CS657 Assignment -2
Predicting Fake Job Posting

Dev Divyendh Dhinakaran
G01450299
Tejaswi Samineni
G01460925

1. Missing value Percentages

job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_logo	has_questions
0.0	0.0	0.01986325592361193	0.6506542496758222	0.8403866556642697	0.1889661676293764	0.0	0.15153837085936578	0.4095838736296122	0.0	0.0	0.0

has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
0.0	0.19291524224920428	0.3934339266768832	0.451550159141813	0.2750795709065189	0.36296121655074853	0.0

2. Cleaned Dataset

job_id	title	description	telecommuting	has_company_logo	has_questions	fraudulent
1	food a fastgrowin...	food a fastgrowin...	0	1	0	0
2	organised focused...	organised focused...	0	1	0	0
3	our client locate...	our client locate...	0	1	0	0
4	the company esri ...	the company esri ...	0	1	0	0
5	job title itemiza...	job title itemiza...	0	1	1	0

3. Sampled Dataset

```
[0, 1]
+----------+-----+
|fraudulent|count|
+----------+-----+
|         0| 4779|
|         1|  886|
+----------+-----+

+------+--------------------+--------------------+-------------+----------------+-------------+----------+
|job_id|               title|         description|telecommuting|has_company_logo|has_questions|fraudulent|
+------+--------------------+--------------------+-------------+----------------+-------------+----------+
|     1|food a fastgrowin...|food a fastgrowin...|            0|               1|            0|         0|
|     2|organised focused...|organised focused...|            0|               1|            0|         0|
|     4|the company esri ...|the company esri ...|            0|               1|            0|         0|
|     5|job title itemiza...|job title itemiza...|            0|               1|            1|         0|
|     8|who is airenvyhey...|who is airenvyhey...|            0|               1|            1|         0|
+------+--------------------+--------------------+-------------+----------------+-------------+----------+
only showing top 5 rows
```

4. Vectorized Features (Title, Description and along with other columns)

```
root
 |-- job_id: long (nullable = true)
 |-- telecommuting: integer (nullable = true)
 |-- has_company_logo: integer (nullable = true)
 |-- has_questions: integer (nullable = true)
 |-- fraudulent: integer (nullable = true)
 |-- title_features: vector (nullable = true)
 |-- description_features: vector (nullable = true)

+------+-------------+----------------+-------------+----------+------------------------------------------
|job_id|telecommuting|has_company_logo|has_questions|fraudulent|title_features
+------+-------------+----------------+-------------+----------+------------------------------------------
|1     |0            |1               |0            |0         |(10000,[21,43,116,188,222,360,468,573,585,738,742,761,780,786,944,1121,1151,1241,1281,1562,1755,1807,1820,2023,2706,2762,2942,310
|2     |0            |1               |0            |0         |(10000,[30,157,289,319,320,377,447,468,511,533,594,630,655,760,1080,1117,1169,1241,1273,1277,1287,1295,1339,1413,1494,1499,1563,16
|4     |0            |1               |0            |0         |(10000,[31,32,114,116,182,193,298,320,373,419,468,524,576,591,619,715,735,736,823,827,895,942,989,1012,1020,1040,1107,1208,1209,12
|5     |0            |1               |1            |0         |(10000,[15,153,157,178,193,423,513,557,573,585,883,942,992,1121,1136,1188,1208,1241,1413,1486,1574,1611,1641,1736,1944,2002,2012,2
|8     |0            |1               |1            |0         |(10000,[1,48,73,288,375,384,429,447,468,507,553,569,585,592,621,649,740,815,853,855,942,1020,1024,1143,1174,1241,1273,1381,1413,15
+------+-------------+----------------+-------------+----------+------------------------------------------
only showing top 5 rows

+---------------------------------------------------------------------------------------------------------
|title_features
+---------------------------------------------------------------------------------------------------------
|(10000,[21,43,116,188,222,360,468,573,585,738,742,761,780,786,944,1121,1151,1241,1281,1562,1755,1807,1820,2023,2706,2762,2942,3100,3324,3354,3540,3772,3942,3988,4534,4588,4808,4867,4926,5081,54
|(10000,[30,157,289,319,320,377,447,468,511,533,594,630,655,760,1080,1117,1169,1241,1273,1277,1287,1295,1339,1413,1494,1499,1563,1611,1652,1671,1690,1764,1790,1807,1851,2110,2116,2198,2230,2252,
|(10000,[31,32,114,116,182,193,298,320,373,419,468,524,576,591,619,715,735,736,823,827,895,942,989,1012,1020,1040,1107,1208,1209,1226,1230,1241,1364,1541,1559,1574,1586,1591,1603,1606,1611,1680,
|(10000,[15,153,157,178,193,423,513,557,573,585,883,942,992,1121,1136,1188,1208,1241,1413,1486,1574,1611,1641,1736,1944,2002,2012,2036,2103,2207,2249,2252,2353,2537,2543,2673,2703,3060,3158,3332
|(10000,[1,48,73,288,375,384,429,447,468,507,553,569,585,592,621,649,740,815,853,855,942,1020,1024,1143,1174,1241,1273,1381,1413,1514,1590,1611,1626,1671,1698,1731,1951,1975,2025,2049,2116,2165,
+---------------------------------------------------------------------------------------------------------
```

5. ML model outputs

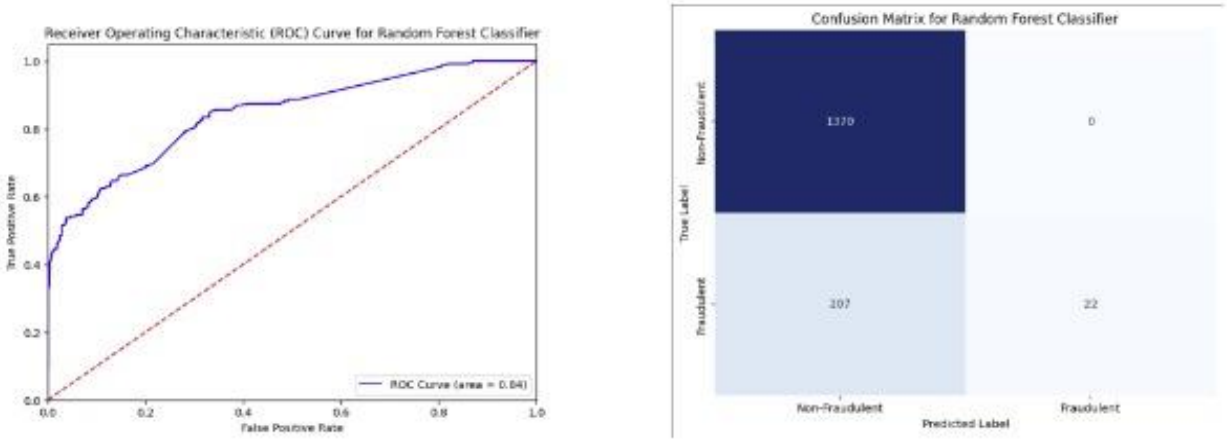
```
Starting training for Random Forest Classifier...
Completed training for Random Forest Classifier.
Best parameters for Random Forest Classifier: {Param(parent='RandomForestClassifier_3da1e832b17a', name='bootstrap', doc='Whether bootstrap sampl
Accuracy: 0.8705440900562852
F1 Score: 0.8217094195262513
```

```
Starting training for Linear SVC...
Performing cross-validation for Linear SVC...
Best parameters for Linear SVC: {Param(parent='LinearSVC_345434355962', name='aggregationDepth', doc='suggested depth for treeAggregate (>= 2).'): 2, Param(parent='LinearSVC_345434355962', name=
Completed training for Linear SVC. Accuracy: 0.9280575539568345, F1 Score: 0.9242883321375246
```

```
Starting training for Logistic Regression...
Performing cross-validation for Logistic Regression...
Best parameters for Logistic Regression: {Param(parent='LogisticRegression_ac4ae20efbc1', name='aggregationDepth', doc='suggested depth for treeAggregate (>= 2).'): 2, Param(parent='LogisticRegr
Completed training for Logistic Regression. Accuracy: 0.9292565947242206, F1 Score: 0.9230732915213902
```

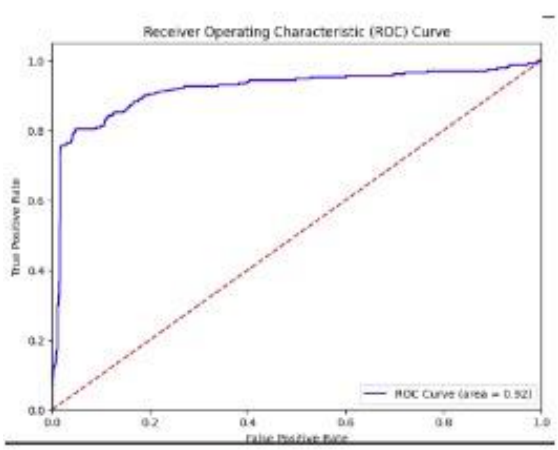
```
Starting training for Multilayer Perceptron Classifier...
Number of features in the dataset: 20003
Starting training for Multilayer Perceptron Classifier...
Completed training for Multilayer Perceptron Classifier.
Best parameters for Multilayer Perceptron Classifier: {Param(parent='MultilayerPerceptronClassifier_b1ee631137fd', name='blockSize', doc='block s
Accuracy: 0.9424640400250156
F1 Score: 0.9413588972559905
```

6. Metrics Graphs

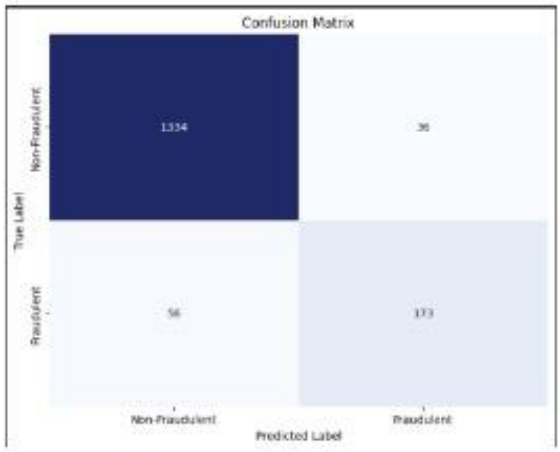


(a) ROC Curve (b) Confusion Matrix

Figure 8: Roc and Comfusion matix for Random Forest



(a) ROC Curve



(b) Confusion Matrix

Figure 10: Roc and Comfusion matix for Multi-Layer Perceptron