road accidents in the city of Austin. We also dived deeper to predict the number of road accidents at larger spatial resolution. In addition, the average road accidents each hour of any day in the city of Austin were also predicted. We performed the random forest (RF) regressor and extreme gradient boosting (XGB) models.

### Prediction of the daily accidents occurred throughout the city:

*Random Forest Regressor Model:*

We predicted for the test set of the data and compared the predicted data to the actual test set of target values. We applied random search method to tune the parameters and for the best set of parameters the random forest regressor yield the mean absolute error ~ 0.38 (i.e. accuracy ~ 62%)

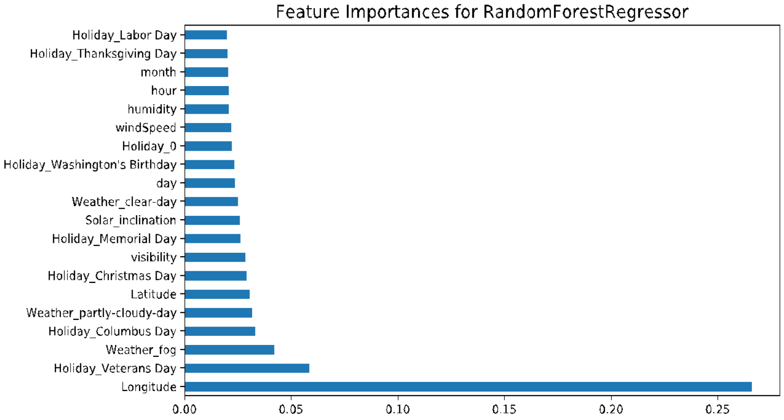


Figure 11. Feature importance plot for Random Forest model. Longitude was found to be the most important feature out of all other features.

From the feature importance of the estimators, one can state that the number of daily accidents in the city of Austin is mainly influenced by spatial location. Longitude was found to be the most important which was not surprising because the city area is extended from North to South (along the longitude) as can be seen on map shown in Fig. 3.

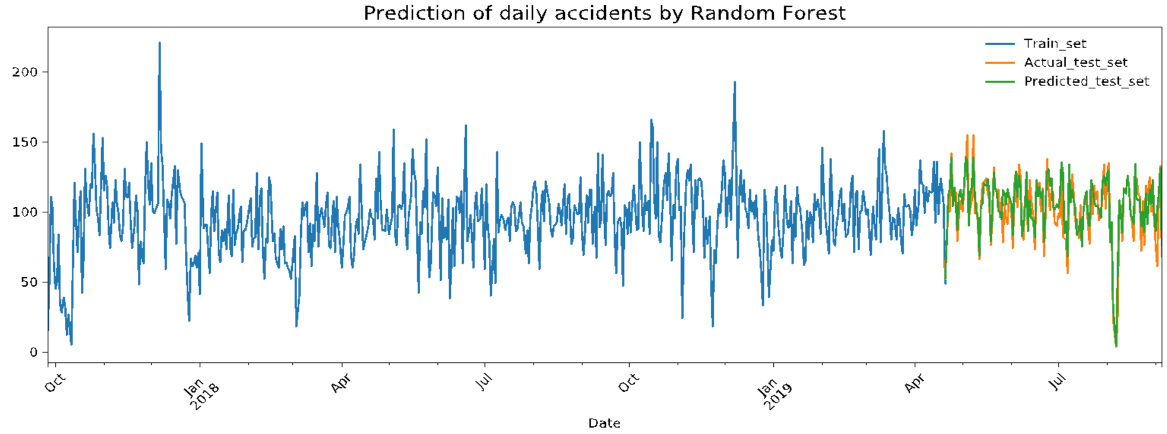


Figure 12: Actual daily sum of accidents throughout the city for training set (blue), predicted daily sum of accidents throughout the city for the test set (green) and actual daily sum of accidents throughout the city for the test set (orange). Random Forest model.

*eXtreme Gradient Boosting Regressor Model:*

For the XGB model, geographical coordinates were found to be the most important features. Other important features include solar inclination, wind speed, humidity, hour, dew point, temperature, day, visibility, etc. From the analysis of feature importance plot, it was found that holidays, year, month, and other features rarely affect the total number of daily accidents.

