Lab 3.4 - Student Notebook

Overview

This lab is a continuation of the guided labs in Module 3.

In this lab, you will split the data into three separate datasets:

- *Training Set* - This will be used to train the model.
- *Validation Set* - This will be used during training to validate the model.
- *Test Set* - This will be held back and used to produce metrics after the model is trained. You will use this dataset in an upcoming lab.

With the split data, you will train a XGBoost model by using Amazon SageMaker.

Introduction to the business scenario

You work for a healthcare provider, and want to improve detection of abnormalities in orthopedic patients.

You are tasked with solving this problem by using machine learning (ML). You have access to a dataset that contains six biomechanical features and a target of *normal* or *abnormal*. You can use this dataset to train an ML model to predict if a patient will have an abnormality.

About this dataset

This biomedical dataset was built by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopaedics (GARO) of the Centre Médico-Chirurgical de Réadaptation des Massues, Lyon, France. The data has been organized in two different, but related, classification tasks.

The first task consists in classifying patients as belonging to one of three categories:

- *Normal* (100 patients)
- *Disk Hernia* (60 patients)
- *Spondylolisthesis* (150 patients)

For the second task, the categories *Disk Hernia* and *Spondylolisthesis* were merged into a single category that is labeled as *abnormal*. Thus, the second task consists in classifying patients as belonging to one of two categories: *Normal* (100 patients) or *Abnormal* (210 patients).

Attribute information:

Each patient is represented in the dataset by six biomechanical attributes that are derived from the shape and orientation of the pelvis and lumbar spine (in this order):

- Pelvic incidence
- Pelvic tilt
- Lumbar lordosis angle
- Sacral slope
- Pelvic radius
- Grade of spondylolisthesis

The following convention is used for the class labels:

- DH (Disk Hernia)
- Spondylolisthesis (SL)
- Normal (NO)
- Abnormal (AB)

For more information about this dataset, see the Vertebral Column dataset webpage.

Dataset attributions

This dataset was obtained from: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). Irvine, CA: University of California, School of Information and Computer Science.

Lab setup

Because this solution is split across several labs in the module, you must run the following cells so that you can load the data.

Importing the data

By running the following cells, the data will be imported and ready for use.

Note: The following cells represent the key steps in the previous labs.

In [27]:
```
import warnings, requests, zipfile, io
warnings.simplefilter('ignore')
import pandas as pd
from scipy.io import arff
import boto3
```

In [28]:
```
f_zip = 'http://archive.ics.uci.edu/ml/machine-learning-databases/00212/vertebral_c
r = requests.get(f_zip, stream=True)
Vertebral_zip = zipfile.ZipFile(io.BytesIO(r.content))
Vertebral_zip.extractall()
```

In [29]:
```
data = arff.loadarff('column_2C_weka.arff')
df = pd.DataFrame(data[0])
```

In [30]:
```
class_mapper = {b'Abnormal':1,b'Normal':0}
df['class']=df['class'].replace(class_mapper)
```

Step 1: Exploring the data

You will start with a quick reminder of the data in the dataset.

To get the most out of this lab, carefully read the instructions and code before you run the cells. Take time to experiment!

First, use **shape** to examine the number of rows and columns.

In [31]:
```
df.shape
```

Out[31]:
```
(310, 7)
```

Next, get a list of the columns.

In [32]:
```
df.columns
```

Out[32]:
```
Index(['pelvic_incidence', 'pelvic_tilt', 'lumbar_lordosis_angle',
       'sacral_slope', 'pelvic_radius', 'degree_spondylolisthesis', 'class'],
      dtype='object')
```

You can see the six biomechanical features, and that the target column is named *class*.

Step 2: Preparing the data

For this lab, you must split the data into three datasets.

An internet search will show many different ways to split datasets. Many code samples that you might find will split the dataset into the *target* and the *features*. Then, they will split each of those two datasets into three subsets, which results in a total of six datasets to track.

Moving the target column position

XGBoost requires the training data to be in a single file. The file must have the target value be the first column.

Get the target column and move it to the first position.

In [33]:
```
cols = df.columns.tolist()
cols = cols[-1:] + cols[:-1]
df = df[cols]
```

You should see that the **class** is now the first column.

In [34]:
```
df.columns
```

Out[34]:
```
Index(['class', 'pelvic_incidence', 'pelvic_tilt', 'lumbar_lordosis_angle',
       'sacral_slope', 'pelvic_radius', 'degree_spondylolisthesis'],
      dtype='object')
```

Splitting the data

You will start by splitting the dataset into two datasets. You will use one dataset for training, and you will split the other dataset again for use with validation and testing.

You will use the *train_test_split function* from the *scikit-learn library*, which is a free machine learning library for Python. It has many algorithms and useful functions, such as the one you will use.

- For more information about the function, see the Train_test_split documentation.
- For more information about scikit-learn, see the scikit-learn guide

Because you don't have a lot of data, you want to make sure that the split datasets contain a representative amount of each class. Thus, you will use the *stratify* switch. Finally, you will use a random number so that you can repeat the splits.

In [35]:
```
from sklearn.model_selection import train_test_split
train, test_and_validate = train_test_split(df, test_size=0.2, random_state=42, str
```

Next, split the *test_and_validate* dataset into two equal parts.

In [36]:
```
test, validate = train_test_split(test_and_validate, test_size=0.5, random_state=42
```

Examine the three datasets.

In [37]:
```
print(train.shape)
print(test.shape)
print(validate.shape)
```

```
(248, 7)
(31, 7)
(31, 7)
```

Now, check the distribution of the classes.

In [38]:
```
print(train['class'].value_counts())
print(test['class'].value_counts())
print(validate['class'].value_counts())
```

```
1    168
0     80
Name: class, dtype: int64
1    21
0    10
Name: class, dtype: int64
1    21
0    10
Name: class, dtype: int64
```

Uploading the data to Amazon S3

XGboost will load the data for training from Amazon Simple Storage Service (Amazon S3). Thus, you must write the data to a comma-separated values (CSV) file, and then upload the file to Amazon S3.

Start by setting up some variables to the S3 bucket, then create a function to upload the CSV file to Amazon S3. You can reuse this function.

First, explore the function.

Note the following line:

`dataframe.to_csv(csv_buffer, header=False, index=False)`

This line writes the pandas DataFrame (which was passed into the function) into the IO buffer that's named *csv_buffer*. You use a buffer because you don't need to write the file locally.

To stop the column headers from being written out, use `header=False`. To stop the pandas index from being output, use `index=False`.

To write the csv_buffer to Amazon S3 as an object, use the `put` operation on the `object`, which is a property of the `bucket`.

In [39]:
```
bucket='c83500a1779279l4126500t1w114409743725-labbucket-1qoiv2d4kxx'

prefix='lab3'

train_file='vertebral_train.csv'
test_file='vertebral_test.csv'
validate_file='vertebral_validate.csv'

import os

s3_resource = boto3.Session().resource('s3')
def upload_s3_csv(filename, folder, dataframe):
```

```
    csv_buffer = io.StringIO()
    dataframe.to_csv(csv_buffer, header=False, index=False)
    s3_resource.Bucket(bucket).Object(os.path.join(prefix, folder, filename)).put(B
```

Use the function that you created to upload the three datasets.

In [40]:
```
upload_s3_csv(train_file, 'train', train)
upload_s3_csv(test_file, 'test', test)
upload_s3_csv(validate_file, 'validate', validate)
```

Step 3: Training the model

Now that the data in Amazon S3, you can train a model.

The first step is to get the XGBoost container URI.

In [41]:
```
import boto3
from sagemaker.image_uris import retrieve
container = retrieve('xgboost',boto3.Session().region_name,'1.0-1')
```

Next, you must set some *hyperparameters* for the model. Because this is the first time you are training the model, you can use some values to get started.

In [42]:
```
hyperparams={"num_round":"42",
             "eval_metric": "auc",
             "objective": "binary:logistic"}
```

Use the **estimator** function to set up the model. Here are a few parameters of interest:

- **instance_count** - This defines how many instances will be used for training. You will use *one* instance.
- **instance_type** - This defines the instance type for training. In this case, it's *ml.m4.xlarge*.

In [43]:
```
import sagemaker
s3_output_location="s3://{}/{}/output/".format(bucket,prefix)
xgb_model=sagemaker.estimator.Estimator(container,
                                        sagemaker.get_execution_role(),
                                        instance_count=1,
                                        instance_type='ml.m4.xlarge',
                                        output_path=s3_output_location,
                                         hyperparameters=hyperparams,
                                         sagemaker_session=sagemaker.Session())
```

The estimator needs *channels* to feed data into the model. For training, the *train_channel* and *validate_channel* will be used.

In [44]:
```
train_channel = sagemaker.inputs.TrainingInput(
    "s3://{}/{}/train/".format(bucket,prefix,train_file),
    content_type='text/csv')
```

```
validate_channel = sagemaker.inputs.TrainingInput(
    "s3://{}/{}/validate/".format(bucket,prefix,validate_file),
    content_type='text/csv')

data_channels = {'train': train_channel, 'validation': validate_channel}
```

Running **fit** will train the model.

Note: This process can take up to 5 minutes.

In [ ]:
```
xgb_model.fit(inputs=data_channels, logs=False)
```

```
INFO:sagemaker:Creating training-job with name: sagemaker-xgboost-2023-05-25-10-12
-01-462
2023-05-25 10:12:02 Starting - Starting the training job....
2023-05-25 10:12:28 Starting - Preparing the instances for training......
```

After the training is complete, you are ready to test and evaluate the model. However, you will do testing and validation in later labs.

Congratulations!

You have completed this lab, and you can now end the lab by following the lab guide instructions.

In [4]:
```
import pandas as pd
import seaborn as sns
```

In [5]:
```
df = pd.read_csv("insurance.csv")
```

In [6]:
```
df.head()
```

Out[6]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In [8]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['sex'] = le.fit_transform(df['sex'])
df['smoker'] = le.fit_transform(df['smoker'])
```

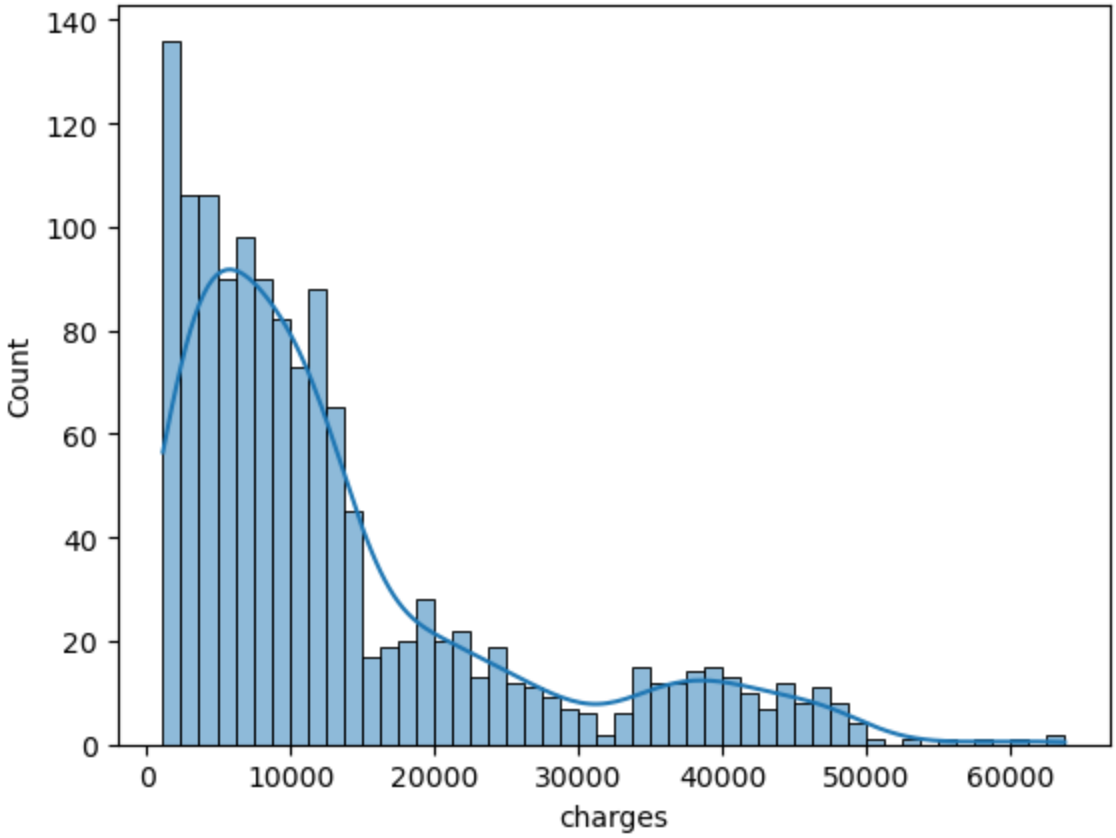
In [9]:
```
df = pd.get_dummies(df, columns = ['region'])
```

In [10]:
```
df.head()
```

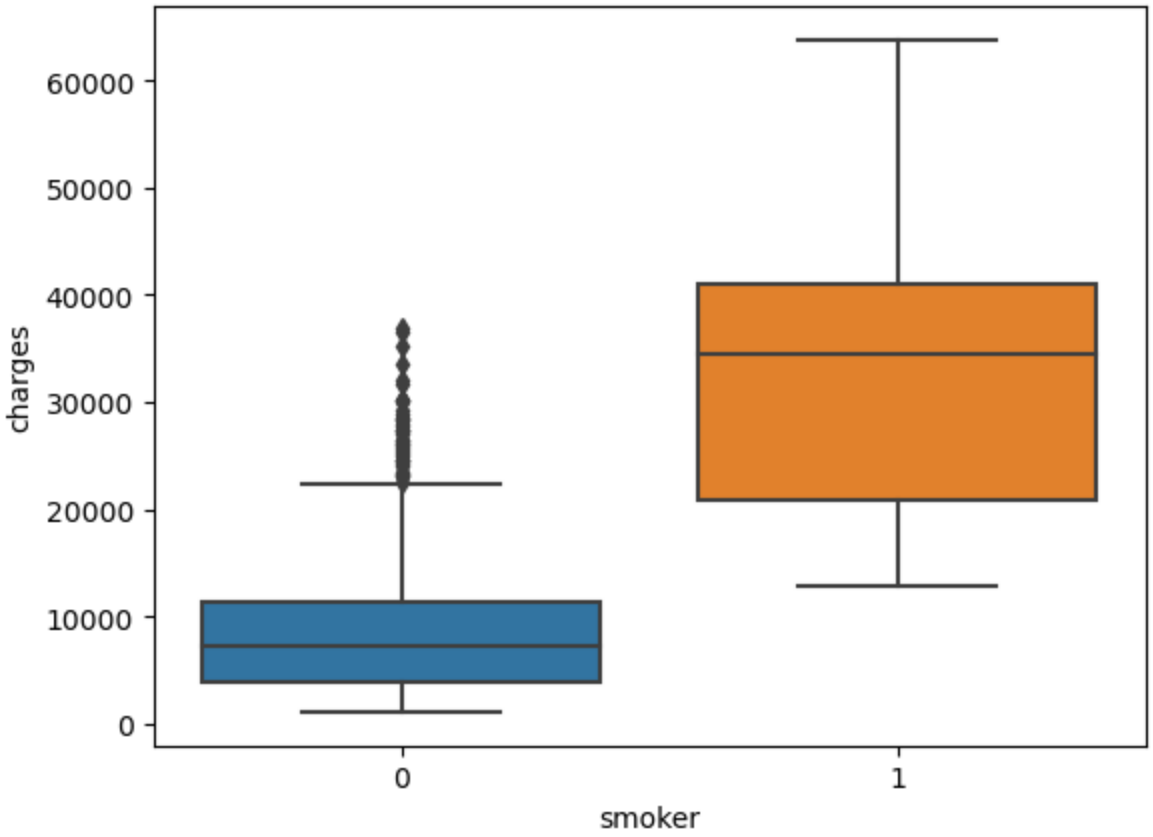
Out[10]:

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_so
0	19	0	27.900	0	1	16884.92400	0	0	
1	18	1	33.770	1	0	1725.55230	0	0	
2	28	1	33.000	3	0	4449.46200	0	0	
3	33	1	22.705	0	0	21984.47061	0	1	
4	32	1	28.880	0	0	3866.85520	0	1	

In [11]:
```
sns.histplot(data=df, x='charges', bins=50 ,kde=True);
```



In [12]:
```
sns.boxplot(data=df, x='smoker', y='charges');
```



In [13]:
```
x = df[['age','sex','bmi','children','smoker','region_northeast','region_northwest'
y = df['charges']
```

In [14]:
```
print(x)
print("-----------------")
print(y)
```

```
      age  sex     bmi  children  smoker  region_northeast  region_northwest  \
0      19    0  27.900         0       1                 0                 0
1      18    1  33.770         1       0                 0                 0
2      28    1  33.000         3       0                 0                 0
3      33    1  22.705         0       0                 0                 1
4      32    1  28.880         0       0                 0                 1
...   ...  ...     ...       ...     ...               ...               ...
1333   50    1  30.970         3       0                 0                 1
1334   18    0  31.920         0       0                 1                 0
1335   18    0  36.850         0       0                 0                 0
1336   21    0  25.800         0       0                 0                 0
1337   61    0  29.070         0       1                 0                 1

      region_southeast  region_southwest
0                    0                 1
1                    1                 0
2                    1                 0
3                    0                 0
4                    0                 0
...                ...               ...
1333                 0                 0
1334                 0                 0
1335                 1                 0
1336                 0                 1
1337                 0                 0

