Ambulance Location Prediction to Serve Future Accident Hotspots

Amol Ravatale   
Student  
Great Lakes InstitutePune, India  
amolravatal@gmail.com

Sanku Lahiri  
Student  
*Great Lakes Institute*Pune, India  
hisanku@gmail.comDevendra Bharadwaj  
Student  
Great Lakes InstitutePune, India  
vyomdev@gmail.com

Kushal Thakral  
Student  
Great Lakes InstitutePune, India  
ramanthakral@gmail.com

**Abstract—** Road accidents contribute to approximately 1.35 million deaths [*source: WHO]* worldwide annually. A major factor in road accidents which decide mortality is post trauma care. This is dependent on how fast medical attention is provided through ambulances after accident. The accessibility of ambulance service is fragmented based on the development index of a nation. Wealthier and developed economies have very fast and reliable ambulance care services while developing and weaker economies lag drastically. Research shows that the survival rate of accident victim increases if an ambulance is in the vicinity of 3-4 miles of accident site to provide post trauma care. This means that if an ambulance could be placed beforehand in the vicinity of all such accident locations it would be able to serve victims faster. Our paper aims to develop Machine Learning based models using two approaches to predict locations of 6 ambulances for 8 timeslots in a day (24 Hrs.) based on the history of accident locations. For our analysis we have taken data set from “Uber Nairobi Ambulation Pre-ambulation” challenge which provides accident history, locations etc. for Nairobi.

Approach 1 attempts to find the locations to park ambulances based on ARIMA time-series forecasting model. Approach 2 builds a predictive model using Classification learning where it attempts to predicts possibility of accident at a particular location and accordingly finds an optimum location to park ambulance(s).

Finally, both the approaches have shown promising results through highly accurate predictions for 1 to 6 locations for future accident hotspots validated through test data.

**Keywords**— ARIMA, Future Predictions, Ambulance

# Introduction

The Uber Nairobi Ambulation Pre-Ambulation challenge envisions to find a solution to the problem of quick access to accident victims by placing ambulances at locations so that ambulances have to cover minimum distance from accident hotspots.

Every year Nairobi observes road accidents where fatality rate is higher due to non-accessibility of ambulance or huge time taken by ambulance to reach accident spot. One way to address the issue is to predict future accident hotspots and place ambulances in the vicinity for faster turn-around to reach victims. The problem is unique in its own way as the number of accidents observed vary with influence factors such as weather conditions, quality of roads, traffic and road safety signs, volume of traffic, category of roads – highway or street and accident location in urban or rural area. A comprehensive study on the differentiation of Emergency Medical service rescue time in fatal road accidents in US was done by Brodsky Harold in 1988[1]. This paper clearly shows that EMS response time in rural set up is almost 4 times than that in urban environment. Another important aspect in access to ambulance is the nearby hospitals or community health center as the ambulances are usually provided by these facilities. Another study by Dong, Shao and Xiong[2] predict the accident hotspots based on various influence factors like Pavement Factors, Environmental variables (Rain, Fog, Sun, Temperature etc), Roadway geometry, Traffic factors and historical accident locations in urban setting. Both the above papers use conventional machine learning approaches, like Unsupervised learning, regression trees etc to predict an accident location. Another paper by Zhang, Yang and Wushour [3] attempts a different approach using LSTM and Gradient Boosting Regression Trees to predict whether a location could be a probable accident hotspot. A very different approach is taken by Kumar and Toshniwal [4] in which they attempt to predict an accident hotspot for Indian roads using Clustering and Representative Time-series algorithms.

Our solution goes one step beyond the solution of only predicting the accident hotspots. Through both approaches, 1 & 2 we attempt to find 6 optimum locations in all of the 8 timeslots in a day where ambulances could be placed so that they can reach in fastest time.

# DATA COLLECTION AND ANALYSIS

We used Uber Nairobi accident dataset which is consolidated data from multiple sources, namely weather dataset and Segment info (road conditions, characteristics as well as behavioral characteristic) from Uber Challenge Website. We also focused on collecting data on other features which could contribute to accidents e.g. Sun elevation, Humidity etc. Overall, data was collected for the period 1-Jan 2018 to 30-June 2019. The entire data was also sampled per 8 time slots in a day where each slot contributed 3hrs starting 12AM every day.

**Approach 1: Prediction of Ambulance Loc using ARIMA**

The entire training dataset is split as per 8 different time slots namely, 12-3AM, 3– 6AM, 6 – 9AM, 9AM – 12 Noon, 12 Noon – 3PM, 3 – 6 PM, 6 – 9 PM and 9PM – 12 AM.

