ReneWind Case Study

ReneWind Model Tuning Course

June 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model performance summary for hyperparameter tuning.
- Model building with pipeline

What and How

Executive Summary

- The goal is to help identify failures so that the generators could be repaired before failing/breaking to reduce the overall maintenance cost
- Utilized data to build a classification model that will help identify which factors should be focused on that will lead to a failure
- Identify factors that influenced failures
- Focused on data from
  - A variety of sensors
  - If the sensors had failed or not
  - Relating to environmental factors
  - Relating to various parts of the wind turbine

Conclusions

Executive Summary

- Most of the data was normally distributed or with slight skews
- The Model
  - Does a good job of in the data
  - 85% true positive rate
  - Is good for prediction

Recommendations

Executive Summary

- Focus on the factors from
  - V36
  - V30
  - V18
  - V12
  - V9
  - V35
  - V37
- Continue to utilize the data and the model to notice trends and changes in failure and non-failure rates
- Collect more data to continue to improve the model
- Collect different data sources to see if other factors lead to failures

How can we discover the most important sensors

Business Problem Overview and Solution Approach

- The Approach
  - Developed the questions to explore data with
  - Perfome data overview
  - Exploratory Data Analysis
  - Data Preprocessing
  - Missing value imputations
  - Model Building
  - Hyperparameter Tuning
  - Compaired models and chose the best fit
  - Created Pipeline
  - Developed recomendations
- Find the sensors that are most important to repair
- What does the data tell us?

Data Overview

- 20,000 Rows – Training Set
- 41 Columns – Training Set
- 5,000 Rows – Test Set
- 41 Columns – Test Set
- V1 – V40 (float64)
- Target (int64)
- Float – 40 values
- Int64 – 1 value
- V1 and V2 have 18 missing values
- No duplicate values

Data – Average, Max, Min

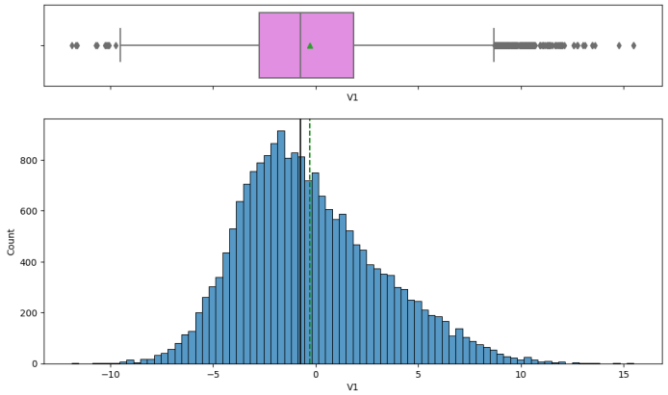
Exploritory Data Analysis

	count	mean	std	min	25%	50%	75%	max
V1	19982.000	-0.272	3.442	-11.876	-2.737	-0.748	1.840	15.493
V2	19982.000	0.440	3.151	-12.320	-1.641	0.472	2.544	13.089
V3	20000.000	2.485	3.389	-10.708	0.207	2.256	4.566	17.091
V4	20000.000	-0.083	3.432	-15.082	-2.348	-0.135	2.131	13.236
V5	20000.000	-0.054	2.105	-8.603	-1.536	-0.102	1.340	8.134
V6	20000.000	-0.995	2.041	-10.227	-2.347	-1.001	0.380	6.976
V7	20000.000	-0.879	1.762	-7.950	-2.031	-0.917	0.224	8.006
V8	20000.000	-0.548	3.296	-15.658	-2.643	-0.389	1.723	11.679
V9	20000.000	-0.017	2.161	-8.596	-1.495	-0.068	1.409	8.138
V10	20000.000	-0.013	2.193	-9.854	-1.411	0.101	1.477	8.108
V11	20000.000	-1.895	3.124	-14.832	-3.922	-1.921	0.119	11.826
V12	20000.000	1.605	2.930	-12.948	-0.397	1.508	3.571	15.081
V13	20000.000	1.580	2.875	-13.228	-0.224	1.637	3.460	15.420
V14	20000.000	-0.951	1.790	-7.739	-2.171	-0.957	0.271	5.671
V15	20000.000	-2.415	3.355	-16.417	-4.415	-2.383	-0.359	12.246
V16	20000.000	-2.925	4.222	-20.374	-5.634	-2.683	-0.095	13.583
V17	20000.000	-0.134	3.345	-14.091	-2.216	-0.015	2.069	16.756
V18	20000.000	1.189	2.592	-11.644	-0.404	0.883	2.572	13.180
V19	20000.000	1.182	3.397	-13.492	-1.050	1.279	3.493	13.238
V20	20000.000	0.024	3.669	-13.923	-2.433	0.033	2.512	16.052
V21	20000.000	-3.611	3.568	-17.956	-5.930	-3.533	-1.266	13.840
V22	20000.000	0.952	1.652	-10.122	-0.118	0.975	2.026	7.410
V23	20000.000	-0.366	4.032	-14.866	-3.099	-0.262	2.452	14.459
V24	20000.000	1.134	3.912	-16.387	-1.468	0.969	3.546	17.163
V25	20000.000	-0.002	2.017	-8.228	-1.365	0.025	1.397	8.223
V26	20000.000	1.874	3.435	-11.834	-0.338	1.951	4.130	16.836
V27	20000.000	-0.612	4.369	-14.905	-3.652	-0.885	2.189	17.560
V28	20000.000	-0.883	1.918	-9.269	-2.171	-0.891	0.376	6.528
V29	20000.000	-0.986	2.684	-12.579	-2.787	-1.176	0.630	10.722
V30	20000.000	-0.016	3.005	-14.796	-1.867	0.184	2.036	12.506
V31	20000.000	0.487	3.461	-13.723	-1.818	0.490	2.731	17.255
V32	20000.000	0.304	5.500	-19.877	-3.420	0.052	3.762	23.633
V33	20000.000	0.050	3.575	-16.898	-2.243	-0.066	2.255	16.692
V34	20000.000	-0.463	3.184	-17.985	-2.137	-0.255	1.437	14.358
V35	20000.000	2.230	2.937	-15.350	0.336	2.099	4.064	15.291
V36	20000.000	1.515	3.801	-14.833	-0.944	1.567	3.984	19.330
V37	20000.000	0.011	1.788	-5.478	-1.256	-0.128	1.176	7.467
V38	20000.000	-0.344	3.948	-17.375	-2.988	-0.317	2.279	15.290
V39	20000.000	0.891	1.753	-6.439	-0.272	0.919	2.058	7.760
V40	20000.000	-0.876	3.012	-11.024	-2.940	-0.921	1.120	10.654
Target	20000.000	0.056	0.229	0.000	0.000	0.000	0.000	1.000

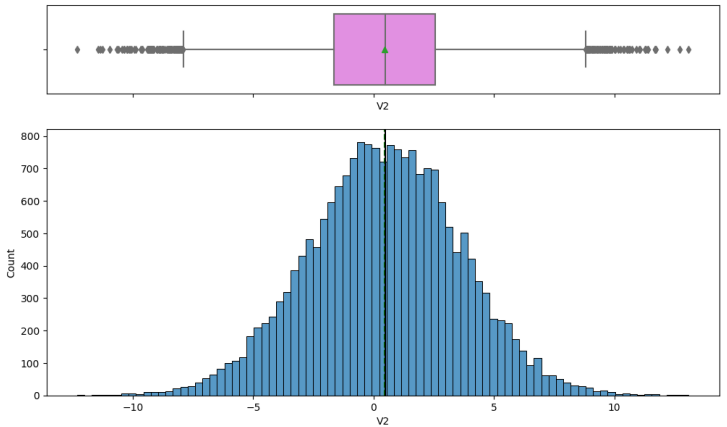
Variables

Exploritory Data Analysis

- V1 is right skewed
- Large numbers of outliers



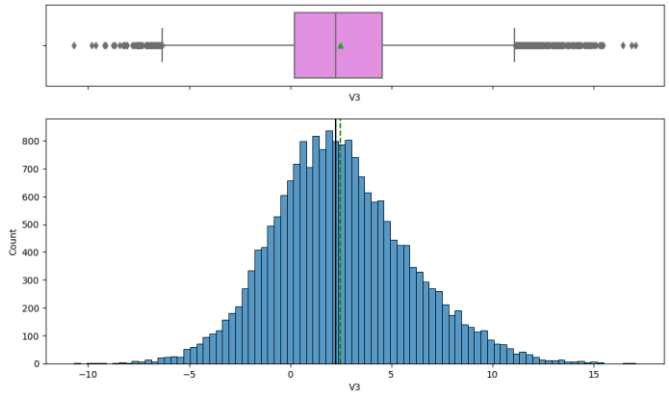
- The V2 is normally distributed
- Large number of outliers



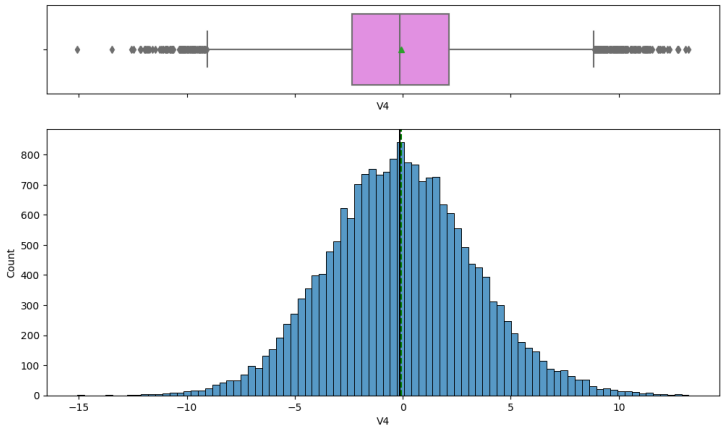
Variables

Exploritory Data Analysis

- V3 is slightly right skewed
- Large numbers of outliers



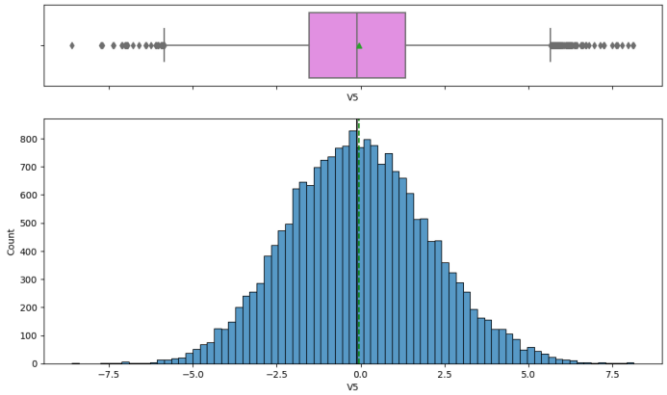
- The V4 is normally distributed
- Large number of outliers



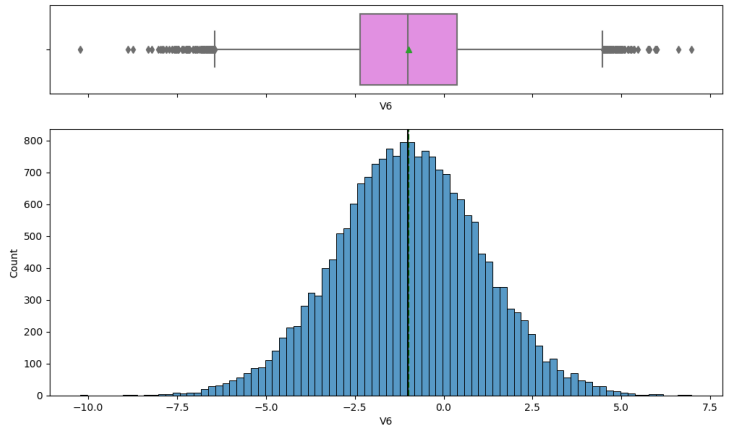
Variables

Exploritory Data Analysis

- V5 is normally distributed
- Large numbers of outliers



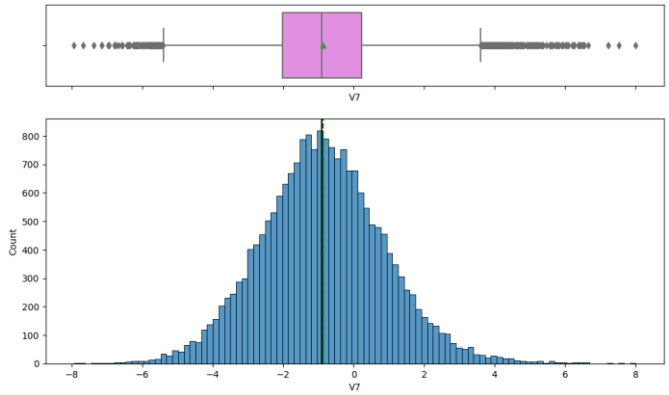
- The V6 is normally distributed
- Large number of outliers



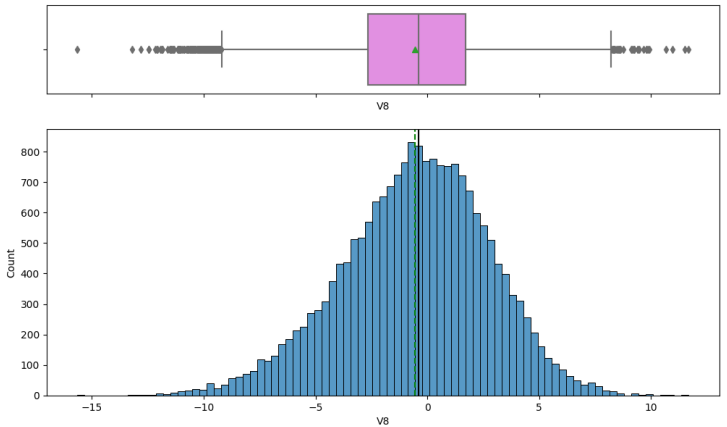
Variables

Exploritory Data Analysis

- V7 is normally distributed
- Large numbers of outliers



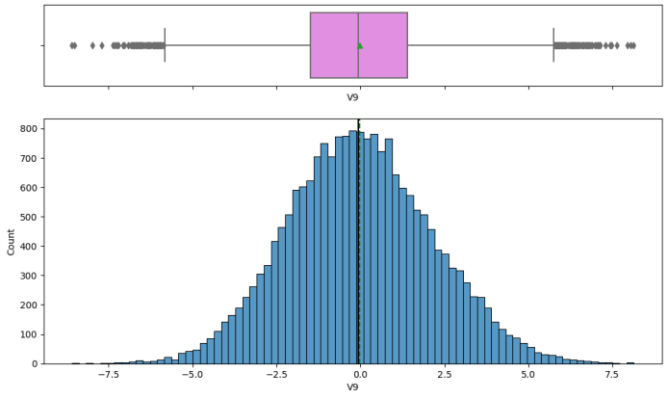
- The V8 is slightly left skewed
- Large number of outliers



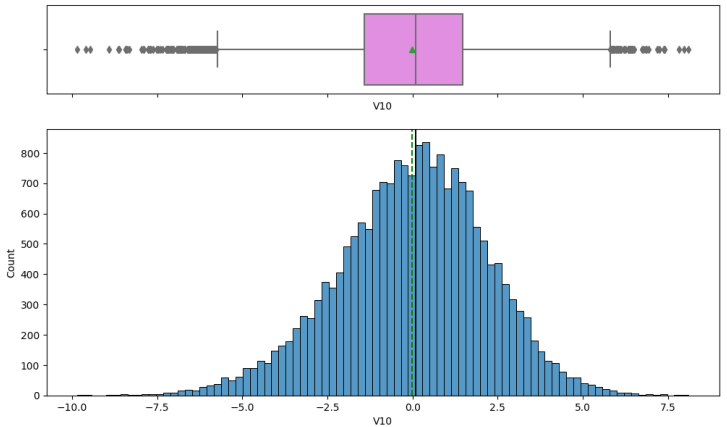
Variables

Exploritory Data Analysis

- V9 is normally distributed
- Large numbers of outliers



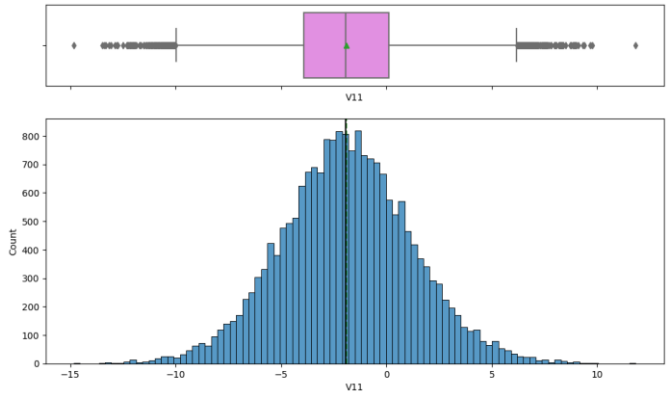
- The V10 is slightly left skewed
- Large number of outliers



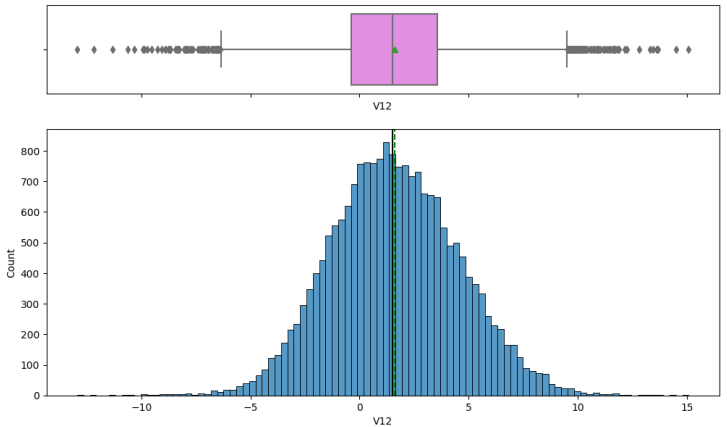
Variables

Exploritory Data Analysis

- V11 is normally distributed
- Large numbers of outliers



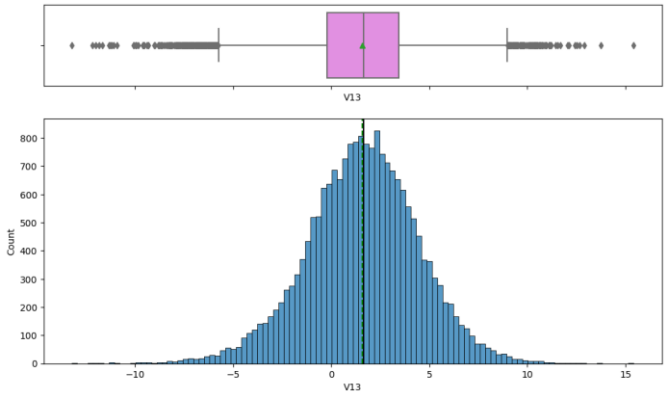
- The V12 is normally distributed
- Large number of outliers



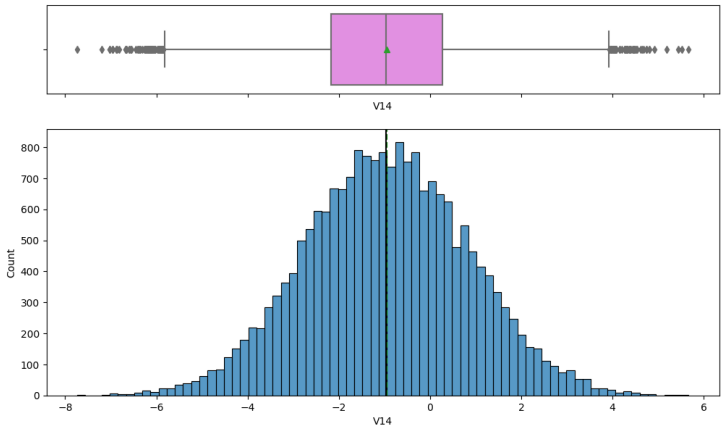
Variables

Exploritory Data Analysis

- V13 is normally distributed
- Large numbers of outliers



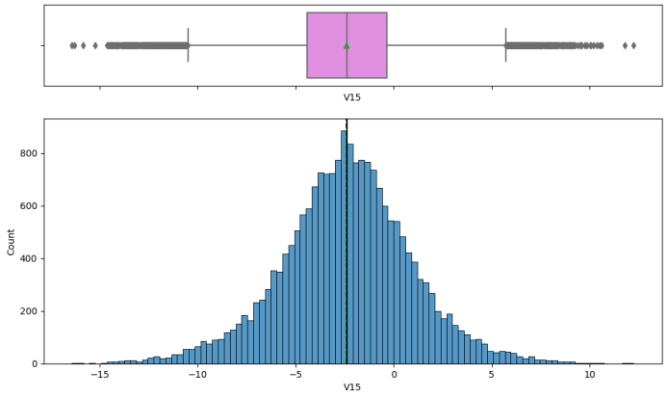
- The V14 is normally distributed
- Large number of outliers



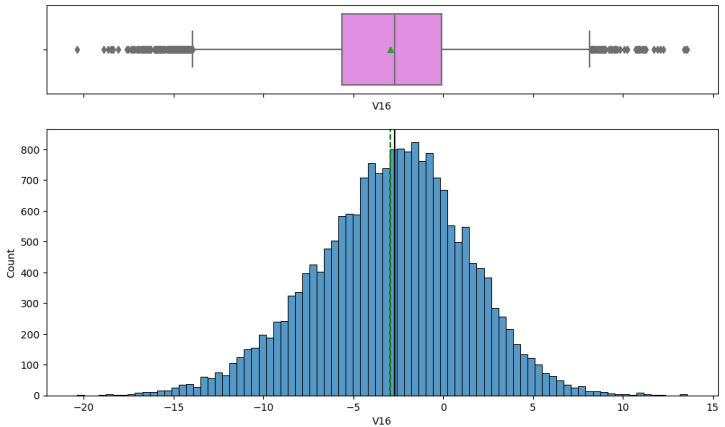
Variables

Exploritory Data Analysis

- V15 is normally distributed
- Large numbers of outliers



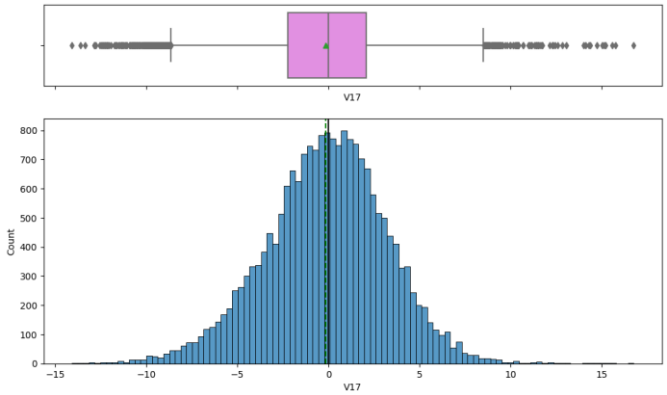
- The V16 is slightly left skewed
- Large number of outliers



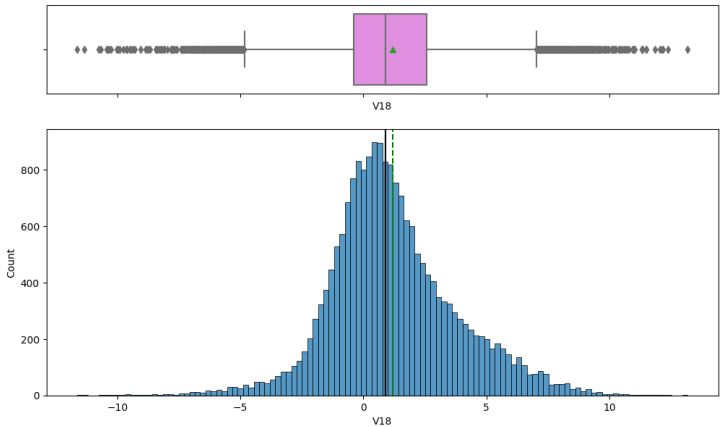
Variables

Exploritory Data Analysis

- V17 is normally distributed
- Large numbers of outliers



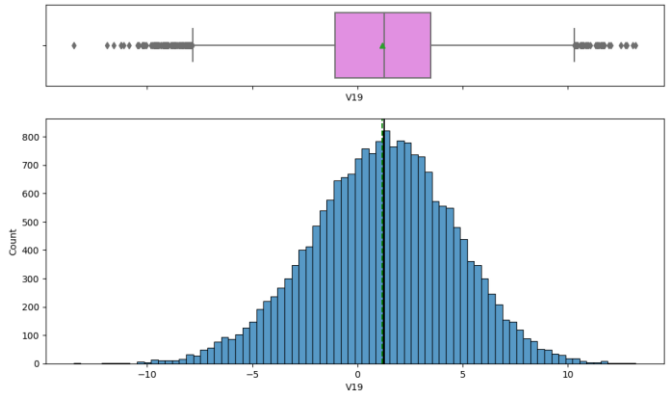
- The V18 is slightly right skewed
- Large number of outliers



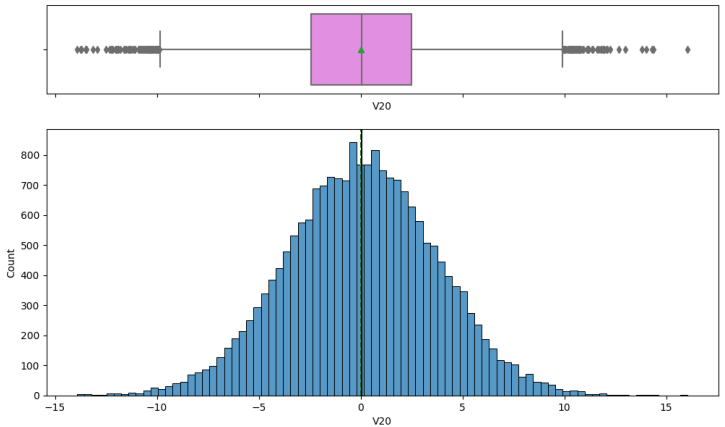
Variables

Exploritory Data Analysis

- V19 is normally distributed
- Large numbers of outliers



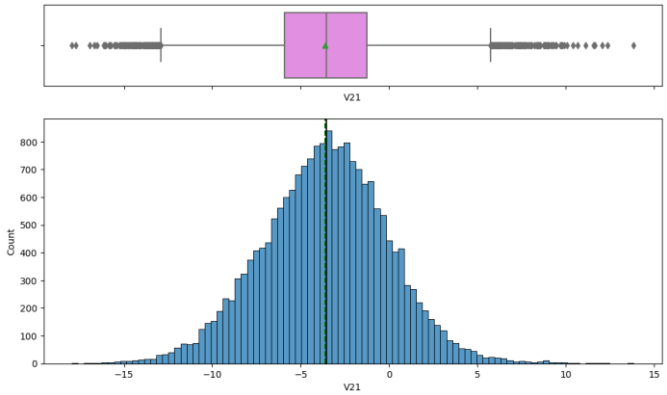
- The V20 is normally distributed
- Large number of outliers



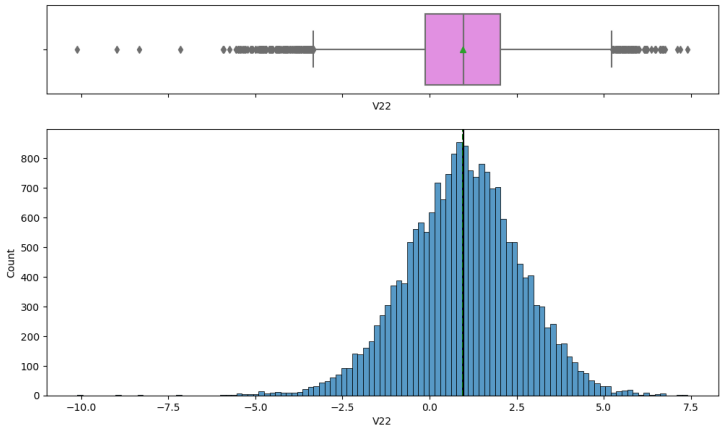
Variables

Exploritory Data Analysis

- V21 is normally distributed
- Large numbers of outliers



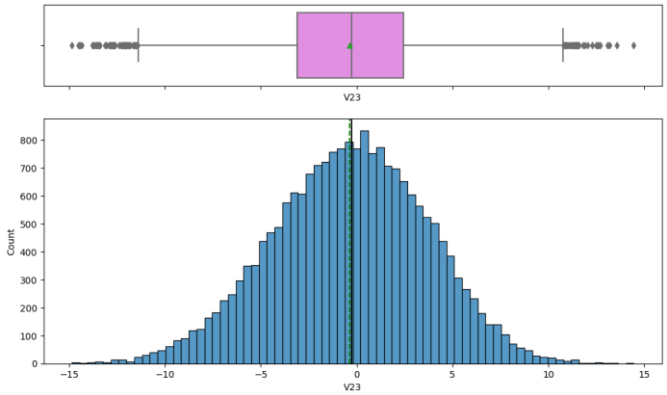
- The V22 is normally distributed
- Large number of outliers



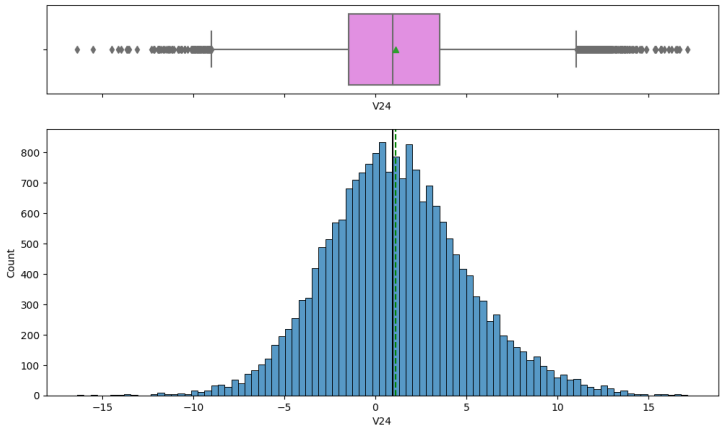
Variables

Exploritory Data Analysis

- V23 is normally distributed
- Large numbers of outliers



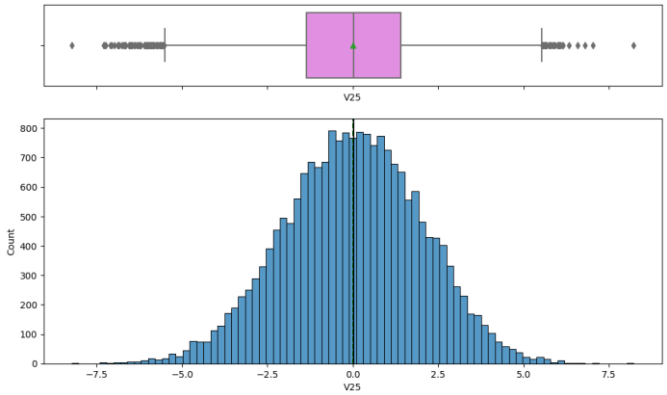
- The V24 is right skewed
- Large number of outliers



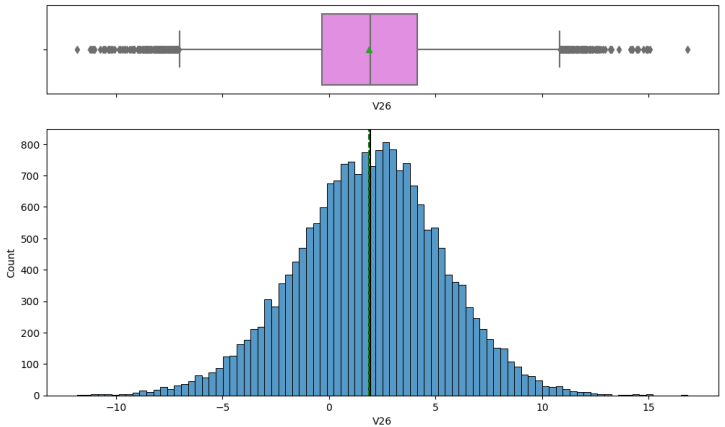
Variables

Exploritory Data Analysis

- V25 is normally distributed
- Large numbers of outliers



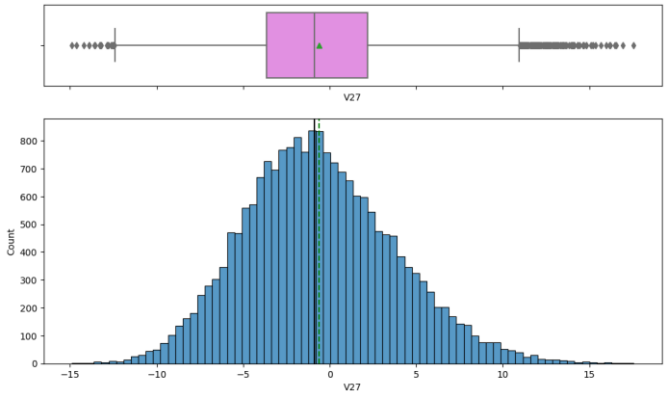
- The V26 is right skewed
- Large number of outliers



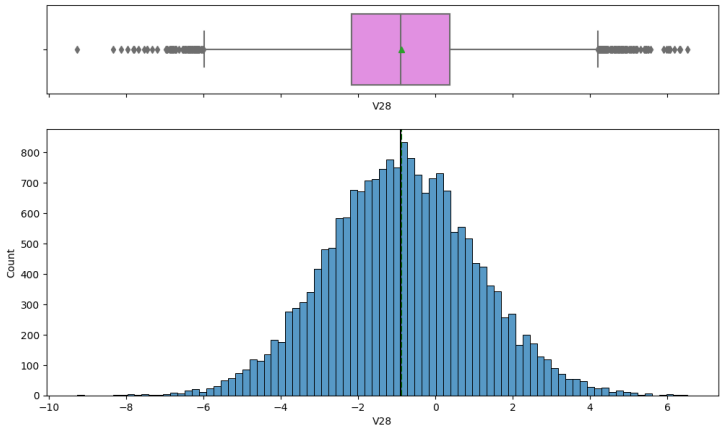
Variables

Exploritory Data Analysis

- V27 is right skewed
- Large numbers of outliers



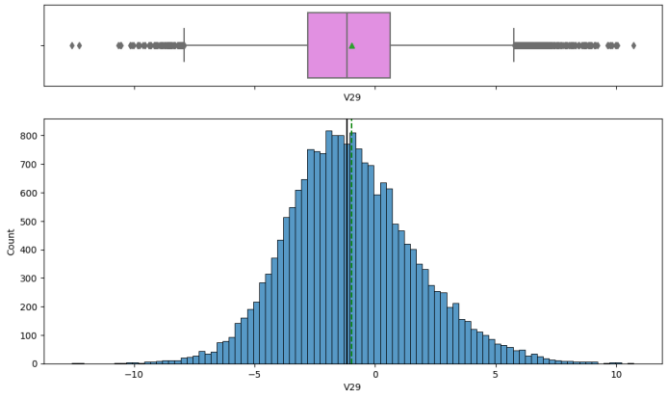
- The V28 is normally distributed
- Large number of outliers



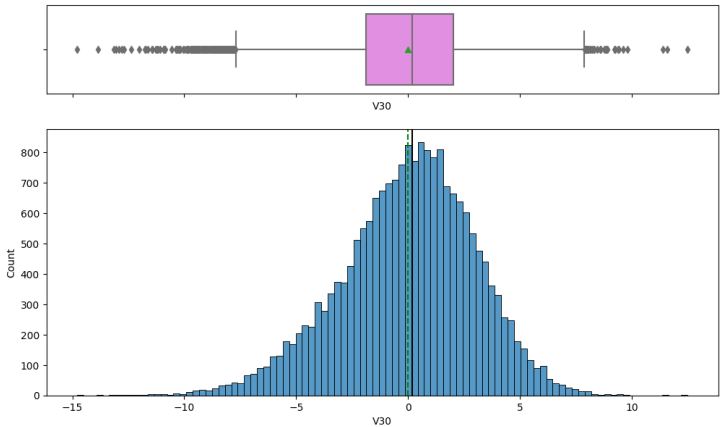
Variables

Exploritory Data Analysis

- V29 is right skewed
- Large numbers of outliers



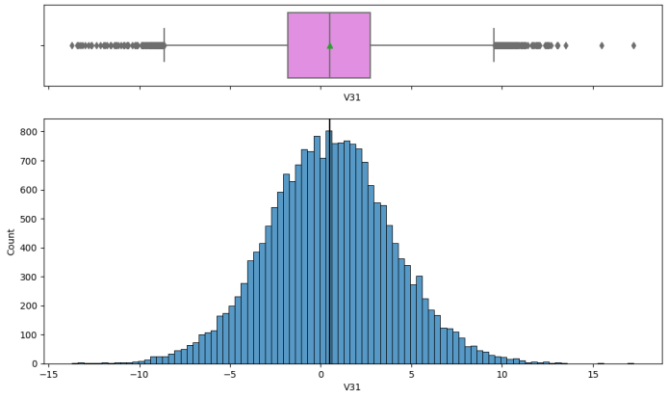
- The V30 is left skewed
- Large number of outliers



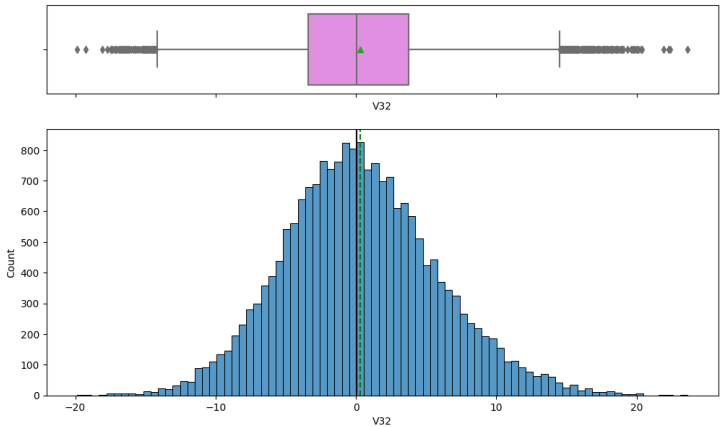
Variables

Exploritory Data Analysis

- V31 is normally distributed
- Large numbers of outliers



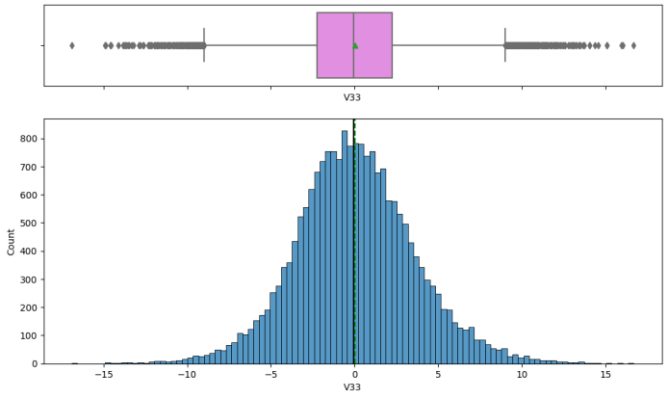
- The V32 is normally distributed
- Large number of outliers



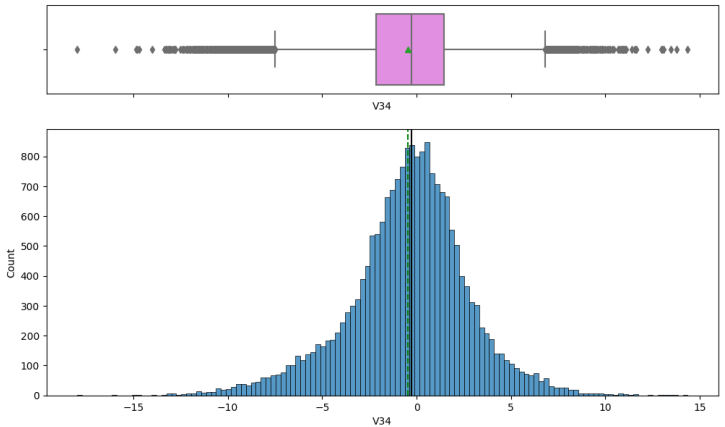
Variables

Exploritory Data Analysis

- V33 is normally distributed
- Large numbers of outliers



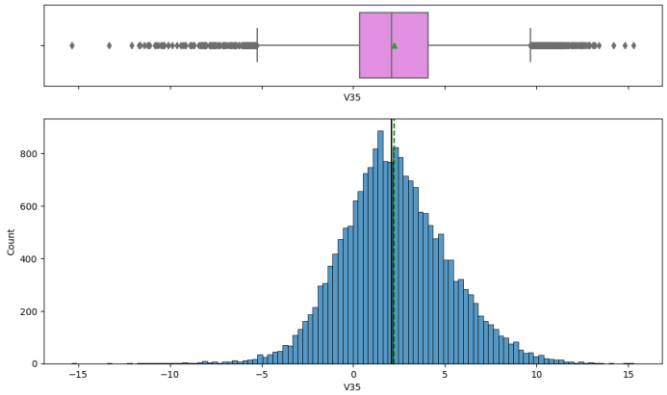
- The V34 is normally distributed
- Large number of outliers



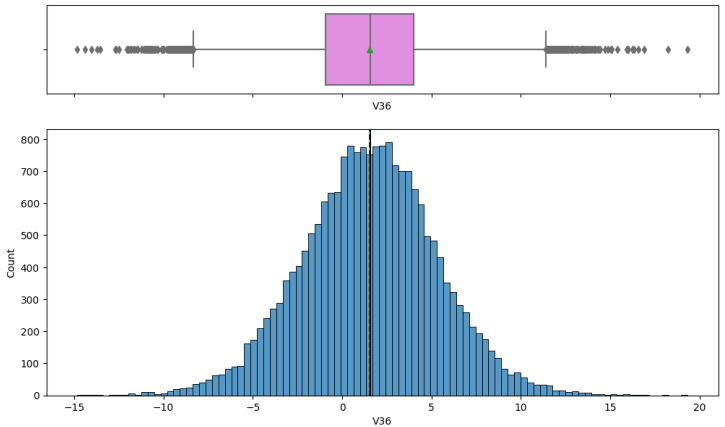
Variables

Exploritory Data Analysis

- V35 is normally distributed
- Large numbers of outliers



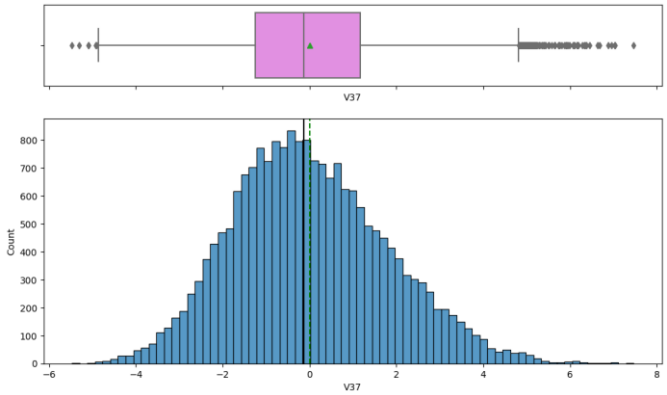
- The V36 is normally distributed
- Large number of outliers



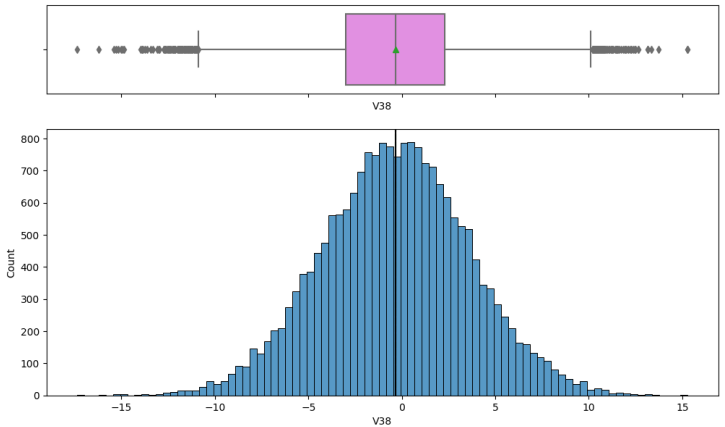
Variables

Exploritory Data Analysis

- V37 is right skewed
- Large numbers of outliers



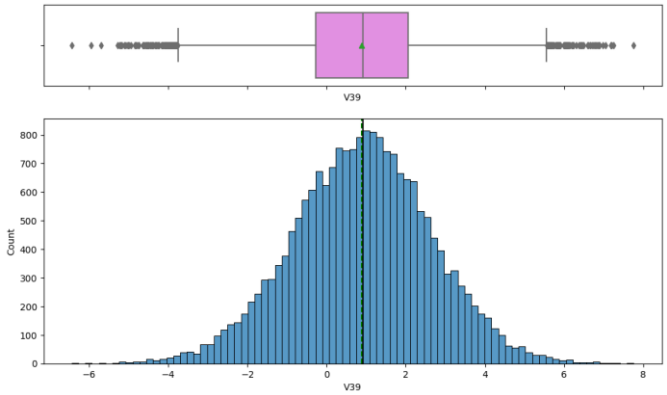
- The V38 is normally distributed
- Large number of outliers



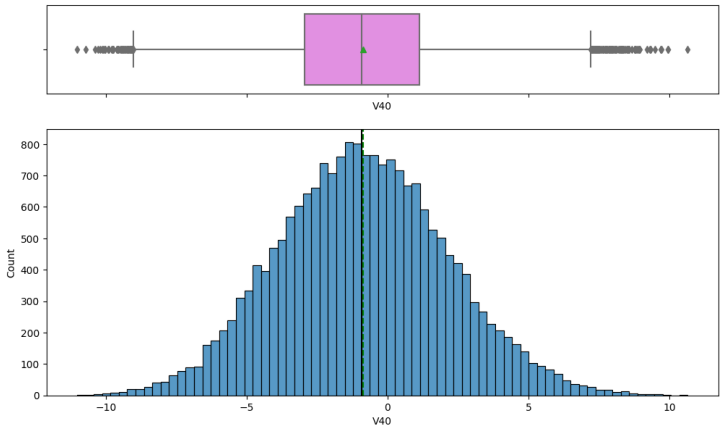
Variables

Exploritory Data Analysis

- V39 is right skewed
- Large numbers of outliers



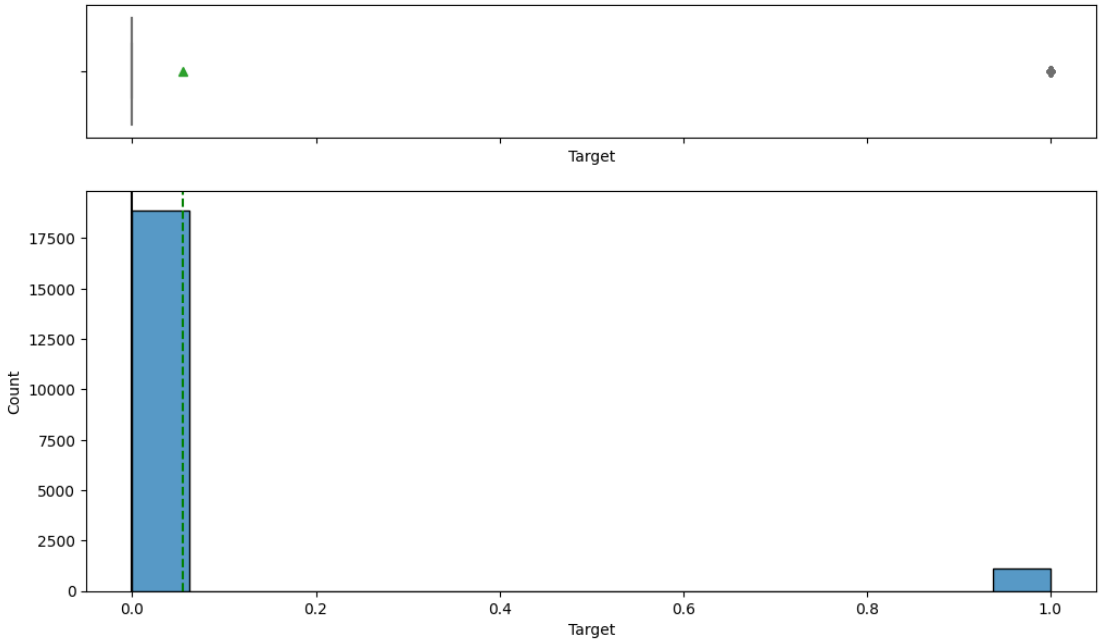
- The V40 is normally distributed
- Large number of outliers



Variables

Exploritory Data Analysis

- Target is majority no failure
- On the training data
  - 18,890 no failure
  - 1,110 failures
- On the test data
  - 4,718 no failure
  - 282 failures



Feature Engineering

Data Preprocessing

- No duplicate value
- Split train data set into train and test
- Split new training data in to training and validation into 75:25 ratio
    - 14,000 Rows – Train
    - 40 Columns – Train
    - 6,000 Rows – Validation
    - 40 Rows - Validation
- Dropped Target variable from test data
    - 5,000 Rows
    - 40 Columns
- Large amount of outliers,
    - No treatment

Missing Value Treatment

Data Preprocessing

- Created an instance of the imputer to be used
- Fit and transformed the train, validation and test data
- There were 36 missing values

Model Evaluation Criterion

Model Building

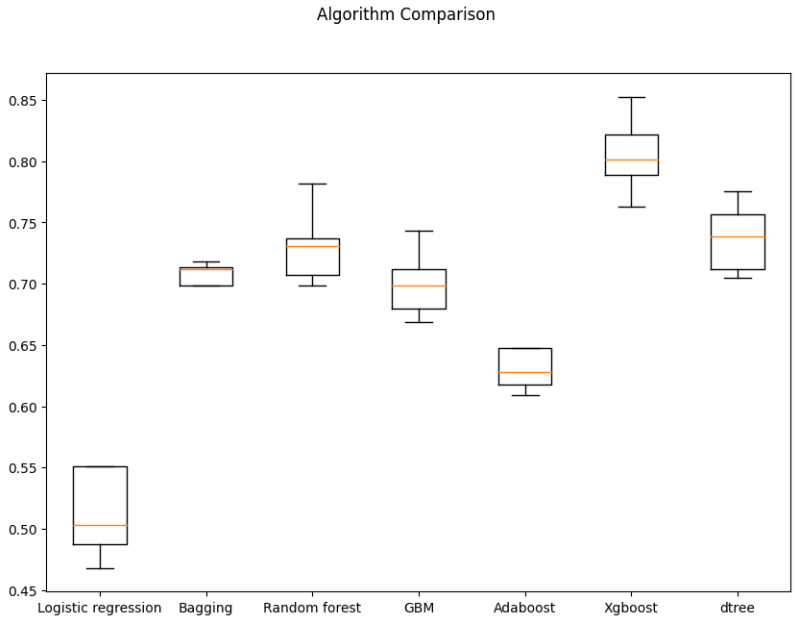
- True positives (TP) are failures correctly predicted by the model.
- False negatives (FN) are real failures in a generator where there is no detection by model.
- False positives (FP) are failure detections in a generator where there is no failure.
- Created a function to compute different metrics to check performance of a classification model using sklearn
- Will try and maximize Recall
- Used Recall as a scorer in cross-validation and hyperparameter tuning

Original Data

Model Building

	Logistic Regression	Bagging	Random Forest	GBM	Ada Boost	Xg Boost	DTree
CV Training	0.512	0.708	0.731	0.700	0.631	0.805	0.738
CV Validation	0.447	0.705	0.708	0.684	0.581	0.805	0.717

CV Training Set

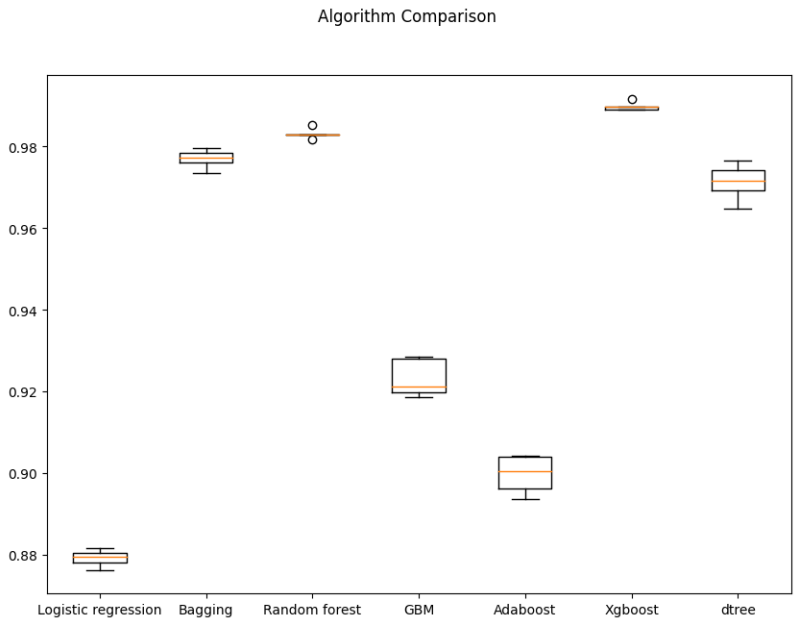


Oversampled Data

Model Building

- Before Oversampling non failures: 13,129
- Before Oversampling failures: 781
- After Oversampling non-failures: 13,219
- After Oversampling failures: 13,219
- After Oversampling shape of train_x: (26438,40)
- After Oversampling shape of train_y: (26,438,)

	Logistic Regression	Bagging	Random Forest	GBM	Ada Boost	Xg Boost	DTree
CV Training	0.879	0.977	0.983	0.923	0.899	0.989	0.971
CV Validation	0.836	0.812	0.839	0.844	0.854	0.857	0.784



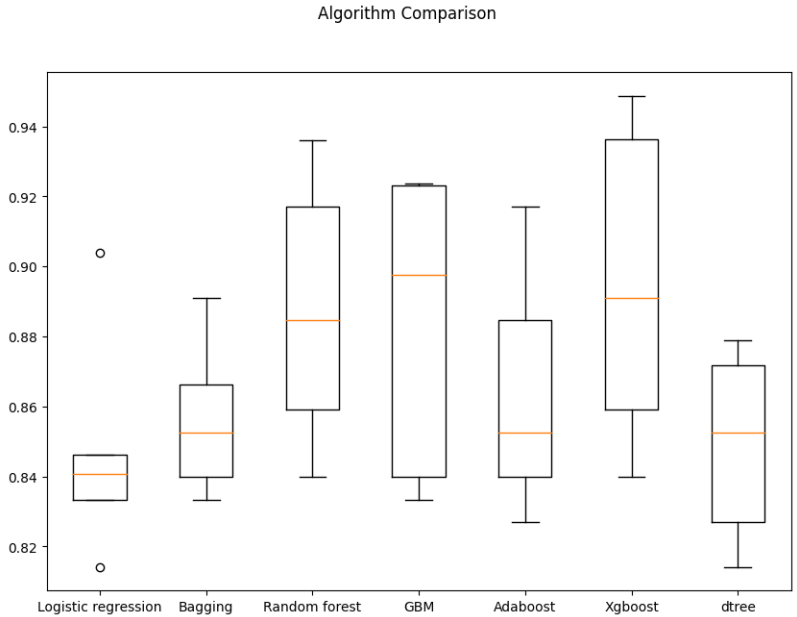
Undersampled Data

Model Building

- Before Under sampling non failures: 13,129
- Before Under sampling failures: 781
- After Under sampling non-failures: 781
- After Under sampling failures: 781
- After Under sampling shape of train_x: (1,562,40)
- After Under sampling shape of train_y: (1,562,)

	Logistic Regression	Bagging	Random Forest	GBM	Ada Boost	Xg Boost	DTree
CV Training	0.848	0.857	0.887	0.883	0.864	0.895	0.849
CV Validation	0.836	0.848	0.881	0.891	0.854	0.891	0.839

CV Training Set



AdaBoost using Oversampled Data

Hyperparameter Tuning

There is overfitting in recall, precision and F1

Training Set

Accuracy	Recall	Precision	F1
0.993	0.990	0.996	0.993

Base Estimator: Decision Tree Classifier

N Estimators	Learning Rate	Base Estimator: Dtree	CV Score
200	0.2	Max Depth: 3	0.974

Validation Set

Accuracy	Recall	Precision	F1
0.984	0.872	0.837	0.854

Gradient Boosting using Undersampled Data

Hyperparameter Tuning

There is overfitting in recall, precision and F1

N Estimators	Subsample	Learning Rate	CV Score	Max Features
125	0.7	1	0.970	0.5

Training Set

Accuracy	Recall	Precision	F1
0.995	0.995	0.995	0.995

Validation Set

Accuracy	Recall	Precision	F1
0.968	0.827	0.663	0.736

XGBoost using Oversampled Data

Hyperparameter Tuning

There is overfitting in recall, precision and F1

N Estimators	Learning Rate	Gamma	CV Score	Sub Sample	Scale POS Weight
150	0.1	3	0.997	0.8	10

Training Set

Accuracy	Recall	Precision	F1
0.994	1.000	0.989	0.994

Validation Set

Accuracy	Recall	Precision	F1
0.969	0.888	0.659	0.756

Model Performance Comparison

Model Performance Summary

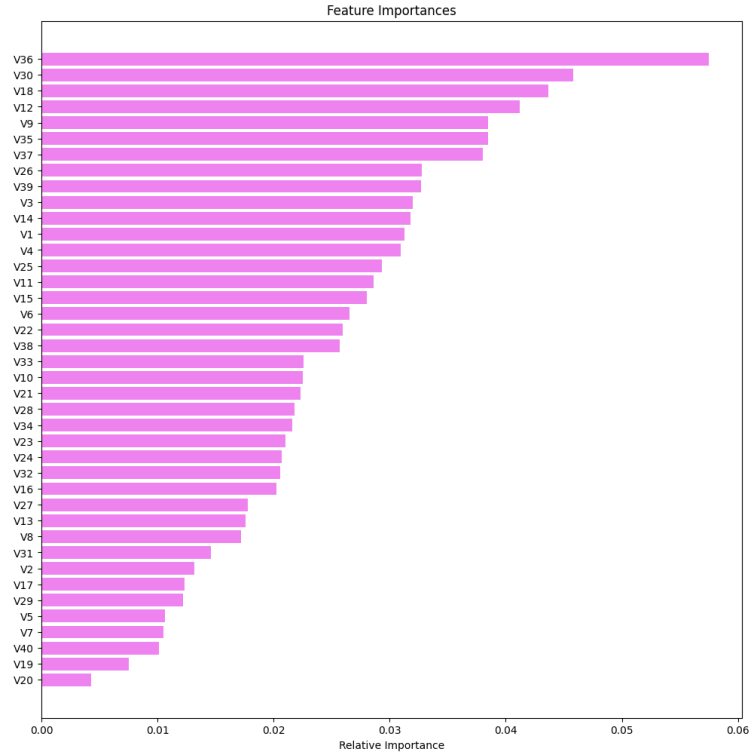
Training Set	Gradient Boosting	AdaBoost	Random Forest	XG Boost
Accuracy	0.995	0.993	0.960	0.994
Recall	0.995	0.990	0.928	1.000
Precision	0.995	0.996	0.990	0.989
F1	0.995	0.993	0.958	0.998
Validation Set	**Gradient Boosting**	**AdaBoost**	**Random Forest**	**XG Boost**
Accuracy	0.968	0.984	0.949	0.969
Recall	0.827	0.872	0.888	0.888
Precision	0.663	0.837	0.522	0.659
F1	0.736	0.854	0.658	0.756

Model Performance Comparison

Model Performance Summary

- Chose AdaBoost as the best fit model
- Did not overfit on the Test data
- Most important features
  - V36, V30, V18, V12, V9, V35, V37

AdaBoost	Test Data	Training Data
Accuracy	0.980	0.984
Recall	0.844	0.872
Precision	0.810	0.837
F1	0.826	0.854



Productionization of final model

Final Model Pipeline

- Created a pipeline for AdaBoost
- Dropped Target variable from the train and test sets
- Ran SMOTE
- Fit model to data
    - Adaboost classifier
    - Base Estimator: Decision Tree Classifier

AdaBoost SMOTE	Test Data
Accuracy	0.978
Recall	0.851
Precision	0.774
F1	0.811

Great Learning

Happy Learning !