Call Prediction Research Update

Xing Li
li64@mcmaster.ca

January 8, 2021

Abstract

This is an update on the research progress of call prediction project. After I got the new data set, I've consolidated them into one data set, finished auto-correlation analysis, reshaped data and introduced 57 features. I also built up the model to analyze the feature importance and trained the model to reach an accuracy of 92.21%.

1 Data Consolidation

I listed the issues found in the consolidation process, as well as the disposals.

- The columns are different. I use the common columns in the first file.
- The date format is different. I write a function to convert data to the same format.
- Some data of year 2020 overlap. In that case, I use the data in the new data set.
- The data between Jan 7, 2009 and Jan 18, 2009 are missing. I remove all data before Jan 19, 2009.

In the end, a new file is generated which include all data available. The file size is 28.4MB

2 Auto-Correlation

I failed to extract the exact coefficients, but it does not block the following steps.

Because now we have enough data, the auto correlation shows significant relationship with 95% confidence intervals. Here period 0 is the period before pandemic which starts from Jan 19, 2009 and ends at Mar 16, 2020.

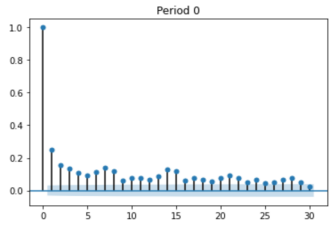

Figure 1: Daily Autocorrelation Graph for Period 0

Significant auto-correlation is found between d_0 *and* d_{-1}, d_0 *and* d_{-2}, d_0 *and* d_{-7}

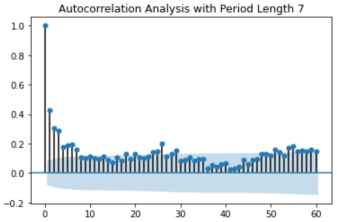

Figure 2: Weekly Autocorrelation Graph for Period 0

Similarly, significant auto-correlation is found between w_0 *and* w_{-1}, w_0 *and* w_{-2}, w_0 *and* w_{-3}

Significant auto-correlation is also found between m_0 *and* m_{-1}, m_0 *and* m_{-12}

No significant auto-correlation is found for annual data because of the small sample number.

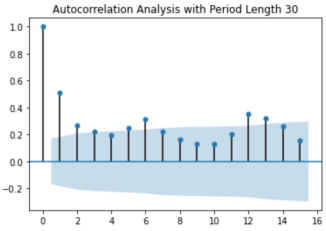

Figure 3: Monthly Autocorrelation Graph for Period 0

In summary, with the autocorrelation analysis, we need to add features of $d_{-1}, d_{-2}, d_{-7}, w_{-1}, w_{-2}, w_{-3}, m_{-1}, m_{-12}$ for prediction.

For d_{-1}, d_{-2}, d_{-7}, we need 6 features for 6 periods with value 1/0 to specify whether it belongs to either of the periods.

For $w_{-1}, w_{-2}, w_{-3}, m_{-1}, m_{-12}$ and target period, we need 6 features with a percentage value showing the percentage of days which fall in either of the periods. Here the target period is the period we want to predict. It is an adjustable variable, but in the following process, its length is set as 7 for days from d0 to d6.

We also need one feature to show whether it is influenced by the pandemic for re-sampling purpose.

In total we need to create 57 features, from which we will pick top features with cumulative importance exceeding 95

3 Modeling

3.1 Re-sampling

Because we have relatively fewer data in pandemic periods, the model may overlook the minority class to reach high accuracy. In order to avoid training with imbalance data, we have at least five solutions[1]. Using tree-based algorithms is one of them.

If we are going to use other models such as SVC, we need to develop some codes to address this concern.

3.2 Feature Importance

To reach cumulative 95% importance, we only need to include the top 8 features. However, I will still include 10 features as tg-1 and m-1_p4 shows pandemic impact. Here tg-1 is the percentage of days in the target period which falling in period 1. m-1_p4 is the percentage of days in the m_{-1} period which falling in period 4.

```
m-1:0.229;  cumulative 22.9%
w-1:0.174;  cumulative 40.3%
m-12:0.154;  cumulative 55.7%
w-2:0.134;  cumulative 69.1%
w-3:0.0964;  cumulative 78.7%
d-1:0.0806;  cumulative 86.8%
d-2:0.0656;  cumulative 93.4%
d-7:0.0499;  cumulative 98.4%
tg_1:0.00599;  cumulative 99.0%
m-1_p4:0.00108;  cumulative 99.1%
```

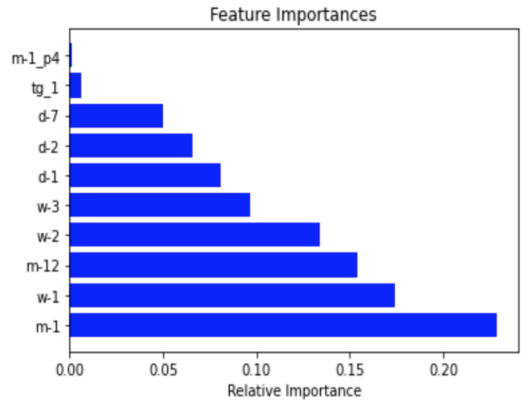

Figure 4: Top 10 important features

3.3 Training

With random forest regressor, we train the model with 80% of the data and test with the rest 20%. The average absolute error is 2.33 incidents, and the accuracy is 92.21%.

The result will serve as the base line for the further fine tuning. The accuracy of over 90% looks good, but I still have concerns on the imbalanced data. After looking into the details, I find that all test data are not from pandemic period. Some re-

sampling is needed to check the performance at pandemic period. This will be included in the things to do next.

4 Next to do

- I will do re-sampling first to see whether the model works for the pandemic data.
- I will use scaler to see whether it helps to improve accuracy.
- I will use grid search with different models and different hyper parameters to look for the better model for our case.
- I will predict at lower levels of different dimensions (property type, incident type) to see whether it's better to predict at lower level for some dimension.

References

[1] *How to handle imbalanced classes in machine learning*. [Online]. Available: `https://elitedatascience.com/imbalanced-classes`.