Business Presentation ReneWind

Contents

- Problem Definition
- Key findings and Insights
- Solution
- Potential benefits of implementing solution

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. However, 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.

”ReneWind” is a company 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. To ensure the effectiveness of their business process their objective is to build various classification models, tune them, and find the best one that will help identify failures so that the generators could be repaired before failing/breaking to reduce the overall maintenance cost.

Business Problem Overview and Solution Approach

We will be majorly focusing on ensuring identifying:

- True positives (TP) are failures correctly predicted by the model. These will result in repairing costs.
- False negatives (FN) are real failures where there is no detection by the model. These will result in replacement costs.
- False positives (FP) are detections where there is no failure. These will result in inspection costs.

status

Given that the cost of repairing a generator is much less than the cost of replacing it, and the cost of inspection is less than the cost of repair.

Cost of inspecting generator < Cost of repairing generator<Cost of replacing generator

Data Overview

Training Data

Observations	Variables
20000	41

Test Data

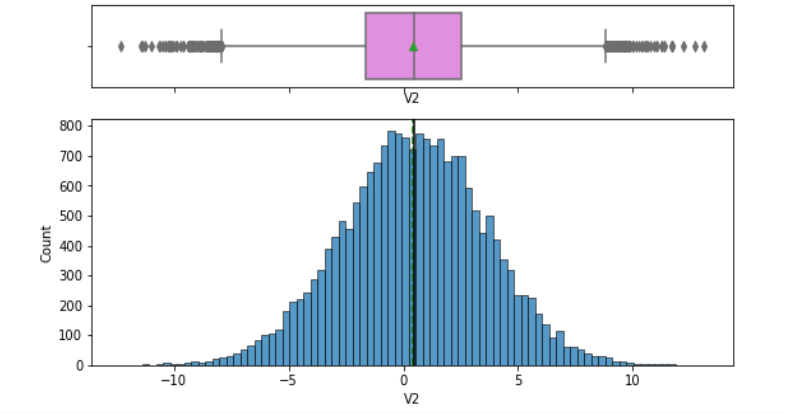
Observations	Variables
5000	41

Note:

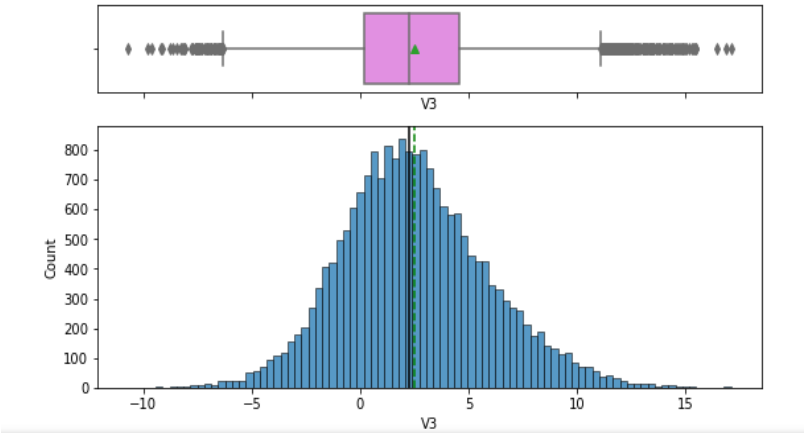
- There are no missing values in the dataset.
- There are no duplicate entries in the dataset
- Number of employees had 33 negative values which have been converted to absolute values
- There are 6 continents represented in the data
- There are 4 education component names Bachelor's, Master's, High School and Doctorate
- Case id column was dropped from the data

