# RENEWIND – PREDICTING WIND TURBINE GENERATOR FAILURES

## MACHINE LEARNING PIPELINES/HYPERPARAMETER TUNING

ARPAN DINESH

07/12/2023

# CONTENTS

- Executive Summary
- Business Problem
- Solution Approach
- Data Overview
- EDA
- Model Building
- Hyper-Tuning
- Final Model Selection
- Pipelines
- Business Recommendations

# EXECUTIVE SUMMARY

In response to the increasing significance of renewable energy sources, ReneWind has been working diligently to enhance wind energy production efficiency through the implementation of machine learning techniques. The primary objective is to devise predictive maintenance practices using sensor data to identify potential generator failures in advance and minimize operational and maintenance costs.

To achieve this, six classification models are developed using various data sampling techniques, yielding an impressive minimum true generator failure prediction rate of 85%. Subsequently, four models are selected based on recall performance on the validation set, with the Random Forest model on undersampled data emerging as the top-performing solution, predicting 87% of actual generator failures on unseen sensor data.

The analysis also highlighted crucial features that facilitated accurate predictions, laying the groundwork for a robust production pipeline model for predicting true machine failures and reducing maintenance costs.

# BUSINESS PROBLEM

As renewable energy sources gain prominence in the global energy mix, wind energy stands out as a highly developed technology. However, a pressing business problem arises concerning the maintenance of wind turbines: maintenance costs due to failures of wind generator turbines.

To optimize operational efficiency and reduce costs, there is a need for predictive maintenance practices. The challenge lies in effectively utilizing sensor data to predict component degradation and failure accurately.

By implementing predictive maintenance techniques, the aim is to minimize operational and maintenance expenses associated with wind turbine failures, ensuring sustainable energy production.

RENEWIND
We provide revolutionary solutions
that reduces the cost up to 35%!

# SOLUTION APPROACH

1. Exploratory Data Analysis: Data Overview, Univariate Analysis.

2. Data Pre-Processing: Missing/Duplicate Value Check & Treatment, Train/Test Split.

3. Model Building: Evaluation Criterion, Original/Oversampled/Undersampled Data Model.

4. Hyperparameter Tuning: Random Search CV.

5. Performance Evaluation: Model Comparison, Final Model Selection.

6. Pipelines: Productionized Models.

7. Business Recommendations: Actionable Insights, Profitable Suggestions.

# DATA OVERVIEW

The data is a transformed version of original data retrieved from sensors. There are two sets of data:

1) Train Set: Will be used to train and tune the models. Contains 20,000 records and 41 features.

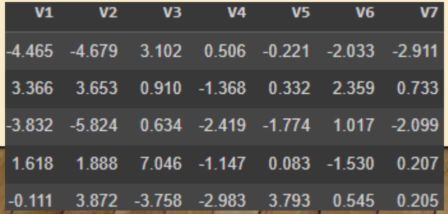
2) Test Set: Will only be used to testing the performance of the final model. Contains 5,000 records and 41 features.

- There are no duplicated records, but features V1 & V2 have 18 missing values.

- Target Variables: ‘1’: Failure; ‘0’: No Failure.

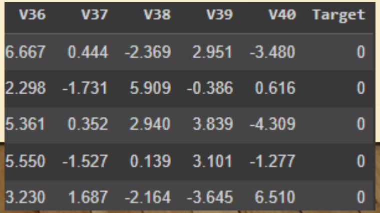
- Proportion of Classes in Train & Test Set: **Train**: 94.5% No Failures & 5.5% Failures. **Test**: 94.4% No Failures & 5.6% Failures.

***Top 5 Rows of Train Set***



| V1 | V2 | V3 | V4 | V5 | V6 | V7 |
|---|---|---|---|---|---|---|
| -4.465 | -4.679 | 3.102 | 0.506 | -0.221 | -2.033 | -2.911 |
| 3.366 | 3.653 | 0.910 | -1.368 | 0.332 | 2.359 | 0.733 |
| -3.832 | -5.824 | 0.634 | -2.419 | -1.774 | 1.017 | -2.099 |
| 1.618 | 1.888 | 7.046 | -1.147 | 0.083 | -1.530 | 0.207 |
| -0.111 | 3.872 | -3.758 | -2.983 | 3.793 | 0.545 | 0.205 |

