Activity_Build a Naive Bayes model

November 22, 2023

1 Activity: Build a Naive Bayes model

1.1 Introduction

In this activity, you will build your own Naive Bayes model. Naive Bayes models can be valuable to use any time you are doing work with predictions because they give you a way to account for new information. In today's world, where data is constantly evolving, modeling with Naive Bayes can help you adapt quickly and make more accurate predictions about what could occur.

For this activity, you work for a firm that provides insights for management and coaches in the National Basketball Association (NBA), a professional basketball league in North America. The league is interested in retaining players who can last in the high-pressure environment of professional basketball and help the team be successful over time. In the previous activity, you analyzed a subset of data that contained information about the NBA players and their performance records. You conducted feature engineering to determine which features would most effectively predict a player's career duration. You will now use those insights to build a model that predicts whether a player will have an NBA career lasting five years or more.

The data for this activity consists of performance statistics from each player's rookie year. There are 1,341 observations, and each observation in the data represents a different player in the NBA. Your target variable is a Boolean value that indicates whether a given player will last in the league for five years. Since you previously performed feature engineering on this data, it is now ready for modeling.

1.2 Step 1: Imports

1.2.1 Import packages

Begin with your import statements. Of particular note here are `pandas` and from `sklearn`, `naive_bayes`, `model_selection`, and `metrics`.

[1]:
```
# Import relevant libraries and modules.
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
```

```
from sklearn.metrics import recall_score, precision_score, f1_score,␣
 ↪accuracy_score
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
```

1.2.2 Load the dataset

Recall that in the lab about feature engineering, you outputted features for the NBA player dataset along with the target variable `target_5yrs`. Data was imported as a DataFrame called `extracted_data`. As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

[2]:
```
# RUN THIS CELL TO IMPORT YOUR DATA.
# Load extracted_nba_players_data.csv into a DataFrame called extracted_data.

extracted_data = pd.read_csv('extracted_nba_players_data.csv')
```

1.2.3 Display the data

Review the first 10 rows of data.

[3]:
```
# Display the first 10 rows of data.

### YOUR CODE HERE ###
extracted_data.head(10)
```

[3]:

	fg	3p	ft	reb	ast	stl	blk	tov	target_5yrs	total_points	\
0	34.7	25.0	69.9	4.1	1.9	0.4	0.4	1.3	0	266.4	
1	29.6	23.5	76.5	2.4	3.7	1.1	0.5	1.6	0	252.0	
2	42.2	24.4	67.0	2.2	1.0	0.5	0.3	1.0	0	384.8	
3	42.6	22.6	68.9	1.9	0.8	0.6	0.1	1.0	1	330.6	
4	52.4	0.0	67.4	2.5	0.3	0.3	0.4	0.8	1	216.0	
5	42.3	32.5	73.2	0.8	1.8	0.4	0.0	0.7	0	277.5	
6	43.5	50.0	81.1	2.0	0.6	0.2	0.1	0.7	1	409.2	
7	41.5	30.0	87.5	1.7	0.2	0.2	0.1	0.7	1	273.6	
8	39.2	23.3	71.4	0.8	2.3	0.3	0.0	1.1	0	156.0	
9	38.3	21.4	67.8	1.1	0.3	0.2	0.0	0.7	0	155.4	

	efficiency
0	0.270073
1	0.267658
2	0.339869
3	0.491379
4	0.391304

```
5    0.324561
6    0.605505
7    0.553398
8    0.242424
9    0.435294
```

1.3 Step 2: Model preparation

1.3.1 Isolate your target and predictor variables

Separately define the target variable (`target_5yrs`) and the features.

[4]:
```
# Define the y (target) variable.

### YOUR CODE HERE ###
y = extracted_data['target_5yrs']

# Define the X (predictor) variables.

### YOUR CODE HERE ###
X = extracted_data.copy()
X = X.drop('target_5yrs', axis = 1)
```

Hint 1

Refer to the content about splitting your data into X and y.

Hint 2

In `pandas`, subset your DataFrame by using square brackets `[]` to specify which column(s) to select.

Hint 3

Quickly subset a DataFrame to exclude a particular column by using the `drop()` function and specifying the column to drop.

1.3.2 Display the first 10 rows of your target data

Display the first 10 rows of your target and predictor variables. This will help you get a sense of how the data is structured.

[5]:
```
# Display the first 10 rows of your target data.

### YOUR CODE HERE ###
y.head(10)
```

[5]:
```
0    0
1    0
2    0
```

```
3    1
4    1
5    0
6    1
7    1
8    0
9    0
Name: target_5yrs, dtype: int64
```

Question: What do you observe about the your target variable?

The target variable is a binary variable where its value is 1 when the player has career duration is longer than 5 years, and 0 otherwise.

[6]:
```
# Display the first 10 rows of your predictor variables.

### YOUR CODE HERE ###
X.head(10)
```

[6]:

	fg	3p	ft	reb	ast	stl	blk	tov	total_points	efficiency
0	34.7	25.0	69.9	4.1	1.9	0.4	0.4	1.3	266.4	0.270073
1	29.6	23.5	76.5	2.4	3.7	1.1	0.5	1.6	252.0	0.267658
2	42.2	24.4	67.0	2.2	1.0	0.5	0.3	1.0	384.8	0.339869
3	42.6	22.6	68.9	1.9	0.8	0.6	0.1	1.0	330.6	0.491379
4	52.4	0.0	67.4	2.5	0.3	0.3	0.4	0.8	216.0	0.391304
5	42.3	32.5	73.2	0.8	1.8	0.4	0.0	0.7	277.5	0.324561
6	43.5	50.0	81.1	2.0	0.6	0.2	0.1	0.7	409.2	0.605505
7	41.5	30.0	87.5	1.7	0.2	0.2	0.1	0.7	273.6	0.553398
8	39.2	23.3	71.4	0.8	2.3	0.3	0.0	1.1	156.0	0.242424
9	38.3	21.4	67.8	1.1	0.3	0.2	0.0	0.7	155.4	0.435294

Question: What do you observe about the your predictor variables?

All of the precitro variables seem to be continuous variables.

1.3.3 Perform a split operation on your data

Divide your data into a training set (75% of data) and test set (25% of data). This is an important step in the process, as it allows you to reserve a part of the data that the model has not observed. This tests how well the model generalizes—or performs—on new data.

[7]:
```
# Perform the split operation on your data.
# Assign the outputs as follows: X_train, X_test, y_train, y_test.

### YOUR CODE HERE ###
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .25,␣
↪stratify = y, random_state = 42)
```

Hint 1

Refer to the content about splitting your data between a training and test set.

Hint 2

Call the function in the `model_selection` module of `sklearn` on the features and target variable, in order to perform the splitting.

Hint 3

Call the `model_selection.train_test_split()` function, passing in both `features` and `target`, while configuring the appropriate `test_size`.

Assign the output of this split as `X_train`, `X_test`, `y_train`, `y_test`.

1.3.4 Print the shape of each output

Print the shape of each output from your train-test split. This will verify that the split operated as expected.

[8]:
```
# Print the shape (rows, columns) of the output from the train-test split.

# Print the shape of X_train.

### YOUR CODE HERE ###

print('X_train shape:', X_train.shape)

# Print the shape of X_test.

### YOUR CODE HERE ###
print('X_test shape:', X_test.shape)


# Print the shape of y_train.

### YOUR CODE HERE ###
print('y_train shape:', y_train.shape)


# Print the shape of y_test.

### YOUR CODE HERE ###
print( 'y_test shape:', y_test.shape)
```

```
X_train shape: (1005, 10)
X_test shape: (335, 10)
y_train shape: (1005,)
y_test shape: (335,)
```

Hint 1

Call the attribute that DataFrames in `pandas` have to get the number of rows and number of columns as a tuple.