Univarate Analysis: Histogram and Box plots

- V2 has a normal distribution with outliers

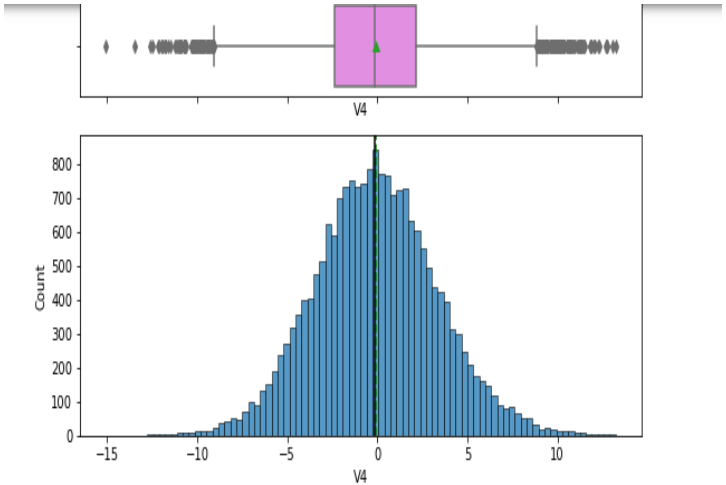


- V3 has a normal distribution slightly right skewed with outliers

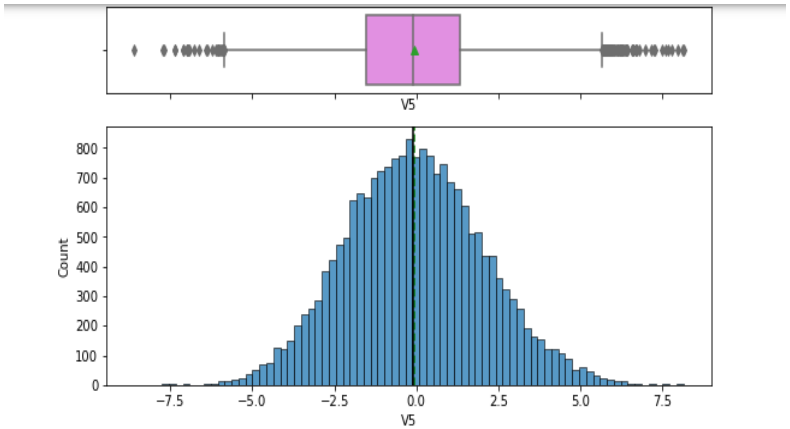


Univarate Analysis: Data collected through Sensors

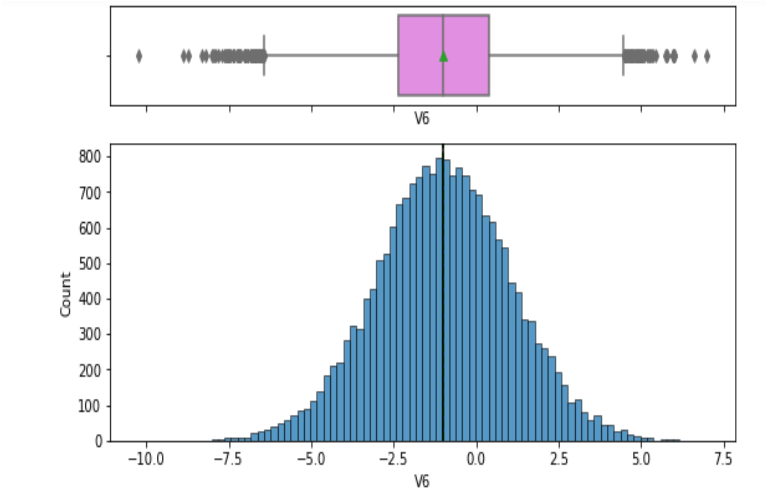
- V4 is normally distributed with outliers on the left and right



