Can we predict traffic congestion using past data?

Amit Pandey

Data Science Department, Saint Peter's University

apandey@saintpeters.edu / +1(732) 986-6045

Abstract

The urban traffic congestion is transforming into an epidemic all over the world. As a result of the constant traffic congestion, transportation cost has increased significantly due to all the time wasted on the road and the corresponding fuel cost. A real smart city will only be smart if we can use the past data to predict the future traffic congestion and take necessary steps to mitigate the issues. We will use traffic congestion prediction models using supervised learning and time series analysis. Subsequently, these models will be cross validated to see which model performs the best in both training and testing environment.

Introduction

Diverse, forefront and fun, Aarhus is a champion among a delighted urban zones on Earth. 13 % of Aarhus' people are understudies, making Aarhus the most energetic city in Denmark. For the most part, in any case, it's a standout amongst the most settled. 315,000 people live in Aarhus and 1.2 million people live in the more significant Aarhus area, so it's the perfect size for an end of the week break or family getaway. It is crucial to forecast traffic for a smart city and play a fundamental role for any city in planning and development of traffic management and control. [4] The goal is to predict traffic conditions in a transportation network based on its past behavior. Using Road Traffic data from CityPulse dataset collection for city of Aarhus in Denmark, this project will explore different ways of predicting traffic congestion in future.[2] This dataset is collection of traffic data between two points for certain duration of time.

Method

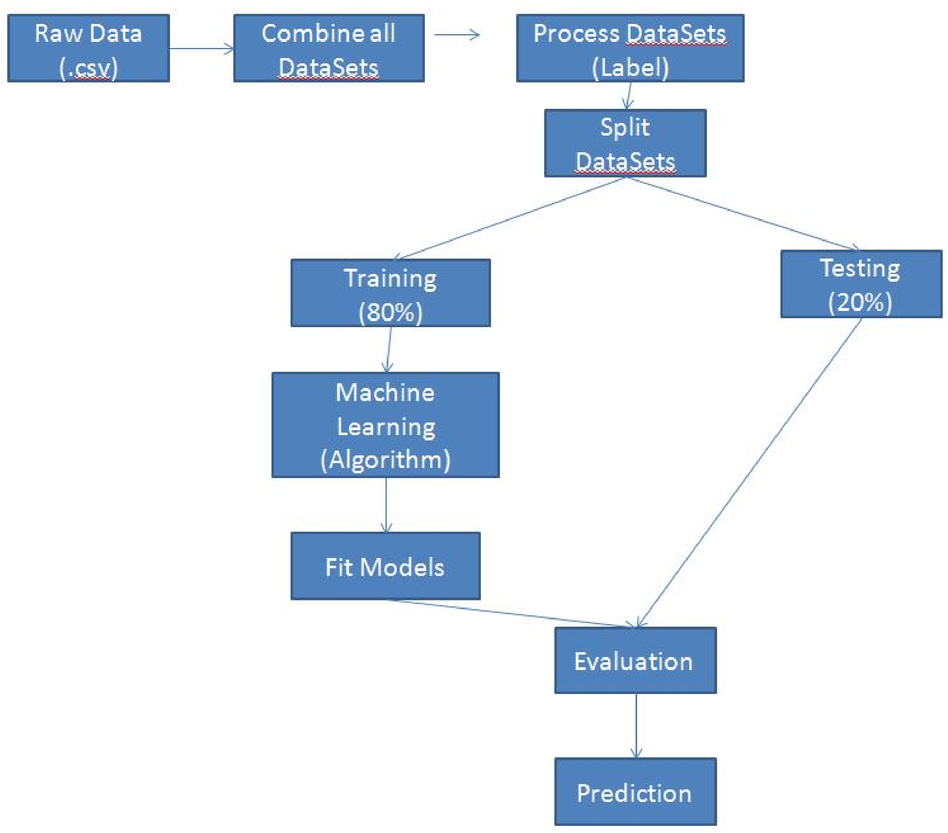

Figure 1: The flow chart summarizes the process that I intend to use for the capstone project. The idea is to split the data set into training (80 %) and test (20 %) in order to test the algorithm on the training and apply on the test data set.

Classification Model

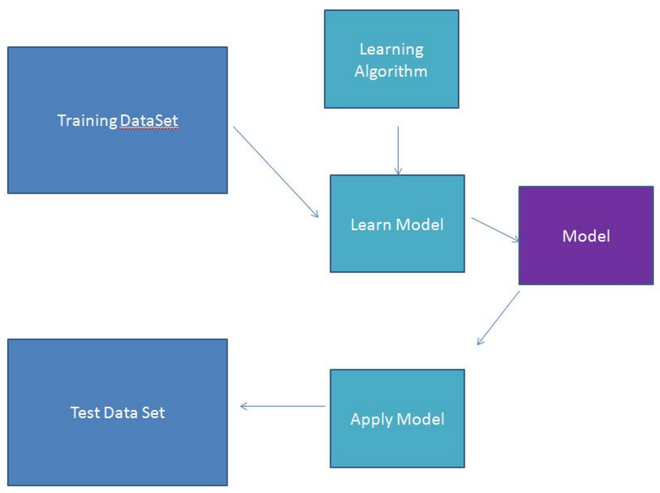

Figure 2: The flow chart summarizes the classification model I intend to use to train and test the data set to put the traffic congestion can be broken down in 3 different levels.

