University of Padua

Department of Mathematics
Master Degree on Data Science

Business, Economics and Financial Data

Behavior of the urban traffic of the city of São Paulo in Brazil

Author:
Piero Romare,
University of Padua

Supervisor:
Prof. Omar Paccagnella,
University of Padua

Contents

Abstract

São Paolo du Brazil is the most populated city of Brazil with over than 10 millions of citizen. Furthermore, São Paulo is the largest city in all of South America and among the largest cities in the world. Over the years the information has had an increase in movement speed. On the same wavelength, the needs of large districts need to increase the movement of goods and people both in terms of quantity and quality. Like online traffic, urban city traffic also needs efficiency and speed. Traffic behavior in cities requires optimization to fulfil what globalization processes and requests of the market are.

In this project is developed a comparison of different machine learning regression models and Neural Networks to predict the traffic behavior. Using the dataset "Behavior of the Urban Traffic of the City of São Paolo in Brazil", various models are available that can predict vehicles routing problem.

1 Introduction

In recent years with the exponential growth of the economic system, the movement of goods and people has become essential. The modern society accompanied by an increase in mobility. Carrying out a movement with any mode of transport entails for the user a series of costs of use and travel times of a network element which depend on:

- the levels of use of the element itself and possibly other elements which interact with it;
- the characteristics of the network element.

About costs, we can distinguish measurable charges:

- monetary type;
- non-monetary (essentially timing charges not measurable).

Examples of measurable monetary charges are the various costs related to the possession and use of a private vehicle, parking fees, motorway tolls, rates for use of collective transport systems. Measurable non-monetary charges are the various components of travel time, which for the user represents a cost since it is subtracted from other activities, for example, travel times onboard a vehicle, waiting times at a stopped bus, parking search times. Non-measurable charges are essentially non-quantifiable factors that the user can perceive as negative elements when making a move (i.e. the risk of accidents).

Urban traffic congestion is an open issue. It is characterized by the network of cars and its costs are both monetary (i.e. oil) and non-monetary (i.e. time). The field that deals with this type of problem and offers some solutions is called Intelligent Transportation Systems (ITS). Citizens suffer from the decreasing in travel efficiency due to the growing number of people on the move. So, traffic flow's temporal dependence is crucial to the effectiveness of a prediction regarding the shifting traffic conditions [1]. In other words, if we can observe traffic's evolution with a temporal pattern we should better predict traffic pattern [2]. Estimation and prediction of traffic behavior open different application suggestions for human, driver assistance and autonomous driving.

This project focused on modelling some "non-monetary" variables proposed in the dataset. It is proposed some descriptive analysis about the

accident and the characteristic of São Paolo traffic. It is demonstrated the utilization of three machine learning regression models and two different Neural Networks to achieve different methodologies and see how the performance indexes change.

1.1 Contribution

As can be seen in the following section 2 (e.g. Related Work) there are no studies which provide or analyze the traffic of São Paolo with regressive machine learning and Neural Networks techniques. So, the contribution in this project has to aim to compare different regression and Neural Networks techniques to verify the possibility to develop other kinds of prediction with the dataset proposed. As mentioned in the introduction section, it is investigated the models considering all variables and considering only hour variable.

1.2 Organization

This project is organized as follows: Related Work Section discusses previous work and reports; Method Section describes the system for achieving results; Results and Evaluation Section gives experimental results, in Conclusion and Practical Suggestions Sections have summarized the project and discuss the future possibility.

2 Related Work

To predict traffic behaviour, different variables are involved in a practical scenario. Not only about security, which related to the industrial efficiency aim, but also about safety, which is related to human prevention. That last point includes, for instance, the decision making that must follow any kind of situation and context with Dynamic Bayesian Network [3]. Fortunately, in terms of safety and security, many data are easily available and to this, fuel and time could be two important components if minimized for the human. In any case, the objective remains the minimization of accidents [4]. Human, in a high-level point of view, has the choice of following or not the rules of the street, like the observation of the traffic light. It is used to separate in time the flow of the various vehicular and pedestrian currents that cross the intersection, assigning to them free time appropriately sized based on the intensity of the currents themselves and, if violated, red-traffic light has an impact in accidents [5]. Also, frameworks of air pollution and traffic emissions are studied [6, 7].

In particular, here we are focused on papers which used the same dataset that is described in section 3.2: "Behaviour of the Urban Traffic of the City of São Paolo". Ferreira et. al provide a Neuro-Fuzzy Network to forecast the behavior of the traffic of São Paolo with three levels of routing: Strategic, Tactical and Operational. Here, the most important aspect is the maximum efficiency in transportation concerning the path that a pilot of a vehicle will select. Authors used a MultiLayer Perception (MLP) algorithm with backpropagation give alternative models for the traffic routing and dynamic vehicle routing. The model takes into account spatial and temporal aspects [8]. Another work is provided by the same authors using Artificial Neural Network (ANN) to improve effectiveness and productivity for the metropolitan area of São Paolo. It is applied a Rough-Fuzzy Sets to use Dynamic Routing without human supervision. This framework contains two main characteristics as Rough Sets that select the features and the Neuro-Fuzzy Network which generates the surface outcome (e.g. Input/Output). The authors obtain a conclusion in which they create an inference possibility for Traffic Flow that can be correctly characterized by Fuzzy Sets as mentioned before. They conclude also that a better ability to generalize reduced rule basis and that for each inference rule it is possible to create an automatic inference mechanism without human supervision to reduce the number of membership functions [9, 10].