The following bar graph shows the number of days for which the data points are available for each time slot. On an average out of 545 days (Jan 01, 2018 – June 30, 2019) data is available only for 288 days. To build ARIMA model, we imputed the missing dates and values using forward fill approach

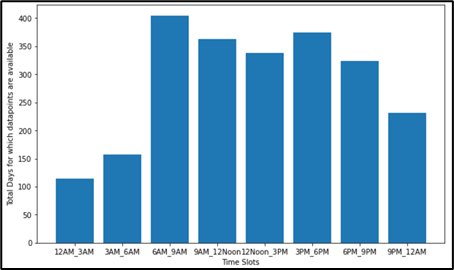


Figure 1: Accident Data available for Number of days for each slot

The below graph shows data for timeslot (12AM-3AM)

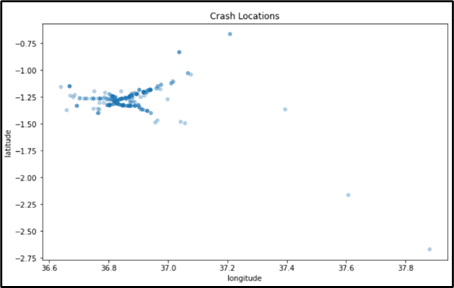


Figure 2: Crash Locations for 12 - 3 AM for Jan 2018 – June 2019

**Approach 2: Prediction – Feature Engg & Classification**

This approach is based on feature engineering and creating classifier to predict crash location. The primary data source and features which are used to build model are depicted in the below figure:

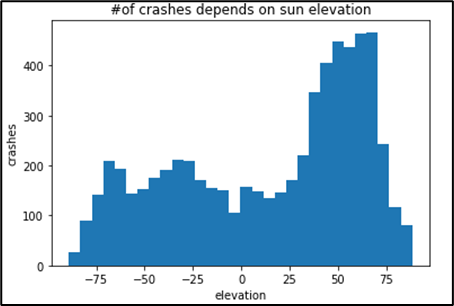


Figure 3: Number of Crashes Vs Sun Elevation

# 

Figure 4: Number of Crashes Vs Day of Week

# 

Figure 5:Number of Crashes Per Hr

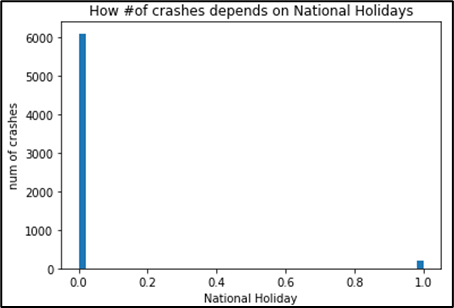


Figure 6: Number of Crashes Vs National Holiday

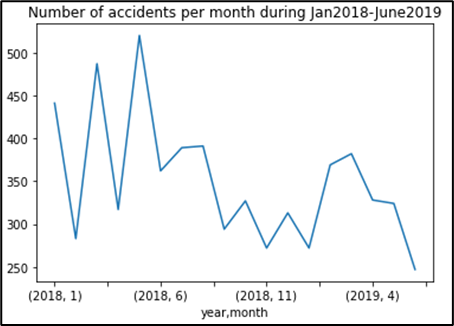


Figure 7: Number of Crashes Per Month

# SOLUTION APPROACH

## **Prediction Using Time Series(ARIMA)**

The below figure depicts the brief workflow of the solution using ARIMA for prediction ambulance locations.

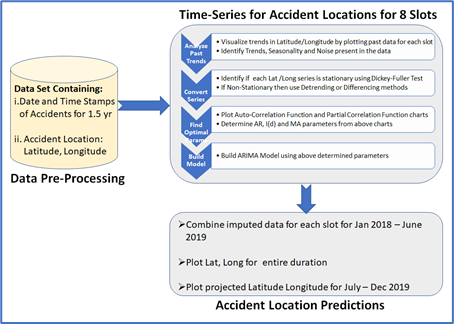


Figure 8: Approach 1 - Workflow

The first approach using time series analysis is formulated in the following steps:

1. Filter data as per respective 8 time slots for entire duration (Jan 2018 – June 2019) of provided data.
2. Segregate data for Latitude and Longitude separately for each slot. We’ll create time series separately for Latitude and Longitude.
3. Apply “centroid algorithm” (using *Shoelace formula*) to find a centroid for a particular slot on a given day for e.g. if the data in ‘b’ above has 4 accident locations for slot 3 PM – 6PM then this centroid algorithm will find 1 representative centroid for that slot. Similarly, repeat for all the slots across entire period.
4. Derive location centroids for each timeslot in a day individually for Latitude and Longitude for the entire period.
5. Segregate dataset with date and location centroid derived above in train and test sets
6. Apply MSE/RMSE to derive accuracy on test set.
7. Use ARIMA time series to predict the next location (centroid) for the period (July 2019 – Dec 2019) for Latitude as well as Longitude.
8. Finally, for every slot in each day assign the centroid derived in ‘g’ above to all the 6 ambulances locations.

The below figure depicts Latitude and Longitude of original accident locations for slot 12 AM-3AM:

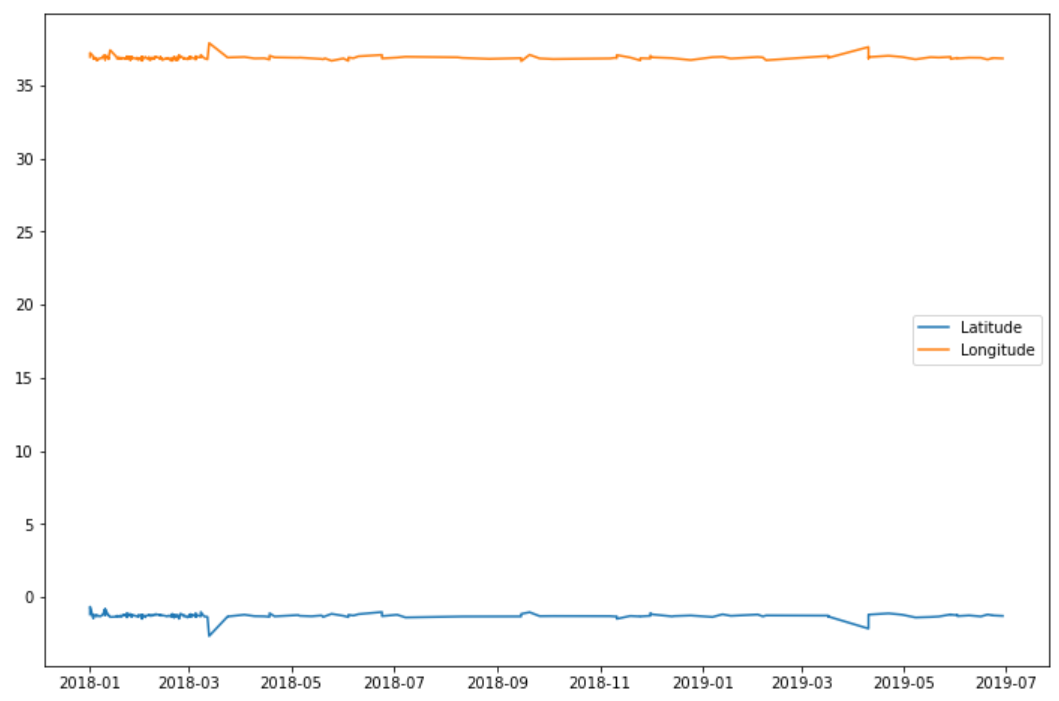


Figure 9: Accident Location: Lat & Long, 12-3AM

The below graph shows latitude and longitude for entire period:

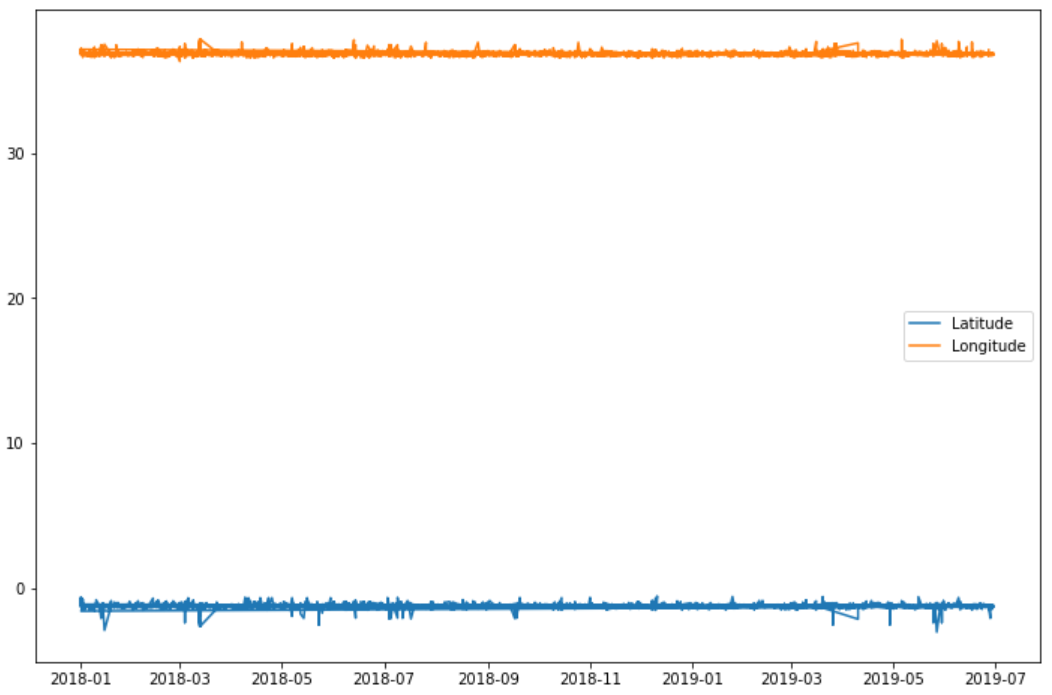


Figure 10: Accident Locations: Lat & Long for Jan’18 – June’19

## **Predictions with Feature Engg & Classification Model**

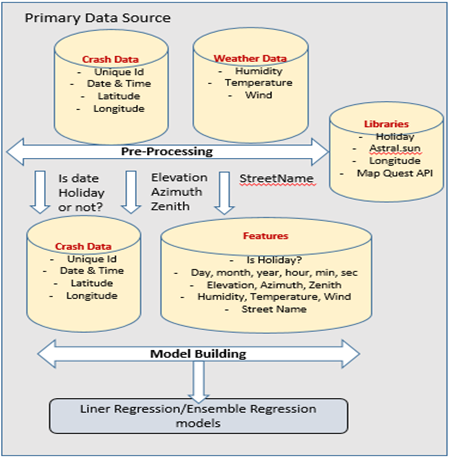


Figure 11: Approach 2: Workflow

Classification approach involves below workflow steps

1. Given data is Crash date & time, Crash location (latitude & longitude) from 1-Jan-2018 to 30-June-2019 [Total - 6318 records]. All these records are accidents happened.
2. Build features – Date-time features (Day, month, year, hour, min, sec), Weather features (humidity, temperature, wind speed), Sun elevation (elevation, azimuth, zenith), Day of week, is day off or not, is day national holiday or not.
3. Target label (prediction): Create target field ‘Accident’ (happened=1 & not happened=0). The given records (6318) will be marked as Accident=1.
4. Create 3 new records for each given record which represents Accident=0.