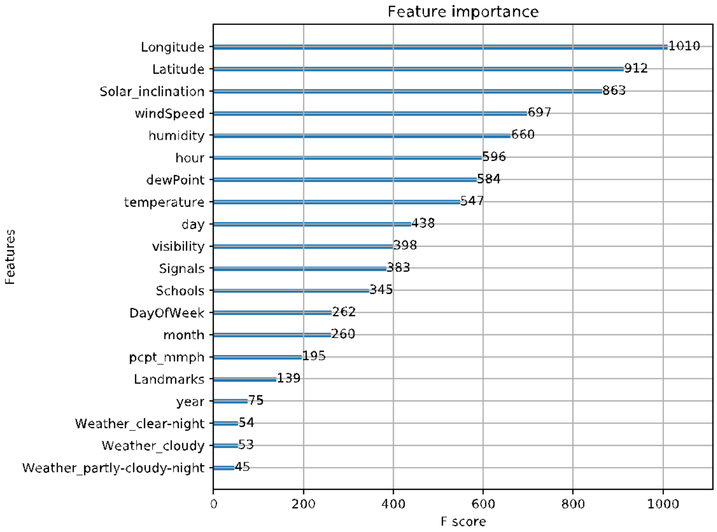


Figure 13. Feature importance for XGB model. Longitude is the most important feature. Note that other features are also found to be very important unlike the results from the RF model.

Figure 14 shows the actual value of total daily accidents from the training set, actual and predicted values for the test set of the data throughout the city. Comparing Figure 12 and 14, we can guess that the XGB model performed better for predicting the data.

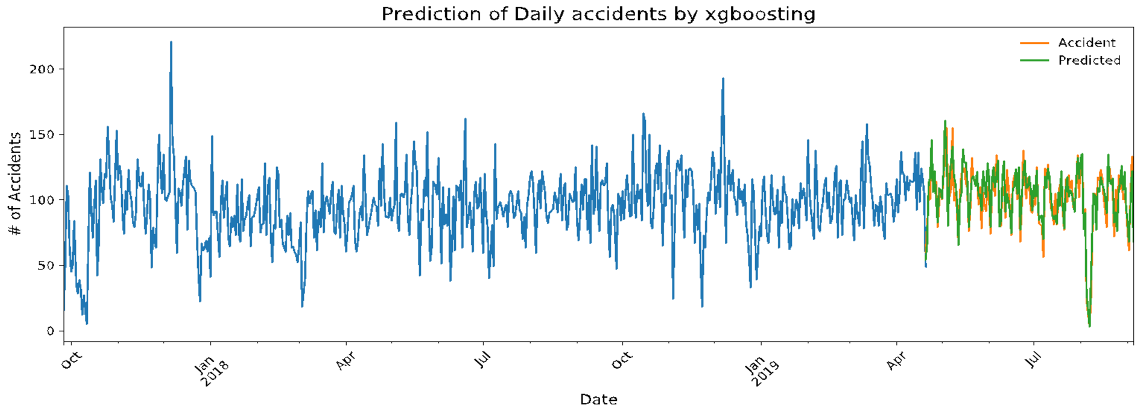


Figure 14: Actual daily sum of accidents throughout the city for training set (blue), predicted daily sum of accidents throughout the city for the test set (green) and actual daily sum of accidents throughout the city for the test set (orange). XG Boosting model.