3 Method

3.1 Dataset

The dataset that is used in this project is called "Behaviour of the Urban Traffic of the City of São Paolo", created by Ricardo Pinto Ferreira, Andrea Martiniano and Renato Jose Sassi [8]. It takes into account 19 variables of the traffic and streets features (Table 1).

Table 1: Variables

Day	Hour (Coded)
Immobilized bus	Broken Truck
Vehicle excess	Accident victim
Running over	Fire vehicles
Occurrence involving freight	Incident involving dangerous freight
Lack of electricity	Fire
Point of flooding	Manifestations
Defect in the network of trolleybuses	Tree on the road
Semaphore off	Intermittent Semaphore
Slowness in traffic (%)	

It contains 135 rows, surveys divided into 5 days from Monday, December 14, 2009, till Friday, December 18, 2009. The surveys are 27 per day. The collection provides a survey every 30 minutes, from 7.00 am to 8.00 pm.

- So, the everyday sum of variables is proposed in Figure 7;
- The correlation between the variables in the dataset is proposed in Figure 8, it can be seen that the independent variable Hour correlated with the dependent variable Slowness in traffic has the maximum value with concerning the other correlation;
- Some descriptive analysis of the variables in the dataset is proposed in Figure 9;
- To figure out how the dependence variable "Slowness in traffic (%)" differs during the hours observed (Figure 10);

- In Figure 11 it is shown the mean of incidents during the time;
- In Figure 12 it is shown the causes (variables) of incidents at 7.00 pm which as it can be seen in the previous Figure 11 is the hour when the number of incidents is maximum;
- It's possible to see also how many incidents caused by a single independent variable during the day (Figure 13);
- Figure 14 takes into account the difference of the slowness traffic with or without incidents: for at some times, the slowness is greater without incident than when there are incidents. However, taking into account the data, it is possible that until 11:00 am the occurrence of incidents does not affect traffic slowness so much. And from 14:30, the incidents have a greater impact on the slow traffic.

3.2 Data Processing

To give correct values for the models, which are discussed in the next section, some transformation is used.

First of all, the dataset does not present any Null values. The hour variable is split into 27 different binary columns: the column called 1 represent 7.00 am, the column called 2 represent 7.30 am, the column called 3 represent 8.00 am and so far and so on.

At this point, the dataset is divided into X vectors (e.g. a cluster of all independent variables) and y vector (e.g. a cluster of the dependent variable). The dataset is split in train and test sets with an arbitrarily train the size of 0.8 (e.g. in the training set the observation from Monday to Thursday and in the test set the observation of Friday). The train set of independent variables contains 108 vectors which contain 43 values. The test set of independent variables contains 27 vector with 43 values. Now the data are ready to be fitted in the models.

The models are fitted two independent times: one with X vectors which contain all independent variables and one with X vectors which contain only hour independent variable. The y vector instead it always remains to contain Slowness in traffic dependent variable.

3.3 Models

3.3.1 Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an n^{th} degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted $E(y|x)$. Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y|x)$ is linear in the unknown parameters that are estimated from the data.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

3.3.2 Decision Tree Regression

Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. [11]

Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, it is needed to trim it down. All the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

$$costs = sum(y - prediction)^2$$

To stop the splitting it is set a minimum number of training inputs to use on each leaf or set maximum depth which refers to the length of the longest path from a root to a leaf.

3.3.3 Random Forest Regression

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a

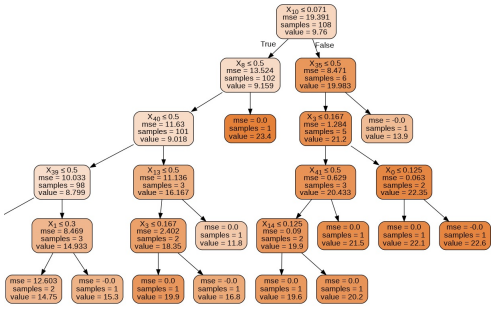

Figure 1: Visualize Decision Tree

technique called Bootstrap Aggregation, commonly known as bagging. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees [12].

3.3.4 MLP Regression

A MultiLayer Perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable [13].
In order to replicate what is described in the main paper [8] where it is used the MLP Regressor, I would mention also this Neural Network regressive model. The parameters used are the following: number of hidden layers = 10, learning state start from 0.01 with constant rate, momentum = 0.75, max iteration (epochs) = 150.

3.3.5 LSTM

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making

predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications [14].
The parameters used are the following: lags = 2, the optimizer choose is Adam optimizer with a learning rate = 0.000005 and the loss for the model is the Mean Squared Error. The model layers start with a BatchNormalization layer, the second layer has 512 neurons of LSTM, the third layer is another BatchNormalizer, the fourth layer is a Dense layer with 128 neurons, the fifth and sixth layers are other two BatchNormalizer and the output layer is a Dense layer with a single neuron with a hard-sigmoid activation. It is trained with early-stopping (validation_loss) and maximum epochs set on 500.

4 Results and Evaluation

4.1 Evaluation Order

The standard index of evaluation of models are provide as follow:

- R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- MAE is a measure of difference between two continuous variables

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

- MSE measures the average of the squares of the errors, the average squared difference between the estimated values and the actual value.

