VEHICLE-TO-EVERYTHING COMMUNICATION (V2X)

Candidate: Aditya Agrawal

August 2021

1 Introduction

The network dataset contains 10 features and those are:

1. Timestep time
2. Vehicle slope
3. Vehicle lane
4. Vehicle angle
5. Vehicle type
6. Vehicle pos
7. Vehicle y coordinate
8. Vehicle x coordinate
9. Vehicle speed
10. Vehicle id

The first 100 nodes in the dataset were taken for the analysis. Using the above dataset I derived some additional datasets containing useful information:

1. Duration dataset: Contains duration of connection of each node with one another and also contains the start time and the end time of each type of connection.
2. Degree of Wifi connections: Contains degree of each node at each time interval between 0 and 600 seconds
3. Degree of IEEE connections: Contains degree of each node at each time interval between 0 and 600 seconds
4. Degree of Cellular connections: Contains degree of each node at each time interval between 0 and 600 seconds

I analysed the data of the first 100 nodes for the time duration of 3600 seconds (0 to 60 minutes).

A valid connection between two nodes is the one in which the distance between the two vehicle nodes at a particular time is less than 400 metres (i.e. they are connected via a Wifi connection or a IEEE connection). Figure 1 below depicts the plot for total number of connections in the network at a certain time frame starting from t=0 sec and goes till 3600 sec. The graph is L-shaped as within the first 500 seconds, the number of connections dropped significantly before it saturates to an almost fixed value. The number of connections are not zero at the end of the simulation time because of the fact that vehicles might be going parallely to each other.

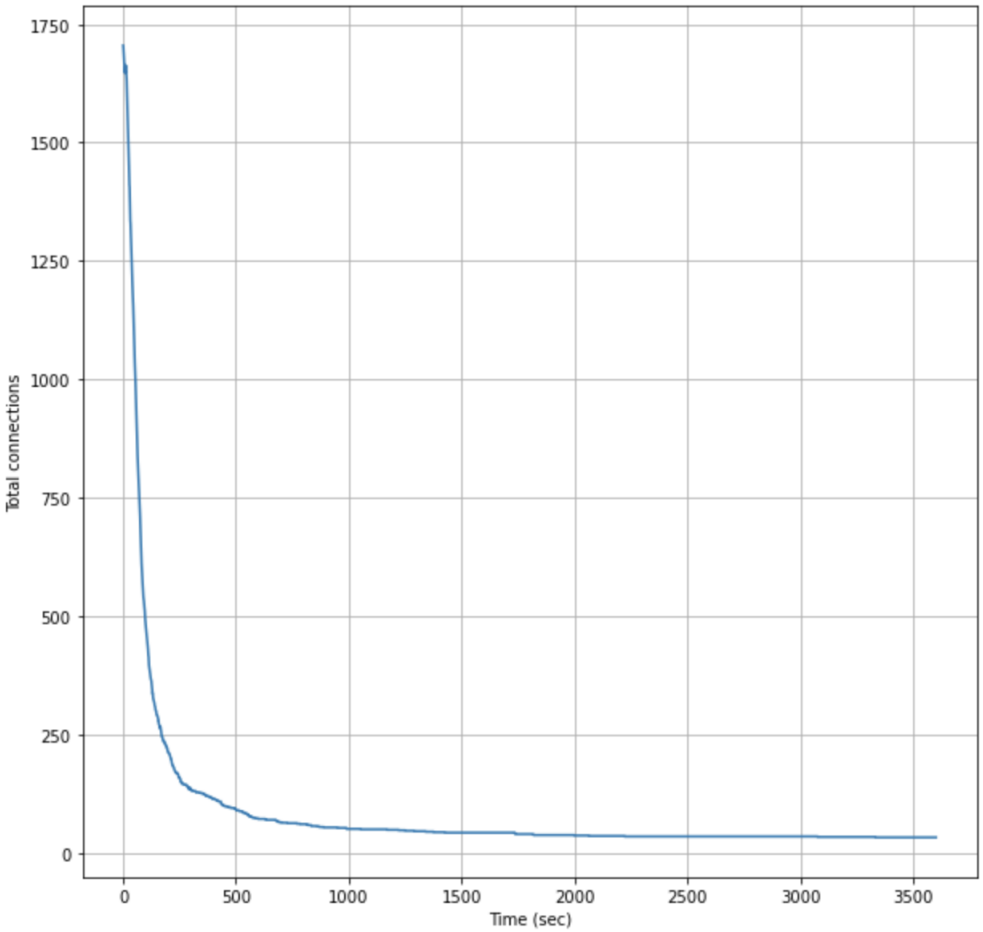
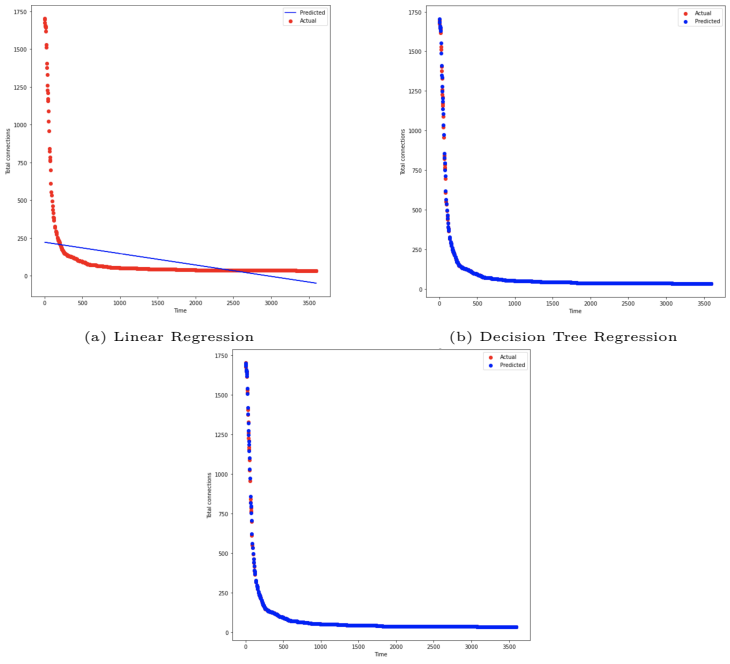


