Team : The Powerpuff Girls

Ramisa Alam, Mushtari Sadia, Mashiat Mustaq, Date : 25-04-2021

1 Descriptive Analysis

The challenge was to predict the hourly taxi trips of different zones in a city. For that, we analysed the data (weather, neighbours, trip counts) to find out their trend, property and relevance.

- The ACF and PACF plots for trip count showed higher correlation with recent previous hours and has almost a periodic tendency of 24 hours. The plots are given in Figure 1
- Two peak traffic periods (06:00:00-10:00:00 & 16:00:00-20:00:00) were noticed in week days and one (16:00:00-20:00:00) in weekends.
- Neighbour zones are not highly correlated with each other. In fact in many cases they showed low correlation.
- Correlation analysis of weather parameters showed that most of them had very little correlation with traffic. The most correlated features we could find were **snow, snow depth, fog, heavy fog**.
- Some zones had higher average traffic than others.

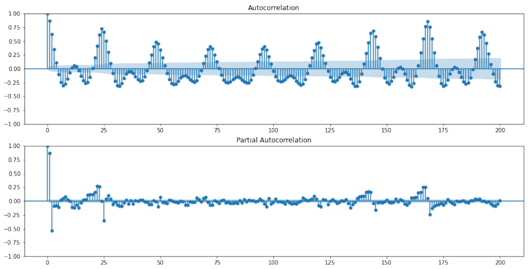

Figure 1: ACF and PACF of hourly trip count - Zone 1

2 Feature Selection

We used the insights obtained in EDA to select the following features for our models.

1. **Traffic demand in the past:** From our findings in the ACF and PACF plots, we used traffic demands in the last 24 hours for indicating today's demand and traffic demands of the previous 2 hour window for the last 30 days for indicating that hour's overall demand.
2. **Weekday/Weekend:** By plotting traffic against day-of-the-week, we observed higher traffic on weekdays. So we added a feature indicating whether the given day is a weekday.
3. **Peak hour:** Based on the peak hours mentioned above we created a new feature by combining "Weekday" and "hour" features.
4. **High Traffic Zone:** Based on our analysis, we added a feature indicating whether the given zone usually has heavy traffic.
5. We tried using the weather parameters most correlated to traffic (snow, fog etc.). But none of these impacted the performance of our models much. So we decided to not include any of them as features.
6. The given information on neighbours data did not provide any useful insight. So we did not use neighbour data as features.

3 Models and Parameters

We used the first four months of the dataset as training data, rest was used for validation. We also cross validated our model using special cross validation techniques for time series models. Initially, we trained some time-series forecasting models such as: ARIMA, Prophet etc. but the results obtained were unsatisfactory. So we trained the data on regression models such as Xgboost, LightGBM, Multilayer Perceptron(MLP) and Random Forest. Among these models the performance of MLP and Random Forest were great with both yielding satisfactory MAE. Even though LightGBM wasn't the best in terms of accuracy, its speed helped us with testing.

Along with trying various models, we also tried parameter tuning with LightGBM, Xgboost, MLP and Random Forest by increasing the number of estimators and max iterations. Increasing these further tended to overfit the model. Finally we used an ensemble of **Xgboost, LightGBM, Multilayer Perceptron(MLP) and Random Forest** with **Voting Regressor**.

4 Results & Discussion

Model	**Tuned Parameters and Other Techniques**	**Mean Absolute Error**	**Findings**
Voting Regressor Ensemble	Models: LGB,XGB, MLP and Random Forest	**14.63**	Best Performance
LightGBM	Larger max bin, smaller learning rate	15.37	Very fast, useful for testing. Accuracy satisfactory.
Xgboost	Increased number of estimators	15.17	Performed well.
Multilayer Perceptron	Increased max iteration and hidden layer dimension	15.05	Performed very well. However, very slow.
Random Forest	Increased number of estimators	15.19	Performed very well. However, very slow and high number of estimators consumes too much memory.
Prophet	Default	>Baseline	Poor performance
Decision Tree Regressor	Changed max depth	>Baseline	Poor performance
Support Vector Regression	Standard Scaling	>Baseline	"
Ridge Regression	Default	>Baseline	"
Gradient Boosting Regressor	Default	>Baseline	"

Our best performing model yielded an MAE score of 14.63. The performance of other models are described in the above table.

From the experiments, the observations that proved to be most helpful were: the traffic varies highly depending on time of the day, whether it is a weekday or weekend and traffic load in zones. Weekdays have more traffic than weekends. The time between 6 to 10 am and 4 to 8 pm showed more traffic in weekdays which corresponds to the start and end of office hours. Some zones showed more traffic than others which might be due to more commercial areas than the rest. Especially zone 44,43,24,31 showed an average trip count over 400. All these information resulted in a periodic tendency of the traffic which gave a better prediction.