Team

Prakhar Raj (Team Leader)

Rajneesh Jha

Syed Saad

Divyam

Abstract

This document is a brief overview of the solution approach for solving the problem statement of “Predicting the traffic volume for given time duration and climate conditions.”

**Sangam 2019 - ML Hackathon by IITMAA**

**Solution Approach**

**Solution Approach**

**Preprocessing Steps:**

1. In order to extract meaningful inferences from date\_time column, it was converted to datetime format.
2. Columns like rain\_ph and snow\_ph had numerical value 0 whereas their corresponding weather\_type or weather\_description had rain/snow details hence for such values mean imputation of the particular categorical variable i.e. (weather\_type or weather\_description) was used to replace such zeros.
3. In the given dataset for both train and test, in the is\_holiday column, every holiday was populated for only 0th hour hence as a preprocessing step we populated the same holiday to rest of the 23 hours for the day.

**Feature Engineering:**

Here is the list of features used and their brief description:

1. **Date Time features:**

These features are extracted from date\_time column.

* 1. hour: contains hour of every data record.
  2. month: contains month to which corresponding data record belongs to.
  3. day of week: tells what day of week i.e. (Monday, Tuesday, etc.) it was for every data record date.

1. **Boolean Features:**
   1. is\_weekend\_boolean: this feature tells if the data record corresponds to a weekend (Saturday and Sunday) day or not.
   2. is\_holiday\_boolean: this feature tells if the corresponding data record is from a day which was a Holiday or not.
2. **Target encoding/ Mean encoding features:**

From a mathematical point of view, mean encoding uses the concept of conditional averaging where **average of your target variable, is calculated conditional on each unique value of the feature**.

Such features give an intuition an average as to what was the trend of output variable with respect to each unique value of a categorical variable.

* **One level Target Encoding**: Only one categorical variable is taken at a time.
  + 1. is\_holiday\_mean\_encoding: this feature calculates conditional average of traffic\_volume for each record on being a holiday or not.
    2. day\_of\_week\_mean\_encoding: this feature calculates conditional average of traffic\_volume for each day of the week i.e. (Monday, Tuesday, etc.) in the past.
    3. weather\_type\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different values of weather type i.e. (Clouds, Clear, etc.) in the past.
    4. weather\_description\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different values of weather description i.e. (scattered clouds, sky is clear, etc.) in the past.
* **Two level Target Encoding:** Two categorical variables are taken at a time.

1. hour\_weather\_type\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different possible combination of values of hour and weather type i.e. (0th hour-Clouds, 1st hour-Clear, etc.)

This feature gives an intuition that at a particular hour and with a particular weather type what was the average traffic volume in the past.

1. hour\_day\_of\_week\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different possible combination of values of hour and day of week i.e. (0th hour-Monday, 1st hour-Monday, etc.)

This feature gives an intuition that at a particular hour and for a particular day of week what was the average traffic volume in the past.

1. month\_weather\_description\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different possible combination of values of month and weather\_description i.e. (January-scattered clouds, February-sky is clear, etc.)

This feature gives an intuition that for a particular month and for a particular weather description as in seasonality what was the average traffic volume in the past.

* **Three level Target Encoding**: Three categorical variables are taken at a time.

1. day\_of\_week\_hour\_weather\_type\_mean\_encoding: this feature calculates conditional average of traffic\_volume with respect to different possible combination of values of day of week, hour and weather\_type i.e. (Monday-0th hour-Clouds, Tuesday-0th hour-Clear, etc.)

This feature gives an intuition that for a particular day of week at every hour for a particular weather type what was the average traffic volume in the past.

1. **Shift Features:**

Shift or lag features are the classical way that time series forecasting problems are transformed into supervised learning problems where the approach is to predict the value at the next time (t+1) given the value at the previous time (t-1)

* 1. weather\_type\_mean\_encoding\_1lag: One level conditional average value of weather type is shifted by 1 hour which creates a new time series with 24 hours of lag values to predict the current observation of mean traffic volume.
  2. weather\_type\_mean\_encoding\_2lag: One level conditional average value of weather type is shifted by 2 hour which creates a new time series with 24 hours of lag values to predict the current observation of mean traffic volume.

1. **Raw features:**

Few raw columns that were considered as features and were used as is were:

* 1. temperature
  2. wind\_direction
  3. humidity
  4. visibility\_in\_miles
  5. dew\_point

1. **Other Features:**
   1. no\_of\_days\_till\_next\_holiday: this numerical feature tries to capture number of days remaining for the upcoming holiday which includes both weekends and festival. Trends in the dataset suggest that traffic volume is likely to go up just a few days before a holiday.
   2. no\_of\_days\_till\_next\_festival: this numerical feature tries to capture number of days remaining for the upcoming festival. Trends in the dataset suggest that traffic volume is likely to go up just a few days before a festival.

This was the exhaustive list of features that were considered for modelling but best set of features were selected using feature selection techniques.

**Feature Selection**

In order to decide best set of the features for the model various techniques were implemented, few of them are mentioned below:

1. Correlation metric
2. Forward feature selection
3. Group Variation Inflation Factor (Group VIF)
4. Feature selection using genetic algorithm

**Modelling**

We have used **ensemble technique** to improve predictive performance of our model as compared to a single model by bagging the predictions from two tree-based model algorithms i.e. Random Forest Regressor and LightGBM Regressor.

**Hyperparameter Tuning**

We have used several techniques to determine optimal set of parameters for learning algorithm:

1. GridSearchCV from scikit-learn.
2. Scikit-Optimize

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END OF DOCUMENT\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*