For this we took 1 record, changed hour & minute part of it randomly and created 3 records with same crash location denoting accident not happened.

E.g. given record: 01-01-2018 00:25:46 & crash location (L1) create 1) 01-01-2018 02:12:46 & same crash location L1 2) 01-01-2018 06:45:46 & same crash location L1 3) 01-01-2018 01:54:46 & same crash location L1

1. As per above step (iv) total record count =25272 consisting of accident happened and not happened.
2. Crash data + features + Label -> split data into 75-25 train test split.
3. Build classification model with different models (like SVM, Random forest etc.), check train & test accuracy.
4. Take future date, create features and generate random crash location and using model built above predict if accident will happen or not.

Libraries used for feature extraction

1. Holidays – To determine whether a day is a holiday or not
2. Astral - is python package for calculating the time of various aspcts of Sun and phases of the moon. We have used it to extrtact elevation (the time when the sun is at a specific elevation for either a rising or a setting sun), azimuth (The number of degrees clockwise from North at which the sun can be seen) and zenith (The angle of the sun down from directly above the observer) of Sun.
3. Geosptial API – Used API provided by “open.mapquestapi.com” to extract streen name for given crash location (lattitude & longitude)

# EVALUATION AND RESULTS

## **Approach-1**

At first the data collected for each slot was converted toa format suitable for application of time-series. For this the date ranges were made consistent for each slot ranging fromJan 2018 – June 2019. Missing dates were filled and missing latitude and Longitude data was imputed. Next, eaach series was plotted to observe Trend and Seasoanlity. Figure below shows one such observation for Slot , 12AM-3AM for Latitude data.

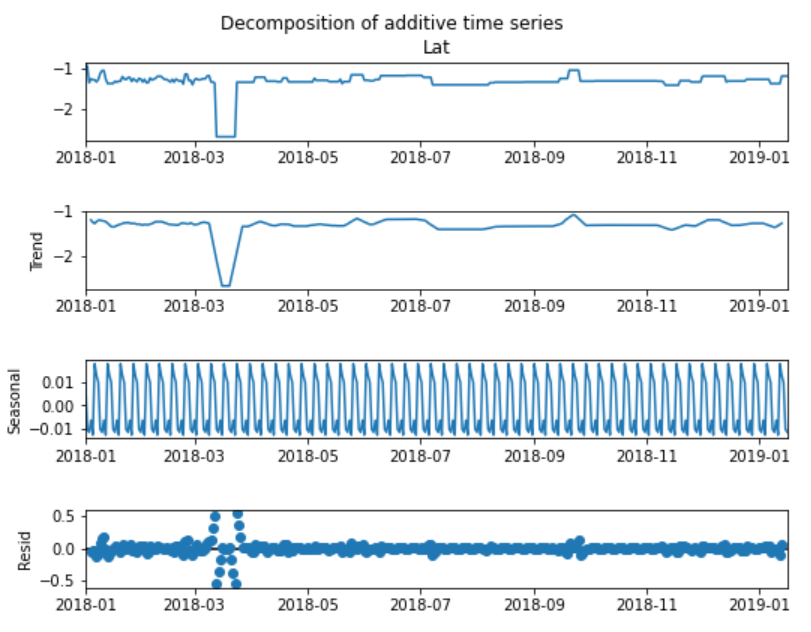


Figure 12: Checking characteristics of Time-Series

The stationarity of training data was evaluated through ADF test and p,d and q parameters were derived to be used in building ARIMA model.

Chart, histogram

Description automatically generated

Figure 13: ACF for Slot 12-3AM for Latitude data

Chart, box and whisker chart

Description automatically generated

Figure 14: PACF for Slot 12-3AM for Latitude data

Further the data was segredated in training and test sets in 70:30 ratio.

