



```
In [ ]: import os
import sys
from tempfile import NamedTemporaryFile
from urllib.request import urlopen
from urllib.parse import unquote, urlparse
from urllib.error import HTTPError
from zipfile import ZipFile
import tarfile
import shutil

CHUNK_SIZE = 40960
DATA_SOURCE_MAPPING = 'telecom-churn-datasets:https%3A%2F%2Fstorage.googleapis.com/telecom-churn-datasets'

KAGGLE_INPUT_PATH='/kaggle/input'
KAGGLE_WORKING_PATH='/kaggle/working'
KAGGLE_SYMLINK='kaggle'

!umount /kaggle/input/ 2> /dev/null
shutil.rmtree('/kaggle/input', ignore_errors=True)
os.makedirs(KAGGLE_INPUT_PATH, 0o777, exist_ok=True)
os.makedirs(KAGGLE_WORKING_PATH, 0o777, exist_ok=True)

try:
    os.symlink(KAGGLE_INPUT_PATH, os.path.join(".", 'input'), target_is_directory=True)
except FileExistsError:
    pass
try:
    os.symlink(KAGGLE_WORKING_PATH, os.path.join(".", 'working'), target_is_directory=True)
except FileExistsError:
    pass

for data_source_mapping in DATA_SOURCE_MAPPING.split(','):
    directory, download_url_encoded = data_source_mapping.split(':')
    download_url = unquote(download_url_encoded)
    filename = urlparse(download_url).path
    destination_path = os.path.join(KAGGLE_INPUT_PATH, directory)
    try:
        with urlopen(download_url) as fileres, NamedTemporaryFile() as tfile:
            total_length = fileres.headers['content-length']
            print(f'Downloading {directory}, {total_length} bytes compressed')
            dl = 0
            data = fileres.read(CHUNK_SIZE)
            while len(data) > 0:
                dl += len(data)
                tfile.write(data)
                done = int(50 * dl / int(total_length))
                sys.stdout.write(f"\r[{'=' * done}{' ' * (50-done)}] {dl} bytes compressed")
                sys.stdout.flush()
                data = fileres.read(CHUNK_SIZE)
            if filename.endswith('.zip'):
                with ZipFile(tfile) as zfile:
                    zfile.extractall(destination_path)
            else:
```

```

        with tarfile.open(tfile.name) as tarfile:
            tarfile.extractall(destination_path)
        print(f'\nDownloaded and uncompressed: {directory}')
    except HTTPError as e:
        print(f'Failed to load (likely expired) {download_url} to path {destination_path}')
        continue
    except OSError as e:
        print(f'Failed to load {download_url} to path {destination_path}')
        continue

print('Data source import complete.')

```

Failed to load (likely expired) [https://storage.googleapis.com/kaggle-data-sets/255093/535845/bundle/archive.zip?X-Goog-Algorithm=G00G4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20241010%2Fauto%2Fstorage%2Fgoog4\\_request&X-Goog-Date=20241010T051751Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=06c803ef95f7e8b0a294fa862426ffc5b913eff6b872597530c7a3e5d647887bb8572f29e13a03ddc6fa5f99f041134c2742b611cc99228d9e51819a09c4d4355d511c3cee9e2d59959fd69c958cfb17bf7614887e76d2c4f1b588eedc4a844f3db3dcc8a44c3c59cb7a282dc66371ea3395feb45168008c9d29322b195c8193bb29cdf46eaa3c257bfad7db08605c49764b21a0a46e57ccb4d9e0ba40d7febcf256d76fa6b7a0e66a8d17d95acd6cd8b45e2f5988a02e4ff94d6632b059919ebe429e5ff59852ae55048668795e67f898fa62fddb695e7830ad308b600423e5af7cfa41a165e1a05bd61b9b145a8c5a130bde163f103d2a7d52d2b40d0908f2](https://storage.googleapis.com/kaggle-data-sets/255093/535845/bundle/archive.zip?X-Goog-Algorithm=G00G4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20241010%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20241010T051751Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=06c803ef95f7e8b0a294fa862426ffc5b913eff6b872597530c7a3e5d647887bb8572f29e13a03ddc6fa5f99f041134c2742b611cc99228d9e51819a09c4d4355d511c3cee9e2d59959fd69c958cfb17bf7614887e76d2c4f1b588eedc4a844f3db3dcc8a44c3c59cb7a282dc66371ea3395feb45168008c9d29322b195c8193bb29cdf46eaa3c257bfad7db08605c49764b21a0a46e57ccb4d9e0ba40d7febcf256d76fa6b7a0e66a8d17d95acd6cd8b45e2f5988a02e4ff94d6632b059919ebe429e5ff59852ae55048668795e67f898fa62fddb695e7830ad308b600423e5af7cfa41a165e1a05bd61b9b145a8c5a130bde163f103d2a7d52d2b40d0908f2) to path /kaggle/input/telecom-churn-datasets  
Data source import complete.

```

In [ ]: import numpy as np
import pandas as pd

import os
for dirname, __, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

```

```

In [ ]: data = pd.read_csv('/kaggle/input/telecom-churn-datasets/churn-bigml-80.csv')
data.sample(5)

```

```

-----
FileNotFoundError                                Traceback (most recent call last)
<ipython-input-24-6ca4a4ae7695> in <cell line: 1>()
----> 1 data = pd.read_csv('/kaggle/input/telecom-churn-datasets/churn-bigml-8
0.csv')
      2 data.sample(5)

/usr/local/lib/python3.10/dist-packages/pandas/io/parsers/readers.py in read_csv
v(filepath_or_buffer, sep, delimiter, header, names, index_col, usecols, dtype,
engine, converters, true_values, false_values, skipinitialspace, skiprows, skip
footer, nrows, na_values, keep_default_na, na_filter, verbose, skip_blank_line
s, parse_dates, infer_datetime_format, keep_date_col, date_parser, date_format,
dayfirst, cache_dates, iterator, chunksize, compression, thousands, decimal, li
neterminator, quotechar, quoting, doublequote, escapechar, comment, encoding, e
ncoding_errors, dialect, on_bad_lines, delim_whitespace, low_memory, memory_ma
p, float_precision, storage_options, dtype_backend)
    1024         kwds.update(kwds_defaults)
    1025
-> 1026     return _read(filepath_or_buffer, kwds)
    1027
    1028

/usr/local/lib/python3.10/dist-packages/pandas/io/parsers/readers.py in _read(f
ilepath_or_buffer, kwds)
    618
    619     # Create the parser.
--> 620     parser = TextFileReader(filepath_or_buffer, **kwds)
    621
    622     if chunksize or iterator:

/usr/local/lib/python3.10/dist-packages/pandas/io/parsers/readers.py in __ini
t__(self, f, engine, **kwds)
    1618
    1619         self.handles: IOHandles | None = None
-> 1620         self._engine = self._make_engine(f, self.engine)
    1621
    1622     def close(self) -> None:

/usr/local/lib/python3.10/dist-packages/pandas/io/parsers/readers.py in _make_e
ngine(self, f, engine)
    1878         if "b" not in mode:
    1879             mode += "b"
-> 1880         self.handles = get_handle(
    1881             f,
    1882             mode,

/usr/local/lib/python3.10/dist-packages/pandas/io/common.py in get_handle(pat
h_or_buf, mode, encoding, compression, memory_map, is_text, errors, storage_opt
ions)
    871         if ioargs.encoding and "b" not in ioargs.mode:
    872             # Encoding
--> 873             handle = open(
    874                 handle,
    875                 ioargs.mode,

```

```
FileNotFoundError: [Errno 2] No such file or directory: '/kaggle/input/telecom-churn-datasets/churn-bigml-80.csv'
```

## EDA

```
In [ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2666 entries, 0 to 2665
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   State                                2666 non-null   object
1   Account length                       2666 non-null   int64
2   Area code                           2666 non-null   int64
3   International plan                   2666 non-null   object
4   Voice mail plan                     2666 non-null   object
5   Number vmail messages               2666 non-null   int64
6   Total day minutes                   2666 non-null   float64
7   Total day calls                     2666 non-null   int64
8   Total day charge                    2666 non-null   float64
9   Total eve minutes                   2666 non-null   float64
10  Total eve calls                     2666 non-null   int64
11  Total eve charge                    2666 non-null   float64
12  Total night minutes                 2666 non-null   float64
13  Total night calls                   2666 non-null   int64
14  Total night charge                  2666 non-null   float64
15  Total intl minutes                  2666 non-null   float64
16  Total intl calls                    2666 non-null   int64
17  Total intl charge                   2666 non-null   float64
18  Customer service calls              2666 non-null   int64
19  Churn                              2666 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 398.5+ KB
```

```
In [ ]: df.shape
```

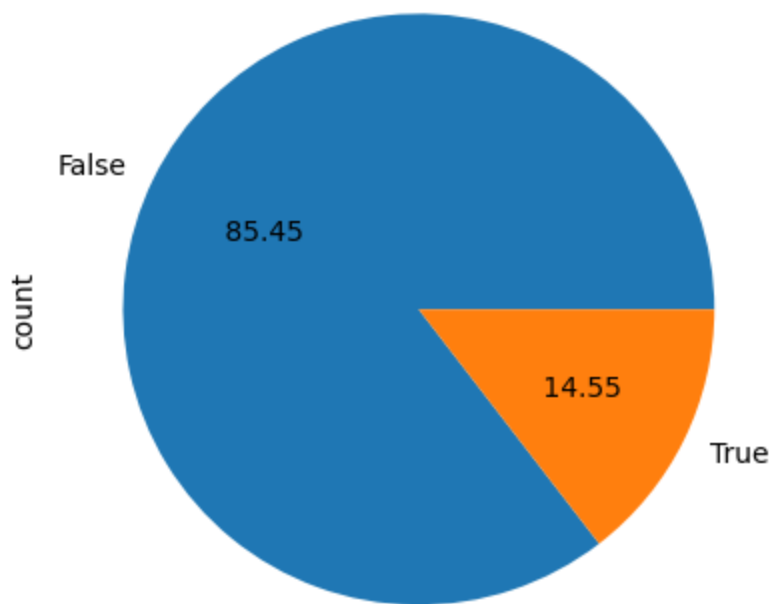
```
Out[ ]: (2666, 20)
```

```
In [ ]: df['State'].nunique()
```

```
Out[ ]: 51
```

```
In [ ]: df['Churn'].value_counts().plot(kind='pie', autopct='% .2f')
```

```
Out[ ]: <Axes: ylabel='count'>
```



## \*\*Data Preprocessing

```
In [ ]: df1 = df.drop(columns=['State', 'Account length', 'Area code'])
df1.head()
```

```
Out[ ]:
```

	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge
0	No	Yes	25	265.1	110	45.07	197.4	99	16.0
1	No	Yes	26	161.6	123	27.47	195.5	103	16.0
2	No	No	0	243.4	114	41.38	121.2	110	10.0
3	Yes	No	0	299.4	71	50.90	61.9	88	5.0
4	Yes	No	0	166.7	113	28.34	148.3	122	12.0

```
In [ ]: X=df1.iloc[:, :-1]
y=df1.iloc[:, -1]
```

```
In [ ]: X.head()
```

```
Out[ ]:
```

	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge
0	No	Yes	25	265.1	110	45.07	197.4	99	16.1
1	No	Yes	26	161.6	123	27.47	195.5	103	16.1
2	No	No	0	243.4	114	41.38	121.2	110	10.1
3	Yes	No	0	299.4	71	50.90	61.9	88	5.1
4	Yes	No	0	166.7	113	28.34	148.3	122	12.1

```
In [ ]: from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.2, random_s
X_train.shape, X_test.shape
```

```
Out[ ]: ((2132, 16), (534, 16))
```

```
In [ ]: from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
transformer = ColumnTransformer(transformers=[
    ('tnfl',OneHotEncoder(sparse_output = False,dr
    ['International plan','Voice mail plan'])],rem
```

```
In [ ]: X_train_new = transformer.fit_transform(X_train)
X_test_new = transformer.transform(X_test)
X_train_new.shape
```

```
Out[ ]: (2132, 16)
```

```
In [ ]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
le.fit(y_train)
le.classes_
```

```
Out[ ]: array([0, 1])
```

```
In [ ]: y_train = le.transform(y_train)
y_test = le.transform(y_test)
```

## **\*\*Model Training**

```
In [ ]: from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train_new)
X_train_scaled = scaler.transform(X_train_new)
X_test_scaled = scaler.transform(X_test_new)
```

```

-----
NameError                                Traceback (most recent call last)
<ipython-input-18-2fc08d7d63a2> in <cell line: 3>()
      1 from sklearn.preprocessing import StandardScaler
      2 scaler = StandardScaler()
----> 3 scaler.fit(X_train_new)
      4 X_train_scaled = scaler.transform(X_train_new)
      5 X_test_scaled = scaler.transform(X_test_new)

NameError: name 'X_train_new' is not defined

```

```

In [ ]: from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(max_iter=2000, penalty = 'l2')
lr.fit(X_train_scaled,y_train)
y_pred_lr = lr.predict(X_test_scaled)

```

```

-----
NameError                                Traceback (most recent call last)
<ipython-input-19-7a74d883a891> in <cell line: 3>()
      1 from sklearn.linear_model import LogisticRegression
      2 lr = LogisticRegression(max_iter=2000, penalty = 'l2')
----> 3 lr.fit(X_train_scaled,y_train)
      4 y_pred_lr = lr.predict(X_test_scaled)

NameError: name 'X_train_scaled' is not defined

```

```

In [ ]: from google.colab import drive
drive.mount('/content/drive')

```

```

In [ ]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test,y_pred_lr)
print("Logistic Regression Accuracy: ",acc)

```

Logistic Regression Accuracy: 0.8595505617977528

## Decision Tree

```

In [ ]: from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(max_depth = 7)
dt.fit(X_train_new,y_train)
y_pred_dt = dt.predict(X_test_new)

```

```

In [ ]: acc_dt = accuracy_score(y_test,y_pred_dt)
print("Logistic Regression Accuracy: ",acc_dt)

```

Logistic Regression Accuracy: 0.9288389513108615

# Grid Search with Random Forest

```

In [ ]: from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()

```

```
In [ ]: n_estimators = [20,60,100,120]
max_features = [0.2,0.6,1.0]
max_depth = [2,8,None]
max_samples = [0.5,0.75,1.0]
param_grid = {'n_estimators': n_estimators,
              'max_features': max_features,
              'max_depth': max_depth,
              'max_samples': max_samples
              }
print(param_grid)

{'n_estimators': [20, 60, 100, 120], 'max_features': [0.2, 0.6, 1.0], 'max_depth': [2, 8, None], 'max_samples': [0.5, 0.75, 1.0]}
```

```
In [ ]: from sklearn.model_selection import GridSearchCV

rf_grid = GridSearchCV(
    estimator = rf,
    param_grid = param_grid,
    cv=5,
    verbose = 2)

rf_grid.fit(X_train_new,y_train)
```





[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

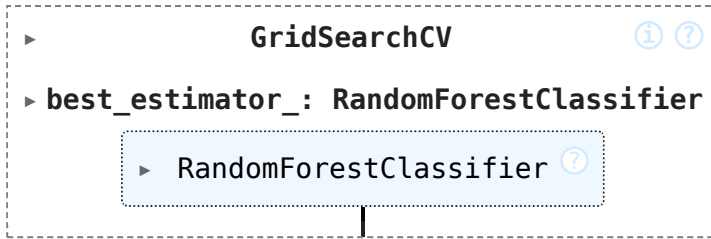
[illegible]

[illegible]

[illegible]

otal time= 1.9s

Out[ ]:



In [ ]: rf\_grid.best\_params\_

Out[ ]: {'max\_depth': None,  
'max\_features': 0.6,  
'max\_samples': 1.0,  
'n\_estimators': 120}

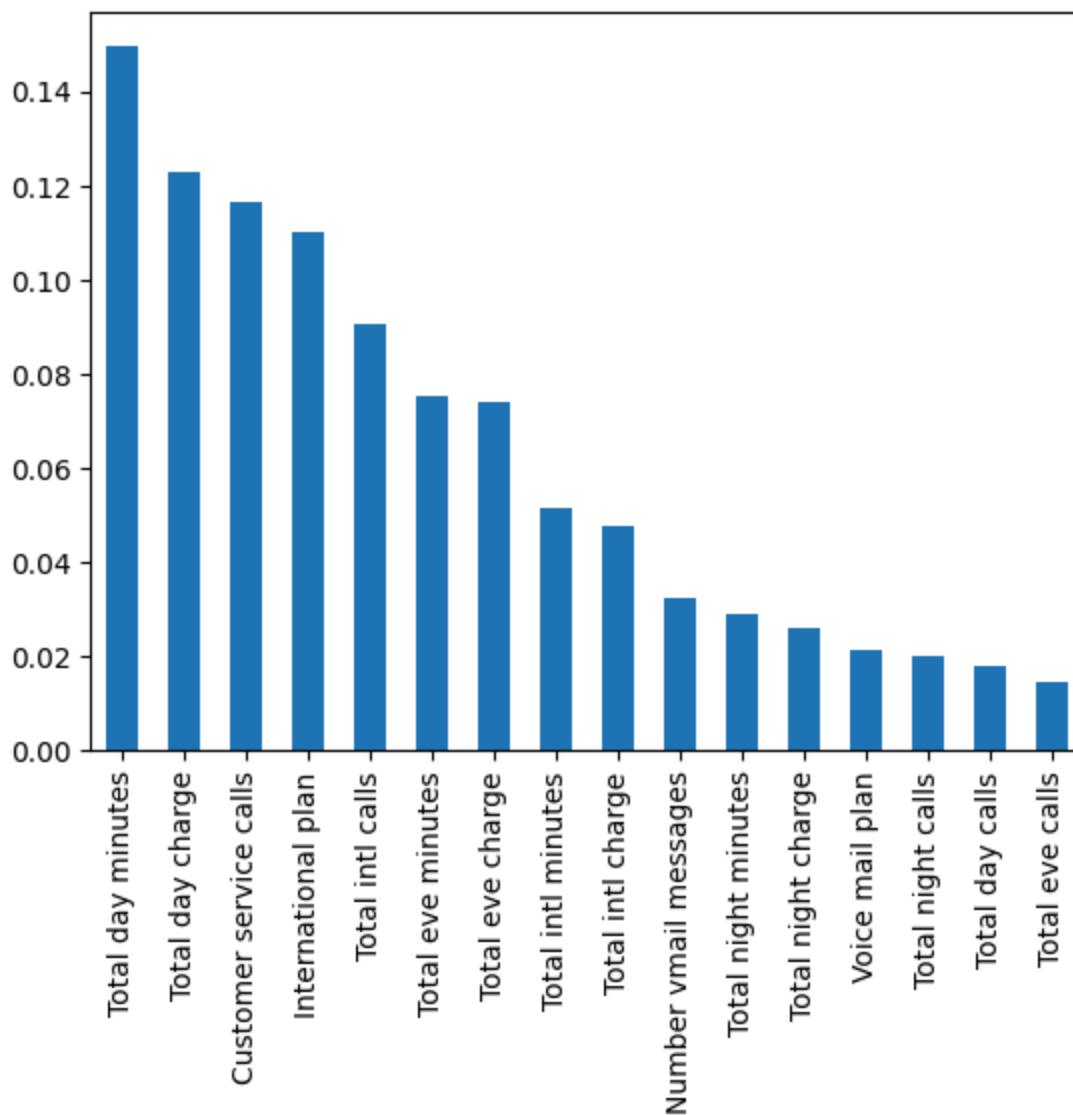
```
In [ ]: rf1 = RandomForestClassifier(max_depth=8, max_features=0.6, max_samples=0.75,  
rf1.fit(X_train_new,y_train)  
y_pred_rf = rf1.predict(X_test_new)  
acc_rf = accuracy_score(y_test,y_pred_rf)  
print("Random Forest Accuracy: ",acc_rf)
```

Random Forest Accuracy: 0.9531835205992509

### Feature Importance

```
In [ ]: cols = X_train.columns  
r=pd.Series(np.abs(rf1.feature_importances_),index = cols)  
r.sort_values(ascending=False).plot(kind='bar')
```

Out[ ]: <Axes: >



## Validation with Random Forest

```
In [ ]: val = pd.read_csv("/kaggle/input/telecom-churn-datasets/churn-bigml-20.csv")
val1 = val.drop(columns=['State', 'Account length', 'Area code'])
X_val= val1.iloc[:, :-1]
y_val= val1.iloc[:, -1]
X_val_new = transformer.transform(X_val)
```

```
In [ ]: y_pred_val = rfl.predict(X_val_new)

acc_val = accuracy_score(y_val, y_pred_val)

print("Validation Accuracy: ", acc_val)
```

Validation Accuracy: 0.9580209895052474

```
In [ ]: from sklearn.metrics import accuracy_score, confusion_matrix, classification_r
```

```
print("Accuracy:", accuracy_score(y_val,y_pred_val))
print("Confusion Matrix:\n", confusion_matrix(y_val,y_pred_val))
print("Classification Report:\n", classification_report(y_val,y_pred_val))
```

Accuracy: 0.9580209895052474

Confusion Matrix:

```
[[568  4]
 [ 24 71]]
```

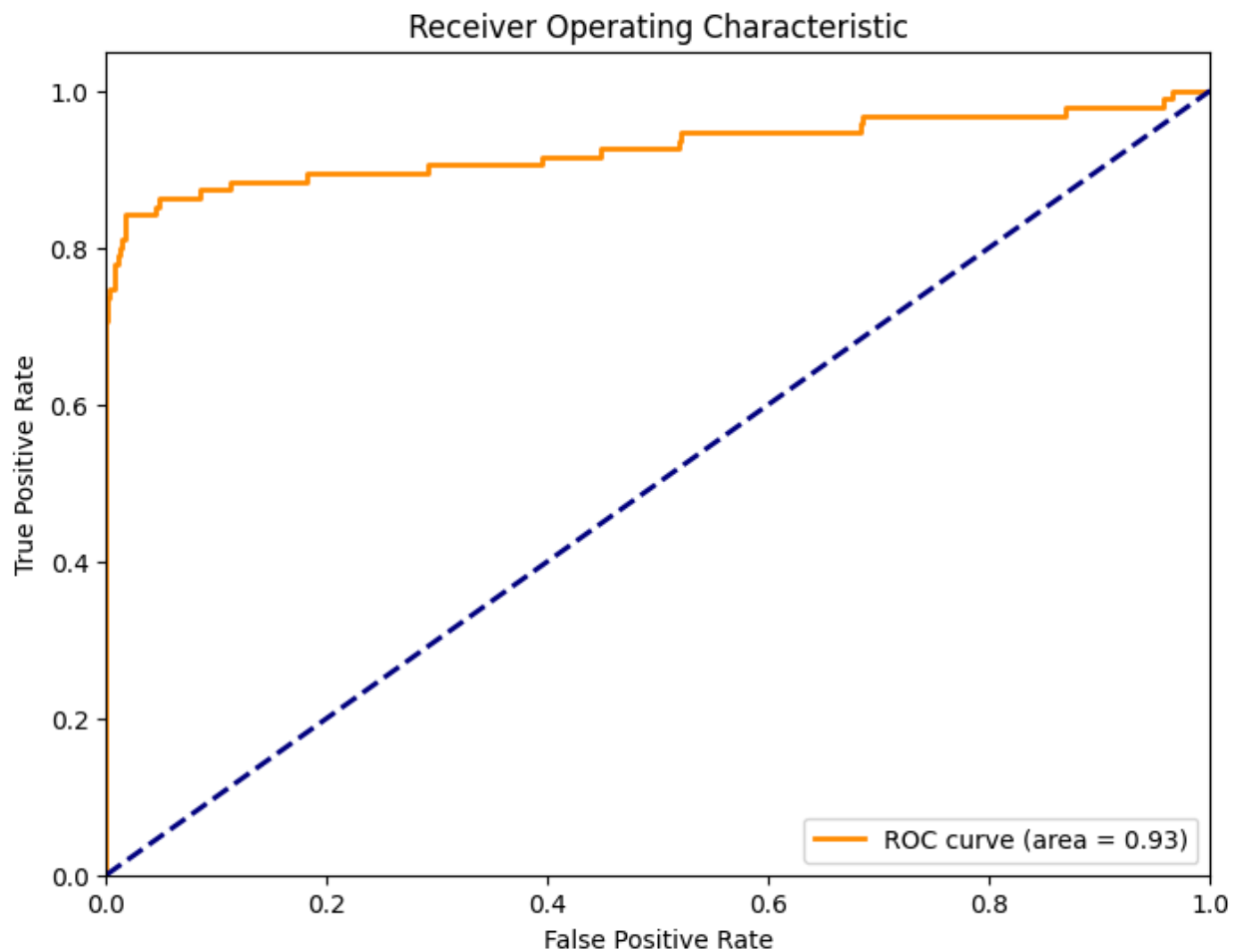
Classification Report:

	precision	recall	f1-score	support
False	0.96	0.99	0.98	572
True	0.95	0.75	0.84	95
accuracy			0.96	667
macro avg	0.95	0.87	0.91	667
weighted avg	0.96	0.96	0.96	667

## ROC-AUC

```
In [ ]: from sklearn.metrics import roc_curve, roc_auc_score, auc
import matplotlib.pyplot as plt
y_prob = rfl.predict_proba(X_val_new)[:,-1]
fpr, tpr, thresholds = roc_curve(y_val,y_prob)
roc_auc = auc(fpr,tpr)
```

```
In [ ]: plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)'
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc='lower right')
plt.show()
print("AUC Score:", roc_auc_score(y_val,y_prob))
```



AUC Score: 0.9270887007729114

## Gradient Boosting

```
In [ ]: from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(learning_rate=0.1, n_estimators=100, max_depth=3)
gbc.fit(X_train_new, y_train)

y_pred_gb = gbc.predict(X_test_new)

acc_gb = accuracy_score(y_test, y_pred_gb)

print("Testing accuracy:", acc_gb)
```

Testing accuracy: 0.9550561797752809

```
In [ ]: print("Accuracy:", accuracy_score(y_test, y_pred_gb))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_gb))
print("Classification Report:\n", classification_report(y_test, y_pred_gb))
```



Accuracy: 0.9550561797752809

Confusion Matrix:

```
[[454  1]
 [ 23 56]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.95	1.00	0.97	455
1	0.98	0.71	0.82	79
accuracy			0.96	534
macro avg	0.97	0.85	0.90	534
weighted avg	0.96	0.96	0.95	534

## Stacking

```
In [ ]: from sklearn.neighbors import KNeighborsClassifier
```

```
estimators = [
    ('rfs', RandomForestClassifier(n_estimators=50, random_state=42)),
    ('knn', KNeighborsClassifier(n_neighbors=3)),
    ('gbdt', GradientBoostingClassifier())
]
```

```
-----
NameError                                Traceback (most recent call last)
<ipython-input-13-b9d594bf29dd> in <cell line: 4>()
      2
      3 estimators = [
----> 4     ('rfs', RandomForestClassifier(n_estimators=50, random_state=42)),
      5     ('knn', KNeighborsClassifier(n_neighbors=3)),
      6     ('gbdt', GradientBoostingClassifier())

NameError: name 'RandomForestClassifier' is not defined
```

```
In [ ]: from sklearn.ensemble import StackingClassifier
```

```
clf = StackingClassifier(
    estimators=estimators,
    final_estimator=LogisticRegression(),
    cv=10
)

clf.fit(X_train_new, y_train)
```

```

-----
NameError                                Traceback (most recent call last)
<ipython-input-9-059946556de1> in <cell line: 3>()
      2
      3 clf = StackingClassifier(
----> 4     estimators=estimators,
      5     final_estimator=LogisticRegression(),
      6     cv=10

NameError: name 'estimators' is not defined

```

```

In [ ]: y_pred_st = clf.predict(X_test_new)

acc_st = accuracy_score(y_test, y_pred_st)

print("Testing accuracy:", acc_st)

```

Testing accuracy: 0.9531835205992509

```

In [ ]: print("Accuracy:", accuracy_score(y_test, y_pred_st))
        print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_st))
        print("Classification Report:\n", classification_report(y_test, y_pred_st))

```

Accuracy: 0.9531835205992509

Confusion Matrix:

```

[[451  4]
 [ 21 58]]

```

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.97	455
1	0.94	0.73	0.82	79
accuracy			0.95	534
macro avg	0.95	0.86	0.90	534
weighted avg	0.95	0.95	0.95	534

```

In [ ]: y_prob_st = clf.predict_proba(X_val_new)[: ,1]
        fpr, tpr, thresholds = roc_curve(y_val, y_prob_st)
        roc_auc = auc(fpr, tpr)

```

```

-----
NameError                                Traceback (most recent call last)
<ipython-input-16-5057430a3752> in <cell line: 1>()
----> 1 y_prob_st = clf.predict_proba(X_val_new)[: ,1]
      2 fpr, tpr, thresholds = roc_curve(y_val, y_prob_st)
      3 roc_auc = auc(fpr, tpr)

NameError: name 'clf' is not defined

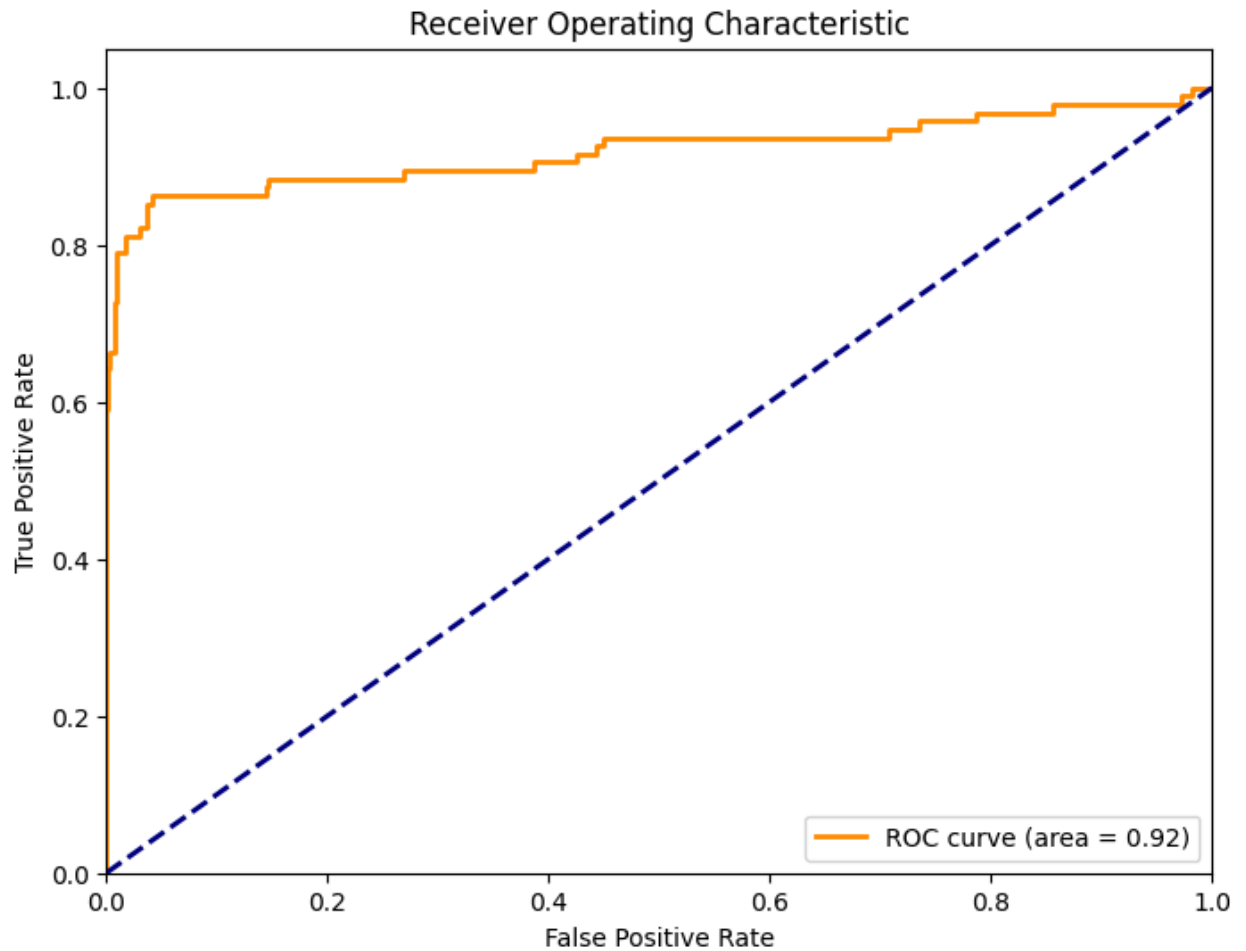
```

```

In [ ]: plt.figure(figsize=(8, 6))
        plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)'
        plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')

```

```
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc='lower right')
plt.show()
print("AUC Score:", roc_auc_score(y_val,y_prob_st))
```



AUC Score: 0.9196724328303276