**Andre Piccolo – 300347025**

**Project – Traffic Volume Prediction**

1. **Introduction and discovery**

This dataset is related to the traffic volume between Minneapolis and Saint Paul in Minnesota (USA), through route 94 W.



This dataset has information about data traffic volume and some other variables that will be explored. The problem to solve with this model is predict the traffic volume based on some data and classify if the transit will be good or not. With this solution, it will be possible to send advice to drivers suggesting the better time to use the interstate 94. This can be used in other routes with the appropriate inputs.

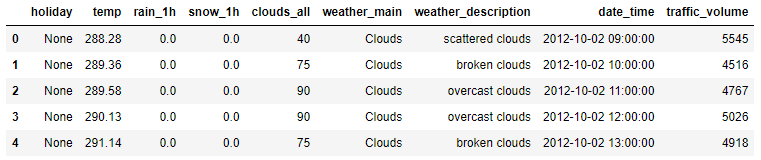
The traffic volume depends on some factors:

* Day of the week: This route can be used for commuters during the weekdays or used by families in weekends. Day of the weeks should give some insights about the typical use of this route and identify the target public to send advice.
* Hour of day: the hour should be a key factor to analyze rush hours. Usually, traffic in weekdays tend to be concentrated in the morning and evening when people are committing to/from work or school. This should be a key factor to predict traffic volume.
* Month: can be affect traffic volume if this route is used for students. In the middle of the year, breaks in classes can reduce the traffic. In December, this route would have less traffic volume because people use have a break in this period to celebrate Christmas and New Year with their families. Or even have a traffic increase in some days because people will travel to reunite and celebrate with their families.
* Weather: In days with snow or some other extreme weather can impact the traffic volume. People usually stay at home in severe weather conditions.

1. **Data Preparation**

The dataset was retrieved from <https://archive.ics.uci.edu/ml/datasets/metro+interstate+traffic+volume>.

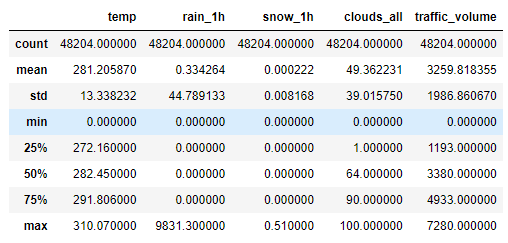
Sample of initial data:



Features description:

* Holiday: US national holidays and regional holidays. “None” for no holiday, otherwise will be the name of holiday.
* Temp: average temperature 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: percentage of cloud cover.
* Weather\_main: short textual description of the current weather
* Weather\_description: longer textual description of the current weather
* Date\_time: date and time of the data collected
* Traffic\_volume: Number of vehicles per hour reported

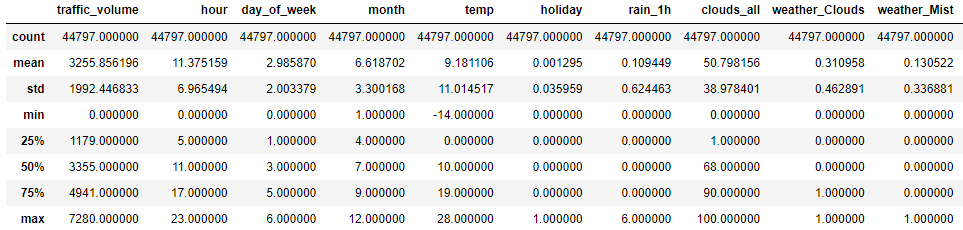
Summary description:

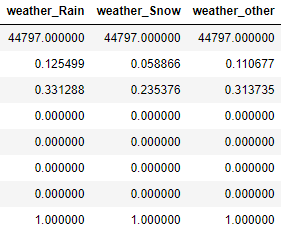


For data preparation, some action was taken:

* Converted temperature from Kelvin to Celsius and round values. Removed values with few entries.
* Split date\_time in date and time features
* Hour has only hour values, without minutes. For this reason, I just removed minutes and seconds from this feature and transformed this feature from object to int.
* Split date in day, month, and year. Year and day are not relevant for our model, so I just dropped year and change day of month to days of week (Sunday to Saturday). This will help to identify weekdays and weekends. Month was kept based in our previous analysis. Month and Day of week was converted from object to int.
* Holidays was converted to yes or no. Just few holidays in our dataset. So, this does not impact our model.
* Weather\_main had a lot of values with few repetitions. I aggregated them in a value “other” and keep only values with more than 2000 entries.
* Weather\_description does not have impact in our model, for this reason I just kept the weather\_main and dropped weather\_description with a full description
* For rain\_1h the values were rounded and change some outliers above 6mm
* Snow\_1h has a few amounts different from 0, so I just drop this column.
* Weather\_main was split in dummies features to convert from object to int

