ROOT2AI Technology Private Limited

A report on the given classification model

Submitted by:

Shristi Bansal

Table of Content

Problem

Create a classification model based on the below mentioned dataset.

Split the dataset in train & test.

Dataset Link:
https://docs.google.com/spreadsheets/d/1DLL6BTXiHHsn1w9NvVi0BZbass0QU0RSvKEXtJlwfCM/

Meta- data

1. Text : contains text from blockchain domain

2. Target : target class

Approach

Taking a look at the dataset, it has 22704 rows comprising of text classified into 11 different target values.

After importing the dataset, step 1 is to check for missing values and handle them if any. The text field is then converted into string datatype so that we don't get any errors. Now, with regex, we must clean the dataset by replacing every character other than A-Z and a-z with ' '(space), converting all the text to lower case.

Using TFIDF, we will convert the text corpus into the vector form so that it can be easily fed to the model we train. Next step is to add the text corpus to an independent variable, say X and the target field to the dependent variable , say y.

The data is then split into the training data(X_train and y_train) and test data(X_test and y_test) in a ratio of 80:20.Then we create a classification model, feed training data to it and train it. After that this trained model will be used to predict the target for X_test.

Evaluation is then performed by constructing a confusion metrix. Accuracy is then monitored.

Model Interpretation

I have used Random Forest Classifier to train my model. A random forest is a meta estimator that fits a number of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

n-estimators = 100 is taken. It is the no. of trees in the forest.

Train & test accuracy score

	Training	Testing
Accuracy	97.65	63.81
Precision	98	66
Recall	98	64
F1 score	98	62

Limitations of the model

TFIDF:

- TF-IDF is based on the bag-of-words (BoW) model, therefore it does not capture position in text, semantics, co-occurrences in different documents, etc.
  For this reason, TF-IDF is only useful as a lexical level feature
- Cannot capture semantics (e.g. as compared to topic models, word embeddings)

Random Forest:

- Overfitting Risk - Although much lower than decision trees, overfitting is still a risk with random forests and something you should monitor.
- Biased towards variables with more levels - If your data has categorical variables with different levels of attributes this can be a big problem because random forest algorithm will favor those with more values which can pose a prediction risk.