ReneWind - ML model to Reduce Maintenance Cost

Prepared By :
Sathya

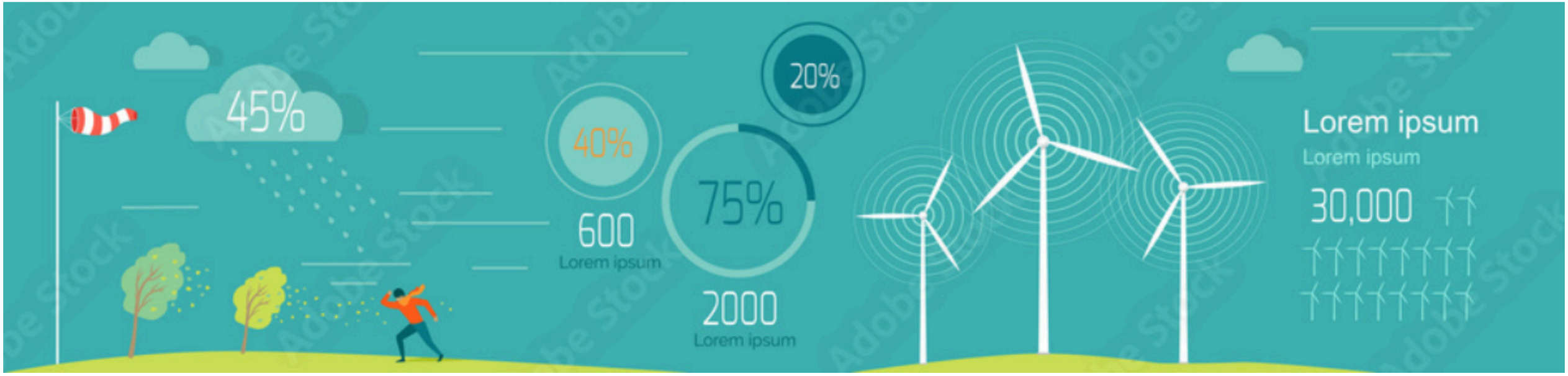
ReneWind - Introduction

ReneWind working on improving the machinery/ processes involved in the production of wind energy using machine learning and has collected data of generator failure of wind turbines using sensors.
By Predicting the generator failure earlier we can able to reduce the maintenance cost.

ReneWind - Data Set

The data set has 40 features ,
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.).

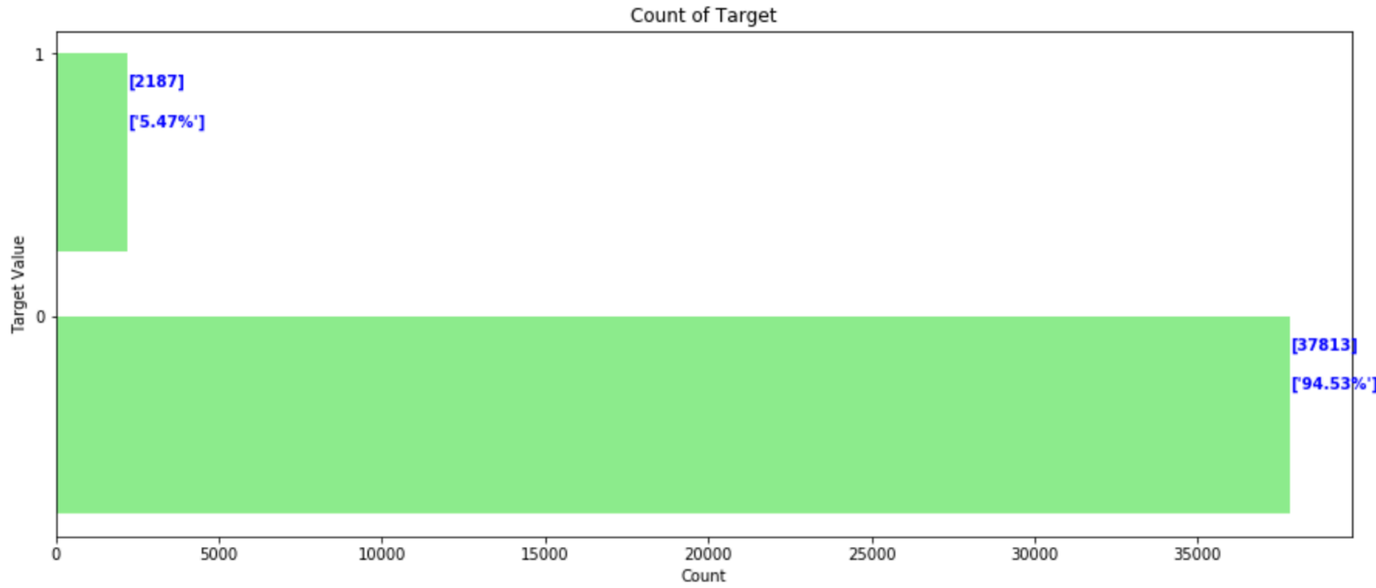
The train Dataset has 40000 records , whereas the test has 10000 records



ReneWind - EDA

The V1, V2 has Null values in the records.

The Data set is imbalanced with the failure record count as 2187 , whereas the Target value 0 record count as 37813

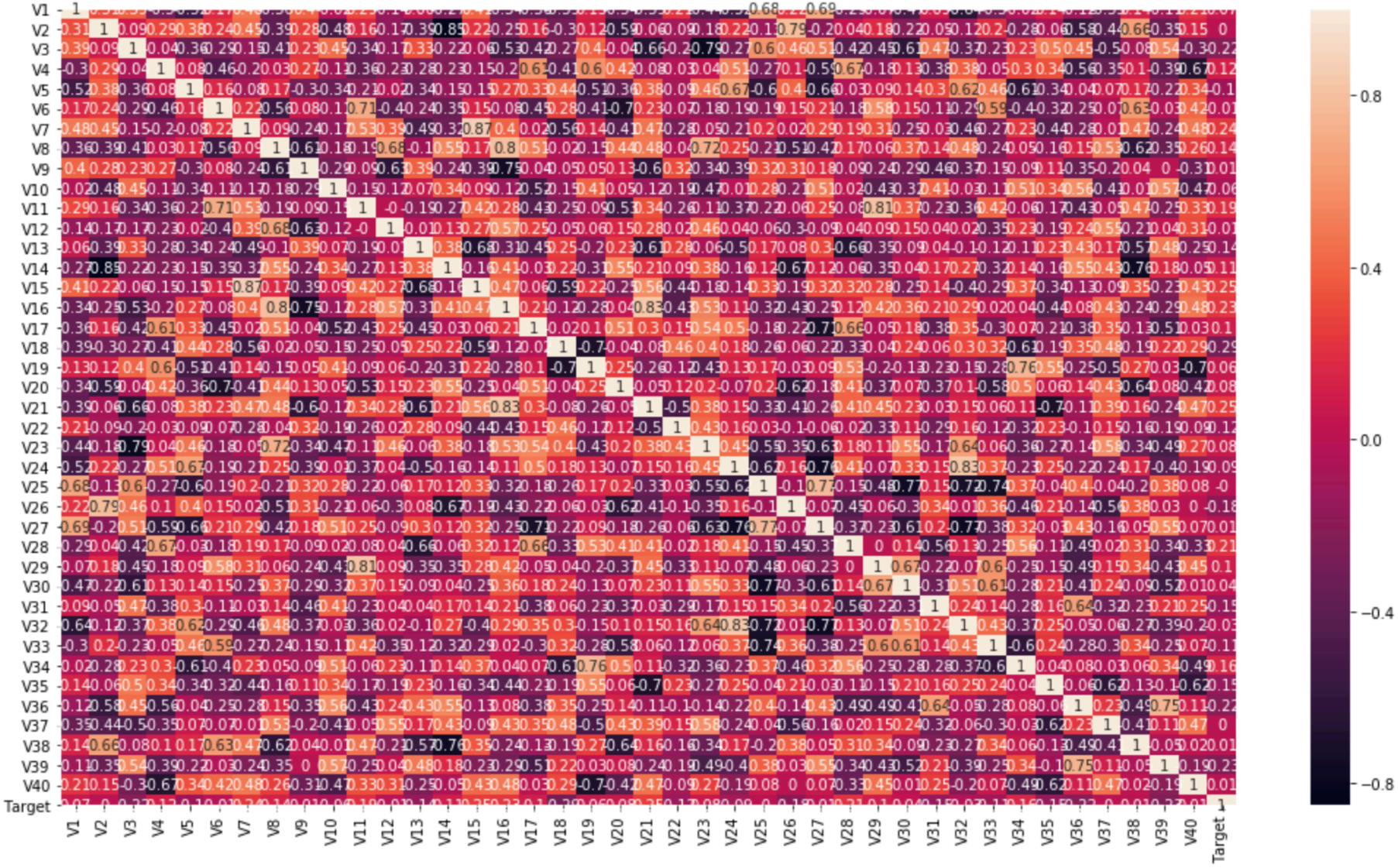


ReneWind - Data Preprocessing

For the Data Preprocessing , the KNNImputer is used to fill the Null values

Based on this V1, V2 values are imputed in the training data set

ReneWind - Heat Map



ReneWind - Model Evaluation Criteria

3 types of cost are associated with the provided problem
1. **Replacement cost** - False Negatives - Predicting no failure, while there will be a failure
2. **Inspection cost** - False Positives - Predicting failure, while there is no failure
3. **Repair cost** - True Positives - Predicting failure correctly

