Traffic Congestion Prediction - Analysis & Prediction model

Date: 15-Dec-2019

1. Without building any model, what are the contributing factors for traffic slowness?

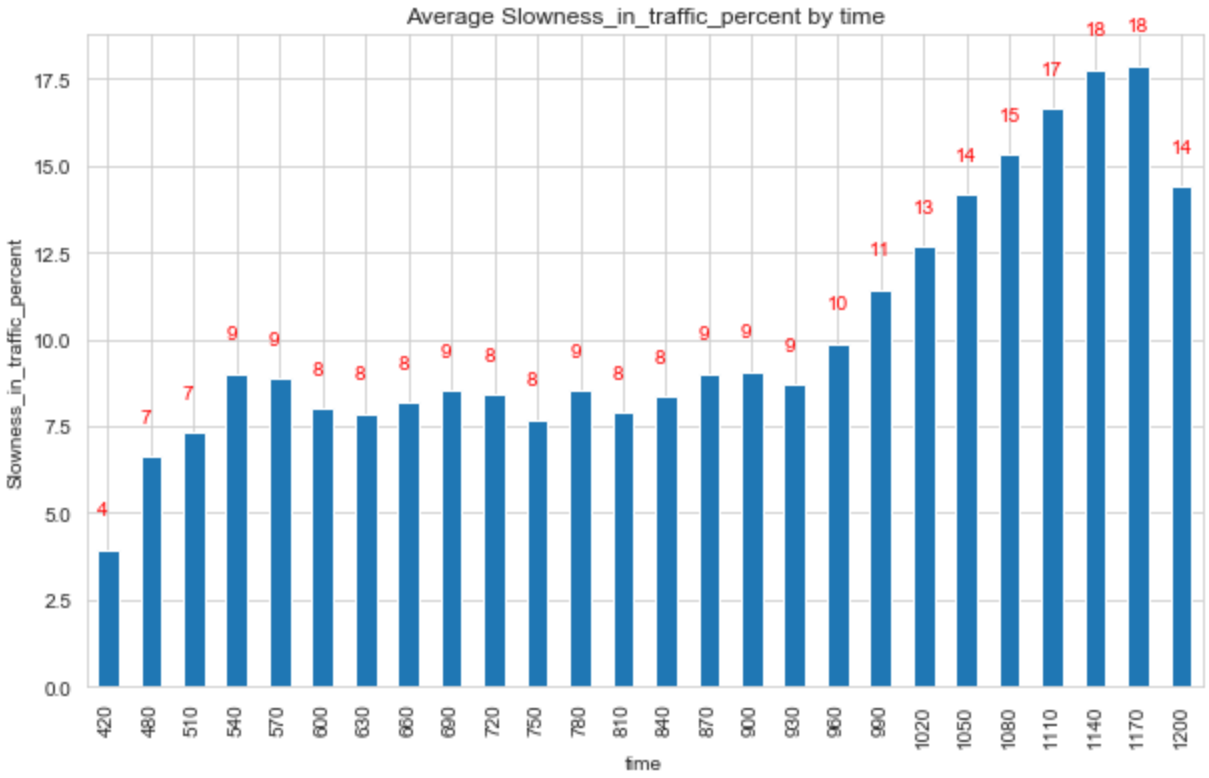
We did exploratory data analysis and have identified 3 categories of factors that influence traffic slowness:

1. **Major factors:** There are 2 of them: time and day
2. **Medium factors:** There are 2 of them: Lack of electricity and point of flooding
3. **Low occurrence factors:** These factors don't play a significant role since their occurrence is very low (Fire_vehicles, Manifestations, Defect_in_the_network_of_trolleybuses, Tree_on_the_road, Semaphore_off, Incident_involving_dangerous_freight, Occurrence_involving_freight, Fire)

Major factors:

1. **Time**

   Time is easily the single most important factor that determines traffic slowness. We have converted time given in format hh:mm into minutes format (e.g. 7:30 is converted to 450). Below is a plot of average slowness seen at any given point of time (On X axis we have time is in terms of mins). In early morning traffic slowness is less and it starts rising steadily and reaches a local peak at 540 mins (9am). From 540 mins to 930 mins (3.30pm) it keeps fluctuating between 8 and 9%. After 930 mins it again gradually rises until 1170 mins (7.30pm). After that it again comes down.



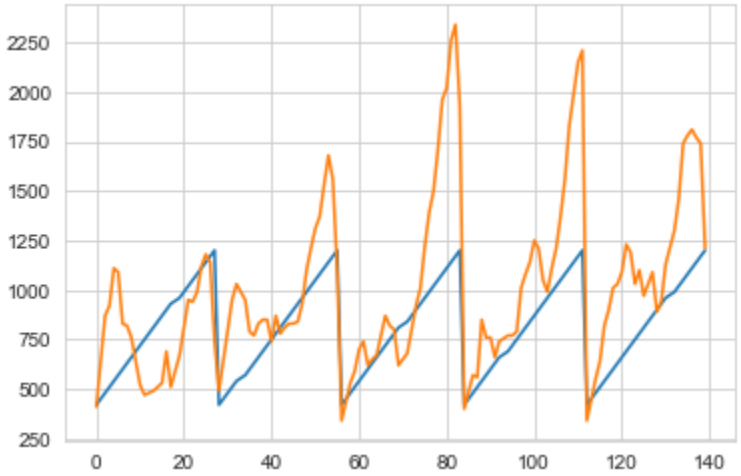
2. Day of week

This was not explicitly provided in the data file. However looking at data we could infer that the records are in sequence of date. After removing duplicates there were 130 records left (there was one duplicate that we did not remove since we would have fallen short of one record in that case). For every half an hour we ended up with 5 observations for each days. Hence we marked first 26 records with day as Monday, next 26 with day as Tuesday and so on.

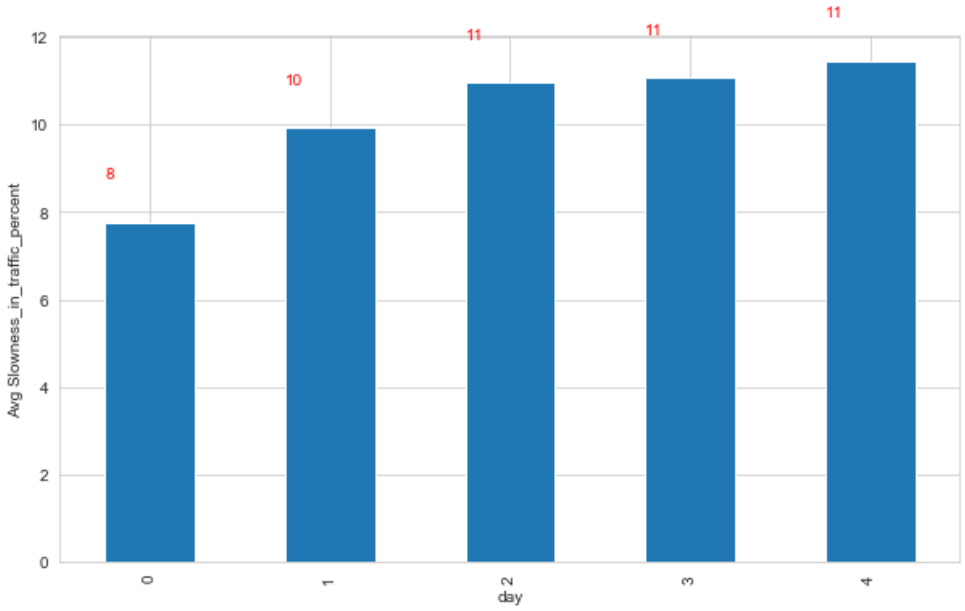
Below is plot that helped us infer that records are ordered by day. We have record number on x axis and on Y axis we have:

i) Time in minutes (blue color)

ii) Scaled version of slowness (orange color)



After generating the day feature, we see a strong correlation between day and slowness. On Mondays traffic is less and then is gradually increases as we move into the week.

