Call Prediction Research Update

Xing Li
li64@mcmaster.ca

December 13, 2020

Abstract

This is a weekly update on the research progress of call prediction project. This week I read through the data set and the related documents, drafted the research plan and finished the autocorrelation analysis on the daily incident numbers.

1 Project Goal

Predict the number of incidents in the next few days based on the past trends during different pandemic phases.

To be more specific, I translate the goal as below.

- Target variable: average incidents number from day t to day t + n. n can be temporarily set as 6 so it is the average incidents number of a week.
- Features: To be discussed in the section of feature engineering.
- Loss function: May use mean square error (MSE) in the machine learning model.
- Performance metric: May use mean absolute percentage error (MAPE).

2 Data Set Overview

- 63 columns in total.
- 14 columns may be dropped immediately because all records have the same value. They have little predicting value. fdid, exp_no, confidencelevel, addressobtained, no_inci, e911_used (1 non-null value), resc_detl, app_supp, fatal_fs, export, export_dt, printed, nfirs4, alm_method.
- 8067 incidents in 296 days from '2020-01-01' to '2020-10-22'. The end date of the five periods are ['2020-03-16','2020-05-18','2020-06-18','2020-07-23','2020-10-18'], so the effective date span for analysis are 292 days in 5 periods.
- The meta data can be found in the file "VFRS Table Field Descriptions.xlsx"
- For column "mutl_aid", 14 means "no mutual aid" and 31 means "with mutl aid." "With mutl aid" are usually for large fires or fires outside the city.

3 Document Overview

The content experts made comparisons on the dimensions below: property type (8 categories), period (4 periods), incident type (medical/vehicle/fire/false fire calls etc.), district, year(no data), geographical location (no data). The first four dimensions might be used to design different predicting levels.

4 Research Plan

4.1 Predicting level

We need to experiment to find the optimal predicting level, but we can start from the top-level – the level of all incidents. The algorithm and the code can be

reused for different predicting levels. We may use design predicting level on three dimensions: property type, incident type and district.

4.2 Feature Engineering

Because we have only 292 days of effective data, the number of predictors should be less than 10. This means we may introduce new columns, such as the average incident number of the last 7 days, to consolidate the info from different columns.

Considering the dimension of period (stage), we may also need to create new columns to represent the period coverage for a given time span. For example, if in the last 30 days, 18 days are in period 2 and 12 days are in period 1, we may set 0.6 for column "percentage_period_2" and 0.4 for column "percentage_period_1".

Several rounds of feature selection are required via tree-based or rule-based machine learning models, such as random forest regressor.

4.3 Autocorrelation

Before consolidating time series data into one feature, we need to check autocorrelation first.

5 Autocorrelation

As different periods may ruin the autocorrelation, the autocorrelation lag is set at 30, which is less than the length of the shortest period. The result shows that there is no autocorrelation on a daily basis.

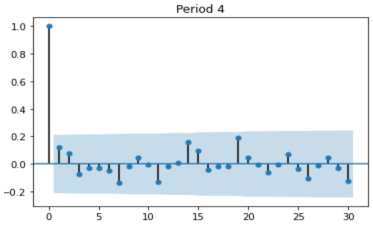

Figure 1: Autocorrelation graph for period 4