Introduction to Data Science 2016

NIST- FINAL REPORT

BY ABHISHEK NIGAM AND ABHISEK MOHANTY

NIST-19

Introduction

In this final iteration of the NIST project we have tried to identify and demonstrate most influential ways in which we can improve the prediction results which we showcased in the previous Lab. The techniques being highlighted in this documents are as mentioned below :-

1. Customizing algorithm to include hour and day calculation in prediction.

2. Using NOAA extreme weather data to identify correlation between extreme weather and traffic incidents.

3. Using NOAA precipitation data to identify correlation between rainfall and traffic incidents.

CONTENT INDEX :-

a. METHODS USED AND THEIR DESCRIPTION
    i. METHOD 1 : - Customizing algorithm to include hour and day calculation in prediction.
    ii. METHOD 2 :- Using NOAA extreme weather data to identify correlation between extreme weather and traffic incidents.
    iii. METHOD 3 :- Using NOAA precipitation data to identify correlation between rainfall and traffic incidents.
b. RESULTS AND OBSERVATIONS
c. CHALLENGES FACED

d. CONCLUSION

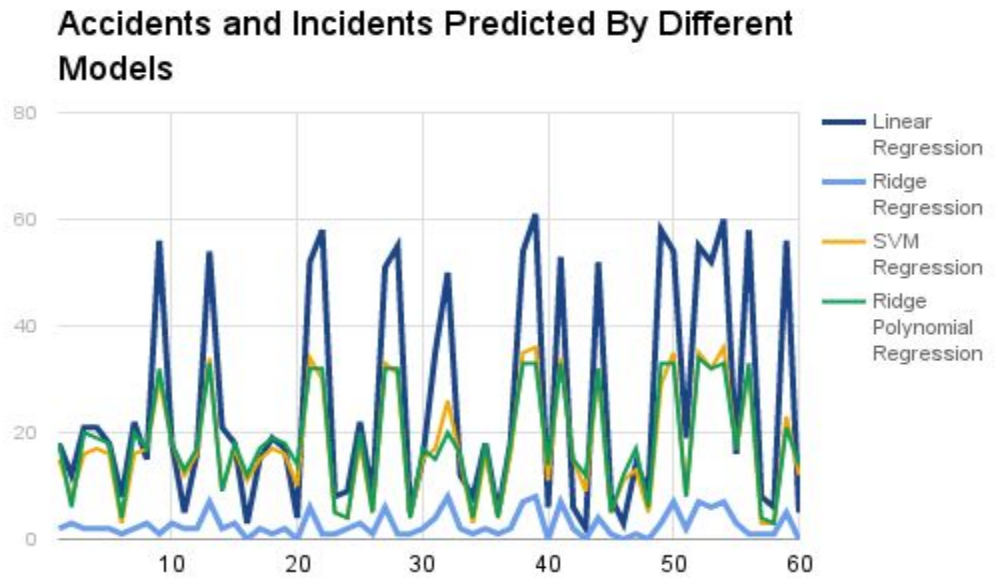
Methods used and their description :-

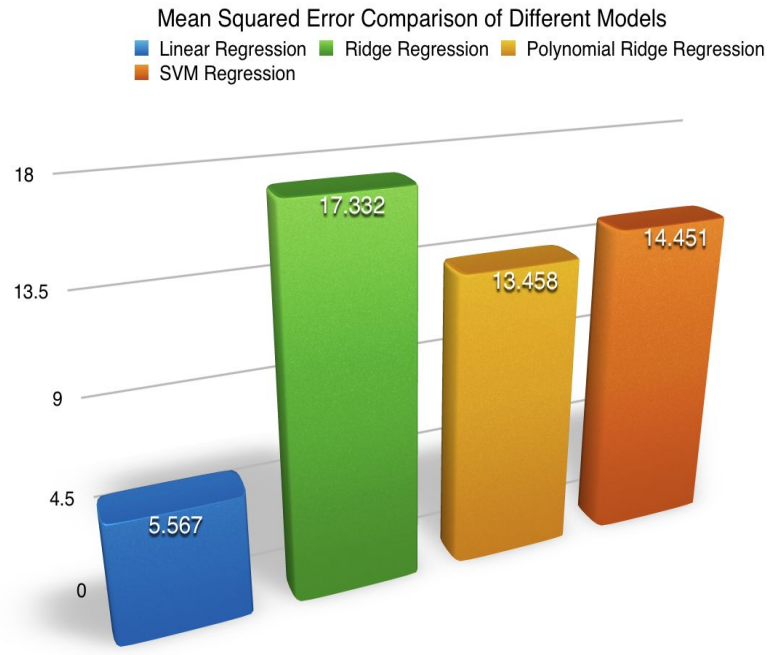
Method -1 : Customizing existing algorithm to include hour and day calculation