## Prediction of the road accidents at different Locations of the city:

The total number of accidents were predicted for the test set of data selected randomly from the original data. Our randomly selected test data set was only 20% of the whole data set. We plotted the actual and predicted data obtained from both of RF and XGB models:

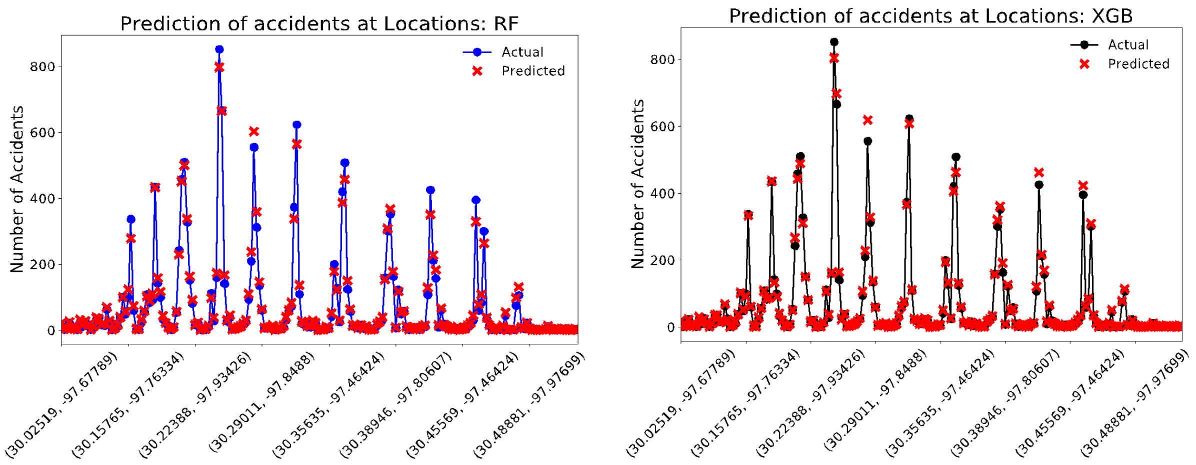


Figure 15. Actual and predicted number of accidents at specific locations of the city. Data shown in the figure contains only the test set of the data: Random Forest model (left panel) and XG Boosting (right panel).

**Prediction of daily road accidents at different Locations of the city:**

We also analyzed the daily accidents at specific locations. The predicted and actual values of the daily accidents at few locations having maximum number of accidents in the test set of data have been shown in the figure below.

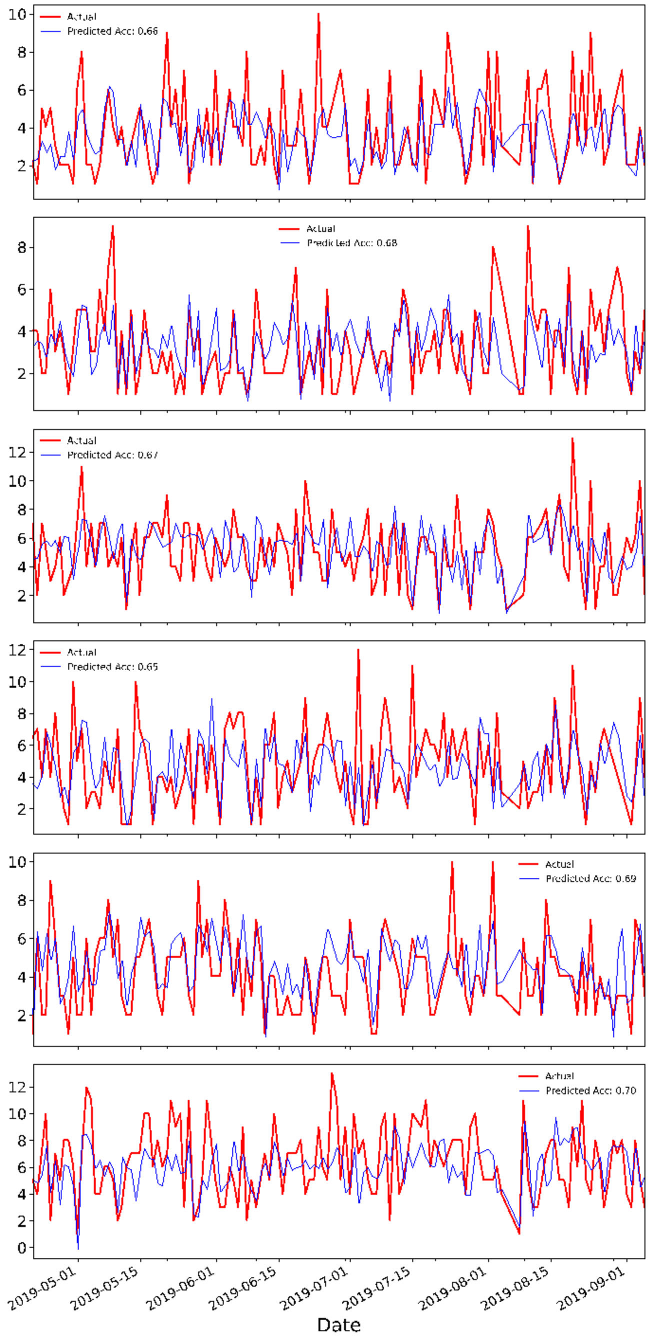


Figure 16. Actual and predicted number of accidents at specific six locations of the city. Both the actual and predicted data shown were from the test set of data. The accuracy (calculated by subtracting the mean absolute percentage error (MAPE) from 1.0) has been shown in the legend of respective window.

**Hourly accidents:**

Features that are important occurring the total daily accidents were not found equally important on analyzing the hourly accidents on any day of the year. We predicted the hourly sum of the accidents throughout the city and also at specific locations. Some of the results are plotted in the following figures.

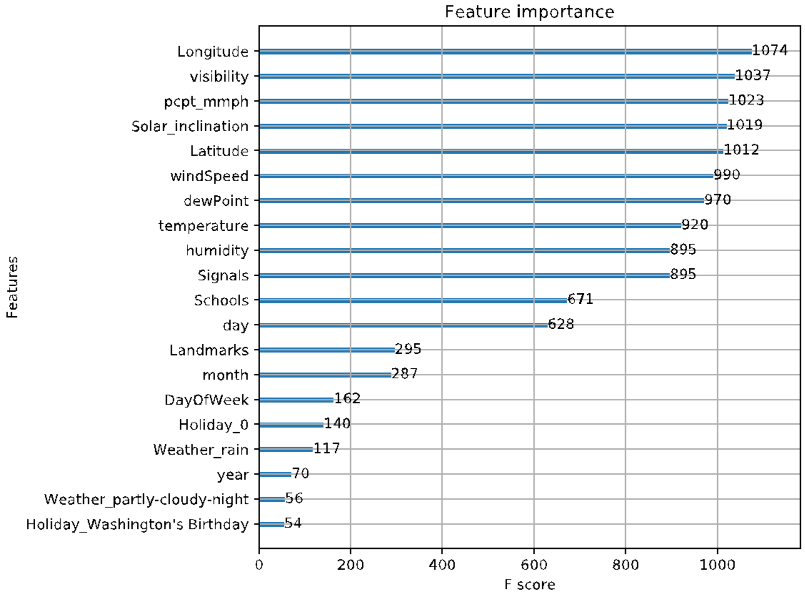


Figure 17. Feature importance for the hourly accidents according to the XGB model.

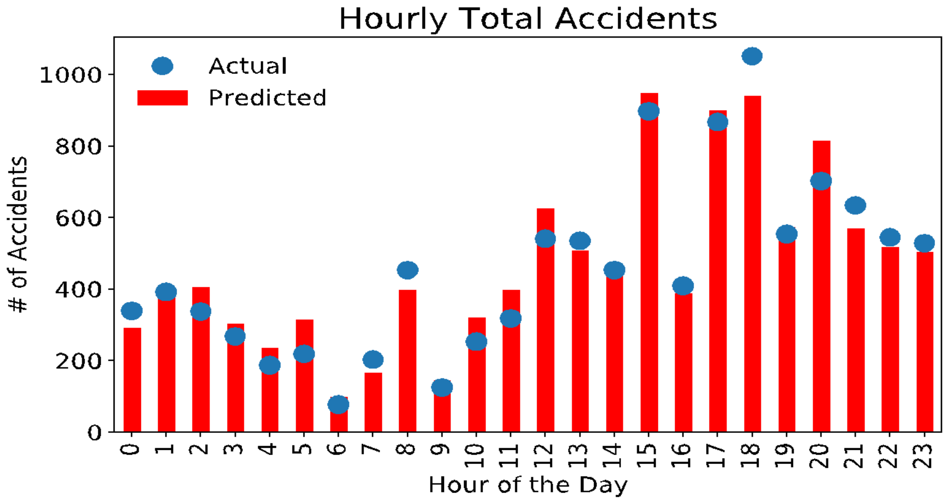


Figure 18. The actual (blue circles) and the predicted (red bar) values of the number of hourly accidents in the whole city of Austin. The plotted data represents only the test set of the data. We used the XGBoosting regressor model to predict the data.

We also analyzed the hourly accidents at specific locations of the city. In Figure 19, we have shown the predicted and actual road accidents at few selected locations.

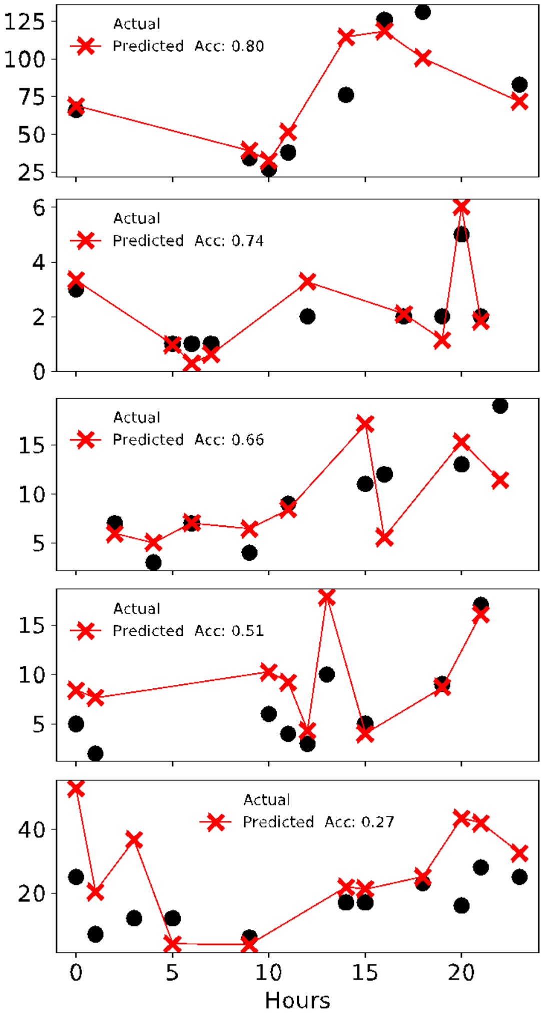


Figure 19. Actual and predicted values of the hourly accidents at five selected locations of the city. The actual and predicted values plotted here were taken only from the test set.

## Cross Validation and Hyper-parameter Tuning:

As mentioned above, during the implementation of decision tree algorithms, we assumed auto accidents as random events. We split train and test data without breaking the time series, however, we applied k-fold cross validation (CV) during training the data for RF and XGB predictive regressors. We found that the $R^2$ score saturated to fix value for CV folds larger than 5. Therefore, we used the 5-fold cross validation. To tune other hyperparameters we used the randomized search algorithms allowing sufficiently large number of iterations. All the predicted results shown in the figures in the previous sections were obtained using the best parameters resulted from the randomized search cross validation tuning. For XG boosting model, we also tuned the parameter ‘number of boosting’ by plotting the RMSE as a function of number of boosting during the cross-validation process. The boosting number was chosen so that the

RMSE becomes saturated its minimum value. In the following figures, we have shown some representative plots of tuning parameters.

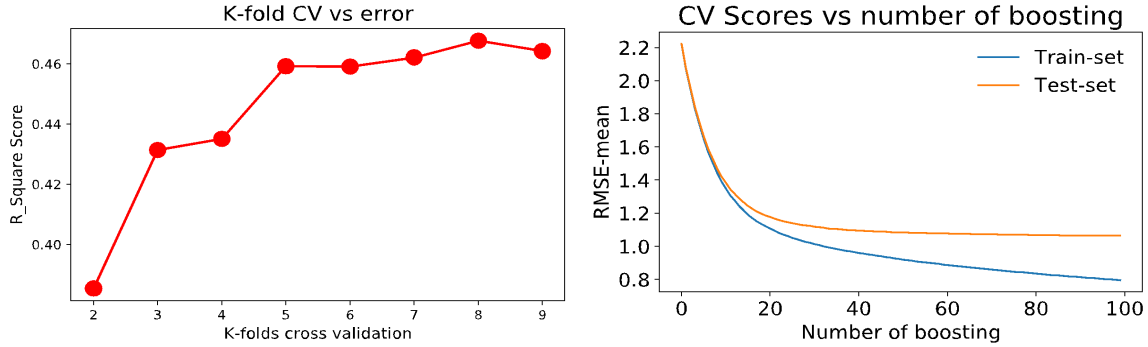


Figure 20. left panel: The $R^2$ score as a function of k-fold cross validation for Random Forest model. Right panel: mean of RMSE as a function of number of boosting for XGB model. Note: $R^2$ is 'higher is better' and RMSE is 'lower is better' types of errors.

## Model Evaluation:

Validity of the model was observed by calculating the error metrics. We calculated three different errors: $R^2$ score, RMSE and MAPE. We have tabulated the error metrics for all models in the following table. All these error metrics were calculated for the test set of data.

Table 1: $R^2$, RMSE and MAPE for various models. 'Accuracy' was calculated as (1-MAPE)

| | **Daily road accidents throughout the city** | | | | | **Daily at Locations** | | **At Locations** | | **Hourly** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Baseline** | **ARIMA** | **ARIMA with ex-variables** | **RF** | **XGB** | **RF** | **XGB** | **RF** | **XGB** | **At Locations (XGB)** | **Total (XGB)** |
| **$R^2$** | -0.88 | 0.03 | 0.06 | 0.75 | 0.80 | 0.55 | 0.61 | 0.99 | 1.0 | 0.93 | 0.94 |
| **RMSE** | 31.09 | 22.0 | 21.44 | 11.4 | 10.0 | 1.2 | 1.1 | 15.0 | 9.6 | 8.7 | 67.8 |
| **MAPE** | 0.21 | 0.17 | 0.15 | 0.09 | 0.08 | 0.38 | 0.34 | 0.11 | 0.06 | 0.26 | 0.10 |
| **Accuracy** | 0.79 | 0.83 | 0.85 | 0.91 | 0.92 | 0.62 | 0.66 | 0.89 | 0.94 | 0.74 | 0.90 |

Using Folium plugin, we have shown the total actual number of daily accidents for days from the test set of data by Green circles. The red circles on the map represent the magnitudes of MAPE. All the mean absolute errors were scaled for visualization purposes.

