**Capstone Project**

**Spam Email Identification**

Name: Isuru Sampath

Company: Dialog Axiata PLC

Email: isa.liyanarachchi@gmail.com

**Introduction**

Project target is to identify a given mail is a spam mail or not after reading the content. Content is given in the form of a CSV which contains how many numbers of times a word is present in an email.

**Data**

Please find the below sample of data to get an idea about the training data.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Email No. | the | to | ect | and | for | of | a | you | hou | Prediction |
| Email 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Email 2 | 8 | 13 | 24 | 6 | 6 | 2 | 102 | 1 | 27 | 0 |
| Email 3 | 0 | 0 | 1 | 0 | 0 | 0 | 8 | 0 | 0 | 0 |
| Email 4 | 0 | 5 | 22 | 0 | 5 | 1 | 51 | 2 | 10 | 0 |
| Email 5 | 7 | 6 | 17 | 1 | 5 | 2 | 57 | 0 | 9 | 0 |
| Email 6 | 4 | 5 | 1 | 4 | 2 | 3 | 45 | 1 | 0 | 1 |
| Email 7 | 5 | 3 | 1 | 3 | 2 | 1 | 37 | 0 | 0 | 0 |
| Email 8 | 0 | 2 | 2 | 3 | 1 | 2 | 21 | 6 | 0 | 1 |
| Email 9 | 2 | 2 | 3 | 0 | 0 | 1 | 18 | 0 | 0 | 0 |
| Email 10 | 4 | 4 | 35 | 0 | 1 | 0 | 49 | 1 | 16 | 0 |

As shown in the above table each row contains a single email. Each column indicates how many times each given word appears in the email. Final column indicates whether it is a spam email or not. There were total of 3001 words. There were 5172 total number of rows in the training data.

**Methodology**

First checked for any missing values and then preprocessed the data. Since there wasn’t any preprocessing required, then I checked the correlation. There wasn’t higher correlation between words and the spam status. Correlation of words with the prediction status was like below.

count 3001.000000

mean 0.020873

std 0.071619

min -0.271433

25% -0.030295

50% 0.015840

75% 0.068911

max 1.000000

So, I used Random Forest classifier and initially I took variables with correlation higher than 0.01. Then when trained observed that there is an overfitting issue. Then used a Randomized search with setting scoring method to ‘roc\_auc’. Then re-iterated the process for a X variables with correlation greater than 0.1.

**Results**

It was possible to get a better accuracy for X variables which are having correlation greater than 0.1.

|  |  |  |  |
| --- | --- | --- | --- |
| X variables selection | Total number of X variables | Accuracy Score | ROC AUC |
| Correlation > 0.01 | 2717 | 0.774 | 0.975 |
| Correlation > 0.1 | 492 | 0.892 | 0.985 |

These models were tested with the below best parameters for random forest model taken using randomized search.

{'criterion': 'entropy',

'max\_depth': 8,

'max\_features': 13,

'min\_samples\_leaf': 6,

'min\_samples\_split': 11,

'n\_estimators': 230}

RoC Curve for the final model is like below.

Chart, line chart

Description automatically generated**Conclusion**

For any classification problem if the accuracy score for the training data is 1 then it is overfitting issue. It would not perform well in test data the model overfits. Hence, I had to conduct a randomized search to get best model with better accuracy while minimizing the overfitting to the training data. Since this data has around 3000 columns had to minimize the number of columns or else model performance will severely degraded. Final model is based on 492 columns which has more correlation with the prediction. As a future development, one can try with minimizing the number X variables further.