As we had observed in the algorithm presented in the Lab 11, there we considerable assumptions about the occurrences of the accidents and incidents as the training was being done on the year-wise occurrences of the events. This created serious faults in the prediction as the number of events which could occur in the provided Geobox time-period would differ significantly from the (predicted number of events over that year / 12).

This can be visualized generally by understanding that during spring break and summer, people tend to travel more and hence the possibility of accidents increases due to more vehicles being on road. Other than this, it has also been shown that weather plays a very important role in the number of traffic incidents being reported. During fog the number of accidents tend to increase due to low visibility.

The result of this monthly prediction was described in our previous report as well and for better comparison we are enlisting it here also :-





For this iteration, we went further and tried to observe what would happen if we included the "Hour-Day" also in our calculations. Though this approach we were trying to minimize the mean squared error which we had observed in our previous models.

The algorithm portrayed in the screenshot below depicts how we have divided the data into the groups based on "YEAR", "MONTH", "DAY", "HOUR". The algorithm is trained using four different algorithms which include :-

1. Linear Regression
2. Ridge Regression
3. Polynomial Ridge Regression
4. SVM

Please find the code below which segregates the data based on the hour-day-month-year calculations :-

```
df1 = pd.DataFrame(row[columns1])
df1 = df1.transpose()
df2 = pd.DataFrame(row[columns2])
df2 = df2.transpose()
df2.columns = columns1
x_predict = df1.append(df2)
intermediate_output = []
m =  df_event[(row["se_lat"] < df_event["latitude"]) & (row["nw_lat"] > df_event["latitude"]) & (row["se_lon"] > df_
if(len(m) > 0):
    data = []
    data.append(m[m["event_type"] == "accidentsAndIncidents"])
    data.append(m[m["event_type"] == "roadwork"])
    data.append(m[m["event_type"] == "precipitation"])
    data.append(m[m["event_type"] == "deviceStatus"])
    data.append(m[m["event_type"] == "obstruction"])
    data.append(m[m["event_type"] == "trafficConditions"])

    for i in range(0,6,1):
        cur_data = data[i]
        cur_data = cur_data.groupby(["year","month",'day','hour'])["count"].count().reset_index()
        columns = ['month', 'year','day', 'hour']
        if (len(cur_data) != 0):
            data_test = cur_data[cur_data['year'] == 2014]
            data_train = cur_data[cur_data['year'] != 2014]
            x_train, y_train, x_test, y_test = data_train[columns] ,data_train['count'], data_test[columns], data_t
            y = linear_regression(x_train, y_train, x_test, y_test, x_predict)
            intermediate_output.append(np.mean(y,axis=0))
        else:
            intermediate_output.append(0);
```

A sample of the predicted output values for Linear Regression and SVM can be found below :-

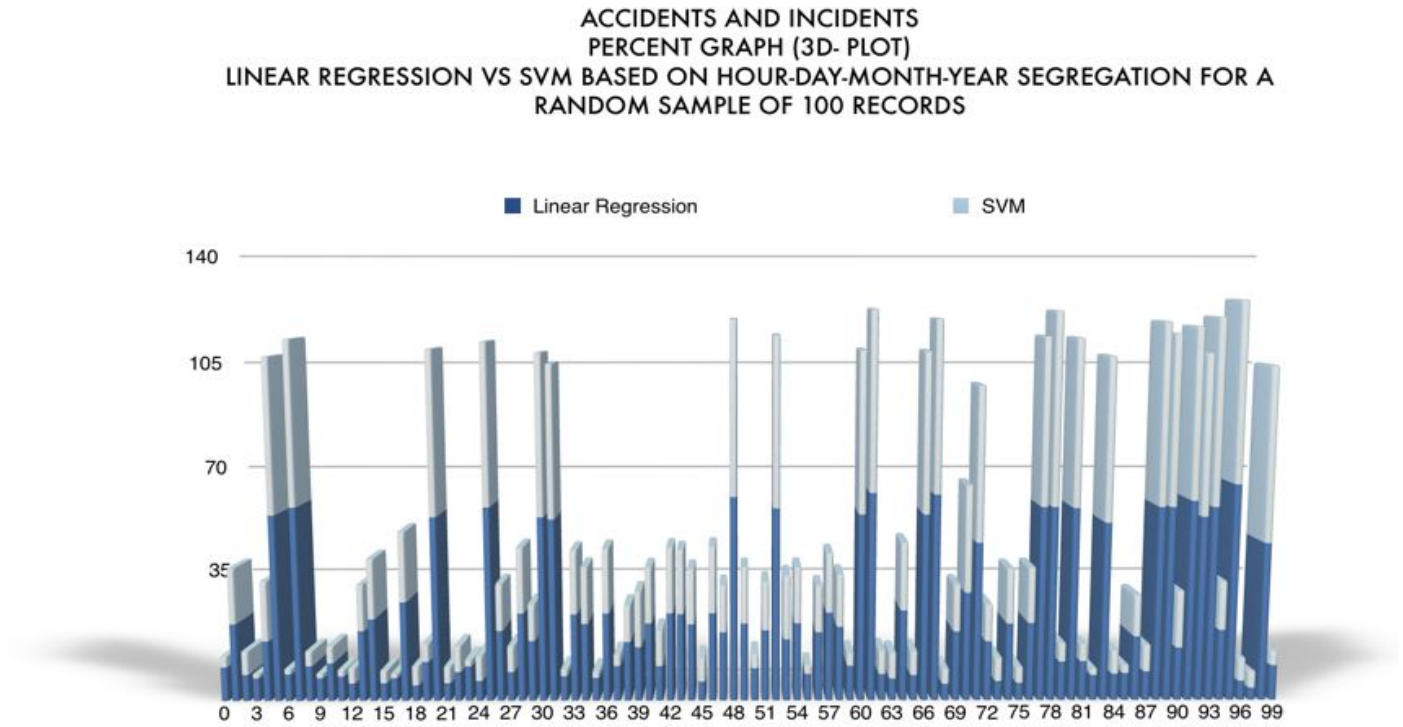
Linear Regression

3.4391844758697516	6.156870918599452	0.0	0.0	0.0
17.748816460813565	13.737595680535378	0.0	0.0	0.0
7.429784313616665	4.603808504129688	0.0	0.0	0.0
1.167173513117632	3.4815496875299914	0.0	0.0	0.0
19.162358111987487	10.230944448904438	0.0	0.0	0.0
49.38281614483094	19.41058106688888	0.0	0.0	0.0
1.559891236182466	4.326129207165877	0.0	0.0	0.0
52.310710402318364	20.500548126484773	0.0	0.0	0.0
4.604955714798507	5.2036647720577776	0.0	0.0	0.0
1.30258350980057	2.89748429813514	0.0	0.0	0.0
5.336521022259149	5.997143682301271	0.0	0.0	0.0
1.8973631592753009	1.6652576946289628	0.0	0.0	0.0
4.73487258864327	3.4564851636586127	0.0	0.0	0.0
14.707638404847671	13.33748110907436	0.0	0.0	0.0
19.156974377561482	17.572151845928147	0.0	0.0	0.0
3.8374697138366827	2.6798110845566896	0.0	0.0	0.0
1.167173513117632	1.0671503273829899	0.0	0.0	0.0
22.335753385191538	19.829212449448278	0.0	0.0	0.0
5.968737960094813	6.4186710729538845	0.0	0.0	0.0
5.721344832688374	7.806498321704339	0.0	0.0	0.0
52.24165574669314	19.61097137919205	0.0	0.0	0.0

SVM Regression

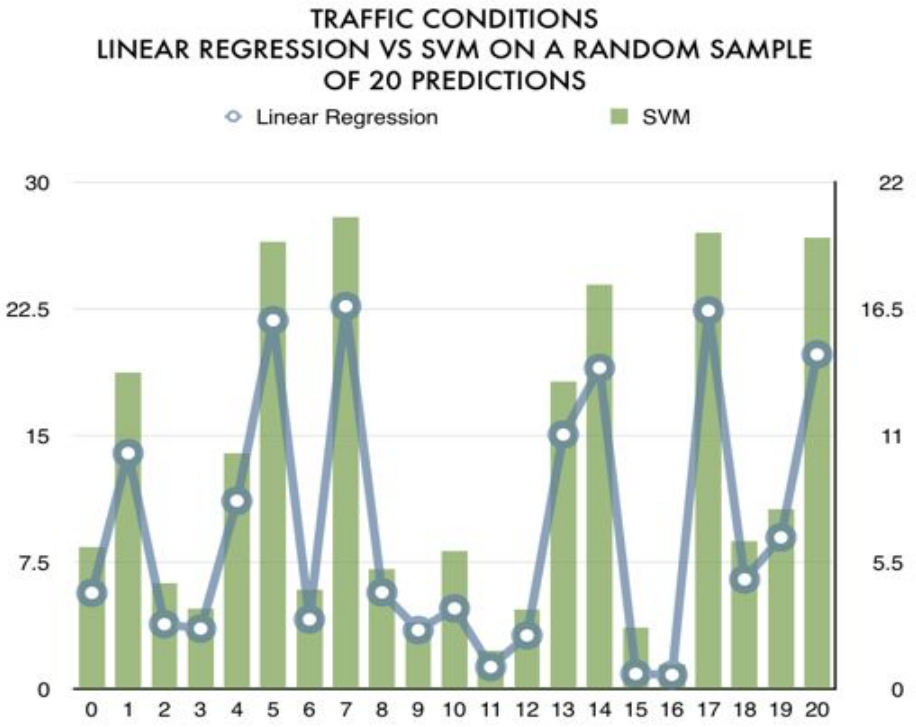
10.91507937	5.7	0	0	0	0
24.78888889	13.96666667	0	0	0	5.78015873
8.483333333	3.85	0	0	0	0
7.387301587	3.583333333	0	0	0	0
19.26666667	11.15	0	0	0	5.562698413
59.43174603	21.83333333	0	0	0	5.926984127
8.642063492	4.133333333	0	0	0	0
61.91666667	22.66666667	0	0	0	5.829365079
11.24285714	5.733333333	0	0	0	0
7.405555556	3.5	0	0	0	0
12.13888889	4.783333333	0	0	0	0
7.942857143	1.316666667	0	0	0	0
5.658333333	3.183333333	0	0	0	0
22.58412698	15.06666667	0	0	0	5.78015873
26.31428571	19.01666667	0	0	0	5.78015873
5.736111111	0.916666667	0	0	0	0
7.387301587	0.85	0	0	0	0
31.8047619	22.41666667	0	0	0	6.022222222
5.033333333	6.5	0	0	0	0
12.61428571	8.983333333	0	0	0	0
58.84444444	19.81666667	0	0	0	5.926984127

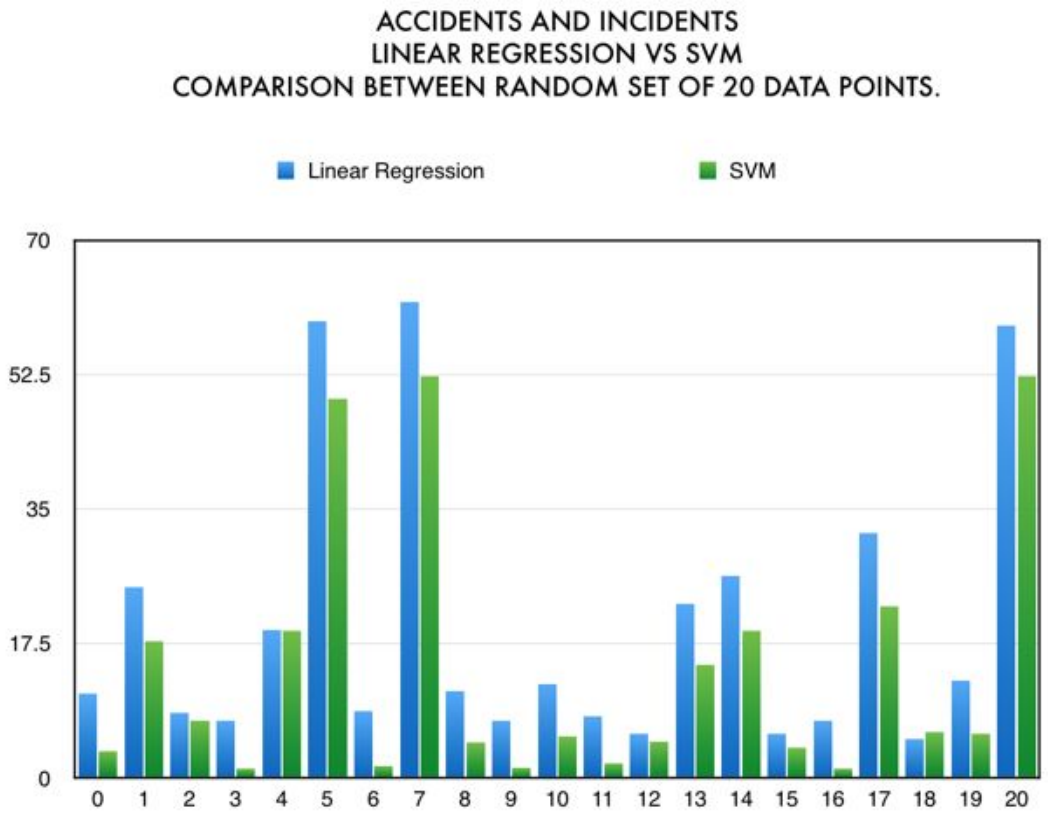
The comparison of the predictions produced by the algorithms can be seen in the below graph. These predictions are from a random set of 100 samples in the data :-



The above plot shows a comparative percentage of the predicted results by the linear regression vs the SVM regression.

Further, We also fine-grained our observations to visualize the data more intuitively. Below you will find a side-by-side comparison of the predicted number of event_subtype “traffic conditions” and “accidents” over the period of the geobox.





INTERESTING RESULTS AND OBSERVATIONS :-

a) We found that the prediction done using the improved algorithm looks to be closer to the data predicted by our previous algorithms but the mean squared error produced by the methods were very close Zero. The mean squared error observed on the sample of 2014 data were as follows:

   Mean Squared Error Values :-

   Linear Regression :- 0.4763

   SVM :- .6333

b) We found the need to normalize the values after prediction, which might have introduced a percentage of error to the prediction. Since we were predicting the data for every single hour between the specified time period of the geobox, we had to club the predictions of every single hour to form a day's prediction and then club the predictions of all the days to form the predictions of the entire month.
c) With even this algorithm, Linear regression seemed to perform better than SVM. It was interesting to see that Linear Regression gave a lower Mean squared error than the SVM.
d) Since, the observed Mean Squared Error was very close to zero, it would be very interesting to compare the predicted output with the actual data.

Method -2 : Using NOAA extreme weather data to identify correlation between extreme weather and traffic incidents :-

Severe weather conditions may have various impacts on the transportation system, involving the impacts on vehicle conditions, road conditions, and driver behavior. These weather events can affect the transportation system both directly and indirectly (Federal Highway Administration Report, 1999). Especially in winter, heavy rains, snow, storms and freezing temperatures can result in a higher frequency of car crashes, and will also have higher opportunities to cause traffic congestion. On the other hand, people's reactions to severe weather conditions may also lead to increased fuel consumption, delays and number of accidents (A. T. Kashani, 2009).[1]

In this method, we have tried to correlate the number of accidents with the "extreme" weather data obtained from NOAA. We have used the weather data from 2003-2013 for District of Columbia, Maryland and Virginia. Similar to lab 11, we find overlapping location bounding boxes between the accidents dataset and the weather dataset. For each bounding box in the accidents dataset, we tried to find the corresponding weather data for the given month and tried to find a correlation between the number of accidents and the type of extreme weather (hail, thunderstorm etc.). We couldn't find any direct correlation with a very low correlation coefficient.

On further contemplation, this makes sense. Mapping weather with accident data is a very broad problem. Consider the following examples, Given a thunderstorm, an I-75 is bound to have more accidents compared to Gainesville city roads, simply because of the speed limit. Also, Archer

Road will have fewer accidents compared to I-75 due to the high traffic and low flow. A road along the coastline/a hilly terrain is prone to have more accidents than I-75 during a hurricane/thunderstorm. This implies that the accidents and weather cannot be simply correlated without considering numerous external factors like the traffic, speed limits of the roads, road damages, amount of flood water on the road etc. and a lot of other factors.

Method -3 : Using NOAA precipitation data to identify correlation between rainfall and traffic incidents :-

In this method we are predicting the accidents and incidents based on the precipitation in the specified month/year. For this, we took a very similar approach to the one specified in the Lab 11. We took the NOAA weather data for Washington DC and its surrounding counties like in Maryland, Virginia. The data provided by NOAA contained data from 2003 up to 2014 only. So, to find the correlation between the traffic data and the weather data we ran regression analysis to predict the amount of precipitation for the year 2015. In this approach, for all the ~3,00,000 geoboxes, we first predicted the precipitation of the geobox alongside the number of accidents happening in the area.