Figure 1: Total connections with time- Morning Dataset

Now, I applied some machine learning models for the above plot.
Training set size-70 %
, Test set size- 30 %



(a) Linear Regression

(b) Decision Tree Regression

(c) Random Forest Regression

Figure 2: Total connections vs Time - Test set results (Morning Dataset)

The mean squared error is formulated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 3: Formula for calculating mean squated error

Model	Mean squared error
Linear Regression	35009.487
Decision Tree Regression	6.197
Random Forest Regression	3.514

Here, we are trying to predict the total connections between the nodes in the network at a specific time. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. The error given by the linear regression is quite large as expected while in the case of random forest and decision tree it is relatively very small. Random forest regression gave the least error.

Vehicle density is the total number of vehicles in a particular area. The reference point was taken as the first node while the radius of the circular area was increased from 1 to 400 metres. The Figure 3 below depics the total number of connections in the area at a particular vehicle denisty. The graph is upward sloping as expected because the total number of connections increases if the vehicle density is increased as more number of vehicles are present in a lesser area. The graph can also be seen having a kink at one point.

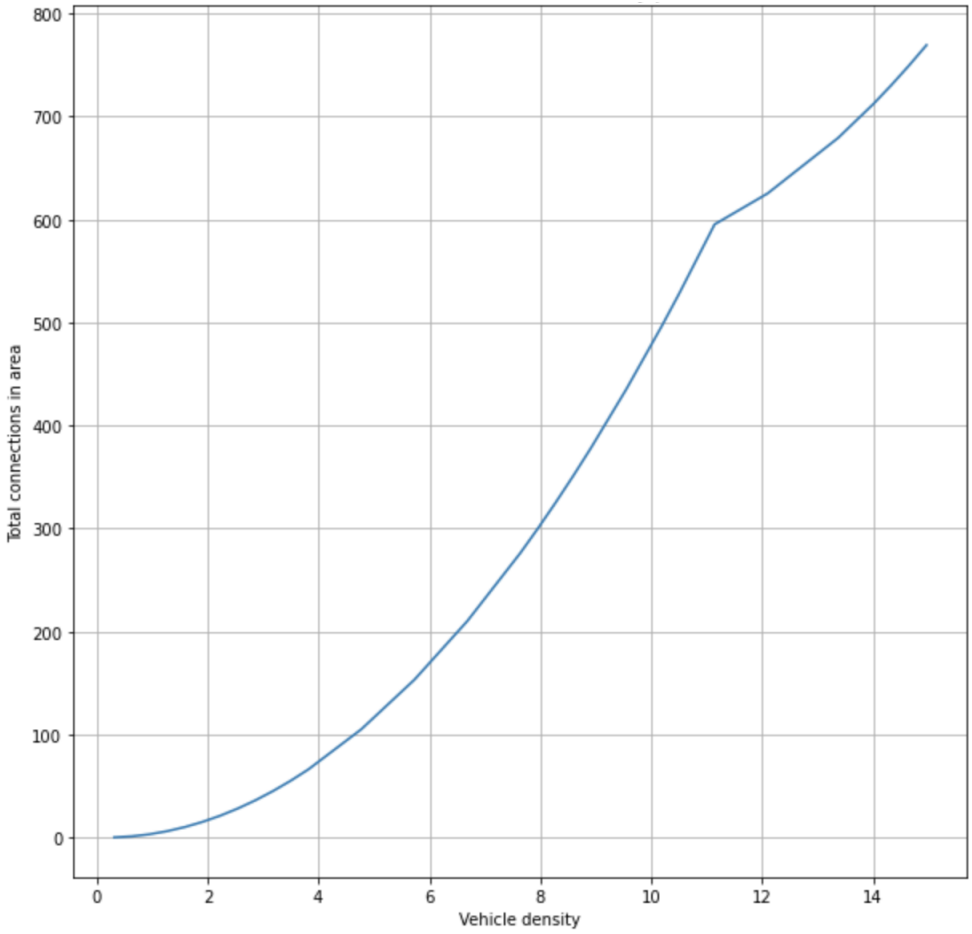
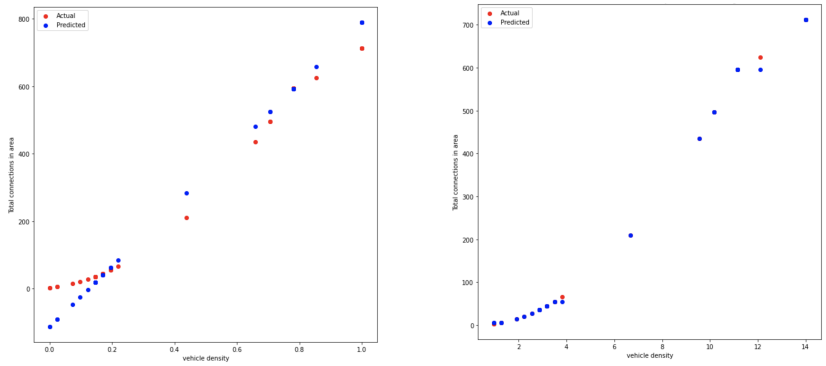


