ReneWind

ReneWind – Model Tunning

14/05/2022

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- As part of the results it shows that six vectors have a higher relevance for the analysis of a failure of a wind turbine. Through analysis using different machine model learning models, it shows that the model can predict the failures in wind turbines. The model was developed with the thought of obtaining the highest amount of True Positive with the lowest False Negatives. These will reduce the amount spent on replacements when it isn't necessary and also the inspection costs.

- The six vectors of data that were responsible for the highest importance can help to check specific sections of the turbines to reduce the cost of stoppage time in the wind turbines with a precise analysis of the repairs needed in the turbine.

Business Problem Overview and Solution Approach

The U.S Department of Energy has put together a guide to achieving operational efficiency using predictive maintenance practices.

To predict maintenance the usage of sensors of information and analysis methods are measured 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.

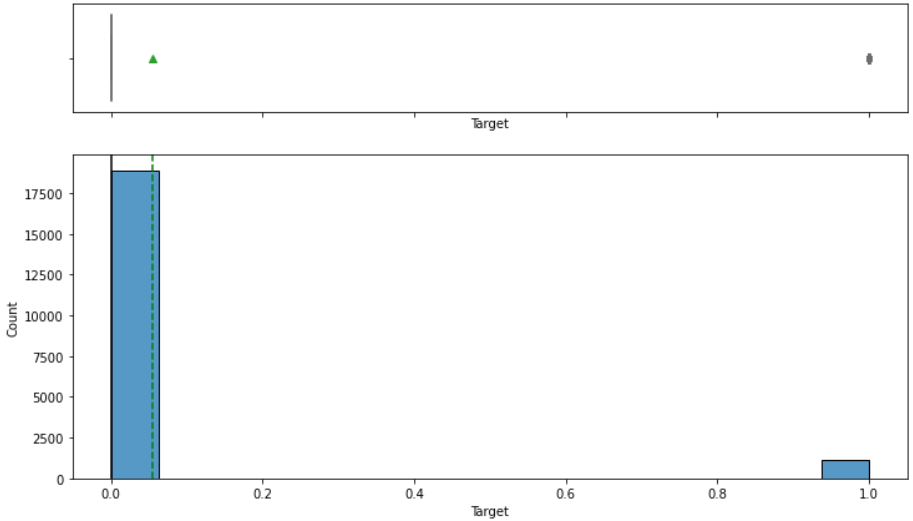
The sensors fitted across different machines involved in the process of energy generation collect data related to various environmental factors (temperature, humidity, wind speed, etc.) and additional features related to various parts of the wind turbine (gearbox, tower, blades, break, etc.).

“ReneWind” is a company working on improving the machinery/processes involved in the production of wind energy using machine learning and has collected data on 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).

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.

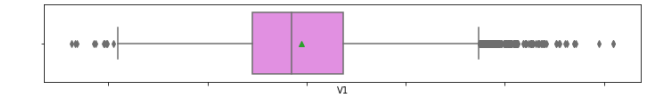
EDA Results

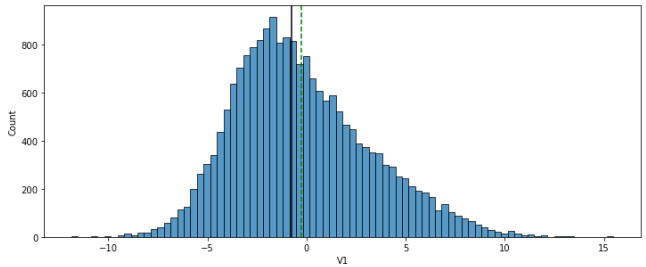
- The results obtained from performing the visual analysis of the 41 variables show a normal behavior for all the variables. The target variable shows that 95% of the measures didn't result in a failure and the rest consist of a failure.
- An example of the univariate analysis of the first four Vectors is in the following slide, but the rest can be found in the appendix.

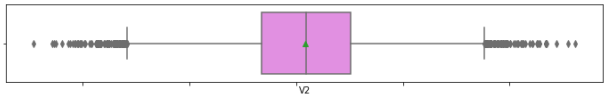


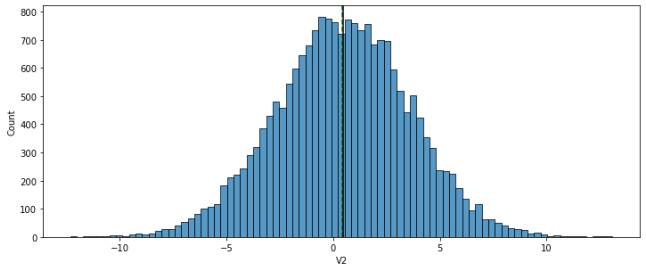
Link to Appendix slide on Data Background and Contents

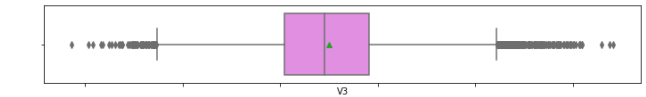
EDA Results

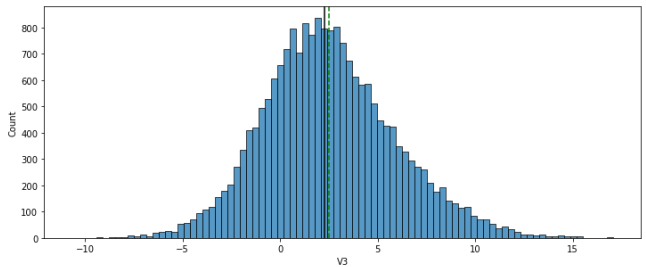


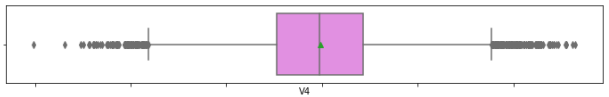


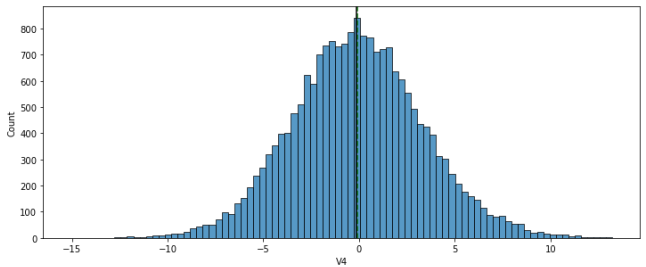










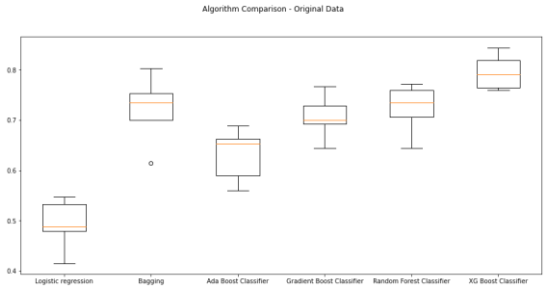


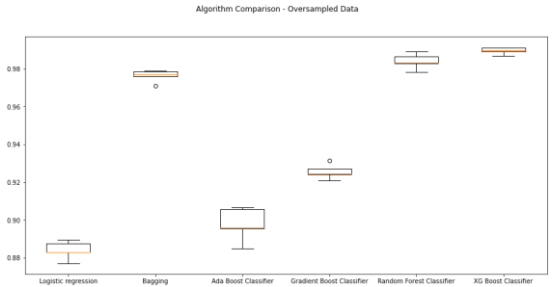
Data Preprocessing

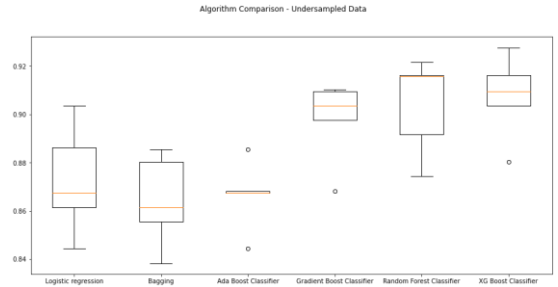
- A duplicated value check was realized, and it was found that none of the values was duplicated.
- A missing value check was done, and it was found that V1 and V2 had 18 missing values each. For the treatment of these values, the missing values were replaced with the median of the values of each variable.
- Due to the normal distribution of all the variables no action was taken as a consideration of the outliers values.
- First the model preparation the train data vas split in a 3 to 1 ratio to have a validation set. This set was used to evaluate the results obtained from the models from the train data set and use the test set for evaluation purposes only. Also, the train and validation sets were stratified so these sets would have a similar distribution of failure cases as the original data set.
- The test set was used only after the model was tunned through the hyperparameters to prevent data leakage.

Model Performance Summary

In order to decide which model would be the best fit for our solution. Following this approach, a randomized grid search was done. The first one used the original data set (image on the left), the second one an oversample data set (image of the center), and the third one an undersample set (image on the right). Based on the results four models were selected: AdaBoost using oversampled data, Random forest using undersampled data, Gradient Boosting using oversampled data and, XGBoost using oversampled data.



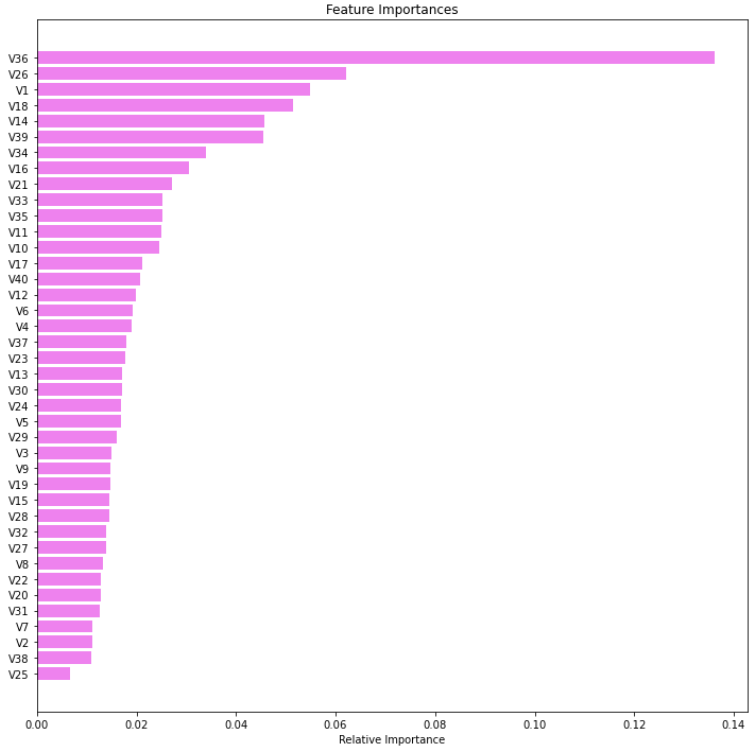




Link to Appendix slide on model assumptions

Model Performance Summary

- The models selected were then tuned the final model will be selected based on the performance of the recall parameter. This is due to the following:
    1. True positive (TP) are failures correctly predicted by the model. These will result in repair costs.
    2. False negatives (FN) are real failures where there is no detection by the model. These will result in replacement costs.
    3. False positives (FP) are detections where there is no failure. These will result in inspection costs.
- Since we want to reduce these costs the recall parameter is the selected parameter to evaluate the parameters.



Link to Appendix slide on model assumptions

Model Performance Summary

	Gradient Boosting tuned with oversampled data	AdaBoost classifier tuned with oversampled data	Random forest tuned with undersampled data	XGBoost tuned with oversampled data
Training Set				
Accuracy	0.993	0.992	0.961	0.929
Recall	0.993	0.988	0.933	**0.998**
Precision	0.994	0.995	0.989	0.877
F1	0.993	0.992	0.960	0.934
Validation Set				
Accuracy	0.971	0.979	0.938	0.855
Recall	0.845	0.853	0.885	**0.914**
Precision	0.693	0.790	0.468	0.266
F1	0.762	0.820	0.612	0.412

Link to Appendix slide on model assumptions

Model Performance Summary

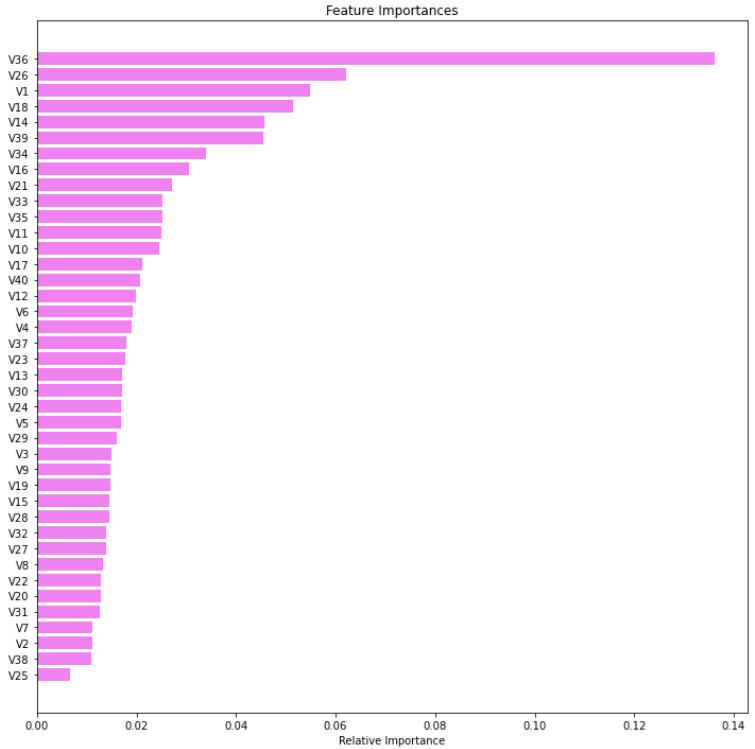
	Gradient Boosting tuned with oversampled data	AdaBoost classifier tuned with oversampled data	Random forest tuned with undersampled data	XGBoost tuned with oversampled data
Test Set				
Accuracy	0.964	0.978	0.988	0.840
Recall	0.844	0.844	0.858	**0.862**
Precision	0.640	0.783	0.920	0.242
F1	0.728	0.812	0.888	0.378

Link to Appendix slide on model assumptions

Model Performance Summary

- Based on the results of the recall parameter the selected model is the XGBoost. As it had the highest recall value over the three data sets.
- The model gave the highest importance to the variable V36 like the one that can predict if a failure is about to occur.
- In order to put this model in production a pipeline was constructed, and it was tested with the model using a SMOTE technique to Over Sample with failures to predict the performance of the model. The results were the following:

Accuracy	Recall	Precision	F1
0.831	0.865	0.232	0.366



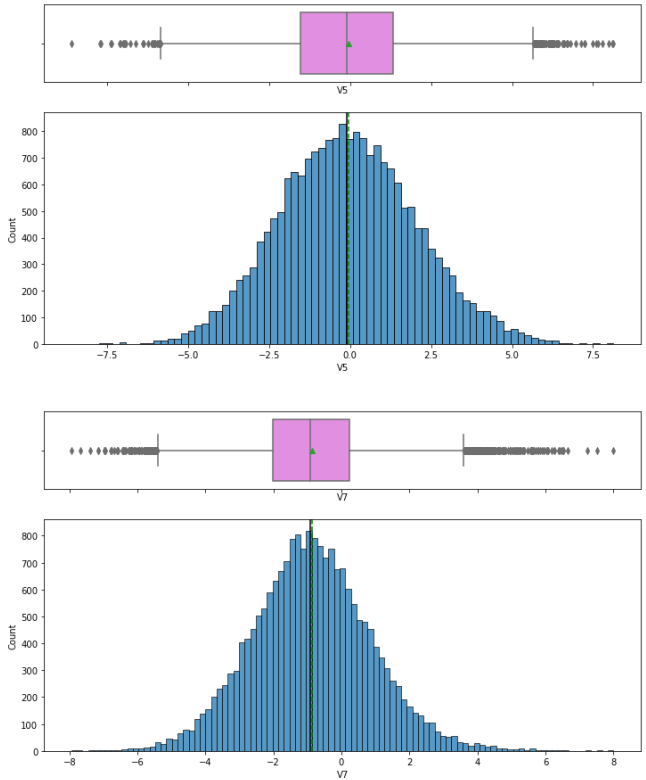
Link to Appendix slide on model assumptions

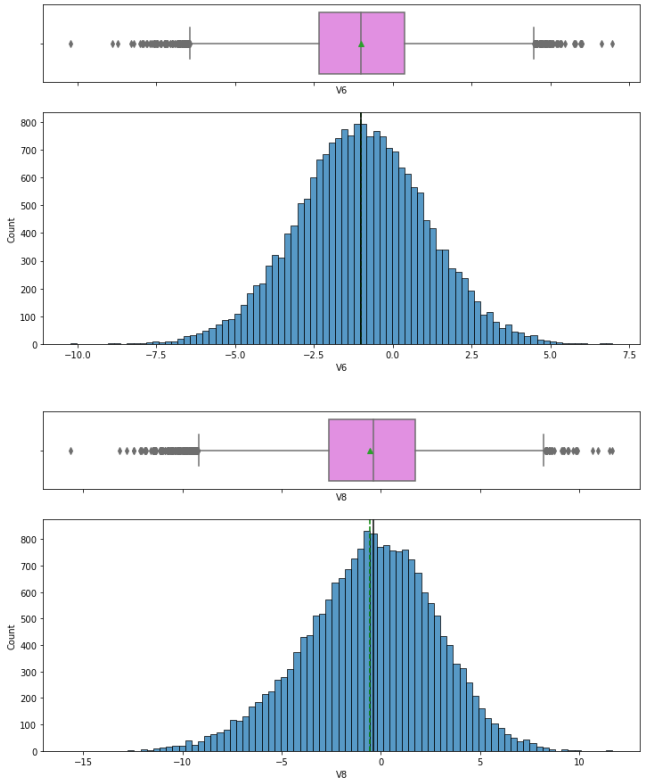
APPENDIX

Data Background and Contents

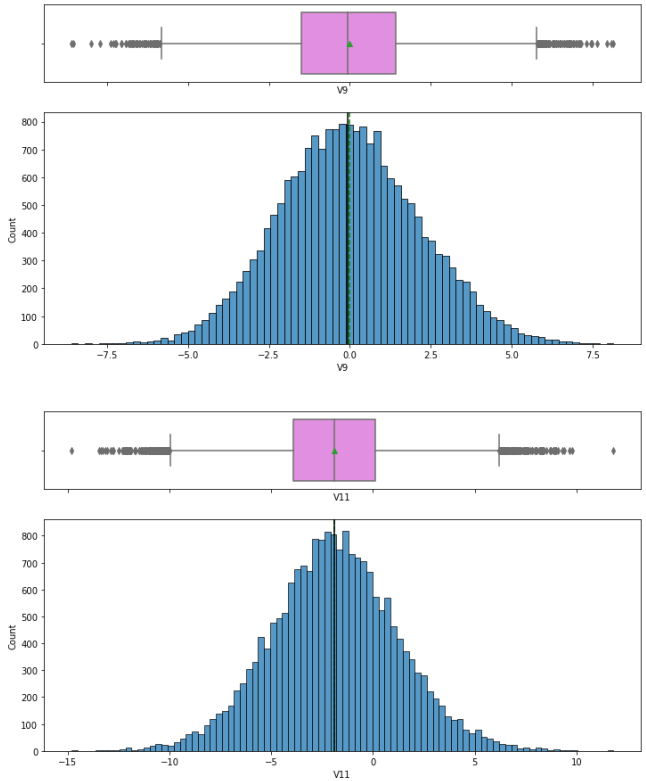
- All the vectors of information come from a float64 type variable, and the target variable is an int64 type variable that has a 1 and 0. The “1” in the target variables should be considered as “failure” and “0” represents “No failure”.

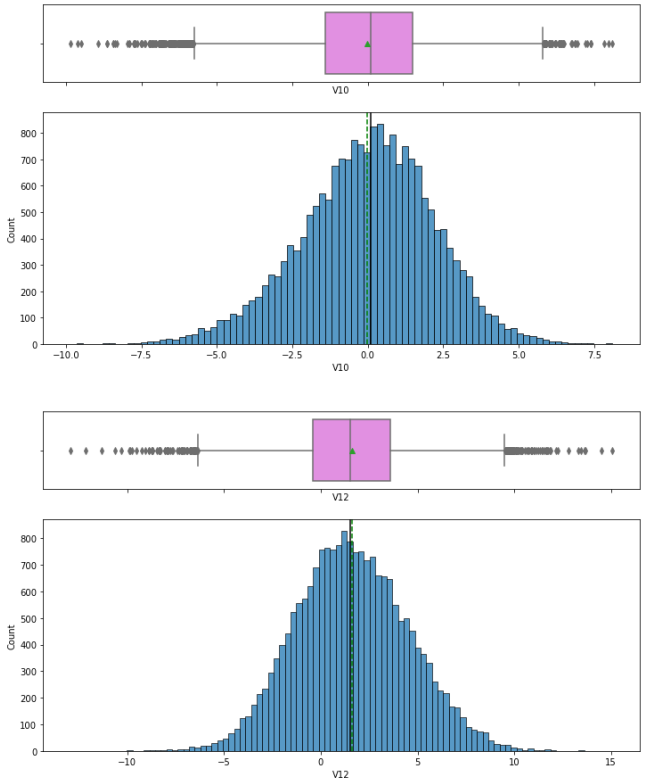
EDA Results - Background and Contents



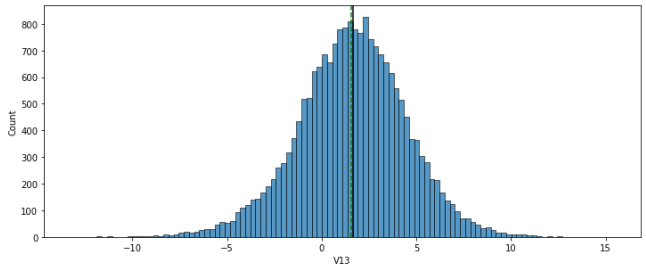


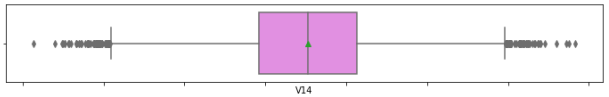
EDA Results - Background and Contents

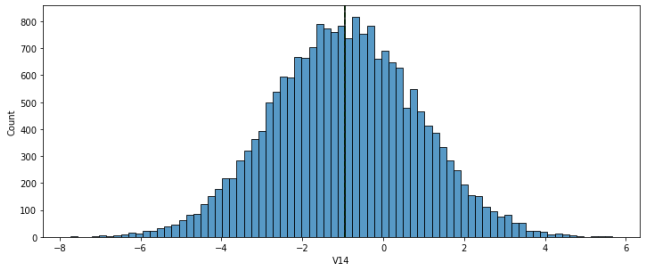


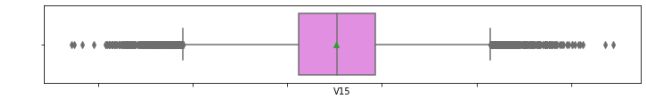


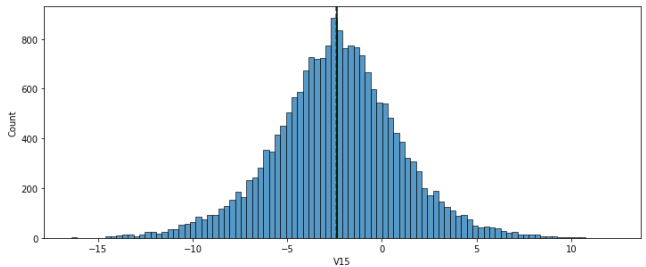
EDA Results - Background and Contents

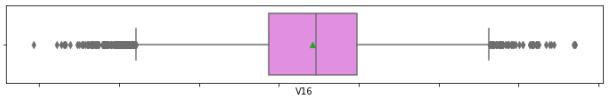


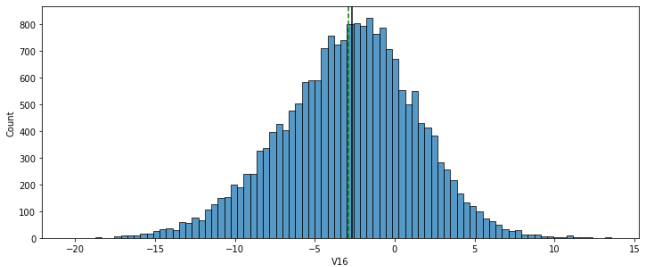




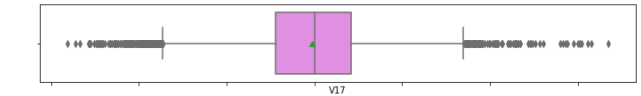


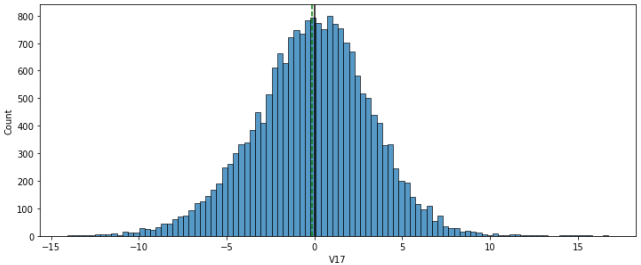


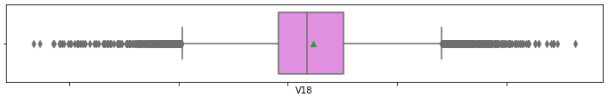


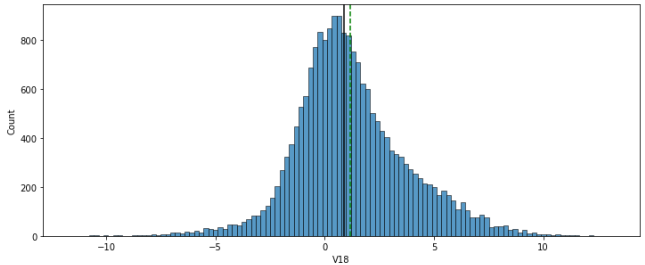


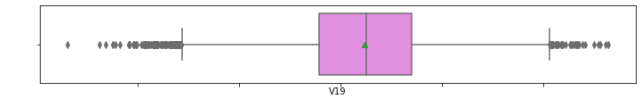
EDA Results - Background and Contents

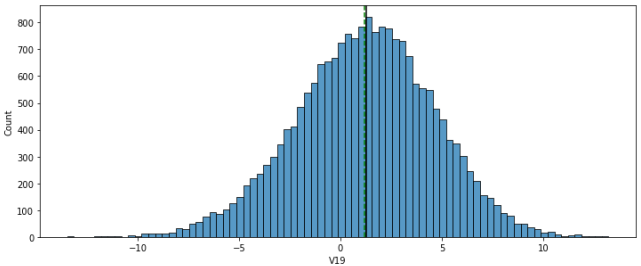


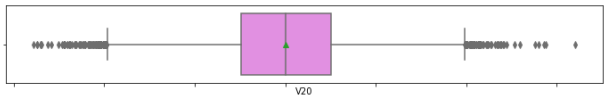


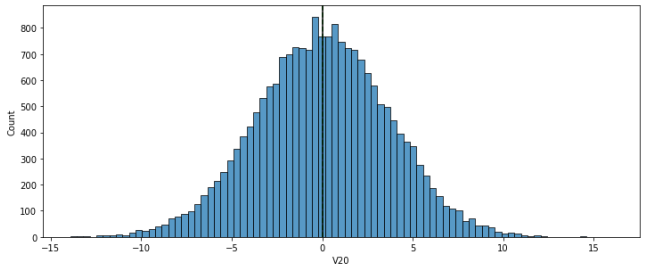




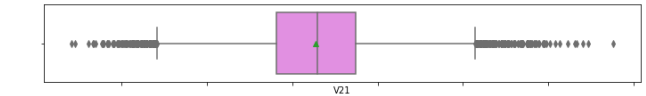


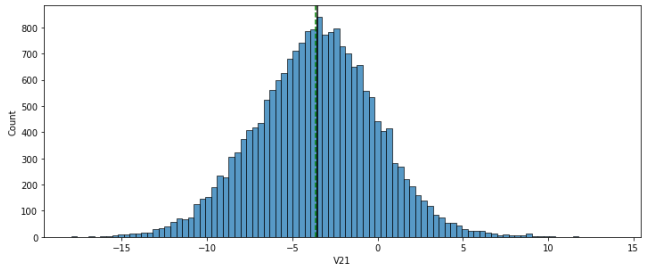


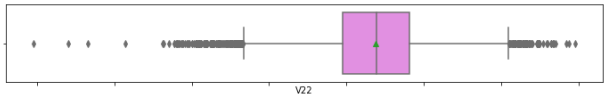


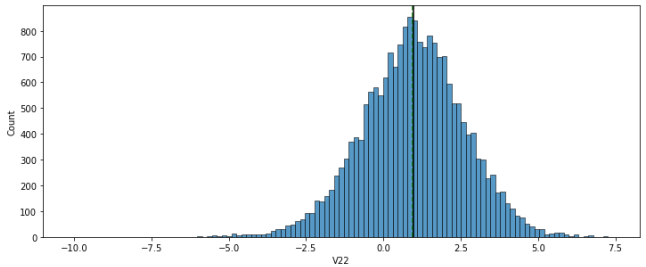


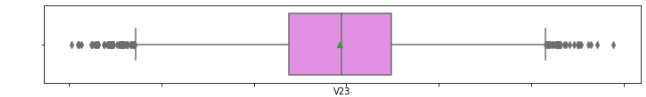
EDA Results - Background and Contents

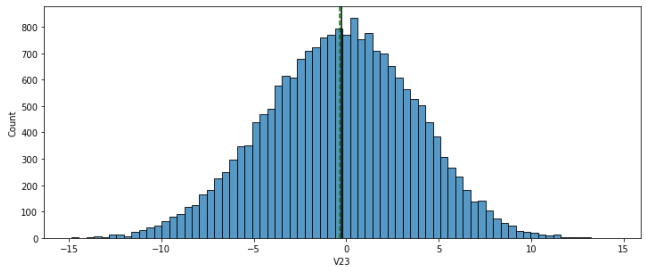


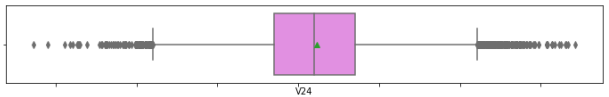


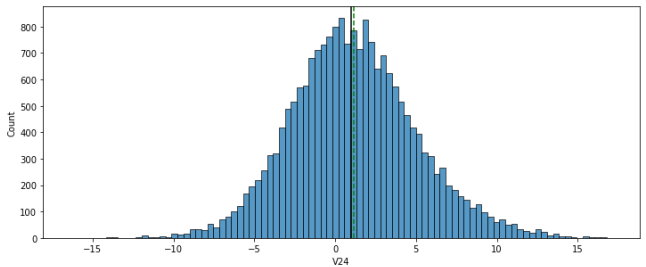




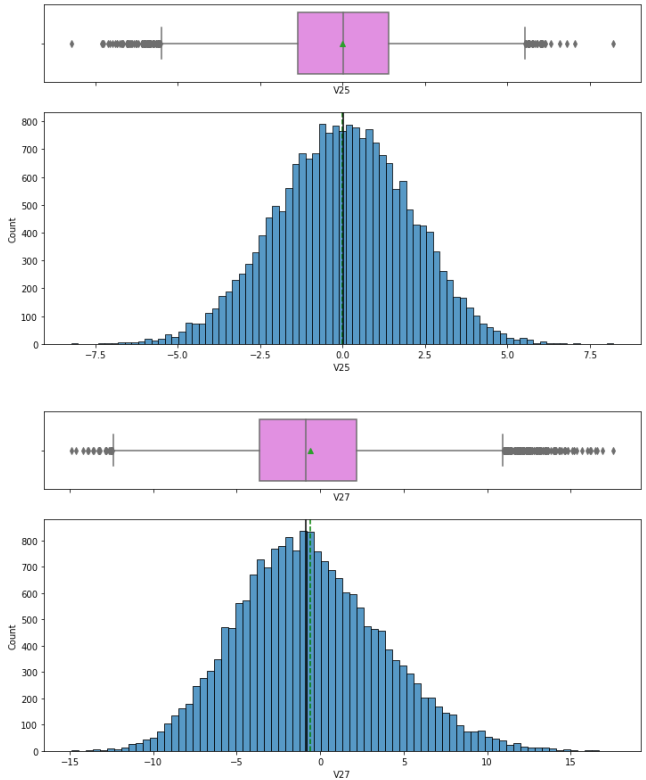


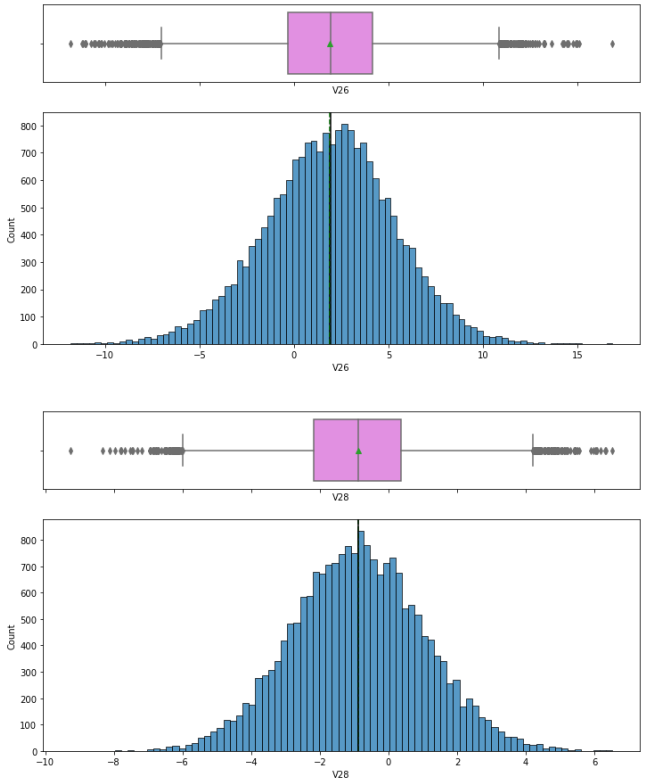




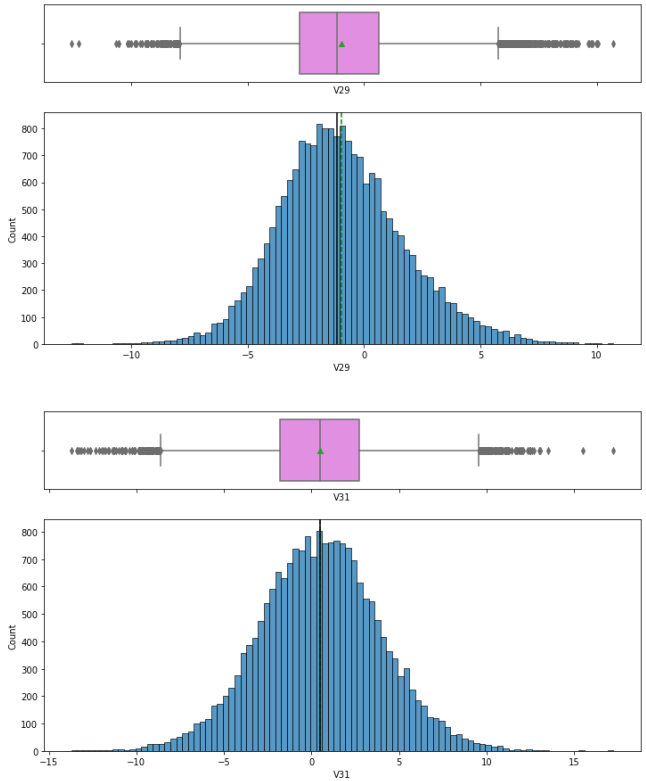


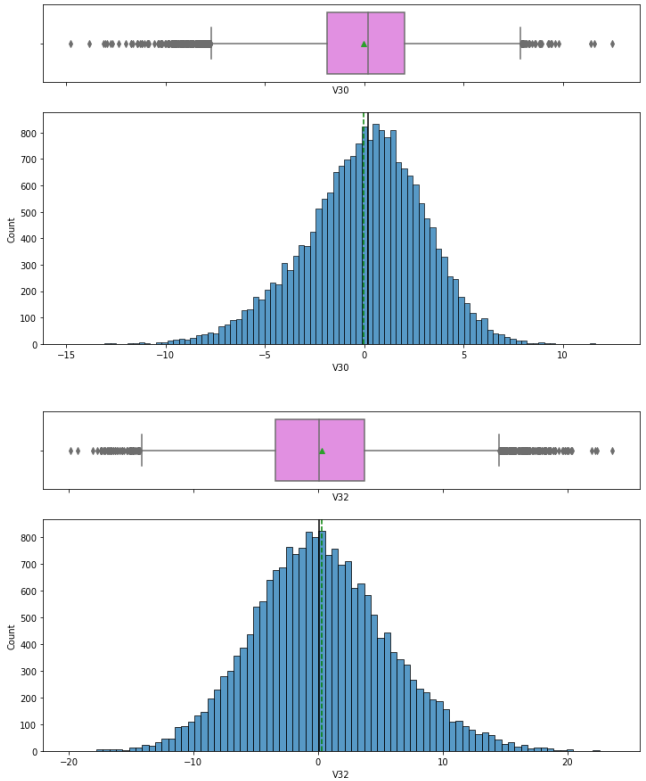
EDA Results - Background and Contents



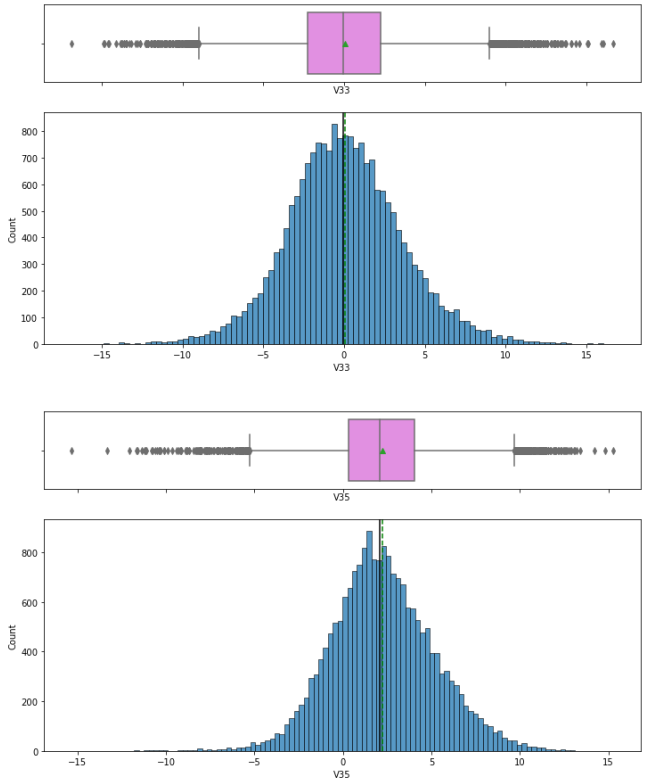


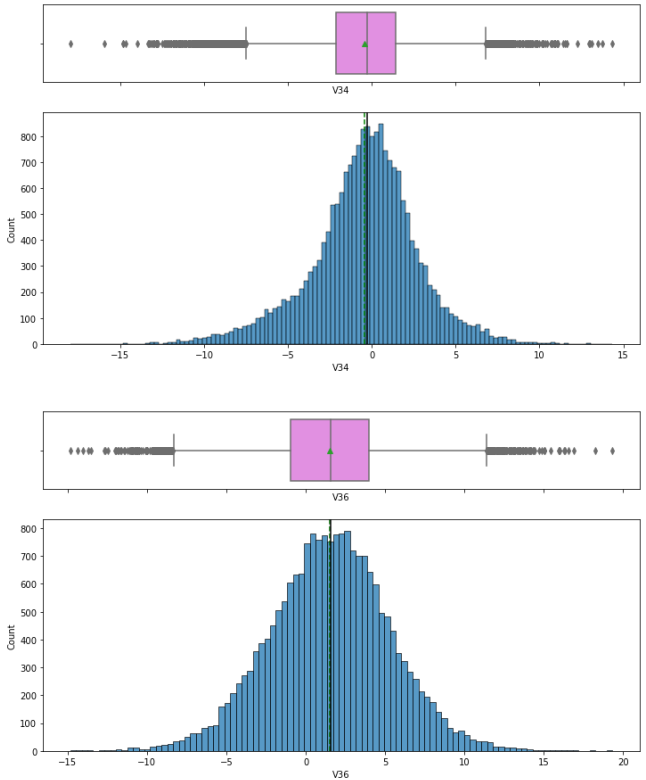
EDA Results - Background and Contents



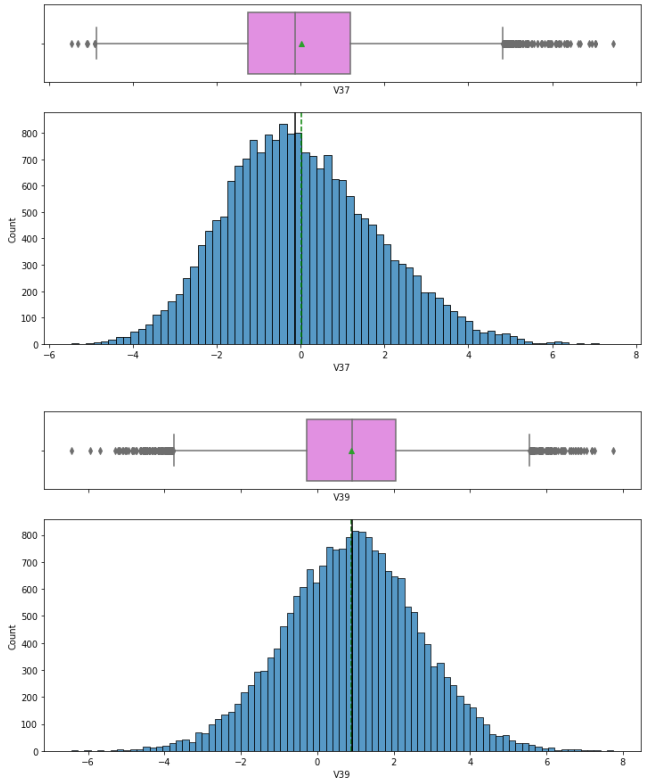


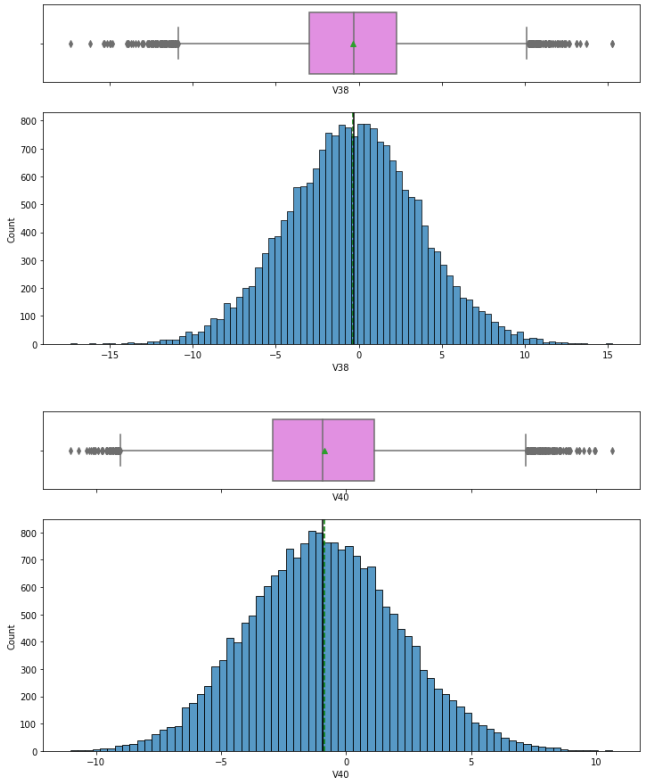
EDA Results - Background and Contents





EDA Results - Background and Contents





Model Assumptions - Model evaluation criterion

The nature of predictions made by the classification model will translate as follows:

1. True positives (TP) are failures correctly predicted by the model.
2. False negatives (FN) are real failures in a generator where there is no detection by the model, these ones are the most expensive for the business.
3. False positives (FP) are failure detections in a generator where there is no failure, it raises the inspection costs but can help prevent the most expensive costs.

The parameter Recall can help to reduce the costs as through the maximization of this one, it minimizes the false negatives.

Great Learning

Happy Learning !