Titanic

July 30, 2022

1 Titanic Kaggle Challenge

1.1 Downloading the Dataset

```
[438]: import os
       import opendatasets as od
       import pandas as pd
       import numpy as np
       import matplotlib.pyplot as plt
       import seaborn as sns
       import plotly.express as px
       %matplotlib inline
       sns.set_style('darkgrid')
```

```
[439]: os.listdir('.')
```

```
[439]: ['gender_submission.csv', 'test.csv', 'Titanic.ipynb', 'train.csv']
```

```
[440]: data_df = pd.read_csv('./train.csv')
       test_df = pd.read_csv('./test.csv')
```

```
[441]: data_df
```

```
[441]:      PassengerId  Survived  Pclass  \
       0              1         0       3
       1              2         1       1
       2              3         1       3
       3              4         1       1
       4              5         0       3
       ..           …         …     …
       886          887         0       2
       887          888         1       1
       888          889         0       3
       889          890         1       1
       890          891         0       3

                                                        Name     Sex   Age  SibSp  \
       0                               Braund, Mr. Owen Harris    male  22.0      1
```

```
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                 …     …   …    …
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

[442]:

```
test_df
```

[442]:

```
     PassengerId  Pclass                                          Name  \
0            892       3                              Kelly, Mr. James
1            893       3              Wilkes, Mrs. James (Ellen Needs)
2            894       2                     Myles, Mr. Thomas Francis
3            895       3                              Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..             …     …                                           …
413         1305       3                            Spector, Mr. Woolf
414         1306       1                  Oliva y Ocana, Dona. Fermina
415         1307       3                  Saether, Mr. Simon Sivertsen
416         1308       3                           Ware, Mr. Frederick
417         1309       3                      Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket     Fare Cabin Embarked
0     male  34.5      0      0              330911     7.8292   NaN        Q
1   female  47.0      1      0              363272     7.0000   NaN        S
2     male  62.0      0      0              240276     9.6875   NaN        Q
3     male  27.0      0      0              315154     8.6625   NaN        S
4   female  22.0      1      1             3101298    12.2875   NaN        S
```

```
..      …   …    …    …                  …       …   …      …
413    male   NaN      0      0            A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0             PC 17758  108.9000  C105        C
415    male  38.5      0      0   SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0               359309    8.0500   NaN        S
417    male   NaN      1      1                 2668   22.3583   NaN        C

[418 rows x 11 columns]
```

1.2 Analyzing Data

[443]:
```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

[444]:
```
data_df.describe()
```

[444]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
```

```
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[445]:

```
data_df.corr()
```

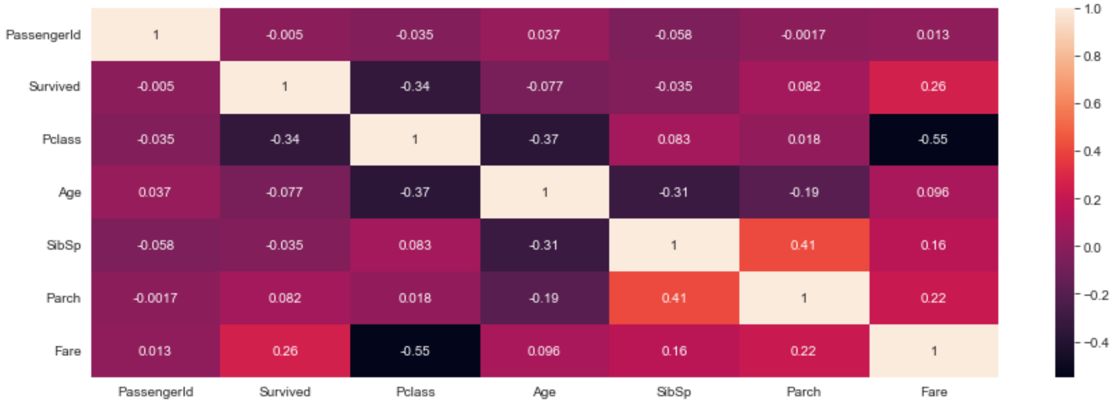
[445]:

```
             PassengerId  Survived    Pclass       Age     SibSp     Parch  \
PassengerId     1.000000 -0.005007 -0.035144  0.036847 -0.057527 -0.001652
Survived       -0.005007  1.000000 -0.338481 -0.077221 -0.035322  0.081629
Pclass         -0.035144 -0.338481  1.000000 -0.369226  0.083081  0.018443
Age             0.036847 -0.077221 -0.369226  1.000000 -0.308247 -0.189119
SibSp          -0.057527 -0.035322  0.083081 -0.308247  1.000000  0.414838
Parch          -0.001652  0.081629  0.018443 -0.189119  0.414838  1.000000
Fare            0.012658  0.257307 -0.549500  0.096067  0.159651  0.216225

                 Fare
PassengerId  0.012658
Survived     0.257307
Pclass      -0.549500
Age          0.096067
SibSp        0.159651
Parch        0.216225
Fare         1.000000
```

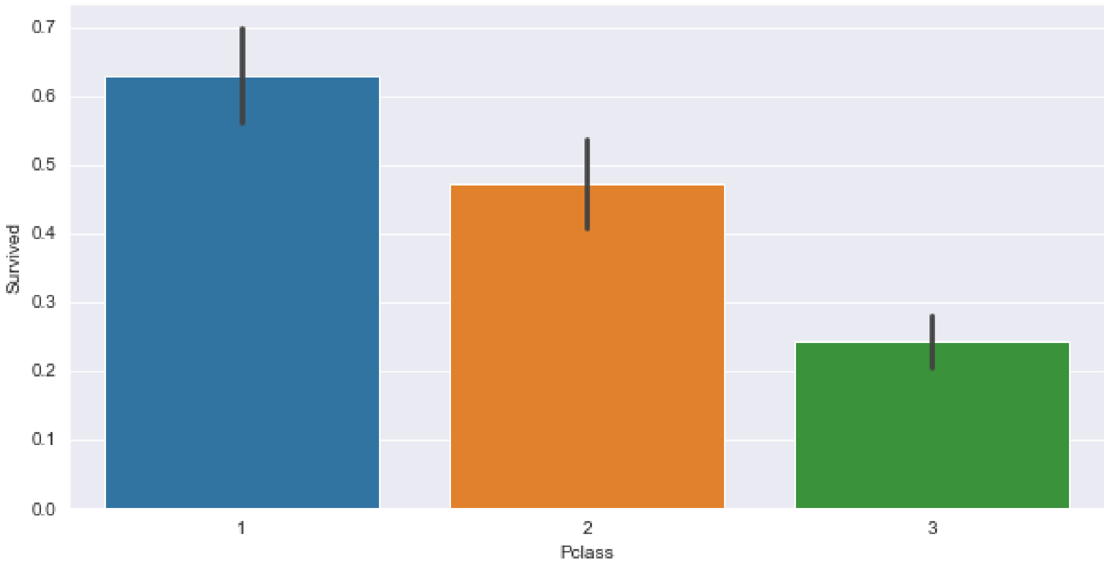
[446]:

```
plt.figure(figsize=(15,5))
sns.heatmap(data_df.corr(), annot=True);
```



[447]:

```
plt.figure(figsize=(10,5))
sns.barplot(data=data_df, x='Pclass', y='Survived');
```



The higher the class, the higher the possibility to survive.

[448]:
```
data_df.Fare.describe()
```

[448]:
```
count    891.000000
mean      32.204208
std       49.693429
min        0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: Fare, dtype: float64
```

[449]:
```
data_df.Fare.value_counts()
```

[449]:
```
8.0500     43
13.0000    42
7.8958     38
7.7500     34
26.0000    31
           ..
35.0000     1
28.5000     1
6.2375      1
14.0000     1
10.5167     1
Name: Fare, Length: 248, dtype: int64
```

[450]:

```
fig = px.histogram(data_df, x='Fare', y='Survived', nbins=100, color='Pclass')

fig.update_layout(bargap=0.1)
fig.show();
```

The higher the class, the higher the price paid for the fee.

1.3 Dataframe Adjustments

[451]:

```
data_df
```

[451]:

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age  SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                …     …   …    …
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
```

```
889      0             111369  30.0000  C148        C
890      0             370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

```
[452]: test_df
```

```
[452]:      PassengerId  Pclass                                          Name  \
       0            892       3                              Kelly, Mr. James
       1            893       3              Wilkes, Mrs. James (Ellen Needs)
       2            894       2                     Myles, Mr. Thomas Francis
       3            895       3                              Wirz, Mr. Albert
       4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
       ..           …     …                                             …
       413         1305       3                            Spector, Mr. Woolf
       414         1306       1                  Oliva y Ocana, Dona. Fermina
       415         1307       3                  Saether, Mr. Simon Sivertsen
       416         1308       3                           Ware, Mr. Frederick
       417         1309       3                     Peter, Master. Michael J

               Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
       0      male  34.5      0      0              330911    7.8292   NaN        Q
       1    female  47.0      1      0              363272    7.0000   NaN        S
       2      male  62.0      0      0              240276    9.6875   NaN        Q
       3      male  27.0      0      0              315154    8.6625   NaN        S
       4    female  22.0      1      1             3101298   12.2875   NaN        S
       ..      …   …    …    …                 …       …   …       …
       413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
       414  female  39.0      0      0            PC 17758  108.9000  C105        C
       415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
       416    male   NaN      0      0              359309    8.0500   NaN        S
       417    male   NaN      1      1                2668   22.3583   NaN        C

       [418 rows x 11 columns]
```

```
[453]: survived_binary = {1: 'yes', 0: 'no'}

       survived_binary_col = data_df['Survived'].map(survived_binary)
```

```
[454]: data_df['SurvivedBinary'] = survived_binary_col
```

```
[455]: data_df
```

```
[455]:      PassengerId  Survived  Pclass  \
       0              1         0       3
       1              2         1       1
       2              3         1       3
```

```
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age  SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                   …     …   …    …
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked SurvivedBinary
0        0         A/5 21171   7.2500   NaN        S             no
1        0          PC 17599  71.2833   C85        C            yes
2        0  STON/O2. 3101282   7.9250   NaN        S            yes
3        0            113803  53.1000  C123        S            yes
4        0            373450   8.0500   NaN        S             no
..     …               …       …   …      …            …
886      0            211536  13.0000   NaN        S             no
887      0            112053  30.0000   B42        S            yes
888      2        W./C. 6607  23.4500   NaN        S             no
889      0            111369  30.0000  C148        C            yes
890      0            370376   7.7500   NaN        Q             no

[891 rows x 13 columns]
```