Here are the different predictive models that will be explored for the study:

KNN:

K closest neighbors (KNN) is a basic algorithm that stores all accessible cases and characterizes new cases in view of a closeness measure. KNN has been utilized as a part of measurable estimation and example acknowledgment as of now in the start of 1970 as a non-parametric strategy. KNN is a non parametric lazy learning calculation. When we say a strategy is non parametric, it implies that it doesn't make any assumption on the data distribution. This is quite helpful, as in practical world, the greater part of the practical information does not comply with the ordinary hypothetical suppositions made.

Support Vector Machines

SVMs also support vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Margin/support to make a group for the training data. A SVM is a discriminative classifier also defined by a separating hyper plane. In other words, in supervised learning, given labeled training data, the algorithm outputs an optimal hyper plane that can categorize new examples.

Random Forest

Random forests or random decision forests are a group learning technique for classification, relapse and different undertakings, that work by developing a huge number of decision trees at preparing time and yielding the class that is the method of the classes (classification) or mean forecast (relapse) of the individual trees. Random decision forests adjust for decision trees' propensity for over fitting to their training set. Forests develops numerous classification trees. To classify another question from an information vector, put the info vector down each of the trees in the timberland. Each tree gives a classification, and we say the tree "votes" for that class. The backwoods picks the classification having the most votes (over every one of the trees in the timberland).

Time Series Analysis

Since the dataset we are looking at is time based we can conduct time series analysis to predict future points in the series. We can also look into seasonality as we may guess that the traffic data changes during the course of the day and the week. We can use the velocity data to predict future velocity or number of vehicles at a given time using time series models like Auto-Regressive Integrated Moving Average (ARIMA) model.

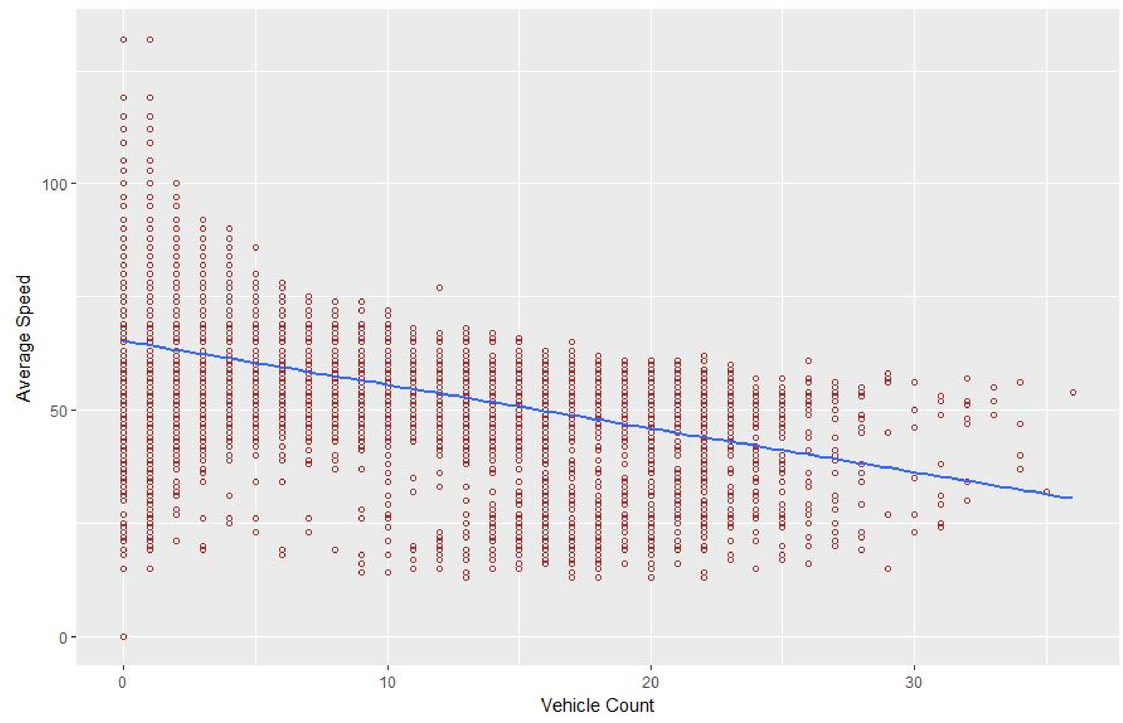

Figure 3: Scatterplot of Vehicle Count vs. Average Speed. As the vehicle count increases in the road average speed decreases.

Conclusions

- Using vehicle count and hour of the day, we will create supervised models and time series analysis for predicting average speed and cross validate the results
- We will determine which model performs the best amongst the models used.

References

[1] Afshin Abadi, Tooraj Rajabioun, and Petros A Ioannou. Traffic flow prediction for road transportation networks with limited traffic data. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):653–662, 2015.

[2] iot.ee.surrey.ac.uk. Dataset colelction, 2016. iot.ee.surrey.ac.uk:8080/datasets.html.

[3] S Vasantha Kumar and Lelitha Vanajakshi. Short-term traffic flow prediction using seasonal arima model with limited input data. *European Transport Research Review*, 7(3):1–9, 2015.

[4] Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

[5] Dawen Xia, Huaqing Li, Binfeng Wang, Yantao Li, and Zili Zhang. A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction. *IEEE access*, 4:2920–2934, 2016.