**Telstra Network Disruption**

Predict service faults on Australia's largest telecommunications network

**Introduction**

Telstra is Australia’s leading telecommunication and technology company, which builds and operates telecommunication networks and market voice, mobile, internet access, and other products and services.

Telstra is on a journey to enhance the customer experience in terms of providing the best network by developing a more advanced predictive model for service disruptions and it helps better serve its customers.

**Business Problem**

The main aim of this case study is to predict the severity of service disruptions on the Telstra network. Based on the information from their service logs (which contain information about the resources in a log file and this information allows for analysis of the resources).

so based on this information we have to form useful features which will allow us to predict whether the disruption in the network is a momentary glitch or a total interruption of connectivity.

Each row in the main dataset (train.csv) represents a location, location time point, and fault severity. They are identified by the ‘location’, ‘id’ ,’fault severity’.

“Fault severity” is the dependent feature it has three categories 0,1,2 (0 means “no-fault”,1 means “only a few faults” and 2 means “many faults” in Telstra network).

**Use of ML/DL**

Here I am using the ML model to solve this problem. The given problem is a multiclass classification task.

**Source of Data**

Telstra Network Disruption is a Kaggle competition that ended on 29th February 2016 and this was a recruiting competition. so the source of Data is provided by Kaggle for this competition.

**Existing Approaches**

Based on the Data obtained from the Kaggle competition, preprocessing is done on this data(such as, from categorical columns like event\_type,log\_feature, resource\_ type,severity\_type numerical values are extracted and a new data frame is formed which contains all numeric values) and then normalizing the train data and then trained the models such as knn, logistic regression, and random forest, from this model's best result is obtained from random forest

logistic regression test loss: 2.3113448041247957

Random forest test loss:0.9396115012417116

Knn test loss: 0.9602964342815254

**Improvements on Existing Approaches**

Improvement in the model is seen when new features are formed as part of feature engineering.

**Case1**

From the preprocessed data frame when the outliers are removed then the model's test losses are:

logistic regression test loss: 0.9392904812772841

Random forest test loss: 0.9460291372214836

Knn test loss: 0.9912332299535925

So after removing the outliers then normalizing the train data and then trained the models. Logistic regression test loss is decreased more than in other models.

**Case2**

From the given dataset One Hot encoding is done on columns “resource\_type” and “severity\_type”. 10 new dummy features are formed for resource\_type and 5 new dummy features are formed for the severity type column. After including these new features, removing outliers, and then normalizing the train data and then training the models. test loss of models are

logistic regression test loss: 0.8600103869755286

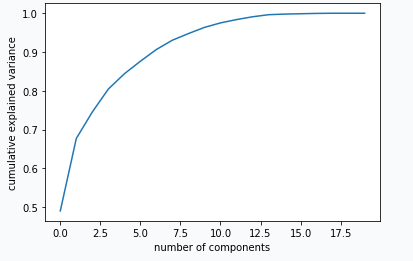
Random forest test loss: 0.9580244208360646

Knn test loss: 0.9553196273337947

From these models, logistic regression is showing a good result. when these new dummy features are included.

**Case3**

Applying PCA on final data and then plotting number of components vs cumulative explained variance (elbow method).we can see that first 15 components contains approx 98% of the variance and rest of the components has only 2% of the variance.



So when the last feature “location “ is removed then logistic regression loss decreased by(0.00004611717) and the test loss is 0.8599642698006663.

**Case4**

Using PCA new features are formed. when I used the first feature which has a maximum variance, in the final data frame, and after normalization train data model is trained. Here logistic regression test loss decreased by 0.03331709217

logistic regression test loss is 0.8266471776303964.

**EDA**

Once the final data frame is obtained after preprocessing. I checked the information of the data frame about any null values present, the type of data in a data frame, and the number of data in each column of the data frame.

so here 61839 data are present in each column, and all values are nonnull, and about the type of data here (total 6, int64 data columns) and (total 15, uint8 data columns).

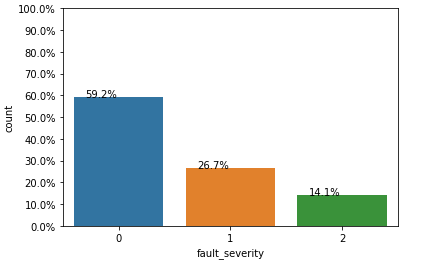
This is a “multiclass classification” problem(with unique classes are 0,1,2).

With “Imbalanced data” (stratify split is done to get the same distribution of data in train,cv and test data)

total 36597 class 0 data is present

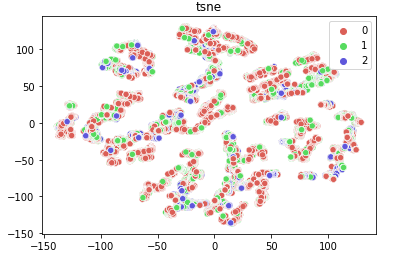
total 16524 class 1 data is present

total 8718 class 2 data is present



tSNE is used for the visualization of data distribution. If clusters are forming that means given features are useful features.

From tSNE at perplexity=200 in some region, class 0 clusters are formed at some region class 1 clusters are formed and in some region class2 cluster is formed with some overlapping.

Most classes are overlapping but in some regions, clusters of classes are formed

**First cut approch**

Based on the Data obtained from the Kaggle competition, preprocessing is done on this data(such as, from categorical columns like event\_type,log\_feature, resource\_ type,severity\_type numerical values are extracted and a new data frame is formed which contains all numeric values) and then normalizing the train data and then trained the models such as knn, logistic regression, and random forest, from this model's best result is obtained from random forest

logistic regression test loss: 2.3113448041247957

Random forest test loss:0.9396115012417116

Knn test loss: 0.9602964342815254

**Models Explanation**

**Code for knn**

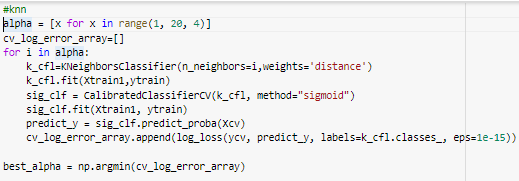
Knn models are trained with different values of n\_neighbors and calculate the log loss in each case. and stored all this log loss value in a list. From this list, the minimum loss value is found and its corresponding alpha value is used as best\_alpha(best value of nearest neighbor).

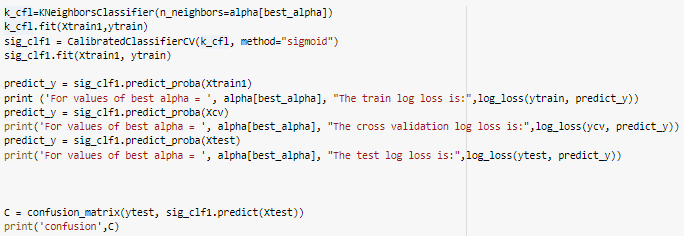
Using this best value of the hyperparameter, the model is trained and calculated logloss.

* So by the first cut approach when the model is trained the

test loss for knn: 0.9912332299535925

* When dummy features are introduced in the data frame after training the model test loss for knn : 0.9553196273337947. As compared with the logistic regression test loss is lesser than knn and random forest after adding new features.





**Random forest classifier**

Random forest classifier trained with different hyperparameters.

After training the model, the best values of hyperparameters are calculated

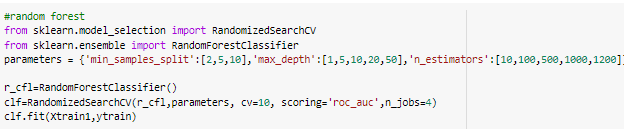
Using this best hyperparameters model is trained and calculates the loss value

* So by the first cut approach when the model is trained the

test loss for the random forest: 0.9396115012417116 .

which is the best compared to other models.

* When dummy features are introduced in the data frame. After training the model the test loss for random forest: 0.9580244208360646 .so when new features are added loss is increased. As compared with the logistic regression test loss is lesser than knn and random forest after adding new features.





**Logistic regression**

The logistic regression classifier is trained with different values of hyperparameter. then find the best values of hyperparameter, using this

hyperparameter model is trained and then calculated the log loss

* So by the first cut approach when the model is trained the

test loss for the logistic regression: 2.3113448041247957.

* When dummy features are introduced in the data frame. After training the model the test loss for logistic regression: 0.8600103869755286. As compared with the logistic regression test loss is lesser than knn and random forest after adding new features.
* From PCA when a new feature is added logistic regression loss is:0.8266471776303964. So when new features are introduced logistic regression loss is decreasing.



**Comparisons of models**

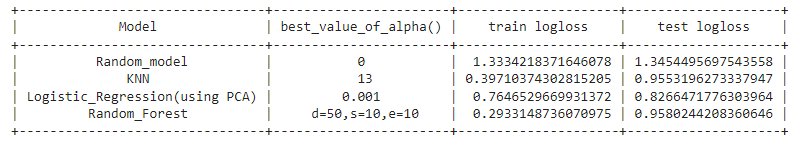
Models used are Knn, logistic regression, random forest classifier and random model is used as a base model.

So in the first cut approach, the random forest gives the best result in comparison with other models. but when dummy features of column “resource type” and “severity type” are added and outliers are removed and then train the model the logistic regression test loss is less as compared to other models.

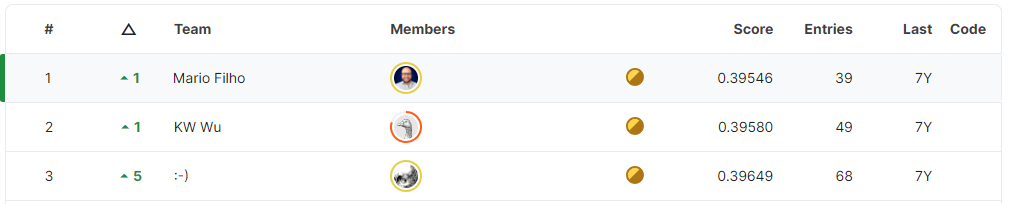
When using PCA a feature is added which as a maximum variance. logistic regression test loss is again decreased as compared to other models.

So in this problem logistic regression is showing the best result in comparison

with other models.



**Kaggle screenshot**

****

**Future work**

**References**

**Link:** [**[2110.12489] A Comprehensive Survey of Logging in Software: From Logging Statements Automation to Log Mining and Analysis (arxiv.org)**](https://arxiv.org/abs/2110.12489)

**Linkdein id**: https://www.linkedin.com/in/surabhi-m-v-511423218/

h