Classification Tool For Chemical Analysis

Namitha Jose(22044511)

2023-08-06

## Overview

The project aims to develop a classification tool for the company, evaluating whether all variables are necessary or if a subset can achieve comparable out-of-sample classification performance.

To achieve this goal, the data is checked for missing values and outliers, followed by exploratory data analysis to identify relationships and patterns.

The dataset is then divided into training, test, and validation sets, allocating 15% of observations to the validation set to ensure proper evaluation of final model’s performance on unseen data.

Five classification approaches (KNN, MclustDA, RF, LDA, and QDA) are evaluated, and optimal model is chosen based on accuracy,features involved,interpretability and processing time.

Finally, the performance of selected model is evaluated using validation data to ensure accurate classification.

## Data Summary

The analysis begins with data loading and basic data validation.

The given dataset comprises of 21 variables (20 numeric and 1 categorical) and 2500 records with 48 missing values. The missing values are attributed to 90.5% of variables.

# Missing data analysis

The dataset has 48 records with 1 missing value which is almost 2% of given data and there are 2452 complete records which represents 98% of data.

The distribution of missing values across different variables is given below.

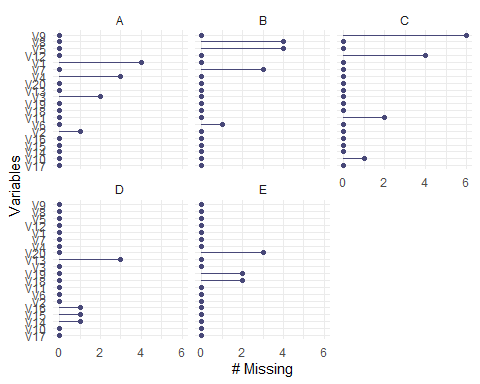
There are 2 variables with 0 missings, comprising 9.52% of variables in the dataset.

There are 6 variables with 1 missing, comprising 28.57% of variables in the dataset.

There are 4 variables with 2 missings,4 variables with 3 missings, 4 variables with4 missings and 1 variable with 6missings.

# Missing values by category.

| **V21** | **variable** | **n\_miss** | **pct\_miss** |
| --- | --- | --- | --- |
| A | V1 | 4 | 0.82 |
| A | V4 | 3 | 0.61 |
| A | V3 | 2 | 0.41 |
| A | V2 | 1 | 0.20 |
| B | V5 | 4 | 0.83 |
| B | V8 | 4 | 0.83 |
| B | V7 | 3 | 0.62 |
| B | V6 | 1 | 0.21 |
| C | V9 | 6 | 1.24 |
| C | V12 | 4 | 0.83 |
| C | V11 | 2 | 0.41 |
| C | V10 | 1 | 0.21 |
| D | V13 | 3 | 0.57 |
| D | V14 | 1 | 0.19 |
| D | V15 | 1 | 0.19 |
| D | V16 | 1 | 0.19 |
| E | V20 | 3 | 0.58 |
| E | V18 | 2 | 0.39 |
| E | V19 | 2 | 0.39 |



## Perform Statistical test for missingness - MCAR test

Missing data can present significant challenges in analysis, as it can lead to biased results. Hence it is important to determine the underlying mechanism of missingness. The Missing Completely at Random (MCAR) test is a statistical tool to assess whether missing data occur in a completely random manner, independent of any observed or unobserved variables.

In statistical terms, the null hypothesis for MCAR test is, missing values in the given data are missing completely at random. The alternate hypothesis is, given data is not missing completely at random.

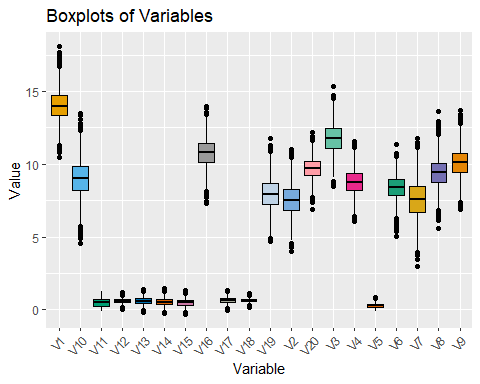
As per the statistical test the p-value is 0.059, which is greater than significance level of 0.05. Hence there is no sufficient evidence that the values are not missing completely at random and fail to reject the null hypothesis.

Hence the statistical test indicates that data is Missing Completely at Random (MCAR), suggesting that missing values are unrelated to both the observed and unobserved data in the given dataset.

In situations where data are MCAR, excluding the missing data is a feasible approach as it is not expected to introduce bias into analysis. Additionally, the proportion of missing data is minimal, accounting for less than 2% of total records. Hence, it has been decided to consider only the complete cases (2452 records) for further analysis, as it provides a sufficient subset of data to draw meaningful conclusions.

## Exploratory Data Analysis(EDA)

The primary objective of EDA is to gain initial insights from data, understand its distribution, and identify patterns or anomalies.



The boxplots displays distribution of all numeric variables.

For variables V11 and V5 the boxplots are asymmetrical and it indicates skewness (concentration of values towards one side) in distribution of data.

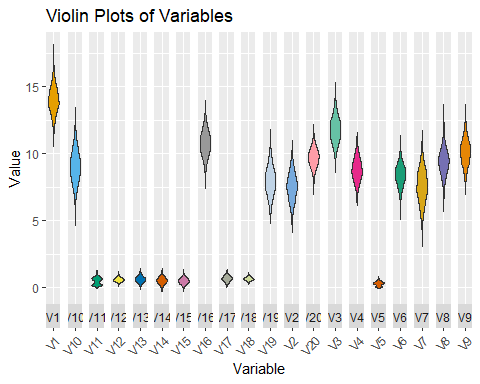
For variables V13, V14 and V15 the boxplots are almost overlapping indicating some relationship between variables.

The presence of extreme values or outliers in most variables can also be noted.

## Extreme Outliers

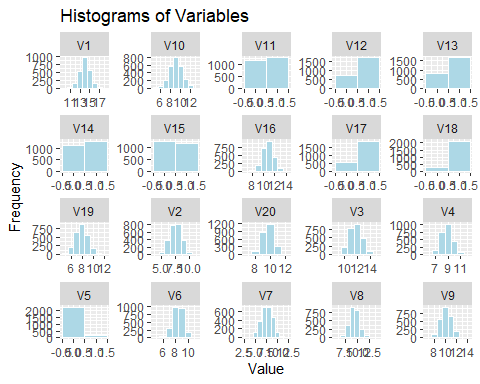
The values that are far away from rest of the data (Mahalanobis distance greater than 4 times the standard deviation) are considered as extreme outliers for each variable and are highlighted for further investigation.

The extreme outliers for each variable is avaialable in Appendix.



The Violin plots displays density distribution of numeric variables. For variables V11 and V5 it is evident that distribution is multimodal which means the data falls in multiple groups.

The histogram shows frequency distribution of each numeric variable. Most of the variables follow a normal distribution while few variables are bimodal.



## Feature Selection

### Descriptive Statistics by Category

The descriptive statistics helps to compare the impact of numeric variables on categorical variable.

In the dataset, the presence of negative values are identified in certain numeric variables, namely V5, V11, V13, V14, and V15. Below is a summary of the negative values:

1. Variable V5: Negative values are present for categories D and E
2. Variable V11: Negative values are present for categories A, C, and E
3. Variable V13: Negative values are present for categories B, C, D, and E
4. Variables V14 and V15: Negative values are present across all categories of the categorical variable.

It is important to investigate whether these negative values are valid data points or if they indicate data entry errors or other issues.

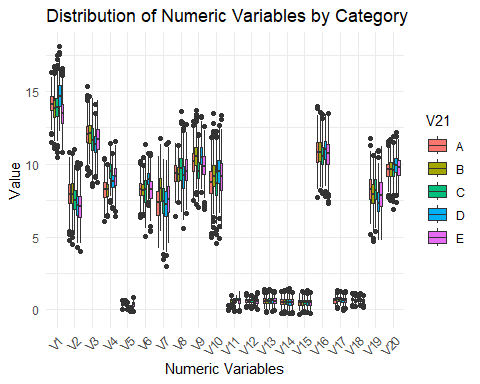
The detailed table is given in appendix.

### Number of samples by category

| V21 | Samplesize | Percentage |
| --- | --- | --- |
| A | 480 | 19.58 |
| B | 470 | 19.17 |
| C | 471 | 19.21 |
| D | 524 | 21.37 |
| E | 507 | 20.68 |

### Boxplots by category

The boxplots by category helps to understand the distribution of variables by category and to get an initial impression of whether there is a potential relationship between the numerical variables and categorical variable. It also guide towards feature selection.



