Machine Learning Model Development for Traffic Speed and Flow Prediction

ENCE 688V Project I

Fall 2022

Ainur Abilbayeva

1. Introduction

The scope of this research is to study the traffic sensor detector data through physics-guided machine learning. The dataset is obtained from the Freeway Performance Management System (PeMS) database from 7th to 20th January 2019. The two machine learning models: neural networks and decision trees have been used to predict the speed and flow captured by sensors. Although, the model developed can take as an input larger data than presented.

2. Performance Freeway Management System (PeMS) Dataset

2.1 Overview of PeMS dataset

The traffic information (such as traffic speed, traffic flow and travel time, etc.) is collected by multiple sensors sources (such as inductive loops, radars, cameras, Global Positioning System (GPS), social media, etc.). The performance freeway management system (PeMS) dataset is provided by the Department of Transportation and is available to the public. PeMS is also an Archived Data User Service (ADUS) that provides over ten years of data for historical analysis. It integrates a wide variety of information from Caltrans and other local agency systems including traffic detectors, incidents, lane closures, toll tags, roadway inventory, vehicle classification and etc. The dataset includes traffic flow and speed data from sensors recorded every five minutes. The dataset contains information for 14 days, and there are five items in each data file, which are the recording from certain sensor. Even though only 2 items were chosen to construct the data for the machine learning model in the first task. The time and location of sensor are input attributes/variables and 2 items (speed and flow) are the targets. The attributes are divided into two categories: numerical and categorical. The 2 targets are the traffic speed and flow at a certain sensor location, which is a continuous numerical value. A summary of the attributes and targets in all the tasks (1 to 3) is shown in Table 1.

Table 1. Attributes and targets

Categorical Attributes	Numerical Attributes	Targets
Time (5 mins)	Distance	PeMS Flow
Day of week	iPeMS Flow	PeMS Speed
	iPeMS Speed	

2.2 Exploratory Data Analysis

Most of the real-world datasets come with lots of missing data, but in this case a well-organized and complete dataset from the PeMS database is analyzed. There are two segments given with six and seven sensor recordings in a dataset (the first segment S375 - S393, with the original point as S375; the second segment S401 - S420, with the original point as S401). Since all the sensor recordings data have same shape and attributes, in order to simplify the analysis, the data files have been merged in

each of the segments in Task 1. The targets of the machine learning algorithms are the traffic flow and speed. Figure 2 shows the distribution of the two in terms of occurrence.

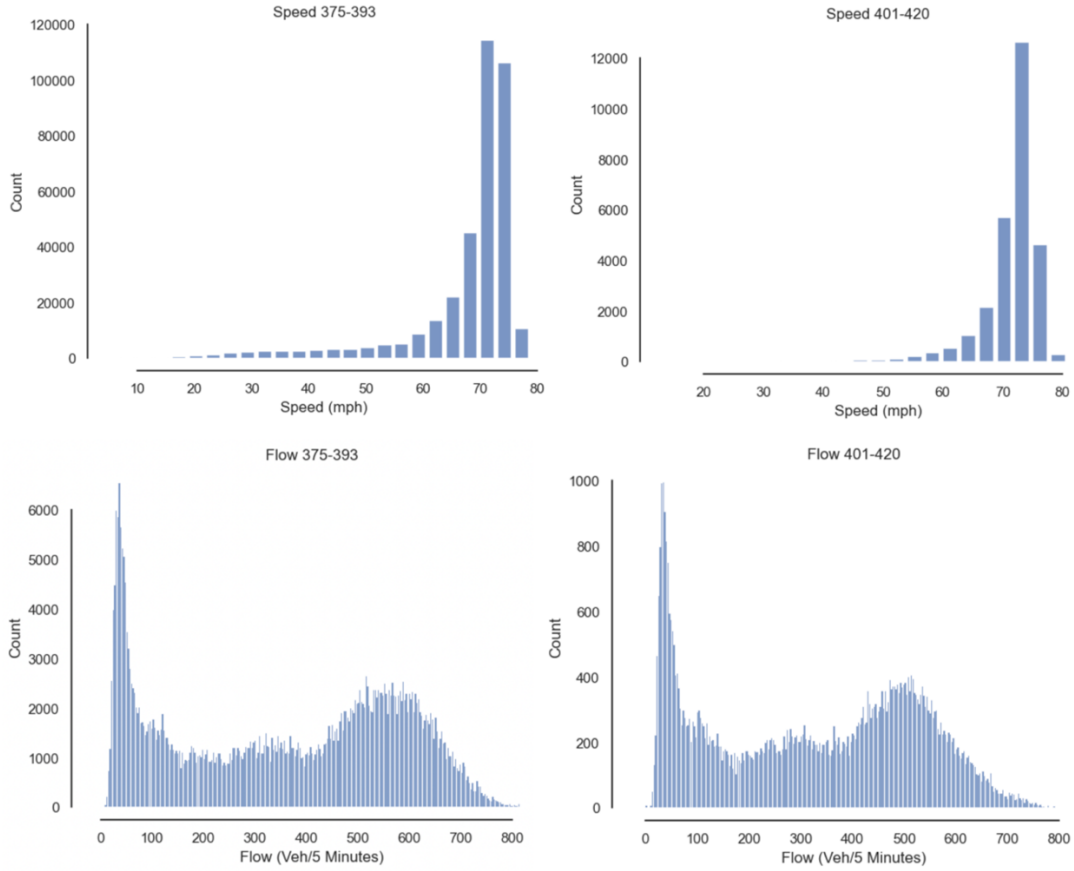

Figure 2. Distribution of traffic flow and speed (Task 1)

Figures 3 and 4 summarize the main statistics of the dataframe in each of the segments.

	Time of day	Week	Distance	Flow (Veh/5 Minutes)	Speed (mph)
count	24192.000000	24192.000000	24192.000000	24192.000000	24192.000000
mean	143.500000	4.000000	2.443333	351.813657	67.400599
std	83.139656	2.000041	1.638332	224.413025	10.882657
min	0.000000	1.000000	0.000000	8.000000	6.700000
25%	71.750000	2.000000	0.990000	118.000000	67.100000
50%	143.500000	4.000000	2.480000	376.000000	71.200000
75%	215.250000	6.000000	3.840000	552.000000	73.200000
max	287.000000	7.000000	4.870000	815.000000	78.600000

Figure 3. Statistical summary of dataframe for segment 375 – 393.

	Time of day	Week	Distance	Flow (Veh/5 Minutes)	Speed (mph)
count	28224.00000	28224.000000	28224.000000	28224.000000	28224.000000
mean	143.50000	4.000000	2.334286	320.120784	71.118211
std	83.13941	2.000035	1.447554	203.688538	5.733879
min	0.00000	1.000000	0.000000	0.000000	14.700000
25%	71.75000	2.000000	0.950000	110.750000	70.200000
50%	143.50000	4.000000	2.250000	337.500000	72.600000
75%	215.25000	6.000000	3.930000	497.000000	74.100000
max	287.00000	7.000000	4.320000	805.000000	82.000000

Figure 4. Statistical summary of dataframe for segment 401 – 420.

For the second task, the speed and flow from iPeMS and PeMS datasets have been analyzed. The graphs of probe and sensor data detected in one day (7th of January) with 5 mins interval for speed is shown in Figure 5 below for each of the sensor locations.

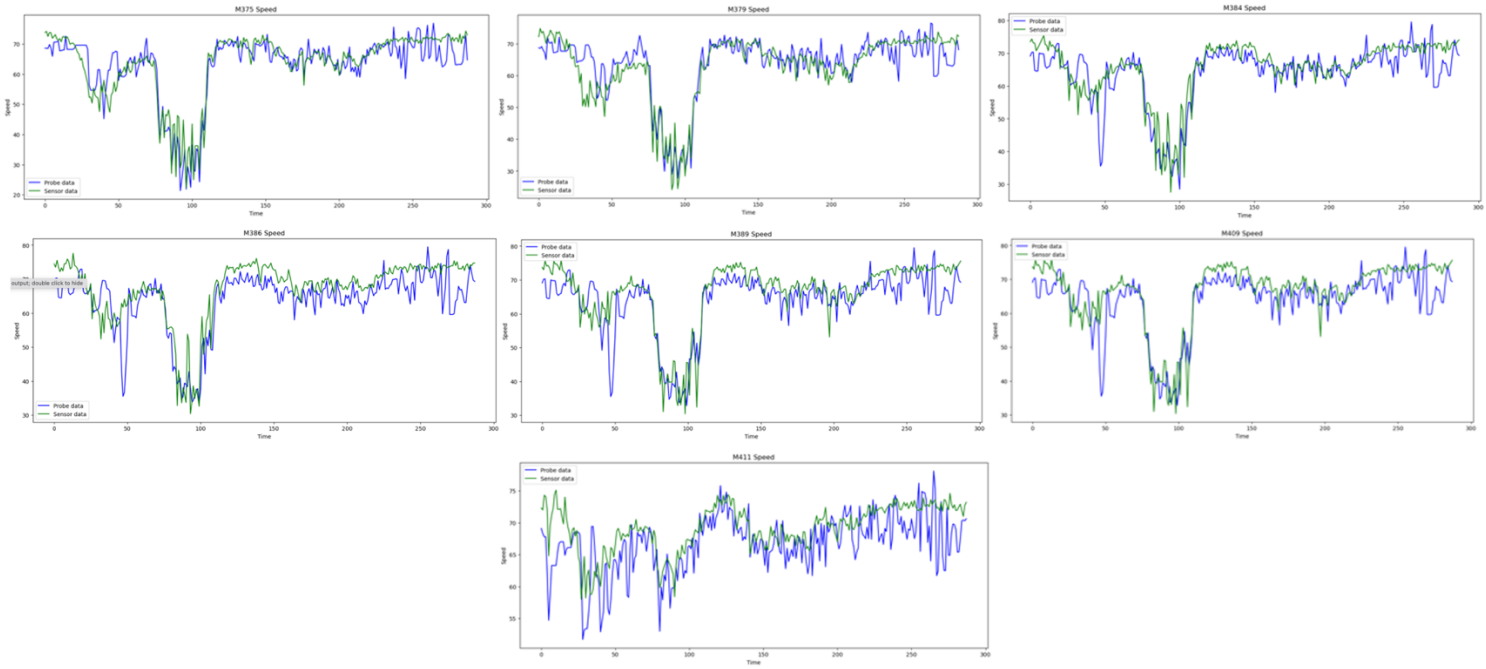

Figure 5. iPeMS and PeMS Speed data

3. Data Preprocessing

3.1 Data Normalization

Machine learning algorithms perform best with normalized data. Data may come in different scales. In this study, for all the 3 tasks the normalization was performed by converting the raw speed and flow traffic data to range from 0 to 1 with the following formula:

$$x_n = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

where x_n and x_i denote the normalized and raw speed and flow data; x_{max} and x_{min} are the minimum and maximum of the raw speed and flow data.

3.2 Training, Validation, and Testing Sets

The PeMS dataset was split into 3 sets: training, validation, and testing. Since this is a supervised learning problem (the datasets are comprised of both location and time as an input and speed and flow detected by sensor as an output), the model should be trained and validated by comparing the results of the set. Training data is the data used to fit the model, which has been chosen as 70% for the models. In order to provide an unbiased evaluation of a final model fit on a training dataset the testing dataset is set aside. For tuning of model hyperparameters in machine learning algorithm, the sample of data called validation dataset is set aside. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The splits of data for this study are shown in Figure 6.

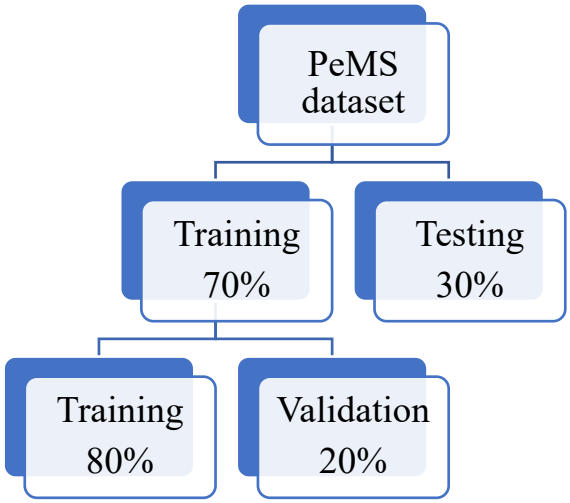

Figure 6. Splits of training, validation, and testing

4. Machine Learning Models

4.1 Overview of The Neural Networks

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data. Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. For all of 3 tasks, the neural network is constructed using a sequential model, with linearly stacked dense layers. To predict the continuous values of traffic speed and flow, this project is defined as a linear regression problem. The number of nodes in output layer is given as 1 due to continuous range of variables. The rectified linear unit (relu) activation function is chosen to be the activation function for the input and hidden layers. For a linear regression problem, the Root Mean Squared Error (RSME) loss function is selected to be the most suitable. While, in a similar way optimizer ‘adam’ gives the most accurate results. To evaluate the models, Mean Absolute Percentage Error (MAPE) as well as RSME metrics have been estimated from each of the epochs (both training and validation loss).