[456]:

```python
name_col = data_df.pop('Name')
```

[457]:

```python
data_df.insert(2, 'Name', name_col)
```

[458]:

```python
data_df
```

[458]:

```
     PassengerId  Survived                                               Name  \
0              1         0                            Braund, Mr. Owen Harris
1              2         1  Cumings, Mrs. John Bradley (Florence Briggs Th…
2              3         1                             Heikkinen, Miss. Laina
```

```
3              4         1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
4              5         0                           Allen, Mr. William Henry
..           …       …                                                 …
886          887         0                              Montvila, Rev. Juozas
887          888         1                       Graham, Miss. Margaret Edith
888          889         0           Johnston, Miss. Catherine Helen "Carrie"
889          890         1                              Behr, Mr. Karl Howell
890          891         0                                Dooley, Mr. Patrick

     Pclass     Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
0         3    male  22.0      1      0         A/5 21171   7.2500   NaN
1         1  female  38.0      1      0          PC 17599  71.2833   C85
2         3  female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
3         1  female  35.0      1      0            113803  53.1000  C123
4         3    male  35.0      0      0            373450   8.0500   NaN
..      …     …   …    …    …                …      …   …
886       2    male  27.0      0      0            211536  13.0000   NaN
887       1  female  19.0      0      0            112053  30.0000   B42
888       3  female   NaN      1      2        W./C. 6607  23.4500   NaN
889       1    male  26.0      0      0            111369  30.0000  C148
890       3    male  32.0      0      0            370376   7.7500   NaN

    Embarked SurvivedBinary
0          S             no
1          C            yes
2          S            yes
3          S            yes
4          S             no
..       …             …
886        S             no
887        S            yes
888        S             no
889        C            yes
890        Q             no

[891 rows x 13 columns]
```

[459]:

```
ticket = data_df.pop('Ticket')
cabin = data_df.pop('Cabin')
```

[460]:

```
data_df.insert(3, 'Ticket', ticket)
data_df.insert(4, 'Cabin', cabin)
```

[461]:

```
data_df
```

[461]:

```
     PassengerId  Survived                                               Name  \
0              1         0                            Braund, Mr. Owen Harris
```

```
1              2         1  Cumings, Mrs. John Bradley (Florence Briggs Th…
2              3         1                             Heikkinen, Miss. Laina
3              4         1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
4              5         0                           Allen, Mr. William Henry
..           …       …                                                …
886          887         0                              Montvila, Rev. Juozas
887          888         1                       Graham, Miss. Margaret Edith
888          889         0           Johnston, Miss. Catherine Helen "Carrie"
889          890         1                              Behr, Mr. Karl Howell
890          891         0                                Dooley, Mr. Patrick

               Ticket Cabin  Pclass     Sex   Age  SibSp  Parch     Fare  \
0           A/5 21171   NaN       3    male  22.0      1      0   7.2500
1            PC 17599   C85       1  female  38.0      1      0  71.2833
2    STON/O2. 3101282   NaN       3  female  26.0      0      0   7.9250
3              113803  C123       1  female  35.0      1      0  53.1000
4              373450   NaN       3    male  35.0      0      0   8.0500
..                …   …     …     …   …    …    …      …
886            211536   NaN       2    male  27.0      0      0  13.0000
887            112053   B42       1  female  19.0      0      0  30.0000
888        W./C. 6607   NaN       3  female   NaN      1      2  23.4500
889            111369  C148       1    male  26.0      0      0  30.0000
890            370376   NaN       3    male  32.0      0      0   7.7500

    Embarked SurvivedBinary
0          S             no
1          C            yes
2          S            yes
3          S            yes
4          S             no
..       …            …
886        S             no
887        S            yes
888        S             no
889        C            yes
890        Q             no

[891 rows x 13 columns]
```

[462]:

```
sex_binary = {'male':0, 'female':1}

sex_binary_col=data_df.Sex.map(sex_binary)
```

[463]:

```
data_df.insert(10, 'SexBinary', sex_binary_col)
```

[464]:

```
data_df.pop('Sex')
```

[464]:
```
0        male
1      female
2      female
3      female
4        male
        …
886      male
887    female
888    female
889      male
890      male
Name: Sex, Length: 891, dtype: object
```

[465]:
```
data_df
```

[465]:
```
     PassengerId  Survived                                               Name  \
0              1         0                            Braund, Mr. Owen Harris
1              2         1  Cumings, Mrs. John Bradley (Florence Briggs Th…
2              3         1                             Heikkinen, Miss. Laina
3              4         1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
4              5         0                           Allen, Mr. William Henry
..           …       …                                                …
886          887         0                              Montvila, Rev. Juozas
887          888         1                       Graham, Miss. Margaret Edith
888          889         0           Johnston, Miss. Catherine Helen "Carrie"
889          890         1                              Behr, Mr. Karl Howell
890          891         0                                Dooley, Mr. Patrick

               Ticket Cabin  Pclass   Age  SibSp  Parch  SexBinary     Fare  \
0           A/5 21171   NaN       3  22.0      1      0          0   7.2500
1            PC 17599   C85       1  38.0      1      0          1  71.2833
2    STON/O2. 3101282   NaN       3  26.0      0      0          1   7.9250
3              113803  C123       1  35.0      1      0          1  53.1000
4              373450   NaN       3  35.0      0      0          0   8.0500
..                …   …     …   …    …    …        …      …
886            211536   NaN       2  27.0      0      0          0  13.0000
887            112053   B42       1  19.0      0      0          1  30.0000
888        W./C. 6607   NaN       3   NaN      1      2          1  23.4500
889            111369  C148       1  26.0      0      0          0  30.0000
890            370376   NaN       3  32.0      0      0          0   7.7500

    Embarked SurvivedBinary
0          S             no
1          C            yes
2          S            yes
3          S            yes
4          S             no
```

```
..       …            …
886         S              no
887         S             yes
888         S              no
889         C             yes
890         Q              no

[891 rows x 13 columns]
```