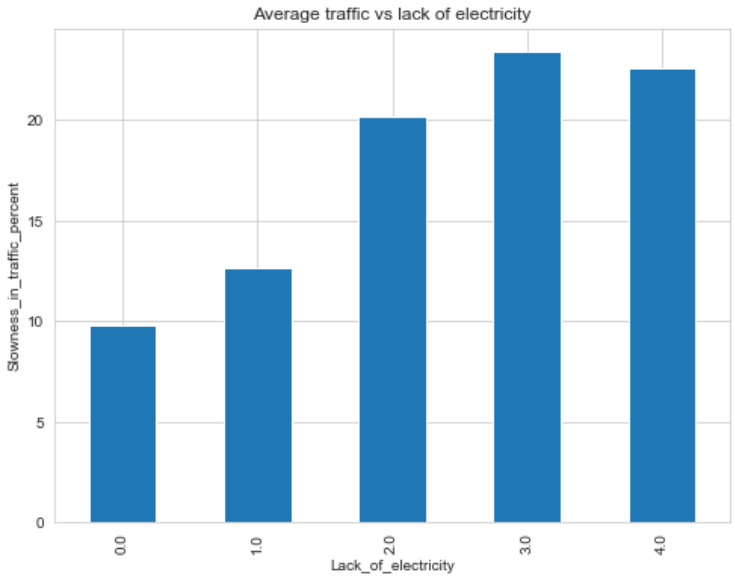


Medium factors:

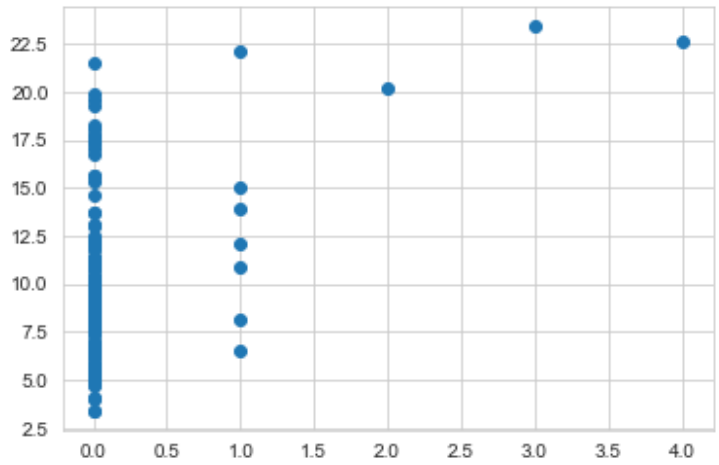
1. **Lack of electricity**

Lack of electricity also plays an important role in slowness. We see that slowness is higher whenever we have non zero number in lack of electricity and it rises as this number goes up. We assume that these numbers most likely mean number of areas in which electricity is not there.

Plot of electricity vs avg slowness:

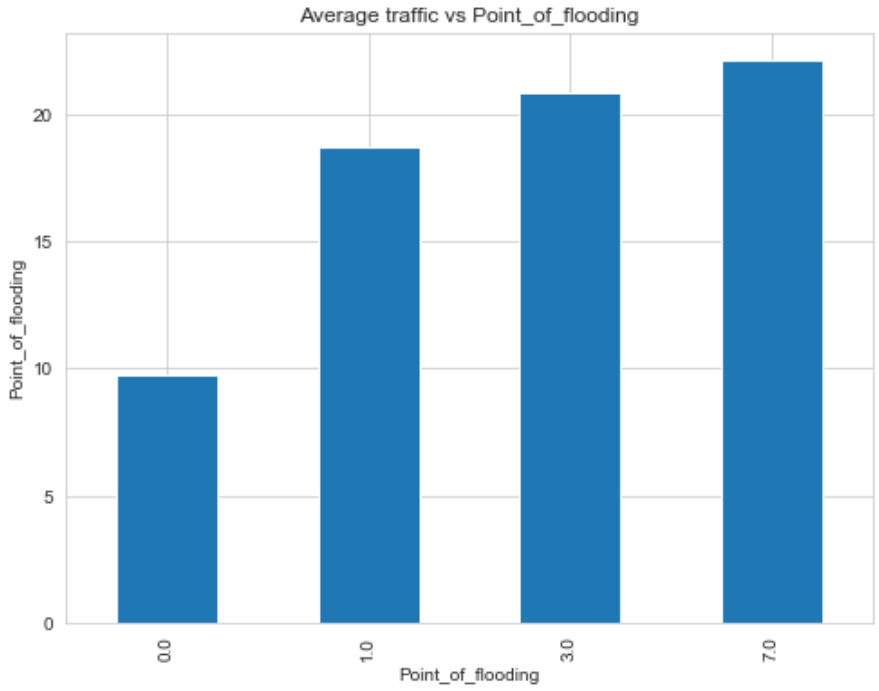


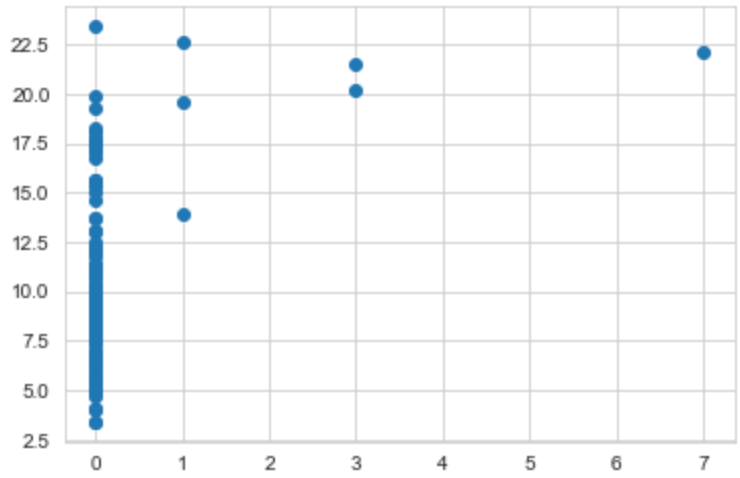
Scatter plot of lack of electricity and slowness.



2. Point of flooding

Point of flooding also has influence on traffic slowness which is very similar to lack of electricity. Following are average slowness and scatter plot of point of flooding and slowness:





Low occurrence factors

These are important factors, however as per given data set their occurrence is low hence we don't expect them to play a significant role in prediction of slowness. These factors are:
Fire_vehicles, Manifestations, Defect_in_the_network_of_trolleybuses, Tree_on_the_road, Semaphore_off, Incident_involving_dangerous_freight, Occurrence_involving_freight, Fire

2. Are you able to confirm the above findings using any two modelling techniques? Give appropriate explanation for the same.

Brief about models we have built:

Following are the 2 models we have built to predict traffic slowness.

1. **Linear regression (treating problem as regression type of problem)**
   Since slowness percentage is numeric, we have first modelled this problem as a regression type of problem and built a Linear Regression predictor.
   We observed that since data is less, model accuracy was fluctuating based on set of records we were using for training vs testing. To overcome this problem we have used k-fold validation using 6 folds.
   We kept aside 30% of data for doing final evaluation post k-fold validation.
   We could achieve accuracy of 85% on test data while accuracy for training data was 87%. We have used RFE to recursively eliminate features while building this mode. We could successfully eliminate 13 features without negatively impacting accuracy of the model.

2. **Randomforest classifier (treating problem as classification type of problem)**
   This is the second of the two models we have built to solve this problem. We decided to convert this problem into classification type because of following reasons:
   1. Number of records are less.
   2. It would be easier to predict slowness class instead of accurately predicting the exact slowness percentage.

   Initially we saw that RandomForest classifier was performing better then Linear regression. However after extensive feature transformation was done to fine tune data for Linear regression, LR started performing better. Hence we focused most of our energy on fine tuning Linear regression and implemented RFE with cross validation for Linear regression only. Eventually we could get this model also to predict correct class of slowness with 87% accuracy.

   Binning of target is done in custom way to create 4 classes of traffic congestion:

   Class # 1: <=7% → No or minimal traffic
   Class # 2: <=12% → Medium traffic
   Class # 3: <=20% → High traffic
   Class # 4: > 20 → Very high traffic

Although both models have similar accuracies, Linear regression is recommended, since it predicts exact slowness % (as compared to broader class of slowness as predicted by classifier) which can be used to make fine-tuned decisions. However if slowness class prediction is sufficient, then the classification model can also be used.

Please see: We also tried building a 3rd model using RandomForest Regressor to compare results between Linear regreesion and RadomForest regressor. Due to lack of time we could do in depth analysis of this model

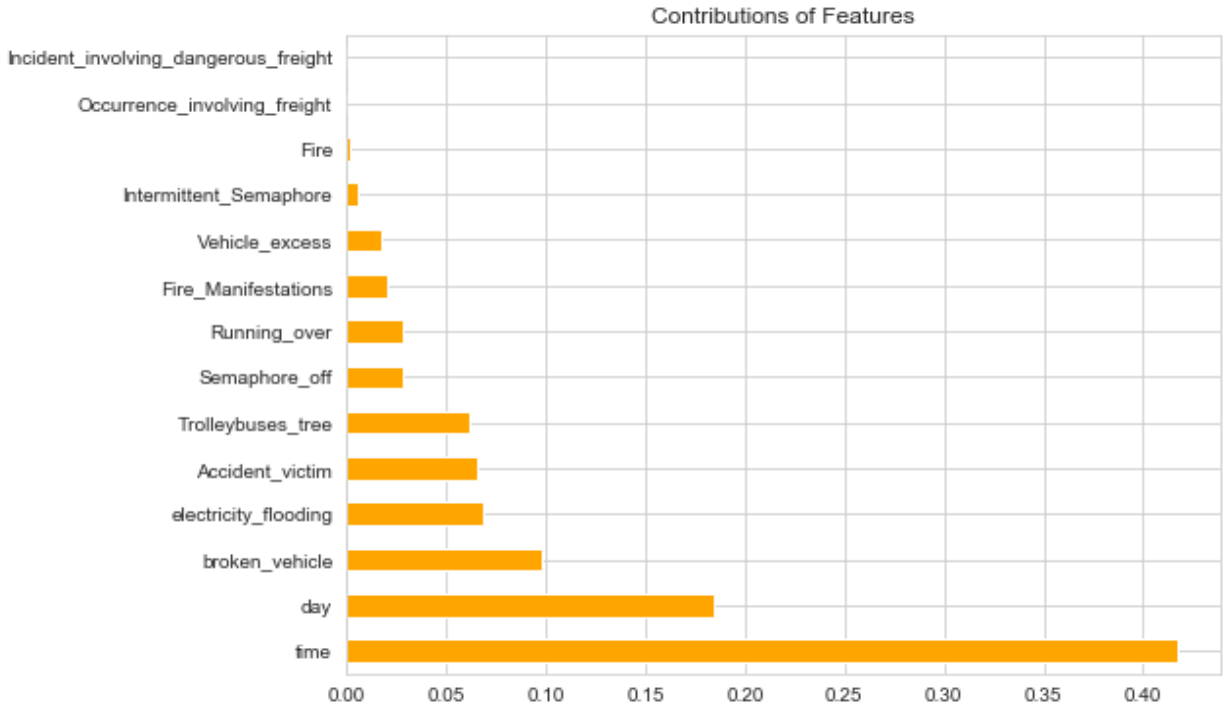
Answer to the question on whether we could confirm our findings:

Yes, we are able to confirm above findings using the modeling techniques. Linear regression model with RFE has given us following set of most optimal features:

1. Time (and its non linear forms)
2. Day
3. Lack of electricity and Flooding point (we had combined these into one feature)
4. Intermittent_Semaphore

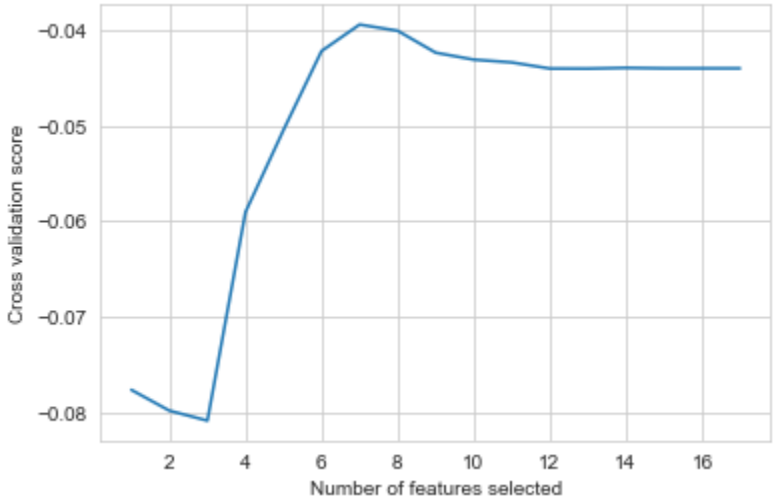
We did not expect Intermittent_Semaphore to feature in this list, however LR model has picked it.