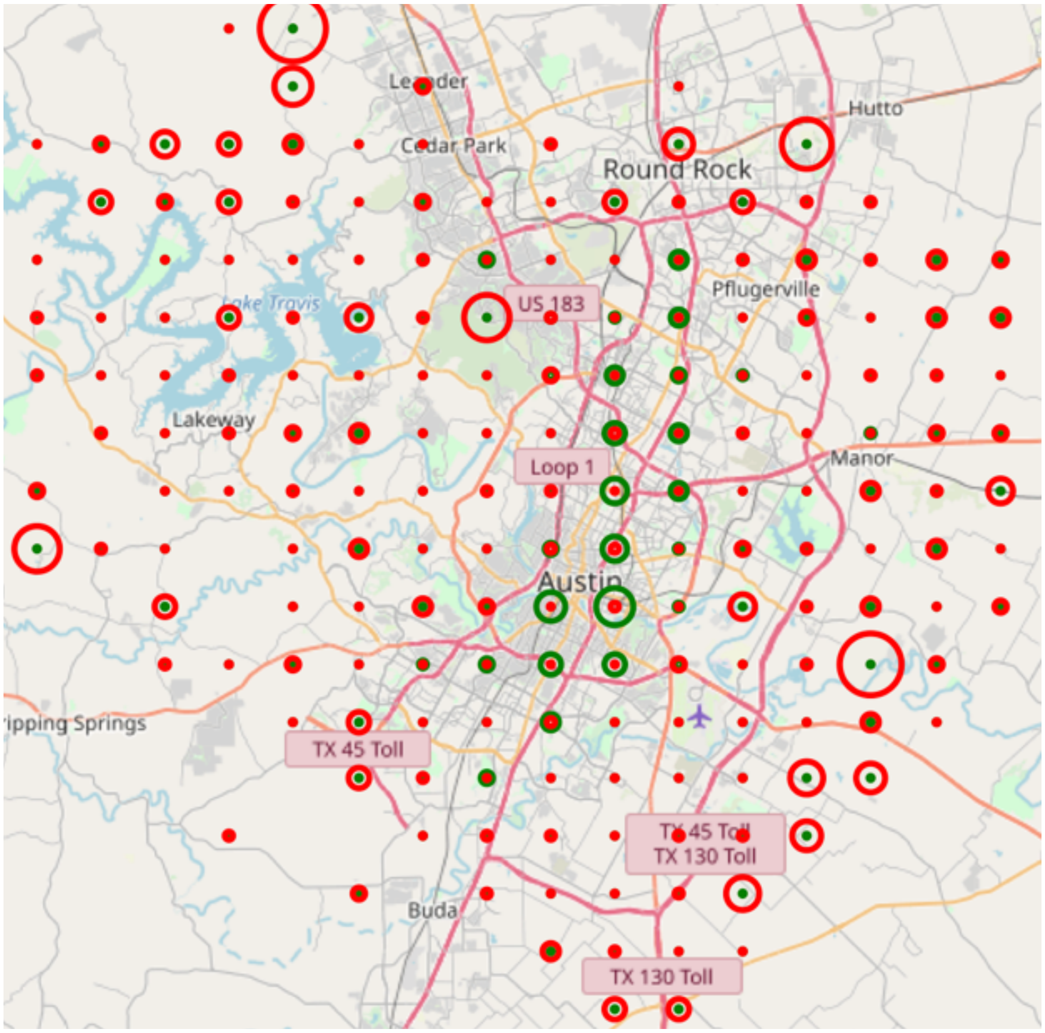


Fig. 21: The green circles: actual accident count around particular location, red circles: scaled absolute errors.

## Space for Future Works and Limitations:

We applied three different models, such as general time series analysis, random forest regressor and extreme gradient boosting model to predict the total number of accidents in the city of Austin. Our regression models predicted the spatiotemporal counts of accidents in the city. The work can be extended applying models to classify the severity of accidents. Prediction of binary classification of accident or 'no accident' also can be modeled by adding the imaginary random samples for 'no accident' corresponding to the time and space other than the actual accidents time and spot.

The time series data for road accidents are neither seasonal nor with some trends, which causes the time series analysis more challenging. Some of the important features such as road structures, population density, socio-economic levels of local people limit the performance of models. Those variables change over time and space; therefore, collection and implementation of those variables are challenging. Another variable that limits the performance is behavior of the involved people, which is very complex and difficult to quantify.

## Conclusions:

The number of road accidents in the city of Austin at various spatiotemporal resolutions were predicted using various machine learning models including time series analysis. Extreme gradient boosting model outperforms the other models that were used in this project. Random forest model performs far better than the autoregressive time series model ARIMA, although ARIMA itself performed better than the baseline model for time series analysis. The spatial accident counts have been predicted with performance better than the daily accident counts.

## Recommendations:

- The number of road accidents in the city of Austin is independent of day, month, day of the week and holidays.
- Hourly counts of accidents depend on the time slots of the day.
- We recommend drivers to avoid the accident risk zones when driving in the afternoon and evening, because most of the accidents occur in the afternoon and reached a peak in the evening. We also suggest drivers avoiding the north-south highways at certain longitude with accident risk spots.
- During the combination of high rainfall at very cold temperatures, we recommend skipping the driving in the city area, because at that conditions we found peaks on the daily accident curve.
- Traffic authorities are also suggested to pay their attention to the above-mentioned localities and conditions for the effective management of road safety and regulations.