[1338 rows x 9 columns]
-----------------
0       16884.92400
1        1725.55230
2        4449.46200
3       21984.47061
4        3866.85520
           ...
1333    10600.54830
1334     2205.98080
1335     1629.83350
1336     2007.94500
1337    29141.36030
Name: charges, Length: 1338, dtype: float64
```

In [15]:
```
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)
```

LINEAR REGRESSION

In [16]:
```
from sklearn.metrics import r2_score,mean_squared_error
from sklearn.linear_model import LinearRegression
```

```
le = LinearRegression()
le_model = le.fit(x_train,y_train)
le_Y_predict = le_model.predict(x_test)
```

In [17]:
```
print('MSE of Linear Regression : ' + str(mean_squared_error(y_test,le_Y_predict)))
print('r2 score of Linear Regression : ' + str(r2_score(y_test,le_Y_predict)))
```

```
MSE of Linear Regression : 40703200.680649884
r2 score of Linear Regression : 0.7071483213790977
```

DecisionTreeRegressor

In [18]:
```
from sklearn.tree import DecisionTreeRegressor

dtr = DecisionTreeRegressor()
dtr_model = dtr.fit(x_train,y_train)
dtr_y_predict = dtr_model.predict(x_test)
```

In [19]:
```
print('MSE of DecisionTreeRegressor : ' + str(mean_squared_error(y_test,dtr_y_predi
print('r2 score of DecisionTreeRegressor : ' + str(r2_score(y_test,dtr_y_predict)))
```

```
MSE of DecisionTreeRegressor : 53295327.90553596
r2 score of DecisionTreeRegressor : 0.6165503945932336
```

KNN

In [20]:
```
from sklearn.neighbors import KNeighborsRegressor
knn = KNeighborsRegressor()
knn_model = knn.fit(x_train,y_train)
knn_y_predict = knn_model.predict(x_test)
```

In [21]:
```
print('MSE of KNeighborsRegressor : ' + str(mean_squared_error(y_test,knn_y_predict
print('r2 score of KNeighborsRegressor : ' + str(r2_score(y_test,knn_y_predict)))
```

```
MSE of KNeighborsRegressor : 109886658.75254203
r2 score of KNeighborsRegressor : 0.20938668371053537
```

XG BOOSTING

In [23]:
```
!pip install xgboost
from xgboost import XGBRegressor
xgb = XGBRegressor()
xgb_model = xgb.fit(x_train,y_train)
# predicting values for test data
xgb_y_predict = xgb_model.predict(x_test)
```

```
Looking in indexes: https://pypi.org/simple, https://pip.repos.neuron.amazonaws.co
m
Collecting xgboost
  Downloading xgboost-1.7.5-py3-none-manylinux2014_x86_64.whl (200.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 200.3/200.3 MB 3.7 MB/s eta 0:00:000
0:0100:01
Requirement already satisfied: numpy in /home/ec2-user/anaconda3/envs/python3/lib/
python3.10/site-packages (from xgboost) (1.22.3)
Requirement already satisfied: scipy in /home/ec2-user/anaconda3/envs/python3/lib/
python3.10/site-packages (from xgboost) (1.10.0)
Installing collected packages: xgboost
Successfully installed xgboost-1.7.5
```

In [24]:
```
print('MSE of AdaBoostRegressor : ' + str(mean_squared_error(y_test,xgb_y_predict))
print('r2 score of AdaBoostRegressor : ' + str(r2_score(y_test,xgb_y_predict)))
```

```
MSE of AdaBoostRegressor : 30612675.697358616
r2 score of AdaBoostRegressor : 0.7797477025114962
```

RANDOM FOREST

In [25]:
```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor()
rf_model = rf.fit(x_train,y_train)
rf_y_predict = rf_model.predict(x_test)
```

In [26]:
```
print('MSE of RandomForestRegressor : ' + str(mean_squared_error(y_test,rf_y_predic
print('r2 score of RandomForestRegressor : ' + str(r2_score(y_test,rf_y_predict)))
```

```
MSE of RandomForestRegressor : 27827563.541509792
r2 score of RandomForestRegressor : 0.7997860473185067
```

In [ ]: