IITM Sangam Solution

Team dumb_terminals :

1. G. Kranthi Kiran
2. G. Mothy

Approach

- The aim of the problem was to predict the traffic_volume for indian metro city given features like the date-time, weather description, etc
- The data contained many duplicate time-indexes with different feature values but same traffic_volume.
- So we removed the duplicate time-indexes and then trained our model and while submission merged the same date values with same traffic_volume prediction.
- Baselined various statistical models like ARIMA, Prophet but regression models proved better than them as we had many weather related-features which we can’t use in the statistical methods.
- Basic Date-Time features + some features were created.
- LightGBM proved best from all the models test locally.
- Evaluation was done using the metric given on the competition page.

Quality Checks/Errors Found

- The data contained many duplicate time-indexes with different feature values but same traffic_volume..
- So we removed the duplicate time-indexes and then trained our model and while submission merged the same date values with same traffic_volume prediction.

Data Preprocessing

- Removed the duplicate time-indexes from both train and test sets and used the remaining data to train and test our model.
- The predictions for the removed indexes would be same as we have unique date-time feature which would let us merge on that to get the predictions for the duplicate time-indexes.

Feature Engineering

- Date-Time Features
  a. Hour
  b. Day of week
  c. Day of month
  d. Month
  e. Year
  f. Week Number
  g. Is_month_end
  h. Is_week_end
- Date_Time was encoded in a way to capture the temporal sense of data or else LGB would normally treat this as a regression problem.
- Target Based Aggregate Features like mean, std, min, max, didn't increase the score so didn't include.
- Target Mean Encoding for Categorical features didn't increase the score too, so used LabelEncoding for Categorical Features.

Model Choice Explanation

- My past experience in using Boosting methods like LightGBM and XGBoost made me take the decision of using it as a model and as it Gradient Boosting Implementation.
- I tried other models linear and probabilistic but lightgbm gave better results than others both locally and on LB.
- As lightgbm trained and fitted faster I had more time to experiment stuff and get results quicker so chose LightGBM over Xgboost due to Xgboost’s higher training time.

Important Features

1. Hour
2. Day of Week
3. Day of Year
4. Date_Time
5. Day of Month
6. Temperature
7. Week of Year
8. Wind Direction

Expected Error for submission

- Due to limited number of submission per day I didn’t have time to properly tune the parameters of LightGBM model.

  Metric according to the competition page i.e max(0, 100-rmse):

Train Score :	99.7584
Validation Score :	99.8087
Public LeaderBoard Score :	99.9763

Thank You.