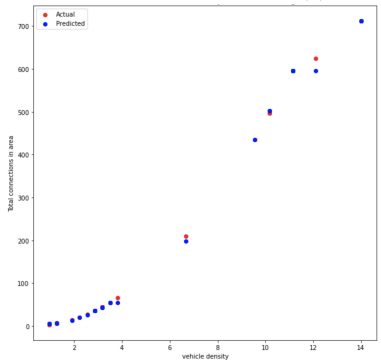
Figure 4: Total connections vs vehicle density

Now, I applied some machine learning models for the above plot.
Training set size-70 %
Test set size- 30 %



(a) Linear Regression

(b) Decision Tree Regression



(c) Random Forest Regression

Figure 5: Total connections vs Vehicle Density - Test set results (Morning Dataset)

The mean squared error is formulated as:

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 6: Formula for calculating mean squated error

Model	Mean squared error
Linear Regression	1666.515
Decision Tree Regression	17.033
Random Forest Regression	20.926

Here, we are trying to predict the total number of connections in the network at a certain vehicle density. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. Again, as expected, the linear regression gave the highest error while the Decision Tree Regression and Random Forest Regression performed well.

Connectivity is defined as the number of nodes a particular node is connected to at a certain timeframe. Here, the number of nodes which are connected to 5 or more than 5 nodes at once at a time or in other nodes, the number of nodes which has connectivity greater than or equal to 5 are plotted against time. The graph is downward sloping because of the fact that more number of nodes are moving farther away from one another than the number of nodes coming closer.

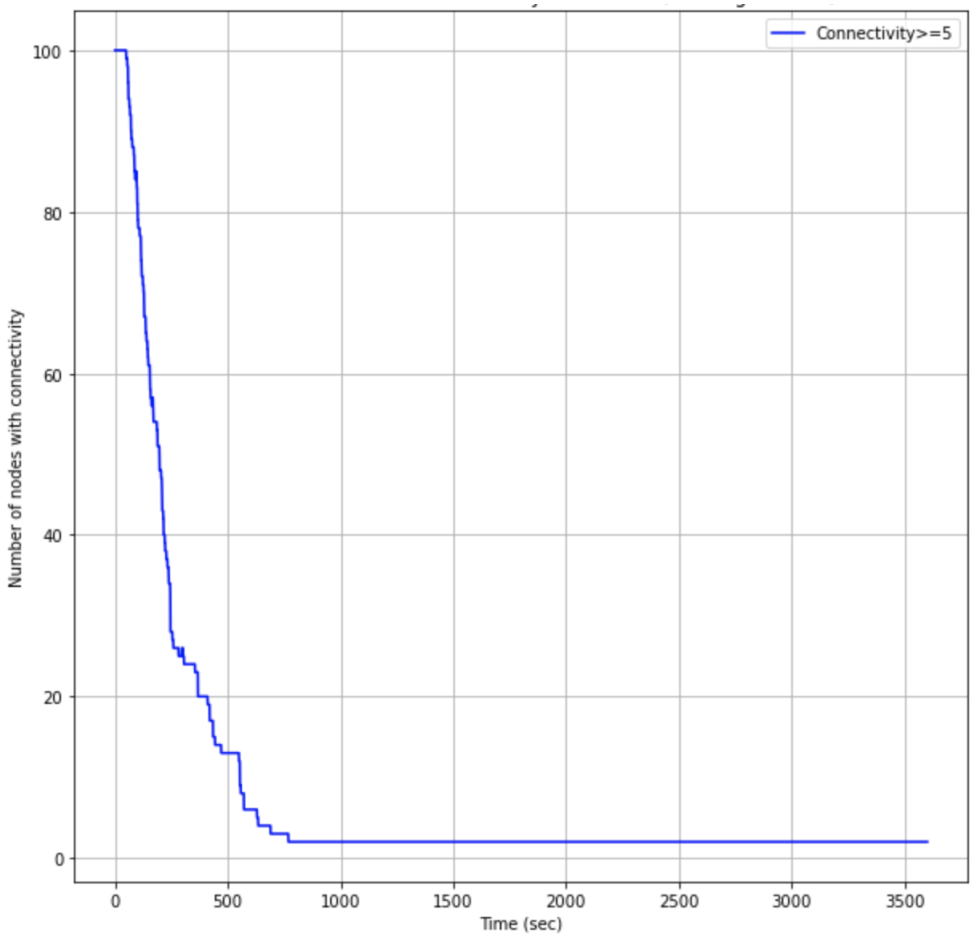
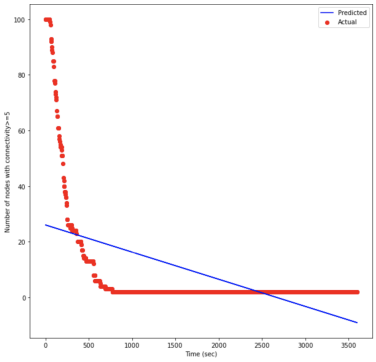
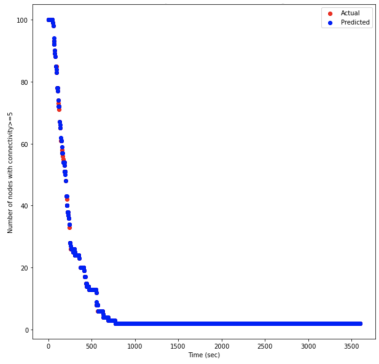


Figure 7: Number of nodes with connectivity>=5 vs Time (sec)

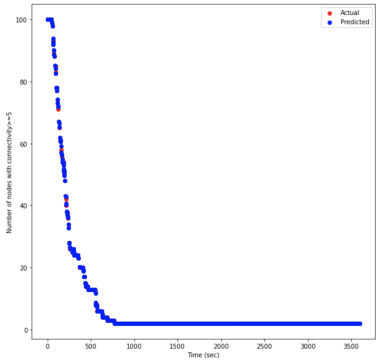
Now, I applied some machine learning models for the above plot.
Training set size-70 %
Test set size- 30 %



(a) Linear Regression



(b) Decision Tree Regression



(c) Random Forest Regression

Figure 8: Connectivity>=5 vs Time - Test set results (Morning Dataset)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 9: Formula for calculating mean squated error

Model	Mean squared error
Linear Regression	259.999
Decision Tree Regression	0.0175
Random Forest Regression	0.0166

Here we are trying to predict the total number of nodes in the network who have connectivity greater than 5 at a certain time. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. In this case, the linear regression has the highest error while the random forest and decision tree has almost the same error.

Results overview (Morning Dataset)

Plot	Model	Mean squared error
Total connections vs Time (sec)	Decision Tree Regression	6.197
	Random forest Regression	3.514
Total connections vs Vehicle Density	Decision Tree Regression	17.033
	Random forest Regression	20.926
Number of nodes with connectivity=5 vs Time (sec)	Decision Tree Regression	0.0175
	Random forest Regression	0.0166

The same is done for the evening dataset: I analysed the data of the first 100 nodes for the time duration of 3600 seconds (0 to 60 minutes). Figure 10 below depicts the plot for total number of connections in the network at a certain time frame starting from t=0 sec and goes till 3600 sec. The graph is L-shaped as within the first 500 seconds, the number of connections dropped significantly before it saturates to an almost fixed value. The number of connections are not zero at the end of the simulation time because of the fact that vehicles might be going parallely to each other.

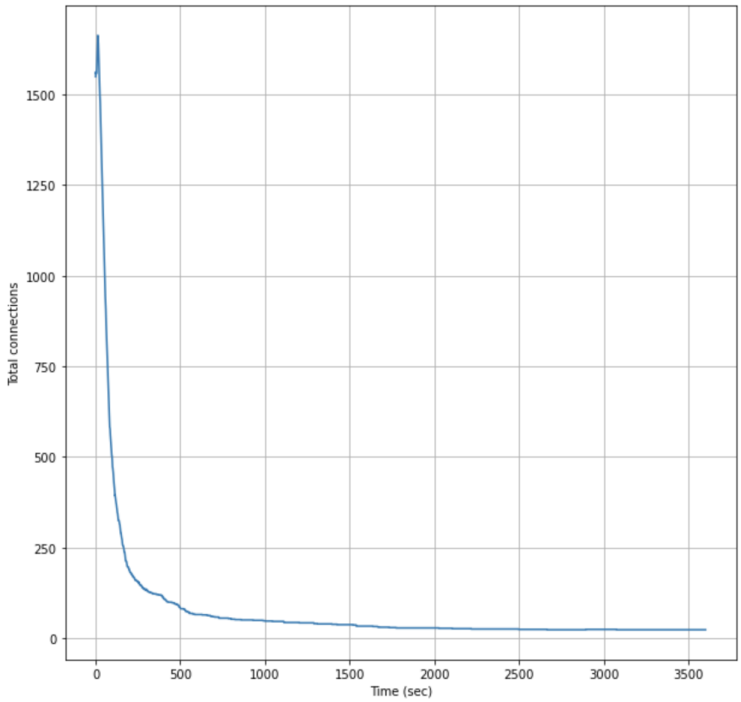
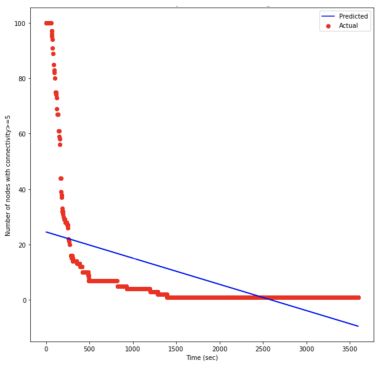
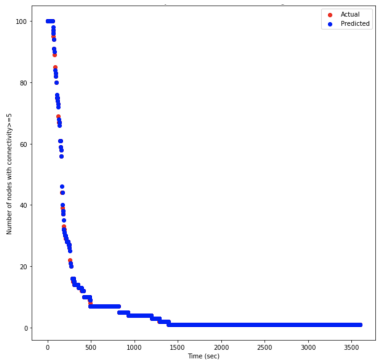


Figure 10: Total connections with time- Evening Dataset

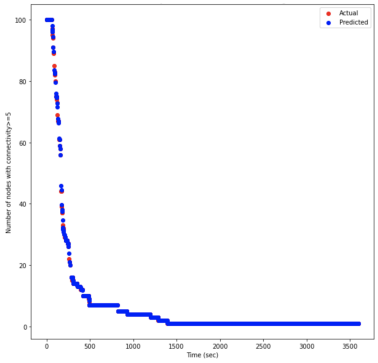
Now, I applied some machine learning models for the above plot.
Training set size-70 %
Test set size- 30 %



(a) Linear Regression



(b) Decision Tree Regression



(c) Random Forest Regression

Figure 11: Total connections vs Time - Test set results (Evening Dataset)

The mean squared error is formulated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 12: Formula for calculating mean squated error

Model	Mean squared error
Linear Regression	34303.688
Decision Tree Regression	7.567
Random Forest Regression	5.129

Here, we are trying to predict the total connections between the nodes in the network at a specific time. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. The error for the linear regression is quite large as expected while in the case of random forest and decision tree it is relatively very small. Random forest regression gave the least error.