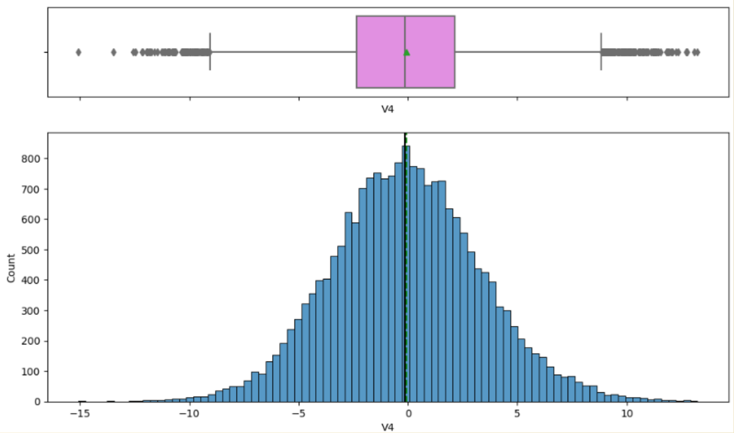
……

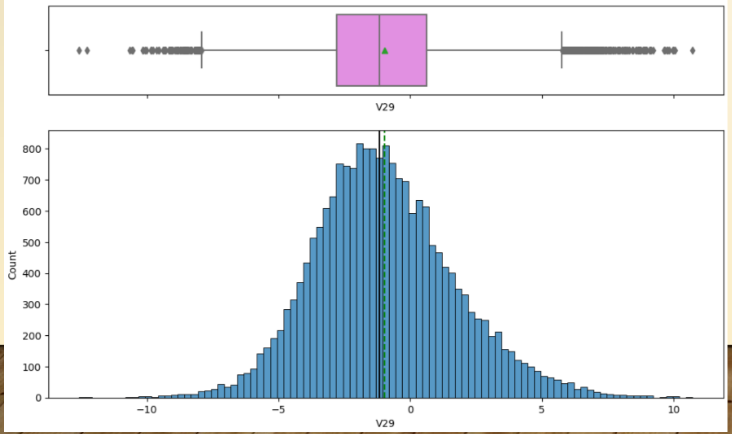


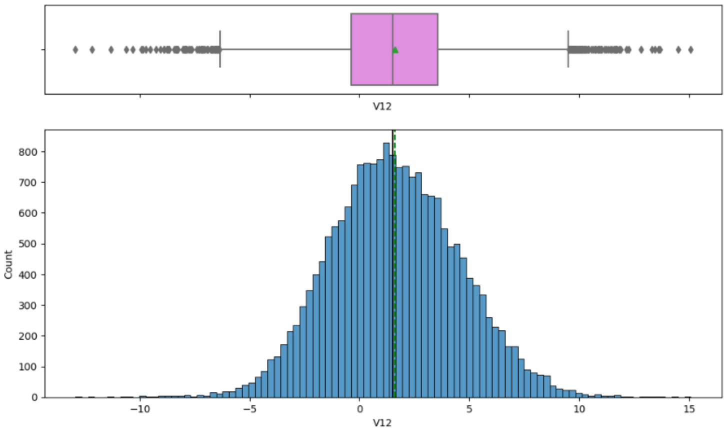
| V36 | V37 | V38 | V39 | V40 | Target |
|---|---|---|---|---|---|
| 6.667 | 0.444 | -2.369 | 2.951 | -3.480 | 0 |
| 2.298 | -1.731 | 5.909 | -0.386 | 0.616 | 0 |
| 5.361 | 0.352 | 2.940 | 3.839 | -4.309 | 0 |
| 5.550 | -1.527 | 0.139 | 3.101 | -1.277 | 0 |
| 3.230 | 1.687 | -2.164 | -3.645 | 6.510 | 0 |

# EDA - UNIVARIATE(VN)

- Histograms & Boxplots are plotted to visualize the distributions of the ‘Vn’ features.
- Most of the features have a normal distribution, while some a slight left/right skew.
- Only a sample of the 41 features(V4,V12,V29) are shown.







# DATA PRE-PROCESSING

- The train set is split into a train and validation set with a 75:25 ratio.

- The train set now has 15,000 records and the validation set has 5,000 records.

- The test set also has 5,000 records.

- The missing values in the 3 sets are imputed using the ‘SimpleImputer’ function in Scikit-Learn: The median of the existing records is imputed into the missing values while ensuring no data leakage.

| Set | Records |
|---|---|
| Train | 15,000 |
| Validation | 5,000 |
| Test | 5,000 |

# MODEL EVALUATION CRITERION

Nature of Predictions made by the model:

1. Failures correctly predicted by the model, resulting in repair costs for ReneWind. (True Positives)

2. Failures not captured by the model, resulting in replacement costs for ReneWind. (False Negatives)

3. Predicted failures when there is no failure, resulting in inspection costs for ReneWind. (False Positives)

Given that inspection costs < repair costs < replacement costs, the model must minimize the replacement costs(minimize false negatives). ‘**Recall**’ is the metric of concern that must be maximized for a robust predictive model for Wind Turbine Generator Failures.

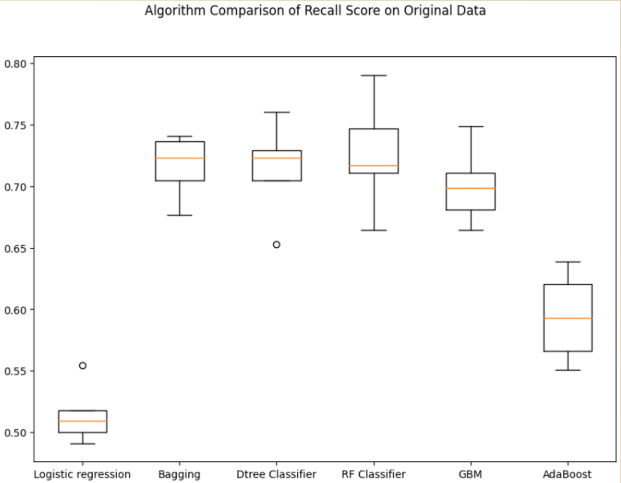
# MODEL BUILDING – ORIGINAL DATA

- 6 Models are built on the original data: Logistic Regression, Bagging, Decision Tree Classifier, Random Forest Classifier, Gradient Boost Classifier, AdaBoost Classifier.

- A k-fold Cross Validation with 5 splits is conducted on each of the models and the CV-recall scores on the training and validation test are summarized in the table to the right.

- Bagging, Decision Tree, Random Forest and Gradient Boosting models all have a good recall score(>70%) and generalize well on the validation set.

- The imbalance of classes may induce bias so under and over sampling will be performed.



CV Recall Scores - Original Data

| Model | Train Set | Validation Set |
|---|---|---|
| Logistic Regression | 0.51 | 0.42 |
| Bagging | 0.72 | 0.69 |
| Decision Tree | 0.71 | 0.7 |
| Random Forest | 0.73 | 0.69 |
| Gradient Boosting | 0.7 | 0.66 |
| Adaboost Classifier | 0.59 | 0.53 |

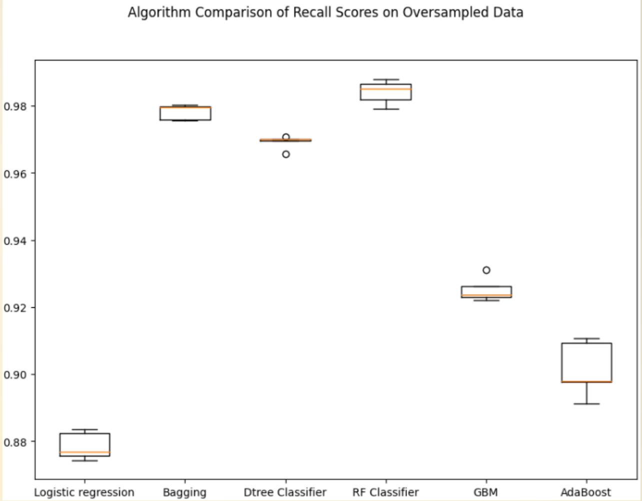
# MODEL BUILDING – OVERSAMPLED DATA

- The SMOTE technique with the k-nearest neighbor algorithm is used to oversample the minority class.

- Post oversampling, all the model's recall scores have increased drastically.

- Gradient Boosting seems to be the model that both performs and generalizes extremely. Recall: train set(93%); validation set(88%).

- Bagging and Random Forests perform the best on training set but their recall scores fall a lot on the validation set implying overfitting.


Algorithm Comparison of Recall Scores on Oversampled Data

| CV Recall Scores - Oversampled Data | | |
|---|---|---|
| Model | Train Set | Validation Set |
| Logistic Regression | 0.88 | 0.84 |
| Bagging | 0.98 | 0.80 |
| Decision Tree | 0.97 | 0.78 |
| Random Forest | 0.98 | 0.82 |
| Gradient Boosting | 0.93 | 0.88 |
| Adaboost Classifier | 0.90 | 0.82 |

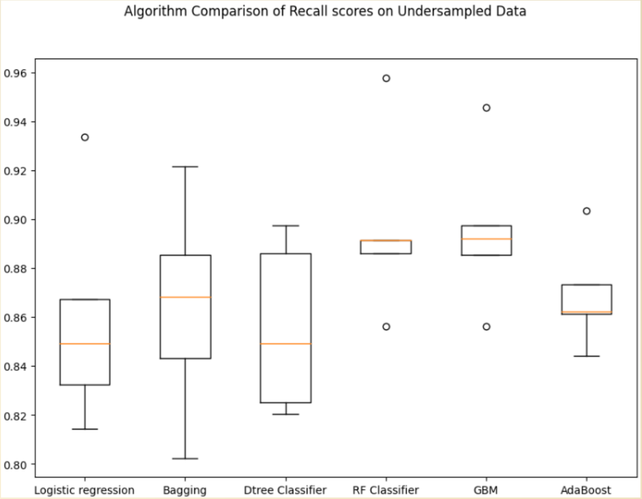
# MODEL BUILDING – UNDER-SAMPLED DATA

- The Random Under Sampler algorithm is used to under-sample the majority class.

- The recall scores have increased drastically from the original data, but have fallen slightly from the oversampled data.

- Unlike on the oversampled data, none of the models are overfitting on under-sampled data.

- The Random Forest, Gradient Boosting and Adaboost Models have high recall scores on the train set and also generalize extremely well on the validation set.



CV Recall Scores – Under-sampled Data

| Model | Train Set | Validation Set |
|---|---|---|
| Logistic Regression | 0.86 | 0.85 |
| Bagging | 0.86 | 0.86 |
| Decision Tree | 0.86 | 0.83 |
| Random Forest | 0.89 | 0.88 |
| Gradient Boosting | 0.90 | 0.88 |
| Adaboost Classifier | 0.87 | 0.87 |

# MODEL SELECTION FOR TUNING

Considering the results of the 6 models on original, oversampled and undersampled data, 4 models will be chosen to hyper tune and hopefully improve model performance and reduce overfitting.

Since all the models perform and generalize relatively similarly, 1 of the models chosen will have a short run time, enabling ReneWind to deploy these into production when time is a constraint. The other 3 models will take longer to run, but may end up performing better than the models with a short run time.

The idea is to provide ReneWind with options to pick a model based on business needs and constraints. The 4 chosen models & the dataset:

- Decision Tree on undersampled data. (Least time to run)
- AdaBoost on oversampled data.
- Gradient Boosting on undersampled data.
- Random Forests on undersampled data.

**It is recommended that ReneWind explore other combinations of models and datasets to check if they perform better than the ones analyzed above.*

# HYPERTUNED MODELS

- The 4 Models chosen are tuned using Random Search CV only due to time constraints.*
- All the 4 models are performing extremely well on the train set with a recall score > 0.89.
- All the models seem to overfit the data but Adaboost’s recall difference between train and validation set is the highest.
- Although Adaboost performs the best on the train set, it overfits the data the most amongst the 4 models while tested on the validation set.
- The Decision Tree has the lowest recall score on the train set, but overfits the least amongst the 4 models while tested on the validation set.

| Hypertuned Model Recall Scores | | |
|---|---|---|
| Model/Data | Train Set | Validation Set |
| Decision Tree/Undersampled Data | 0.89 | 0.82 |
| Adaboost/Oversampled Data | 0.99 | 0.84 |
| Gradient Boost/Undersampled Data | 0.98 | 0.89 |
| Random Forests/Undersampled Data | 0.97 | 0.88 |

_*It is recommended that ReneWind explore Grid Search CV as well for a more comprehensive search of optimal parameters._

# FINAL MODEL SELECTION & PERFORMANCE

- The 2 models with the highest recall performance on the validation set are chosen to proceed for final model selection: Random Forest or Gradient Boosting on undersampled data.

- Both these models' performance on the completely unseen test set is checked and the results are shown in the table below.

- Both model's are capable of predicted 87% of actual failures of turbine generators.

- Random Forest is chosen as the best model since it's difference between the validation and test scores is slightly lesser than that of Gradient Boosting's.

Recall Scores of Final 2 Models

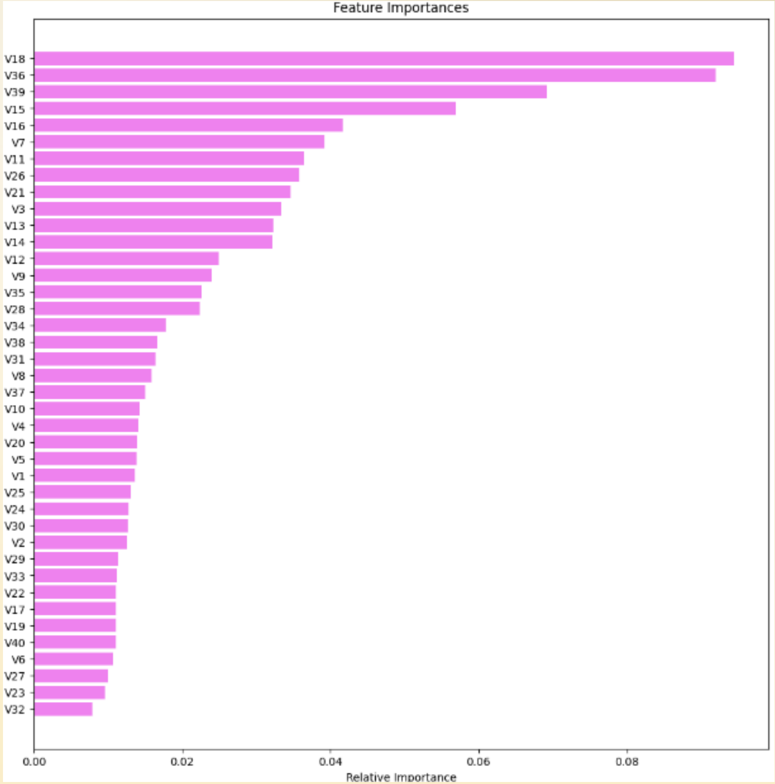
| Model | Validation Set | Test Set |
|---|---|---|
| Gradient Boost on Undersampled Data | 0.89 | 0.87 |
| Random Forest on Undersampled Data | 0.88 | 0.87 |

**Final Model** → Random Forest on Undersampled Data

# FEATURE IMPORTANCE – FINAL MODEL(RANDOM FOREST)

- V18 & V36 are the most important features for making accurate predictions of turbine generator failures.

- V27,V23 & V32 are the least important features in predicting turbine failures.



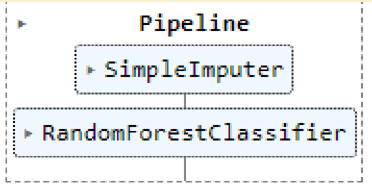
# PIPELINES – PRODUCTIONIZING FINAL MODEL

- A pipeline is constructed to productionize the final model chosen: Random Forest Classifier.

- The pipeline first uses the simple imputer to impute the median into the missing values.

- The Random Classifier is then fit into the undersampled data.

- This pipeline may be used by ReneWind for a simple and easy-to-implement model for predicting wind turbine generator failures.



# BUSINESS RECOMMENDATIONS

- Deploy the Top-Performing Model: Given the Random Forest model's exceptional performance on undersampled data, we recommend ReneWind to integrate this model into their predictive maintenance strategy. Its high recall rate ensures a proactive approach to identify true generator failures, reducing repair costs, and improving turbine reliability.

- Emphasize Feature Importance: Prioritize the maintenance of components represented by important features, namely V18, V36, and V39. By focusing on these critical aspects, ReneWind can further optimize their maintenance efforts, effectively extending the lifespan of wind turbines and mitigating potential failure-related expenses.

- Implement the Production Pipeline: Adopt the modeled production pipeline to streamline the process of predictive maintenance implementation. This will enable real-time monitoring of wind turbine health, allowing for timely interventions and reducing downtime, thereby enhancing overall energy generation efficiency.

- Continuously Update Models: As the wind energy sector evolves, ReneWind should regularly reevaluate and update the predictive maintenance models. Incorporating new data and improvements in machine learning techniques will ensure the company stays at the forefront of wind energy production efficiency and cost-effectiveness.

# THANK YOU

ARPAN DINESH