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

- $RMSE$ is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

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

4.2 Results of Models

Plotting results:

1. Figure 2: Polynomial Regression of both X fitted plot
2. Figure 3: Decision Tree Regression of both X fitted plot
3. Figure 4: Random Forest Regression of both X fitted plot
4. Figure 5: MLP Regression of both X fitted plot
5. Figure 6: LSTM of train and test sets fitted plot with all variables

Table 2 and Table 3 shows the results of R^2, MAE, MSE and $RMSE$ of all models, Table 2 with all variables fitted and Table 3 with only Hour variable fitted.

Results comparison plot:

1. In Figure 15 it is possible to visualize the R^2 of models used;
2. In Figure 16 it is possible to visualize the MAE of models used;
3. In Figure 17 it is possible to visualize the MSE of models used;
4. In Figure 18 it is possible to visualize the $RMSE$ of models used;

Table 2: Evaluation of Models with all variables sorted by R^2

Model	$R2$	MAE	MSE	$RMSE$
LSTM			0.018329	
Polynomial Regression	0.742658	1.683098	3.916365	1.978981
Random Forest Regression	0.518525	2.288580	7.327352	2.706908
MLP Regression	0.499408	2.392366	7.618276	2.760122
Decision Tree Regression	0.244945	2.753704	11.490833	3.389813

Table 3: Evaluation of Models with Hour variable sorted by R^2

Model	R^2	MAE	MSE	$RMSE$
Polynomial Regression	0.749250	1.660185	3.816052	1.953472
MLP Regression	0.619040	2.064076	5.797664	2.407834
Random Forest Regression	0.606292	2.080489	5.991664	2.447788
Decision Tree Regression	0.603257	2.084259	6.037847	2.457203

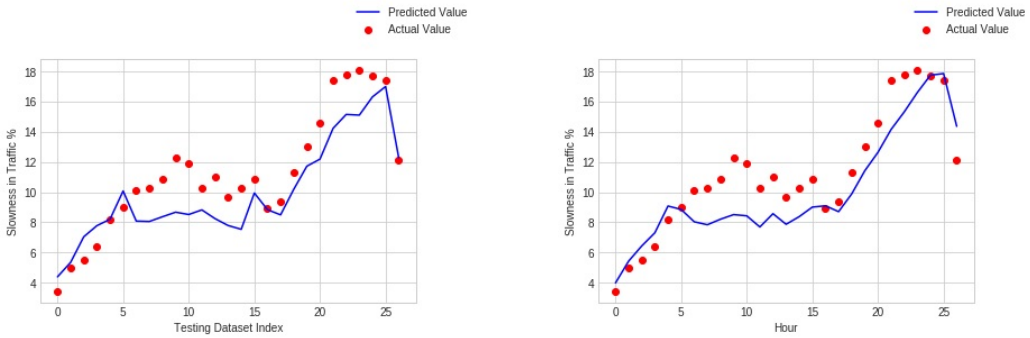

Figure 2: Polynomial Regression

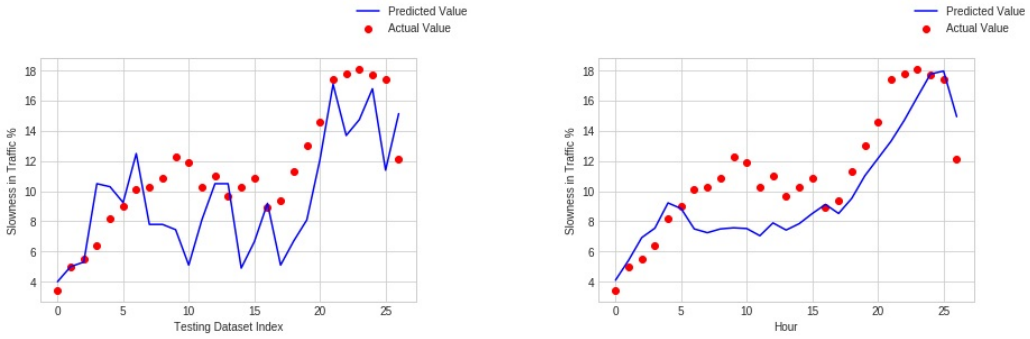

Figure 3: Decision Tree Regression

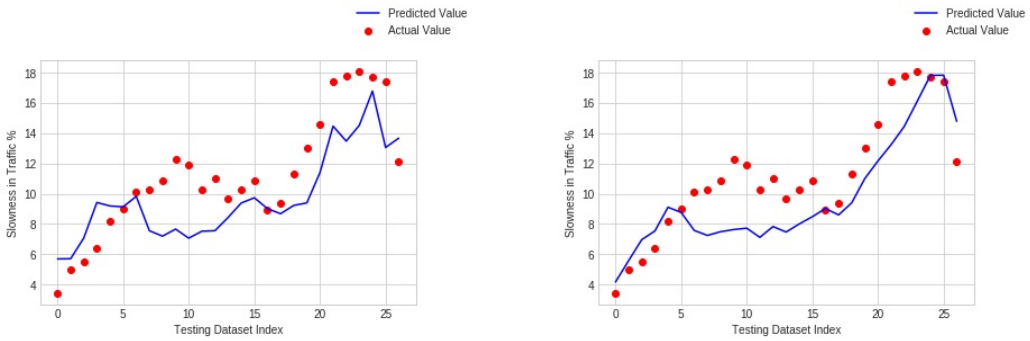

Figure 4: Random Forest Regression

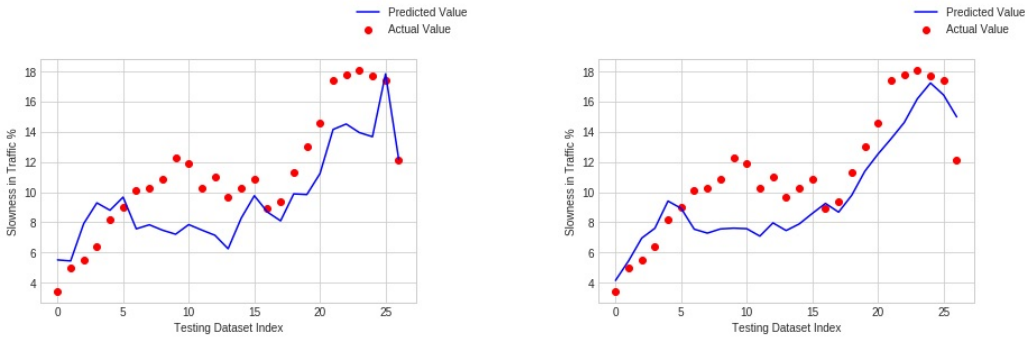

Figure 5: MLP Regression

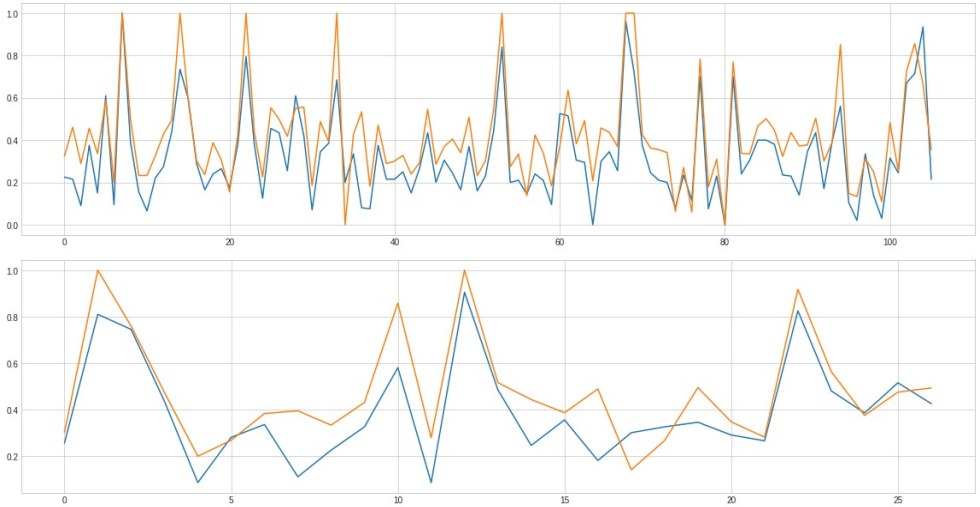

Figure 6: LSTM train and test

5 Conclusion

In this project, it is evaluated the prediction of the behavior of urban traffic in the city of São Paolo. The problem of traffic involve every city. The air pollution, the price of oil and the time spent in the car are some of the problems related to the congestion referred to the urban traffic. After some transformation of the dataset, it is proposed, three different regression models and two Neural Networks. The best one result is achieved by the Polynomial Regression with 0.75 of R^2, 1.66 of MAE and 3.81 of MSE fitted only with the X vectors which contain the Hour independent variable. With different results, it is another time the Polynomial Regression the best model with the X vectors which contain all independent variables. I also try with the same parameters described in the main paper [8] with MLP Regression, without using Fuzzy-Sets, but the best results are achieved by the LSTM model with a MSE of 1.83%. With this kind of analysis, I conclude that, in general, the best regressive model continues to be the Polynomial Regression and the best Neural Networks model is the LSTM.

6 Practical Suggestions

Prediction and estimation of traffic behavior could have an impact on driver assistant and autonomous driving which are in the nearly future. As mentioned above, even if there is an increase in available data, the main problem remains the homogenization of the same and the scalability of the models that are used, be they neural networks, regressive or classification models.

It is known that more data in that a model is fitted more accuracy it's possible to achieve and avoid the problem of overfitting. A higher collection of values could be interesting to apply that's kind of prediction in a real scenario. The hyperparameters tuning is not used in this project because of the few data available and, to this, arbitrary, it is chosen to use default sklearn choice of parameters for what concern the regressive models, while for what concern the Neural Networks the parameters is choosen arbitrary.

