RENEWIND PROJECT

COURSE TITLE: MODEL TUNING

DATE: JULY 22nd, 2022

CONTENTS

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

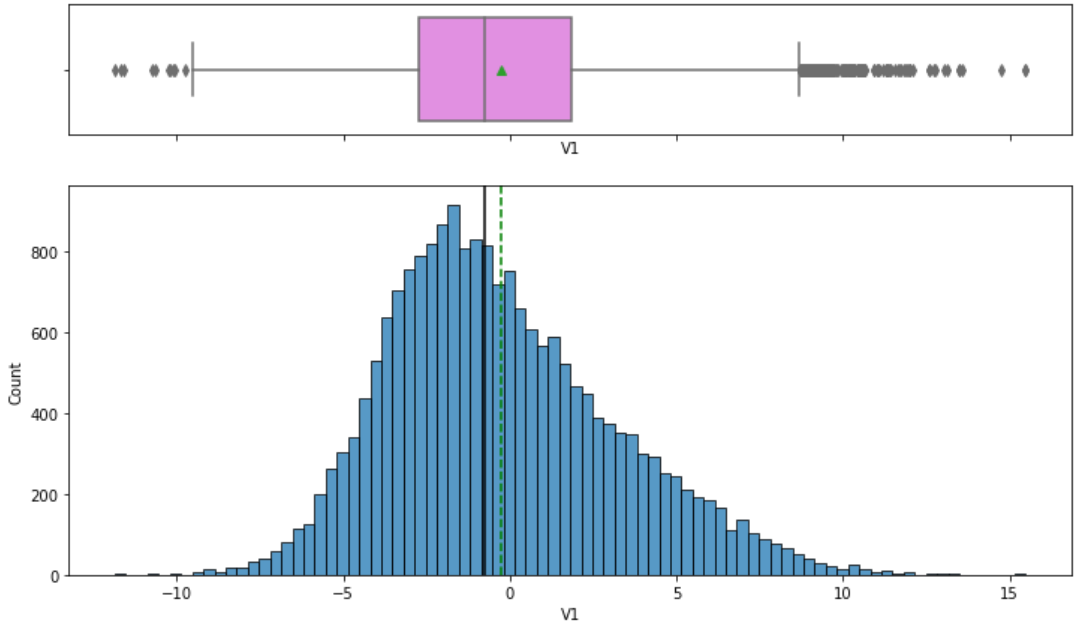
- Renewable energy sources play an important role in the global energy mix, as the effort to reduce the environmental impact of energy production increases.
- Wind energy is one of the most developed technologies worldwide and the U.S Department of Energy has put together a guide to achieving operational efficiency using predictive maintenance practices.
- Predictive maintenance means 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.
- Predictive maintenance uses sensor information 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.) to measure and predict degradation and future component capability.
- 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.
- They have shared a ciphered version of the data, as the data collected through sensors is confidential (the type of data collected varies with companies). Data has 40 predictors, 20000 observations in the training set and 5000 in the test set.
- 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.