The algorithm for our approach is highlighted in the below screenshot :-

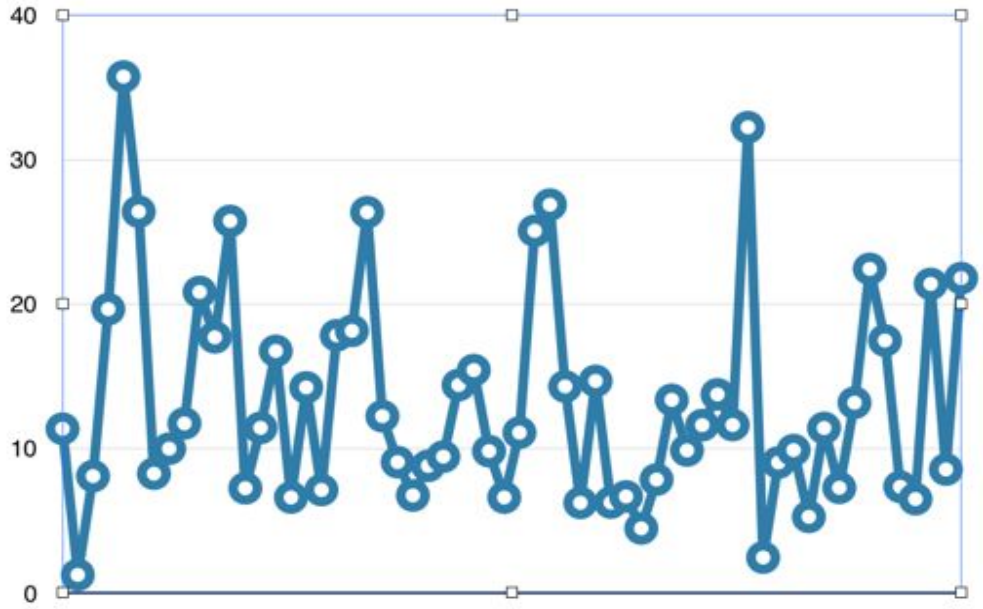
```
def create_precipitation_Dict(row):
    columns1 = ['month1', 'year1']
    columns2 = ['month2', 'year2']

    df1 = pd.DataFrame(row[columns1])
    df1 = df1.transpose()
    df2 = pd.DataFrame(row[columns2])
    df2 = df2.transpose()
    df2.columns = columns1
    x_predict = df1.append(df2)

    startMonth = int(row["start"].month)
    endMonth = int(row["end"].month)
    prec_data =  df_precipitation[((df_precipitation["month"] == startMonth) | (df_precipitation["month"] == endMonth))
    prec_data_group = prec_data.groupby(["year","month"])["HPCP"].sum().reset_index()
    data_test = prec_data_group[prec_data_group['year'] == 2014]
    data_train = prec_data_group[prec_data_group['year'] != 2014]
    columns = ['month', 'year']
    x_train, y_train, x_test, y_test = data_train[columns] ,data_train['HPCP'], data_test[columns], data_test['HPCP']
    y_ = train_regression_model(x_train, y_train, x_test, y_test, x_predict)
    final_prec_result.append(np.mean(y_,axis=0))

    global count
    global start
    count = count + 1
    if count%10000 == 0:
        end = time.time()
        print (str(count) + '\t' + str(end - start))
        start = time.time()

    return len(prec_data)

trial_data = df_prediction_trial
trial_data["precipitation_count"] = trial_data.apply(lambda row:create_precipitation_Dict(row), axis= 1)
print('svm reg mse : ' + str(math.sqrt(np.mean(total_error_svr,axis=0))))
```

On a raw level, we got the following snapshot of data.

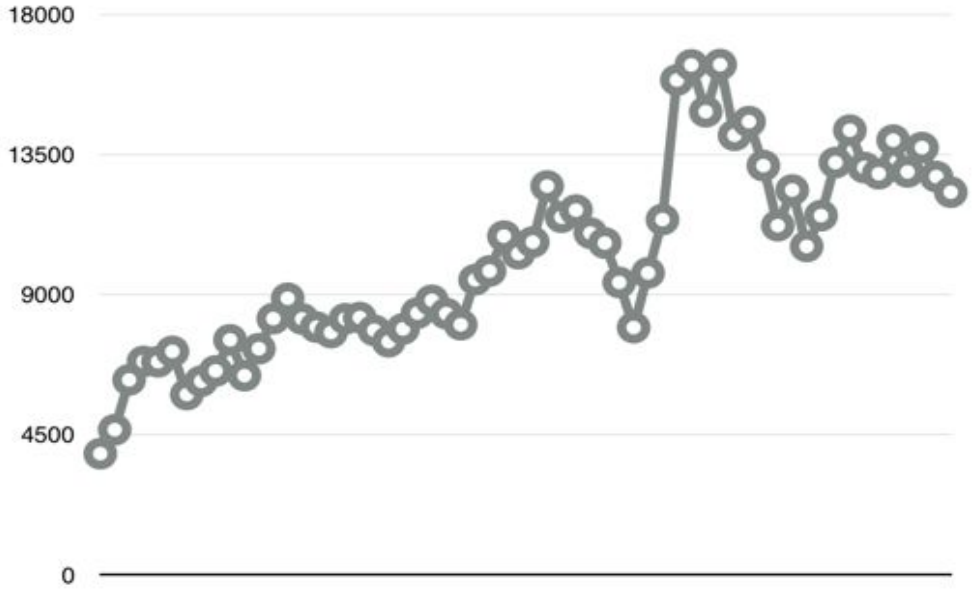
	A	B	C	D
1	year	month	HPCP	Accidents
2	2009	1	11.35	3879
3	2009	2	1.23	4649
4	2009	3	8.07	6252
5	2009	4	19.63	6844
6	2009	5	35.74	6835
7	2009	6	26.39	7166
8	2009	7	8.24	5780
9	2009	8	9.97	6209
10	2009	9	11.72	6539
11	2009	10	20.82	7538
12	2009	11	17.68	6386
13	2009	12	25.74	7253
14	2010	1	7.23	8221
15	2010	2	11.41	8874
16	2010	3	16.74	8210
17	2010	4	6.59	7947
18	2010	5	14.21	7774
19	2010	6	7.11	8214
20	2010	7	17.82	8254
21	2010	8	18.13	7844
22	2010	9	26.35	7477
23	2010	10	12.22	7889
24	2010	11	9.04	8398
25	2010	12	6.73	8808
26	2011	1	8.78	8380
27	2011	2	9.42	8024
28	2011	3	14.38	9470
29	2011	4	15.41	9751
30	2011	5	9.8	10871
31	2011	6	6.59	10308
32	2011	7	11.08	10682
33	2011	8	25.06	12482
34	2011	9	26.87	11481
35	2011	10	14.28	11709

The HPCP here represents the monthly grouped precipitation of the city and its surrounding counties. The accidents here refer to the accidents which have occurred during this period in the city. This way we first tried to correlate on a broader scale, as to how the monthly precipitation affected the number of accidents. To understand if there is any correlation, we charted the below mentioned graphs.

GRAPH REPRESENTING PRECIPITATION(HPCP) FROM 2009 TO 2013



GRAPH REPRESENTING ACCIDENTS ON ROAD FROM 2009 TO 2013



Through these visualizations we found the generalization that there is no immediate correlation between the number of accidents occurring during a period of month. For e.g, it is evident from general intuition and many papers that with bad weather, the number of accidents tend to increase. **As, we try to verify this fact from the NOAA data, we can see for October of 2011 the precipitation was 14.28 and the number of accidents were 11709 . Also, the amount of precipitation in September of 2011 is 26.87 and the number of accidents in the same month is 11481. If we follow the proposed theory, we should have seen a positive trend in the number of accidents i.e they should have increased with the increase in amount of precipitation. But, this is not the case.**

So, we went though a couple of Journal Papers to verify this claim. We wanted to verify that the null hypothesis which we had rejected was indeed false or not. After reading through the content of papers like "**WEATHER IMPACT ON ROAD ACCIDENT SEVERITY IN MARYLAND**" and "**The influence of weather on road safety**" we understood that precipitation as the only factor cannot be used to determine the source of accidents. There are several other contributing factors like weather factors, relative humidity, average wind speed etc. which significantly affect the outcome of our predictions.

Moreover, after visualizing the precipitation data for all the ~3,00,000 geoboxes which we had found through our code, we got a correlation factor of .01 which is very weak to suggest any significant result. We also have the .tsv file available for all the geoboxes available for review.