7 Appendix

Figure 7: Variables count

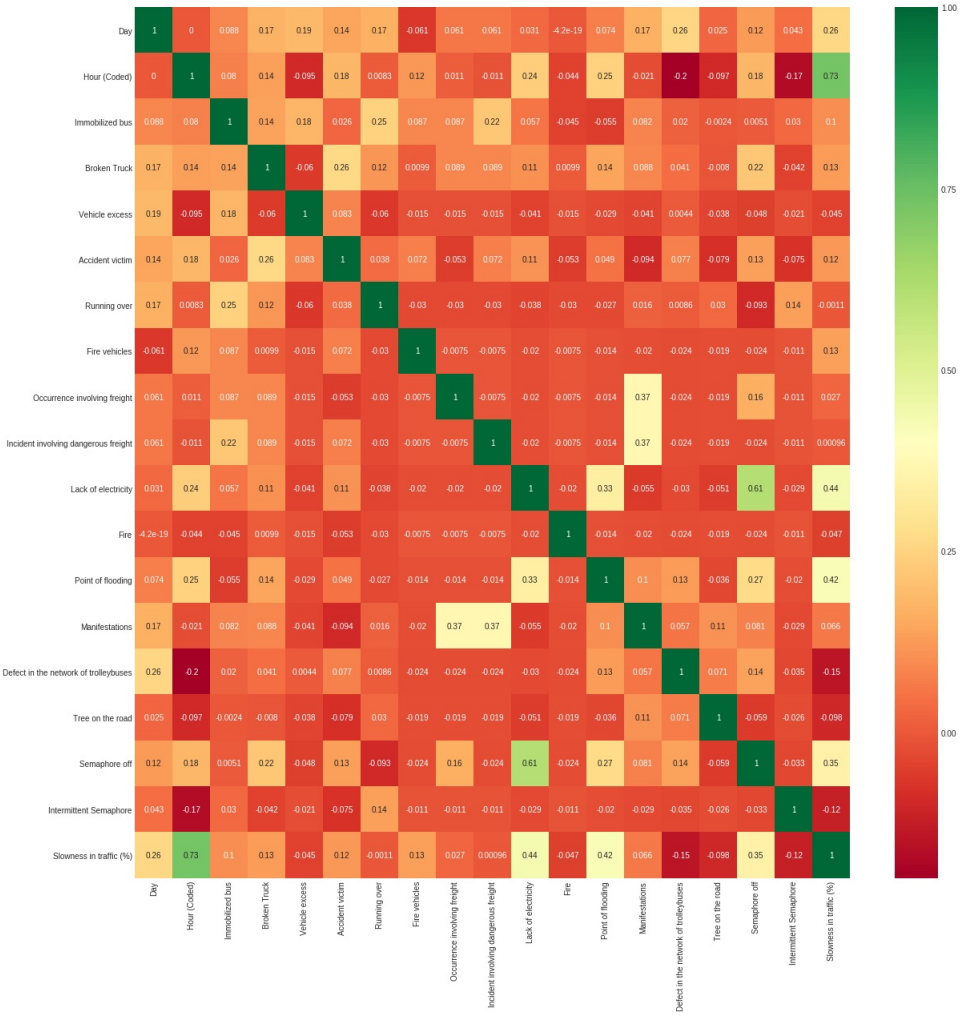

Figure 8: Correlation

	count	mean	std	min	25%	50%	75%	max
Day	135.0	3.000000	1.419481	1.0	2.0	3.0	4.00	5.0
Hour (Coded)	135.0	14.000000	7.817890	1.0	7.0	14.0	21.00	27.0
Immobilized bus	135.0	0.340741	0.659749	0.0	0.0	0.0	1.00	4.0
Broken Truck	135.0	0.874074	1.102437	0.0	0.0	1.0	1.00	5.0
Vehicle excess	135.0	0.029630	0.170195	0.0	0.0	0.0	0.00	1.0
Accident victim	135.0	0.422222	0.696116	0.0	0.0	0.0	1.00	3.0
Running over	135.0	0.118519	0.346665	0.0	0.0	0.0	0.00	2.0
Fire vehicles	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Occurrence involving freight	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Incident involving dangerous freight	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Lack of electricity	135.0	0.118519	0.504485	0.0	0.0	0.0	0.00	4.0
Fire	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Point of flooding	135.0	0.118519	0.712907	0.0	0.0	0.0	0.00	7.0
Manifestations	135.0	0.051852	0.222554	0.0	0.0	0.0	0.00	1.0
Defect in the network of trolleybuses	135.0	0.229630	0.818998	0.0	0.0	0.0	0.00	8.0
Tree on the road	135.0	0.044444	0.206848	0.0	0.0	0.0	0.00	1.0
Semaphore off	135.0	0.125926	0.464077	0.0	0.0	0.0	0.00	4.0
Intermittent Semaphore	135.0	0.014815	0.121261	0.0	0.0	0.0	0.00	1.0
Slowness in traffic (%)	135.0	10.051852	4.363243	3.4	7.4	9.0	11.85	23.4

Figure 9: Descriptive Analysis

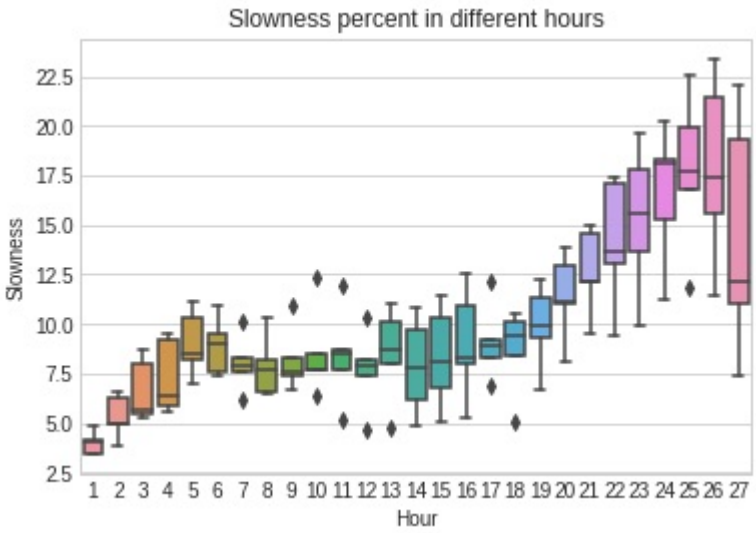

Figure 10: Slowness By Time

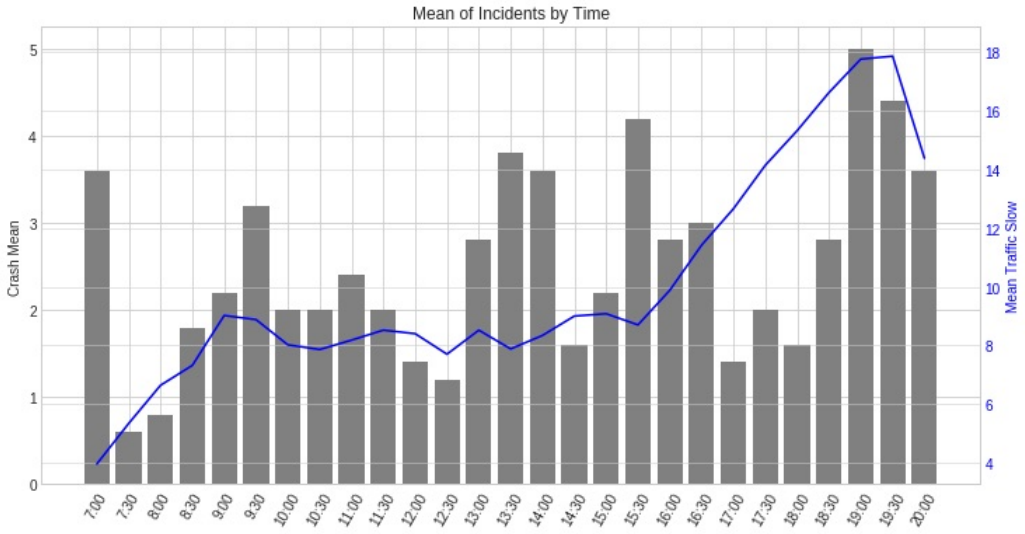

Figure 11: Mean of Incidents by Time

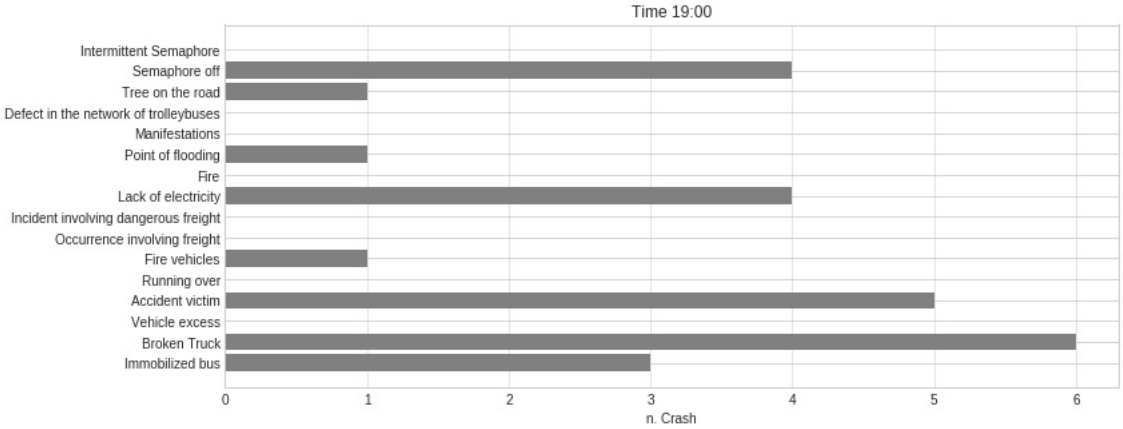

Figure 12: Incidents at 19.00pm

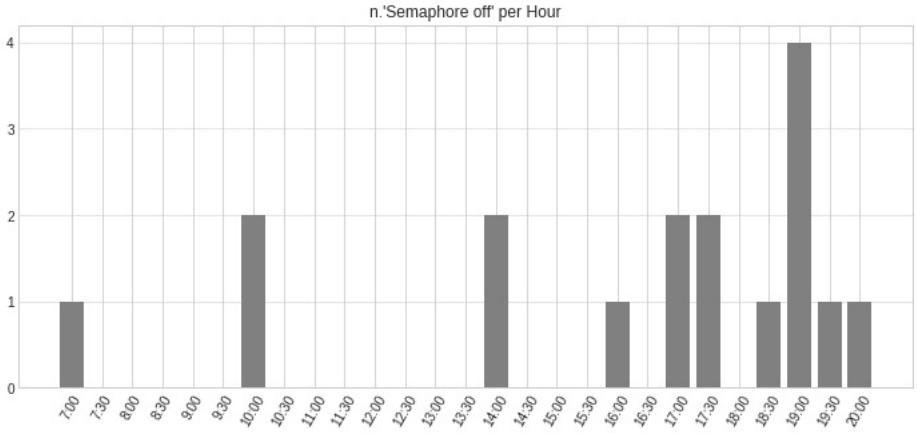

Figure 13: Semaphore off incidents

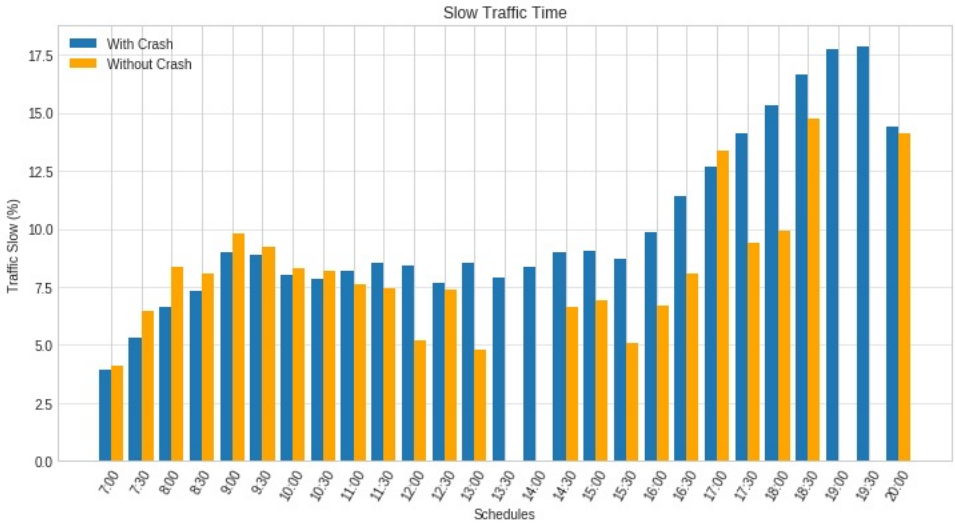

Figure 14: Slowness with or without

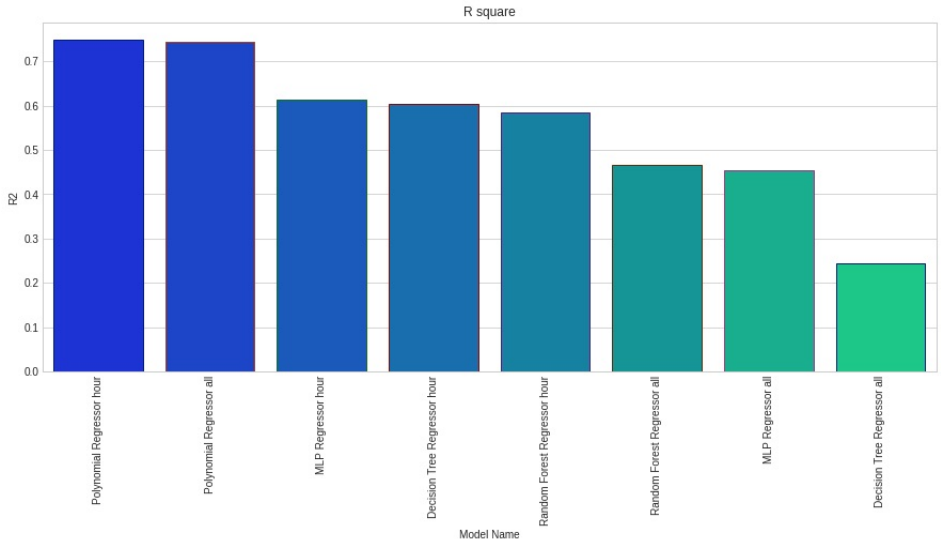

Figure 15: $R2$

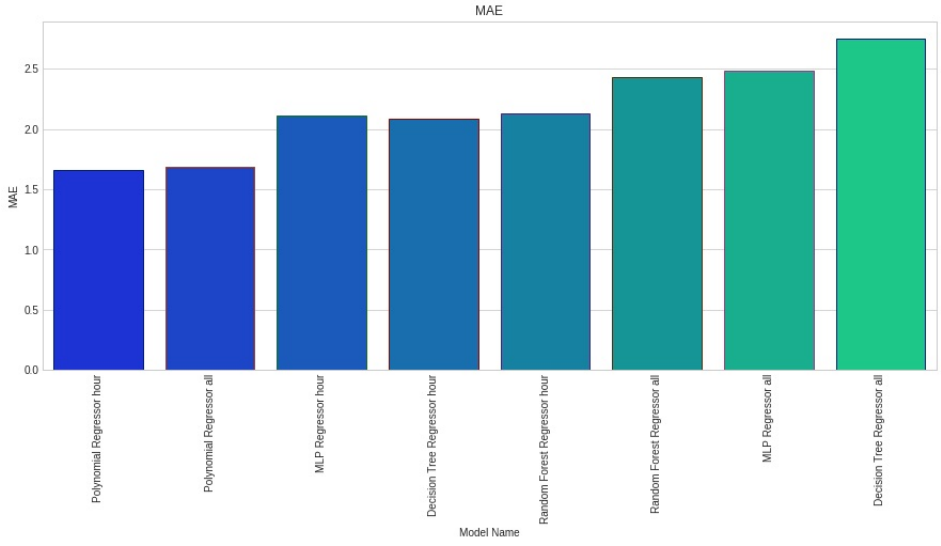

Figure 16: MAE

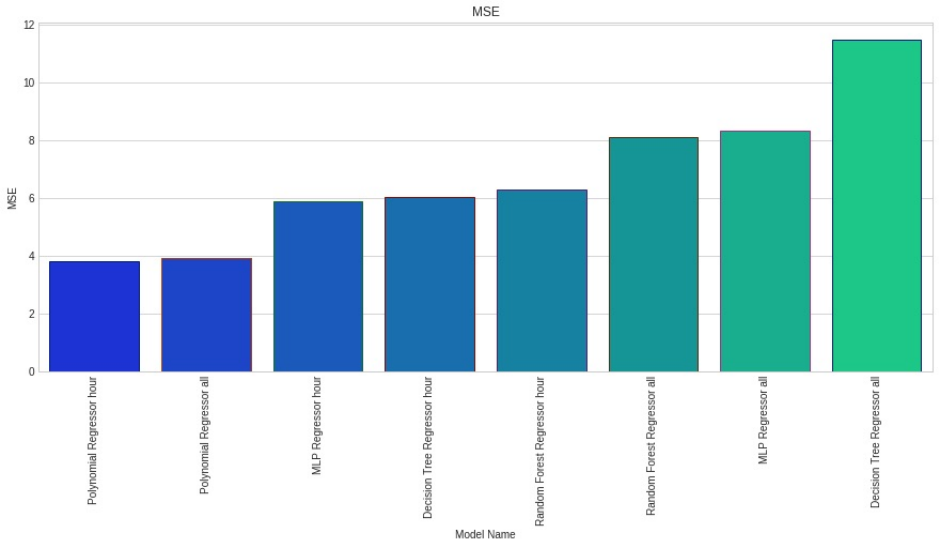

Figure 17: MSE

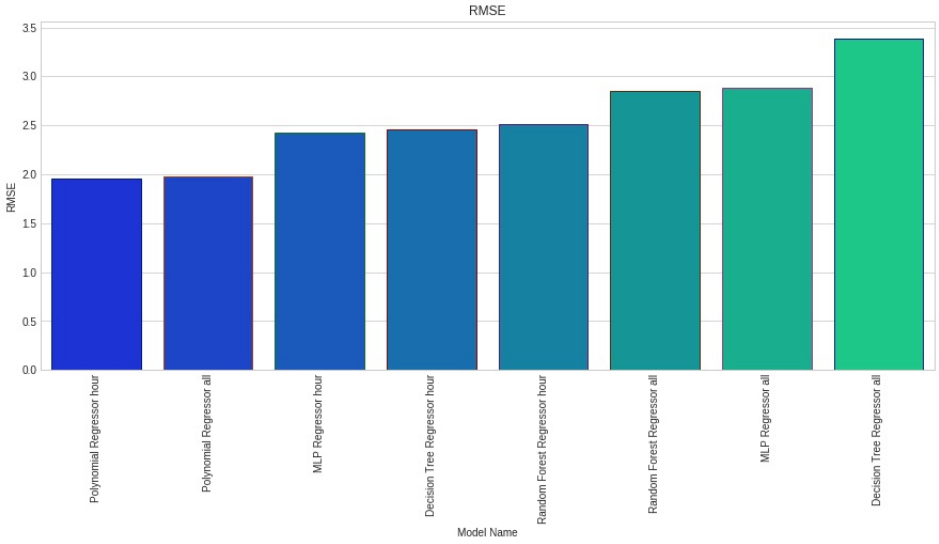

Figure 18: $RMSE$

References

[1] Smith, B. L. & Oswald, R. K. (2003), Meeting real-time traffic flow forecasting requirements with imprecise computations, Computer Aided Civil and Infrastructure Engineering, 18(3), 201– 13.

[2] Kerner, B. S. (2004a), Three-phase traffic theory and highway capacity, Physica A, 333, 379– 440.

[3] T. Gindele, S. Brechtel and R. Dillmann, "A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments," 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, 2010, pp. 1625-1631.

[4] B. Pan, U. Demiryurek and C. Shahabi, "Utilizing Real-World Transportation Data for Accurate Traffic Prediction," 2012 IEEE 12th International Conference on Data Mining, Brussels, 2012, pp. 595-604. doi: 10.1109/ICDM.2012.52

[5] Bryan E. Porter, Kelli J. England, Predicting Red-Light Running Behavior: A Traffic Safety Study in Three Urban Settings, Journal of Safety Research, Volume 31, Issue 1, 2000, Pages 1-8, ISSN 0022-4375, https://doi.org/10.1016/S0022-4375(99)00024-9.

[6] G. Gualtieri, M. Tartaglia, Predicting urban traffic air pollution: A gis framework, Transportation Research Part D: Transport and Environment, Volume 3, Issue 5, 1998, Pages 329-336, ISSN 1361-9209, https://doi.org/10.1016/S1361-9209(98)00011-X.

[7] Anil Namdeo, Gordon Mitchell, Richard Dixon, TEMMS: an integrated package for modelling and mapping urban traffic emissions and air quality, Environmental Modelling & Software, Volume 17, Issue 2, 2002, Pages 177-188, ISSN 1364-8152, https://doi.org/10.1016/S1364-8152(01)00063-9.

[8] Ferreira, R. P., Affonso, C., & Sassi, R. J. (2011, November). Combination of Artificial Intelligence Techniques for Prediction the Behavior of Urban Vehicular Traffic in the City of São Paulo. In 10th Brazilian Congress on Computational Intelligence (CBIC) - Fortaleza, Brazil. (pp.1-7), 2011.

[9] Sassi, R. J., Affonso, C., & Ferreira, R. P. (2011, August). Rough Neuro-Fuzzy Network Applied to Traffic Flow Breakdown in the City of São Paulo. In Management and Service Science (MASS), International Conference on (pp. 1-5). IEEE, 2011.

[10] Affonso, C., Sassi, R. J., & Ferreira, R. P. (2011, July). Traffic flow breakdown prediction using feature reduction through rough-neuro fuzzy networks. In Neural Networks (IJCNN), The International Joint Conference Neural Networks (pp. 1943-1947). IEEE, 2011.

[11] https://en.wikipedia.org/wiki/Decision_tree_learning

[12] https://medium.com/datadriveninvestor/random-forest-regression-9871bc9a25eb

[13] https://en.wikipedia.org/wiki/Multilayer_perceptron

[14] https://en.wikipedia.org/wiki/Long_short-term_memory