The initial inferences are:

Variables V13, V14, and V15: The boxplots for these variables shows significant overlap between categories (A, B, C, D, and E), and the mean values are also similar. This suggests that these variables may not be very effective in distinguishing between the five categories. It is likely that these variables have similar distributions across all categories, indicating a weak or no relationship between the categorical variable (V21) and these numeric variables (V13, V14, and V15).

Variables V5 and V11: The plot shows distinct and non-overlapping boxes for different categories (A, B, C, D, and E). Additionally, the mean values for these variables vary between categories.This suggests that these variables have different distributions across the five categories, indicating a potential relationship or dependency between the categorical variable (V21) and numeric variables (V5 and V11).

Individual boxplots are available in appendix.

Further statistical analysis is needed to validate and quantify the relationship between the categorical variable and the numeric variables. In order to do that ANOVA is used.

## ANOVA (Analysis of Variance)

ANOVA is a statistical technique used to compare the means of two or more groups to determine if there are any significant differences among them.

The goal of ANOVA is to test the null hypothesis that the means of all the groups are equal, against the alternative hypothesis that at least one group’s mean is different from the others. If the p-value from the ANOVA test is below a chosen significance level (e.g., 0.05), you can reject the null hypothesis, indicating that there is a significant difference in means among the groups.

Here One way ANOVA is used.

One-Way ANOVA: This is used when we have one categorical independent variable with two or more groups and one continuous dependent variable. It tests for differences in means among the groups.

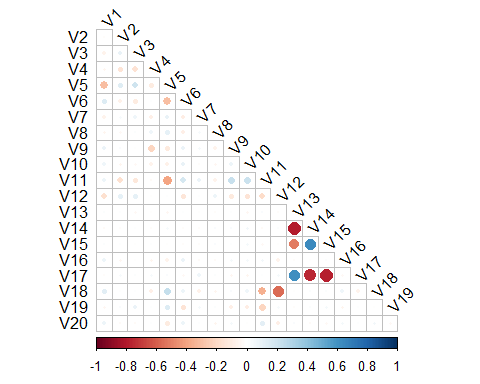
The ANOVA results shows that for:

Variables V11 and V5 : The F-statistic is high and p value is less than 0.05. Hence there is significant evidence to reject the null hypothesis, indicating that there is significant difference in means among the 5 categories for these variables. This suggests that these variables are important for classifying the data into the five categories. Based on ANOVA the top 10 important variables include V4,V1,V6,V9,V2,V12,V18 and V7.

Variables V13, V14 and V15: The F-statistic is very minimal and p value is greater than 0.05. Hence we fail to reject the null hypothesis, indicating that there is no significant difference in means among the 5 categories for each of these variables. This implies that these variables may not be helpful in classifying the data into the five categories.

As a further analysis for feature selection the variable importance will be calculated using random-forest modeling.

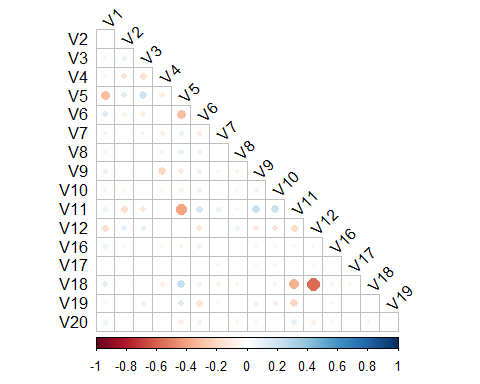
# Assessing the relationship between variables



The correlation plot visualizes the relationships between numeric variables and helps us assess potential multicollinearity issues. Multicollinearity refers to situation where two or more variables are highly correlated with each others, which can create challenges.

The plot indicates that variables V13,V14 and V15 are correlated with other variables.

## Correlation plot after removing least influential variables V13,V14 and V15



Now, the correlation plot shows that numeric variables have relatively low correlation with each other. This is a positive finding as it indicates that multicollinearity is minimal among variables.

## Classification Modelling

### Splitting Dataset

The dataset is divided into training (60% data,1471 records) for training the model , test (25% data,613 records)for testing the performance of various models and validation (15% data,368 records) for evaluating the best model.

## Calculate the variable importance using random forest

Variable importance measures the relative contribution of each predictor variable (feature) in a classification model. It helps to identify which features have the most significant impact on the model’s predictive performance.

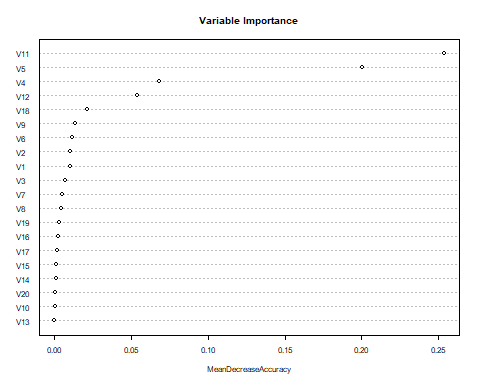
To determine the variable importance the MeanDecreaseGini values are used, higher values indicate more important variables for classification, while lower values suggest less influential variables.

Variables V13,V14,V15 have lowest MeanDecreaseGini indicating that they are least influential.

Variables V11 and V5 have highest MeanDecreaseGini suggesting that they are important variables for classification.

These findings are consistent with observations made during EDA and ANOVA.

The plot for variable importance is given below.



# Classification Model Performance Comparison

Five classification models (K-Nearest Neighbours (KNN), Model based Discriminant analysis (MclustDA), Random forest (RF), Linear discriminant analysis (LDA), and Quadratic discriminant analysis (QDA)) are trained using training dataset. Predictions are made on the test data using each model, and the accuracy was calculated by comparing the predicted classes with the actual classes in the test data. All models are trained using training data with all features, and additional experiments are performed by removing the least influential variables and also using principal components.

The model performance comparison is shown below:

| ModelName | K | Accuracy | Variables | AccuracyPct |
| --- | --- | --- | --- | --- |
| KNN | 29 | 0.5856444 | ALL | 58.56 |
| KNN | 24 | 0.5921697 | RemovedV13,V14,V15 | 59.22 |
| MclustDA | 0 | 0.9168026 | ALL | 91.68 |
| MclustDA | 0 | 0.9070147 | Removed V13,V14,V15 | 90.70 |
| MclustDA | 0 | 0.9233279 | Removed least 7 Var | 92.33 |
| MclustDA | 0 | 0.9282219 | Removed least 9 Var | 92.82 |
| MclustDA | 0 | 0.7014682 | PCA components | 70.15 |
| MclustDA | 0 | 0.9216966 | Removed least 10 Var | 92.17 |
| RF | 0 | 0.8499184 | ALL | 84.99 |
| RF | 0 | 0.8678630 | Removed V13,V14,V15 | 86.79 |
| LDA | 0 | 0.8156607 | ALL | 81.57 |
| LDA | 0 | 0.8107667 | Removed V13,V14,V15 | 81.08 |
| LDA | 0 | 0.6802610 | PCA components | 68.03 |
| QDA | 0 | 0.9347471 | ALL | 93.47 |
| QDA | 0 | 0.9282219 | Removed V13,V14,V15 | 92.82 |
| QDA | 0 | 0.9282219 | Removed 9 variables | 92.82 |
| QDA | 0 | 0.7063622 | PCA components | 70.64 |

The highest accuracy is observed for QDA model(93.47%), followed by MClustDA (92.8%), Random Forest (86.7%), LDA (81.5%), and KNN (59.2%). The accuracy reduced drastically for all models when pca is applied. After removing 9 least influential variables (V13,V14,V20,V15,V10,V19,V16,V3 and V17), QDA’s accuracy slightly reduced to 92.8%, while MClustDA’s increased to the same value. By using a subset of 11 features, the QDA model predicted without significant accuracy loss (0.6% increase when using all variables). The QDA model offers better interpretability and faster processing time compared to MClustDA.

Hence after considering interpretability, accuracy, and processing time, the QDA model with a subset of features(top 11 features from Variable Importance chart) is recommended for the data, and its performance will be further evaluated on the validation dataset.

## Confusion Matrix for Validation Data

The confusion matrix shows number of correct and incorrect predictions made by the model for each class. It has rows representing true classes and columns representing predicted classes. Below is the confusion matrix for the prediction made by the model on validation data:

predicted  
Actual A B C D E  
 A 66 0 4 0 0  
 B 0 72 0 0 2  
 C 3 0 67 0 2  
 D 0 0 0 85 1  
 E 0 2 3 0 61

Accuracy: 95.38 %

The selected QDA model achieves an accuracy of 95.38% on unseen data, showing its strong predictive capability. This indicates that the model is capable of correctly predicting the category for 95.38% of cases in the validation set. The 2.56% increase in accuracy compared to the test data indicates the model’s reliability and generalizability, and it confirms that the model’s predictions are not overfit to the training data. This demonstrates the model’s suitability for making accurate predictions on new data.

Overall, the QDA model with feature selection(V11,V5,V4,V12,V18,V9,V1,V6,V2,V8 and V7) appears to be the optimal choice for accurate classification.

## Appendix

## External Validation

### Confusion Matrix and Accuracy

predicted  
Actual A B C D E  
 A 467 0 28 0 0  
 B 0 491 1 0 21  
 C 21 0 435 0 17  
 D 0 0 0 459 6  
 E 1 26 22 3 452

Accuracy: 0.9404082

The model achieved an accuracy of 94% on a new dataset. There is no much accuracy loss and hence the selected model can be considered as an optimal one.

### Descriptive Statistics

| **variable** | **V21** | **n** | **mean** | **median** | **sd** | **min** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| V1 | A | 480 | 14.14 | 14.08 | 0.77 | 11.49 | 16.31 |
| B | 470 | 13.83 | 13.82 | 0.94 | 11.07 | 16.71 |
| C | 471 | 13.93 | 13.89 | 0.97 | 10.49 | 17.48 |
| D | 524 | 14.71 | 14.68 | 1.00 | 12.23 | 18.12 |
| E | 507 | 13.47 | 13.48 | 1.02 | 10.79 | 16.61 |
| V2 | A | 480 | 7.91 | 7.91 | 0.96 | 4.76 | 10.80 |
| B | 470 | 7.95 | 7.91 | 1.01 | 4.48 | 10.75 |
| C | 471 | 7.57 | 7.55 | 0.98 | 5.09 | 11.01 |
| D | 524 | 7.15 | 7.17 | 0.98 | 4.30 | 10.16 |
| E | 507 | 7.11 | 7.09 | 1.00 | 4.03 | 10.05 |
| V3 | A | 480 | 12.08 | 12.08 | 0.92 | 9.23 | 15.32 |
| B | 470 | 12.10 | 12.13 | 0.91 | 9.37 | 14.64 |
| C | 471 | 11.66 | 11.64 | 1.09 | 8.51 | 14.53 |
| D | 524 | 11.38 | 11.36 | 0.77 | 9.10 | 14.11 |
| E | 507 | 11.71 | 11.71 | 0.99 | 8.77 | 14.37 |
| V4 | A | 480 | 8.26 | 8.27 | 0.77 | 6.07 | 10.38 |
| B | 470 | 8.23 | 8.24 | 0.63 | 6.45 | 9.98 |
| C | 471 | 9.50 | 9.48 | 0.70 | 7.38 | 11.46 |
| D | 524 | 8.83 | 8.82 | 0.65 | 6.97 | 10.73 |
| E | 507 | 9.13 | 9.15 | 0.85 | 6.45 | 11.56 |
| V5 | A | 480 | 0.35 | 0.35 | 0.11 | 0.01 | 0.66 |
| B | 470 | 0.37 | 0.37 | 0.09 | 0.16 | 0.64 |
| C | 471 | 0.26 | 0.26 | 0.08 | 0.01 | 0.53 |
| D | 524 | 0.05 | 0.05 | 0.05 | -0.11 | 0.18 |
| E | 507 | 0.38 | 0.38 | 0.17 | -0.10 | 0.86 |
| V6 | A | 480 | 8.27 | 8.28 | 0.63 | 6.32 | 10.22 |
| B | 470 | 8.21 | 8.19 | 0.64 | 6.49 | 10.44 |
| C | 471 | 8.16 | 8.28 | 1.02 | 5.03 | 11.38 |
| D | 524 | 9.00 | 8.97 | 0.60 | 7.14 | 10.75 |
| E | 507 | 8.23 | 8.25 | 0.86 | 5.40 | 10.71 |
| V7 | A | 480 | 7.34 | 7.35 | 1.18 | 4.23 | 10.50 |
| B | 470 | 8.29 | 8.31 | 1.15 | 5.38 | 11.77 |
| C | 471 | 7.48 | 7.47 | 1.24 | 3.44 | 10.35 |
| D | 524 | 7.30 | 7.26 | 1.21 | 2.97 | 11.38 |
| E | 507 | 7.55 | 7.57 | 1.28 | 4.56 | 11.50 |
| V8 | A | 480 | 9.35 | 9.35 | 0.75 | 6.42 | 11.45 |
| B | 470 | 9.30 | 9.32 | 0.71 | 7.37 | 11.31 |
| C | 471 | 9.69 | 9.73 | 1.17 | 6.69 | 13.64 |
| D | 524 | 9.13 | 9.22 | 1.15 | 5.58 | 12.77 |
| E | 507 | 9.60 | 9.52 | 1.07 | 6.59 | 12.72 |
| V9 | A | 480 | 10.14 | 10.18 | 0.92 | 7.50 | 12.92 |
| B | 470 | 10.54 | 10.51 | 0.94 | 7.27 | 13.69 |
| C | 471 | 9.54 | 9.48 | 0.87 | 7.00 | 12.02 |
| D | 524 | 10.55 | 10.48 | 0.95 | 8.05 | 13.28 |
| E | 507 | 9.86 | 9.85 | 0.95 | 6.91 | 12.60 |
| V10 | A | 480 | 8.72 | 8.73 | 1.20 | 4.97 | 11.81 |
| B | 470 | 9.20 | 9.14 | 1.20 | 5.25 | 13.48 |
| C | 471 | 8.64 | 8.65 | 1.16 | 4.58 | 12.25 |
| D | 524 | 9.49 | 9.51 | 1.21 | 4.93 | 12.68 |
| E | 507 | 9.16 | 9.18 | 1.30 | 5.55 | 13.37 |
| V11 | A | 480 | 0.15 | 0.15 | 0.05 | -0.02 | 0.31 |
| B | 470 | 0.62 | 0.63 | 0.12 | 0.19 | 1.00 |
| C | 471 | 0.26 | 0.26 | 0.11 | -0.09 | 0.62 |
| D | 524 | 0.71 | 0.72 | 0.10 | 0.45 | 0.97 |
| E | 507 | 0.63 | 0.62 | 0.25 | -0.14 | 1.28 |
| V12 | A | 480 | 0.62 | 0.62 | 0.15 | 0.16 | 1.12 |
| B | 470 | 0.60 | 0.60 | 0.13 | 0.24 | 0.93 |
| C | 471 | 0.63 | 0.63 | 0.13 | 0.11 | 1.02 |
| D | 524 | 0.49 | 0.49 | 0.09 | 0.23 | 0.83 |
| E | 507 | 0.59 | 0.60 | 0.19 | 0.01 | 1.19 |
| V13 | A | 480 | 0.61 | 0.60 | 0.23 | 0.04 | 1.41 |
| B | 470 | 0.60 | 0.60 | 0.24 | -0.20 | 1.35 |
| C | 471 | 0.60 | 0.59 | 0.24 | -0.05 | 1.37 |
| D | 524 | 0.63 | 0.63 | 0.23 | -0.10 | 1.29 |
| E | 507 | 0.61 | 0.60 | 0.22 | -0.06 | 1.26 |
| V14 | A | 480 | 0.52 | 0.53 | 0.25 | -0.24 | 1.23 |
| B | 470 | 0.53 | 0.52 | 0.26 | -0.28 | 1.32 |
| C | 471 | 0.53 | 0.52 | 0.27 | -0.23 | 1.37 |
| D | 524 | 0.51 | 0.51 | 0.28 | -0.26 | 1.44 |
| E | 507 | 0.52 | 0.53 | 0.25 | -0.28 | 1.18 |
| V15 | A | 480 | 0.49 | 0.50 | 0.24 | -0.33 | 1.14 |
| B | 470 | 0.48 | 0.48 | 0.22 | -0.16 | 1.28 |
| C | 471 | 0.50 | 0.50 | 0.24 | -0.21 | 1.18 |
| D | 524 | 0.49 | 0.48 | 0.23 | -0.18 | 1.35 |
| E | 507 | 0.48 | 0.48 | 0.23 | -0.27 | 1.20 |
| V16 | A | 480 | 10.83 | 10.82 | 0.97 | 7.75 | 13.98 |
| B | 470 | 10.83 | 10.84 | 0.96 | 8.17 | 13.56 |
| C | 471 | 10.58 | 10.55 | 0.92 | 7.99 | 13.43 |
| D | 524 | 10.98 | 11.02 | 1.01 | 7.76 | 13.42 |
| E | 507 | 10.68 | 10.71 | 1.02 | 7.33 | 13.50 |
| V17 | A | 480 | 0.62 | 0.63 | 0.22 | 0.06 | 1.24 |
| B | 470 | 0.70 | 0.70 | 0.20 | 0.11 | 1.36 |
| C | 471 | 0.70 | 0.71 | 0.21 | 0.20 | 1.27 |
| D | 524 | 0.63 | 0.63 | 0.21 | -0.01 | 1.22 |
| E | 507 | 0.63 | 0.63 | 0.20 | 0.07 | 1.29 |
| V18 | A | 480 | 0.71 | 0.72 | 0.16 | 0.25 | 1.13 |
| B | 470 | 0.61 | 0.61 | 0.15 | 0.22 | 1.14 |
| C | 471 | 0.62 | 0.62 | 0.08 | 0.40 | 0.93 |
| D | 524 | 0.68 | 0.68 | 0.14 | 0.30 | 1.13 |
| E | 507 | 0.61 | 0.61 | 0.14 | 0.17 | 1.01 |
| V19 | A | 480 | 8.26 | 8.24 | 1.01 | 5.17 | 11.79 |
| B | 470 | 7.93 | 7.94 | 1.00 | 4.69 | 10.63 |
| C | 471 | 8.25 | 8.23 | 1.01 | 5.39 | 10.85 |
| D | 524 | 7.43 | 7.34 | 1.01 | 4.77 | 10.98 |
| E | 507 | 7.91 | 7.88 | 1.20 | 4.79 | 11.20 |
| V20 | A | 480 | 9.61 | 9.63 | 0.66 | 7.69 | 11.57 |
| B | 470 | 9.61 | 9.63 | 0.73 | 7.54 | 11.72 |
| C | 471 | 9.58 | 9.60 | 0.69 | 6.88 | 11.91 |
| D | 524 | 9.93 | 9.90 | 0.72 | 7.38 | 12.18 |
| E | 507 | 9.75 | 9.81 | 0.73 | 7.74 | 11.70 |