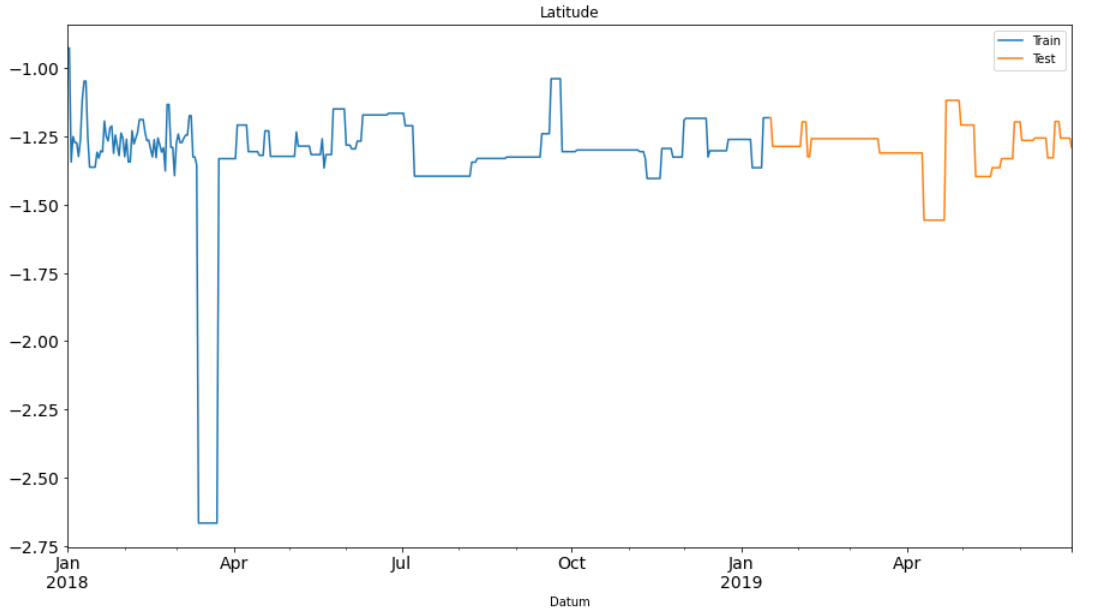


Figure 15: Train-Test Data for Slot 12-3AM

ARIMA model was trained on the training data set. The accuracy of the model was verified through prediction on test set. The figure below shows example of application of ARIMA on Latitude for Slot 01 and the results obtained on test set. The model has very high accuracy.

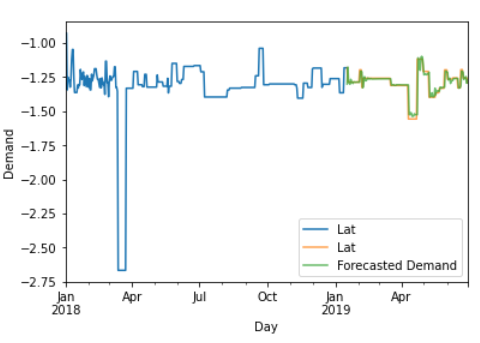


Figure 16: Validation on Test Set, Slot 12-3AM

Once validated the trained ARIMA model was applied to forecast Latitude data for Slot 12 – 3 AM for the period July 2019 – Dec 2019. The below figure shows the results achieved:

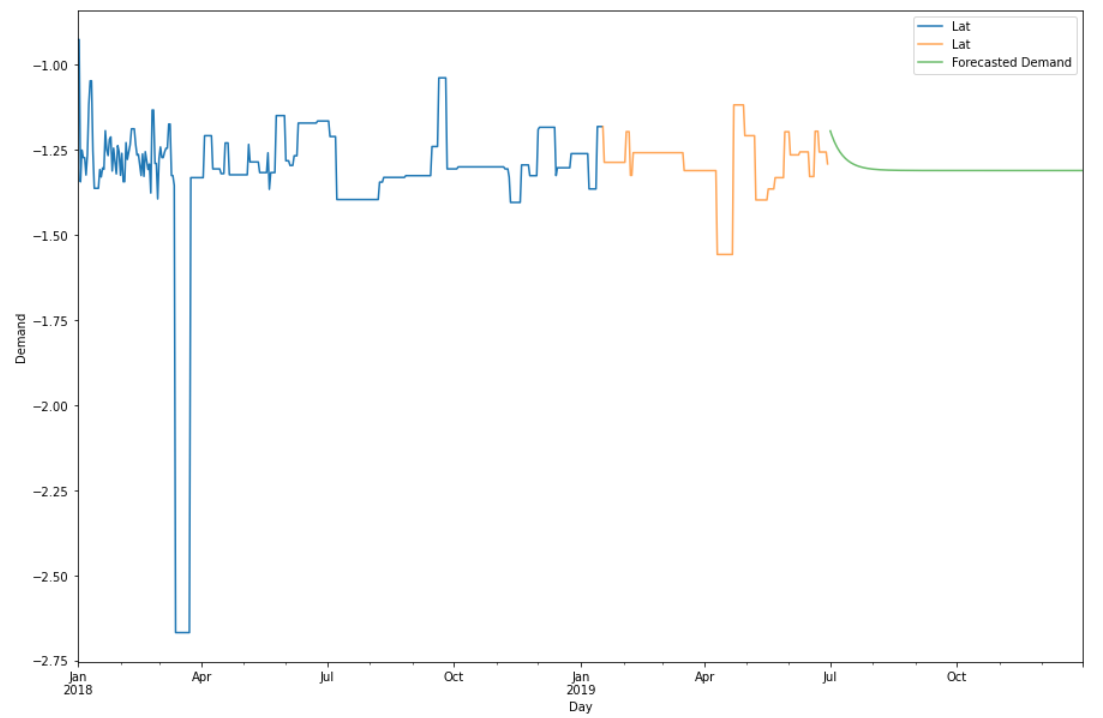


Figure 17: Predicted Latitude Co-ordinates for July – Dec 2019 for Slot 12-3AM

Similarly, longitude co-ordinates were also derived. The below figure shows the predicted ambulance locations in yellow against the backdrop of current accident co-ordinates.

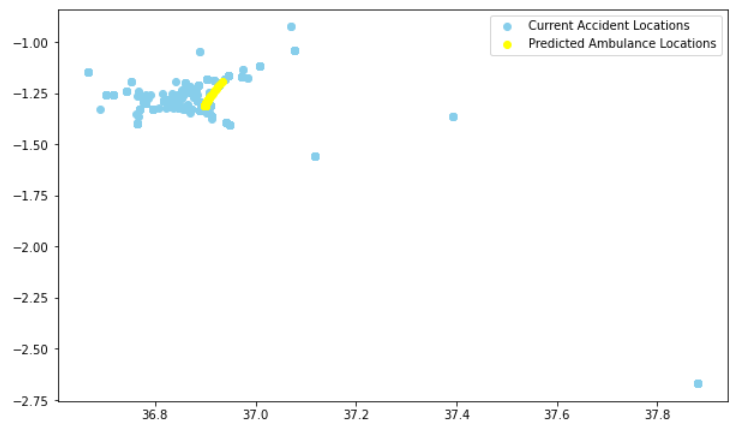


Figure 18: Predicted ambulance locations for July-Dec’19 for 12-3AM

Similarly, predictions for remaining 7 slots were derived for the period July -Dec 2019. MSE and RMSE were derived for all ARIMA predictions. In total, 8 ARIMA models for Latitude and 8 ARIMA models for Longitude for each slot were developed and tested. The table below shows the summary of model performance for all 8 slots:

Table 1: ARIMA test accuracy for all slots

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Slot** | **Latitude** | | **Longitude** | |
| MSE | RMSE | MSE | RMSE |
| **01**: 12AM-3AM | .002 | .049 | .002 | .039 |
| **02**: 3AM – 6AM | .003 | .057 | .003 | .058 |
| **03**: 6AM-9AM | .005 | .071 | .006 | .077 |
| **04**:9AM -12PM | .026 | .160 | .014 | .120 |
| **05**:12PM -3PM | .006 | .075 | .006 | .076 |
| **06**:3PM-6PM | .005 | .071 | .006 | .078 |
| **07**:6PM-9PM | .039 | .197 | .016 | .128 |
| **08**:9PM-12AM | .005 | .072 | .005 | .073 |

The final result depicting ambulance locations for the entire period July – Dec 2019 is shown below:

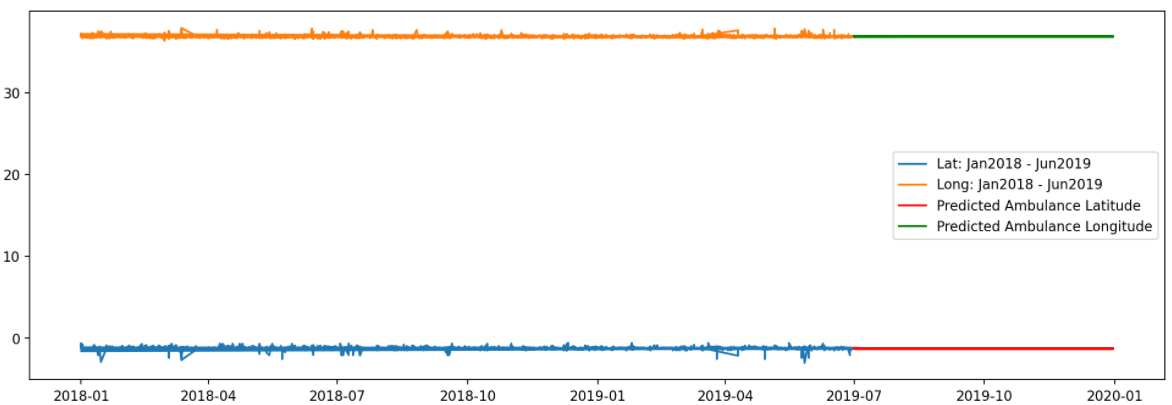


Figure 19: Predicted Locations – Lat & Long for July-Dec 2019

As observed the developed ARIMA models are highly accurate. However, sufficient data or benchmark is not available to compare the outcome of this study. The closest approaches which this model compares with in terms of accuracy is “A geographical location prediction method based on continuous time series Markov model, Du, Wang, Zhao, Guo[5] with precision in the range of 0.10 to 0.82.

## **Approach-2**

The feature importance metrics is shown below

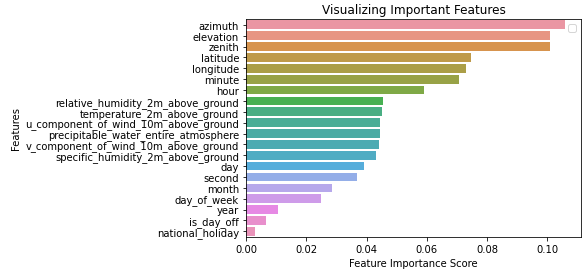


Figure 20: Feature Importance

Classification Model Results (Train-test data: 1-Jan-2018 to 30-June-2019)

Table 2: Classification Model Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Training-Test Data** | **SVM** | **Random Forest** | **Dense**  **Network** |
| Accuracy (Test) % | 75.82 | 81 | 77.9 |
| F1 Score % | 17.83 | 52 | 39 |
| Test Precision % | 72 | 78 | 63 |
| Test Recall % | 10 | 39 | 28 |

**Prediction for future date**

Step (1): In order to predict if an accident will happen or not in future date, we considered one date and generated random crash locations (the range used is based on train data). While generating crash locations randomly, we used latitude & longitude range (min, max) of actual crash locations.

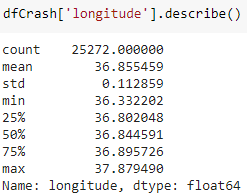
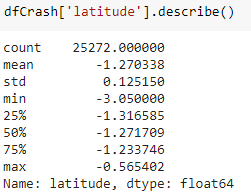


Figure 21: Latitude, Longitude Data Distribution

Table 3: Accident Predictions (For Date: 05-Sep-19, for 3000 Records)

|  |  |  |
| --- | --- | --- |
| **Time Zone** | **No of accidents Predicted-SVM** | **No of accidents Predicted-Random Forest** |
| 12AM-3AM | 0 | 0 |
| 3AM-6AM | 44 | 89 |
| 6AM-9AM | 34 | 275 |
| 9AM-12PM | 5 | 3 |
| 12PM-3PM | 0 | 3 |
| 3PM-6PM | 0 | 4 |
| 6PM-9PM | 0 | 0 |
| 9PM-12AM | 0 | 0 |
| Total | 83 | 374 |

Step (2): Used clustering model (KNN) to create maximum 6 locations, for each time zone, to place ambulance

1. Ambulance Location for Time zone -3 using Random forest classifier

|  |
| --- |
| **Ambulance Location For – Time zone 3** |
| [-2.74473827, 36.73875905] |
| [-0.98552441, 37.36197124] |
| [-1.70586475, 37.45590102] |
| [-0.94985487, 36.67220806] |
| [-2.63965247, 37.50083687] |
| [-1.89652329, 36.66578275] |

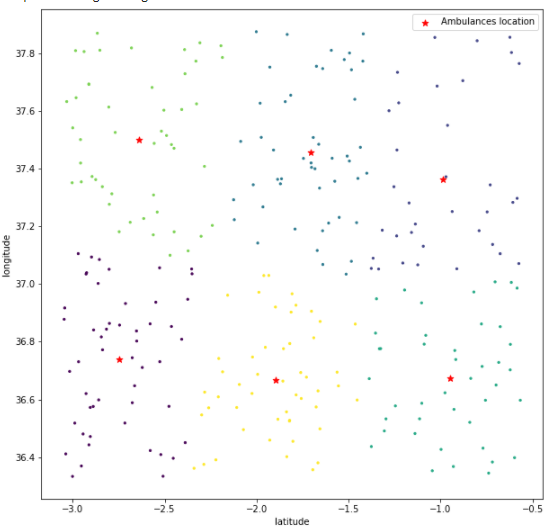
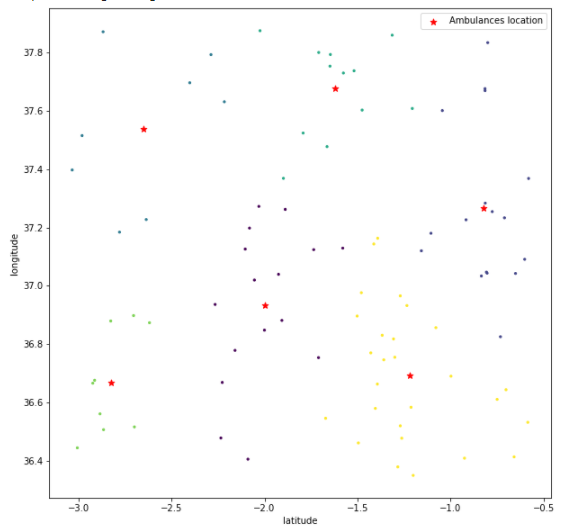


Figure 21: Ambulance Placement (marked in red)

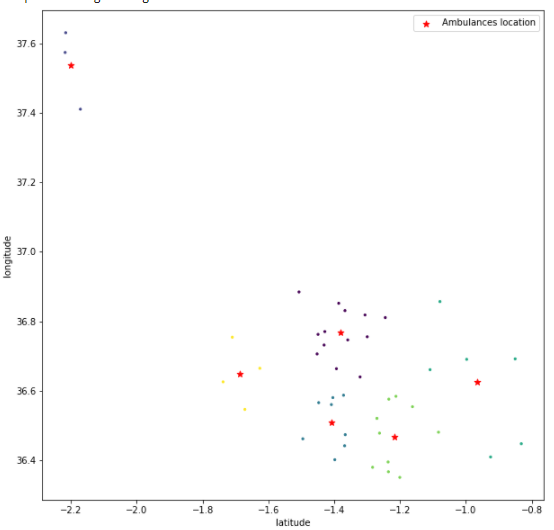
1. Ambulance Location for Time zone -2 using Random forest classifier

|  |
| --- |
| **Ambulance Location For – Time zone 2** |
| [-1.99900765, 36.93278538] |
| [-0.82084953, 37.26684971] |
| [-2.65011818, 37.54000612] |
| [-1.6222563 , 37.67812828] |
| [-2.82647136, 36.66900329] |
| [-1.21971844, 36.6930236 ] |

  
Figure 22: Ambulance Placement (marked in red)

1. Ambulance Location for Time zone -2 using SVM classifier

|  |
| --- |
| **Ambulance Location For – Time zone 2** |
| [-1.38015108, 36.76673074] |
| [-2.201255 , 37.53928814] |
| [-1.40723179, 36.50829414] |
| [-0.96501831, 36.62561554] |
| [-1.21774543, 36.46775386] |
| [-1.68590506, 36.64723214] |

  
Figure 23: Ambulances Placement (marked in red)

1. Ambulance Location for Time zone -3 using SVM classifier

|  |
| --- |
| **Ambulance Location For – Time zone 3** |
| [-1.31706048, 36.83578258] |
| [-2.01499278, 37.49369612] |
| [-1.02857965, 36.49216989] |
| [-1.66476216, 36.6438078 ] |
| [-1.35349569, 36.55504617] |
| [-2.21232559, 37.28651609] |

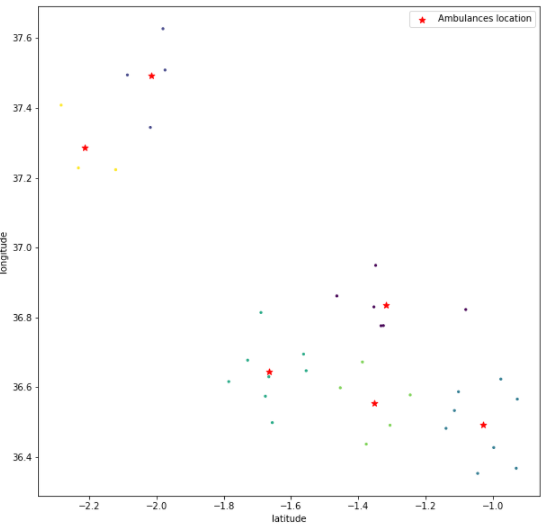


Figure 24: Ambulance Placement (marked in red)

Predictions for one month period

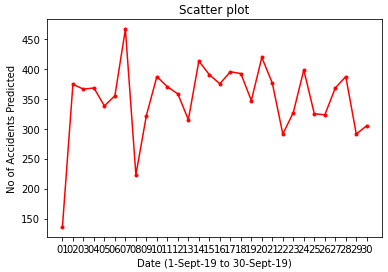


Figure 25: Accidents predicted for 1 month

As demonstrated above this approach can be used to randomly generate thousands of crash locations and predict if accident can happen or not. Such crash location can be clustered and accordingly ambulance location can be set.

# CONCLUSION

With Uber Nairobi challenge we attempted to provide a solution to an acute problem of Road Traffic Safety (RTA) through machine learning application. Though there are multiple solution approaches from machine learning perspective the problem goes beyond theoretical derivations and studies. Few implementations of machine learning have already been implemented and have achieved tremendous success in reducing fatalities by predicting accident hotspots as in the case of FLARE Emergency Response Technologies using automated ambulance dispatch centralized system. However, these kinds of systems are solely based on some local study based on patterns and are not based on machine learning due to which they may lose their efficacy with changing traffic behaviour. Machine learning application in such scenarios definitely add value and ensures agility of implementation as well. Moreover, the platforms using Machine learning could easily be scaled, optimized and customized with rules based on road topographies, local traffic behavior and public compliance of traffic rules. We could hence say that technology could really come to the rescue of human life.

# ACKNOWLEDGEMENT

We are highly grateful to Dr. Narayana, Director, Great Lakes for his guidance and support for our work. We are also indebted to our guide Mrs. Sulekha Aloor Ravi, Mentor, Great Lakes for her valuable support, feedback, and advice throughout project tenure.

# REFERENCES

|  |  |
| --- | --- |
| [1] | Brodsky, H. (n.d.). Emergency medical service rescue time in fatal road accidents. Retrieved July 12, 2021, from Trb.org website: http://onlinepubs.trb.org/Onlinepubs/trr/1990/1270/1270-011.pdf |
| [2] | Dong, C., Shao, C., Li, J., & Xiong, Z. (2018). An improved deep learning model for traffic crash prediction. Journal of Advanced Transportation, 2018, 1–13. |
| [3] | Zhang, Z., Yang, W., & Wushour, S. (2020). Traffic accident prediction based on LSTM-GBRT model. Journal of Control Science and Engineering, 2020, 1–10. |
| [4] | Kumar, S., & Toshniwal, D. (2016). A novel framework to analyze road accident time series data. Journal of Big Data, 3(1). doi:10.1186/s40537-016-0044-5 |
| [5] | Du, Y., Wang, C., Qiao, Y., Zhao, D., & Guo, W. (2018). A geographical location prediction method based on continuous time series Markov model. PloS One, 13(11), e0207063. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |