FLIP ROBO

Internship 32

PROJECT REPORT
ON

FAKE NEWS DETECTION

SUBMITTED BY : SHAHLA M

SME : KHUSHBOO GARG

ACKNOWLEDGEMENT

I would like to express my sincere thanks of gratitude to my SME as well as “Flip Robo Technologies” team for letting me work on “ Fake News Detection ” project also huge thanks to my academic team “DataTrained”. Their suggestions and directions have helped me in the completion of this project successfully. This project also helped me in doing lots of research wherein I came to know about so many new things.

Finally, I would like to thank my family and friends who have helped me with their valuable suggestions and guidance and have been very helpful in various stages of project completion.

References:

I have also used few external resources that helped me to complete this project successfully. Below are the external resources that were used to create this project.

1. https://www.google.com/
2. https://scikit-learn.org/stable/index.html
3. https://github.com/
4. www.ijcseonline.org
5. https://www.analyticsvidhya.com/

INTRODUCTION

Business Problem Framing:

Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

Conceptual Background of the Domain Problem:

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

Review of Literature:

There are two datasets one for fake news and one for true news. We are combined both datasets using pandas built-in function. Machine learning data only works with numerical features so we have to convert text data into numerical columns. So we have to preprocess the text by steaming, lemmatization, remove stopwords, remove specialsymbols and numbers, etc.

Motivation for the Problem Undertaken:

We have to detect that the news are published on websites these are fake news or not. For this we analyze our data and then apply model to get better prediction regarding the news.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

We use Statistical techniques and analytics modeling in our projects, such as:

- describe() : use to calculate the statistical values that are mean, standard deviation, quantile deviation, minimum and maximum values.
- corr(): use to calculate the relation between feature variable with the target variable.

Data Source and their formats:

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news.

Importing necessary libraries:

```
#importing libraries
import pandas as pd
import numpy as np
import os
import seaborn as sns
import matplotlib.pyplot as plt
import nltk
from nltk.tokenize import regexp_tokenize
from nltk.corpus import stopwords
from sklearn.preprocessing import OrdinalEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix, roc_curve, roc_auc_score

from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import warnings
warnings.filterwarnings('ignore')
```

Importing the Datasets :

The imported datasets containing two different dataset namely fake news and true news . The fake news dataset contains 23481 rows and 4 columns .

Fake News Dataset

The fake news dataset contains list of fake news posted on various platforms including the title , subject ,content and published date. These are the datas collected from different sources.

```
fake_data  = pd.read_csv(r"C:\Users\sahal\OneDrive\Documents\Flip Robo Asgmnt\Fake News Project\Fake.csv")
```

```
fake_data
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016

23481 rows × 4 columns

True News Dataset

```
true_data= pd.read_csv(r"C:\Users\sahal\OneDrive\Documents\Flip Robo Asgmnt\Fake News Project\True.csv")
```

```
true_data.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

The True News dataset contains 21417 rows and 5 columns .

We insert one label column zero for fake news and one for true news then we can add a column named label and assign zero as values in Fake News dataset.

Then we combined both datasets using pandas built-in function.

The data set of the Fake News Project as show in the fig:

```
true_data=true_data.assign(label=1)
true_data.head()
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

```
true_data.shape
```

```
(21417, 5)
```

```
data=data.append(true_data,ignore_index=True)
data.head()
```

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0

```
data.shape
```

```
(44898, 5)
```

Data Pre-processing

In fake dataset there are 168 columns having NaN, so we drop all of them. In text preprocess we are cleaning our text by steaming, remove stopwords, remove special symbols and numbers, etc. After cleaning the data we have to convert this text data into numerical features using encoding technique.

Data Pre-processing

In fake dataset there are 168 columns having NaN, so we drop all of them. In text preprocess we are cleaning our text by steaming, remove stopwords, remove special symbols and numbers, etc. After cleaning the data we have to convert this text data into numerical features using encoding technique.

Data inputs-Logic-Output Relationships

The 'label' column is our target variable. In this problem label column is weak correlated with the title column. The text, date and subject columns have good correlation with the target variable.

Hardware and Software Requirements and Tools used

Hardware:

- Memory 16GB minimum
- Hard Drive SSD is preferred 500GB
- Processor intel i5 minimum
- Operating system Windows 10

Software:

- Jupyter notebook (Python)

Libraries:

- Pandas (used to create the data and read the data)
- Numpy (used with the mathematical function)
- Seaborn (used to create a different types of graphs)
- Matplotlib (used to plot the graph)
- Regexp_tokenize (used to remove numbers and symbols)
- Stopwords (used to remove the unnecessary words)
- Train_test_split (used to split the data into train and test data)
- Accuracy_score (used to calculate accuracy score for train and test)
- Classification_report (to display precision, f1 score)
- Confusion_matrix (form the matrix)
- Roc_curve (used to plot the area under curve)

Model/s Development and Evaluation

Identification of possible problem: We approach to both statistical and analytical problem

❖ Plot a bar graph for nominal data and distribution graph for continuous data.

❖ describe () use to calculate mean, standard deviation, minimum, maximum and quantile deviation.

❖ corr() used to calculate the correlation of input variable with the output variables.

❖ Scatter plot between target variable to the feature variables.

Testing of Identified Approaches:

Here we work on the classification problem so the machine learning models are:

- Logistic Regression
- K Neighbors Classifier
- Random Forest Classifier
- Decision Tree Classifier.

Run and Evaluate selected models:

➢Logistic Regression.

```
#Train Test Split
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=46)
```

```
lr.fit(x_train,y_train)
y_pred1 = lr.predict(x_test)
accuracy = accuracy_score(y_test,y_pred1)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 63.03637713437268
```

```
cm= confusion_matrix(y_test,y_pred1)
print(cm)
```

```
[[3193 3752]
 [1227 5298]]
```

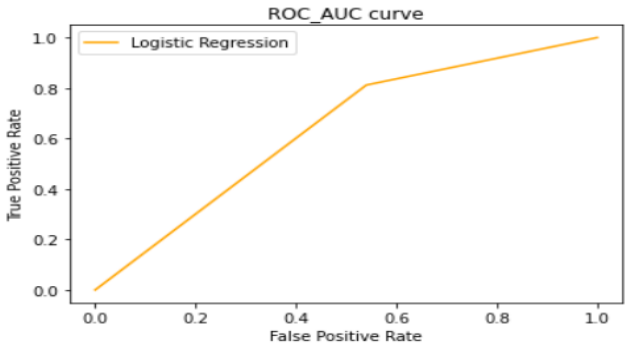
```
clr=classification_report(y_test,y_pred1)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.72      0.46      0.56      6945
           1       0.59      0.81      0.68      6525

    accuracy                           0.63     13470
   macro avg       0.65      0.64      0.62     13470
weighted avg       0.66      0.63      0.62     13470
```

The accuracy score of logistic regression is 63.03%.And the precision is 72 , recall is 46 and f1-score is 52 . The sum of true negative and false negative is 5862 and the area under the curve is 56.64.

ROC – CURVE:

```
fpr,tpr,thresholds = roc_curve(y_test,y_pred1)

plt.plot(fpr,tpr,color='orange',label='Logistic Regression')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC_AUC curve')
plt.legend()
plt.show()

auc_score = roc_auc_score(y_test,y_pred1)*100
print("AUC_score",auc_score)
```




```
AUC_score 63.58546212854696
```

➢ K Neighbors Classifier

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.30,random_state=68)
```

```
knn.fit(x_train,y_train)
y_pred2 = knn.predict(x_test)
accuracy = accuracy_score(y_test,y_pred2)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 90.65330363771344
```

Confusion Matrix:

Confusion Matrix :

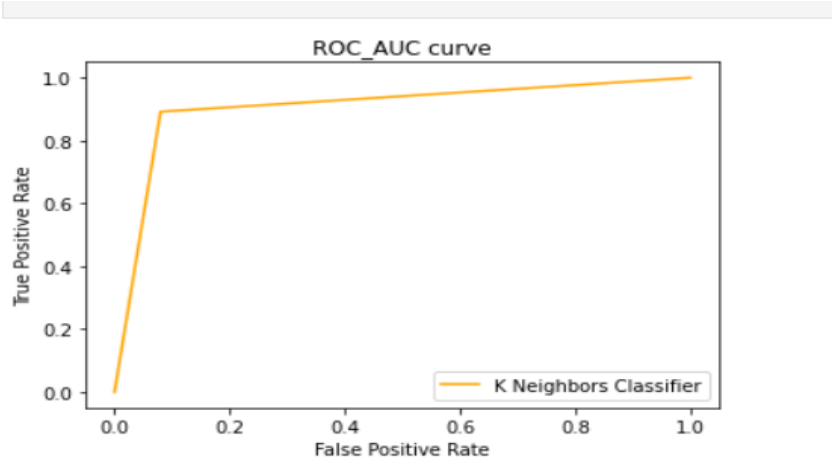
```
cm= confusion_matrix(y_test,y_pred2)
print(cm)
```

```
[[6432  558]
 [ 701 5779]]
```

```
clr=classification_report(y_test,y_pred2)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      6990
           1       0.91      0.89      0.90      6480

    accuracy                           0.91     13470
   macro avg       0.91      0.91      0.91     13470
weighted avg       0.91      0.91      0.91     13470
```

The accuracy score of K Neighbors classification is 91%. And the precision is 90, recall is 92 and f1-score is 91. The sum of true negative and false negative is 1649 and the area under the curve is 91.



AUC_score 90.59963307370317

➢ Random Forest Classifier :

```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=93)
```

```
rfc.fit(x_train,y_train)
y_pred3 = rfc.predict(x_test)
accuracy = accuracy_score(y_test,y_pred3)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 99.24276169265033
```

```
cm= confusion_matrix(y_test,y_pred3)
print(cm)
```

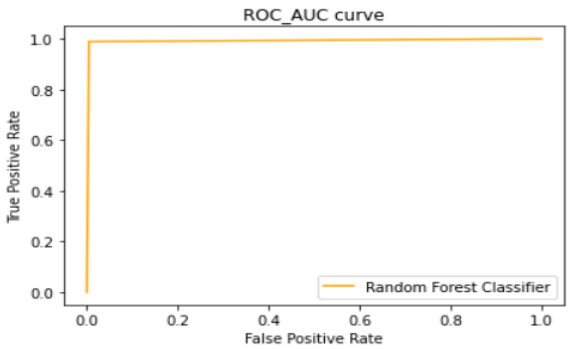
```
[[7012   32]
 [  70 6356]]
```

```
clr=classification_report(y_test,y_pred3)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      7044
           1       0.99      0.99      0.99      6426

    accuracy                           0.99     13470
   macro avg       0.99      0.99      0.99     13470
weighted avg       0.99      0.99      0.99     13470
```

The accuracy score of Random Forest classification is 99.45%. And the precision is 99, recall is 100 and f1-score is 100. The sum of true negative and false negative is 63 and the area under the curve is 99.45.

ROC CURVE :




```
AUC_score 99.22819402226156
```

➢ Decision Tree Classifier :

```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=40)
```

```
clf.fit(x_train,y_train)
y_pred4 = clf.predict(x_test)
accuracy = accuracy_score(y_test,y_pred4)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 98.30734966592428
```

```
cm= confusion_matrix(y_test,y_pred4)
print(cm)
```

```
[[6894  111]
 [ 117 6348]]
```

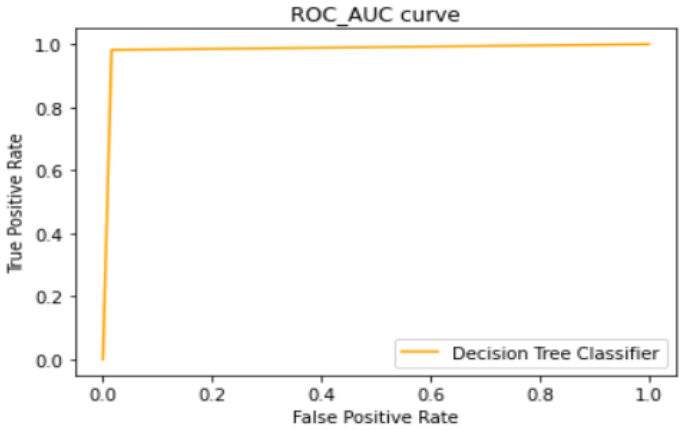
```
clr=classification_report(y_test,y_pred4)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      7005
           1       0.98      0.98      0.98      6465

    accuracy                           0.98     13470
   macro avg       0.98      0.98      0.98     13470
weighted avg       0.98      0.98      0.98     13470
```

The accuracy score of Decision Tree classification is 98 %. And the precision is 98, recall is 98 and f1-score is 100. The sum of true negative and false negative is 16 and the area under the curve is 99.94. The Decision Tree Classifier gives better accuracy score, precision score, recall and f1-score. The total of True Negative and False Negative in the confusion matrix is less in the same model and area under the curve is also higher for the testing data.

ROC CURVE

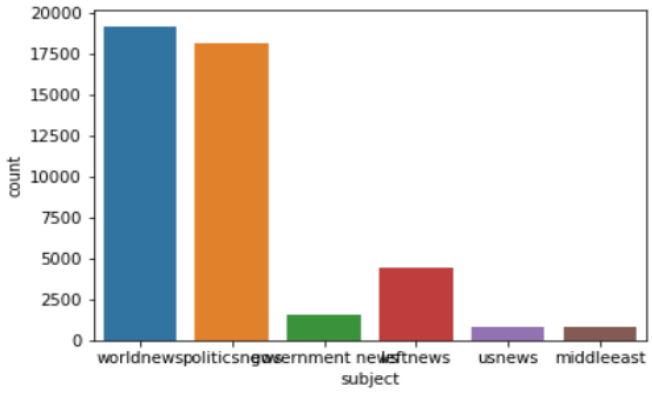
```
fpr,tpr,thresholds = roc_curve(y_test,y_pred4)

plt.plot(fpr,tpr,color='orange',label='Decision Tree Classifier')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC_AUC curve')
plt.legend()
plt.show()

auc_score = roc_auc_score(y_test,y_pred4)*100
print("AUC_score",auc_score)
```




```
AUC_score 98.30283638965207
```

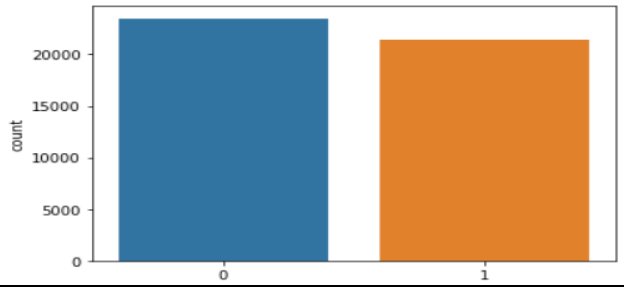
Visualization:

On visualizing the continuous data we see that our target variable is balance. In subject column there are two subject that are politics, politics news and News, world news both are same so we replace politics by politics news and News by world news.

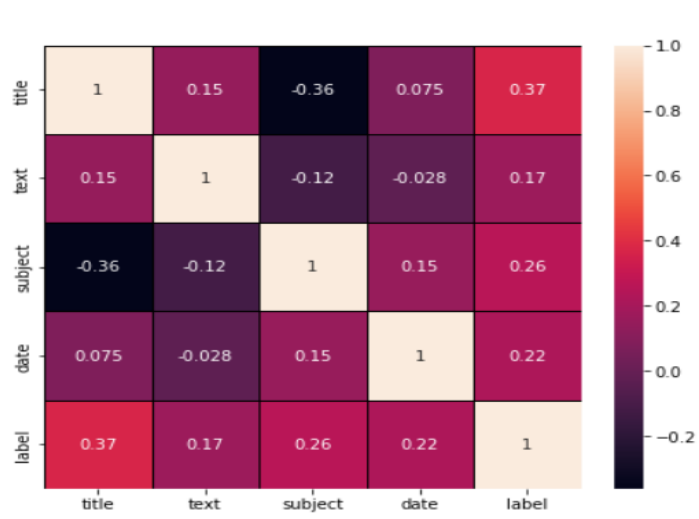


```
ax= sns.countplot(x='label',data=data)
print(data['label'].value_counts())
```

```
0    23481
1    21417
Name: label, dtype: int64
```



Correlation :



Interpretation of the Results:

On our analysis basis we go through various models and then we conclude better model on the basis of various classifications. Our data is balance so we do consider accuracy score for model testing. Then we go with the precision, f1-score, confusion matrix and area under the curve. After that we will predict the test data on the basis of train data.

Conclusion :

Key Findings and Conclusions of the Study:

On study the fake data we see that there are 168 columns having missing values so we drop them and add fake data with the true data. We see that target variable is balanced. The relation of feature variables are good with the target variable but not good with each other.

Learning Outcomes of the Study in respect of Data Science:

Here we first clean the data by dropping the columns from dataset whom having huge null values. Removing the unnecessary word, symbols from the title and text. In analysis we do describe the statistical values and correlation. Fit some classification models and find the better one i.e. Decision Tree Classifier Model. Calculate accuracy score, confusion metrics, classification report and ROC curve and these are better in the same model.

Advantage in Future:

1.Advertisers take the Advantages of Fake News

2. Influencers also take benefits of Fake News

3. Political Warfare

4. Fun and Entertainmen