### Extreme Outliers list

$V1  
 [1] 502 544 577 658 666 792 980 1095 1111 1229 1264 1292 1297 1398 1422  
[16] 1463 1530 1624 1630 1640 1655 1702 1713 1714 1789 1805 1810 1821 1824 1885  
[31] 1936 1969 1977 2015 2035 2038 2040 2160 2199 2228 2342 2362 2365 2366 2389  
  
$V2  
 [1] 275 316 356 429 445 493 595 604 896 909 912 921 930 943 1251  
[16] 1387 1428 1448 1459 1581 1647 1821 1841 1846 1880 1886 1901 1995 2112 2140  
[31] 2221 2242 2337 2359 2448  
  
$V3  
 [1] 28 83 189 207 220 241 373 405 427 437 544 572 592 662 680  
[16] 759 779 802 821 969 1025 1052 1067 1077 1102 1105 1147 1187 1204 1225  
[31] 1231 1329 1368 1433 1620 1655 1663 1759 2138 2176 2193 2206 2347 2452  
  
$V4  
 [1] 16 24 37 67 149 202 214 227 252 373 432 437 441 598 716  
[16] 760 820 853 976 1022 1026 1101 1103 1117 1130 1184 1211 1214 1234 1311  
[31] 1342 1967 2000 2040 2065 2079 2147 2169 2181 2223 2244 2308 2316 2422  
  
$V5  
 [1] 233 1843 1946 2011 2027 2040 2043 2059 2062 2080 2084 2089 2116 2121 2132  
[16] 2153 2173 2179 2185 2230 2238 2244 2321 2330 2339 2358 2362 2369 2377 2391  
[31] 2407 2430 2433 2440  
  
$V6  
 [1] 345 535 953 959 974 994 1005 1022 1027 1035 1077 1108 1117 1130 1134  
[16] 1157 1162 1198 1231 1253 1268 1299 1318 1319 1323 1330 1333 1392 1408 1638  
[31] 1705 1840 1938 1956 1977 2006 2037 2065 2079 2085 2283 2378 2410 2423 2447  
  
$V7  
 [1] 45 159 246 460 470 487 493 541 560 565 578 595 652 760 777  
[16] 782 879 917 928 934 1028 1035 1128 1263 1526 1592 1606 1724 1830 1852  
[31] 1903 1990 1997 2000 2003 2105 2122 2198 2254 2291  
  
$V8  
 [1] 461 965 978 999 1026 1132 1133 1143 1182 1212 1230 1252 1302 1328 1339  
[16] 1381 1413 1459 1462 1503 1522 1524 1532 1537 1553 1654 1701 1708 1762 1780  
[31] 1793 1801 1866 1869 1880 1930 1975 2049 2055 2138 2249 2256 2287 2293 2322  
[46] 2381  
  
$V9  
 [1] 89 98 242 298 395 591 625 631 652 698 764 868 872 878 882  
[16] 899 947 966 977 1016 1028 1083 1130 1294 1312 1427 1491 1500 1544 1581  
[31] 1783 1808 1895 1896 1992 2115 2191 2202 2275 2359 2373 2376 2417 2448  
  
$V10  
 [1] 12 37 210 379 412 419 505 508 564 668 730 755 771 880 984  
[16] 987 1065 1191 1192 1255 1259 1278 1400 1432 1470 1472 1492 1494 1561 1668  
[31] 1707 1790 1813 1996 2001 2016 2054 2083 2117 2155 2207 2249 2364 2386 2427  
  
$V11  
 [1] 1228 1378 1954 1971 1977 1987 2033 2064 2090 2099 2112 2125 2130 2145 2152  
[16] 2154 2172 2184 2198 2208 2211 2231 2249 2251 2279 2355 2356 2364 2376 2378  
[31] 2380 2393 2446 2450  
  
$V12  
 [1] 130 155 184 185 186 245 382 388 455 1011 1102 1279 1410 1947 1953  
[16] 1954 1984 2030 2047 2055 2062 2096 2104 2128 2151 2161 2176 2207 2211 2223  
[31] 2244 2267 2290 2305 2342 2347 2356 2362 2436 2440 2444  
  
$V13  
 [1] 29 244 465 479 498 511 530 559 615 682 697 726 775 817 852  
[16] 970 973 1043 1145 1227 1229 1232 1282 1403 1477 1493 1528 1545 1576 1626  
[31] 1632 1675 1858 1927 1988 1992 2094 2116 2266 2402 2409  
  
$V14  
 [1] 29 181 417 437 445 530 602 654 678 684 726 775 800 839 970  
[16] 1011 1045 1051 1084 1103 1119 1262 1286 1316 1396 1464 1493 1545 1565 1576  
[31] 1623 1660 1661 1663 1675 1738 1886 1905 1917 1919 1935 1981 1988 1992 2048  
[46] 2113 2305 2438  
  
$V15  
 [1] 20 27 43 69 98 267 330 371 538 602 678 698 970 1017 1039  
[16] 1044 1045 1067 1116 1173 1179 1326 1340 1385 1420 1563 1608 1660 1675 1693  
[31] 1717 1786 1871 1880 1905 2046 2048 2283 2351 2438  
  
$V16  
 [1] 4 61 75 89 109 141 204 222 240 289 408 421 490 571 633  
[16] 668 746 794 824 865 950 1148 1166 1232 1328 1429 1457 1486 1515 1524  
[31] 1566 1616 1687 1734 1757 1810 1848 1928 2025 2109 2133 2147 2239 2258 2283  
[46] 2338 2404 2420 2434 2446  
  
$V17  
 [1] 19 20 48 56 69 224 244 273 275 361 430 453 530 602 654  
[16] 656 734 800 970 1111 1145 1160 1316 1333 1396 1463 1503 1509 1603 1646  
[31] 1656 1660 1663 1669 1675 1717 1779 1798 1905 1912 1927 2048 2086 2137 2283  
[46] 2342 2402 2409  
  
$V18  
 [1] 20 21 26 71 83 98 123 130 155 170 182 185 225 233 245  
[16] 273 274 360 392 405 523 590 698 706 759 774 881 882 887 1452  
[31] 1754 1843 1916 2045 2171 2195 2198 2207 2213 2254 2354 2357 2413 2440  
  
$V19  
 [1] 93 99 112 130 132 253 303 332 359 379 645 646 767 828 927  
[16] 994 1045 1181 1281 1421 1438 1501 1746 1887 1904 1930 1946 1949 1989 2074  
[31] 2098 2132 2295 2302 2306 2357 2370 2401  
  
$V20  
 [1] 173 189 316 478 490 575 634 654 718 753 773 791 795 1052 1056  
[16] 1083 1118 1154 1182 1223 1243 1317 1344 1486 1528 1580 1697 1717 1722 1729  
[31] 1741 1783 1924 1934 1973 1996 2018 2075 2329 2334 2397 2451

### Variable Importance based on meanDecreaseGini

V11 V5 V4 V12 V18 V9 V6 V1   
284.71271 246.02900 113.99409 61.46026 48.23010 45.04854 43.68995 42.10865   
 V2 V7 V8 V3 V19 V17 V10 V16   
 36.51510 30.71486 30.19591 28.71354 26.70593 23.52451 21.77663 20.55815   
 V20 V14 V15 V13   
 19.98343 17.88709 17.82503 15.96831

### Individual boxplots for variables V13,V4,V15,V5 and V11