RandomForest classifier (we converted the problem into classification problem by binning slowness % into 4 classes) has given us following ranking of important features. Top 2 factors match with what we expected.



3. Using Recursive Feature Elimination technique, what are the contributing factors for traffic slowness?

Using Recursive Feature elimination technique total 7 features were selected. Below is plot of grid score v/s number of features selected. We observe that score was pretty low with 1,2 and 3 features and then it rises exponentially with 4,5 and 6 features before maxing out at 7 features. Beyond 7 features we don't see much improvement in accuracy. So we can safely eliminate remaining 13 features without impacting model accuracy. This improves model performance as we are able to bring down number of features by 13.



Following are the 7 features selected along with their coefficients:

Feature	Coeff
time_2	14.329800
time_3	-14.197912
Intermittent_Semaphore	-0.221617
day	0.318823
electricity_flooding	0.638682
sin_time	2.174323
cos_time	4.503101

4 of these features are derived time features (time_2, time_3, sin_time and cos_time). They are having the highest influence on the model.
Next we have day and combined feature of electricity_flooding having next highest influence.
There is further scope to improve this model by increasing influence of day in the model. This will need further transformation of day feature possibly in combination with time and using time series.
Positive coefficient indicates that the corresponding feature has a positive influence on the slowness %. Magnitude of coefficient indicates the magnitude of influence feature has on slowness.

Hence we could eliminate following features without significantly impacting the model accuracy. This will lead to a better performance for this model for large volume of data.

1. Immobilized_bus
2. Broken_Truck
3. Vehicle_excess
4. Accident_victim
5. Running_over
6. Fire_vehicles
7. Occurrence_involving_freight
8. Incident_involving_dangerous_freight
9. Fire
10. Manifestations
11. Defect_in_the_network_of_trolleybuses
12. Tree_on_the_road
13. Semaphore_off

Hence we conclude that based on RFE following are the contributing factors for traffic slowness:

1. Time of day
2. Day of week
3. Lack of electricity
4. Flooding points
5. Intermittent semaphore

4. Suggestions on how to improve delivery schedule?

1. **Schedule maximum deliveries before 9am as much as possible**
   We observe that traffic is lighter before 9am, so we should try to do maximum possible deliveries before 9am. May be start shift of delivery boys early in morning, so that they can optimize on this window of light traffic.
2. **Schedule maximum deliveries earlier in the week as much as possible**
   We observe that as week progresses, slowness in traffic increases. Roaster should be prepared in such a way that maximum delivery boys are available in earlier part of the week.
3. **Use this model to predict heavy and very heavy traffic slowness**
   Using this model, we can accurately predict traffic slowness which can be used to send instructions to delivery boys so that their time is optimally utilized instead of getting stuck in the traffic.

5. Assumptions:

1. We assume that there are some errors in the file which had to be corrected before we could import the file using arff loader. Following are the changes done programmatically in the file:
   i) Replaced STRING by INTEGER as we saw that the two features which were marked as STRING had integer values in them
   ii) One feature had small case f, we converted that to F
   iii) Value of NO for Fire_vehicles feature in file row number 34 was swapped. We moved it to correct position. We initially thought of deleting this record but later realized that this is one of the 5 records for 12:30 time and hence we cannot delete it.
   iv) We changed empty values with '?' so that they correctly get represented as nan values.
2. We are assuming that the file has observations recorded by day i.e. first 26 records belong to Monday, next 26 belong to Tuesday and so on (after removal of appropriate duplicates)