The Figure 13 below depics the total number of connections in the area at a particular vehicle denisty. The graph is upward sloping as expected because the total number of connections increases if the vehicle density is increased as more number of vehicles are present in a lesser area.

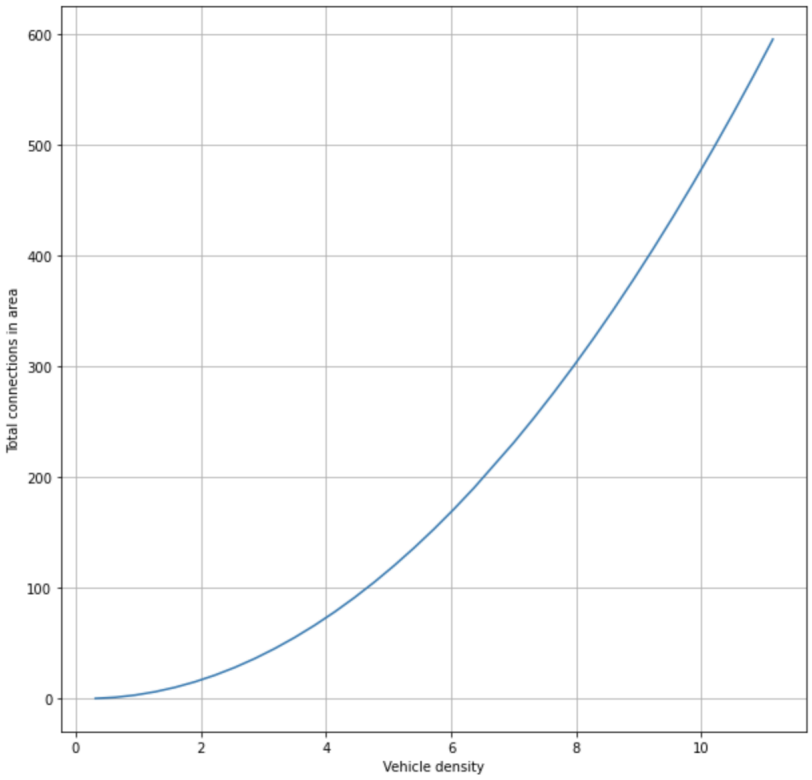
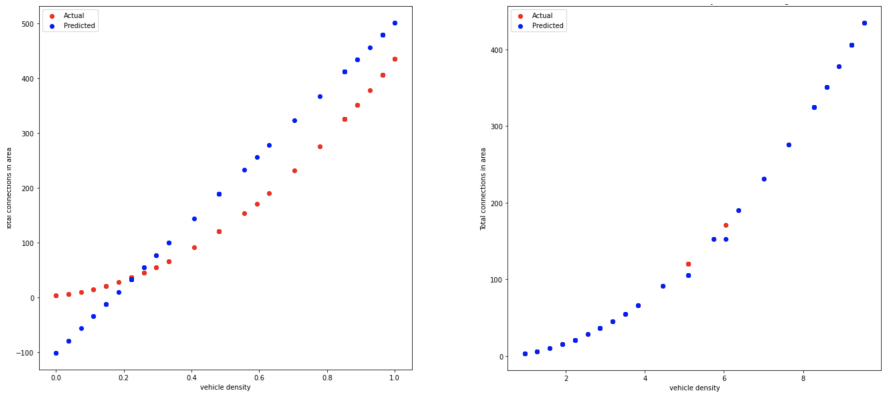


Figure 13: Total connections vs vehicle density

Now, I applied some machine learning models for the above plot.
Training set size-70 %
Test set size- 30 %



(a) Linear Regression

(b) Decision Tree Regression

(c) Random Forest Regression

Figure 14: Total connections vs Vehicle Density - Test set results (Evening Dataset)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 15: Formula for calculating mean squated error

Model	Mean squared error
Linear Regression	3740.423
Decision Tree Regression	20.065
Random Forest Regression	10.164

Here, we are trying to predict the total number of connections in the network at a certain vehicle density. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. Again, as expected, the linear regression gave the highest error while the Decision Tree and Random Forest performed well.

Connectivity is defined as the number of nodes a particular node is connected to at a certain timeframe. Here, the number of nodes which are connected to 5 or more than 5 nodes at once at a time or in other nodes, the number of nodes which has connectivity greater than or equal to 5 are plotted against time. The graph is downward sloping because of the fact that more number of nodes are moving farther away from one another than the number of nodes coming closer.

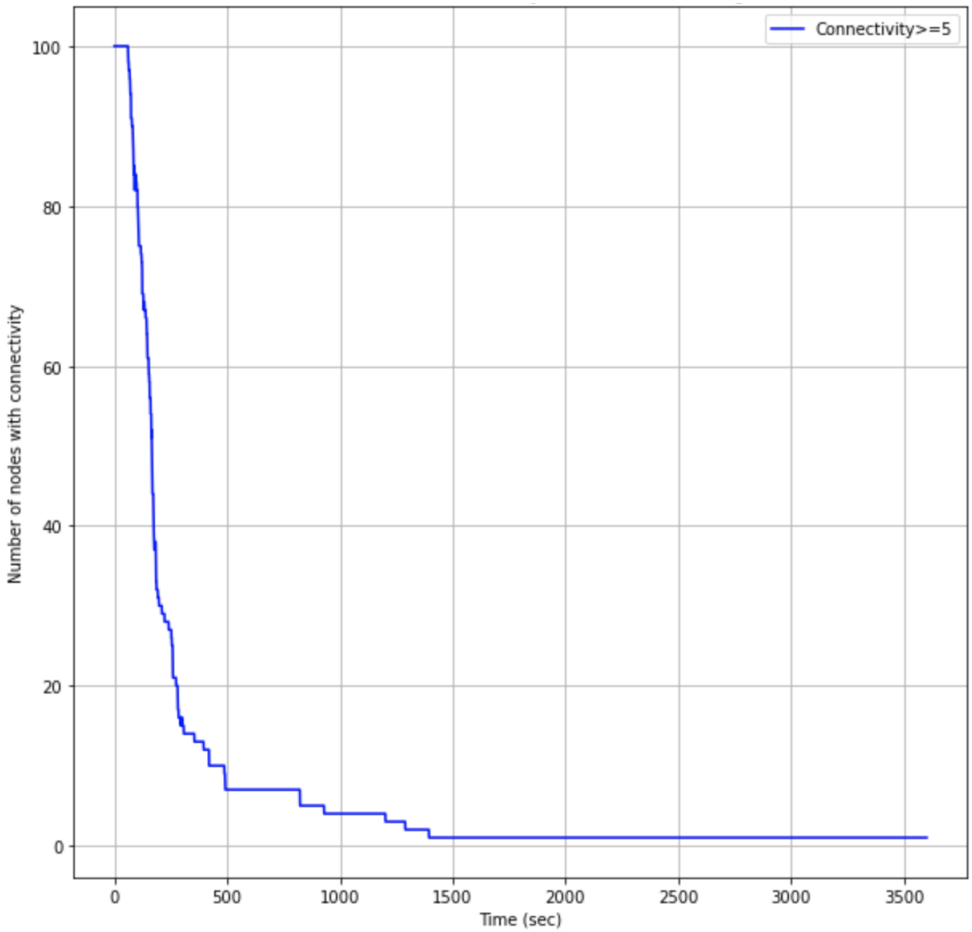
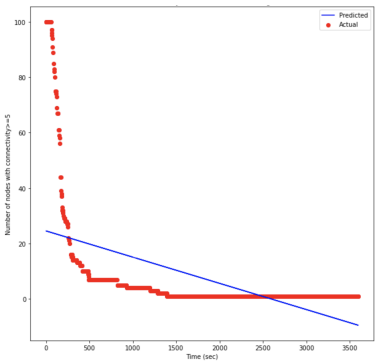
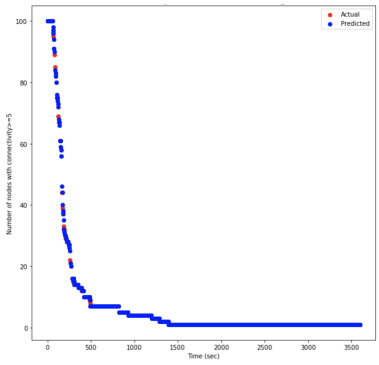


Figure 16: Number of nodes with connectivity>=5 vs Time (sec)

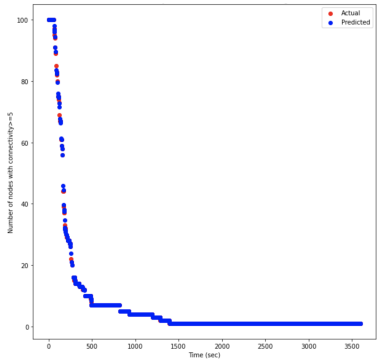
Now, I applied some machine learning models for the above plot.
Training set size-70 %
Test set size- 30 %



(a) Linear Regression



(b) Decision Tree Regression



(c) Random Forest Regression

Figure 17: Connectivity>=5 vs Time - Test set results (Evening Dataset)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 18: Formula for calculating mean squated error

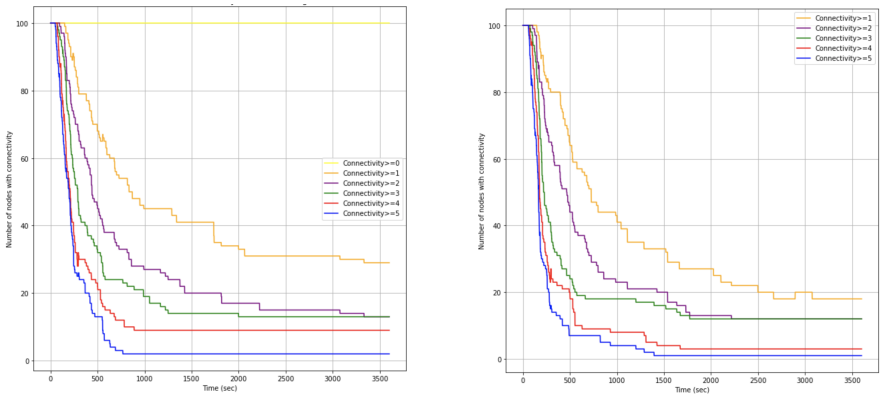
Model	Mean squared error
Linear Regression	244.047
Decision Tree Regression	0.0407
Random Forest Regression	0.028

Here we are trying to predict the total number of nodes in the network who have connectivity greater than 5 at a certain time. So, we trained the models on the training set and above are the errors in the prediction of the 30 % test data that we divided our data into. In this case, the linear regression has the highest error while the random forest and decision tree has almost the same error.

Results overview (Evening Dataset)

Plot	Model	Mean squared error
Total connections vs Time (sec)	Decision Tree Regression	7.567
	Random forest Regression	5.129
Total connections vs Vehicle Density	Decision Tree Regression	20.065
	Random forest Regression	10.164
Number of nodes with connectivity=5 vs Time (sec)	Decision Tree Regression	0.0407
	Random forest Regression	0.028

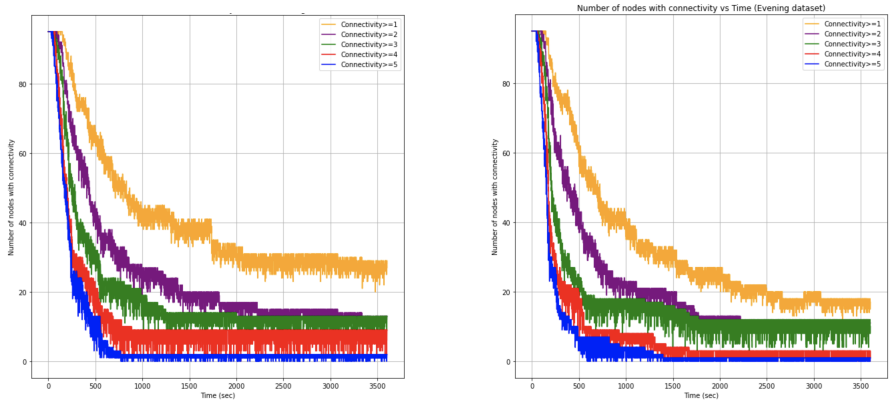
Now, we compare the plots of number of vehicluar nodes having different connectivities against the time for both morning and evening dataset. The connctivities ranges between 0 to 5. The connectivity greater than or equal to 1 signifies whether a node is connected to any other node and resembles the plot of number of connected nodes vs time. It can be seen that as the connectivity increases, the number of nodes having higher or equal connectivity than that increases which makes sense as the number of nodes having higher connectivity threshold are also included in the ones with lower connectivity limit.



(a) Morning Dataset (b) Evening Dataset

Figure 19: Number of nodes having different connectivities vs Time (Morning vs Evening)

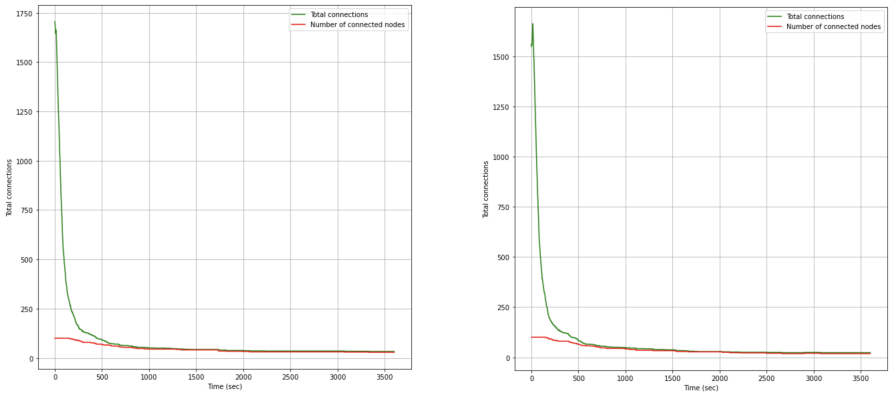
After this, we will also consider the case when 5% of the random nodes sre removed at every iteration of the time interval. Figure 20 below shows the plot for both the morning and evening dataset and it can be seen the that the removal of random nodes has introduced fluctuations in the plot.



(a) Morning Dataset (b) Evening Dataset

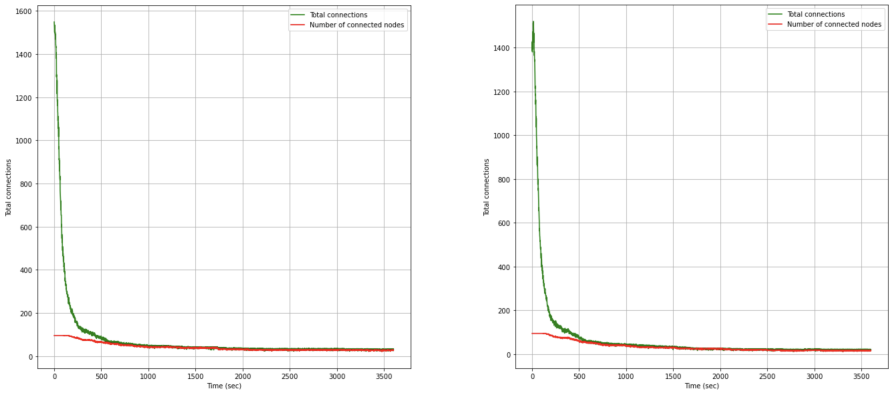
Figure 20: Number of nodes having different connectivities vs Time (Morning vs Evening)- Removing 5% random nodes

Now the total number of connections at a given time and the number of nodes connected are plotted against time for both the morning and evening dataset. Figure 21 below is the plot for the same but without removing any random nodes while the Figure 22 is the plot when removing 5% random nodes at once at avery iteration of the time interval. Again, in the Figure 22, some fluctuations are there because of the randomness.



