← Go Back to Elective Project

Course Content

Project FAQ - Potential Customers Prediction

Frequently Asked Questions
Practical Data Science Project

1. Getting errors when I'm trying to predict the training data. How to fix the below error?

```
In [59]: # Fitting the random forest tree classifier on the training data
         rf_estimator = RandomForestClassifier(criterion='entropy', random_state=7)

         rf_estimator.fit(X_train,y_train)

Out[59]: RandomForestClassifier(criterion='entropy', random_state=7)
```

Let's check the performance of the model on the training data

```
In [61]: # Checking performance on the training data
         y_pred_train3 = rf_estimator(X_train)

         metrics_score(y_train, y_pred_train3)

---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_30100/2150125279.py in <module>
      1 # Checking performance on the training data
----> 2 y_pred_train3 = rf_estimator(X_train)
      3
      4 metrics_score(y_train, y_pred_train3)

TypeError: 'RandomForestClassifier' object is not callable
```

The rf_estimator itself is the model object and it is not callable. The model object has several other functions to invoke for different purposes. To fix the error, please invoke the appropriate function using the model object. In the above case, you need to invoke the **predict()** method through the rf_estimator object.

2. While checking the decision tree model performance on the test data, receiving an error "Found input variables with inconsistent numbers of samples: [3228, 1384]". In code, my functions are correct so far but not sure why I am getting this error.

The error you are getting is because you are passing the y_train instead of y_test to the metrics_score() function in the code. This is causing inconsistencies in the shape of the data during the prediction. Please pass the y_test as the parameter in metrics_score() function.

metrics_score(y_train, y_pred_test5) # Change it to y_test

3. In the Project, I noticed that we set drop_first=true when encoding the categorical features. In the case of this project, it results in dropping a seemingly important category value for the profile_completed feature from the data before the model is built. Are we expected to work with this as configured or is the expectation that we decide whether to change that?

The drop_first is a parameter that is used to drop the first level to get k-1 dummies out of k categorical levels. It does not drop any information instead it will create k-1 dummies for k categories.

For example - If we have 2 categories, say yes and no, instead of having values 0 1, 1 0 as 2 columns we can have one single column that represents yes or no values. Please refer to this source to understand the concept well.

4. While during the Decision Tree and Random Forest Project: Predicting Potential Customers, I wondered why it was not required to normalize or standardize the dataset.

The Tree-based algorithms such as Decision Tree, Random forest, and gradient boosting, are not sensitive to the magnitude of variables. So standardization or normalization is not needed before fitting the tree-based models.

note: If you would like to perform scaling on the data, you can do it. But it may vary with our expected results for this problem solution.

5. The data set has around 28% of positive outcomes. Do we need to use class weights? Could you provide some detailed information on how much-unbalanced data needs weights?

If the class imbalance is greater than or equal to 70%, we would go for either class imbalance handling techniques or giving more weight to the minor class among the classes. In this use case, we are interested in class 1 rather than class 0, so we are giving higher weights to class 1.

6. What if we have three categories in the imbalance datasets, how to handle them?

The main idea is based on the problem statement, we have to go for imbalancing handling techniques and class weights. If the three classes are in equal distribution, it is fine. Let's say class 1 has 20 %, class 2 has 40% and class has 40% of data, it is also acceptable. But if any class is less than 20, it means it is highly imbalanced then we need to handle it either by using imbalance techniques or providing higher class weight to the minority classes.

7. When I plot the heatmap with numerical columns it is working fine, and gives an error when the data has categorical columns. How to plot the heatmap for all variables?

We can able to plot the heatmap for only numerical variables because it takes the correlation matrix of the data and visualizes it through the heatmap. In order to include the categorical variables in the heatmap, first, we need to encode them into numerical and compute the correlation matrix that can be further visualized through the heatmap.

8. What is the difference between a single square bracket [] and a double square bracket [[]]?

The single bracket [] is used to access one feature at a time. If you want to access multiple features/columns from the data frame, we need to use double brackets [[ ]].

Previous

Next