← Go Back to Model Tuning

Course Content

FAQ - ReneWind

1. How should one approach the ReneWind project?

Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric

Then you should start with an exploratory analysis of the data.

This understanding will help you identify the need for pre-processing the data.

Once the data is ready, you can start with the steps that need to be followed as mentioned in the rubric

Build 6 models (at least) with original data
Build 6 models (at least) with oversampled data
Build 6 models (at least) with undersampled data
Choose 3 best models (at least) among the models built in the previous 3 steps
Tune the chosen models
Choose one best model and productionize it using pipelines

It is important to close the analysis with key findings and recommendations to the business.

2. How to split the train and test datasets provided?

In this project, two separate datasets have been provided: Train.csv and Test.csv. The Train.csv file can be split into training and validation sets to train, tune, and check how models perform. The Test.csv file should be used for testing the performance of the final model only. So, there is no need to split the Test.csv file.

3. Do we have to use both random search and grid search?

Any one of the search algorithms can be used for hyperparameter tuning. As random search tries random combinations of the hyperparameters, it is computationally less expensive and often executes faster than grid search.

4. I am trying to tune the decision tree with the pipeline and getting this error:

```
ValueError: Invalid parameter classifier for estimator Pipeline(steps=[('standardsc
Check the list of available parameters with `estimator.get_params().keys()`.
```

How to resolve it?

The parameter grid passed for tuning is not defined properly. Please find out the correct names of hyperparameters that need to be passed using **pipe.get_params()**

5. I am getting this error while importing SMOTE even after successful installation of imblearn library:

```
ImportError: cannot import name 'delayed' from 'sklearn.utils.fixes' (C:\Users\anac
```

How to resolve it?

One can follow the following steps:

1. Run **!pip install delayed** in your Jupyter notebook.
2. Restart the kernel and try importing SMOTE again.

6. I am getting this error while trying to tune random forest:

```
NotFittedError: All estimators failed to fit
```

How to resolve it?

The Numpy library might not be updated. You can update the Numpy library to the latest version using

!pip install numpy==1.20.3 in your Jupyter notebook

OR

pip install numpy==1.20.3 in Anaconda prompt

7. Do we oversample/undersample for the validation dataset?

We oversample/undersample the training data only to ensure that the model is trained on a balanced dataset. After building the model, we check its performance on the validation/test data to understand how well the model generalizes.

8. Do we need to tune the hyperparameters of all the models or only a select few?

You need to build 6 (at least) classification models with each of the original, undersampled, and oversampled datasets, giving a total of 18(at least) models. From this pool of initial models, 3 (at least) best models have to be chosen and their hyperparameter tuned to try and improve performance.

For example, if one finds two models as the best models from oversampled data and one from undersampled data by comparing the pool of initial models, then these 3 best models will be selected for hyperparameter tuning.

9. The statistical summary of the dataset has a lot of numerical columns that have negative values. Should we fix them or just leave them?

The negative values are values obtained from sensors placed on the generators and are not anomalous data. So, they need not be treated (imputed/dropped) and can be used as it is.

10. I am trying to understand the features in the data for this project. What does V stand for in V1 to V40?

V has been used to denote different variables in the data. All the variables present in the data are ciphered versions of sensor data related to various environmental factors (temperature, humidity, wind speed, etc.). The data has been ciphered in order to maintain confidentiality.

11. How to give business recommendations as data has been ciphered?

For business recommendations, one can mention the key features used by the model for making predictions and some pointers around how they affect the target.

12. For EDA, if we plot histograms and boxplots for all the variables, do we include all in the presentation or can we show only the important variables identified in feature importance for the best model?

You can add the plots for a few attributes and then mention observations for the rest in the presentation.

13. Do we need to do an EDA for the test dataset?

You only need to do EDA on the training dataset. The test dataset is an unseen dataset and should be used to check the performance of the final model.

14. I am getting this error while trying to check the performance of the model on the test dataset:

```
ValueError: Input contains NaN, infinity or a value too large for dtype('float64').
```

How to resolve it?

The error may have occurred due to missing values present in the test dataset. The missing values present in the test dataset should be treated and then the performance of the model checked on the test dataset.

15. I am getting this error while checking the performance of the model:

```
Value error: Found input variables with inconsistent numbers of the samples
```

How to resolve it?

The error occurred due to the mismatch in the shapes of the dataset. Please make sure that the number of rows present in both datasets passed to the function used to check the model performance should be the same.

16. I am getting this error while importing the imblearn library:

```
Attribute error: module imblearn not found
```

How to resolve it?

The imblearn library has to be installed before importing it. To install imblearn, the following code can be run in a Jupyter notebook.

```
!pip install imbalanced-learn==0.8.0
```

17. How to approach the pipeline section of the project?

Step 1 - Create a pipeline with a simple imputer. This basic structure of the pipeline is as follows:

```
Model = Pipeline([('Name of preprocessing step', preprocessor method()), ('model_na
```

For example:

```
Model = Pipeline([('imputer', SimpleImputer()), ('rf2', 'RandomForestClassifier(all
```

Step 2 - Separate the target variable and other variables into X1(independent variables) and Y1(target) for train data. Repeat the same for the test data.

Step 3 - We can't oversample/undersample data without doing missing value treatment, so first, we have to treat missing values in the train and test sets.

For example,

```
imputer = SimpleImputer(strategy="median")
X1 = imputer.fit_transform(X1)
```

Step 4 - Oversample/undersample the train data and create necessary variables for them (if needed)

Step 5 - Fit the model on the train data

Step 6 - Check the performance of the model on the test data

‹ Previous

Next ›