(a) Morning Dataset (b) Evening Dataset

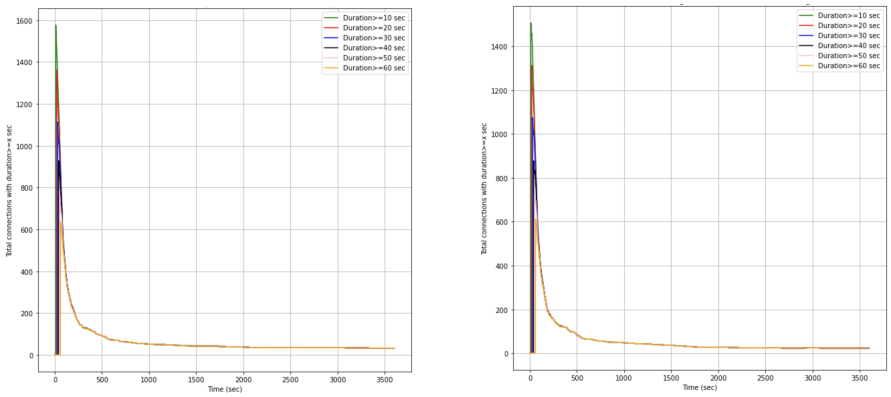
Figure 21: Total Connections with time and number of nodes connected vs Time (Morning vs Evening)



(a) Morning Dataset (b) Evening Dataset

Figure 22: Total Connections with time and number of nodes connected vs Time (Morning vs Evening)

The below plots (Figure 23) depicts the total number of connections which are active for more than a certain time (10, 20, 30, 40, 50, 60 seconds) at a given time frame. If the connection between two nodes is active for the given time duration continously without breaking, it counts towards a successful connection corresponding to the time duration. The lesser is the time duration, the higher the graph goes as there are more chances of a successful connection.

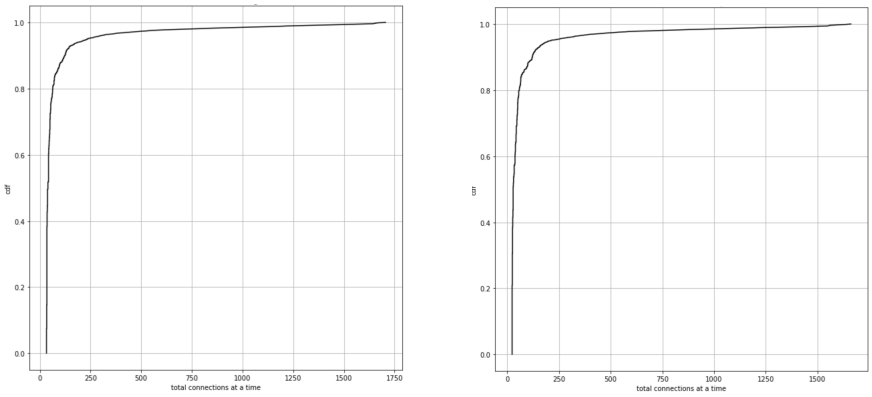


(a) Morning Dataset (b) Evening Dataset

Figure 23: Total connections having duration of more than a certain time vs Time (Morning vs Evening)

Cumulative distribution function is defined as a function whose value is the probability that a corresponding continuous random variable has a value less

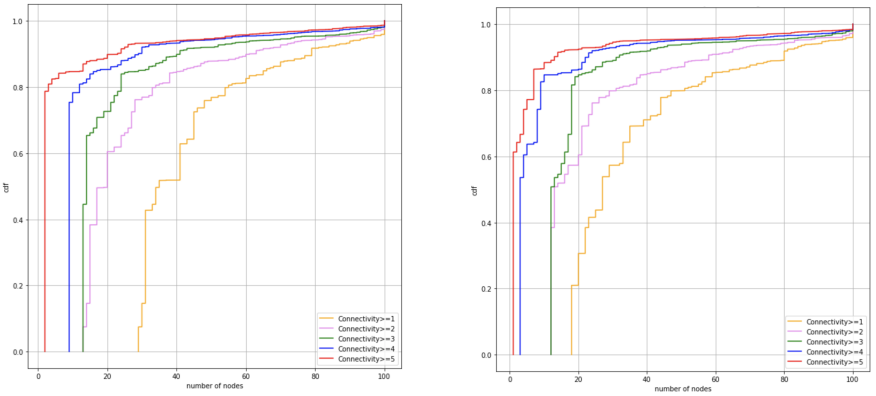
than or equal to the argument of the function. In this case, the random variable is the total connections in the network. Figure 24 below shows the plot for both the morning and evening dataset.



(a) Morning Dataset (b) Evening Dataset

Figure 24: Cumulative distribution function for total number of connections at a time (Morning vs Evening)

Figure 25 below shows the cumulative distribution function of number of nodes which have connectivity higher than equal to the given number for both the morning and evening dataset.



(a) Morning Dataset (b) Evening Dataset

Figure 25: Cumulative distribution function for number of nodes having different connectivity(Morning vs Evening)

The correlations between relevant features in the morning and evening dataset

are given below. The highest correlation in both the datasets is between vehicle speed and vehicle angle.

	vehicle_angle	vehicle_pos	vehicle_speed
vehicle_angle	1.000000	-0.226580	-0.451956
vehicle_pos	-0.226580	1.000000	0.241447
vehicle_speed	-0.451956	0.241447	1.000000

(a) Morning Dataset

	vehicle_angle	vehicle_pos	vehicle_speed
vehicle_angle	1.000000	-0.214979	-0.426248
vehicle_pos	-0.214979	1.000000	0.218127
vehicle_speed	-0.426248	0.218127	1.000000

(b) Evening Dataset

Figure 26: Correlation matrix (Morning vs Evening)