ReneWind

Eugenie Seholm

Contents

Business Problem - Overview

Problem:

- *ReneWind* is working to improve machinery/processes involved in the production of wind energy using ML and has collected data of generator failure of wind turbines using sensors.
- There are sensors fitted across different machines that collect data (e.g., temperature, humidity, wind speed) and additional features related to various parts of the wind turbine (e.g., gearbox, tower, blades, break, etc.)
- Use of predictive maintenance methods can potentially predict component failure accurately so that components can be replaced before failure, keeping operation and maintenance costs down.

Objective:

- Build various classification models, tune them, and find the best one that will help identify failures so that the generator may be repaired prior to failure/breakage and the overall maintenance cost of the generators can be brought down.

Business Problem - Financial Implications

- True positives (TP) are failures correctly predicted by the model
- False negatives (FN) are real failures in the generator of wind turbine where there is no detection by the model
- False positives (FP) are failure detections in the generator of the wind turbine where there is no actual failure

While we would like to minimize FN and FP, our main objective is to **minimize maintenance cost** associated with the model:

Maintenance cost = TP*(Repair cost) + FN*(Replacement cost) + FP*(Inspection cost),

where replacement cost = $40,000; repair cost = $15,000; inspection cost = $5,000.

Business Problem - Solution Approach

- Data preparation
- Split data into train, validation, and test sets
- Train models using training set on:
  - Original dataset
  - Oversampled dataset
  - Undersampled dataset
- Test performance on validation set, maximizing *minimum vs model cost* (higher values → lower maintenance cost)
- Select best 3 performing models
- Improve 3 models with hyperparameter tuning
- Select top performing model
- Create pipeline to productionize the model

Data Overview

- Train set
  - 40,000 observations
  - 41 columns (40 numerical variables + 1 target)
- Test set
  - 10,000 observations
  - 41 columns (40 numerical variables + 1 target)
- All numerical columns encoded with a cipher
- No duplicated rows
- Two columns (V1 & V2) with missing values in both train and test sets
  - V1 - 0.12% missing values in train set, 0.11% missing values in test set
  - V2 - 0.1% missing values in train set, 0.07% missing values in test set
- V1 through V40 have varying range of values including negative and positive float values between -24 to 25

Data Manipulations

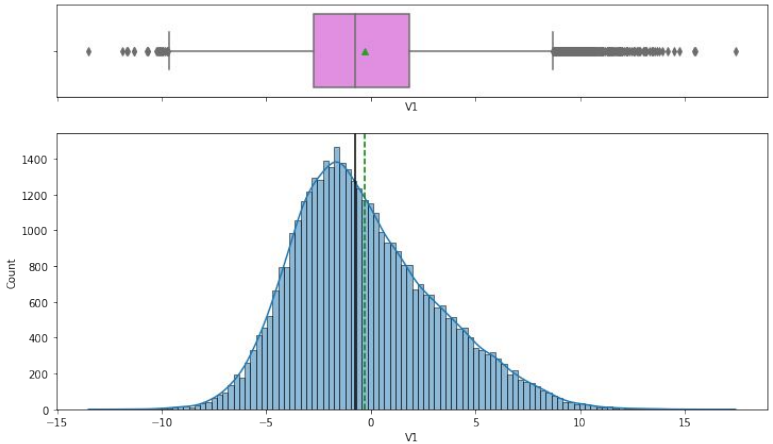
- Split train dataset into train and validation sets
- Use simple imputation to fill missing values with the median
- Original data set with simple imputation for one set of models
- Oversampled data for one set of models
- Undersampled data for one set of models

Exploratory Data Analysis (EDA)

Exploratory Data Analysis - V1 & V2

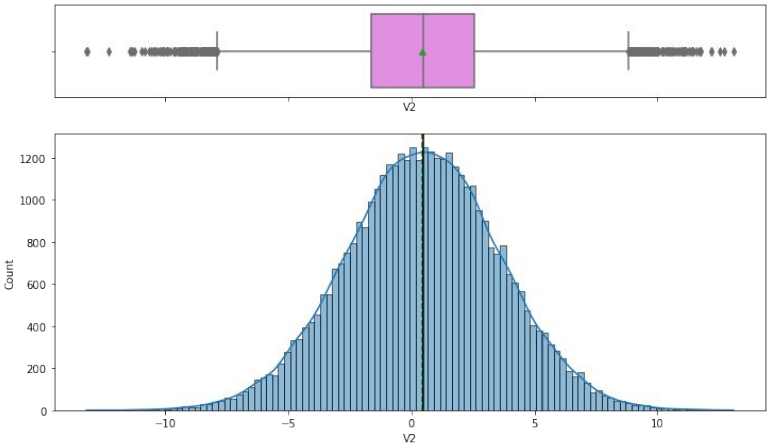
V1 Observations

- V1 is slightly skewed in the positive direction but otherwise looks like an approximately normal distribution
- Outliers in both directions
- Negative mean and median



V2 Observations

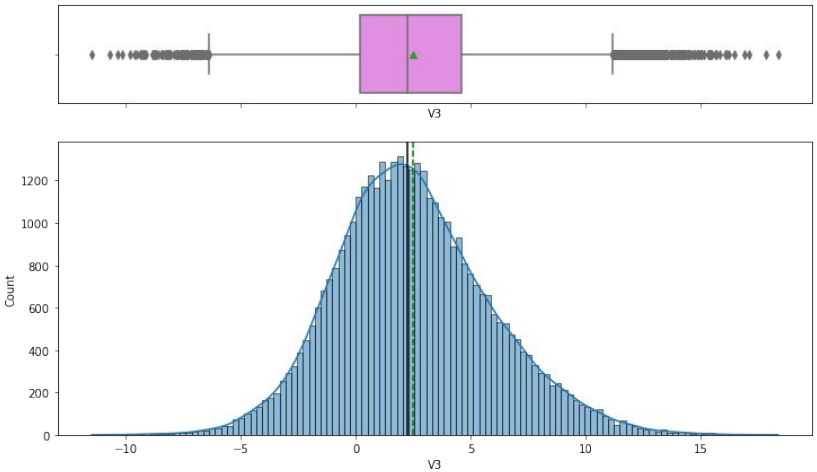
- Approximately normal distribution
- Outliers in both directions
- Positive mean and median



Exploratory Data Analysis - V3 & V4

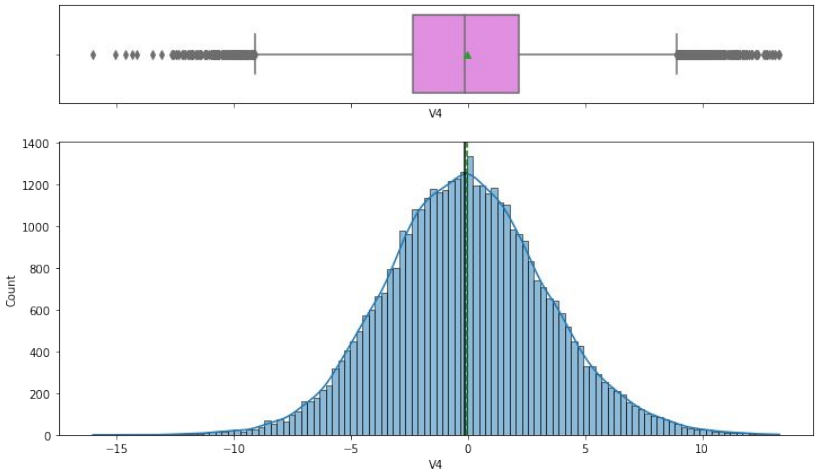
V3 Observations

- V3 is slightly negatively skewed, but otherwise has an approximately normal distribution
- Outliers in both directions
- Positive mean and median



V4 Observations

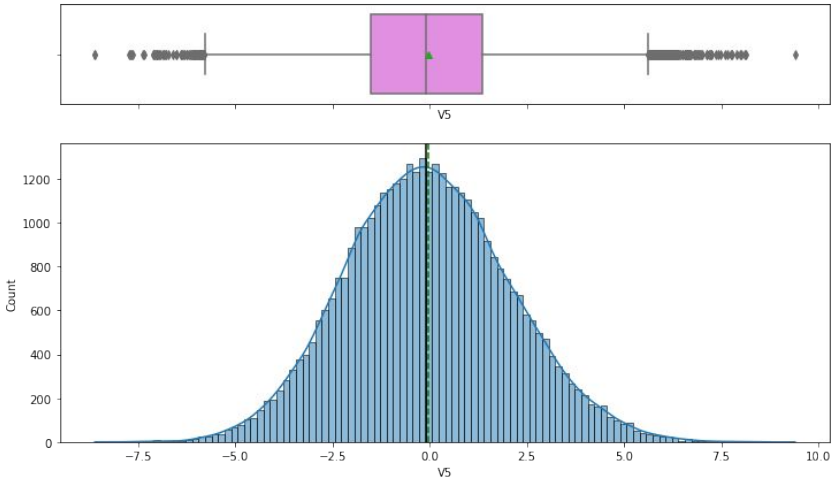
- V4 has an approximately normal distribution
- Outliers in both directions
- Negative mean and median



Exploratory Data Analysis - V5 & V6

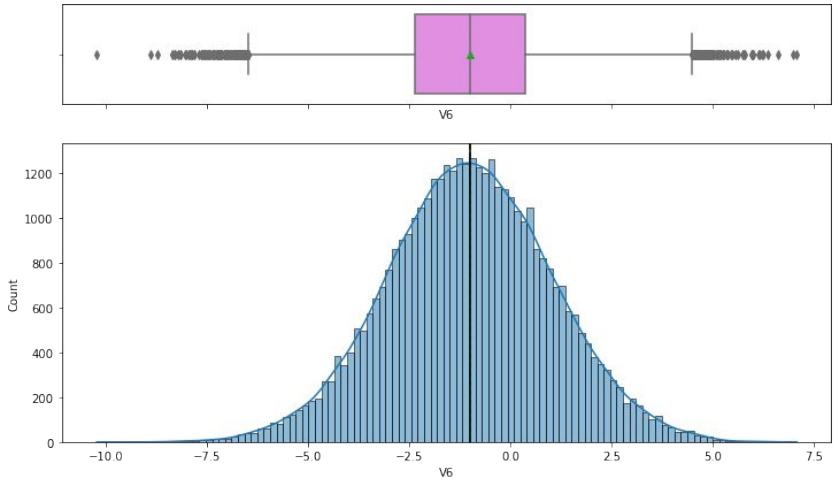
V5 Observations

- V5 has an approximately normal distribution
- Outliers in both directions
- Mean and median close to zero, but negative