DATA OVERVIEW

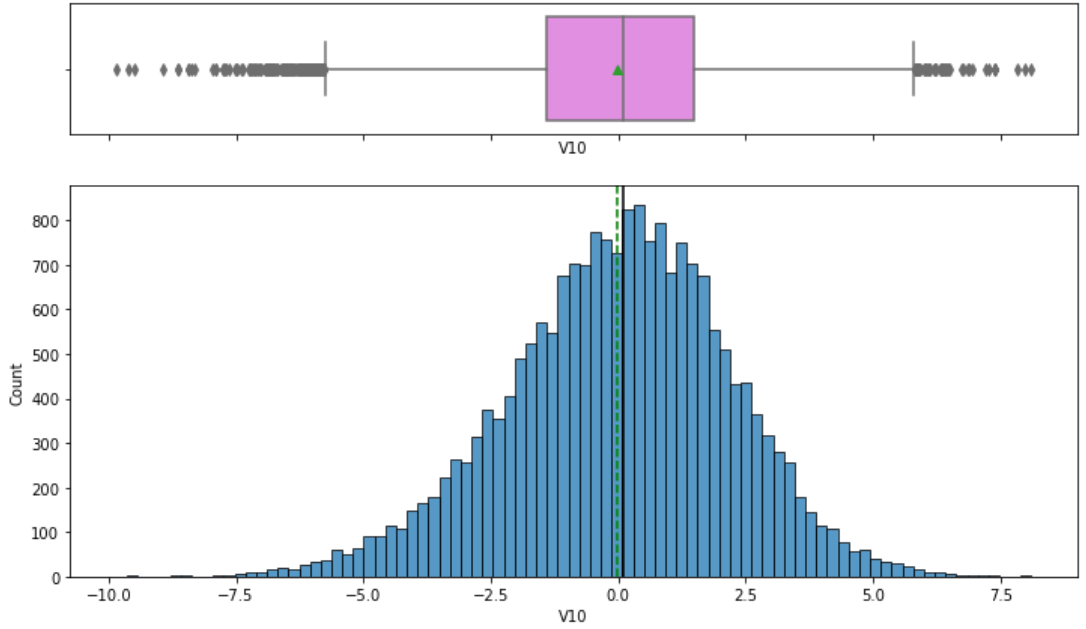
- The data is divided into two, the train and test set
- The train set consist of 20,000 rows, 41 columns
- The test set consist of 5,000 rows, 41 columns
- The data set consist of 40 predictor variable and 1 target variable
- The target variable has two unique numbers 0 & 1, identifying as not failure and failure respectively
- The train data for target variable had the following distribution
  - A total of 18890 machines in good condition
  - a total of 1110 failure which means it needs repair before breaking down
- The test data for target variable had the following distribution
  - A total of 4718 machines in good condition
  - a total of 282 failure which means it needs repair before breaking down

EXPLORATORY DATA ANALYSIS(EDA)

- The average V1 predictor variable is higher than the median for V1 predictor variable indicating the distribution is skewed to the right
- The V1 predictor variable is distributed between -10 to 15
- there are outliers in the distribution

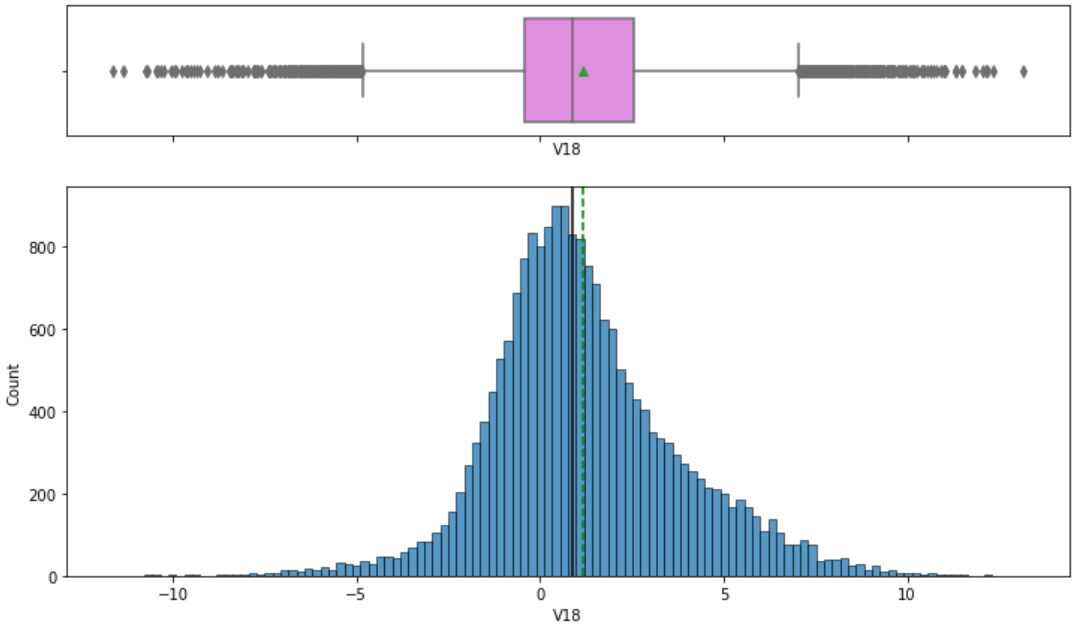


- The average V10 predictor is almost the same with the median V10 predictor indicating the median is nearly symmetrical
- The V10 predictor variable is almost evenly distributed between -7.5 to 7.5
- there are outliers in the distribution

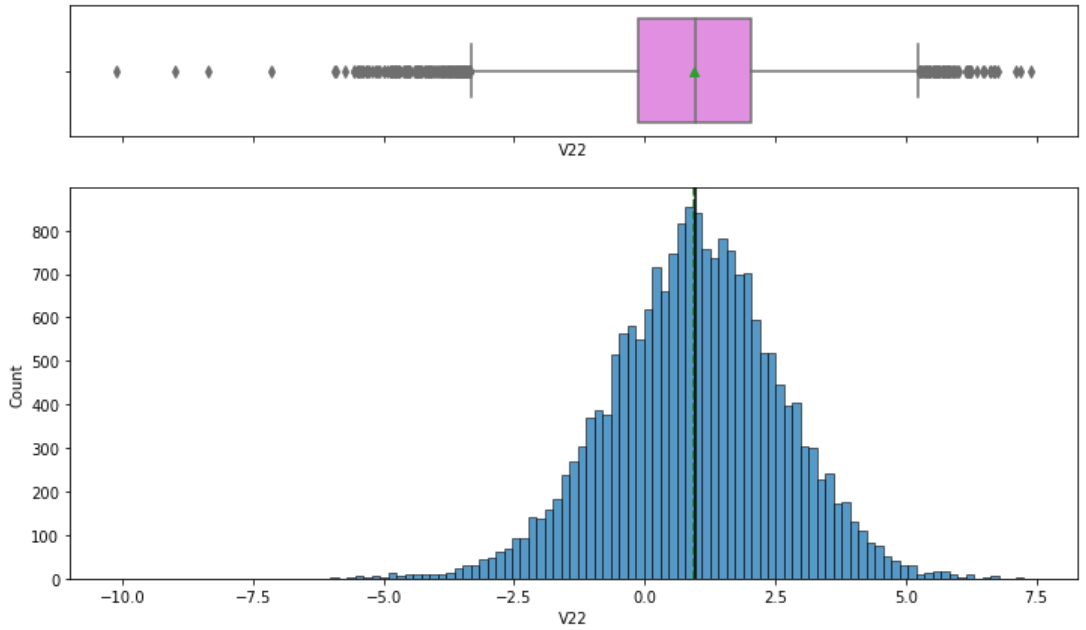


EXPLORATORY DATA ANALYSIS(EDA)

- The average V18 predictor variable is higher than the median for V18 predictor variable indicating the distribution is skewed to the right
- The V1 predictor variable is almost evenly distributed between -10 to 10
- there are outliers in the distribution

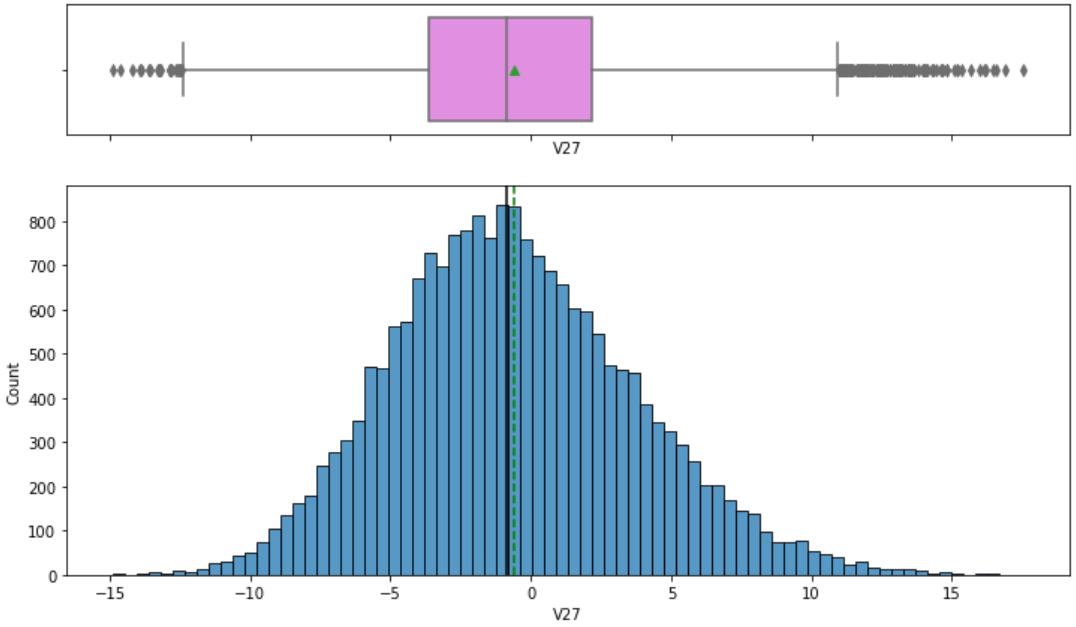


- The average V22 predictor is the same with the median V22 predictor indicating the median is symmetrical
- there are outliers in the distribution

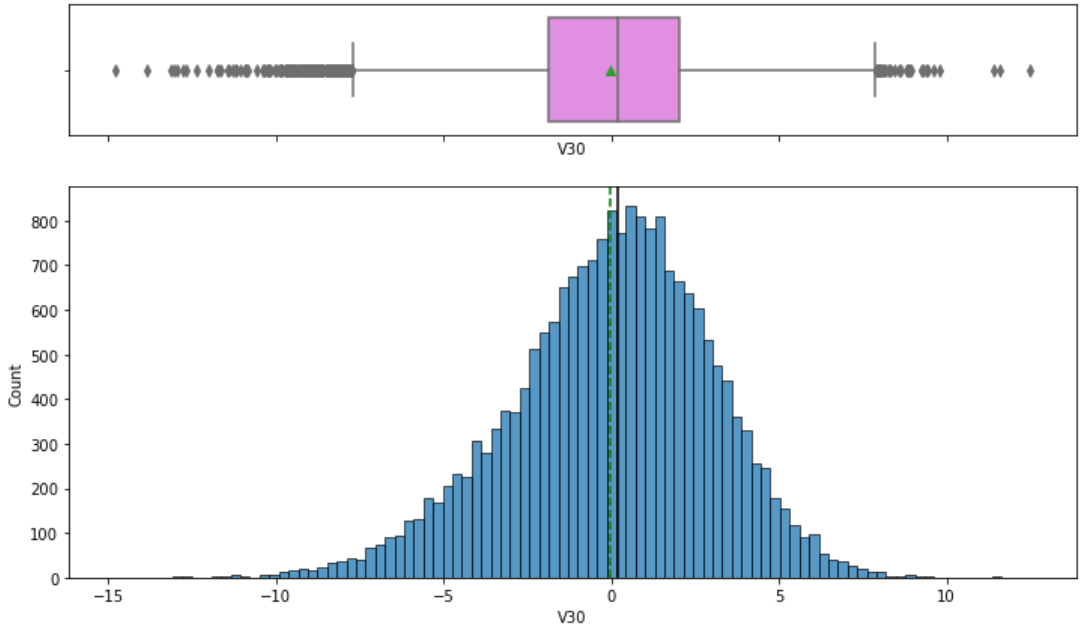


EXPLORATORY DATA ANALYSIS(EDA)

- The average V27 predictor variable is higher than the median for V27 predictor variable indicating the distribution is slightly skewed to the right
- there are outliers in the distribution

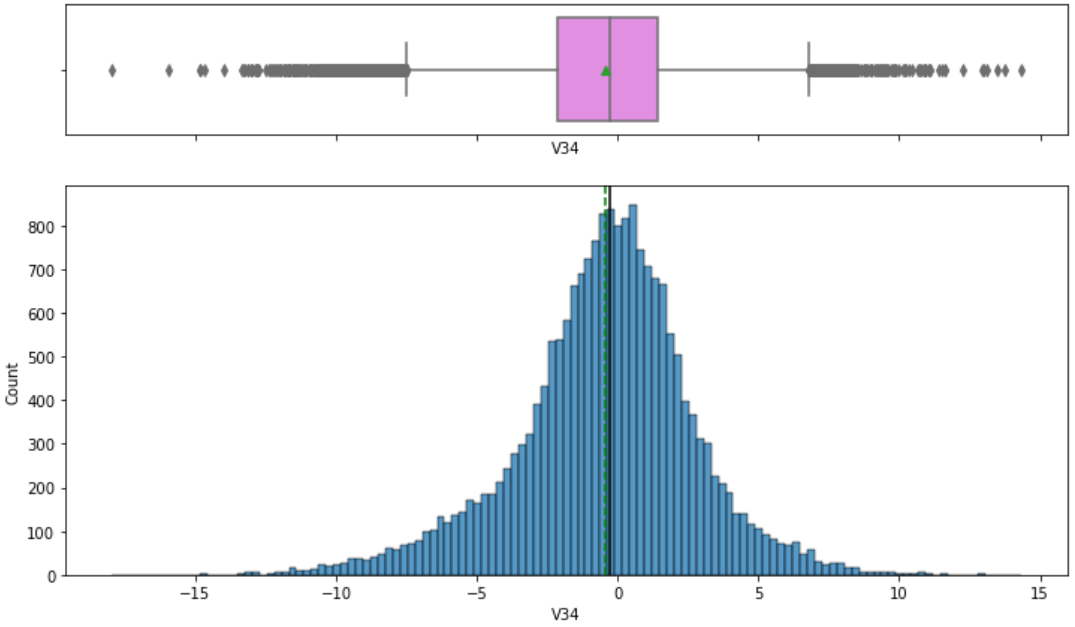


- The average V30 predictor variable is lower than the median for V30 predictor variable indicating the distribution is skewed to the left
- there are outliers in the distribution

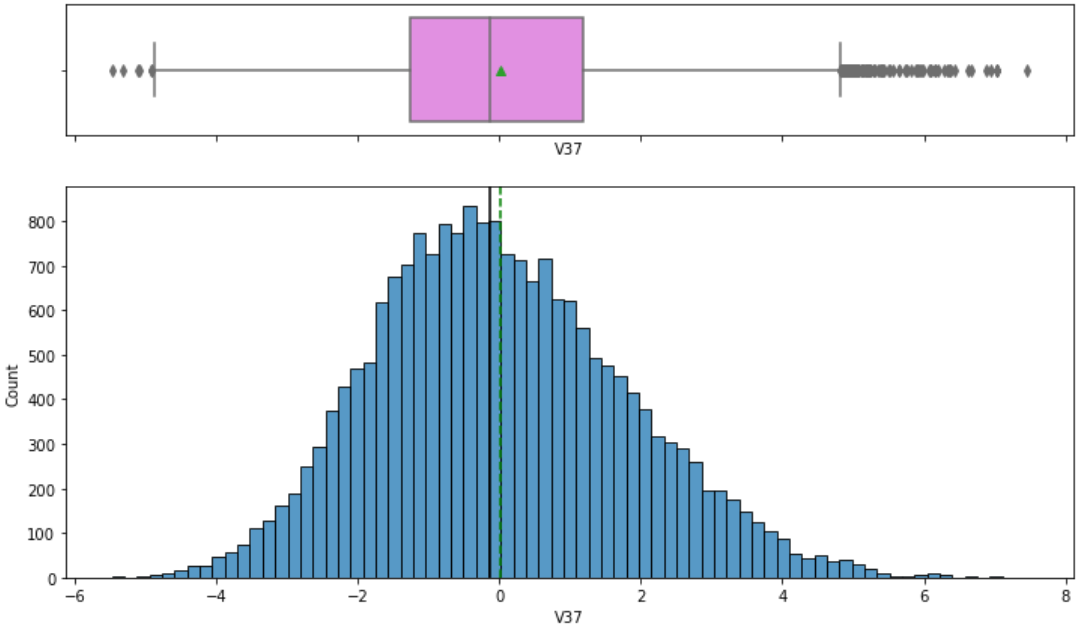


EXPLORATORY DATA ANALYSIS(EDA)

- The average V34 predictor variable is almost the same to the median for V34 predictor variable indicating the distribution is nearly symmetrical
- The V1 predictor variable is almost evenly distributed between -7.5 to 7.5
- there are outliers in the distribution

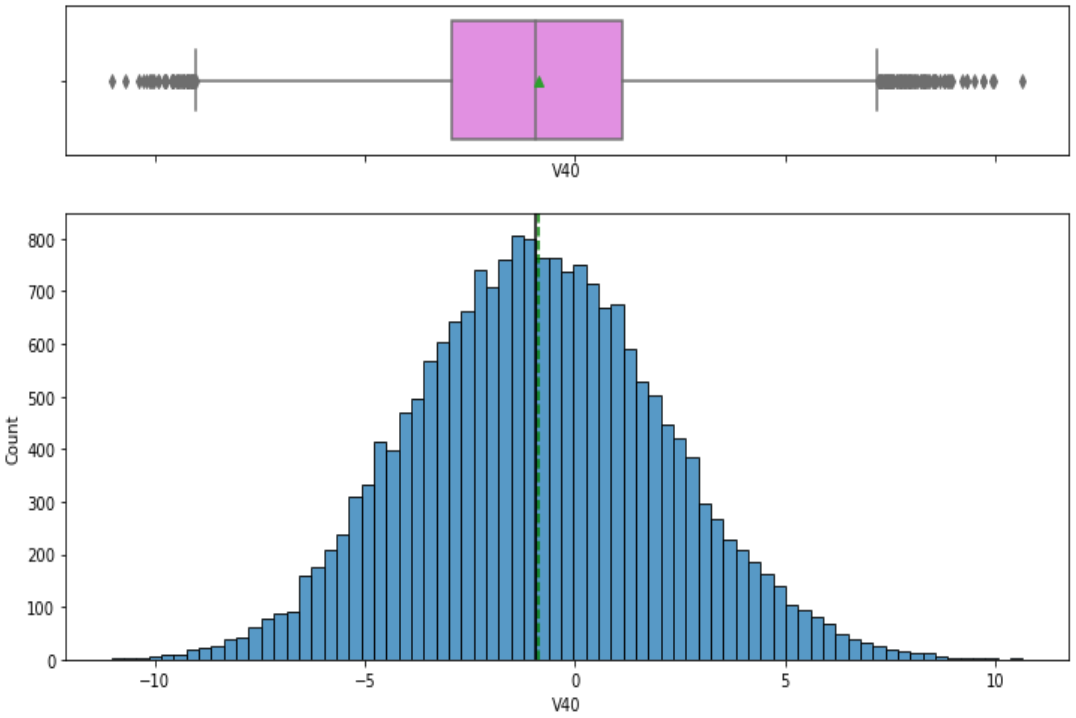


- The average V37 predictor variable is higher than the median for V37 predictor variable indicating the distribution is skewed to the right
- there are outliers in the distribution

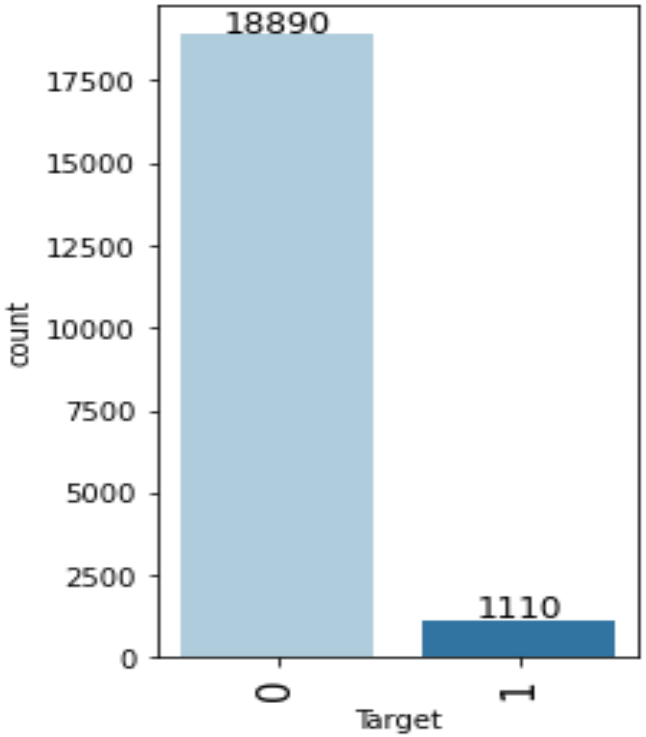


EXPLORATORY DATA ANALYSIS(EDA)

- The average V40 predictor variable is almost the same to the median for V40 predictor variable indicating the distribution is nearly symmetrical
- The V40 predictor variable is almost evenly distributed between -10 to 10
- there are outliers in the distribution



- The target variable shows 18890 predictor are in good condition
- The distribution predicts 1110 predictors need repair



DATA PREPROCESSING

- There are no duplicates in the data set
- There are missing value in both the train and test data indication predictor V1 and V2 with missing value
- The data preparation will be used 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 Performance summary

- 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
- Recall will be used as performance metric of evaluation,
  - A high recall means failures correctly predicted by the model. These will result in repairing costs.
  - The lower the recall means 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.
  - the greater the Recall, the higher the chances of minimizing false negatives.
- The most significant predictors variable to identify failure
  - V30
  - V9
  - V18
  - V12
  - V36
  - V3

Model Performance summary

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY	TRAIN RECALL	VALIDATION RECALL	TRAIN PRECISION	VALIDATION PRECISION	TRAIN F1	VALIDATION F1
Gradient Boosting tuned with oversampled data	0.99	0.97	0.99	0.84	0.99	0.69	0.99	0.76
AdaBoost classifier tuned with oversampled data	0.99	0.97	0.98	0.85	0.99	0.79	0.99	0.82
Random forest tuned with undersampled data	0.96	0.93	0.93	0.88	0.98	0.46	0.96	0.61

Model Building with Pipeline

Steps taken to create a pipeline for the final model

- Create pipeline with the best model which was adaboost classifier tuned with oversampled data
- Next, separate target variable and other variables into independent variable and target for train data
- 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.
- Oversample/undersample the train data and create necessary variables for them (if needed)
- Fit the model on train data
- Check the performance on test set

The performance of the model built with pipeline on the test dataset

- The test recall is 79% which is low compared to the precision which is 98%, this cant fully predict failures and can lead to maximizing replacement cost. This is not a great model to predict failures

The most important factors used by the model built with pipeline for prediction are;

V30,V9, & V18.

MODEL PERFORMANCE SUMMARY (original data)

- We can see that the xgboost is giving the highest cross-validated recall followed by random forest and bagging
- The boxplot shows that the performance of xgboost, random forest and bagging is consistent and their performance on the validation set is also good

Model	Cross validation cost	Validation performance
Logistic regression	0.49	0.48
Bagging	0.72	0.73
Random forest	0.72	0.72
Gradient boost	0.70	0.72
Adaboost	0.63	0.67
Xgboost	0.79	0.82

MODEL PERFORMANCE SUMMARY (oversampled data)

we chose the random search cv for the oversampling method for gradient boost and ada boost classifier.

The gradient boost and adaboost classifier did not have the highest recall in both the cross validation and validation set.

After hyper tuning the data there was a general performance increase which made it a better option in predicting failures with the highest recall

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY	TRAIN RECALL	VALIDATION RECALL	TRAIN PRECISION	VALIDATION PRECISION	TRAIN F1	VALIDATION F1
Gradient Boosting tuned with oversampled data	0.99	0.97	0.99	0.84	0.99	0.69	0.99	0.76
AdaBoost classifier tuned with oversampled data	0.99	0.97	0.98	0.85	0.99	0.79	0.99	0.82

MODEL PERFORMANCE SUMMARY (undersampled data)

we chose the random search cv for the undersampling method for random forest.

The random forest performed well in the original data for both the cross validation cost and validation performance.

After hyper tuning the data the training the performance improved but validation precision was very low and it also gave the highest recall.

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY	TRAIN RECALL	VALIDATION RECALL	TRAIN PRECISION	VALIDATION PRECISION	TRAIN F1	VALIDATION F1
Random forest tuned with undersampled data	0.96	0.93	0.93	0.88	0.98	0.46	0.96	0.61

BUSINESS INSIGHTS AND CONCLUSION

The following are the insights the data displayed

- The logistic regression shows performance with oversampled data on training set varies between 0.83 to 0.86 recall
- The logistic regression shows performance with oversampled data on training set varies between 0.86 to 0.88 recall

The following will be recommended for RENEWIND;

The best model to help identify failures so that the generators could be repaired before failing/breaking to reduce the overall maintenance cost will be the Adaboost classifier tuned with oversampled data, Which presents the best recall of 85% with overall good performance across all set and this will help minimize false negatives.