4.2 Overview of Decision Trees

Decision Trees are a non-parametric supervised learning method used for solving classification and regression problems. The major goal is to generate a model which will predict the target variable values by learning simple rules gathered from data features. A tree can be seen as a piecewise constant approximation. The major advantage of decision trees is that it requires little data preparation and does not require data normalization compared to neural networks in example. However, they are not stable because small differences in the data lead to entirely different tree generation.

5. Speed and Flow Prediction using Neural Network and Decision Trees

5.1 Task 1

In this task, neural network model was constructed using PeMS data only. The location of the sensor and time t have been used as an input to predict the speed and flow for each freeway segment (S375 - S393 and S401 - S420). In general, for this task four models have been developed using the merged files from each sensor datasets. For the speed prediction neural network accuracy has been estimated as 83.15% for the first highway segment (375-393) and 96.56% for the second highway segment (401-420). To reach higher accuracy several models with different epochs and batch sizes have been developed, Figures 7 and 8 present the best results for speed and flow prediction in each of the segments respectively. Table 2 below summarizes mean absolute percentage errors in all machine learning models developed in Task 1.

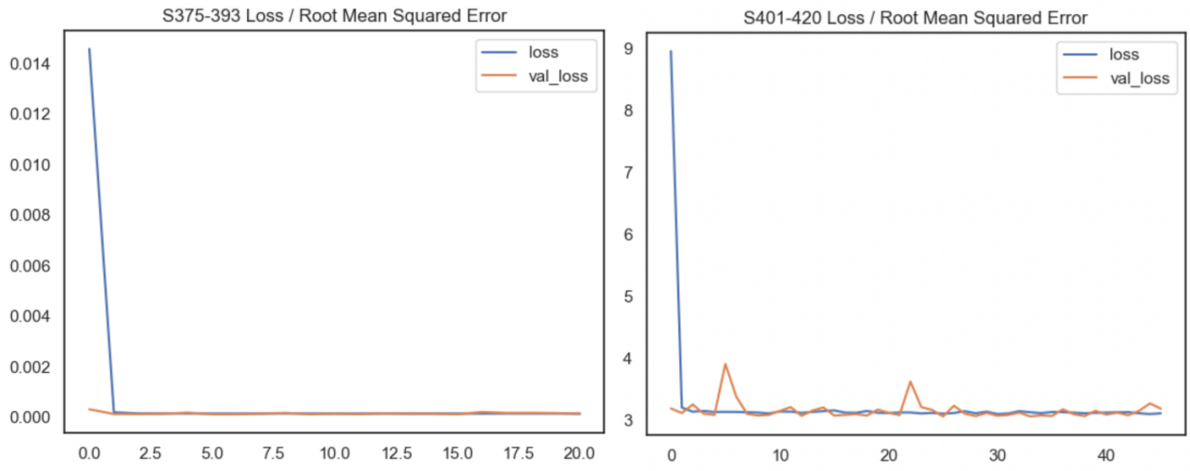

Figure 7. Speed prediction loss for two highway segments

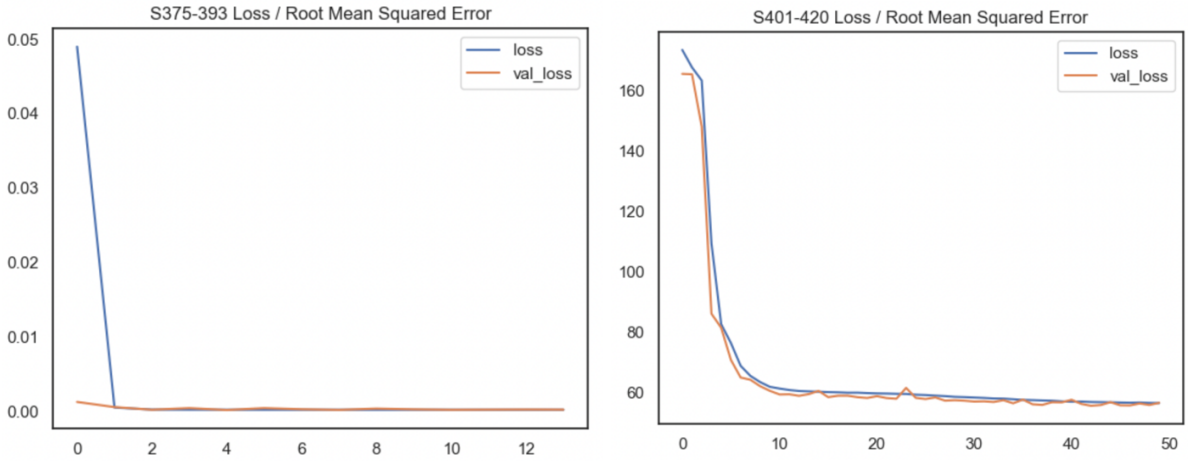

Figure 8. Flow prediction loss for two highway segments

Table 2. Summary of Machine Learning Models for Task 1.

Sensor	ANN MAPE (%)
Speed S375-393	16.85
Speed S401-420	5.34
Flow S375-393	34.43
Flow S401-420	35.30

5.2 Task 2

For task 2, two different machine learning models have been compared. iPeMS data has been taken as an input and PeMS data as an output for each of the freeway segments, at both upstream and downstream stations of the sensor *i*, where only probe vehicle data is given. The mean absolute percentage error of 4.38% in speed prediction has been achieved for sensor 375 using neural networks model, while using decision trees the accuracy was 5.4%. In a similar way, this has been repeated for each of the sensors. To get better results in prediction, several neural network models have been built for each sensor. As an example, neural network summary for speed prediction of sensor 375 is shown in Table 3. The best results were obtained, while constructing a model with 1 intermediate layer with epochs of 50 and batch size 16.

Table 3. Speed prediction neural networks summary for S375

Speed prediction	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Hidden Layers	1	2	1	1	1	1
Layer 1 nodes	128	128	64	64	32	64
Layer 2 nodes	64	64	64	64	64	64
Layer 3 nodes		64				
Epochs	50	100	100	50	50	25
Batch Size	16	32	16	16	32	16
Accuracy	86.4%	87.2%	86.1%	**96.20%**	87.1%	**94.9%**

The results of loss functions for each sensor are indicated in the figure 9 below. The summary of the MAPEs for each of the models is given in Table 4. In general, neural network model performed better for most of the sensors, due to lower MAPE compared to Decision Trees.

Table 4. Summary of Machine Learning Models for speed prediction for Task 2.

Sensor	ANN MAPE (%)	Decision Trees MAPE (%)
S375	4.38	5.40
S379	4.94	5.11
S384	5.29	5.14
S386	4.49	4.97
S389	5.18	6.02
S409	3.02	3.3
S411	3.75	3.16

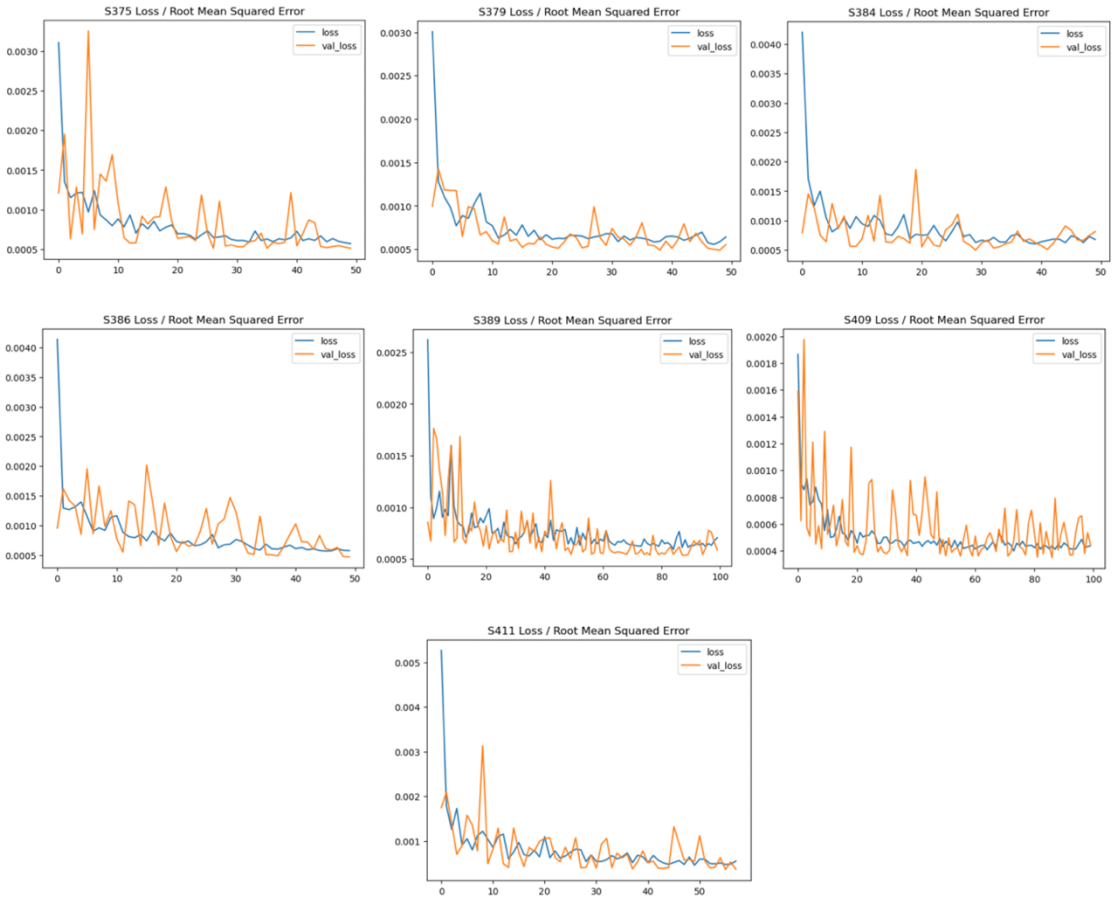

Figure 9. Speed prediction models for each sensor location for Task 2.

In a similar way, the machine learning models have been developed for each sensor location to predict the flow of vehicles in 5 minutes. The results are presented in Figure 11 and Table 5. For flow prediction both machine learning models (neural networks and decision trees) have given same results, since the MAPE values are very similar for each sensor with slight differences only.

Table 5. Summary of Machine Learning Models for flow prediction for Task 2.

Sensor	ANN MAPE (%)	Decision Trees MAPE (%)
S375	20.86	21.01
S379	20.18	20.07
S384	17.64	18.06
S386	20.03	18.25
S389	18.12	20.87
S409	19.22	17.18
S411	17.62	17.53

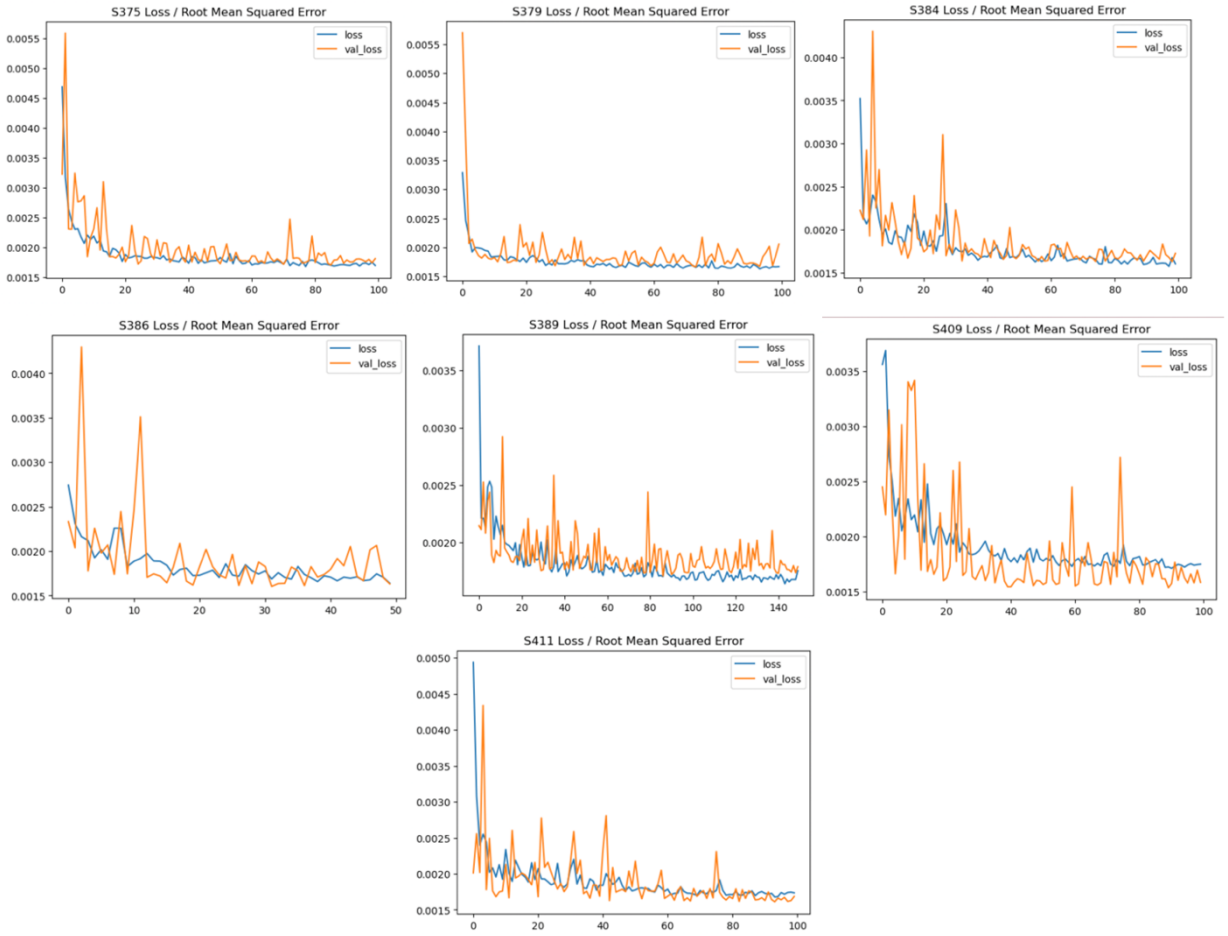

Figure 10. Flow prediction models for each sensor location for Task 2.

The network summary of the sequential model 4 in Task 2 is given in Figure 11 below, which is consisted of input and output layers including one intermediate layer. Overall, for both speed and flow prediction the models were more accurate when iPeMS data was given as an input.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                192

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 4,417
Trainable params: 4,417
Non-trainable params: 0
_________________________________________________________________
```

Figure 11. Network summary of model 2 (Task 2).

5.2 Task 3

For task 3, two machine learning models: neural networks (using MLPRegressor) and Decision Trees have been used to predict the speed and flow for each sensor adding the flow data from Traffic Flow Model as an input. The output values of MAPE are presented for flow and speed prediction for each of the sensor in tables 6 and 7. The prediction accuracies were very similar to Task 2, however for flow prediction MAPE was significantly of high value. The reason might be incorrect set up of the traffic flow model.

Table 6. Summary of Machine Learning Models for speed prediction for Task 3.

Sensor	ANN MAPE (%)	Decision Trees MAPE (%)
S401	5.2	5.01
S406	4.36	4.17
S409	5.84	5.5
S411	5.58	5.38
S414	5.06	5.14
S419	6.49	5.69
S420	7.38	6.33

Table 7. Summary of Machine Learning Models for Flow prediction for Task 3.

Sensor	ANN MAPE (%)	Decision Trees MAPE (%)
S401	42.76	31.96
S406	45.4	33.85
S409	39.72	30.46
S411	20.08	12.21
S414	37.24	33.42
S419	43.57	31.66
S420	35.49	28.53

6. Conclusion

To conclude, two machine learning models have been used to predict speed and flow values from PeMS and iPeMS data. In Task 1 PeMS data from nearby detector was used for training, and selected detector data was tested. For Task 2 iPeMS data was used as input and PeMS as output for training, while in Task 3 one more data was added from traffic flow model as input. As the results have shown in Task 2, using iPeMS data for prediction significantly increases the accuracy of the prediction.

However, in general for linear regression problems neural networks are very hard to get working precisely because the output is unbounded, so it is prone to the exploding gradients problem (the cause of the nans, which was reduced by adjusting different parameters). The major reason for choosing deep neural networks was to learn the use of them through this project.