V6 Observations

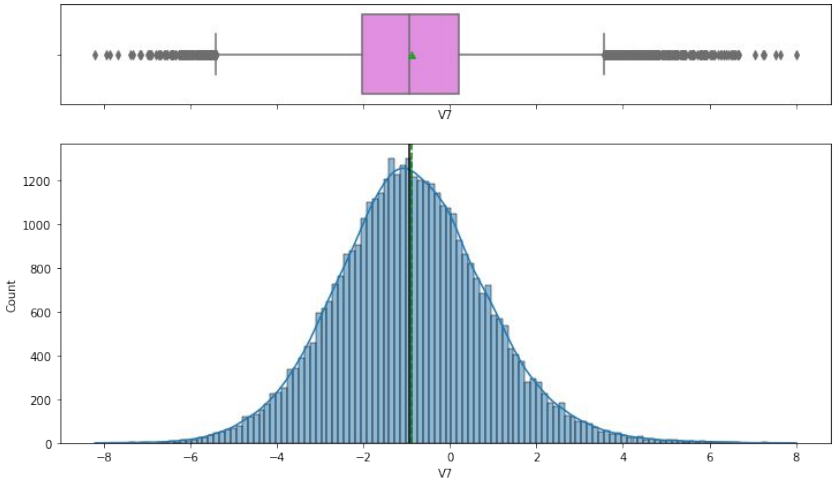
- V6 has an approximately normal distribution
- Outliers in both directions
- Negative mean and median



Exploratory Data Analysis - V7 & V8

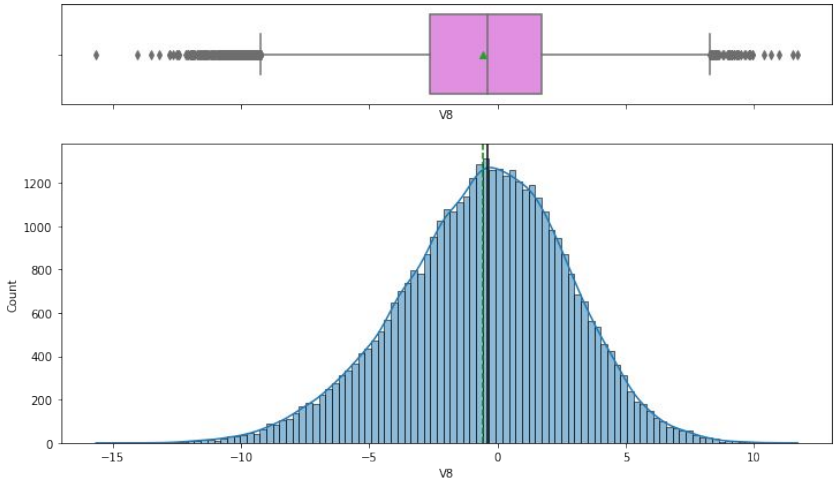
V7 Observations

- V7 has an approximately normal distribution
- Outliers in both directions
- Negative mean and median



V8 Observations

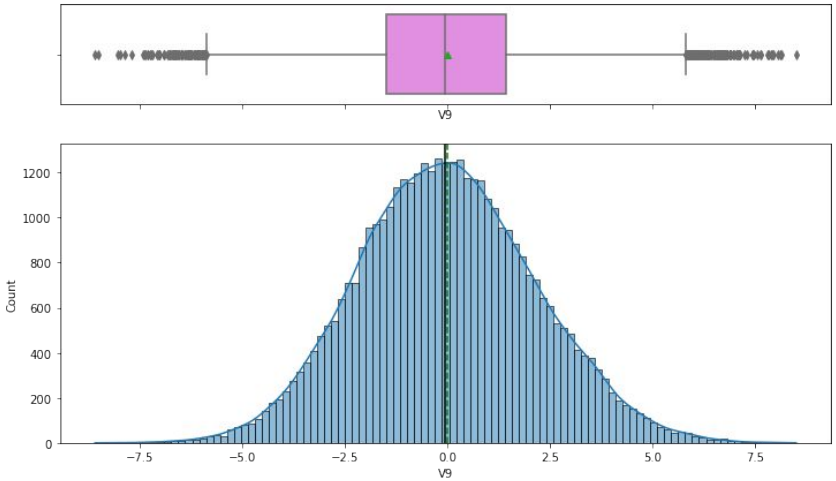
- V8 has a slight negative skew
- Outliers in both directions
- Negative mean and median



Exploratory Data Analysis - V9 & V10

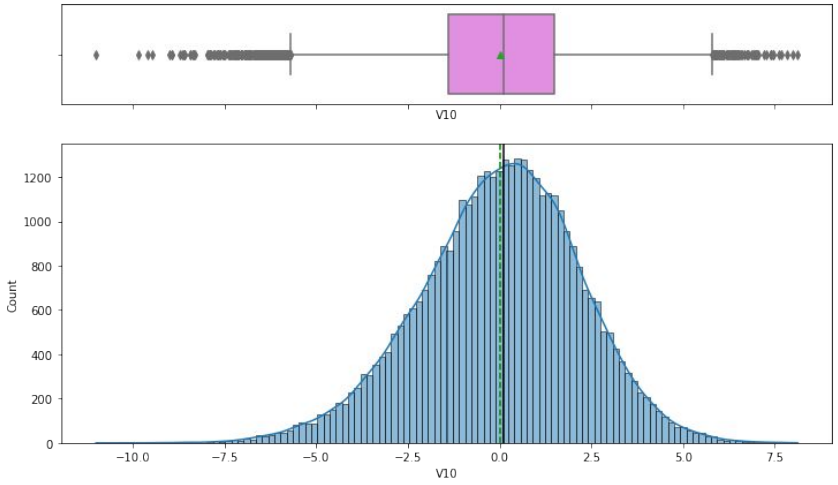
V9 Observations

- V9 has an approximately normal distribution
- Outliers in both directions
- Mean and median approximately zero (negative)



V10 Observations

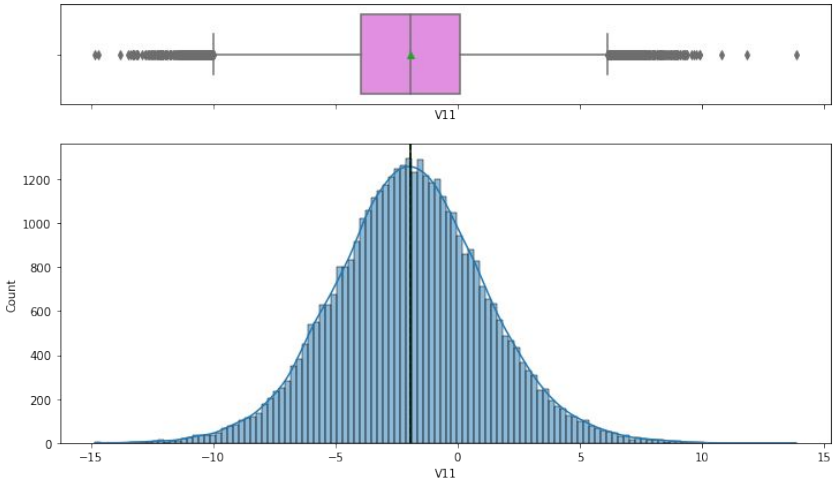
- V10 has a slight negative skew
- Outliers in both directions
- Mean negative (approximately zero) and median positive



Exploratory Data Analysis - V11 & V12

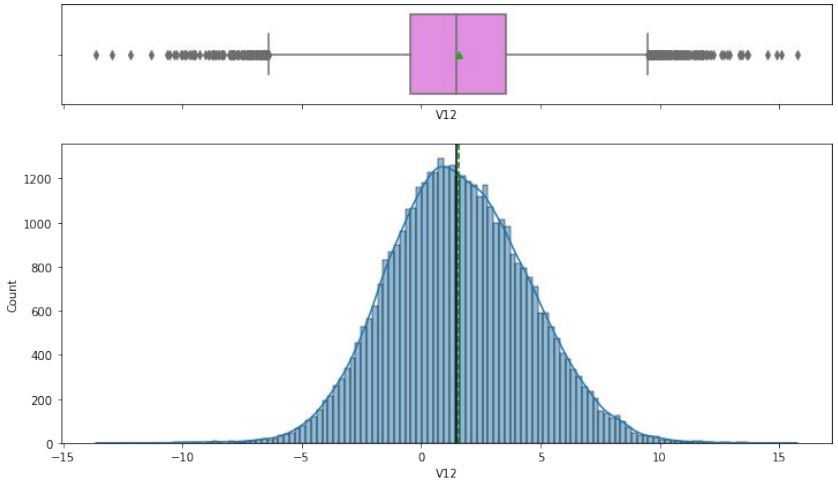
V11 Observations

- V11 has an approximately normal distribution
- Outliers in both directions
- Negative mean and median



V12 Observations

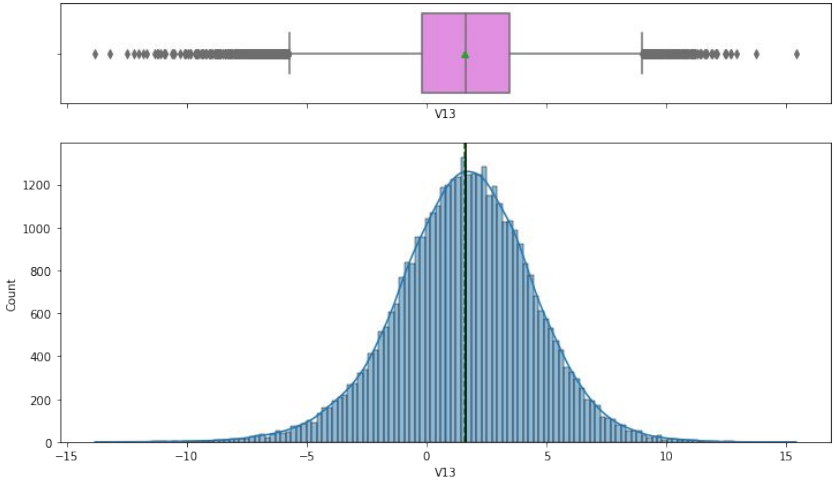
- V12 has a slight positive skew
- Outliers in both directions
- Positive mean and median



Exploratory Data Analysis - V13 & V14

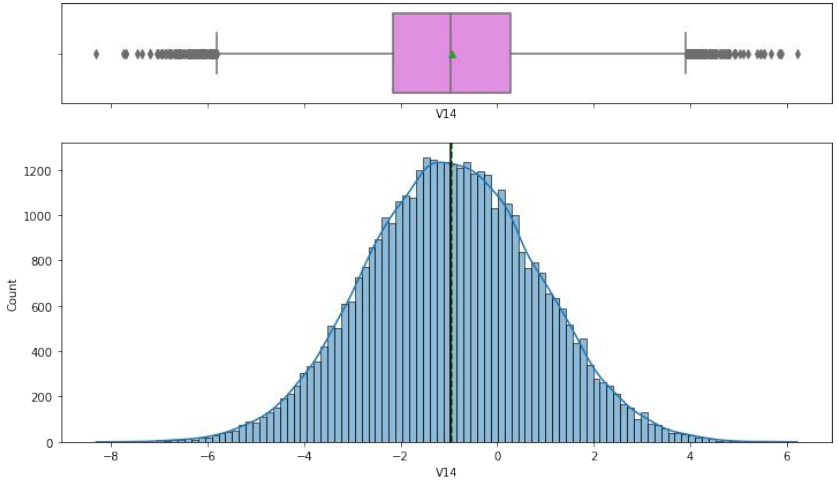
V13 Observations

- V13 has an approximately normal distribution
- Outliers in both directions
- Positive mean and median



V14 Observations

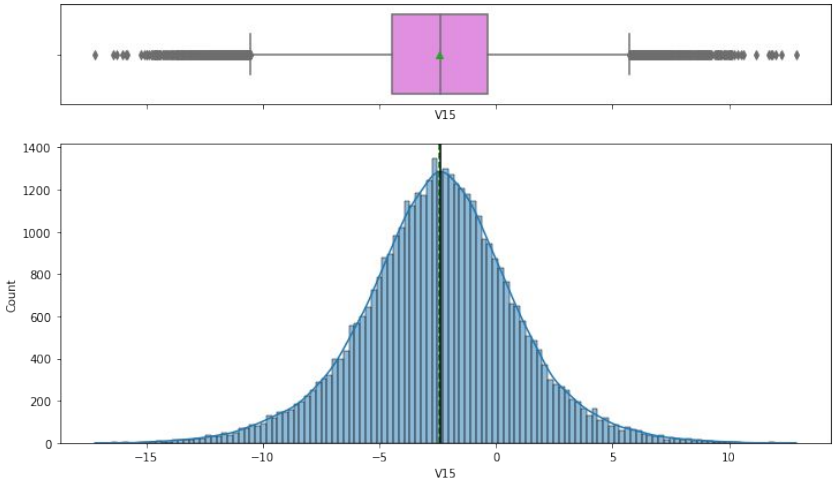
- V14 has an approximately normal distribution
- Outliers in both directions
- Negative mean and median



Exploratory Data Analysis - V15 & V16

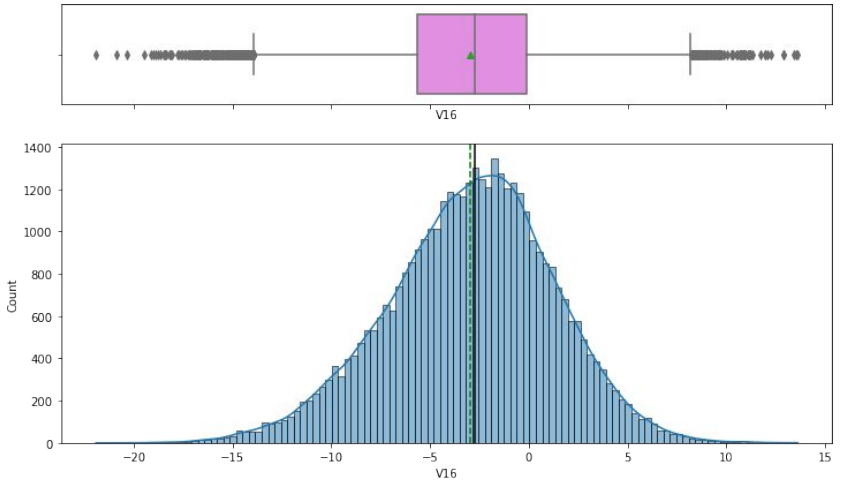
V15 Observations

- V15 has an approximately normal distribution
- Outliers in both directions
- Mean and median negative



V16 Observations

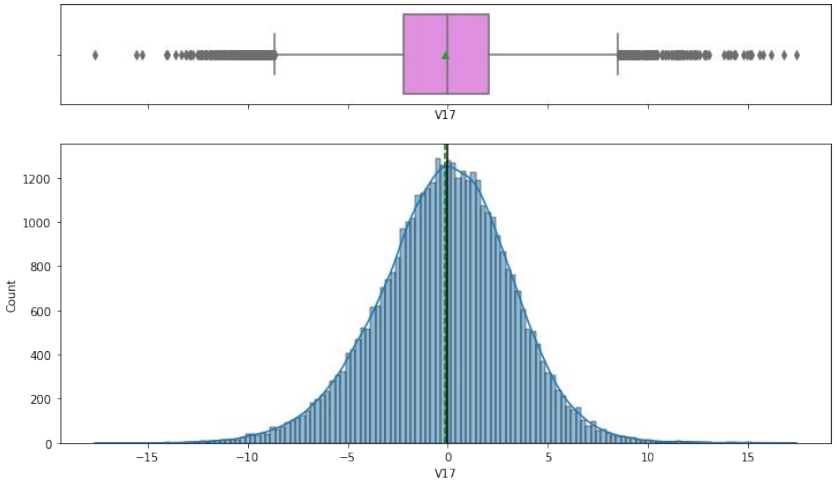
- V16 has a slight negative skew
- Outliers in both directions
- Mean and median both negative



Exploratory Data Analysis - V17 & V18

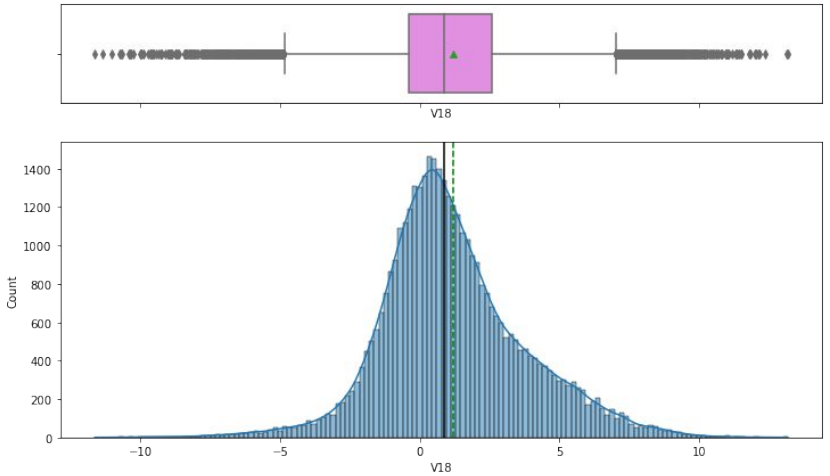
V17 Observations

- V17 has an approximately normal distribution
- Outliers in both directions
- Mean and median approximately zero (negative)



V18 Observations

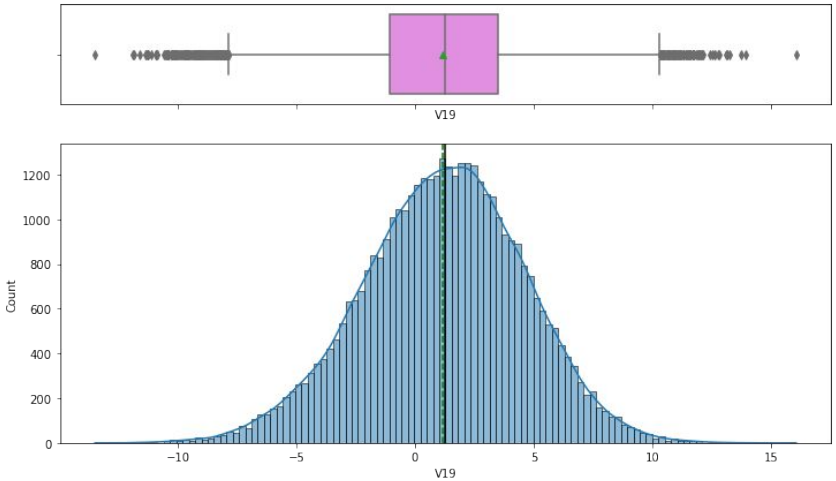
- V18 has a positive skew
- Outliers in both directions
- Mean and median are both positive



Exploratory Data Analysis - V19 & V20

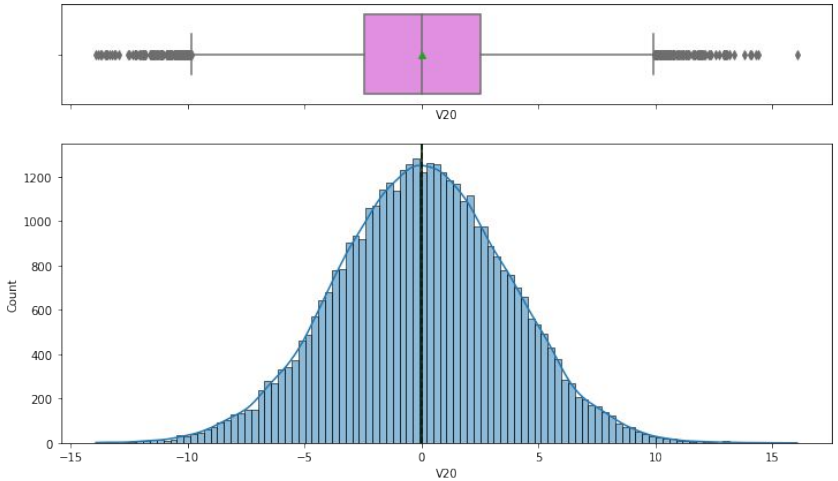
V19 Observations

- V19 has an approximately normal distribution
- Outliers in both directions
- Mean and median both positive



V20 Observations

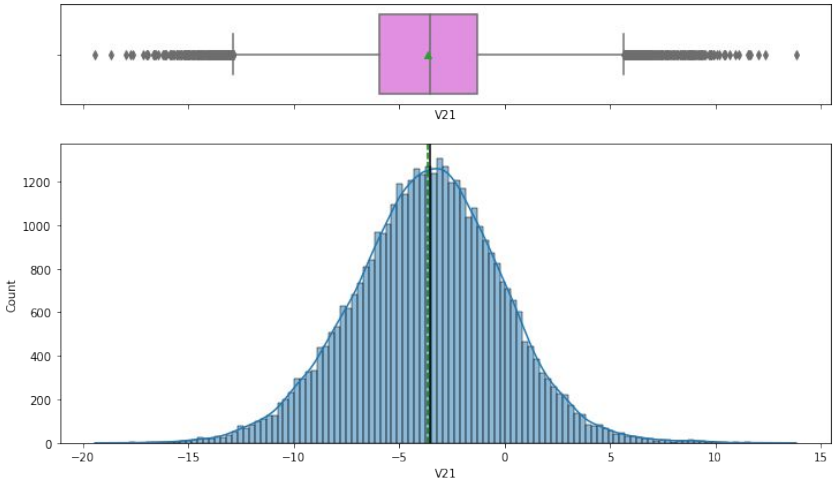
- V20 has an approximately normal distribution
- Outliers in both directions
- Mean and median approximately zero (negative)



Exploratory Data Analysis - V21 & V22

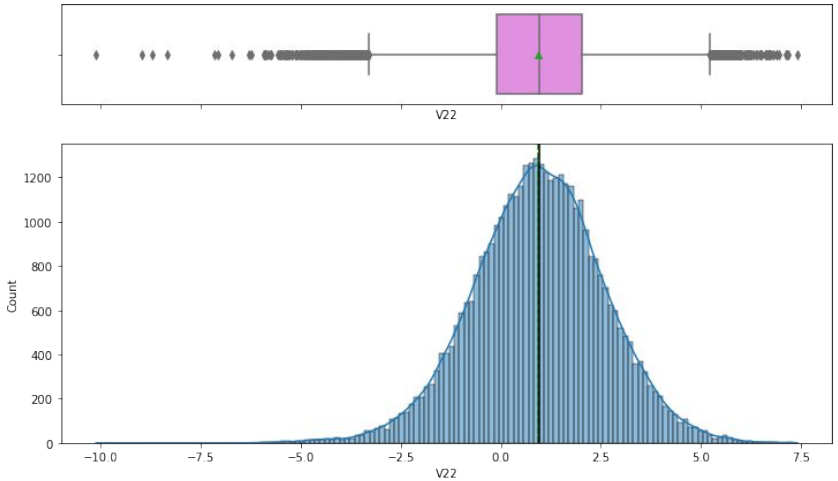
V21 Observations¶

- V21 has an approximately normal distribution
- Outliers in both directions
- Mean and median both negative



V22 Observations

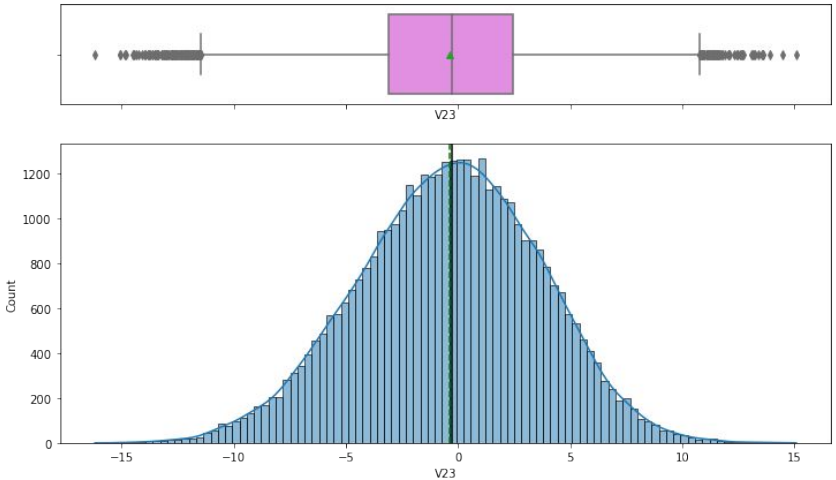
- V22 has an approximately normal distribution
- Outliers in both directions, but negative direction outliers much farther from the measure of center
- Mean and median both positive



Exploratory Data Analysis - V23 & V24

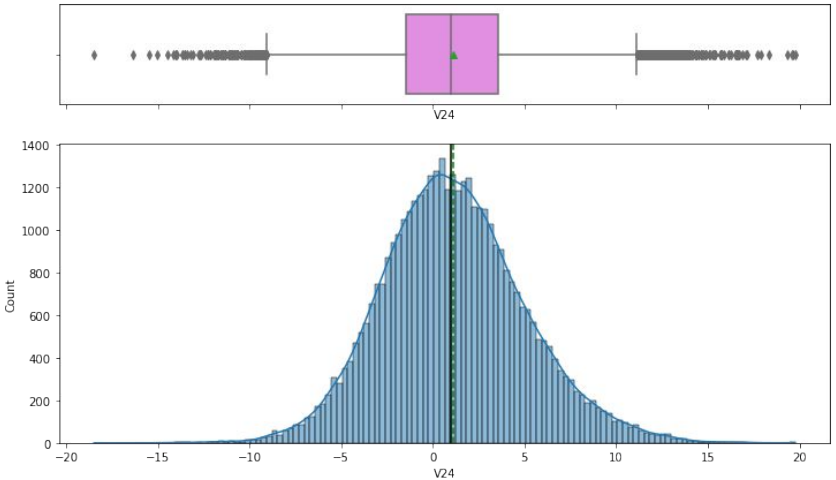
V23 Observations

- V23 has an approximately normal distribution
- Outliers in both directions
- Mean and median are negative, close to zero



V24 Observations

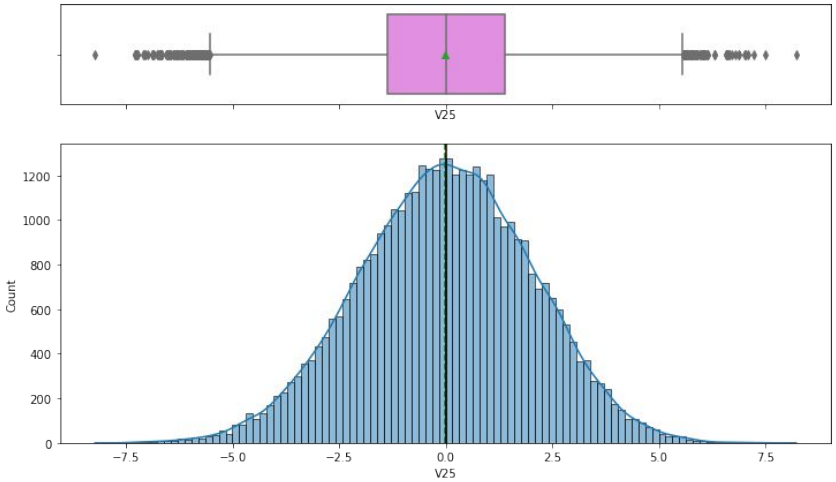
- V24 has an approximately normal distribution with a slight positive skew
- Outliers in both directions
- Mean and median both positive



Exploratory Data Analysis - V25 & V26

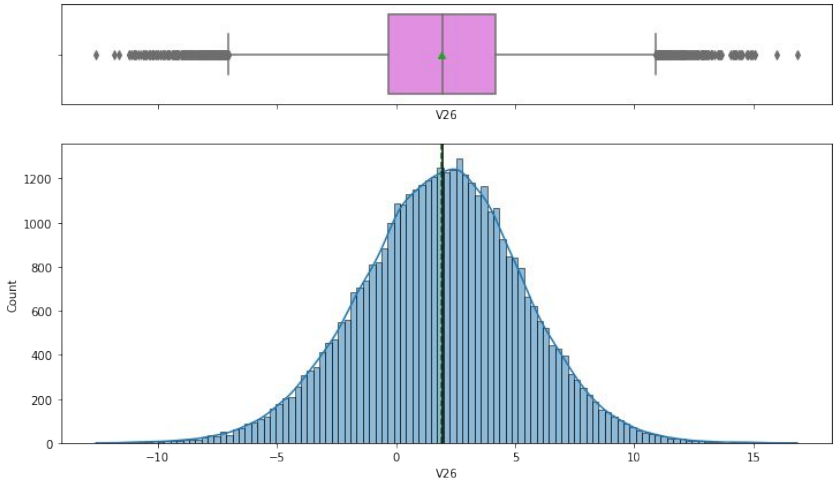
V25 Observations

- V25 has an approximately normal distribution
- Outliers in both directions
- Mean and median approximately zero



V26 Observations

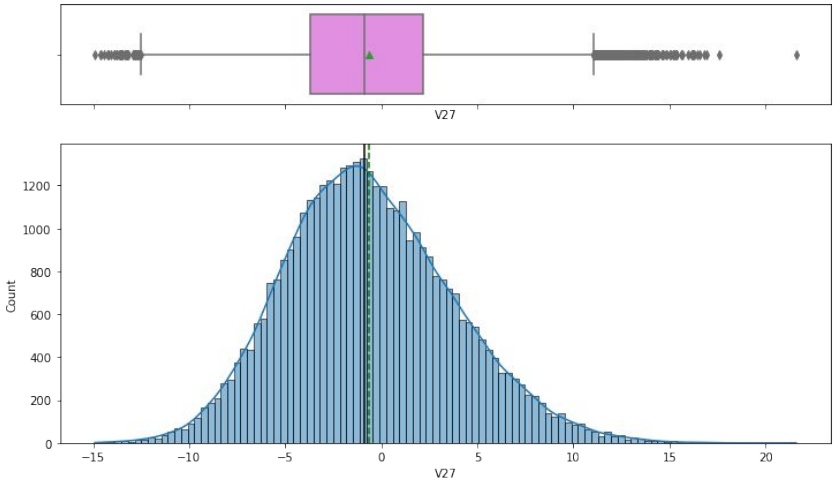
- V26 has an approximately normal distribution
- Outliers in both directions
- Mean and median both positive



Exploratory Data Analysis - V27 & V28

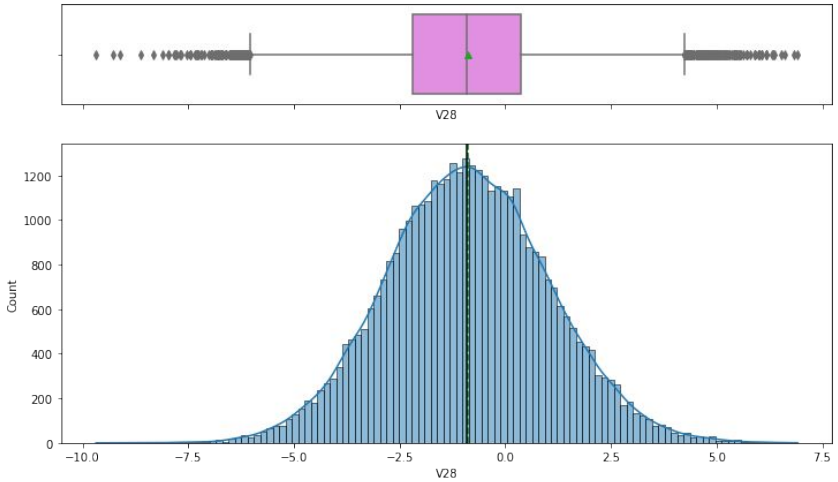
V27 Observations

- V27 has a slight positive skew
- Outliers in both directions
- Mean and median both negative



V28 Observations

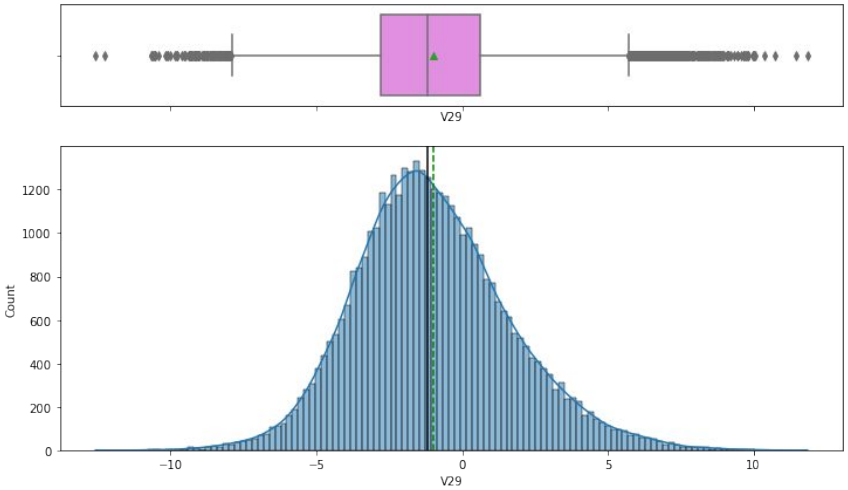
- V28 has an approximately normal distribution
- Outliers in both directions
- Mean and median negative



Exploratory Data Analysis - V29 & V30

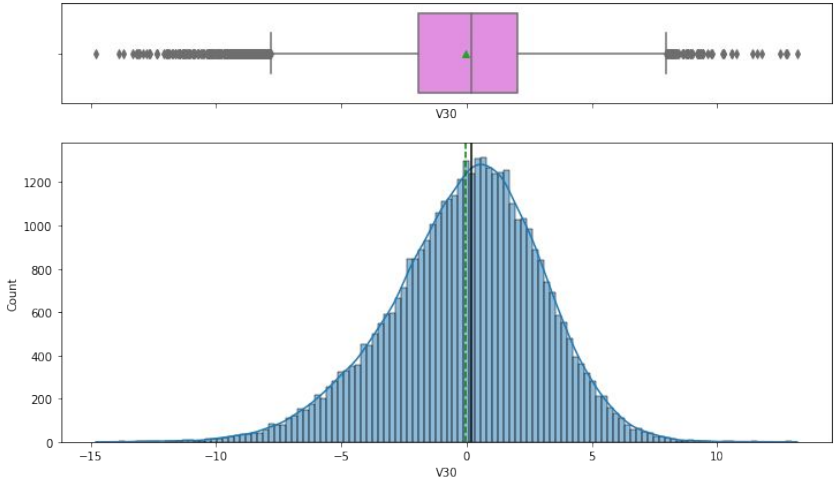
V29 Observations

- V29 has a slight positive skew
- Outliers in both directions
- Mean and median negative



V30 Observations

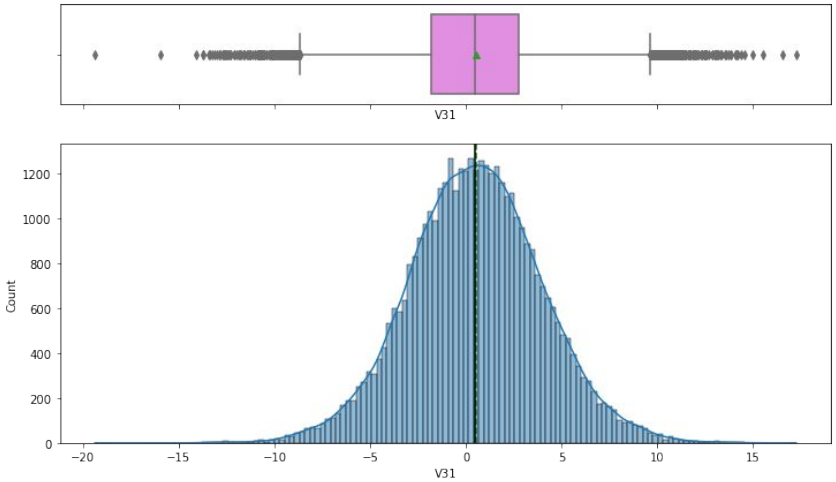
- Slight negative skew
- Outliers in both directions
- Mean negative, median positive



Exploratory Data Analysis - V31 & V32

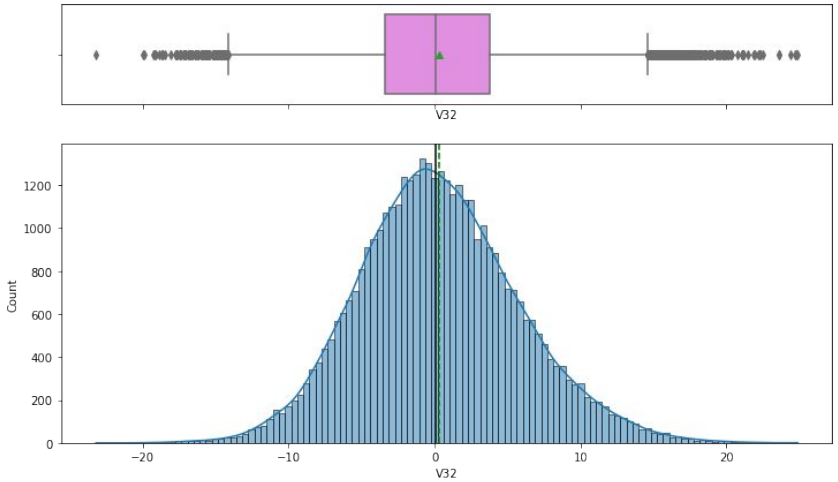
V31 Observations

- V31 has an approximately normal distribution
- Outliers in both directions
- Mean and median both positive



V32 Observations

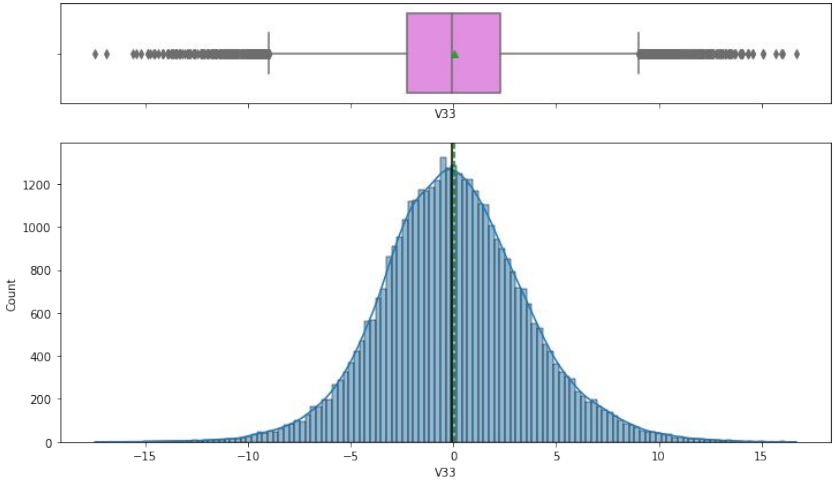
- V32 has a slight positive skew
- Outliers in both directions
- Mean and median positive, close to zero



Exploratory Data Analysis - V33 & V34

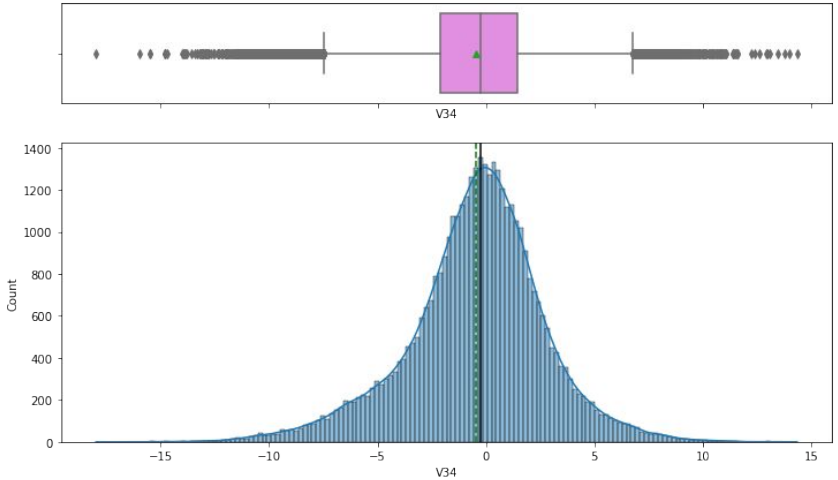
V33 Observations

- V33 has an approximately normal distribution
- Outliers in both directions
- Mean negative, median approximately zero



V34 Observations

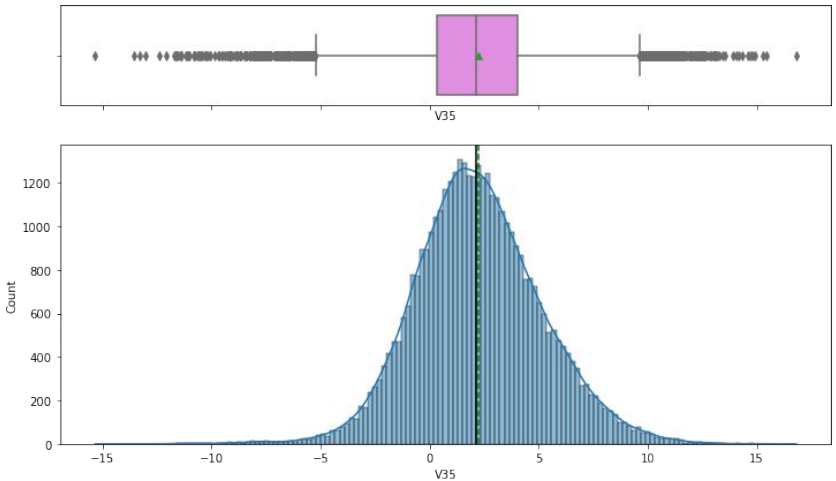
- V34 has a slight negative skew
- Outliers in both directions
- Mean and median negative



Exploratory Data Analysis - V35 & V36

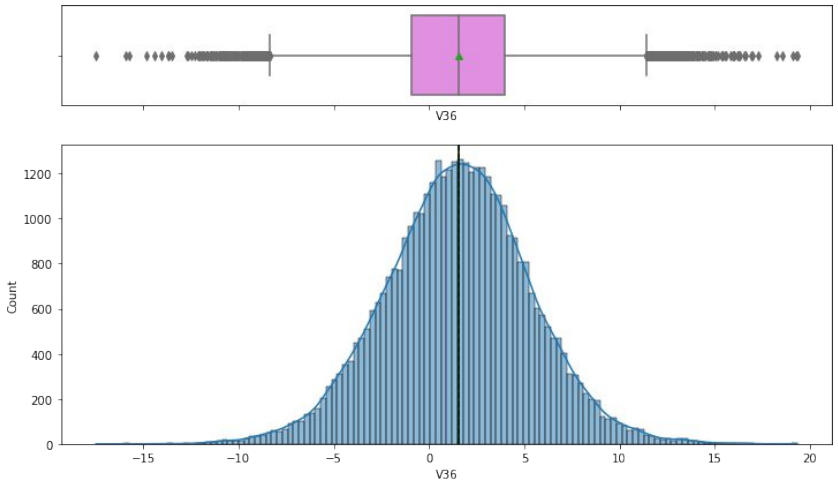
V35 Observations

- V35 has a slight positive skew
- Outliers in both directions
- Mean and median both positive



V36 Observations

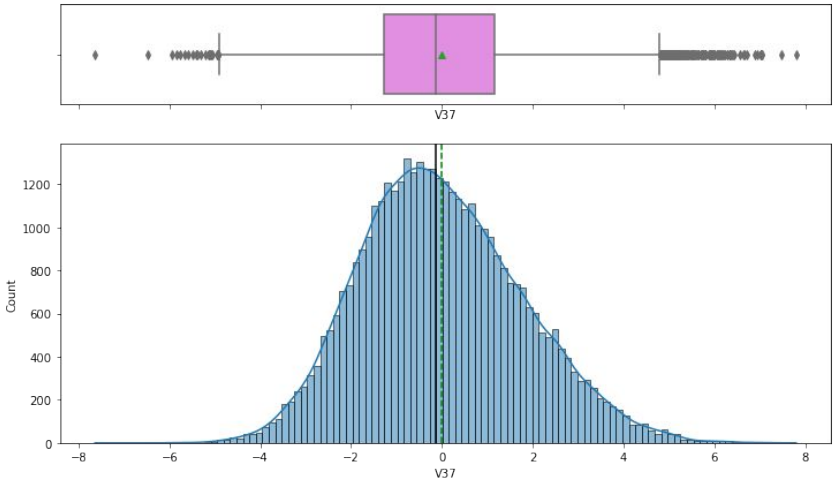
- V36 has an approximately normal distribution
- Outliers in both directions
- Mean and median are positive



Exploratory Data Analysis - V37 & V38

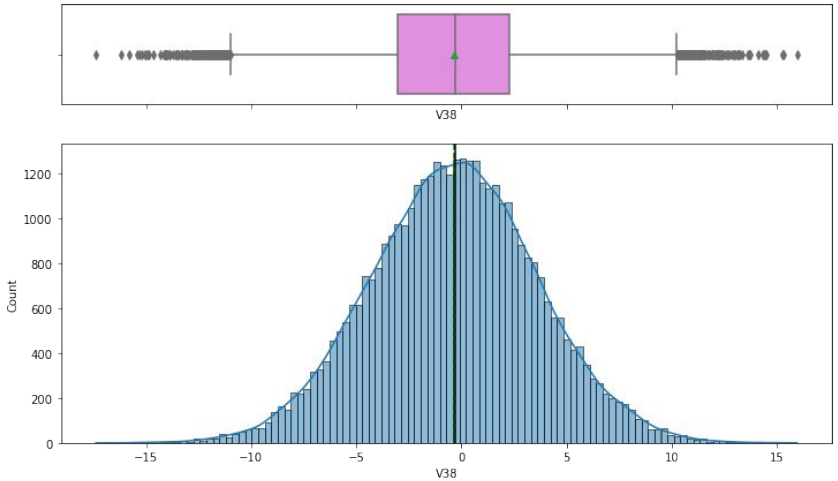
V37 Observations

- V37 has a slight positive skew
- Outliers in both directions
- Mean and median are negative



V38 Observations

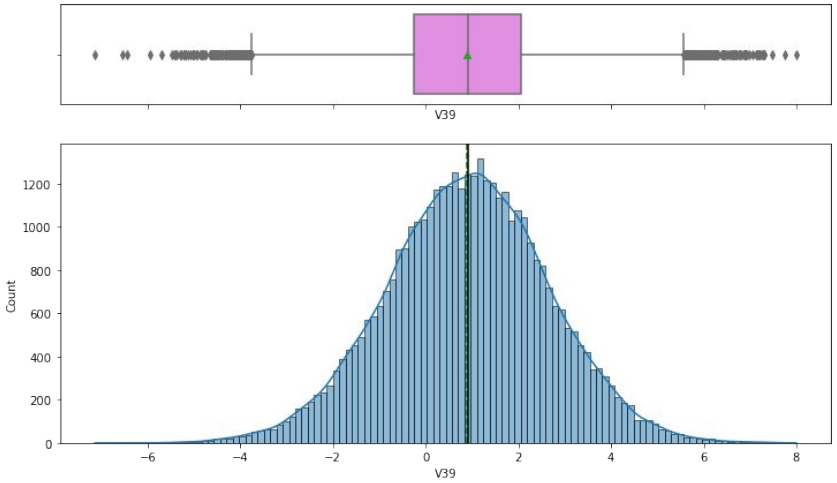
- V38 has an approximately normal distribution
- Outliers in both directions
- Mean and median are negative



Exploratory Data Analysis - V39 & V40

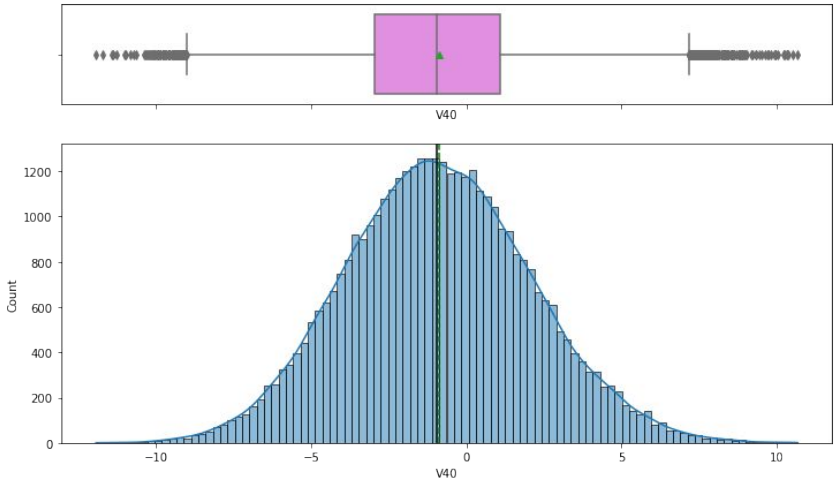
V39 Observations

- V39 has an approximately normal distribution
- Outliers in both directions
- Mean and median are positive



V40 Observations

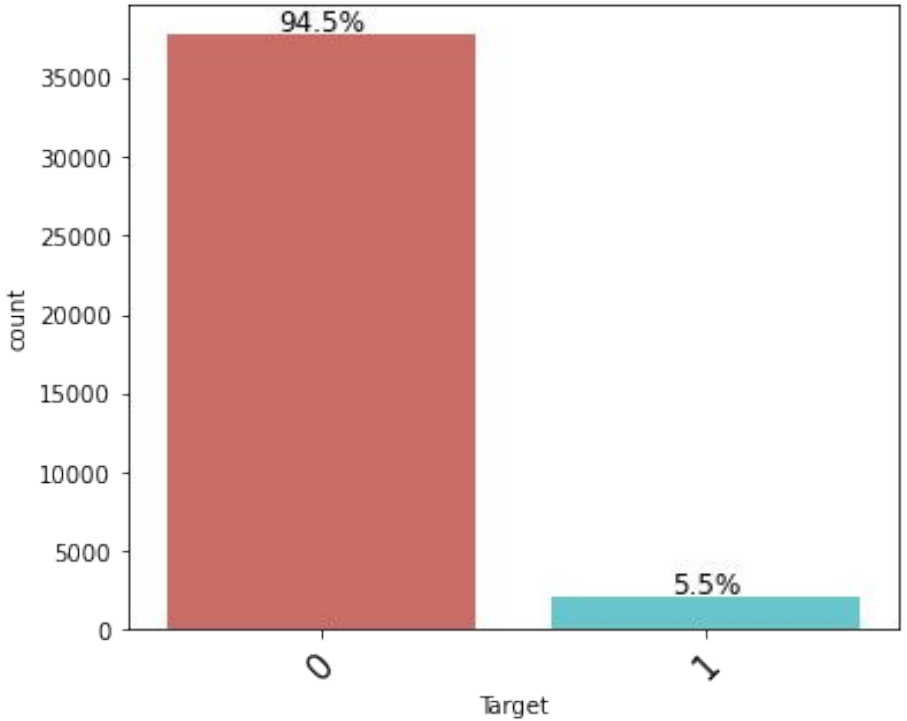
- V40 has an approximately normal distribution
- Outliers in both directions
- Mean and median are negative



Exploratory Data Analysis - Target Variable

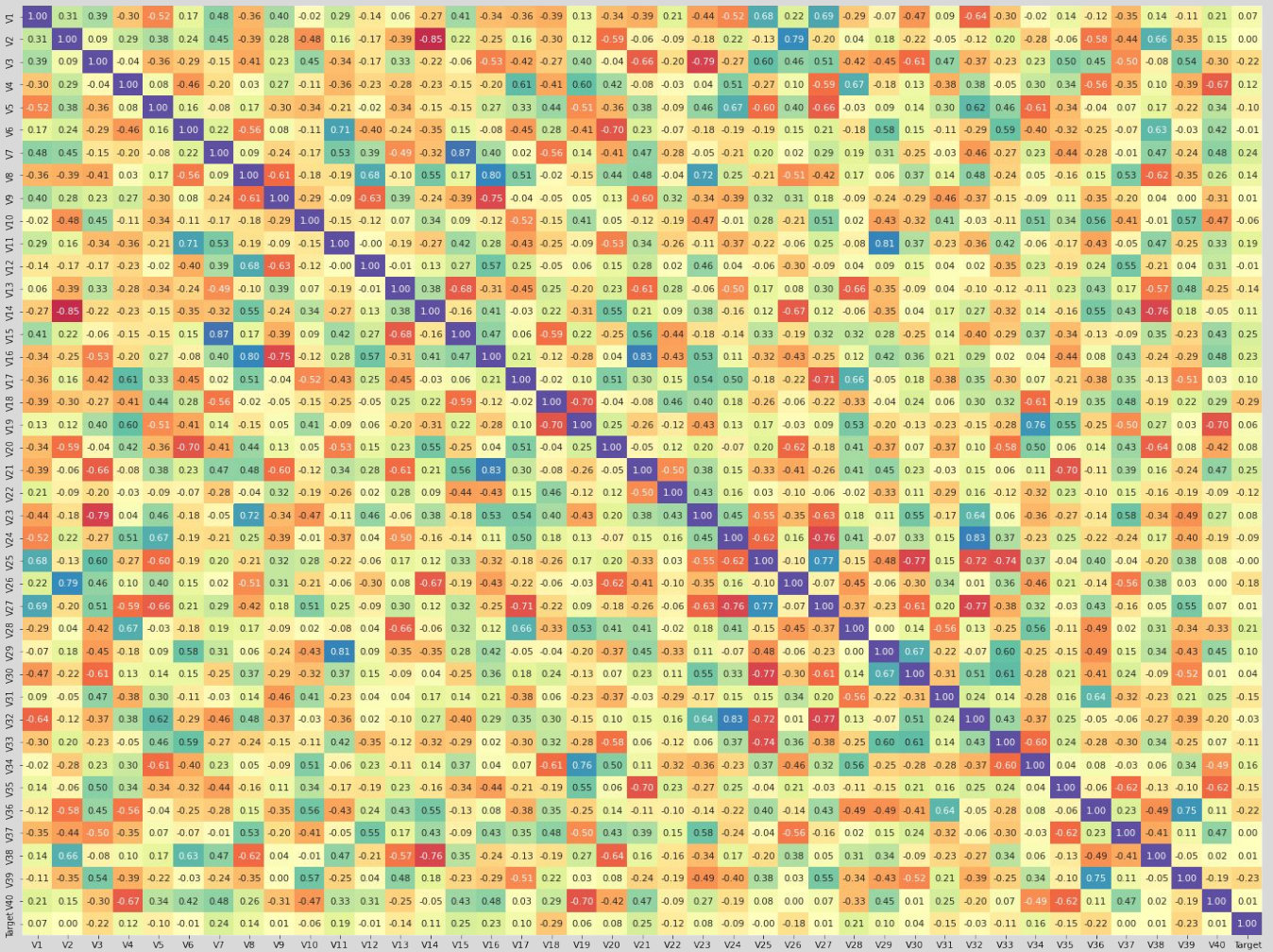
Target Variable Observations

- 94.5% no failure
- 5.5% failure



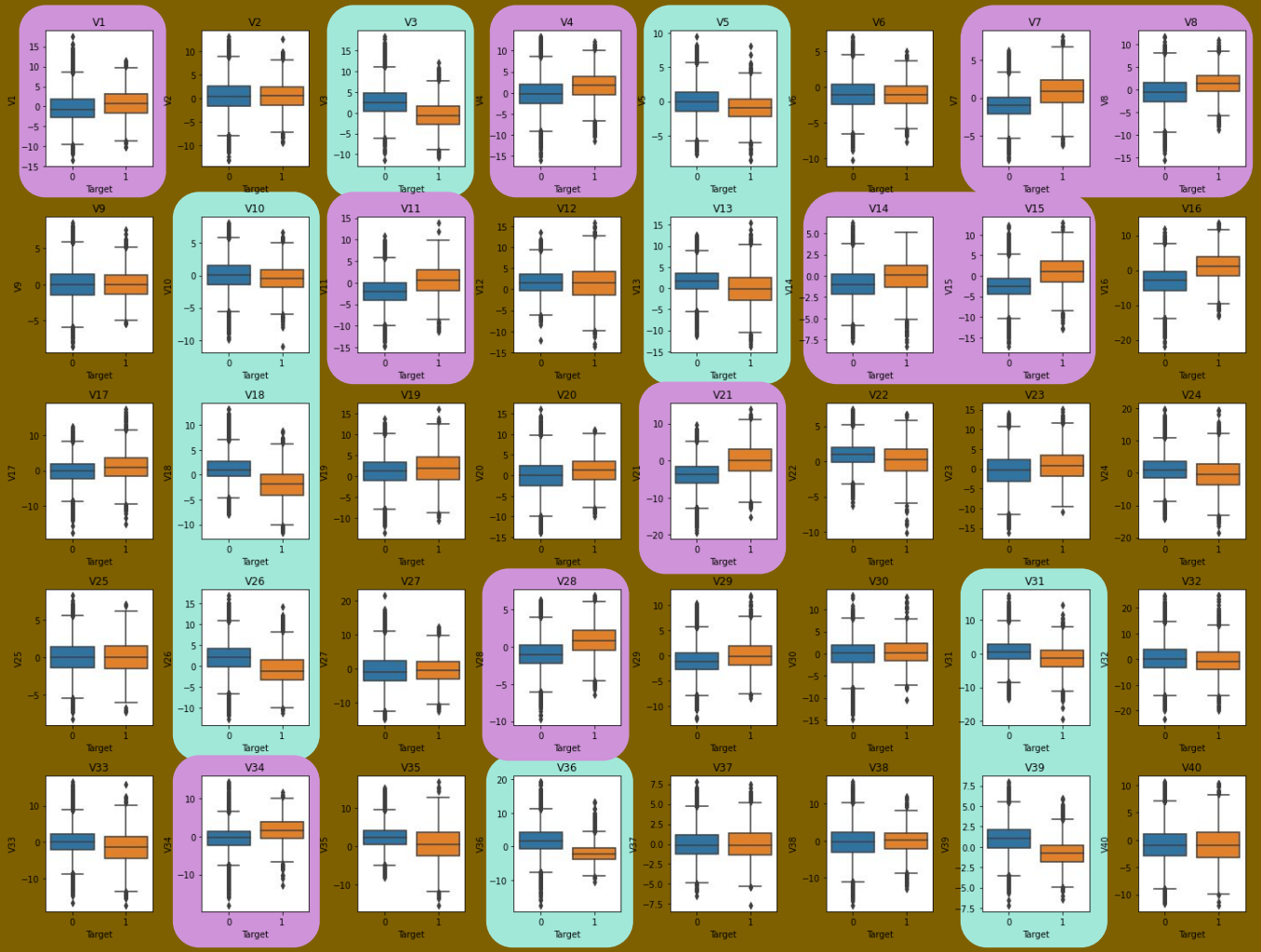
EDA - Correlation Heatmap

- V2 and V14 have strong negative correlation
- V7 and V15 have strong positive correlation
- V11 and V29 have strong positive correlation
- V24 and V32 have strong positive correlation



EDA: Bivariate Analysis - Variables vs Target

- Failure occurs for **higher** values of:
  - V1, V4, V7, V8, V11, V14, V15, V21, V28, V34
- Failure occurs for **lower** values of:
  - V3, V5, V10, V13, V18, V26, V31, V36, V39
- Failure occurs for **similar** values of:
  - V2, V6, V9, V12, V16, V17, V19, V20, V22, V23, V24, V25, V27, V29, V30, V32, V33, V35, V37, V38, V40



EDA Key Findings and Insights

- Some variables appear to be significant in identifying failure:
  - V3, V10, V15, V18, V21, V26, V36, V39, etc
- Build models to determine which variables significantly contribute to failure

Model Overview and Performance

Decision Tree Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	1.000	1.000	1.000	1.000	**1.000**
Validation Original	0.971	0.748	0.728	0.738	**0.661**
Training Upsampled	1.000	1.000	1.000	1.000	**1.000**
Validation Upsampled	0.951	0.814	0.535	0.645	**0.647**
Training Downsampled	1.000	1.000	1.000	1.000	**1.000**
Validation Downsampled	0.866	0.854	0.271	0.411	**0.498**

- Overfit on training set
- Perfect scores for training
- High accuracy scores
- Good recall scores
- Precision and F1 suffered considerably from upsampling & downsampling

Random Forest Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	1.000	1.000	1.000	1.000	**1.000**
Validation Original	0.987	0.766	0.991	0.864	**0.718**
Training Upsampled	1.000	1.000	1.000	1.000	**1.000**
Validation Upsampled	0.991	0.868	0.962	0.913	**0.812**
Training Downsampled	1.000	1.000	1.000	1.000	**1.000**
Validation Downsampled	0.966	0.885	0.639	0.742	**0.736**

- Overfit on training set
- Perfect scores for training
- Good scores for minimum_vs_model_cost (one of three highest)
- High accuracy scores
- Good recall and F1 scores
- Precision score suffered from downsampling

Bagging Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	0.997	0.944	0.999	0.971	**0.914**
Validation Original	0.984	0.735	0.962	0.833	**0.689**
Training Upsampled	0.999	0.998	1.000	0.999	**0.997**
Validation Upsampled	0.984	0.835	0.866	0.850	**0.759**
Training Downsampled	0.989	0.980	0.998	0.989	**0.967**
Validation Downsampled	0.951	0.863	0.529	0.656	**0.674**

- Overfit on training set
- Minimum_vs_model_cost score improved from upsampling
- Accuracy high
- Recall good
- Precision & F1 suffered from downsampling

AdaBoost Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	0.976	0.637	0.890	0.743	**0.613**
Validation Original	0.973	0.614	0.848	0.713	**0.595**
Training Upsampled	0.905	0.894	0.914	0.904	**0.830**
Validation Upsampled	0.905	0.850	0.350	0.496	**0.563**
Training Downsampled	0.906	0.894	0.916	0.905	**0.831**
Validation Downsampled	0.880	0.865	0.295	0.440	**0.523**

- Overfit on training set after upsampling & downsampling
- Minimum_vs_model_cost scores low on validation set
- High accuracy
- Recall improved after upsampling & downsampling
- Precision & F1 suffered from upsampling & downsampling

Gradient Boost Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	0.987	0.780	0.981	0.869	**0.729**
Validation Original	0.983	0.715	0.954	0.817	**0.673**
Training Upsampled	0.944	0.914	0.971	0.942	**0.868**
Validation Upsampled	0.966	0.881	0.636	0.739	**0.732**
Training Downsampled	0.952	0.918	0.985	0.950	**0.877**
Validation Downsampled	0.951	0.888	0.533	0.666	**0.692**

- Overfit on training set after upsampling & downsampling
- Minimum_vs_model_cost improved after upsampling & downsampling
- Accuracy high
- Recall improved after upsampling & downsampling
- Precision & F1 suffered after upsampling & downsampling

XGBoost Classifiers

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	1.000	1.000	1.000	1.000	**1.000**
Validation Original	0.990	0.826	0.987	0.900	**0.773**
Training Upsampled	0.999	0.999	0.999	0.999	**0.998**
Validation Upsampled	0.989	0.879	0.911	0.895	**0.813**
Training Downsampled	1.000	1.000	1.000	1.000	**1.000**
Validation Downsampled	0.969	0.901	0.656	0.760	**0.757**

- Overfit on training set
- Perfect scores or near perfect for training
- Two of three best minimum_vs_model_cost scores
- High accuracy
- Good recall
- Precision & F1 suffered from downsampling

Logistic Regression Models

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training Original	0.967	0.485	0.853	0.619	**0.530**
Validation Original	0.966	0.463	0.849	0.599	**0.520**
Training Upsampled	0.874	0.876	0.874	0.875	**0.800**
Validation Upsampled	0.873	0.839	0.279	0.419	**0.503**
Training Downsampled	0.859	0.855	0.862	0.859	**0.777**
Validation Downsampled	0.864	0.846	0.266	0.405	**0.492**

- Good at predicting no failure
- Not as good at predicting if there will be failure
- Low minimum_vs_model_cost score
- Low scores throughout

Model Selection

The three following models that performed the best according to the minimum_vs_model_cost metric and were not too overfit were:

- XGBoost on original data
- Random Forest on upsampled data
- XGBoost on upsampled data

The aforementioned models were hypertuned using RandomizedSearchCV to save time (GridSearchCV was taking over a day to run so the process was interrupted and RandomizedSearchCV was chosen for all models).

Model Performance Comparison

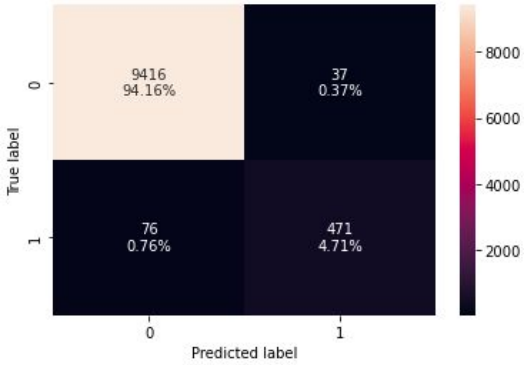
Although minimum_vs_model_cost was higher for tuned XGBoost classifier on original data, it is overfitting on the training data slightly more than the tuned random forest classifier on the upsampled data.

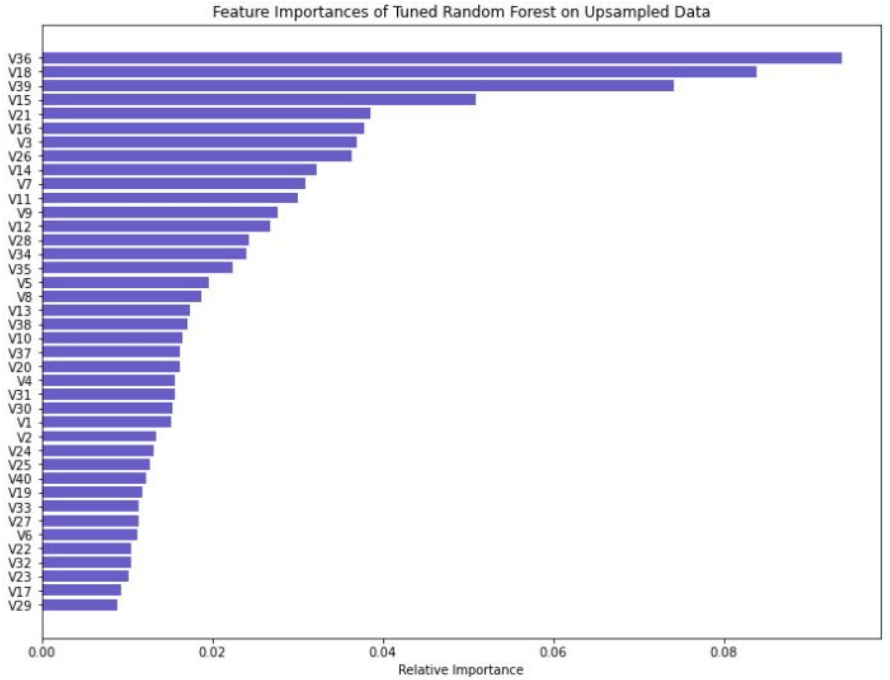
We choose the **tuned random forest classifier on the upsampled data** because it is slightly more generalizable

	Accuracy	Recall	Precision	F1	Minimum_vs_model_cost
Training - Tuned XGBoost (original data)	1.000	1.000	1.000	1.000	1.000
Validation - Tuned XGBoost (original data)	0.991	0.864	0.973	0.917	0.813
Training - Tuned XGBoost (upsampled data)	0.999	1.000	0.999	0.999	1.000
Validation - Tuned XGBoost (upsampled data)	0.985	0.887	0.841	0.863	0.803
Training - Tuned RF (upsampled data)	0.994	0.990	0.999	0.994	0.983
Validation - Tuned RF (upsampled data)	0.990	0.872	0.943	0.906	0.812

Model Test Set Final Performance

- V36 is the most important feature
- V18 is the next most important feature, followed by V39 and V15
- Test performance:
  - 0.797 minimum_vs_model_cost score
  - 0.989 accuracy
  - 0.861 recall
  - 0.927 precision
  - 0.893 F1





Business Recommendations & Conclusions

Top 3 important features of tuned Random Forest Classifier were also identified by tuned XGBoost Classifiers: **V36, V18, V39**

- Use feature importances to focus on most influential factors in that order
  - V36, V18, V39, V15, V21, etc
- Determine why these factors may be the most important
  - Consult domain expert / business stakeholder
- ReneWind may use pipeline / model to identify machinery that will likely fail soon, prior to breakage

Business Recommendations & Conclusions

Consult expert regarding missing data, feature importances, and outliers

- Both train and test files had missing data for V1 & V2. Systematic errors? Sensor errors? What is the cause?
- All influential factors are ciphered so we cannot make in-depth business recommendations without more information about factors
- Expert or business stakeholder may have more information regarding outliers, whether they are true outliers or valid data

Business Recommendations & Conclusions

GridSearchCV

Treat outliers

Hyperparameter tuning with more models

More EDA & more analysis

- If time allowed, we would have additionally tried
  - GridSearchCV - more extensive, systematic search for a model with potentially better results
  - Treat outliers before fitting the model to see if that would improve performance
  - Tune hyperparameters and use a more exhaustive search using more hyperparameters to search for a model with potentially better results
  - More bivariate/multivariate analysis