MT: RENEWIND PROJECT

BY: DR. CYNTHIA OFFOHA

AGENDA

- EXECUTIVE SUMMARY
- BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH
- DATA OVERVIEW
- EDA – UNIVARIATE ANALYSIS & KEY QUESTIONS
- DATA PRE-PROCESSING
- MODEL BUILDING & HYPERPARAMETER TUNING
- MODEL PERFORMANCE AND FINAL PIPELINE
- RECOMMENDATIONS & CONCLUSIONS

EXECUTIVE SUMMARY

- Renewable energy sources play vital role in global energy mix with the aim to decrease environmental impact of energy production increases
- Among the energy renewable alternatives, wind energy is the most developed technology
    - This led to U.S dept of Energy organizing an efficient way for achieving operational efficiency using predictive maintenance practices
- Sensor information and analysis methods are utilized to measure and predict degradation and future component capability
- The main purpose for predictive maintenance is that failure patterns are predictable and if component failure are predicted correctly and component replaced before it fails then operation cost and maintenance will be much lower
- The sensors fitted across different machines involved in procession of energy generation collects its data via different environmental factors (eg. Temperature, wind speed, and humidity); and additional features related to different aspects of wind turbine (eg. Gearbox, tower, blades, break)

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

- Renewind is a company working on enhancing the machinery/processes involved in generation of wind energy with the aid of machine learning
    - The company utilizes the sensors in the collection of data (generator failure of wind turbines)
    - The data collected are kept confidential and the data consist of 40 predictors, 20000 observations in training and 5000 in the test set
- The aim is to develop various classification models, tune them, and find the best will assist in identifying failures to repair generators before failing/breaking to decrease total maintenance cost
- The nature of predictions made by classification model will translate as follows:
    - True positives (TP) are failures correctly predicted by the model. These will result in repair 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
- Thus, it is given that the cost of repairing a generator is much less than the cost of replacing it; and cost of inspection is less than cost of repair
- However, the purpose is to develop various classification models and identify the best model that will assist in specifying failures that generators could be repaired before failing/breaking to decrease the overall maintenance cost

DATA OVERVIEW

- The data provided --> transformed version of original data collected using sensors
- **Train.csv** - To be used for training and tuning of models
- **Test.csv** - To be used only for testing the performance of the final best model
- Both the datasets consist of 40 predictor variables and 1 target variable
- The target variable consist of:
  - "1" as "failure"
  - "0" represents "No failure".

DATA STRUCTURE

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
0	-4.465	-4.679	3.102	0.506	-0.221	-2.033	-2.911	0.051	-1.522	3.762	-5.715	0.736	0.981	1.418	-3.376	-3.047	0.306
1	3.366	3.653	0.910	-1.368	0.332	2.359	0.733	-4.332	0.566	-0.101	1.914	-0.951	-1.255	-2.707	0.193	-4.769	-2.205
2	-3.832	-5.824	0.634	-2.419	-1.774	1.017	-2.099	-3.173	-2.082	5.393	-0.771	1.107	1.144	0.943	-3.164	-4.248	-4.039
3	1.618	1.888	7.046	-1.147	0.083	-1.530	0.207	-2.494	0.345	2.119	-3.053	0.460	2.705	-0.636	-0.454	-3.174	-3.404
4	-0.111	3.872	-3.758	-2.983	3.793	0.545	0.205	4.849	-1.855	-6.220	1.998	4.724	0.709	-1.989	-2.633	4.184	2.245

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
19995	-2.071	-1.088	-0.796	-3.012	-2.288	2.807	0.481	0.105	-0.587	-2.899	8.868	1.717	1.358	-1.777	0.710	4.945	-3.100
19996	2.890	2.483	5.644	0.937	-1.381	0.412	-1.593	-5.762	2.150	0.272	-2.095	-1.526	0.072	-3.540	-2.762	-10.632	-0.495
19997	-3.897	-3.942	-0.351	-2.417	1.108	-1.528	-3.520	2.055	-0.234	-0.358	-3.782	2.180	6.112	1.985	-8.330	-1.639	-0.915
19998	-3.187	-10.052	5.696	-4.370	-5.355	-1.873	-3.947	0.679	-2.389	5.457	1.583	3.571	9.227	2.554	-7.039	-0.994	-9.665
19999	-2.687	1.961	6.137	2.600	2.657	-4.291	-2.344	0.974	-1.027	0.497	-9.589	3.177	1.055	-1.416	-4.669	-5.405	3.720

TRAINING DATA SETS – 20000 ENTRIES

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
0	-0.613	-3.820	2.202	1.300	-1.185	-4.496	-1.836	4.723	1.206	-0.342	-5.123	1.017	4.819	3.269	-2.984	1.387	2.032
1	0.390	-0.512	0.527	-2.577	-1.017	2.235	-0.441	-4.406	-0.333	1.967	1.797	0.410	0.638	-1.390	-1.883	-5.018	-3.827
2	-0.875	-0.641	4.084	-1.590	0.526	-1.958	-0.695	1.347	-1.732	0.466	-4.928	3.565	-0.449	-0.656	-0.167	-1.630	2.292
3	0.238	1.459	4.015	2.534	1.197	-3.117	-0.924	0.269	1.322	0.702	-5.578	-0.851	2.591	0.767	-2.391	-2.342	0.572
4	5.828	2.768	-1.235	2.809	-1.642	-1.407	0.569	0.965	1.918	-2.775	-0.530	1.375	-0.651	-1.679	-0.379	-4.443	3.894

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
4995	-5.120	1.635	1.251	4.036	3.291	-2.932	-1.329	1.754	-2.985	1.249	-6.878	3.715	-2.512	-1.395	-2.554	-2.197	4.772
4996	-5.172	1.172	1.579	1.220	2.530	-0.669	-2.618	-2.001	0.634	-0.579	-3.671	0.460	3.321	-1.075	-7.113	-4.356	-0.001
4997	-1.114	-0.404	-1.765	-5.879	3.572	3.711	-2.483	-0.308	-0.922	-2.999	-0.112	-1.977	-1.623	-0.945	-2.735	-0.813	0.610
4998	-1.703	0.615	6.221	-0.104	0.956	-3.279	-1.634	-0.104	1.388	-1.066	-7.970	2.262	3.134	-0.486	-3.498	-4.562	3.136
4999	-0.604	0.960	-0.721	8.230	-1.816	-2.276	-2.575	-1.041	4.130	-2.731	-3.292	-1.674	0.465	-1.646	-5.263	-7.988	6.480

TEST DATA SETS – 5000 ENTRIES

	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
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

TRAINING DATA – STATISTICAL INFO

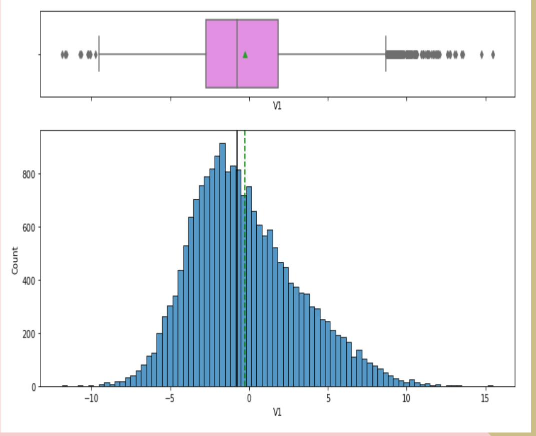
	count	mean	std	min	25%	50%	75%	max
V1	4995.000	-0.278	3.466	-12.382	-2.744	-0.765	1.831	13.504
V2	4994.000	0.398	3.140	-10.716	-1.649	0.427	2.444	14.079
V3	5000.000	2.552	3.327	-9.238	0.315	2.260	4.587	15.315
V4	5000.000	-0.049	3.414	-14.682	-2.293	-0.146	2.166	12.140
V5	5000.000	-0.080	2.111	-7.712	-1.615	-0.132	1.341	7.673
V6	5000.000	-1.042	2.005	-8.924	-2.369	-1.049	0.308	5.068
V7	5000.000	-0.908	1.769	-8.124	-2.054	-0.940	0.212	7.616
V8	5000.000	-0.575	3.332	-12.253	-2.642	-0.358	1.713	10.415
V9	5000.000	0.030	2.174	-6.785	-1.456	-0.080	1.450	8.851
V10	5000.000	0.019	2.145	-8.171	-1.353	0.166	1.511	6.599
V30	5000.000	-0.119	3.023	-12.438	-1.997	0.112	1.946	10.315
V31	5000.000	0.469	3.446	-11.263	-1.822	0.486	2.779	12.559
V32	5000.000	0.233	5.586	-17.244	-3.556	-0.077	3.752	26.539
V33	5000.000	-0.080	3.539	-14.904	-2.348	-0.160	2.099	13.324
V34	5000.000	-0.393	3.166	-14.700	-2.010	-0.172	1.465	12.146
V35	5000.000	2.211	2.948	-12.261	0.322	2.112	4.032	13.489
V36	5000.000	1.595	3.775	-12.736	-0.866	1.703	4.104	17.116
V37	5000.000	0.023	1.785	-5.079	-1.241	-0.110	1.238	6.810
V38	5000.000	-0.406	3.969	-15.335	-2.984	-0.381	2.288	13.065
V39	5000.000	0.939	1.717	-5.451	-0.208	0.959	2.131	7.182
V40	5000.000	-0.932	2.978	-10.076	-2.987	-1.003	1.080	8.698
Target	5000.000	0.056	0.231	0.000	0.000	0.000	0.000	1.000

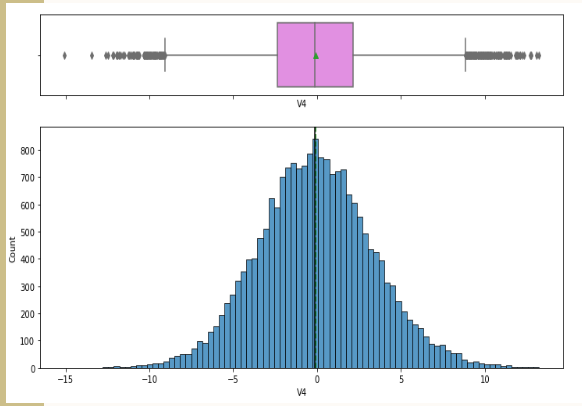
TESTING DATA – STATISTICAL INFO

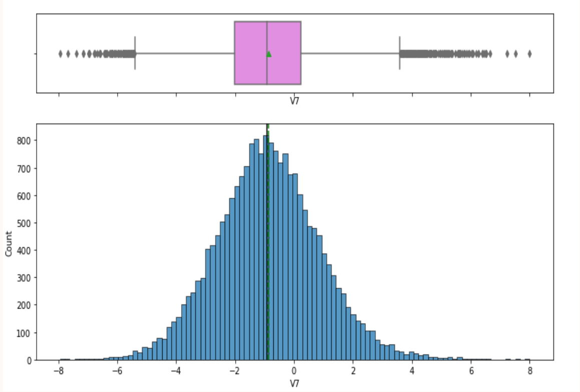
OBSERVATIONS

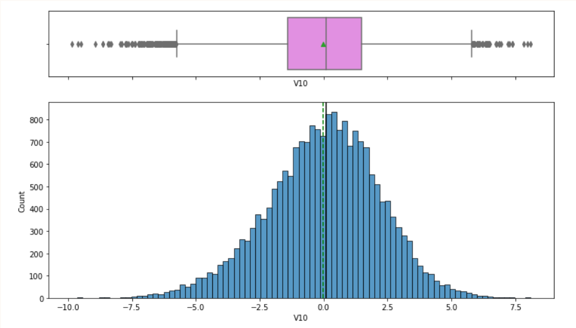
- The data consist of two data sets – Training and Test data sets
- Training data set consist of 20000 rows (entries) and 4 columns; while Test dataset consist of 5000 rows (entries) and 41 columns
- Both training and test data sets consist of variables labelled V1 to V40 with a target variable
    - 40 predictor variable and 1 target variable
- Few negatives were observed in the variables
- Variables seen with the training and test data sets are all floats except the target variable which is an integer
- Training and test data sets consist of means and medians that are very close to each other --> Normal distribution
    - The mean, std, and percentiles seem usual
- Some negative variables were observed but will not be treated since data originate from wind sensor
- No duplications were observed in both training and testing data sets
- 18 null values observed in both V1 and V2 variables within the training data whereas 5 null values in V1 column, and 6 null values in V2 variable within testing data

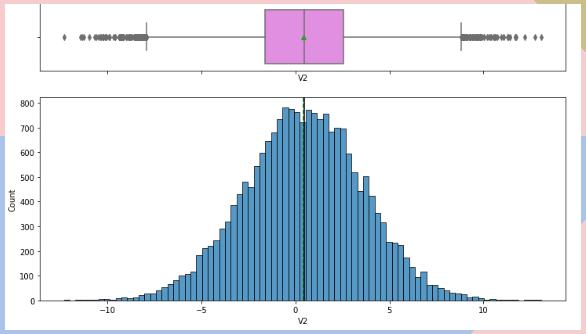
EDA – UNIVARIATE ANALYSIS

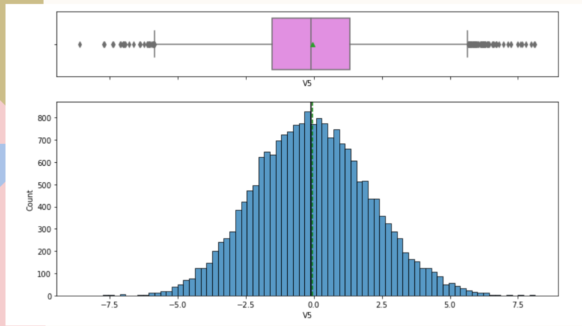


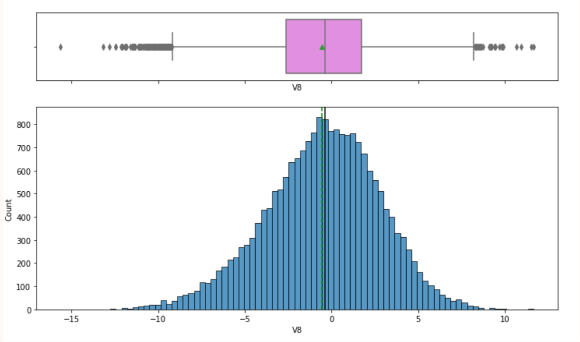


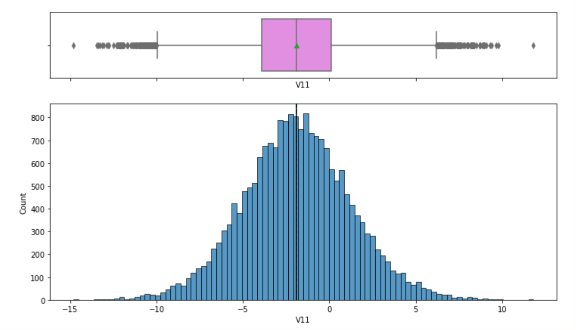


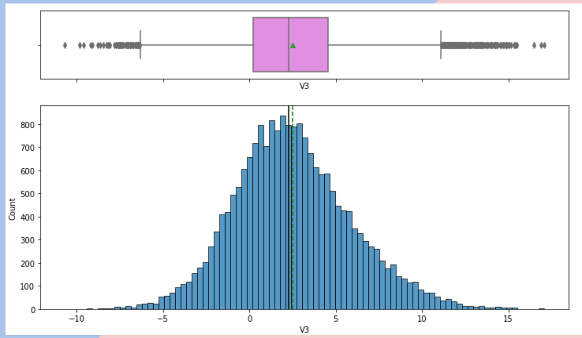


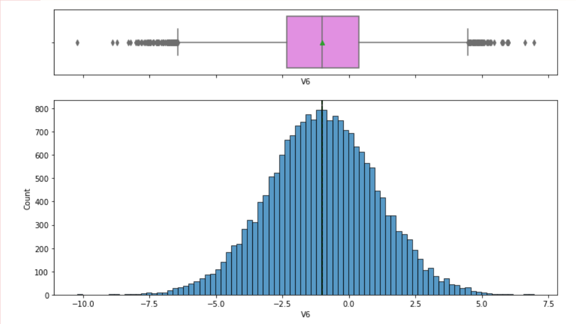


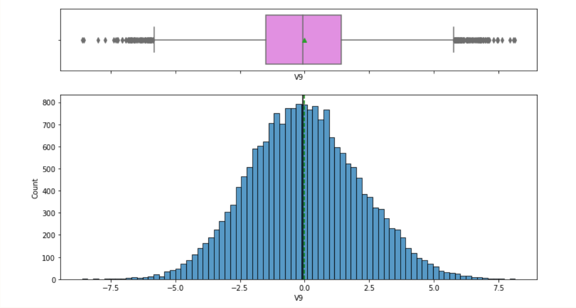


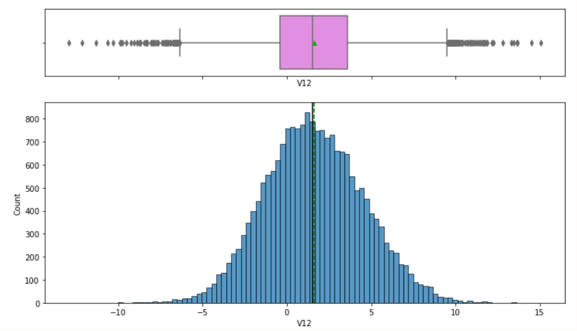


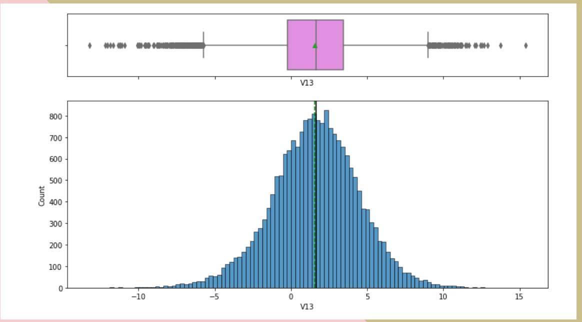


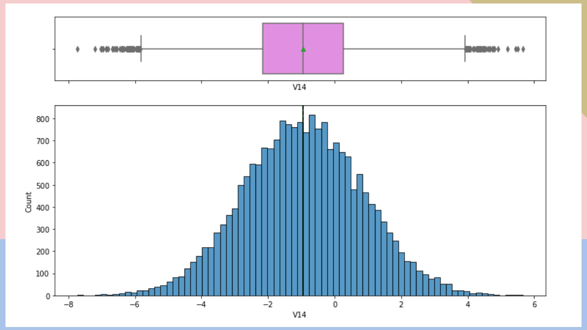


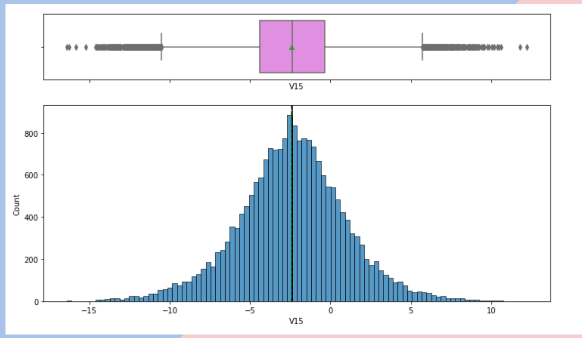


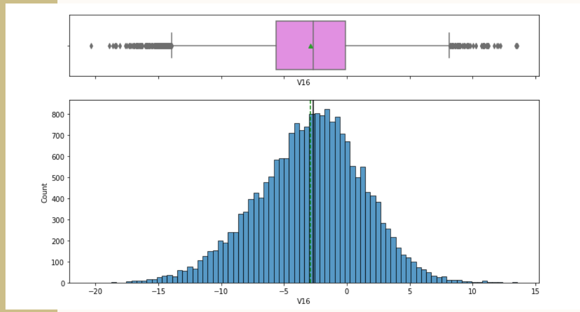


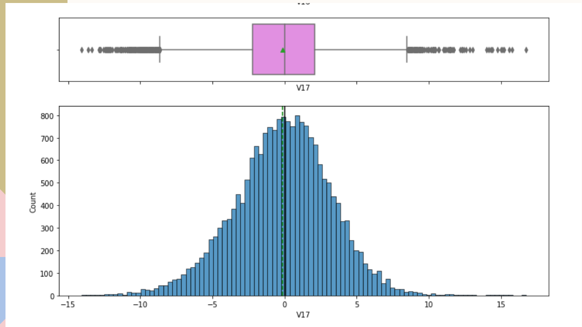


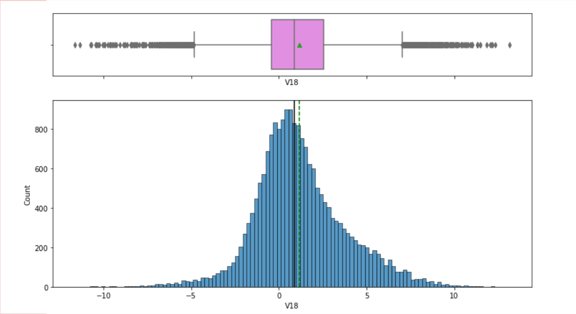

V13
Count
V13

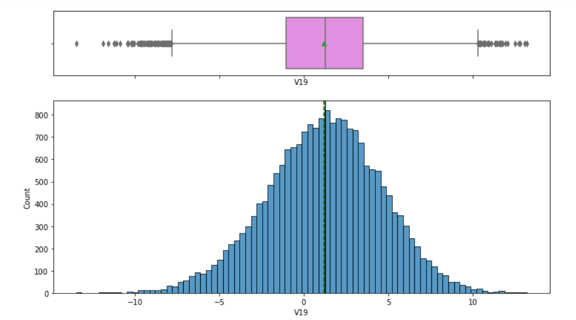

V14
Count
V14

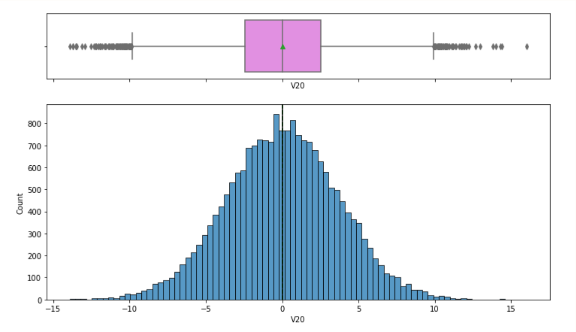

V15
Count
V15

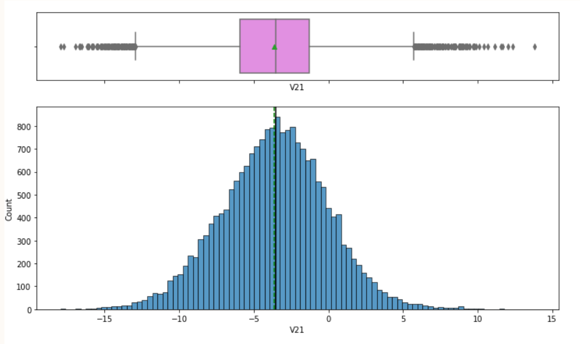

V16
Count
V16

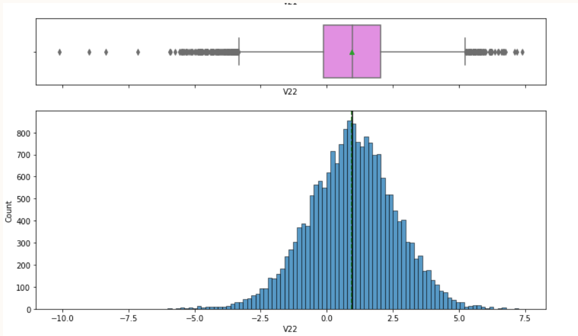

V17
Count
V17

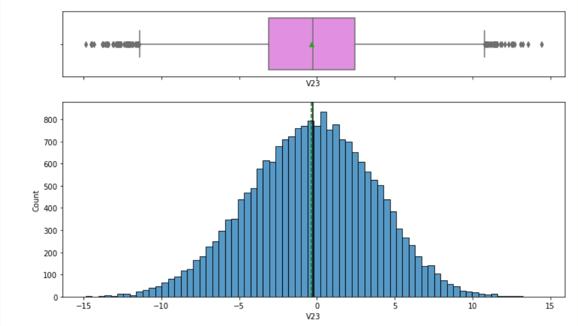

V18
Count
V18

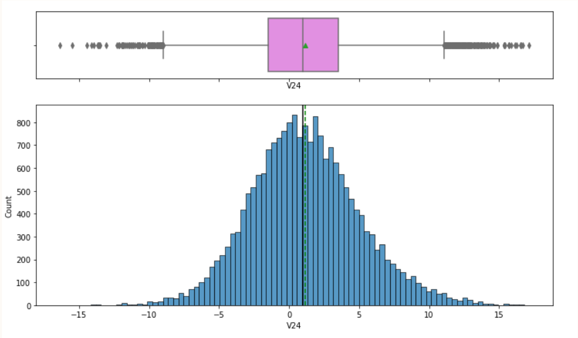

V19
Count
V19

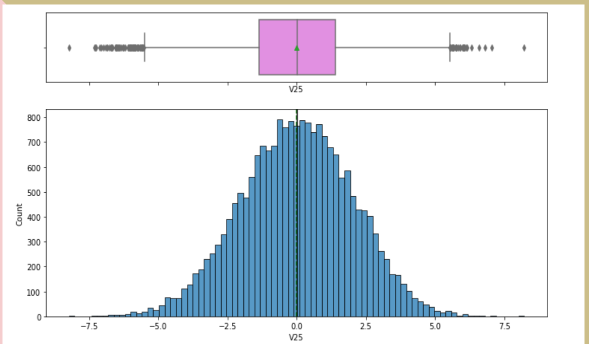

V20
Count
V20

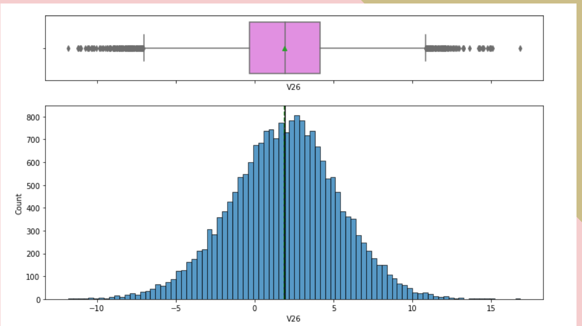

V21
Count
V21

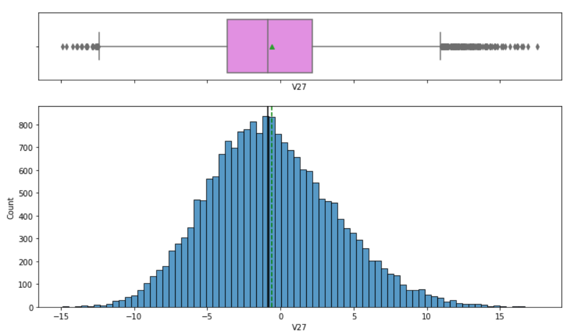

V22
Count
V22

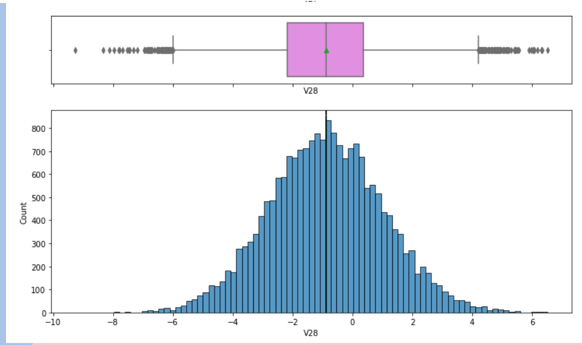

V23
Count
V23

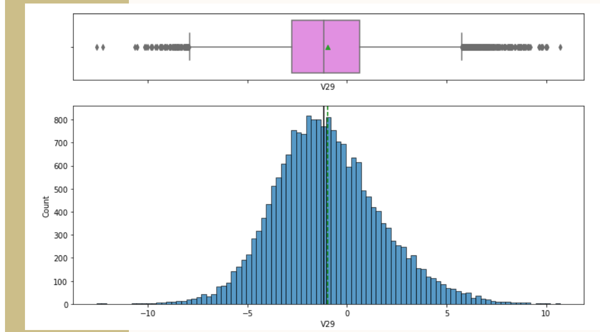

V24
Count
V24

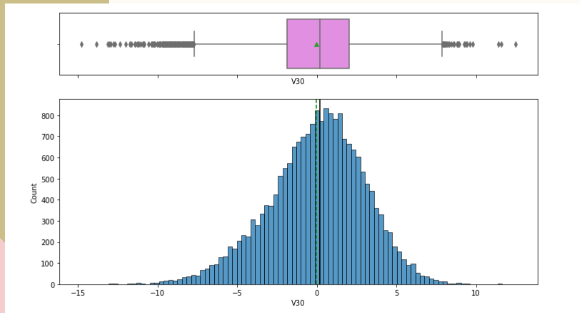

V25
Count

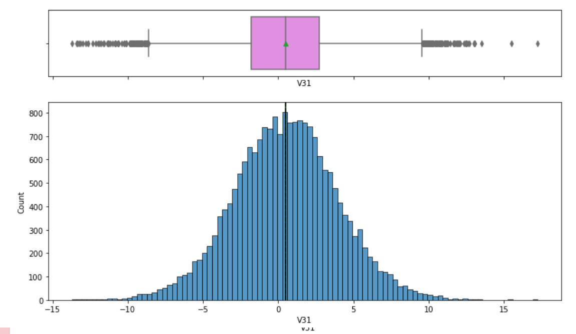

V26
Count

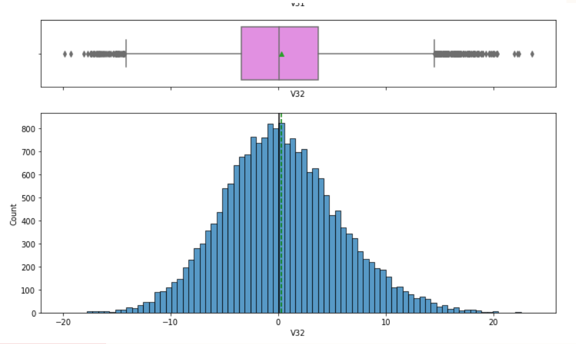

V27
Count

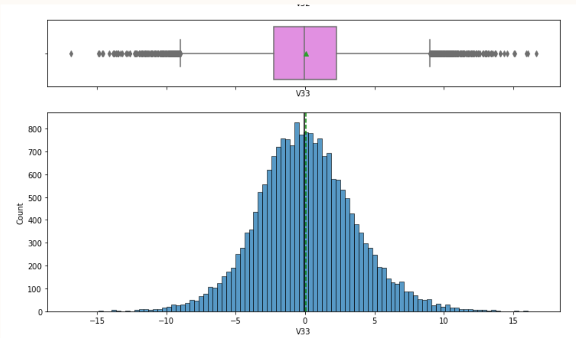

V28
Count

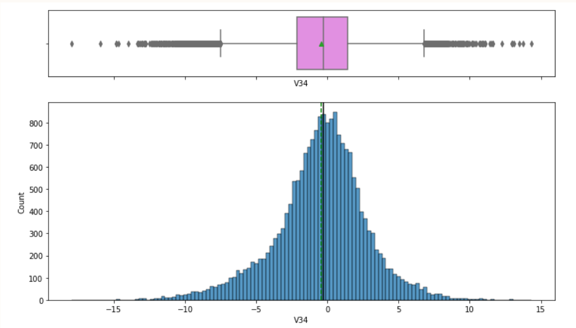

V29
Count

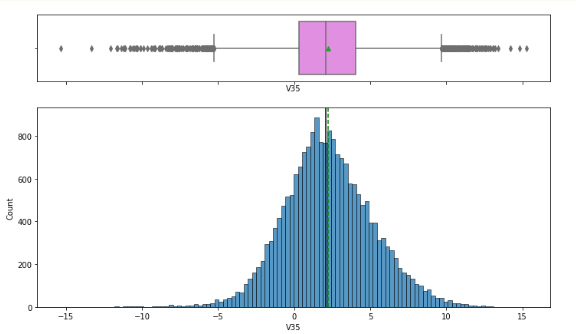

V30
Count

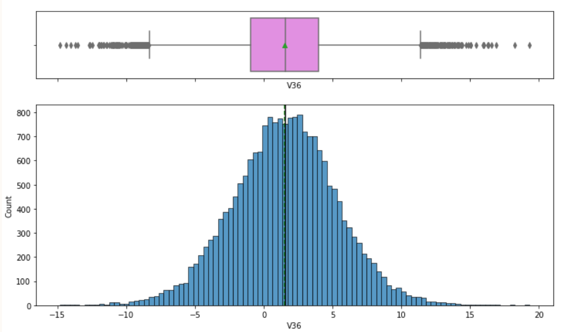

V31
Count

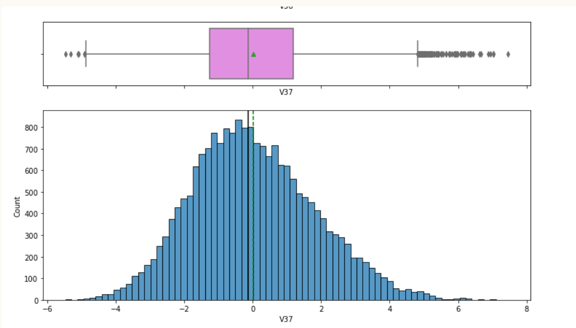

V32
Count

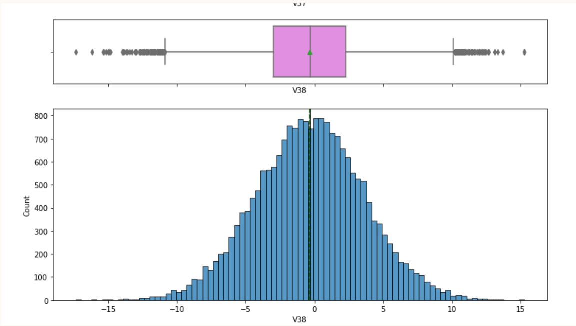

V33
Count

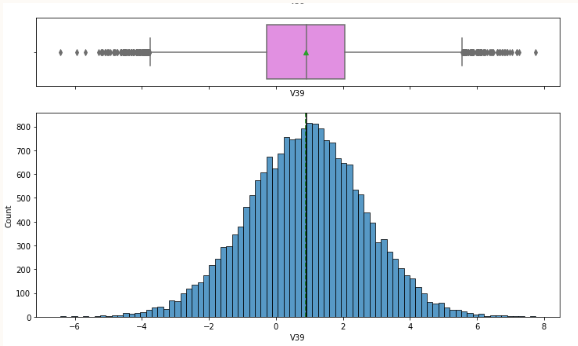

V34
Count

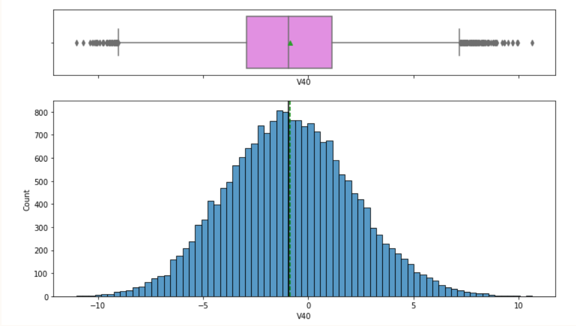

V35
Count

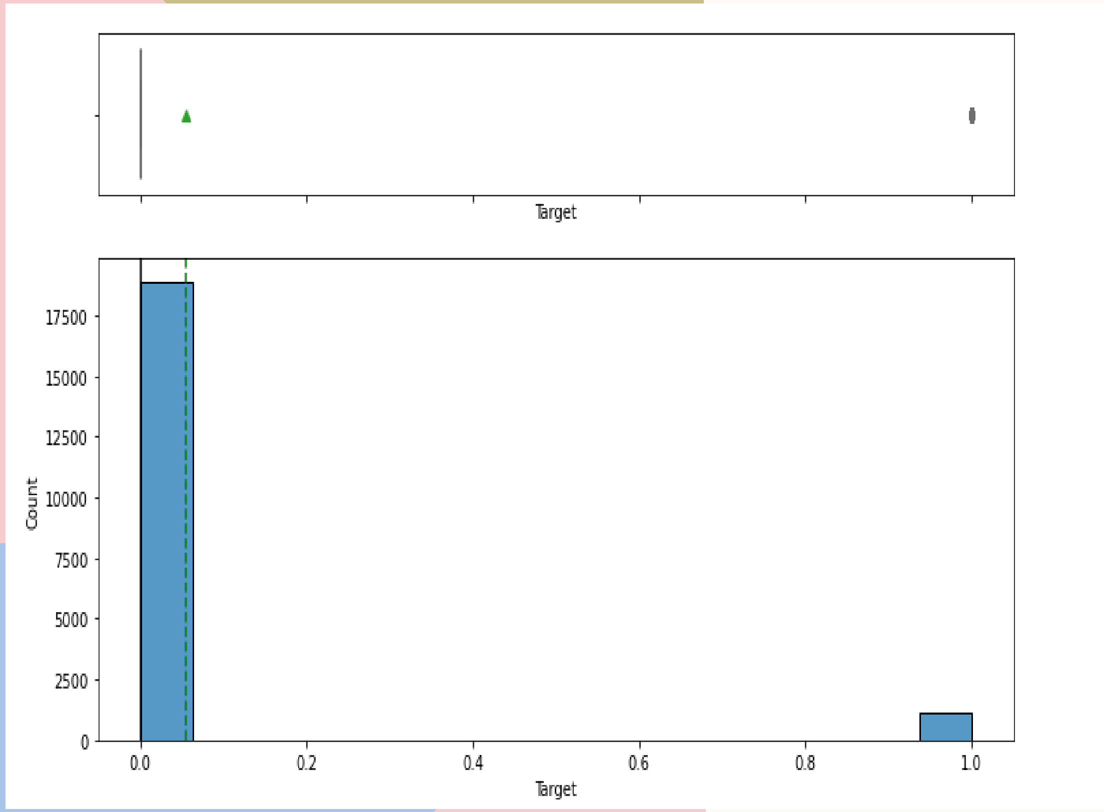

V36
Count

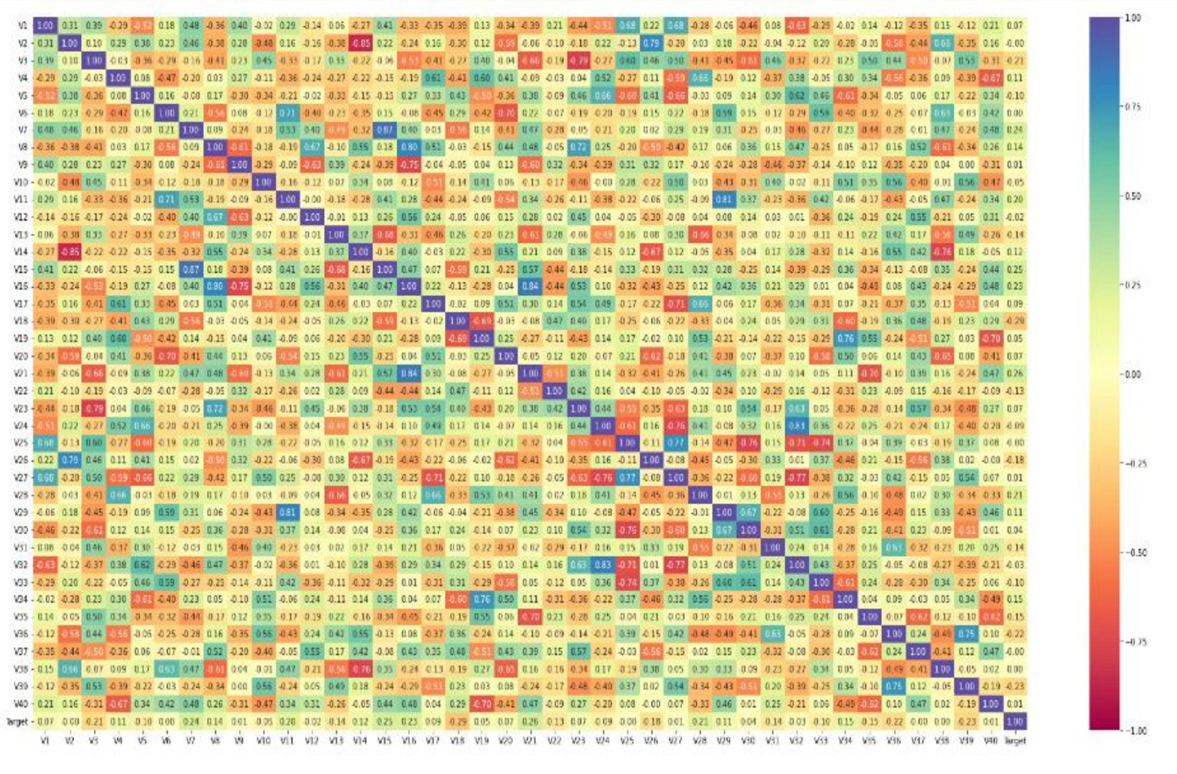

V37
Count


V38
Count


V39
Count


V40
Count

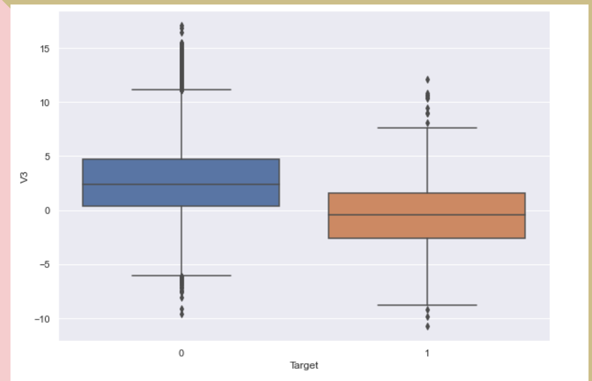

Target
17500
15000
12500
10000
7500
5000
2500
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
Target

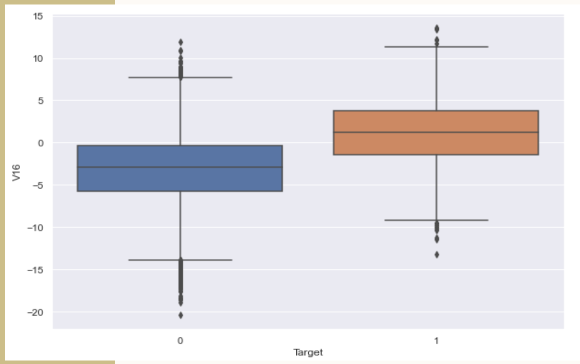

V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
V29
V30
V31
V32
V33
V34
V35
V36
V37
V38
V39
V40
Target

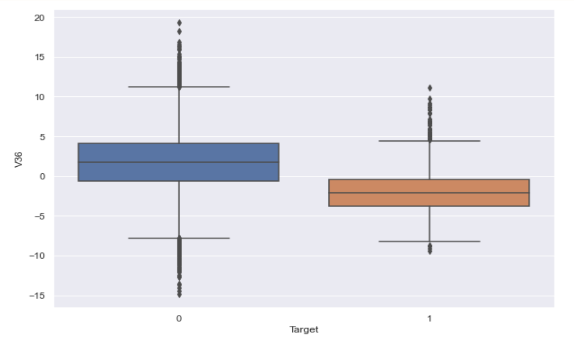
OBSERVATIONS

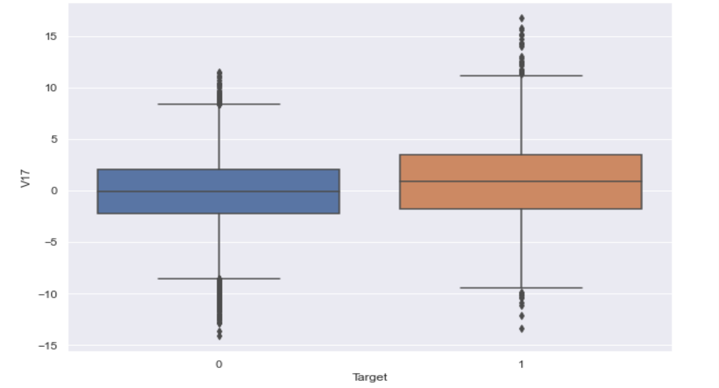
- All variables depicts a normal distribution except V1 with slight right skew demonstration and its target variables
- Outliers were observed based on the box plots indicating that the data collected is from sensors
- Based on the correlation heat map, it is shown that the strongest correlations are within variables V3, V7, V11, V15, V16, V18, V21, V28, V36, and V39
- Few correlations were seen between V11 and V29, V2 and V14, V2 and V26; and others
- Note --> correlations on the target values and other variables can be observed when utilizing some bivariate analysis

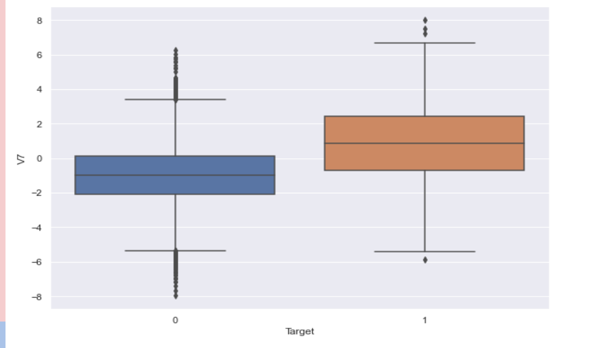
BOX PLOTS OF VARIABLES

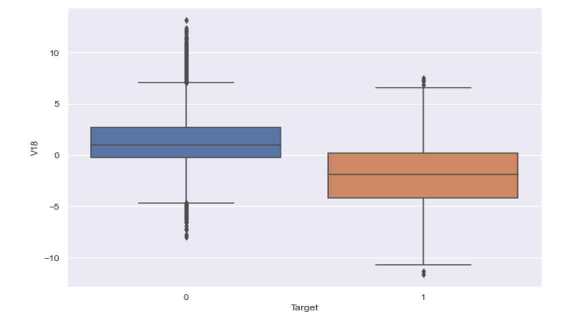


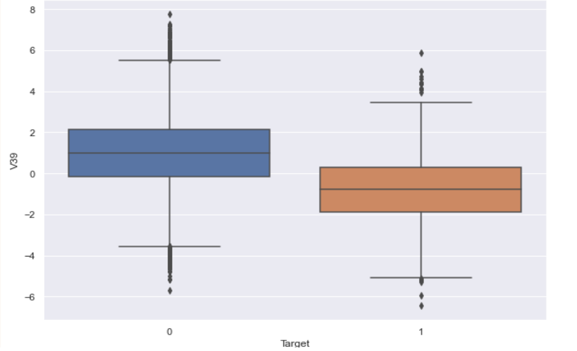


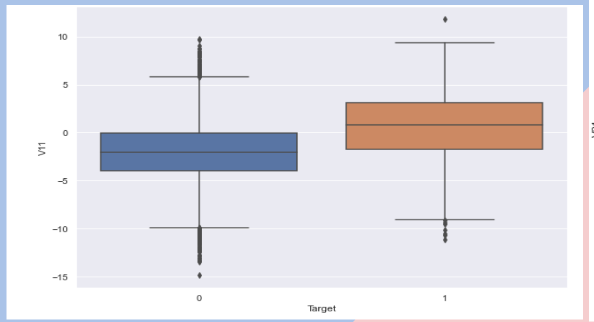


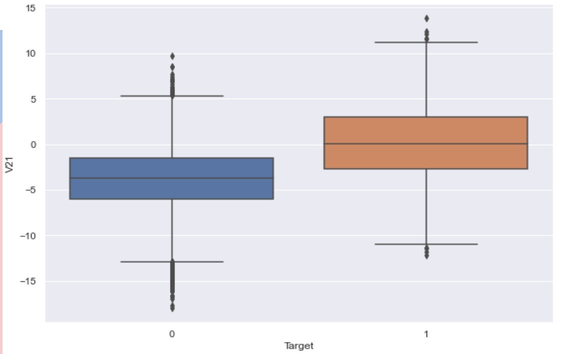


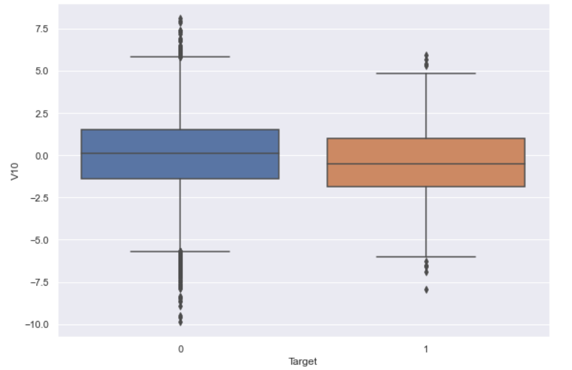


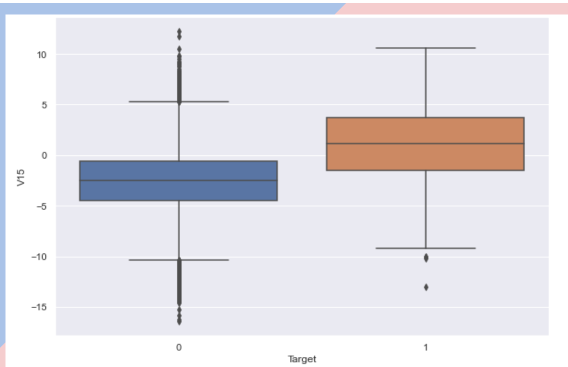


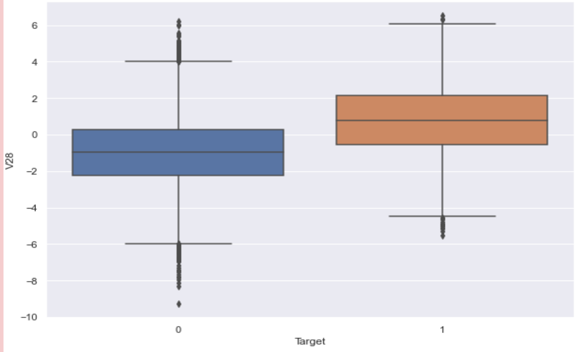


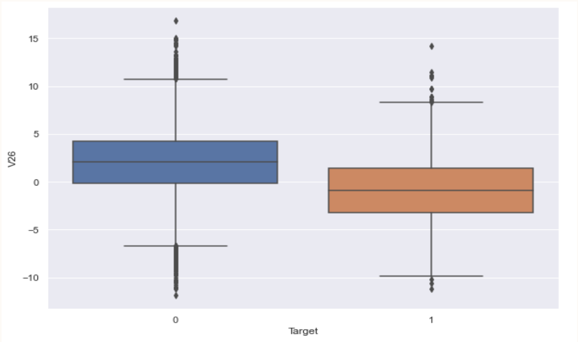


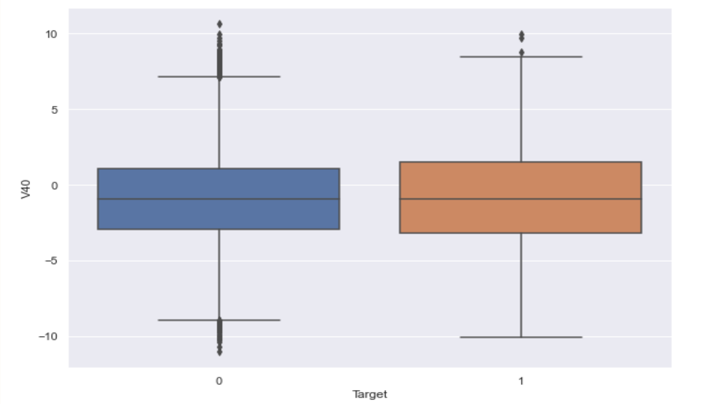








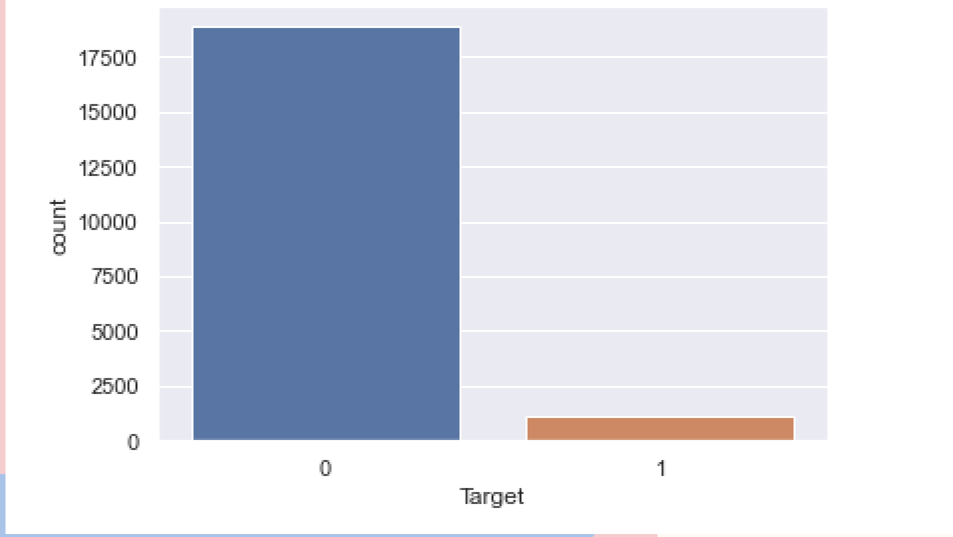


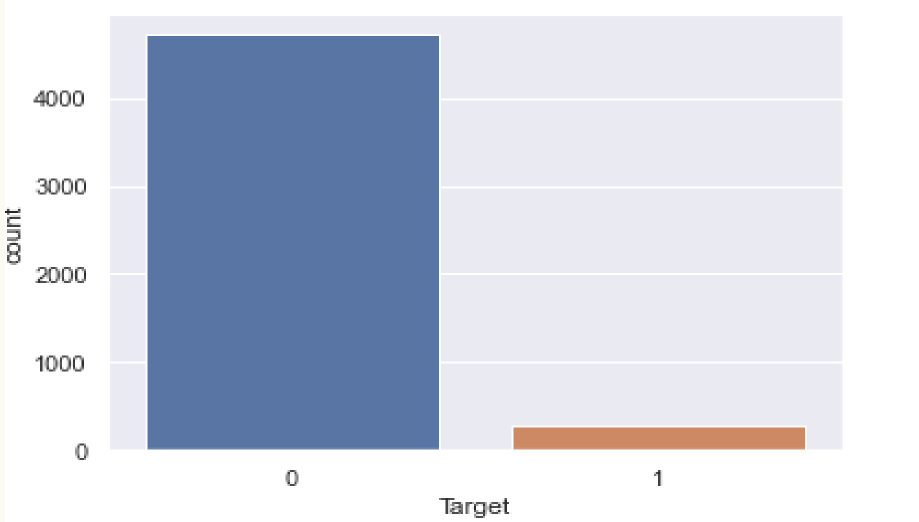


OBSERVATIONS

- Box plots indicates wind turbine generators with higher V3 values are mostly not to fail
    - 2nd and 3rd quartile V3 likely to fail is lower than those unlikely to fail
- Wind turbine generators with lower V7 values are likely not to fail
    - 2nd and 3rd quartile V7 likely to fail is greater than those likely not to fail
- Turbine generators with lesser V11 values are likely not to fail
    - 2nd and 3rd quartile V11 likely to fail is greater than those unlikely to fail
- Wind turbine generators with lesser V15 values are less likely to fail
    - 2nd and 3rd quartile V15 of those likely to fail is greater than those unlikely to fail
- Wind turbine generators with lower V16 values are less likely to fail
    - 2nd and 3rd quartile V16 of those likely to fail is greater than those unlikely to fail
- Wind turbine generators with lower V18 values have higher chances to fail
    - 2nd and 3rd quartile V18 of those likely to fail is lesser than those unlikely to fail
- Wind turbine generators with lower V21 values are less likely to fail
    - 2nd and 3rd quartile V21 values of those likely to fail are greater than those unlikely to fail
- Wind turbine generators with lesser V28 values are less likely to fail
    - 2nd and 3rd quartile V28 values of those likely to fail is greater than those unlikely to fail
- Wind turbine generators with greater V36 are less likely to fail
    - 2nd and 3rd quartile V36 values of those likely to fail is lesser than those unlikely to fail
- Wind turbine generators with lower V39 values are more likely to fail
    - 2nd and 3rd quartile V39 values of those likely to fail is lower than those unlikely to fail
- Wind turbine generators with lower V10 values are more likely to fail
- Wind turbine generators with lesser V26 values are more likely to fail
    - 2nd and 3rd quartile V26 of those likely to fail is lower than those unlikely to fail
- Wind turbine generators with higher V17 values are more likely to fail
- Thus: V40 variable values does not show impact on the failure of wind turbine generators
- Also, target variables value count is imbalanced

DATA PRE-PROCESSING

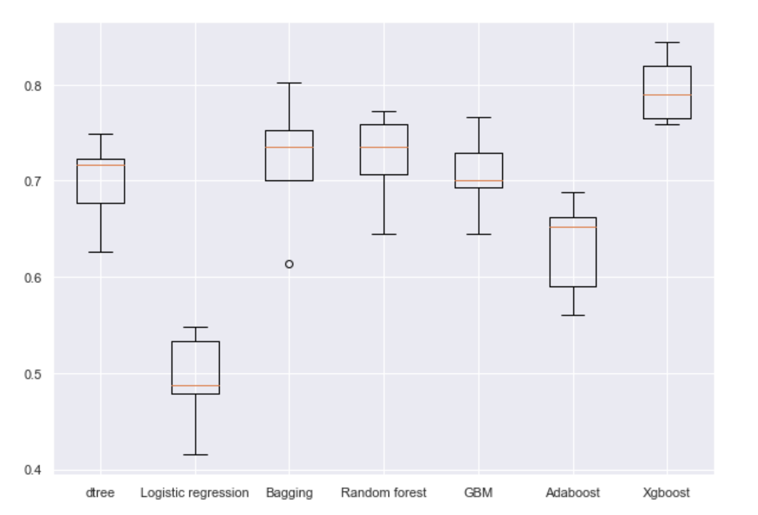




- Observations:
  - 18890 non-failures and 1110 failures w/n target variable of training set
  - 4718 non-failures and 282 failures
  - No duplicates w/n data sets

MODEL BUILDING

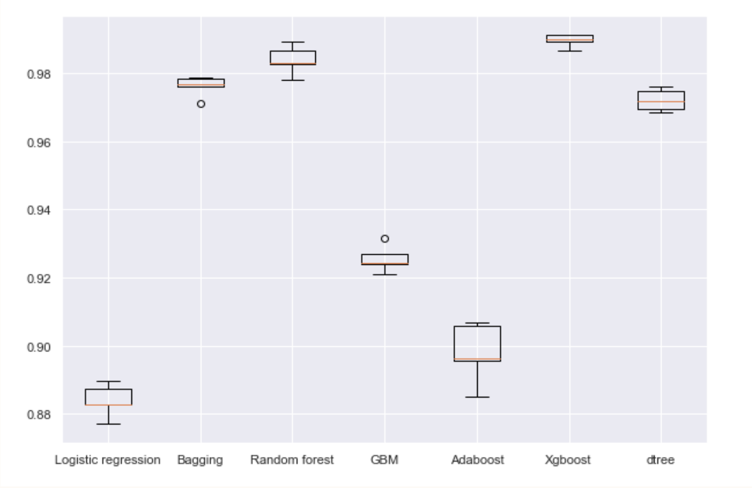
ORIGINAL DATA



Observations:

- XGBoost performed the best
- GBM demonstrates the least variance
- Models not overfitting and performance can be enhanced
- Best 3 models --> XGBoost, Bagging & Random forest

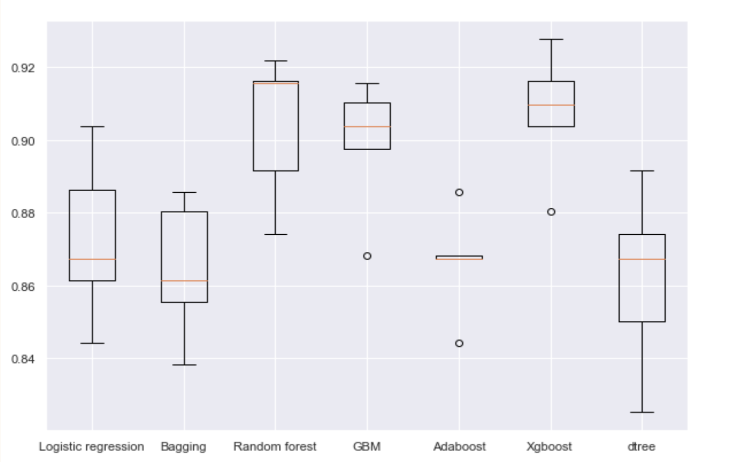
OVERSAMPLED DATA



Observations:

- XGBoost performed the best followed by random forest
- XGBoost shows least variance
- Slight overfitting observed and most overfitting is decision tree
- Logistics regression showed the poorest
- Best 3 models --> GBM, AdaBoost, & XGBoost

UNDERSAMPLED DATA

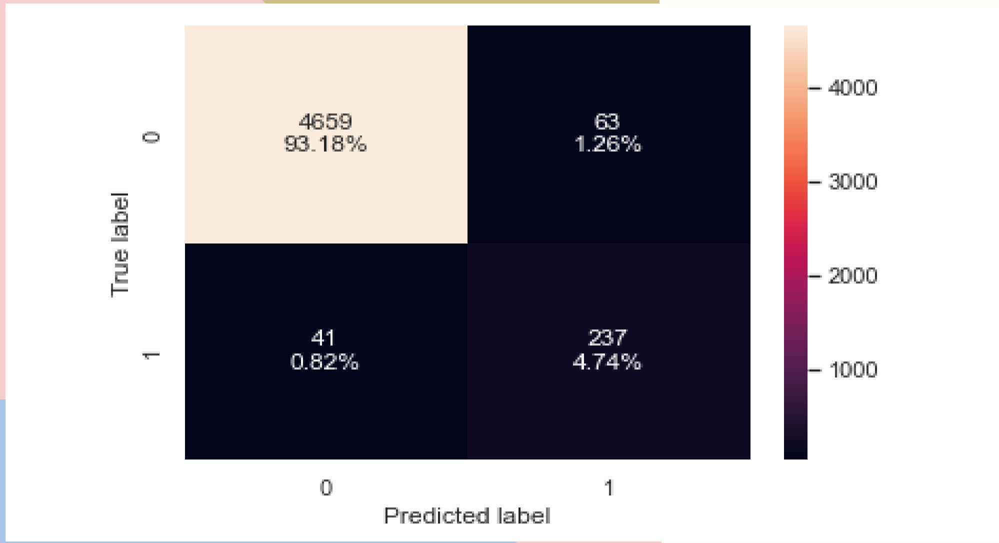


Observations:

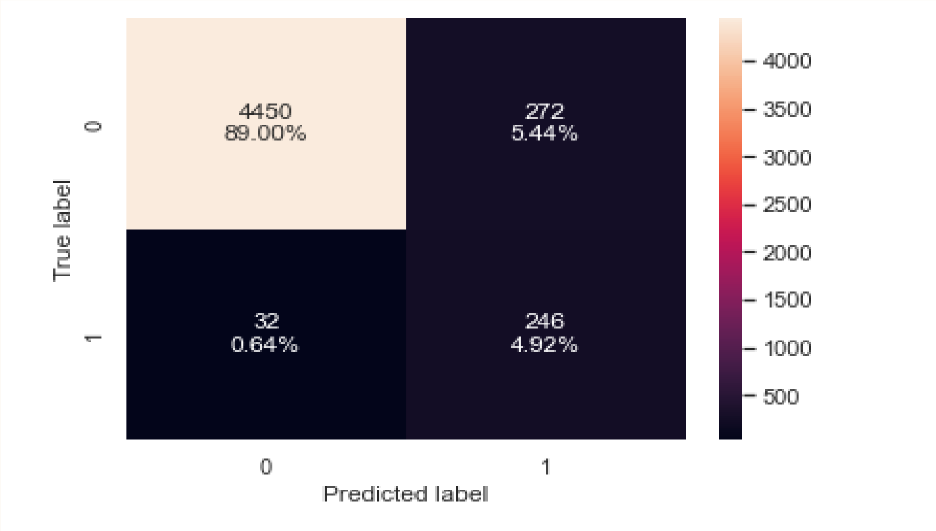
- XGBoost performed the best
- Decision tree consists of lowest recall score
- AdaBoost depicts least variance
- Satisfactory performance w/n Random forest, Bagging, & GBM

HYPERPARAMETER TUNING

AdaBoost – Oversampled Data



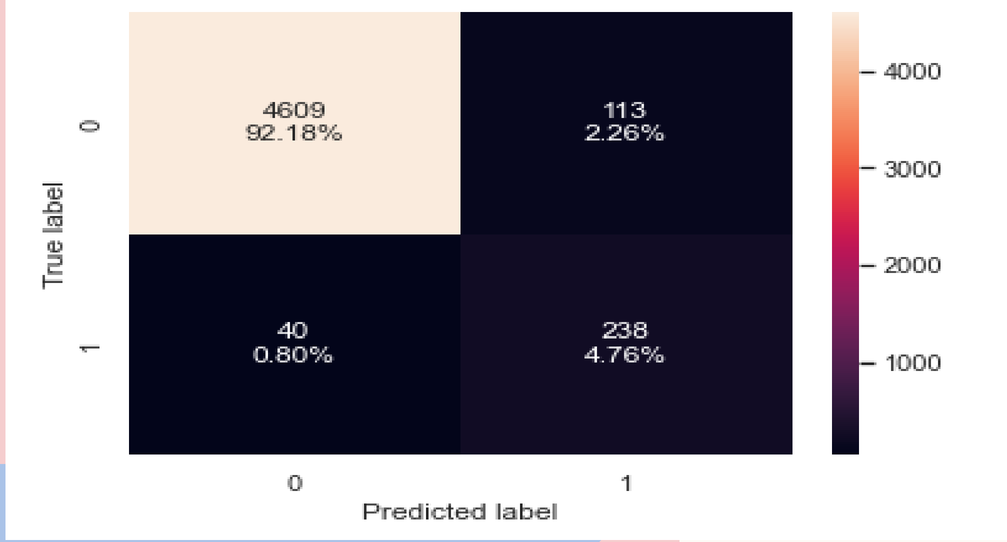
Random Forest – Undersampled Data



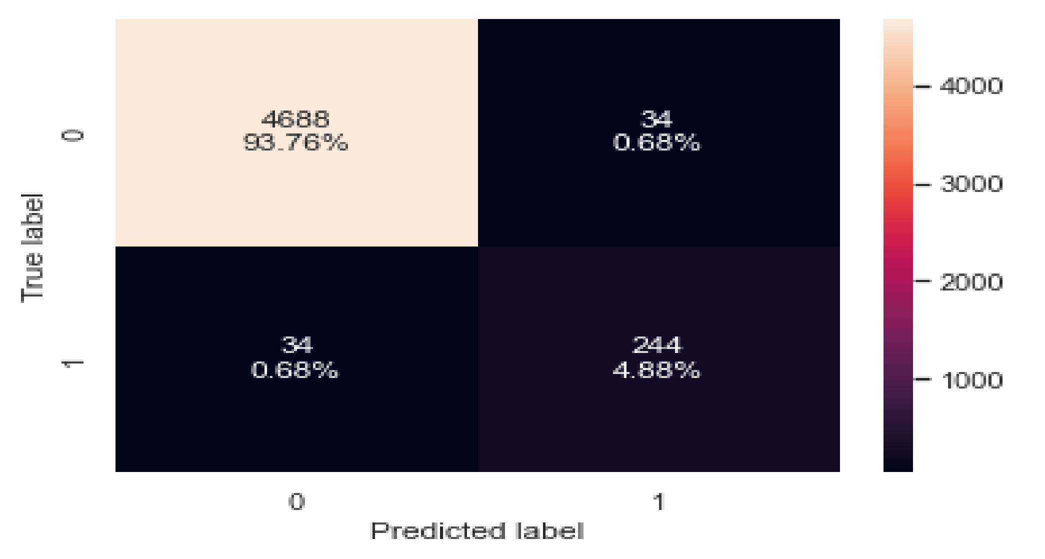
- AdaBoost w/ oversampling data depicts good performance
- Minimal overfitting in precision and recall scores
- 237 failures were accurately predicted

- Random Forest w/ undersampling data depicts good performance in terms of recall
- Consist of poor precision score and F1 score on validation set
- 246 failures were accurately predicted

Gradient Boosting – Oversampled Data



XGBoost – Oversampled Data



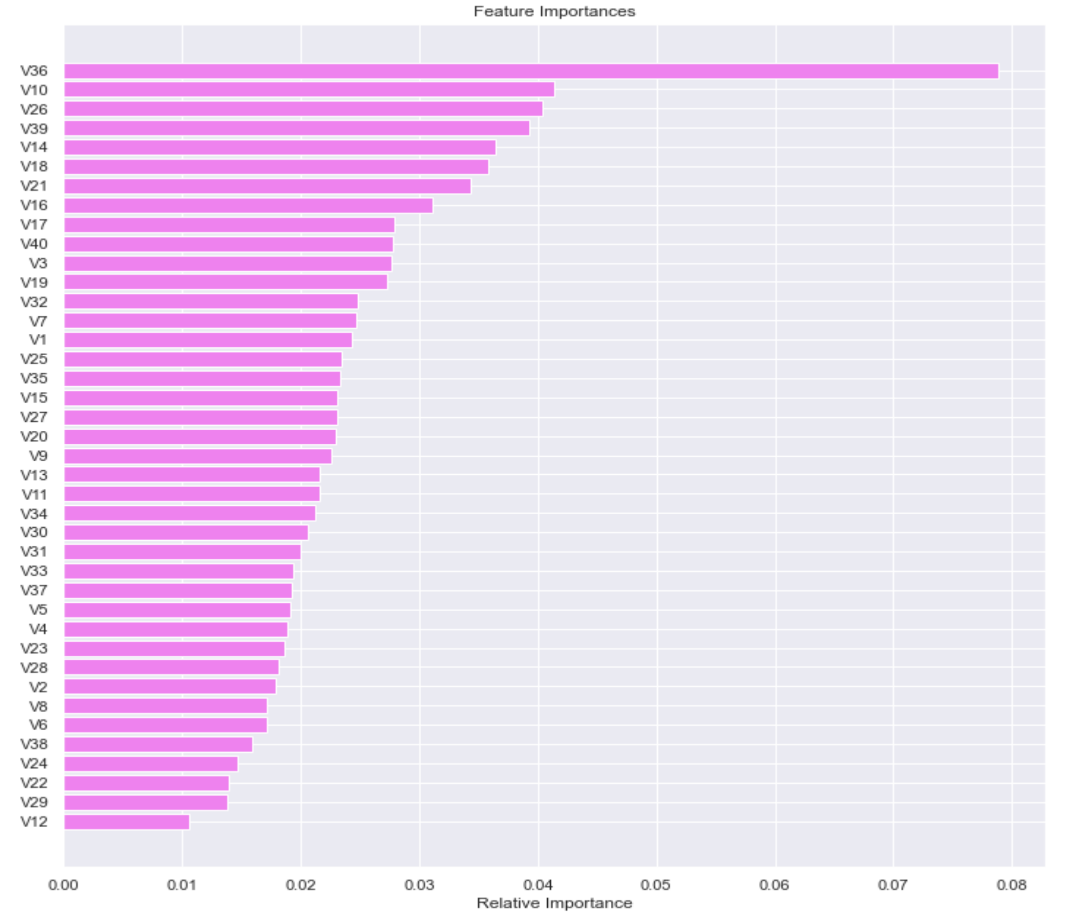
- GBM (Gradient Boosting Model) w/ oversampling data detects good performance in recall w/ 0.856 score on validation set
- Overfitting was shown in precision and F1 scores
- 238 failures were accurately predicted

- XGBoost w/ oversampled data shows good performance with recall score of 0.878 on validation set
- It is observed that precision, accuracy, and F1 scores are shown to be acceptable
- 244 failures were accurately predicted

MODEL PERFORMANCE & FINAL MODEL

	Decision tree tuned with oversampled data	Decision tree tuned with undersampled data	Gradient Boosting tuned with oversampled data	Gradient Boosting tuned with undersampled data	AdaBoost classifier tuned with oversampled data	AdaBoost classifier tuned with undersampled data	Random forest tuned with oversampled data	Random forest tuned with undersampled data
Accuracy	0.843	0.764	0.993	0.995	0.992	0.950	1.000	0.978
Recall	0.917	0.909	0.992	0.992	0.988	0.916	0.999	0.958
Precision	0.799	0.705	0.994	0.998	0.995	0.982	1.000	0.999
F1	0.854	0.794	0.993	0.995	0.992	0.948	1.000	0.978

	Decision tree tuned with oversampled data	Decision tree tuned with undersampled data	Gradient Boosting tuned with oversampled data	Gradient Boosting tuned with undersampled data	AdaBoost classifier tuned with oversampled data	AdaBoost classifier tuned with undersampled data	Random forest tuned with oversampled data	Random forest tuned with undersampled data
Accuracy	0.763	0.609	0.969	0.919	0.979	0.928	0.988	0.939
Recall	0.885	0.888	0.856	0.871	0.853	0.878	0.863	0.885
Precision	0.176	0.114	0.678	0.395	0.790	0.427	0.920	0.475
F1	0.294	0.202	0.757	0.544	0.820	0.574	0.891	0.618



OBSERVATIONS

- XGBoost w/ undersampled data depicts best performance with recall score 0.906 on validation set
    - Shows failures were correctly predicted by the model leading in repairing cost
- The next best models are decision tree with undersampled data, decision tree with oversampled data, and tuned random forest with undersampled data
- Most important feature is V36 followed by V10, V26, V39, V14, V18, V21, V16, V17, and V40
- The least important features are V12, V29, V22, V24, and V38
- Within the testing data, recall score of 0.887 was observed
    - Note: XGBoost undersampled model performed the most on the testing data

CONCLUSIONS

- XGBoost with undersampled data consist of the best performance on validation set and depicts the best performance on the testing data with recall score of 0.887
- The model verified did not show great precision and F1 score but demonstrated highest recall score
- However, model predicted the highest number of True positives and lowest number of false negatives which will be very useful in saving maintenance cost
- Top 10 features for best performing model in order of importance: V36,V10,V26,V39,V14,V18,V21,V16,V17 and V40 while V12, V29, V22, V24 and V38 were seen not to be important features
- Higher values in variables such as V14,V21,V16 and V17 indicates that generators are likely to fail when values are high whereas, when variables V36,V10,V26,V39,V18 are low, are more likely to fail
    - Thus; these variables need to be monitored to save maintenance cost
- Variable V40 depicted no significant difference in values for either failure or non-failure, but variance in the failure range is slightly wider than none failures
    - This signifies an important feature within the model

- However, pipeline was developed to generate the chosen final model

RECOMMENDATIONS

- In the future, in the event there is a change and company move to the direction to prioritize both precision and recall as metrics then oversampled AdaBoost, oversampled rf1 and oversampled XGB will be best to utilized
- It is recommended to perform routine checks on V40 variable by obtaining more data to investigate the pattern while providing more information
- It is also recommended that Renewind utilized timer within the sensors for further analysis to verify the length of time it takes for a sensor to move from safe zone to red zone
    - This will lead to more improvement within the model and further savings in maintenance cost
- Alarm systems or warning messages should be implemented to the sensors to trigger an alert once values are within or near the failing zone

THANK YOU