How to reduce the overall cost?
1. We need to create a customized metric, that can help to bring down the overall cost.
The cost associated with any model = TP * 15000 + FP * 5000 + FN * 40000

2. And the minimum possible cost will be when, the model will be able to identify all failures,
in that case, the cost will be (TP + FN) * 15000

So, we will try to maximize `Minimum cost/Cost associated with model`

ReneWind - Model Comparison

	SNO	Model Name	Accuracy	Recall	Precision	F1	Minimum_Vs_Model_cost
0	1	Logistic Regression	0.965250	0.463415	0.823848	0.593171	0.518851
1	2	Decision Tree	0.969500	0.746951	0.710145	0.728083	0.656438
2	3	Random Forest	0.986417	0.760671	0.988119	0.859604	0.713302
3	4	Ada Boost	0.971667	0.605183	0.830544	0.700176	0.588517
4	5	Gradient Boost	0.981417	0.713415	0.930417	0.807593	0.668705
5	6	Logistic Regression with Over Sampling	0.872667	0.841463	0.279352	0.419453	0.503067
6	7	Decision Tree with Over Sampling	0.947000	0.818598	0.509488	0.628070	0.638961
7	8	Random Forest with Over Sampling	0.990167	0.861280	0.954392	0.905449	0.803265
8	9	Ada Boost with Over Sampling	0.971667	0.605183	0.830544	0.700176	0.588517
9	10	Gradient Boost with Over Sampling	0.981333	0.711890	0.930279	0.806563	0.667571
10	11	Logistic Regression with Under Sampling	0.867167	0.836890	0.269646	0.407875	0.493233
11	12	Decision Tree with Under Sampling	0.852750	0.838415	0.248756	0.383676	0.473191
12	13	Random Forest with Under Sampling	0.965250	0.888720	0.628910	0.736576	0.735151
13	14	Ada Boost with Under Sampling	0.896500	0.862805	0.329453	0.476832	0.551261
14	15	Gradient Boost with Under Sampling	0.949167	0.879573	0.520758	0.654195	0.680028

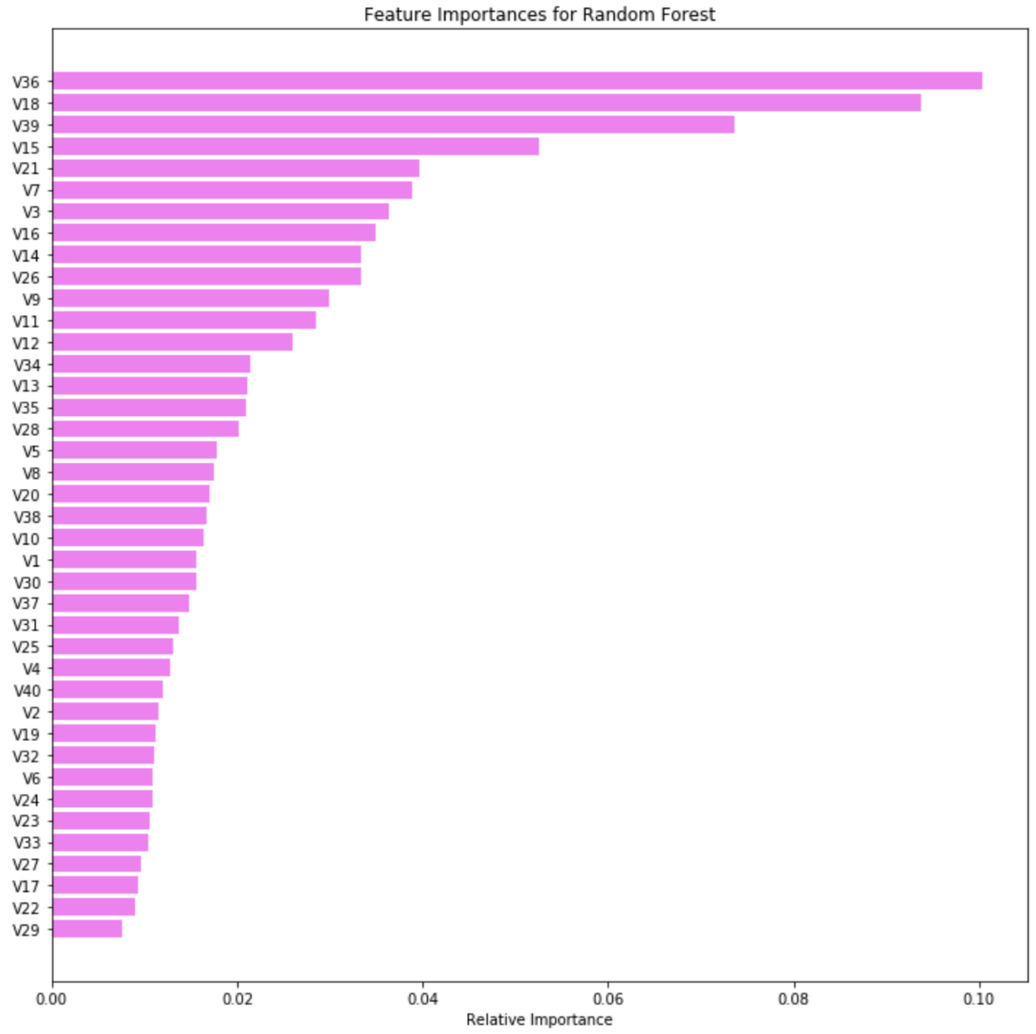
All the Scores are against there validation Data set

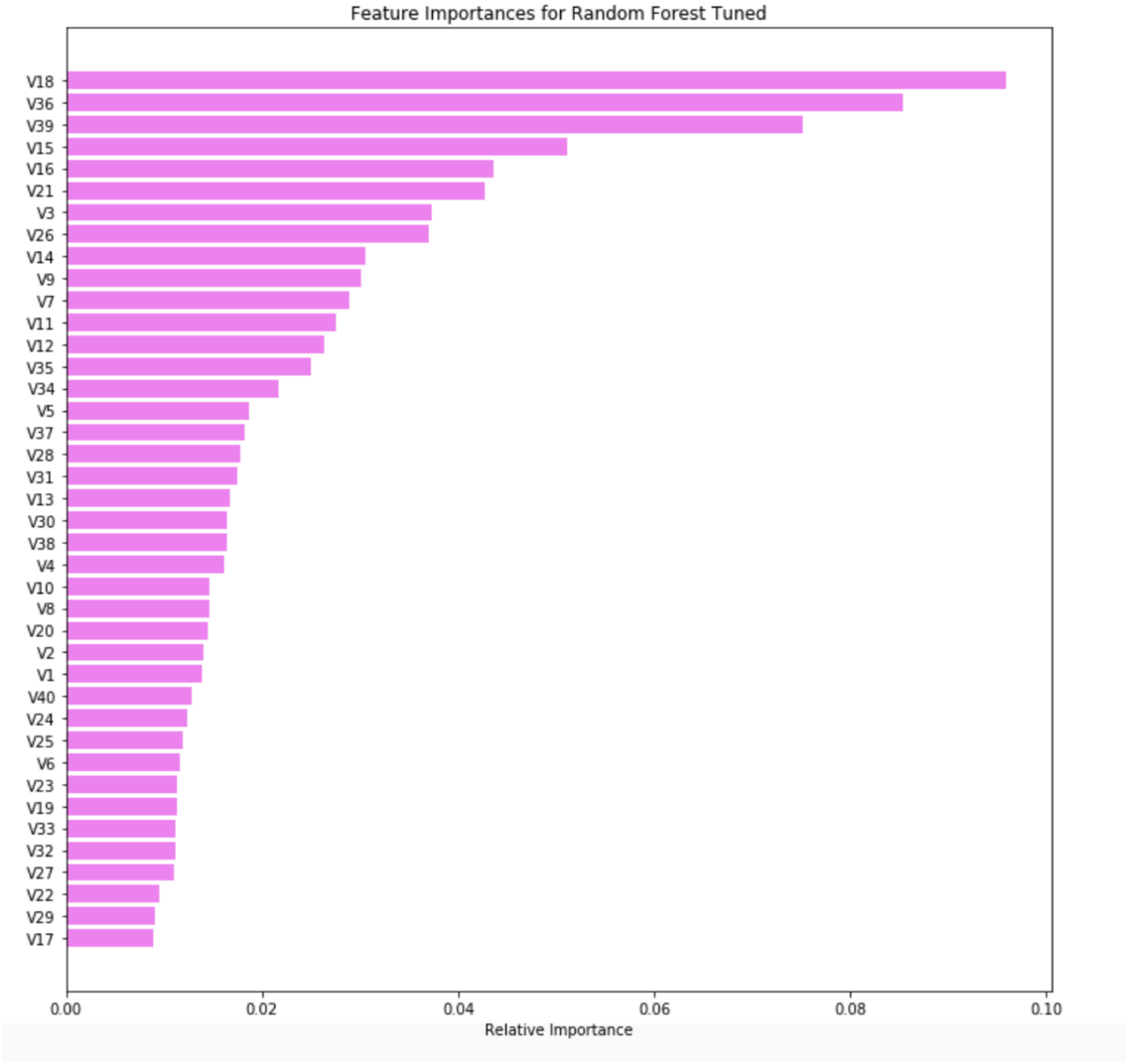
ReneWind - Hyper Parameter Tuning

Random Forest with Over Sampling (SMOTE , gives good F1 Score and lower maintance cost , The hyper parameter tuning is carried on the Random Forest and tested against the test data.

	SNO	Model Name	Accuracy	Recall	Precision	F1	Minimum_Vs_Model_cost
0	8	Random Forest with Over Sampling	0.9899	0.850091	0.960744	0.902037	0.792754
1	16	Random Forest with Over Sampling Tuned Param	0.9893	0.846435	0.952675	0.896418	0.787428

ReneWind - Feature Importance





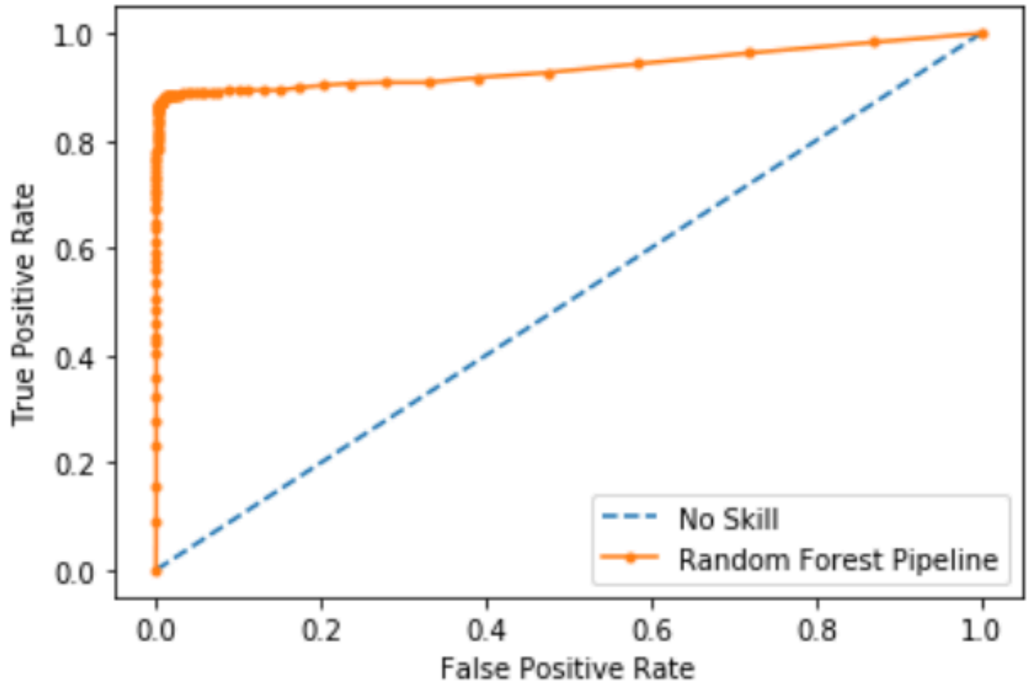
ReneWind - Top 5 feature Analysis

The V36,V18,V39 are negatively corelated with the target value whereas th V15,V21 are positively corelated with the target value.

ReneWind - Pipeline / ROC Curve for Test Data

The Pipeline will impute the data for the Null values based on the KNN imputes and Random Forest Model is used to predict the classification

```
No Skill: ROC AUC=0.500
Random Forest Pipeline : ROC AUC=0.936
```



ReneWind - Business Insights and Recommendations

1) Based on the feature importance , V36,V18,V39 are among top 5 important features and the decrease in the values of the feature leads to the failure of the machines/components for the wind energy production

2) Based on the feature importance , V15,V21 are among top 5 important features and the increase in the values of the feature leads to the failure of the machines/components for the wind energy production

Thank You!