# Airline Random Forest Model churn rate

November 8, 2024

[1]:
```
import numpy as np
import pandas as pd

import pickle as pkl

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, PredefinedSplit,␣
 ↪GridSearchCV
from sklearn.metrics import f1_score, precision_score, recall_score,␣
 ↪accuracy_score
```

[2]:
```
air_data = pd.read_csv("Invistico_Airline.csv")
```

[3]:
```
air_data.head(10)
```

[3]:
```
  satisfaction   Customer Type  Age   Type of Travel     Class  \
0    satisfied  Loyal Customer   65  Personal Travel       Eco
1    satisfied  Loyal Customer   47  Personal Travel  Business
2    satisfied  Loyal Customer   15  Personal Travel       Eco
3    satisfied  Loyal Customer   60  Personal Travel       Eco
4    satisfied  Loyal Customer   70  Personal Travel       Eco
5    satisfied  Loyal Customer   30  Personal Travel       Eco
6    satisfied  Loyal Customer   66  Personal Travel       Eco
7    satisfied  Loyal Customer   10  Personal Travel       Eco
8    satisfied  Loyal Customer   56  Personal Travel  Business
9    satisfied  Loyal Customer   22  Personal Travel       Eco

   Flight Distance  Seat comfort  Departure/Arrival time convenient  \
0              265             0                                  0
1             2464             0                                  0
2             2138             0                                  0
3              623             0                                  0
4              354             0                                  0
5             1894             0                                  0
6              227             0                                  0
7             1812             0                                  0
8               73             0                                  0
```

```
9             1556             0                                  0

   Food and drink  Gate location  …  Online support  Ease of Online booking  \
0               0              2  …               2                       3
1               0              3  …               2                       3
2               0              3  …               2                       2
3               0              3  …               3                       1
4               0              3  …               4                       2
5               0              3  …               2                       2
6               0              3  …               5                       5
7               0              3  …               2                       2
8               0              3  …               5                       4
9               0              3  …               2                       2

   On-board service  Leg room service  Baggage handling  Checkin service  \
0                 3                 0                 3                5
1                 4                 4                 4                2
2                 3                 3                 4                4
3                 1                 0                 1                4
4                 2                 0                 2                4
5                 5                 4                 5                5
6                 5                 0                 5                5
7                 3                 3                 4                5
8                 4                 0                 1                5
9                 2                 4                 5                3

   Cleanliness  Online boarding  Departure Delay in Minutes  \
0            3                2                           0
1            3                2                         310
2            4                2                           0
3            1                3                           0
4            2                5                           0
5            4                2                           0
6            5                3                          17
7            4                2                           0
8            4                4                           0
9            4                2                          30

   Arrival Delay in Minutes
0                       0.0
1                     305.0
2                       0.0
3                       0.0
4                       0.0
5                       0.0
6                      15.0
7                       0.0
```

```
8                        0.0
9                       26.0

[10 rows x 22 columns]
```

[4]:
```
air_data.dtypes
```

[4]:
```
satisfaction                          object
Customer Type                         object
Age                                    int64
Type of Travel                        object
Class                                 object
Flight Distance                        int64
Seat comfort                           int64
Departure/Arrival time convenient      int64
Food and drink                         int64
Gate location                          int64
Inflight wifi service                  int64
Inflight entertainment                 int64
Online support                         int64
Ease of Online booking                 int64
On-board service                       int64
Leg room service                       int64
Baggage handling                       int64
Checkin service                        int64
Cleanliness                            int64
Online boarding                        int64
Departure Delay in Minutes             int64
Arrival Delay in Minutes             float64
dtype: object
```

[5]:
```
air_data.shape
```

[5]:
```
(129880, 22)
```

[6]:
```
air_data.isna().any(axis=1).sum()
```

[6]:
```
393
```

[7]:
```
air_data_subset = air_data.dropna(axis=0)
```

[8]:
```
air_data_subset.head(10)
```

[8]:
```
  satisfaction   Customer Type  Age   Type of Travel     Class  \
0    satisfied  Loyal Customer   65  Personal Travel       Eco
1    satisfied  Loyal Customer   47  Personal Travel  Business
2    satisfied  Loyal Customer   15  Personal Travel       Eco
```

```
3    satisfied  Loyal Customer   60  Personal Travel       Eco
4    satisfied  Loyal Customer   70  Personal Travel       Eco
5    satisfied  Loyal Customer   30  Personal Travel       Eco
6    satisfied  Loyal Customer   66  Personal Travel       Eco
7    satisfied  Loyal Customer   10  Personal Travel       Eco
8    satisfied  Loyal Customer   56  Personal Travel  Business
9    satisfied  Loyal Customer   22  Personal Travel       Eco

   Flight Distance  Seat comfort  Departure/Arrival time convenient  \
0              265             0                                  0
1             2464             0                                  0
2             2138             0                                  0
3              623             0                                  0
4              354             0                                  0
5             1894             0                                  0
6              227             0                                  0
7             1812             0                                  0
8               73             0                                  0
9             1556             0                                  0

   Food and drink  Gate location  …  Online support  Ease of Online booking  \
0               0              2  …               2                       3
1               0              3  …               2                       3
2               0              3  …               2                       2
3               0              3  …               3                       1
4               0              3  …               4                       2
5               0              3  …               2                       2
6               0              3  …               5                       5
7               0              3  …               2                       2
8               0              3  …               5                       4
9               0              3  …               2                       2

   On-board service  Leg room service  Baggage handling  Checkin service  \
0                 3                 0                 3                5
1                 4                 4                 4                2
2                 3                 3                 4                4
3                 1                 0                 1                4
4                 2                 0                 2                4
5                 5                 4                 5                5
6                 5                 0                 5                5
7                 3                 3                 4                5
8                 4                 0                 1                5
9                 2                 4                 5                3

   Cleanliness  Online boarding  Departure Delay in Minutes  \
0            3                2                           0
1            3                2                         310
```

```
2            4                2                           0
3            1                3                           0
4            2                5                           0
5            4                2                           0
6            5                3                          17
7            4                2                           0
8            4                4                           0
9            4                2                          30

   Arrival Delay in Minutes
0                       0.0
1                     305.0
2                       0.0
3                       0.0
4                       0.0
5                       0.0
6                      15.0
7                       0.0
8                       0.0
9                      26.0

[10 rows x 22 columns]
```

[9]:
```
air_data_subset.isna().sum()
```

[9]:
```
satisfaction                         0
Customer Type                        0
Age                                  0
Type of Travel                       0
Class                                0
Flight Distance                      0
Seat comfort                         0
Departure/Arrival time convenient    0
Food and drink                       0
Gate location                        0
Inflight wifi service                0
Inflight entertainment               0
Online support                       0
Ease of Online booking               0
On-board service                     0
Leg room service                     0
Baggage handling                     0
Checkin service                      0
Cleanliness                          0
Online boarding                      0
Departure Delay in Minutes           0
Arrival Delay in Minutes             0
```

```
dtype: int64
```

```
[10]: air_data_subset_dummies = pd.get_dummies(air_data_subset,
                                              columns=['Customer Type','Type of␣
       ↪Travel','Class'])
```

```
[11]: air_data_subset_dummies.head(10)
```

```
[11]:   satisfaction  Age  Flight Distance  Seat comfort  \
      0    satisfied   65              265             0
      1    satisfied   47             2464             0
      2    satisfied   15             2138             0
      3    satisfied   60              623             0
      4    satisfied   70              354             0
      5    satisfied   30             1894             0
      6    satisfied   66              227             0
      7    satisfied   10             1812             0
      8    satisfied   56               73             0
      9    satisfied   22             1556             0

         Departure/Arrival time convenient  Food and drink  Gate location  \
      0                                  0               0              2
      1                                  0               0              3
      2                                  0               0              3
      3                                  0               0              3
      4                                  0               0              3
      5                                  0               0              3
      6                                  0               0              3
      7                                  0               0              3
      8                                  0               0              3
      9                                  0               0              3

         Inflight wifi service  Inflight entertainment  Online support  ...  \
      0                      2                       4               2  ...
      1                      0                       2               2  ...
      2                      2                       0               2  ...
      3                      3                       4               3  ...
      4                      4                       3               4  ...
      5                      2                       0               2  ...
      6                      2                       5               5  ...
      7                      2                       0               2  ...
      8                      5                       3               5  ...
      9                      2                       0               2  ...

         Online boarding  Departure Delay in Minutes  Arrival Delay in Minutes  \
      0                2                           0                       0.0
      1                2                         310                     305.0
```

```
2                2                           0                        0.0
3                3                           0                        0.0
4                5                           0                        0.0
5                2                           0                        0.0
6                3                          17                       15.0
7                2                           0                        0.0
8                4                           0                        0.0
9                2                          30                       26.0

   Customer Type_Loyal Customer  Customer Type_disloyal Customer  \
0                             1                                0
1                             1                                0
2                             1                                0
3                             1                                0
4                             1                                0
5                             1                                0
6                             1                                0
7                             1                                0
8                             1                                0
9                             1                                0

   Type of Travel_Business travel  Type of Travel_Personal Travel  \
0                               0                               1
1                               0                               1
2                               0                               1
3                               0                               1
4                               0                               1
5                               0                               1
6                               0                               1
7                               0                               1
8                               0                               1
9                               0                               1

   Class_Business  Class_Eco  Class_Eco Plus
0               0          1               0
1               1          0               0
2               0          1               0
3               0          1               0
4               0          1               0
5               0          1               0
6               0          1               0
7               0          1               0
8               1          0               0
9               0          1               0

[10 rows x 26 columns]
```

[12]:

```
air_data_subset_dummies.dtypes
```

[12]:

```
satisfaction                          object
Age                                    int64
Flight Distance                        int64
Seat comfort                           int64
Departure/Arrival time convenient      int64
Food and drink                         int64
Gate location                          int64
Inflight wifi service                  int64
Inflight entertainment                 int64
Online support                         int64
Ease of Online booking                 int64
On-board service                       int64
Leg room service                       int64
Baggage handling                       int64
Checkin service                        int64
Cleanliness                            int64
Online boarding                        int64
Departure Delay in Minutes             int64
Arrival Delay in Minutes             float64
Customer Type_Loyal Customer           uint8
Customer Type_disloyal Customer        uint8
Type of Travel_Business travel         uint8
Type of Travel_Personal Travel         uint8
Class_Business                         uint8
Class_Eco                              uint8
Class_Eco Plus                         uint8
dtype: object
```

[13]:

```
y = air_data_subset_dummies["satisfaction"]
X = air_data_subset_dummies.drop("satisfaction", axis=1)
```

[14]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25,␣
 ↪random_state = 0)
X_tr, X_val, y_tr, y_val = train_test_split(X_train, y_train, test_size = 0.25,␣
 ↪random_state = 0)
```

[15]:

```
cv_params = {'n_estimators' : [50,100],
              'max_depth' : [10,50],
              'min_samples_leaf' : [0.5,1],
              'min_samples_split' : [0.001, 0.01],
              'max_features' : ["sqrt"],
              'max_samples' : [.5,.9]}
```

[16]:

```
split_index = [0 if x in X_val.index else -1 for x in X_train.index]
custom_split = PredefinedSplit(split_index)
```

[17]:

```
rf = RandomForestClassifier(random_state=0)
```

[18]:

```
rf_val = GridSearchCV(rf, cv_params, cv=custom_split, refit='f1', n_jobs = -1,␣
 ↪verbose = 1)
```

[19]:

```
%%time

rf_val.fit(X_train, y_train)
```

```
Fitting 1 folds for each of 32 candidates, totalling 32 fits

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.
[Parallel(n_jobs=-1)]: Done  32 out of  32 | elapsed:   40.8s finished

CPU times: user 4.99 s, sys: 87.7 ms, total: 5.08 s
Wall time: 45.5 s
```

[19]:

```
GridSearchCV(cv=PredefinedSplit(test_fold=array([-1, -1, …, -1, -1])),
             error_score=nan,
             estimator=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,
                                              class_weight=None,
                                              criterion='gini', max_depth=None,
                                              max_features='auto',
                                              max_leaf_nodes=None,
                                              max_samples=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weig…
                                              n_estimators=100, n_jobs=None,
                                              oob_score=False, random_state=0,
                                              verbose=0, warm_start=False),
             iid='deprecated', n_jobs=-1,
             param_grid={'max_depth': [10, 50], 'max_features': ['sqrt'],
                         'max_samples': [0.5, 0.9],
                         'min_samples_leaf': [0.5, 1],
                         'min_samples_split': [0.001, 0.01],
                         'n_estimators': [50, 100]},
             pre_dispatch='2*n_jobs', refit='f1', return_train_score=False,
             scoring=None, verbose=1)
```

[20]:

```
rf_val.best_params_
```

[20]:

```
{'max_depth': 50,
 'max_features': 'sqrt',
 'max_samples': 0.9,
 'min_samples_leaf': 1,
 'min_samples_split': 0.001,
 'n_estimators': 50}
```

[21]:
```
rf_opt = RandomForestClassifier(n_estimators = 50, max_depth = 50,
                                min_samples_leaf = 1, min_samples_split = 0.001,
                                max_features="sqrt", max_samples = 0.9,␣
 ↪random_state = 0)
```

[22]:
```
rf_opt.fit(X_train, y_train)
```

[22]:
```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=50, max_features='sqrt',
                       max_leaf_nodes=None, max_samples=0.9,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=0.001,
                       min_weight_fraction_leaf=0.0, n_estimators=50,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

[23]:
```
y_pred = rf_opt.predict(X_test)
```

[24]:
```
pc_test = precision_score(y_test, y_pred, pos_label = "satisfied")
print("The precision score is {pc:.3f}".format(pc = pc_test))
```

```
The precision score is 0.950
```

[25]:
```
rc_test = recall_score(y_test, y_pred, pos_label = "satisfied")
print("The recall score is {rc:.3f}".format(rc = rc_test))
```

```
The recall score is 0.945
```

[26]:
```
ac_test = accuracy_score(y_test, y_pred)
print("The accuracy score is {ac:.3f}".format(ac = ac_test))
```

```
The accuracy score is 0.942
```

[27]:
```
f1_test = f1_score(y_test, y_pred, pos_label = "satisfied")
print("The F1 score is {f1:.3f}".format(f1 = f1_test))
```

```
The F1 score is 0.947
```

[28]:
```
print("\nThe precision score is: {pc:.3f}".format(pc = pc_test), "for the test␣
 ↪set,", "\nwhich means of all positive predictions,", "{pc_pct:.1f}%␣
 ↪prediction are true positive.".format(pc_pct = pc_test * 100))
```

```

The precision score is: 0.950 for the test set,
which means of all positive predictions, 95.0% prediction are true positive.
```

[29]:

```
print("\nThe recall score is: {rc:.3f}".format(rc = rc_test), "for the test␣
↪set,", "\nwhich means of which means of all real positive cases in test␣
↪set,", "{rc_pct:.1f}% are  predicted positive.".format(rc_pct = rc_test *␣
↪100))
```

```

The recall score is: 0.945 for the test set,
which means of which means of all real positive cases in test set, 94.5% are
predicted positive.
```

[30]:

```
print("\nThe accuracy score is: {ac:.3f}".format(ac = ac_test), "for the test␣
↪set,", "\nwhich means of all cases in test set,", "{ac_pct:.1f}% are␣
↪predicted true positive or true negative.".format(ac_pct = ac_test * 100))
```

```

The accuracy score is: 0.942 for the test set,
which means of all cases in test set, 94.2% are predicted true positive or true
negative.
```

[31]:

```
print("\nThe F1 score is: {f1:.3f}".format(f1 = f1_test), "for the test set,",␣
↪"\nwhich means the test set's harmonic mean is {f1_pct:.1f}%.".format(f1_pct␣
↪= f1_test * 100))
```

```

The F1 score is: 0.947 for the test set,
which means the test set's harmonic mean is 94.7%.
```

[32]:

```
table = pd.DataFrame({'Model': ["Tuned Decision Tree", "Tuned Random Forest"],
                        'F1':  [0.945422, f1_test],
                        'Recall': [0.935863, rc_test],
                        'Precision': [0.955197, pc_test],
                        'Accuracy': [0.940864, ac_test]
                      }
                    )
table
```

[32]:

```
                 Model        F1    Recall  Precision  Accuracy
0  Tuned Decision Tree  0.945422  0.935863   0.955197  0.940864
1  Tuned Random Forest  0.947306  0.944501   0.950128  0.942450
```