[466]:

```
test_df
```

[466]:

```
     PassengerId  Pclass                                          Name  \
0            892       3                              Kelly, Mr. James
1            893       3              Wilkes, Mrs. James (Ellen Needs)
2            894       2                     Myles, Mr. Thomas Francis
3            895       3                              Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..             …     …                                             …
413         1305       3                            Spector, Mr. Woolf
414         1306       1                  Oliva y Ocana, Dona. Fermina
415         1307       3                  Saether, Mr. Simon Sivertsen
416         1308       3                           Ware, Mr. Frederick
417         1309       3                      Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0      male  34.5      0      0              330911    7.8292   NaN        Q
1    female  47.0      1      0              363272    7.0000   NaN        S
2      male  62.0      0      0              240276    9.6875   NaN        Q
3      male  27.0      0      0              315154    8.6625   NaN        S
4    female  22.0      1      1             3101298   12.2875   NaN        S
..      …   …    …    …                  …       …   …      …
413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0            PC 17758  108.9000  C105        C
415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0              359309    8.0500   NaN        S
417    male   NaN      1      1                2668   22.3583   NaN        C

[418 rows x 11 columns]
```

[467]:

```
test_df.insert(2, 'Name' ,test_df.pop('Name'))
```

[468]:

```
test_df
```

[468]:

```
     PassengerId  Pclass                                          Name  \
0            892       3                              Kelly, Mr. James
1            893       3              Wilkes, Mrs. James (Ellen Needs)
2            894       2                     Myles, Mr. Thomas Francis
```

```
3            895       3                              Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..           …     …                                           …
413         1305       3                            Spector, Mr. Woolf
414         1306       1                  Oliva y Ocana, Dona. Fermina
415         1307       3                  Saether, Mr. Simon Sivertsen
416         1308       3                           Ware, Mr. Frederick
417         1309       3                      Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0      male  34.5      0      0              330911    7.8292   NaN        Q
1    female  47.0      1      0              363272    7.0000   NaN        S
2      male  62.0      0      0              240276    9.6875   NaN        Q
3      male  27.0      0      0              315154    8.6625   NaN        S
4    female  22.0      1      1             3101298   12.2875   NaN        S
..      …   …    …    …                 …       …   …      …
413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0            PC 17758  108.9000  C105        C
415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0              359309    8.0500   NaN        S
417    male   NaN      1      1                2668   22.3583   NaN        C

[418 rows x 11 columns]
```

[469]:

```
test_df.insert(3, 'Pclass' ,test_df.pop('Pclass'))
```

[470]:

```
test_df
```

[470]:

```
     PassengerId                                          Name     Sex  \
0            892                              Kelly, Mr. James    male
1            893              Wilkes, Mrs. James (Ellen Needs)  female
2            894                     Myles, Mr. Thomas Francis    male
3            895                              Wirz, Mr. Albert    male
4            896  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female
..           …                                           …     …
413         1305                            Spector, Mr. Woolf    male
414         1306                  Oliva y Ocana, Dona. Fermina  female
415         1307                  Saether, Mr. Simon Sivertsen    male
416         1308                           Ware, Mr. Frederick    male
417         1309                      Peter, Master. Michael J    male

     Pclass   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0         3  34.5      0      0              330911    7.8292   NaN        Q
1         3  47.0      1      0              363272    7.0000   NaN        S
2         2  62.0      0      0              240276    9.6875   NaN        Q
3         3  27.0      0      0              315154    8.6625   NaN        S
4         3  22.0      1      1             3101298   12.2875   NaN        S
```

```
..      …   …    …    …                 …       …    …      …
413       3   NaN      0      0           A.5. 3236    8.0500    NaN        S
414       1  39.0      0      0            PC 17758  108.9000   C105        C
415       3  38.5      0      0  SOTON/O.Q. 3101262    7.2500    NaN        S
416       3   NaN      0      0              359309    8.0500    NaN        S
417       3   NaN      1      1                2668   22.3583    NaN        C

[418 rows x 11 columns]
```

[471]:

```
test_df.insert(2, 'Ticket' ,test_df.pop('Ticket'))
test_df.insert(3, 'Cabin', test_df.pop('Cabin'))
test_df.insert(8, 'SexBinary', sex_binary_col)
```

[472]:

```
test_df
```

[472]:

```
     PassengerId                                          Name  \
0            892                              Kelly, Mr. James
1            893              Wilkes, Mrs. James (Ellen Needs)
2            894                     Myles, Mr. Thomas Francis
3            895                              Wirz, Mr. Albert
4            896  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..           …                                           …
413         1305                            Spector, Mr. Woolf
414         1306                  Oliva y Ocana, Dona. Fermina
415         1307                  Saether, Mr. Simon Sivertsen
416         1308                           Ware, Mr. Frederick
417         1309                      Peter, Master. Michael J

                 Ticket Cabin     Sex  Pclass   Age  SibSp  SexBinary  Parch  \
0                330911   NaN    male       3  34.5      0          0      0
1                363272   NaN  female       3  47.0      1          1      0
2                240276   NaN    male       2  62.0      0          1      0
3                315154   NaN    male       3  27.0      0          1      0
4               3101298   NaN  female       3  22.0      1          0      1
..                  …   …     …     …   …    …        …    …
413           A.5. 3236   NaN    male       3   NaN      0          0      0
414            PC 17758  C105  female       1  39.0      0          0      0
415  SOTON/O.Q. 3101262   NaN    male       3  38.5      0          1      0
416              359309   NaN    male       3   NaN      0          1      0
417                2668   NaN    male       3   NaN      1          1      1

         Fare Embarked
0      7.8292        Q
1      7.0000        S
2      9.6875        Q
3      8.6625        S
4     12.2875        S
```

```
..        …      …
413    8.0500         S
414  108.9000         C
415    7.2500         S
416    8.0500         S
417   22.3583         C

[418 rows x 12 columns]
```

[473]:

```
data_df
```

[473]:

```
     PassengerId  Survived                                                Name  \
0              1         0                             Braund, Mr. Owen Harris
1              2         1  Cumings, Mrs. John Bradley (Florence Briggs Th…
2              3         1                              Heikkinen, Miss. Laina
3              4         1        Futrelle, Mrs. Jacques Heath (Lily May Peel)
4              5         0                            Allen, Mr. William Henry
..           …       …                                                 …
886          887         0                               Montvila, Rev. Juozas
887          888         1                        Graham, Miss. Margaret Edith
888          889         0            Johnston, Miss. Catherine Helen "Carrie"
889          890         1                               Behr, Mr. Karl Howell
890          891         0                                 Dooley, Mr. Patrick

               Ticket Cabin  Pclass   Age  SibSp  Parch  SexBinary      Fare  \
0           A/5 21171   NaN       3  22.0      1      0          0    7.2500
1            PC 17599   C85       1  38.0      1      0          1   71.2833
2    STON/O2. 3101282   NaN       3  26.0      0      0          1    7.9250
3              113803  C123       1  35.0      1      0          1   53.1000
4              373450   NaN       3  35.0      0      0          0    8.0500
..                 …   …     …   …    …    …        …       …
886            211536   NaN       2  27.0      0      0          0   13.0000
887            112053   B42       1  19.0      0      0          1   30.0000
888        W./C. 6607   NaN       3   NaN      1      2          1   23.4500
889            111369  C148       1  26.0      0      0          0   30.0000
890            370376   NaN       3  32.0      0      0          0    7.7500

    Embarked SurvivedBinary
0          S             no
1          C            yes
2          S            yes
3          S            yes
4          S             no
..        …            …
886        S             no
887        S            yes
888        S             no
```

```
889        C            yes
890        Q             no

[891 rows x 13 columns]
```

1.4 Input and Target Columns

[474]:

```
input_cols = list(data_df.columns[5:12])
target_col = 'SurvivedBinary'
```

[475]:

```
input_cols
```

[475]:

```
['Pclass', 'Age', 'SibSp', 'Parch', 'SexBinary', 'Fare', 'Embarked']
```

[476]:

```
target_col
```

[476]:

```
'SurvivedBinary'
```

[477]:

```
input_df = data_df[input_cols].copy()
target_df = data_df[target_col].copy()
```

[478]:

```
input_df
```

[478]:

```
     Pclass   Age  SibSp  Parch  SexBinary     Fare Embarked
0         3  22.0      1      0          0   7.2500        S
1         1  38.0      1      0          1  71.2833        C
2         3  26.0      0      0          1   7.9250        S
3         1  35.0      1      0          1  53.1000        S
4         3  35.0      0      0          0   8.0500        S
..      …     …    …      …          …        …      …
886       2  27.0      0      0          0  13.0000        S
887       1  19.0      0      0          1  30.0000        S
888       3   NaN      1      2          1  23.4500        S
889       1  26.0      0      0          0  30.0000        C
890       3  32.0      0      0          0   7.7500        Q

[891 rows x 7 columns]
```

[479]:

```
target_df
```

[479]:

```
0       no
1      yes
2      yes
3      yes
4       no
      …
886     no
887    yes
```

```
888     no
889    yes
890     no
Name: SurvivedBinary, Length: 891, dtype: object
```

1.5 Numerical and Categorical Columns

[480]:
```
numerical_cols = input_df.select_dtypes(include=np.number).columns.tolist()
categorical_cols = input_df.select_dtypes(include='object').columns.tolist()
```

[481]:
```
numerical_cols
```

[481]:
```
['Pclass', 'Age', 'SibSp', 'Parch', 'SexBinary', 'Fare']
```

[482]:
```
categorical_cols
```

[482]:
```
['Embarked']
```

1.6 Feature Engineering

Scaling Numerical Columns

[483]:
```
from sklearn.preprocessing import MinMaxScaler
```

[484]:
```
scaler = MinMaxScaler()
scaler.fit(input_df[numerical_cols])
```

[484]:
```
MinMaxScaler()
```

[485]:
```
input_df[numerical_cols] = scaler.transform(input_df[numerical_cols])
test_df[numerical_cols] = scaler.transform(test_df[numerical_cols])
```

[486]:
```
input_df
```

[486]:
```
     Pclass       Age  SibSp     Parch  SexBinary      Fare Embarked
0       1.0  0.271174  0.125  0.000000        0.0  0.014151        S
1       0.0  0.472229  0.125  0.000000        1.0  0.139136        C
2       1.0  0.321438  0.000  0.000000        1.0  0.015469        S
3       0.0  0.434531  0.125  0.000000        1.0  0.103644        S
4       1.0  0.434531  0.000  0.000000        0.0  0.015713        S
..      …         …     …        …          …       …
886     0.5  0.334004  0.000  0.000000        0.0  0.025374        S
887     0.0  0.233476  0.000  0.000000        1.0  0.058556        S
888     1.0       NaN  0.125  0.333333        1.0  0.045771        S
889     0.0  0.321438  0.000  0.000000        0.0  0.058556        C
890     1.0  0.396833  0.000  0.000000        0.0  0.015127        Q

[891 rows x 7 columns]
```

```
[487]: test_df[numerical_cols]
```

```
[487]:      Pclass       Age  SibSp     Parch  SexBinary      Fare
       0       1.0  0.428248  0.000  0.000000        0.0  0.015282
       1       1.0  0.585323  0.125  0.000000        1.0  0.013663
       2       0.5  0.773813  0.000  0.000000        1.0  0.018909
       3       1.0  0.334004  0.000  0.000000        1.0  0.016908
       4       1.0  0.271174  0.125  0.166667        0.0  0.023984
       ..      …         …      …         …          …         …
       413     1.0       NaN  0.000  0.000000        0.0  0.015713
       414     0.0  0.484795  0.000  0.000000        0.0  0.212559
       415     1.0  0.478512  0.000  0.000000        1.0  0.014151
       416     1.0       NaN  0.000  0.000000        1.0  0.015713
       417     1.0       NaN  0.125  0.166667        1.0  0.043640

       [418 rows x 6 columns]
```

Imputing Missing Values

```
[488]: from sklearn.impute import SimpleImputer
```

```
[489]: input_df.isna().sum()
```

```
[489]: Pclass         0
       Age          177
       SibSp          0
       Parch          0
       SexBinary      0
       Fare           0
       Embarked       2
       dtype: int64
```

```
[490]: input_df.Embarked.value_counts()
```

```
[490]: S    644
       C    168
       Q     77
       Name: Embarked, dtype: int64
```

```
[491]: input_df.Embarked.fillna('Q', inplace=True)
```

```
[492]: input_df.Embarked.value_counts()
```

```
[492]: S    644
       C    168
       Q     79
       Name: Embarked, dtype: int64
```

```
[493]: test_df.isna().sum()
```

```
[493]: PassengerId      0
       Name             0
       Ticket           0
       Cabin          327
       Sex              0
       Pclass           0
       Age             86
       SibSp            0
       SexBinary        0
       Parch            0
       Fare             1
       Embarked         0
       dtype: int64
```

```
[494]: imputer = SimpleImputer(strategy='mean')
       imputer.fit(input_df[numerical_cols])
```

```
[494]: SimpleImputer()
```

```
[495]: input_df[numerical_cols] = imputer.transform(input_df[numerical_cols])
       test_df[numerical_cols] = imputer.transform(test_df[numerical_cols])
```

```
[496]: input_df.isna().sum()
```

```
[496]: Pclass       0
       Age          0
       SibSp        0
       Parch        0
       SexBinary    0
       Fare         0
       Embarked     0
       dtype: int64
```

```
[497]: test_df.isna().sum()
```

```
[497]: PassengerId      0
       Name             0
       Ticket           0
       Cabin          327
       Sex              0
       Pclass           0
       Age              0
       SibSp            0
       SexBinary        0
       Parch            0
       Fare             0
```

```
Embarked         0
dtype: int64
```

One Hot Encoding

[498]:
```
from sklearn.preprocessing import OneHotEncoder
```

[499]:
```
encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')
encoder.fit(input_df[categorical_cols])
```

[499]:
```
OneHotEncoder(handle_unknown='ignore', sparse=False)
```

[500]:
```
encoded_cols = list(encoder.get_feature_names_out(categorical_cols))
encoded_cols
```

[500]:
```
['Embarked_C', 'Embarked_Q', 'Embarked_S']
```

[501]:
```
input_df[encoded_cols] = encoder.transform(input_df[categorical_cols])
test_df[encoded_cols] = encoder.transform(test_df[categorical_cols])
```

[502]:
```
input_df
```

[502]:
```
     Pclass       Age  SibSp     Parch  SexBinary      Fare Embarked  \
0       1.0  0.271174  0.125  0.000000        0.0  0.014151        S
1       0.0  0.472229  0.125  0.000000        1.0  0.139136        C
2       1.0  0.321438  0.000  0.000000        1.0  0.015469        S
3       0.0  0.434531  0.125  0.000000        1.0  0.103644        S
4       1.0  0.434531  0.000  0.000000        0.0  0.015713        S
..      …         …     …         …          …         …      …
886     0.5  0.334004  0.000  0.000000        0.0  0.025374        S
887     0.0  0.233476  0.000  0.000000        1.0  0.058556        S
888     1.0  0.367921  0.125  0.333333        1.0  0.045771        S
889     0.0  0.321438  0.000  0.000000        0.0  0.058556        C
890     1.0  0.396833  0.000  0.000000        0.0  0.015127        Q

     Embarked_C  Embarked_Q  Embarked_S
0           0.0         0.0         1.0
1           1.0         0.0         0.0
2           0.0         0.0         1.0
3           0.0         0.0         1.0
4           0.0         0.0         1.0
..          …           …           …
886         0.0         0.0         1.0
887         0.0         0.0         1.0
888         0.0         0.0         1.0
889         1.0         0.0         0.0
890         0.0         1.0         0.0
```

```
[891 rows x 10 columns]
```

[503]:
```
test_df
```

[503]:
```
     PassengerId                                          Name  \
0            892                              Kelly, Mr. James   
1            893              Wilkes, Mrs. James (Ellen Needs)   
2            894                     Myles, Mr. Thomas Francis   
3            895                              Wirz, Mr. Albert   
4            896  Hirvonen, Mrs. Alexander (Helga E Lindqvist)   
..           …                                            …
413         1305                            Spector, Mr. Woolf   
414         1306                  Oliva y Ocana, Dona. Fermina   
415         1307                  Saether, Mr. Simon Sivertsen   
416         1308                           Ware, Mr. Frederick   
417         1309                      Peter, Master. Michael J   

                 Ticket Cabin     Sex  Pclass       Age  SibSp  SexBinary  \
0                330911   NaN    male     1.0  0.428248  0.000        0.0   
1                363272   NaN  female     1.0  0.585323  0.125        1.0   
2                240276   NaN    male     0.5  0.773813  0.000        1.0   
3                315154   NaN    male     1.0  0.334004  0.000        1.0   
4               3101298   NaN  female     1.0  0.271174  0.125        0.0   
..                  …   …     …     …       …     …       …
413           A.5. 3236   NaN    male     1.0  0.367921  0.000        0.0   
414            PC 17758  C105  female     0.0  0.484795  0.000        0.0   
415  SOTON/O.Q. 3101262   NaN    male     1.0  0.478512  0.000        1.0   
416              359309   NaN    male     1.0  0.367921  0.000        1.0   
417                2668   NaN    male     1.0  0.367921  0.125        1.0   

        Parch      Fare Embarked  Embarked_C  Embarked_Q  Embarked_S  
0    0.000000  0.015282        Q         0.0         1.0         0.0  
1    0.000000  0.013663        S         0.0         0.0         1.0  
2    0.000000  0.018909        Q         0.0         1.0         0.0  
3    0.000000  0.016908        S         0.0         0.0         1.0  
4    0.166667  0.023984        S         0.0         0.0         1.0  
..        …       …       …        …         …         …
413  0.000000  0.015713        S         0.0         0.0         1.0  
414  0.000000  0.212559        C         1.0         0.0         0.0  
415  0.000000  0.014151        S         0.0         0.0         1.0  
416  0.000000  0.015713        S         0.0         0.0         1.0  
417  0.166667  0.043640        C         1.0         0.0         0.0  

[418 rows x 15 columns]
```

2 Creating a Training and a Validation Set

[504]:
```
from sklearn.model_selection import train_test_split
```

[505]:
```
train_inputs, val_inputs, train_targets, val_targets =␣
 ↪train_test_split(input_df[numerical_cols+encoded_cols], target_df,␣
 ↪test_size=0.20, random_state=42)
```

[506]:
```
train_inputs
```

[506]:
```
     Pclass       Age  SibSp     Parch  SexBinary      Fare  Embarked_C  \
331     0.0  0.566474  0.000  0.000000        0.0  0.055628         0.0
733     0.5  0.283740  0.000  0.000000        0.0  0.025374         0.0
382     1.0  0.396833  0.000  0.000000        0.0  0.015469         0.0
704     1.0  0.321438  0.125  0.000000        0.0  0.015330         0.0
813     1.0  0.070118  0.500  0.333333        1.0  0.061045         0.0
..      …         …    …       …         …        …           …
106     1.0  0.258608  0.000  0.000000        1.0  0.014932         0.0
270     0.0  0.367921  0.000  0.000000        0.0  0.060508         0.0
860     1.0  0.509927  0.250  0.000000        0.0  0.027538         0.0
435     0.0  0.170646  0.125  0.333333        1.0  0.234224         0.0
102     0.0  0.258608  0.000  0.166667        0.0  0.150855         0.0

     Embarked_Q  Embarked_S
331         0.0         1.0
733         0.0         1.0
382         0.0         1.0
704         0.0         1.0
813         0.0         1.0
..          …           …
106         0.0         1.0
270         0.0         1.0
860         0.0         1.0
435         0.0         1.0
102         0.0         1.0

[712 rows x 9 columns]
```

[507]:
```
val_inputs
```

[507]:
```
     Pclass       Age  SibSp     Parch  SexBinary      Fare  Embarked_C  \
709     1.0  0.367921  0.125  0.166667        0.0  0.029758         1.0
439     0.5  0.384267  0.000  0.000000        0.0  0.020495         0.0
840     1.0  0.246042  0.000  0.000000        0.0  0.015469         0.0
720     0.5  0.070118  0.000  0.166667        1.0  0.064412         0.0
39      1.0  0.170646  0.125  0.000000        1.0  0.021942         1.0
..      …         …    …       …         …        …           …
```

```
433     1.0  0.208344  0.000  0.000000        0.0  0.013907         0.0
773     1.0  0.367921  0.000  0.000000        0.0  0.014102         1.0
25      1.0  0.472229  0.125  0.833333        1.0  0.061264         0.0
84      0.5  0.208344  0.000  0.000000        1.0  0.020495         0.0
10      1.0  0.044986  0.125  0.166667        1.0  0.032596         0.0

     Embarked_Q  Embarked_S
709         0.0         0.0
439         0.0         1.0
840         0.0         1.0
720         0.0         1.0
39          0.0         0.0
..          …         …
433         0.0         1.0
773         0.0         0.0
25          0.0         1.0
84          0.0         1.0
10          0.0         1.0

[179 rows x 9 columns]
```

[508]:

```
train_targets
```

[508]:

```
331     no
733     no
382     no
704     no
813     no
      …
106    yes
270     no
860     no
435    yes
102     no
Name: SurvivedBinary, Length: 712, dtype: object
```

[509]:

```
val_targets
```

[509]:

```
709    yes
439     no
840     no
720    yes
39     yes
      …
433     no
773     no
25     yes
```

```
84     yes
10     yes
Name: SurvivedBinary, Length: 179, dtype: object
```

```
[510]: X_train = train_inputs[numerical_cols+encoded_cols]
       X_val = val_inputs[numerical_cols+encoded_cols]
       X_test = test_df[numerical_cols+encoded_cols]
```

3 Creating the Model

The chosen model is Random Forest.

```
[511]: from sklearn.ensemble import RandomForestClassifier
```

```
[512]: %%time
       ran_for_model = RandomForestClassifier(n_jobs=-1, random_state=42)
       ran_for_model.fit(X_train, train_targets)
```

```
CPU times: total: 281 ms
Wall time: 161 ms
```

```
[512]: RandomForestClassifier(n_jobs=-1, random_state=42)
```

Accuracies

```
[513]: ran_for_model_base_accs = ran_for_model.score(X_train, train_targets),␣
         ↪ran_for_model.score(X_val, val_targets)
```

```
[514]: ran_for_model_base_accs
```

```
[514]: (0.9803370786516854, 0.8100558659217877)
```

4 Hypertuning

max_depth

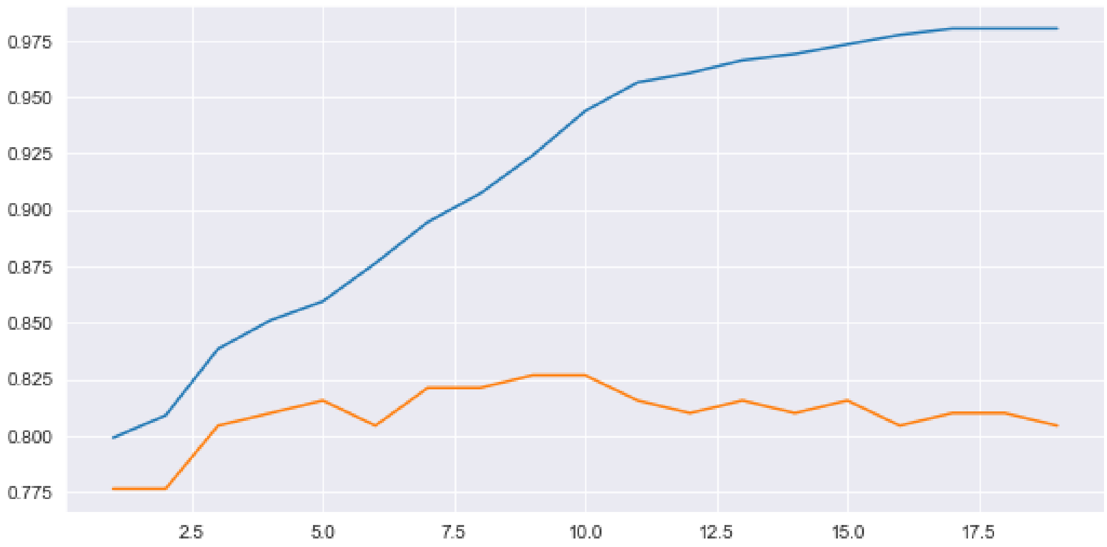
```
[515]: def max_depth_accuracy(md):
           model = RandomForestClassifier(n_jobs=-1, random_state=42, max_depth=md)
           model.fit(X_train, train_targets)
           train_acc = model.score(X_train, train_targets)
           val_acc = model.score(X_val, val_targets)
           return {'MaxDepth': md, 'TrainAccuracy': train_acc, 'ValidationAccuracy':␣
         ↪val_acc}
```

```
[516]: max_depth_acc_df = pd.DataFrame([max_depth_accuracy(md) for md in range(1,20)])
```

```
[517]: max_depth_acc_df.head(10)
```

```
[517]:    MaxDepth  TrainAccuracy  ValidationAccuracy
       0         1       0.799157            0.776536
       1         2       0.808989            0.776536
       2         3       0.838483            0.804469
       3         4       0.851124            0.810056
       4         5       0.859551            0.815642
       5         6       0.876404            0.804469
       6         7       0.894663            0.821229
       7         8       0.907303            0.821229
       8         9       0.924157            0.826816
       9        10       0.943820            0.826816
```

```
[518]: plt.figure(figsize=(10,5))
       plt.plot(max_depth_acc_df['MaxDepth'], max_depth_acc_df['TrainAccuracy'])
       plt.plot(max_depth_acc_df['MaxDepth'], max_depth_acc_df['ValidationAccuracy']);
```



```
[519]: ran_for_model = RandomForestClassifier(n_jobs=-1, random_state=42, max_depth=10)
       ran_for_model.fit(X_train, train_targets)
```

```
[519]: RandomForestClassifier(max_depth=10, n_jobs=-1, random_state=42)
```

```
[520]: ran_for_model_base_accs
```

```
[520]: (0.9803370786516854, 0.8100558659217877)
```

```
[521]: ran_for_model.score(X_train, train_targets), ran_for_model.score(X_val,␣
        ↪val_targets)
```

```
[521]: (0.9438202247191011, 0.8268156424581006)
```

n_estimators

[522]:

```
ran_for_model_estim = RandomForestClassifier(n_jobs=-1, random_state=42,␣
  ↪n_estimators=57, max_depth=9)
ran_for_model_estim.fit(X_train, train_targets)
ran_for_model_estim.score(X_train, train_targets), ran_for_model_estim.
  ↪score(X_val, val_targets)
```

[522]:

```
(0.9199438202247191, 0.8379888268156425)
```

min_impurity_decrease

[523]:

```
ran_for_model_imp = RandomForestClassifier(n_jobs=-1, random_state=42,␣
  ↪n_estimators=57, max_depth=9, min_impurity_decrease=1e-6)
ran_for_model_imp.fit(X_train, train_targets)
ran_for_model_imp.score(X_train, train_targets), ran_for_model_imp.score(X_val,␣
  ↪val_targets)
```

[523]:

```
(0.9199438202247191, 0.8435754189944135)
```

4.0.1 class_weight

[524]:

```
saved=list(ran_for_model_imp.predict(X_train))
```

[525]:

```
len(saved)
```

[525]:

```
712
```

[526]:

```
count = 0
for x in saved:
    if x == 'yes':
        count += 1
```

[527]:

```
count
```

[527]:

```
219
```

[539]:

```
ran_for_model_imp = RandomForestClassifier(n_jobs=-1, random_state=42,␣
  ↪n_estimators=57, max_depth=9, min_impurity_decrease=1e-7,␣
  ↪class_weight={'yes':2, 'no':1})
ran_for_model_imp.fit(X_train, train_targets)
ran_for_model_imp.score(X_train, train_targets), ran_for_model_imp.score(X_val,␣
  ↪val_targets)
```

[539]:

```
(0.9353932584269663, 0.8435754189944135)
```

5 Final Model

[529]:

```
ran_for_model_final = RandomForestClassifier(n_jobs=-1, random_state=42,␣
 ↪n_estimators=57, max_depth=9, min_impurity_decrease=1e-7,␣
 ↪class_weight={'yes':2, 'no':1})
ran_for_model_final.fit(X_train, train_targets)
```

[529]:

```
RandomForestClassifier(class_weight={'no': 1, 'yes': 2}, max_depth=9,
                       min_impurity_decrease=1e-07, n_estimators=57, n_jobs=-1,
                       random_state=42)
```

[530]:

```
preds=ran_for_model_final.predict(X_test)
```