Machine Learning Engineer Nanodegree

Voice call quality prediction analysis

Sagar Chavan 7th Aug 2019

Proposal

Domain Background

This problem is related to telecom domain. Mobile communication is growing the at rapid pace and there are multiple service providers in the market who provides services voice, internet, messages. Now mobiles are used beyond the just voice calls like mainly for internet usages video streaming. Industry is now entering the 5G (5th generation) network services it is utmost important to pay attentions towards the quality of the services.

Overall 96% people are using mobiles phones worldwide and it is growing further [1]

When use telecom services is a need to todays world the quality of these services become main topic for both the parties that is for the customer as well as for service provider.

There are rules and regulations from TRAI related to quality of service located at TRAI website [2]

Data source is taken from Kaggle.

Problem Statement

Considering the growth in mobile usage and available service providers in the market there is a tough competition to catch the customer. Customer also looking for good quality services from the service provider. The quality of the services is different in different areas. It depends on many factors like service provider infrastructure, network type service provider providing, different service provider has different quality of service in the same area.

As part of this project we are going to predict the voice call quality based on give feature. This will help TRAI keep watch on service providers for their service quality. It can also help new customers which service provider is good in their area so they can choose based on it. It will help service providers as well to improve their service qualities.

Datasets and Inputs

This data is captured using customers feedback using TRAI MyCall App. Dataset having 2 files of data for the month of Apr2018 and May2018. Each record contains the columns Operator, call category (indoor, outdoor, travelling), network type, rating, voice call quality, latitude, longitude, state name.

For some records network type can be unknown. latitude and longitude can be -1 or 0 that means user did not disclosed their location while providing the feedback. It is going to be difficult to use the latitude and longitude as it is so it will be either converted to the specific area(city) or those locations will be converted to zones using clustering algorithm. Exact approach will be decided at the time of project implementation.

There are 38487 datapoints present in the dataset. Labels are structured like 'Call dropped' , 'Poor Voice Quality' and 'Satisfactory'. There are only 2 potential classes here one is poor quality and other is good quality.

Data set is quite imbalance if I go to multiclass classification. To avoid that I am considering it as binary classification so 'Call Dropped' and 'Poor voice quality' both mapping to same 'poor quality' class. So now dataset has 27729 datapoints belong class 'good quality' and 10758 datapoints belong to 'poor quality' class.

This dataset taken from Kaggle site [3]

Solution Statement

From the problem statement we can clearly identify that this problem is classification problem where we have to predict the voice call quality.

As part of solution we need to prepare/preprocess the data which should be best suited for our model. I will use the pandas and numpy lib for it. Then we have to design the model to predict the target label for that there are many available algorithms. I will mainly try decision tree, random forest and ensemble methods. Once models are ready we will evaluate the performance of the models based on

the evaluation metrics like F1 score to choose the best model. To design the model I will use scikit learn frame work

Benchmark Model

For such classification problem we can consider decision tree as benchmark model with default parameter. Whatever score we get from it is our benchmark and then will try to improve the it by tuning the parameters and identifying other model with tuned hyper parameters.

Evaluation Metrics

There are multiple evaluation metrics to evaluate the model performances like F1 score , ROC score. Here for this problem I am going to use the F1 score to evaluate the model performance.

F1 score is calculated based on the precision and recall values using below formula.

F1score= 2 * (precision * recall /(precision + recall))

Here precision and recall are calculated based on the confusion matrix with the below formula.

Precision = TP/(TP+FP)

Recall = TP/(TP+FN).

Project Design

To solve the above stated problem we need to follow the below steps as part of the solution.

1. I need to do the data exploration to see how many records are present in the data how many are falling in either categories (yes/no) and what is the percentage
2. I need to do data preparation and preprocessing like transform skewed features, normalize numerical features. We may need to do one hot encoding for non-numeric features.
3. Then I have to split and shuffle the data for training and testing set.

4. I have to transform latitude and longitude in to some other feature either to city or clustering those points in small clusters with unique ids and use it
5. Build models as mentioned in the solution statement.
6. Train and test models
7. Then evaluate model based on F1 score.
8. Choose the best model based on the F1 score
9. Try to improve the F1 score by tuning the hyperparameters
10. Conclusion

References

[1] https://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use

[2] https://main.trai.gov.in/telecom/qos

[3] https://www.kaggle.com/pranaysharma1108/real-time-voice-call-quality-data-from-customers