Renewind Project

Contents

* Business Problem Overview and Solution Approach
* Objective
* Data Overview
* EDA
    Univariate Analysis
* Model Building
* Hyperparameter Tuning
* Business Insights and Recommendations

Business Problem Overview and Solution Approach

Renewable energy sources play an increasingly important role in the global energy mix, as the effort to reduce the environmental impact of energy production increases. Out of all the renewable energy alternatives, wind energy is one of the most developed technologies worldwide. The U.S Department of Energy has put together a guide to achieving operational efficiency using predictive maintenance practices. Predictive maintenance uses sensor information and analysis methods to measure and predict degradation and future component capability. The idea behind predictive maintenance is that failure patterns are predictable and if component failure can be predicted accurately and the component is replaced before it fails, the costs of operation and maintenance will be much lower. The sensors fitted across different machines involved in the process of energy generation collect data related to various environmental factors (temperature, humidity, wind speed, etc.) and additional features related to various parts of the wind turbine (gearbox, tower, blades, break, etc.).

Objective:

The objective is to build various classification models, tune them and find the best one that will help identify failures so that the generator could be repaired before failing/breaking and the overall maintenance cost of the generators can be brought down.
“1” in the target variables should be considered as “failure” and “0” will represent “No failure”.
The nature of predictions made by the classification model will translate as follows:

- True positives (TP) are failures correctly predicted by the model.
- False negatives (FN) are real failures in a wind turbine where there is no detection by model.
- False positives (FP) are detections in a wind turbine where there is no failure.

So, the maintenance cost associated with the model would be:
Maintenance cost = TP*(Repair cost) + FN*(Replacement cost) + FP*(Inspection cost)
Here the objective is to reduce the maintenance cost so, we want a metric that could reduce the maintenance cost.

- The minimum possible maintenance cost = Actual failures*(Repair cost) = (TP + FN)*(Repair cost)
- The maintenance cost associated with model = TP*(Repair cost) + FN*(Replacement cost) + FP*(Inspection cost)

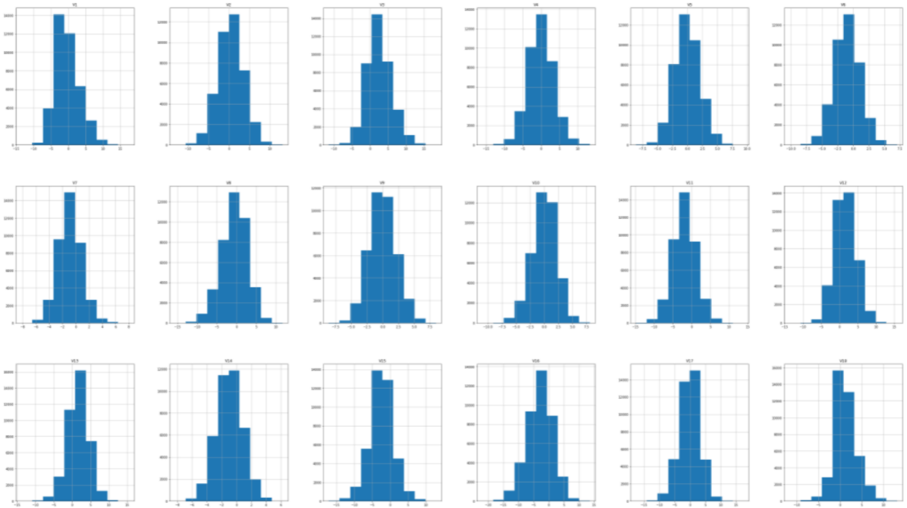
The value of this ratio will lie between 0 and 1, the ratio will be 1 only when the maintenance cost associated with the model will be equal to the minimum possible maintenance cost.

Data Overview

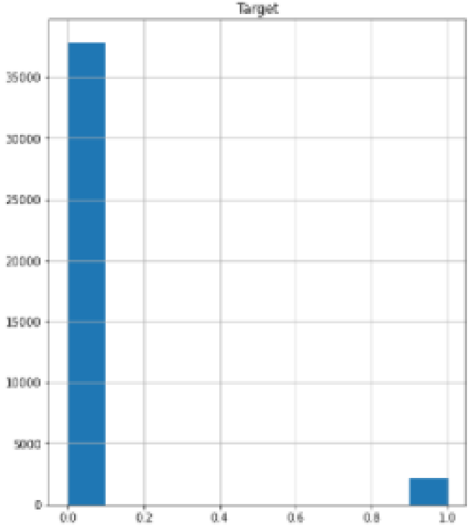
- The data provided is a transformed version of original data which was collected using sensors.
- Train.csv - To be used for training and tuning of models.
- Test.csv - To be used only for testing the performance of the final best model.
- Both the datasets consist of 40 predictor variables and 1 target variable
- There are a total of 41 columns and 40000 observations in the dataset
- We can see that V1 and V2 column have less than 40000 non-null values i.e. column have missing values.
- The values of all the predictor variables are between -23.201 to 24.848.
- We can see that most of the predictor variables are having a Bell shaped/Symmetric distribution.
- Most of the values are concentrated near 0 and in between the ranfe -2 to +2.
- The count of Failure is more than 35000 and count of no Failure is less than 5000.

EDA

- **UNIVARIATE ANALYSIS OF ALL VARIABLES**



Predictor Variables

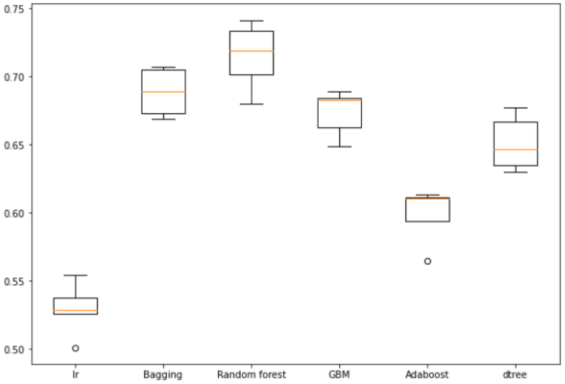


Target

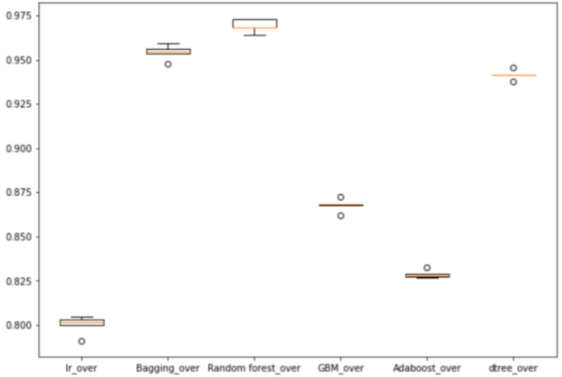
Insights Based on EDA

- The values of all the predictor variables are between -23.201 to 24.848.
- We can see that most of the predictor variables are having a Bell shaped/Symmetric distribution.
- Most of the values are concentrated near 0 and in between the ranfe -2 to +2.
- The count of Failure is more than 35000 and count of no Failure is less than 5000.

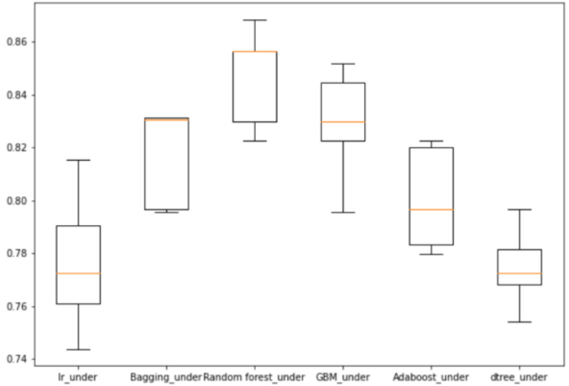
Model Building



Original Data



Oversampling



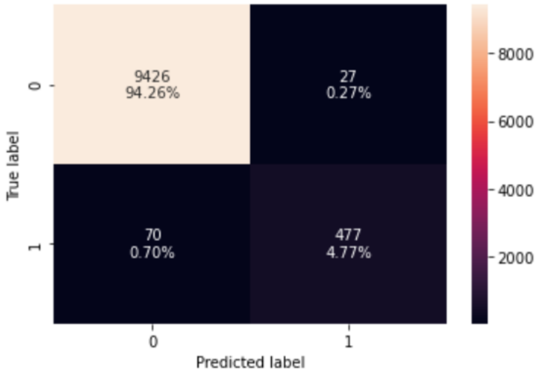
Undersampling

Comapring the Original Data, Oversampled data and Undersampled data, I am selecting the below three models for HyperTuning:

- Random forest_Over
- Bagging_over
- GBM_over

Hyperparameter Tuning using RandomizedSearchCV

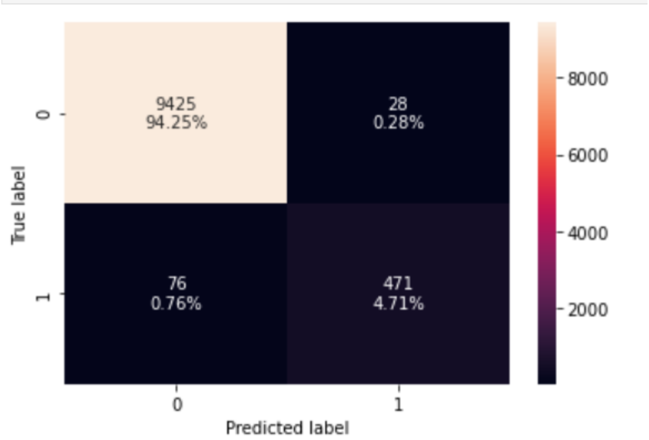
Random Forest is meeting all criteria and it is selected as the final model after Hyperparameter tuning.



- Overfitting has reduced.
- CV score is almost the same ~0.96
- Metrics scores are good

Performance on the test set

- The model provides overall good Metrics scores on Test data
- The model is giving a CV score of 64 on Test Data.
- The important Predictors affecting the Target variable are V18, V36 and V39.



Business Insights and Recommendations

- Company should give more importance to the predictors V18, V36 and V39 as these features greatly affect the Target scores.
- Using the above model the company can reduce their maintenance and Repair Costs.
- It will help identify failures so that the generator could be repaired before failing/breaking and the overall maintenance cost of the generators can be brought down.
- True positives (TP) are failures correctly predicted by the model and it stand at 4.71%
- False negatives (FN) are real failures in a wind turbine where there is no detection by model and its 0.76%
- False positives (FP) are detections in a wind turbine where there is no failure and its 0.28%.

greatlearning

Power Ahead

Happy Learning !