Project 1: Network Congestion Type Prediction

Given dataset from telecom firms has 78560 records with multiple features such as bytes used in various ways, timestamp when data collected, no. of subscribers, antenna direction & tile angle, cell range and vendor name. Objective is to build a model from the dataset to predict the right Congestion type.

Analysis of data is given below as summary.

Summary

1. Collected data are from 3 telecom firms (called vendors or ran_vendor) such as **ERICSSON, HUAWEI** and **NOKIA** in the month of Dec 2018.
2. In the dataset, almost all the variables are numerical except ran_vendor and Congestion_Type. There is no missing value in any column.
    a. "Congestion_Type" is the target variable or dependent variable with values such as 4G_BACKHAUL_CONGESTION, 3G_BACKHAUL_CONGESTION, 4G_RAN_CONGESTION and NC (No Congestion).
    b. "ran_vendor" is the telecom firm name, as mentioned above in point 1.
3. "cell_name" is numerical, but it specifies unique identity for the cellular network as given in problem statement. So this may not have much value towards predicting the target. Similarly as data collected on Dec 2018, "par_year" and "par_month" are same for all the records, so these two also do not give any value towards target. We can discard these 3 variables while training the model.
4. par_day, par_hour, par_min: these may give some value as which day and what time has some congestion or not.
5. From describe() we can see that many columns are right skewed such as subscriber count and many of the bytes_used columns.
6. We can see an almost equal amount of data collected from each vendor and also an almost equal amount of data towards each congestion type.
7. From heatmap of correlation, we can see there is no strong relationship between features.
8. As our dataset contains different units of data (such as bytes, timestamp, count, range etc.), we need to normalize the data to bring all into same scale.
9. Logistic Regression Model trained with default values, gave us ~77% accuracy, whereas with certain tuned values, it gave us accuracy of ~79%.
    a. Values such as solver='lbfgs', multi_class='auto' and C=0.1
10. However accuracy may not be a good performance measure in case of classification problems and especially when data is skewed. So we can see other measures such as Precision, Recall and F1-score.
    a. We can see the measure as given below.

Kishore Chandra Sahoo
kishore.cs@samsung.com

	precision	recall	f1-score	support
0	**0.76**	**0.77**	**0.76**	5849
1	**0.70**	**0.70**	**0.70**	6032
2	**0.84**	**0.78**	**0.81**	5818
3	**0.87**	**0.92**	**0.89**	5869
accuracy			**0.79**	23568
macro avg	0.79	0.79	0.79	23568
weighted avg	0.79	0.79	0.79	23568

- For NC, we got **87% Precision**, which means 87% are actually correct out of total predicted positives. And **92% Recall**, which says 92% are correctly predicted from total positives.
- For other classes, it's little at lower side.
- F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall. We can see F1-score value for each class, 70% - 89%.

Answer for Q.3:

Steps taken to improve the accuracy:

1. Normalizing the data, which brings all the data into same scale.
2. Giving tuned parameters rather than default values.

Answer for Q.4:

Conclusion:

1. Data collected only for a month for the classification, so we may need to collect more data on other months.
2. For congestion type prediction, Logistic Regression gave ~79% accuracy with 79%-89% F1-score.
3. Better algorithms might be there for this type of multiclass classification problem, which may give better performance as compared to Logistic regression.

Answer for Q.5:

1. Accuracy is not a good measure in case of classification problems, moreover when data is skewed.
2. Precision, Recall and F1-score can be better measures to judge this problem.
3. Other measures can also be included for better judgement, such as ROC, AUC, etc.

4. Also there can be better algorithms as compared to Logistic regression for this type of classification problem which may give better performance for the prediction.

Kishore Chandra Sahoo
kishore.cs@samsung.com