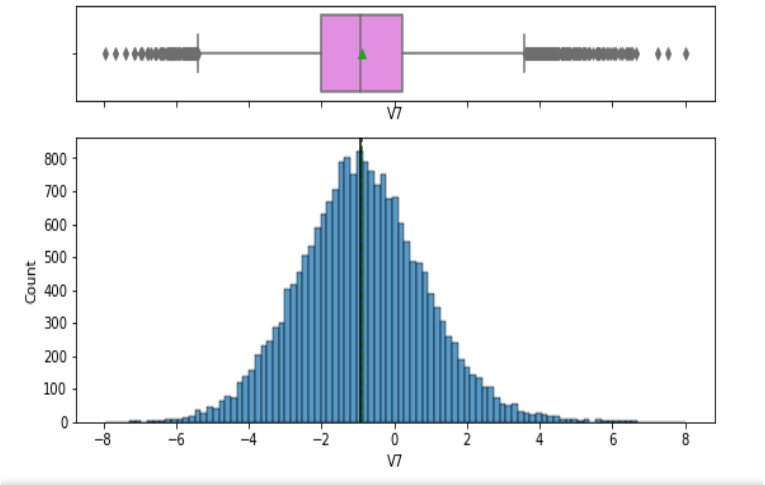
- V5 is normally distributed with more outliers on the right tail



- V6 is normally distributed with outliers on both the right and left tails



- V7 is normally distributed with outliers on both the right and left tails



Values in Target variables

- Training Data

O 0.944
1 0.056
Name: Target, dtype: float 64

- Test Data

0 0.944
1 0.056
Name: Target, dtype: float 64

Model Building on Original Data

- Cross validation performance

```
Cross-Validation Cost:

Logistic regression: 0.4823354838709678
Bagging: 0.6810193548387096
Decision Tree: 0.7034967741935484
Random Forest: 0.7083483870967742
Ada Boost: 0.6168516129032259
Gradient Boosting: 0.7035741935483871
```

- Based on the cross validation cost Random Forest has the highest figure followed by Gradient Boosting

- Validation performance

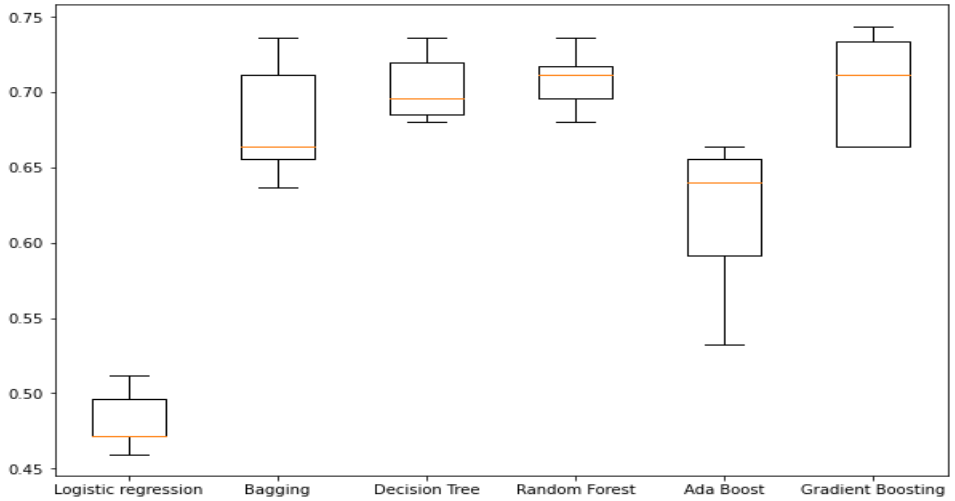
```
Validation Performance:

Logistic regression: 0.5432692307692307
Bagging: 0.7067307692307693
Decision Tree: 0.7211538461538461
Random Forest: 0.75
Ada Boost: 0.6153846153846154
Gradient Boosting: 0.7451923076923077
```

- Based on the validation performance Random Forest has the highest figure followed by Gradient Boosting

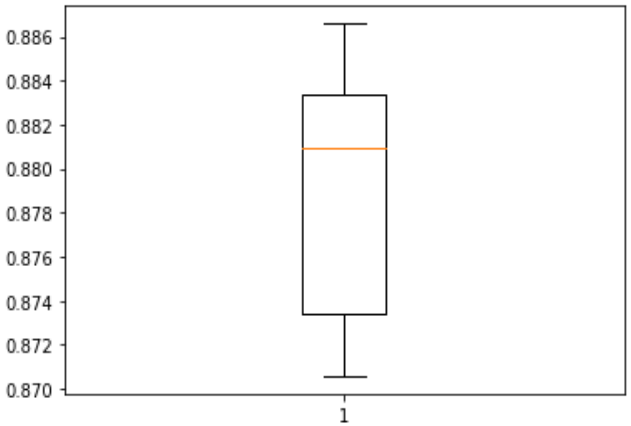
Boxplot for CV scores of all models defined

- We can see that the random forest is giving the highest cross-validated recall followed by gradient boosting
- The boxplot shows that the performance of gradient boosting and random forest is consistent and their performance on the validation set is also good
- We will tune the best two models i.e. gradient boosting and random forest and see if the performance improves




```
<Figure size 432x288 with 0 Axes>
```

Oversampling train data using SMOTE (Synthetic Minority Over Sampling Technique



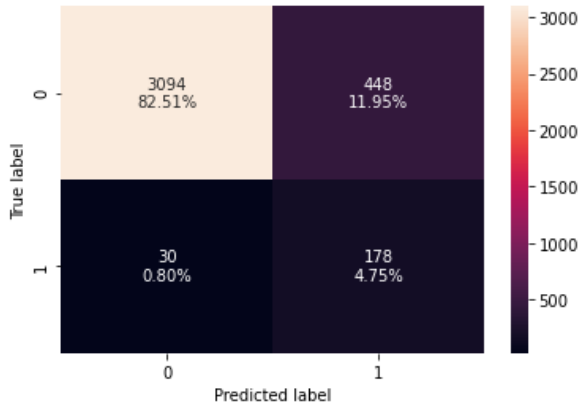
- Performance on training set varies between 0.880 to 0.89 recall

Data Pre-processing: Over-Sampled data Log reg

- Training

Training performance:

	Accuracy	Recall	Precision	F1
0	0.876	0.880	0.874	0.877



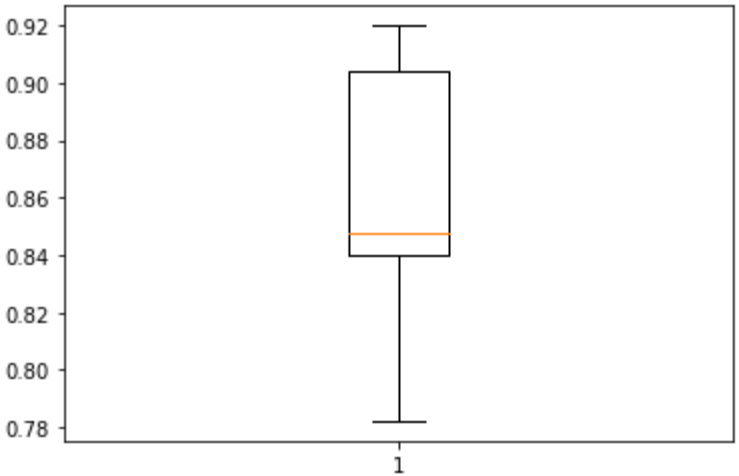
- Validation

validation performance:

	Accuracy	Recall	Precision	F1
0	0.873	0.856	0.284	0.427

- Performance on the training set improved and is able to predict the same on reccal in validation set
- However, there is a significant drop in the precision figure in the validation set thus negating out ability to correctly classify real failures against total failures which could in turn impact replacement cost

Undersampling train data using Random Under Sampler



- Performance of model on training set varies between 0.845 and 0.93 which is an improvement on the initial model

Under sampling train data using Random Under Sampler

- Training

Training performance:

:

	Accuracy	Recall	Precision	F1
0	0.862	0.870	0.856	0.863

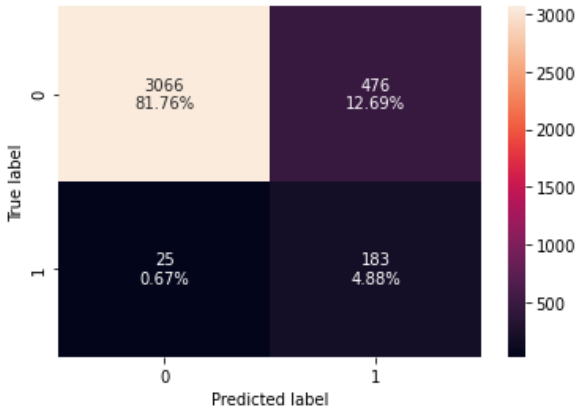
- Model has slight improvement in the figures for Accuracy and Recall
- Model performance has improved using under sampling- Logistic regression is now able to differentiate well between positive and negative classes

- Validation

Validation performance:

:

	Accuracy	Recall	Precision	F1
0	0.866	0.880	0.278	0.422



Model Building

The objective is to build various classification models, tune them, and find the best one that will help identify failures so that the generators could be repaired before failing/breaking to reduce the overall maintenance cost

- **Model evaluation criterion**
  - 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 generator where there is no detection by model.
  - False positives (FP) are failure detections in a generator where there is no failure.

- **Which metric to optimize?**
  - We need to choose the metric which will ensure that the maximum number of generator failures are predicted correctly by the model.
  - We would want Recall to be maximized as greater the Recall, the higher the chances of minimizing false negatives.
  - We want to minimize false negatives because if a model predicts that a machine will have no failure when there will be a failure, it will increase the maintenance cost.

Data Preparation for Modeling

- Training set (11250,40)
- Validation set (3750, 40)
- Test set (5000,40)

- The data is split 70/30 ratio by train and test data
- Training set has 17836 observatins while test set has 7644 observations
- The data type are float 64
- The y variable is case_status

Hyperparameter Tuning: AdaBoost using Oversampled data

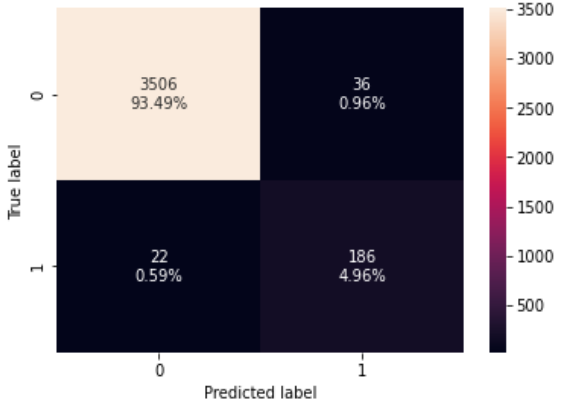
- Training

	Accuracy	Recall	Precision	F1
0	0.996	0.994	0.997	0.996

- Validation

	Accuracy	Recall	Precision	F1
0	0.985	0.894	0.842	0.867

- There is a slight decrease in recall figure from training data to validation data
- The validation recall is still more than 80%i.e the model is good at identifying potential failures correctly thus reducing replacement cost



Hyperparameter Tuning: Random Forest using Under sampled data

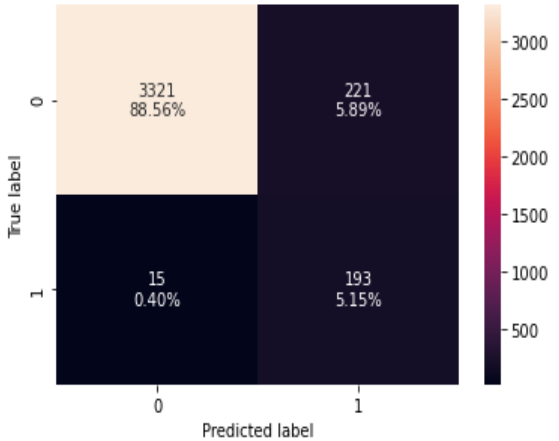
- Training

	Accuracy	Recall	Precision	F1
0	0.990	0.981	1.000	0.990

- Validation

	Accuracy	Recall	Precision	F1
0	0.937	0.928	0.466	0.621

- There has been a significant decrease in the precision and f1 score in the validation model
- Attributing to inability of the validation model to correctly classify generator failures to the total failures predicted
- With the F1 score there is a decrease in equilibrium between the precision and recall
- However the validation model is able to decrease the false negatives thus reducing replacement costs



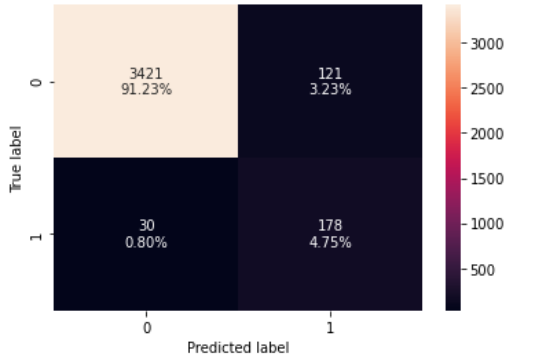
HyperparameterTuning: Gradient Boosting using Oversampled data

- Training

	Accuracy	Recall	Precision	F1
0	0.991	0.991	0.991	0.991

- Validation

	Accuracy	Recall	Precision	F1
0	0.937	0.928	0.466	0.621



- Slight decrease in the recall and accuracy figures
- Precision and f1 has a significant decrease in the validation data
- However, looking at the confusion matrix the FN data is low, showing that there would be a reduction in replacement cost and inspection costs, repair costs are correctly predicted

Hyperparameter Tuning: XGB using oversampled data

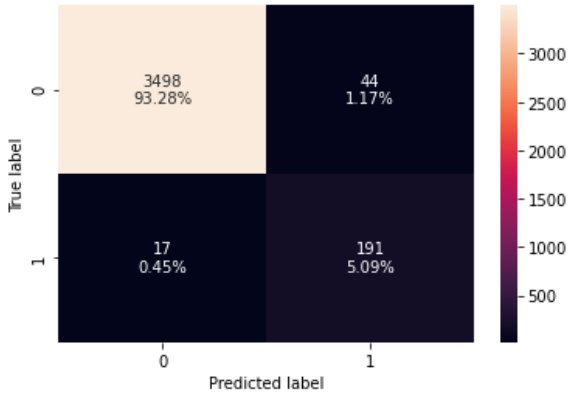
- Training

	Accuracy	Recall	Precision	F1
0	0.999	1.000	0.998	0.999

- There appears to be overfitting on the training data
- There is a very slight reduction in the validation data
- Overall the validation data has a high recall

- Validation

	Accuracy	Recall	Precision	F1
0	0.984	0.918	0.813	0.862



Model performance comparison – Training performance

	Gradient Boosting tuned with oversampled data	AdaBoost classifier tuned with oversampled data	Random forest tuned with undersampled data	XGBoost tuned with oversampled data
Accuracy	0.991	0.996	0.990	0.999
Recall	0.991	0.994	0.981	1.000
Precision	0.991	0.997	1.000	0.998
F1	0.991	0.996	0.990	0.999

- XG Boost has the highest figure on oversampled data for Recall and Precision

Model performance comparison- Validation

	Gradient Boosting tuned with oversampled data	AdaBoost classifier tuned with oversampled data	Random forest tuned with undersampled data	XGBoost tuned with oversampled data
Accuracy	0.960	0.985	0.937	0.984
Recall	0.856	0.894	0.928	0.918
Precision	0.595	0.842	0.466	0.813
F1	0.702	0.867	0.621	0.862

- With validation data it is Random Forest that has the highest figure for recall, though it is just less than 10% higher than the figure for XG Boost
- With respect to Accuracy and Precision the Ada Boost classifier has the highest figure

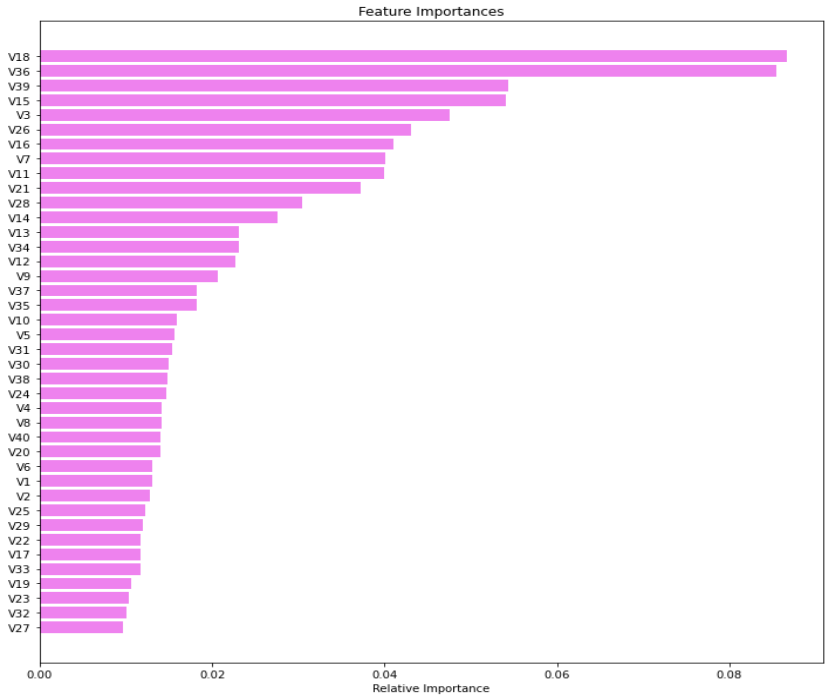
Final model performance on test data

	Accuracy	Recall	Precision	F1
0	0.927	0.888	0.427	0.576

- The performance on test data is generalized
- However, recall and accuracy score are still higher than 50%

Feature Importance

- The most important feature is v18 followed very closely by v36
- The lease most important feature is v27 followed very closely by v32



Final Model using Pipelines

	Accuracy	Recall	Precision	F1
0	0.983	0.705	0.985	0.822

- Recall score is still significant at over 50%
- Accuracy at over 98.3% is good as it highlights our ability to correctly predict accurate generator failures
- Precision at 98.5% us also good as it demonstrates the ability of the model to accurately predict when inspections are need or repairs are needed

Actionable Insights and Recommendations

- Key Takeaways
  - The Random Forest model has the best performance in terms of Recall followed very closely by XG Boost model
  - The most important feature is v18 followed very closely by v36 which are significantly more important than the third and fourth most important features v39 and v15
  - It is important using Recall to ensure that replacement costs are kept at their lowest, thus focusing more on inspection costs and repair costs

- Business Recommendation
  - The most important feature is v18 followed very closely by v36 close attention must be paid to them
  - In addition the features v27 followed very closely by v32 must also be closely monitored
  - To keep close tabs on the False Negatives as replacement cost being the highest cost would have a significant impact not only on profitability but would increase down time thus impacting generation
  - Implement service agreements with key providers

Potential Benefits of Implementing Business Recommendation

- Replacement Cost: Minimize false negatives and replacement cost which if they increase would impact both the profit statement and image of the company
- Feature importance: Close attention should be paid to v18 and v13 as it can have an impact on detection and real failures of generators
- Agreements: Standing agreements for turn around maintenance on generators as this would include inspection costs and also repair costs thus minimizing replacement costs

greatlearning
Power Ahead

Happy Learning !