Therefore, In order to correctly correlate the data between precipitation and number of accidents, there is a need for more detailed weather data which is not directly available on NOAA and other weather monitoring websites. Further, Grouping and Cleaning the data is also very necessary before making any decisions/predictions about such sensitive areas.

CHALLENGES FACED:-

1. As we started working with the data we noticed that there were a lot of ambiguities in the data. This went out for the traffic data as well. To illustrate this, please notice the screenshot below:-

	A	B	C	D
1		year	month	count
2	30	2006	3	64
3	31	2006	4	323
4	32	2006	5	290
5	33	2006	6	8
6	34	2006	7	106
7	35	2006	8	17
8	36	2006	9	709
9	37	2006	10	1183
10	38	2006	11	1020
11	39	2006	12	594
12	40	2007	1	921
13	41	2007	2	1097
14	42	2007	3	827
15	43	2007	4	780
16	44	2007	5	926
17	45	2007	6	727
18	46	2007	7	785
19	47	2007	8	811
20	48	2007	9	847
21	49	2007	10	1075
22	50	2007	11	934
23	51	2007	12	783
24	52	2008	1	541
25	53	2008	2	528
26	54	2008	3	236
27	55	2008	4	347
28	56	2008	5	565
29	57	2008	6	651
30	58	2008	7	989
31	59	2008	8	1243
32	60	2008	9	1924
33	61	2008	10	3245
34	62	2008	11	2771
35	63	2008	12	3566

The data on the left shows the year, month and the count of the number of accidents during that period

Through the data it is evident that the count is gradually increasing along the years.(see 2008 onwards) The difference between the counts is huge and this is clearly misrepresented data. The possibility is that initially the recorded accidents were very less in years 2003-2009. Since 2009, a proper history is maintained for all the accidents. This could justify the dramatic rise in the data over the years.

However, this skews the output of the machine learning algorithms which we use and therefore makes the predictions unreliable.

2. The Weather data was not readily available. For this assignment, it was imperative that a lot of data is available about different weather conditions within the city i.e Washington DC. But the most we got were bare 20,000 data points for last 10 years . This was too less to find a proper correlation. Moreover, We only had data about severe weather and precipitation. Several other very important parameter like relative humidity, average wind speed etc. were totally missing.

3. Since the scope of this one single iteration i.e Lab12 can be huge, we would have been able to perform a much better analysis, had there been more time and computational resources available.

CONCLUSION : -

Though our analysis of 3 different methods, we identified that there are several techniques through which we can increase the accuracy of our predicted results. As identified in the previous labs, one of the approaches which could have proved good would have been an ensemble method.

We could probably use an ensemble method that uses both Linear Regression and one of the Kernel Models/Neural Net model (We would prefer Kernel Models over Neural Networks as the time taken by the kernel models is much much lower and not much difference in the performance) to get the final prediction.

$$Final\ Model = c1 * LinearRegression + c2 * (Kernelized\ Ridge\ Regression\ or\ SVR)$$

where c1 and c2 are the weights of the models.

Weather also has a huge role to play in road accidents. Average wind speed has a significant effect on accident severity on highways. When wind speed is high, vehicles can be hard to control especially for the new drivers, and the visibility usually decreases as well. Lower visibility can cause difficulty on driving, but it does not necessarily have impact on accident severity, because drivers will be more cautious and careful. The accumulation of precipitation, no matter rain or snow, has a significant effect in accident severity, which is intuitive and can be easily understood.

Thus with better data and insights into the ways weather correlates to the traffic conditions, we can provide valuable information to the city planner for making it safe for the drivers. Further,

With the understanding of these tools and techniques we are in a better shape to predict and warn about any disasters for which precautions can be taken.

CITATIONS

1. WEATHER IMPACT ON ROAD ACCIDENT SEVERITY IN MARYLAND http://drum.lib.umd.edu/bitstream/handle/1903/14263/Liu_umd_0117N_14019.pdf.;jsessionid=AFB8587983A2EC87A6B6C6C99476E9B4?sequence=1
2. The influence of weather on road safety https://www.swov.nl/rapport/Factsheets/UK/FS_Influence_of_weather.pdf
3. The influence of rainfall on road accidents in urban areas: A weather radar approach http://www.sciencedirect.com/science/article/pii/S2214367X13000069
4. How Do Weather Events Impact Roads? http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm