Team 14

Network Congestion in Telecom Industry

Index

Introduction

General Overview:

Network congestion has been one of the most significant issues faced by the modern telecom industry. It leads to a decrease in quality or complete failure of services because of over burdening of the network with data. Dealing with congestion is one of the most significant issue faced by the market players of telecom industry. Its solution involves analysing the data , possible prediction and countermeasures in accordance with the predicted situations.

Data:

- Data available for training this model comprised of data from various cell towers as well as from individual users
- It envolved number of bytes consumed segregated against various activities such as audio,video and other applications like weather, health prediction apps as well as time involved
- There were no missing data in the training or test set, thus simplifying overall preprocessing.

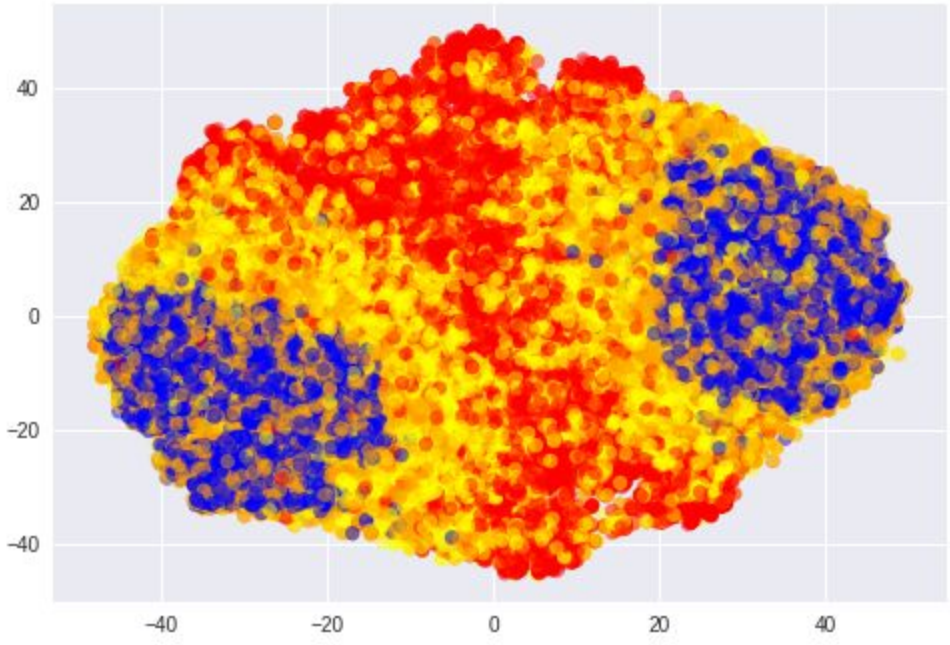
Target :

The target variable was the classification of congestion into :

1) NC (No Congestion)
2) 3G_BACKHAUL_CONGESTION
3) 4G_RAN_CONGESTION
4) 4G_BACKHAUL_CONGESTION

Preprocessing :

- Columns signifying year, month and cell name were removed because of the insignificance in prediction of the final class of congestion.
- Data was normalized to make it a better fit for possible models.

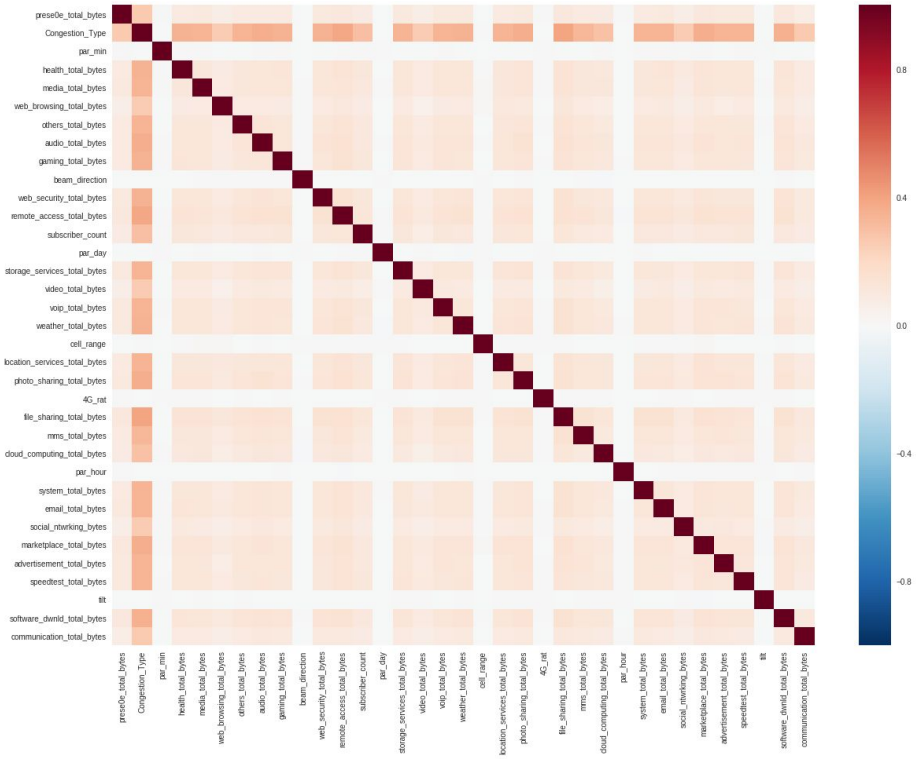


tsne

- t-Distributed Stochastic Neighbor Embedding was used to foresee possible dimensionality reduction, however, no definite correlation was observed

Initial Models:

1) **Naive Bayes** | MCC achieved : **0.66**
    - Initial intuition behind using Naive Bayes was the conditional independence of the available features. This was implied from the **correlation plot** between all the features.



As clearly visible from the plot the features were independent of one another to a large extent.

2) **Adaboost** : MCC achieved : **0.64**
   - Ada-boost classifier combines the weak classifier algorithm to form a strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with a selection of training set at every iteration and assigning right amount of weight in the final voting, we can have good accuracy score for the overall classifier.

     Its **MCC** was **lesser** than that of Naive Bayes specifically because of the data involved.

3) **SVC:** RBF kernel MCC achieved: **0.71**

Linear kernel MCC achieved : **0.72**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space, this hyperplane is a line dividing a plane into two parts wherein each class dimensional lay in either side.

4) **MLP** : **Solver** : Adam Optimizer **Activation :** Relu **MCC** Achieved : 0.72

Solver : LBFGS **Activation :** Relu, **Layer size** : 5 **MCC** Achieved : 0.718

Solver : Adam Optimizer **Activation :** tanh **MCC** Achieved : 0.7201

A **multilayer perceptron** (MLP) is a class of feedforward artificial **neural network**. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.

Measures to prevent overfitting:

Hybrid Approach:

In order to capture various features such as non-dependency (**Naive Bayes**) as well as subtle features from weaker algorithms (**SVC and MLP Neural Networks**) and various nuances from different activation functions such as ReLU **ensembling** was incorporated.

This hybrid approach ensured that better model with all the characteristics pertaining to different algorithmic models are captured. Voting Parameter in Ensembling was set to hard vote to ensure best possible prediction pertaining to all the classes.

The idea behind the **VotingClassifier** is to combine conceptually different machine learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses.

Hybrid Approach was calculated incorporating various models based on their individual accuracy but simultaneously ensuring that overfitting doesn't occur by keeping track of overall cross-validation accuracy.

Final Result:

The MCC achieved by the final model was : 0.7176

Confusion Matrix obtained from Final Hybrid Model :

```
[[ 894  132    0  173]
 [ 151  867  188    0]
 [   9  228  903    0]
 [ 107    0    0 1105]]
```