Hint 2

Call the `shape` attribute.

Question: How many rows are in each of the outputs?

X_train and y_train have 1005 rows each. X_test and y_test have 335 rows each.

Question: What was the effect of the train-test split?

THe train set is larger than the test set, and the split is 75% to 25% ratio.

1.4 Step 3: Model building

Question: Which Naive Bayes algorithm should you use?

We will use GaussianNB since the features are continuous variables. We will assume that these variables are normally distributed.

Hint 1

Refer to the content about different implementations of the Naive Bayes to determine which is appropriate in this situation.

Hint 2

Note that you are performing binary classification.

Hint 3

You can identify the appropriate algorithm to use because you are performing a binary classification and assuming that the features of your model follow a normal distribution.

1.4.1 Fit your model to your training data and predict on your test data

By creating your model, you will be drawing on your feature engineering work by training the classifier on the `X_train` DataFrame. You will use this to predict `target_5yrs` from `y_train`.

Start by defining `nb` to be the relevant algorithm from `sklearn.naive_bayes`. Then fit your model to your training data. Use this fitted model to create predictions for your test data.

[9]:
```
# Assign `nb` to be the appropriate implementation of Naive Bayes.

### YOUR CODE HERE ###
nb = GaussianNB()


# Fit the model on your training data.
```

```
### YOUR CODE HERE ###

nb.fit(X_train, y_train)

# Apply your model to predict on your test data. Call this "y_pred".

### YOUR CODE HERE ###
y_pred = nb.predict(X_test)
```

Hint 1

Refer to the content about constructing a Naive Bayes.

Hint 2

The appropriate implementation in this case is `naive_bayes.GaussianNB()`. Fit this model to your training data and predict on your test data.

Hint 3

Call `fit()`and pass your training feature set and target variable. Then call `predict()` on your test feature set.

1.5 Step 4: Results and evaluation

1.5.1 Leverage metrics to evaluate your model's performance

To evaluate the data yielded from your model, you can leverage a series of metrics and evaluation techniques from scikit-learn by examining the actual observed values in the test set relative to your model's prediction. Specifically, print the accuracy score, precision score, recall score, and f1 score associated with your test data and predicted values.

[10]:
```
# Print your accuracy score.

### YOUR CODE HERE ###
print('Accuracy score: ', '%.3f' % accuracy_score(y_test, y_pred))


# Print your precision score.

### YOUR CODE HERE ###
print('Precision score: ', '%.3f' % precision_score(y_test, y_pred))


# Print your recall score.

### YOUR CODE HERE ###
print('Recall score: ', '%.3f' % recall_score(y_test, y_pred))
```

```
# Print your f1 score.

### YOUR CODE HERE ###
print('F1 score: ', '%.3f' % f1_score(y_test, y_pred))
```

```
Accuracy score:  0.621
Precision score:  0.779
Recall score:  0.543
F1 score:  0.640
```

Hint 1

Refer to the content about model evaluation for detail on these metrics.

Hint 2

The `metrics` module in `sklearn` has a function for computing each of these metrics.

Hint 3

Call `accuracy_score()`, `precision_score()`, `recall_score()`, and `f1_score()`, passing `y_test`, and `y_pred` into each function.

Question: What is the accuracy score for your model, and what does this tell you about the success of the model's performance?

The accuracy score is 0.621, which means that 62.1% of the model's predictions were accurate.

Question: Can you evaluate the success of your model by using the accuracy score exclusively?

Probably not, since the data was mildly imbalanced. Moreover, it is better to use multiple metrics to evaluate model performace.

Question: What are the precision and recall scores for your model, and what do they mean? Is one of these scores more accurate than the other?

The precision score is 0.779, while the recall score is 0.543. This means that 77.9% of predicted positives are true positives. In other words, 77.9% of those who were predicted to have a career longer than 5 years indeed had such a carreer duration. The recall score means that 54.3% of those who actually had a carrer duration longer than 5 years were correctly identified as such.

Question: What is the F1 score of your model, and what does this score mean?

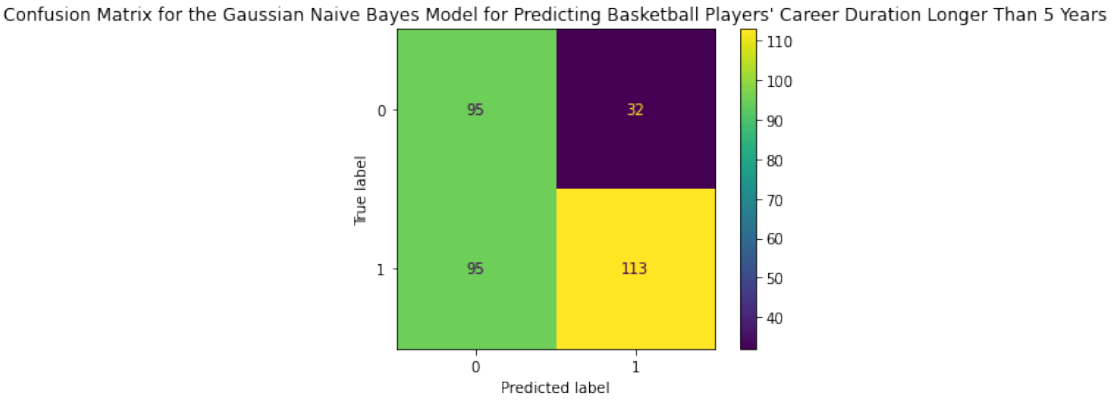
F1 score is 0.64. It is a harmonic mean of precision and recall scores. This metric combines both precision and recall score.

1.5.2 Gain clarity with the confusion matrix

Recall that a confusion matrix is a graphic that shows your model's true and false positives and negatives. It helps to create a visual representation of the components feeding into the metrics.

Create a confusion matrix based on your predicted values for the test set.

[11]:
```
# Construct and display your confusion matrix.

# Construct the confusion matrix for your predicted and test values.

### YOUR CODE HERE ###
cm = confusion_matrix(y_test, y_pred, labels = nb.classes_)


# Create the display for your confusion matrix.

### YOUR CODE HERE ###
disp = ConfusionMatrixDisplay(confusion_matrix = cm, display_labels = nb.
 ↪classes_)
disp.plot(values_format = '')

# Plot the visual in-line.

### YOUR CODE HERE ###
plt.title('''Confusion Matrix for the Gaussian Naive Bayes Model for Predicting␣
 ↪Basketball Players' Career Duration Longer Than 5 Years''')
plt.show()
```



Hint 1

The `metrics` module has functions to create a confusion matrix.

Hint 2

Call `confusion_matrix`, passing in `y_test` and `y_pred`. Then, utilize `ConfusionMatrixDisplay()` to display your confusion matrix.

Question: What do you notice when observing your confusion matrix, and does this correlate to any of your other calculations?

The number of false negatives, 95, is the same as that of true negatives. The number of false postives

is relatively small, which is not surprising given the fact that the precision score is relatively high.

1.6 Considerations

What are some key takeaways that you learned from this lab?

- One can use Naive Bayes (NB) model to predict binary outcomes.
- There are several types of NB model.
- F1 score is a combination of precision and recall scores
-

How would you present your results to your team?

Since all predictors are continuous variable, I used Gaussian NB model. The model has a decent precision score of 77.9%. The accuracy score is lower by being 62.1%. The recall score is even lower by being 54.3%. This is due to the fact that the model has a sizable number of false negatives.

How would you summarize your findings to stakeholders?

This is a decent model for predicting basketball players' career duration. The precision score is 0.779, while the recall score is 0.543. In other words, 77.9% of those who were predicted to have a career longer than 5 years indeed had such a carreer duration. The recall score means that 54.3% of those who actually had a carrer duration longer than 5 years were correctly identified as such. However, the model can be improved to get a better prediction performance.

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged