UNIVERSITY OF SOUTH FLORIDA

MUMA COLLEGE OF BUSINESS

DATA MINING PROJECT REPORT
COURSE: ISM6136.001F22

NOVEMBER 2022

METRO INTERSTATE TRAFFIC VOLUME I-94

Project Members:

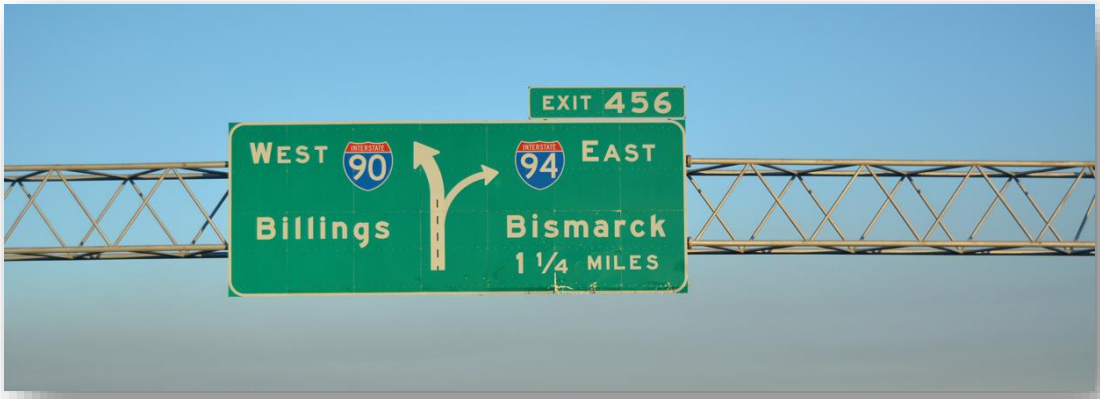
Durga Mohan Bathula
Sai Pavan Banala
Arshad Abdullah Mohammad
Sandhyasree Yarra
Abhishek Vishnubhaktula

Introduction:

Interstate 94 (I-94) is an east–west Interstate Highway connecting the Great Lakes and northern Great Plains regions of the United States. Its western terminus is just east of Billings, Montana, at a junction with I-90; its eastern terminus is in Port Huron, Michigan, where it meets with I-69and crosses the Blue Water Bridge into Sarnia, Ontario, Canada, where the route becomes Ontario Highway 402. It thus lies along the primary overland route from Seattle (via I-90) to Toronto (via Ontario Highway 401), and is the only east–west Interstate highway to have a direct connection to Canada.

I-94 intersects with I-90 several times: at its western terminus; Tomah to Madison in Wisconsin; in Chicago; and in Lake Station, Indiana. Major cities that I-94 connects to are Billings, Bismarck, Fargo, Minneapolis–Saint Paul, Madison, Milwaukee, Chicago, and Detroit.

Azure Machine Learning Studio is used to train the models from the obtained dataset.



The I-94 Traffic Dataset:

This dataset consists of traffic volume between Minneapolis and Saint Paul. The dataset has 48121 rows and 10 columns, all of them with no missing values, right data type and no problems for working with original column labels. No data cleaning necessary at this stage.

Every row traffic and weather condition for a specific hour, containing data from 2012-10-02 09:00:00 until 2018-09-30 23:00:00.

Due to the location where this data was obtained. we will consider the results of this analysis to be pointed to the westbound traffic (cars moving from east to west) in the proximity of the station and not to the entire I-94 highway.

Expected Outcomes:

To increase road safety:

- By using the predicted data generated from the model, traffic can be better regulated with the use of traffic signals.
- From the data of the previous accidents a heatmap can be generated highlighting the accident-prone areas in the highway.

Maintenance of Roads:

- The data model which is generated can be used to predict at what times the traffic volume is low so that maintenance of the roads can be planned accordingly.
- This results in the increase of work efficiency when repairing the roads.

Surge in fare prices:

- Commute and travel apps like Uber and Lyft can increase the fare prices based on the traffic and weather conditions.
- If the traffic volume is more in a particular time, the cab and vehicle rental fares can be surged.
- The prices can also be surged based on the weather conditions.

Dataset Characteristics:

DateTime - Hour of the data collected in local CST time

Junction - The number of the junction on the highway

holiday - US National holidays plus regional holiday

temp - Numeric Average tempature in kelvin

rain_1h - Numeric Amount in mm of rain that occurred in the hour

snow_1h - Numeric Amount in mm of snow that occurred in the hour

clouds_all - Numeric Percentage of cloud cover

weather_main - Categorical Short textual description of the current weather

weather_description - Categorical Longer textual description of the current weather

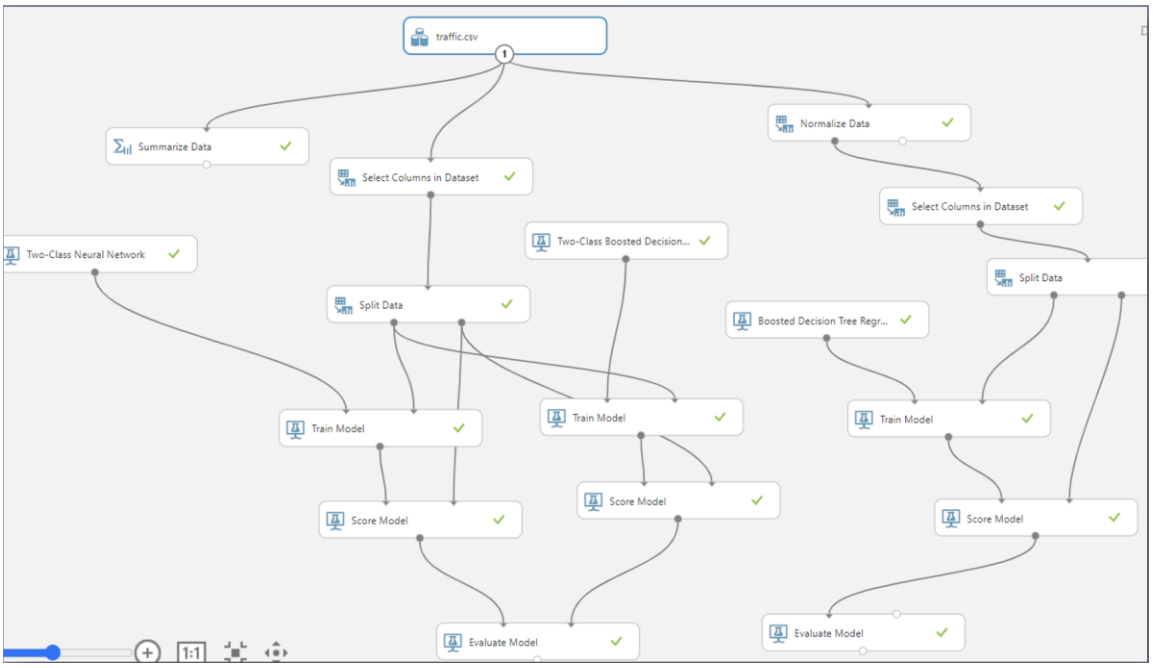
traffic_volume - Numeric Hourly I-94 ATR 301 reported westbound traffic volume

Workflow:

The first item in the workflow is our dataset which is "traffic.csv". After importing the data we used the item 'Select Columns in Dataset' to select columns from the dataset. We disregarded the 'Weather Description' column from the dataset as there is no correlation with the dependent variable.

The item 'Split Data' splits the data into training and test data. The training data is 70 percent and test data is 30 percent of the dataset. The data is fed into the 'Train Model' item. Three models are trained using three different machine learning techniques. The machine learning models used are Two-class neural network, Two-class boosted decision tree, Boosted decision tree regression. All

the three models are scored for efficiency using the ‘Score Model’ item. The test is used to score the models. The models are evaluated and compared using the ‘Evaluate Model’ item.



Machine Learning Models:

1.Two-class neural network:

Classification using neural networks is a supervised learning method, used for model to predict binary outcomes. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes. To compute the output of the network for a particular input, a value is calculated at each node in the hidden layers and in the output layer.

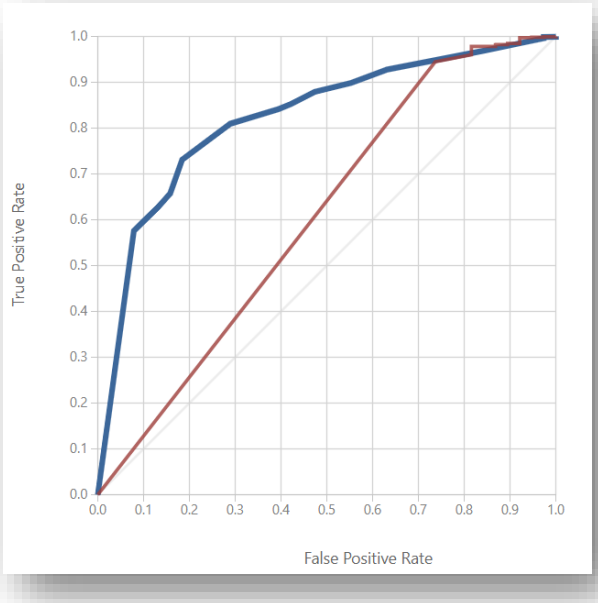
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
11500	11	0.799	0.594	0.5	0.825
False Positive	True Negative	Recall	F1 Score		
37	1	0.534	0.562		

2.Two-class boosted decision tree:

A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction. Generally, when properly configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks.

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
11500	11	0.949	0.949	0.5	0.978
False Positive	True Negative	Recall	F1 Score		
18	20	0.997	0.972		
Positive Label	Negative Label				
3	1				

The above models are evaluated and compared using the 'Evaluate Model' item in Azure ML Studio. An ROC()Receiver Operation Characteristic Curve) is generated to compare the accuracy of the model. We can see that the blue plot which is of the Two-class boosted decision tree has a larger AOC(Area Under the ROC Curve). The red plot is of the two-class neural network which has a lower accuracy compared to the red plot.



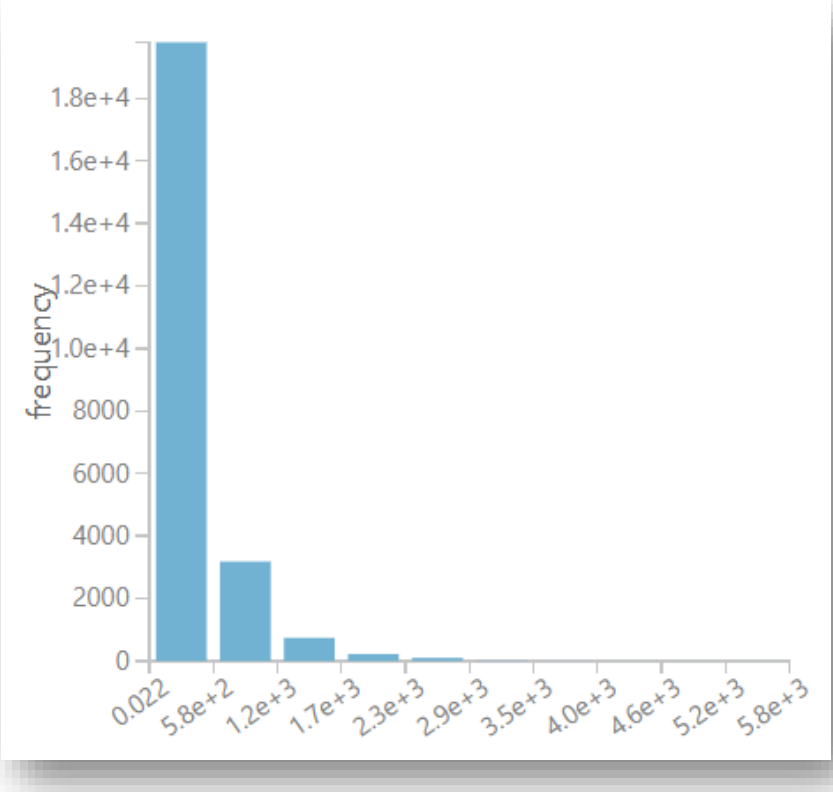
ROC Plot

3.Boosted decision tree regression:

Gradient boosting is a machine learning technique for regression problems. It builds each regression tree in a stepwise fashion, using a predefined loss function to measure the error in each step and correct for it in the next. Thus, the prediction model is actually an ensemble of weaker prediction models. The regression model used here only taken numeric Independent Variables.

Coefficient of determination, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit.

Relative Absolute Error	0.58552
Relative Squared Error	0.422426
Coefficient of Determination	0.577574



Predictive Analysis and web deployment:

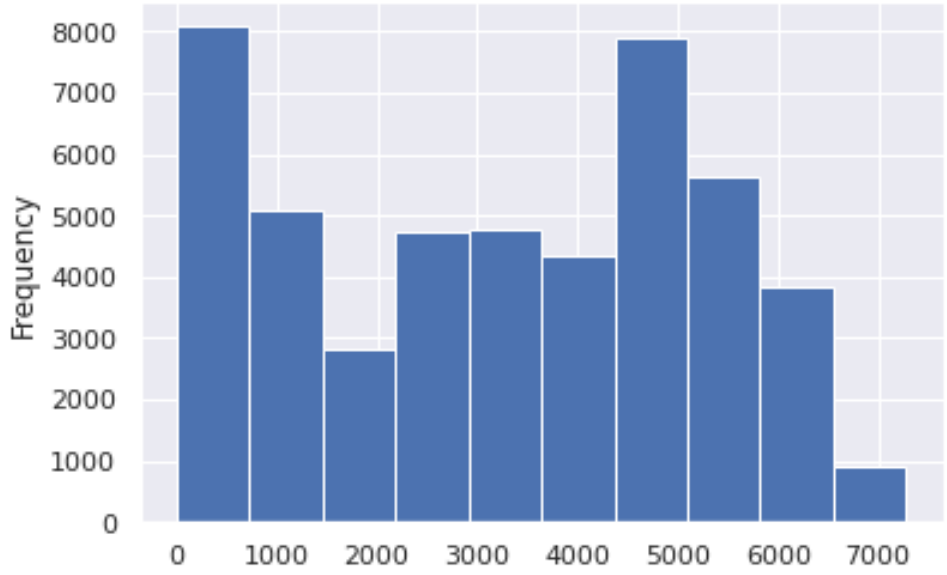
The model which is created using the 'Boosted Decision Tree Regression' method is deployed to the web to test the predicted outputs. A predictive experiment is created in Azure ML Studio using this model. The web service which is created using the predictive experiment is deployed into the web. Using the integration of Azure ML studio and Microsoft Excel, a sample data set of the independent variables is given as an input to the data model created. The model predicted the dependent variable which is the traffic volume and the output is generated in an excel spreadsheet. Below is the sample data which is given as an input and the predicted traffic volume which is generated by the data model.

Junction	holiday	temp	rain_1h	snow_1h	clouds_a	weather_	Traffic Vc	Scored Labels
1	None	288.28	0	0	40	Clouds	585	1582.902222
1	None	289.36	0	0	75	Clouds	507	1730.321411
1	None	289.58	0	0	90	Clouds	390	1718.067993
1	None	290.13	0	0	90	Clouds	273	1618.49585
1	None	291.14	0	0	75	Clouds	351	1777.270996
3	None	333	0	0	66	Clouds	666	603.3262939

Observations:

Based on few factors we have observed the following:

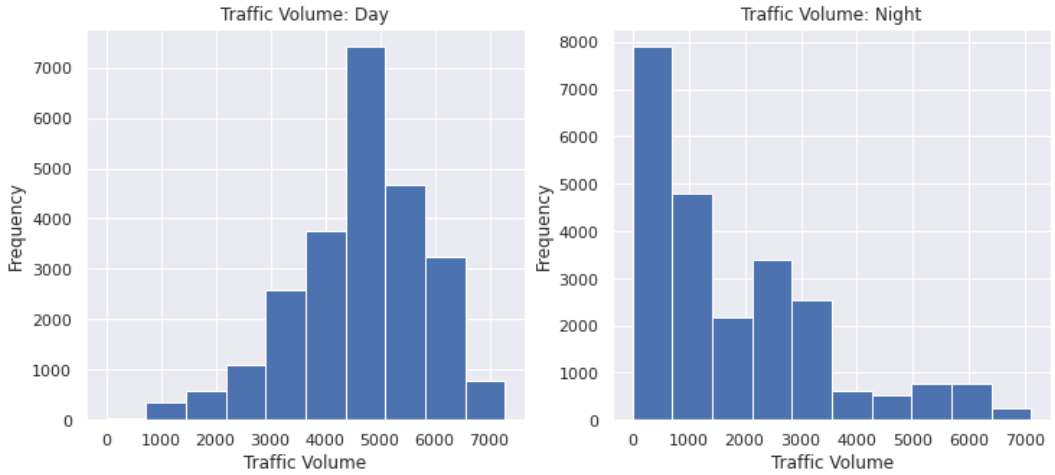
1. Exploring traffic volumes



The average number of cars per hour between 2012-10-02 09:00:00 and 2018-09-30 23:00:00 is 3260. The hourly traffic fluctuates from 0 to 7280 cars per hour. As we can see from the

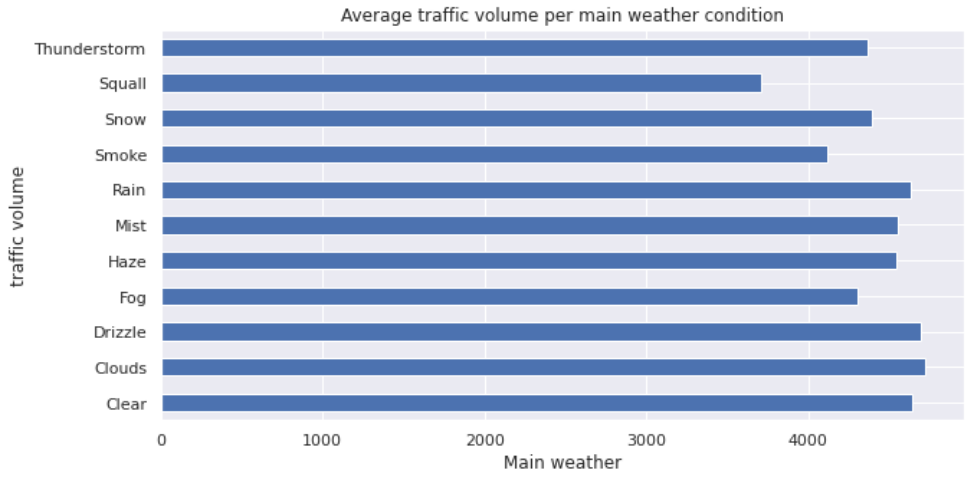
histogram, there are two peaks where the frequency nearly equals one another, suggesting that there may be a time variable explanation. We might consider that there is a correlation between traffic volume and daytime, with less activity occurring at night and more occurring during the day on the same date.

2.Analyzing day time vs night time:



The data on traffic volume during the day is left skewed, which means that there are lower frequencies of lower traffic volume at first and subsequently climbs to a traffic volume of 5000 automobiles per hour registered more than 7000 times between 2002 and 2018. This also implies that the majority of daylight readings are high. Additionally, 4019 or more cars passed the station 75% of the time, indicating that the majority of daytime data have high values. The histogram for the nighttime data is right skewed, behaving differently than it does during the daytime, with smaller values having higher frequencies. 25% of the time, there were even less cars per hour than 517, and 75% of the time, there were fewer than 2775. We will only take into account daytime data for the results because our objective is to discover indicators of heavy traffic.

Climate indicators:



The main weather conditions don't alter all that much. The use of precise meteorological terminologies like "light rain and snow" and "shower and drive" appears to have increased slightly. Although there are other more harsh days where people didn't seem to have the same behavior, one explanation for this behavior is that people may have driven to work on days with bad weather rather than biking or walking. Not much to say in this regard.

Conclusion:
In this project, we aimed to find a few indicators of heavy traffic on the I-94 Interstate highway. We come to the conclusion that datetime indications are more important than weather indicators. Indicators of heavy traffic can be seen in this dataset, including:

- Traffic is heavier in rush hour at morning from 6am to 8amand evenings from 3pm to 6pm.
- Traffic is heavier in weekdays than in weekends.
- Traffic is heavier in Autumn and Spring rather than Summer and Winter.
- No strong correlation between temperature and traffic volume.
- Not much variance between different weather conditions.

With all these data patterns from the trained models, our expected outcomes can be achieved.