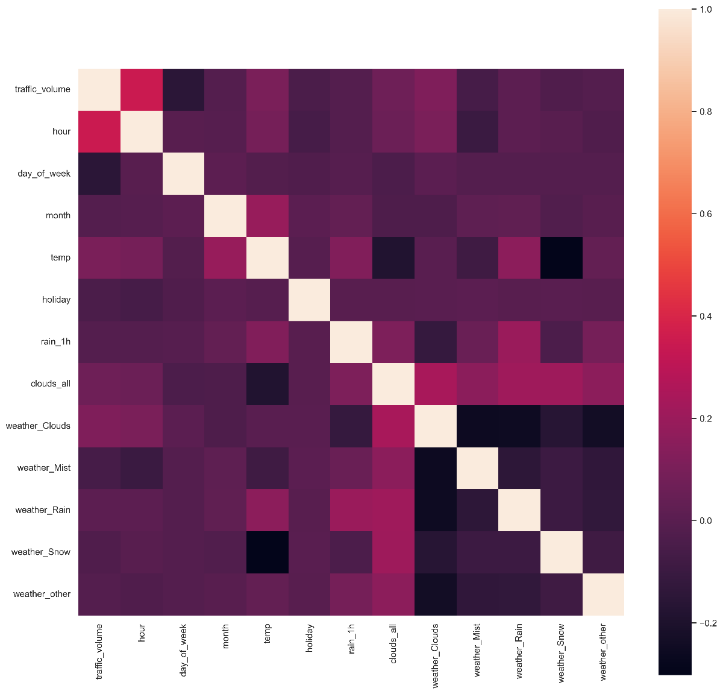
The final dataset:





1. **Model Planning**

Heatmap correlation of features



Heatmap correlation shows weak correlation between features, in this case is necessary to apply a polynomial transformation to improve correlation values. Scaling will not be applied in this model because range values is not distant between each other.

Regression models will be used to predict the traffic volume based on past data.

After chose a regression model to predict the traffic volume, a classification model will be applied to predict if the route is good to use or if it should be avoided. A column will be added in dataset to represent this information. Two different models will be generated to attend requests separately. First to get the traffic volume, and second to give advice to drivers.

1. **Model Implementation**

Three different estimators were implemented to check what will be fit better in model. Variance Threshold using threshold between 0.1 and 0.2, KBest feature selection with 10 features to select, and Random Forest estimator with parameters max leaf nodes 8 and number of estimators 100.

Feature selection for each estimator:

Variance Threshold:

A picture containing website

Description automatically generated

KBest:



Random Forest:

Diagram

Description automatically generated with low confidence

Different regressors was used to identify the best model. Models applied were: Linear Regression, Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor, Ada Boost Regressor, and CatBoost Regressor. To apply more models in an efficient way, a pipeline process was applied, a polynomial degree 2 was applied in dataset to improve the correlation. Resulting in a better analysis to choose the best model for this problem.

For the second part of the problem, an extra column was added to indicate if it is good to use the route or if is better to avoid it. Those predictions are based on the traffic volume that are presented in dataset. Assuming that until 80% of traffic volume available in dataset is considered a good volume and driver can use the route. Over than 80%, the model will predict that this route should be avoided.

Traffic volume value based on data collected:





For this second part, SVC model and KNeighbors Classifier models were used.

1. **Results Interpretation and Implications**

For first part to predict the traffic volume value.

Table below shows all results based in regressor model and estimator:

Table

Description automatically generated

Results shows that different estimators do not have a considerable impact on the model score, they are similar. Different models have a considerable impact in model score. CatBoost Regressor and Gradient Boosting Regressor are the best models for this problem, both are similar between each other. Linear Regression is the worst model in output. Model in position 17 was the selected model, CatBoost Regressor model using Variance Threshold as estimator give a precision near of 95% and value variance of 446 vehicles.

Best model prediction plot is showed below:

Chart, scatter chart

Description automatically generated

For the second part, classification model was used to predict if a driver should use the route or not. Two classification models were used, SVC Classifier and KNeighbors Classifier.

SVC classifier parameters and score result was:



This value is not good because is overfitting the model. For this reason, KNeighbors was used to try a better prediction without overfitting.



Some information about this model using confusion matrix and classification report.

Chart, treemap chart

Description automatically generated

A picture containing table

Description automatically generated

Information shows that overall accuracy is 96% and precision for predict a true negative is 91% and a true positive is 97%.

KNeighbors Classifier was the selected model.

1. **Out-Of-Sample Predictions**

For traffic volume prediction, a sample was created from dataset with better estimator (Variance Threshold).





Applying CatBoost Regressor model, the following result is showed:



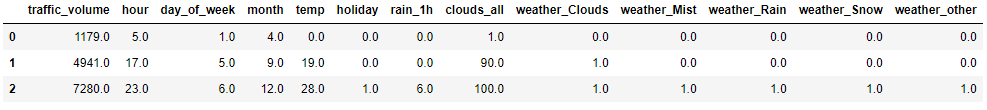
In the original dataset with the same values, traffic volume value is





Original value is 4941 vehicles, the predict value is 4599 vehicles. The accuracy of model is confirmed with those values, precision of almost 95% and range value error from 446 as predicted in the model analysis from item 5.

For traffic advise, the sample below was created.



Based on traffic volume, values over than 5185 vehicles it is considered to avoid the route, under that value, is considered a good route to use.

In this sample, the first two rows prediction model output should be to drivers use this route, the last row model output should be to avoid this route.

Running the best classification model that was KNeighbors Classifier, the result below is showed:

Text

Description automatically generated

1. **Concluding Remarks**

After some transformation, this dataset did not have a good correlation features value. To fix this problem, polynomial transformation with degree 2 was applied to improve the quality of dataset. CatBoost Regressor was the best regression model to predict traffic volume with approximately 95% of precision. KNeighbors Classifier was the selected model to identify if this route is good or not for drivers with 96% of accuracy. For a better prediction, a bigger dataset should be considered, and more data about new routes. For this route 94, both models can be used to predict traffic and help drivers.