Contents

[DSAT Model Building 2](#_Toc491940611)

[Code Flow 2](#_Toc491940612)

[Data Preparation 2](#_Toc491940613)

[Discrete Variables 2](#_Toc491940614)

[Continuous Variables 2](#_Toc491940615)

[Modelling 2](#_Toc491940616)

[Classification Models 2](#_Toc491940617)

[Generative models 2](#_Toc491940618)

[Discriminative models 3](#_Toc491940619)

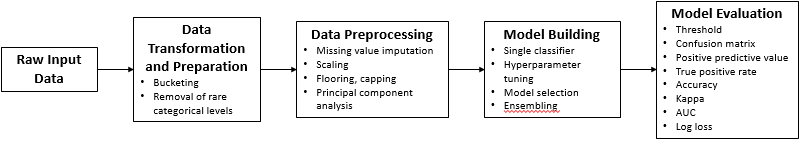
[Parameter Tuning 3](#_Toc491940620)

[Ensembling 3](#_Toc491940621)

[Model Evaluation 3](#_Toc491940622)

# DSAT Model Building

## Code Flow



## Data Preparation

### Discrete Variables

Variables have multiple levels with varying relative frequencies. Modeling with levels that have low relative frequencies (especially with cardinality of 1) leads to overfitting. Also, if a level is present only in one of training or testing set, the model built may not be stable. The following processing is done on categorical variables:

1. Removed variables with only 1 level (0 variation)
2. Removed rows with levels that occur only once
3. Removed rows with levels that occur only in training or testing set

### Continuous Variables

Variables tend to be observed in ranges. The probability of observing data outside a range is usually very low. Values that are observed beyond the range can be considered as outliers and need to be treated either by: a) removal, b) substituting with mean, c) flooring/capping. Variables were treated by flooring to 2.5 percentile and capping to 97.5 percentile values. Variables are also scaled to enable faster convergence of learning algorithms.

Beyond this, variables carry some signal and noise. Usually the variation in noise is lesser in proportion compared to that of the signal. Unsupervised learning techniques like PCA and k-means clustering identify the systematic variation in variables and minimize the amount of noise that may creep into the analysis.

## Modelling

### Classification Models

Probabilistic classification models tend to model either conditional probability or joint probability distribution of outcome with respect to independent variables. The following classification models were used:

### Generative models

1. Naïve Bayes
2. Restricted Boltzmann machine

### Discriminative models

1. Linear discriminant analysis
2. Logistic regression – saturated and stepwise
3. Regularized logistic regression (lasso, ridge and elasticnet)
4. Multivariate adaptive regression splines
5. Recursive partitioning (decision tree)
6. C 5.0 (decision tree)
7. Random forest
8. Extremely randomized trees
9. Extreme gradient boosting – linear and tree
10. Support vector machine – linear, polynomial and Gaussian kernel
11. Bayesian additive regression trees

## Parameter Tuning

Most of the models have parameters that need to be tuned to optimize performance. However, scoring performance on training set will lead to biased estimates. Hence a repeated cross validation approach was used. Multiple bootstrap samples are created for model building and performance estimation. Model is built with certain combination of hyperparameters on one of these samples and performance is scored on remaining samples. Average performance is used to compare the performance of models with different hyperparameters. The best performing model is chosen and predictions are made.

## Ensembling

Models with best performing hyperparameters are fused together using a single stage stack ensemble with low variance model to stack the results of first set of models that are built on base data set. To avoid overfitting, the stack is built using average of models built on bootstrap samples.

## Model Evaluation

The following metrics were used for model evaluation:

1. Positive predictive value: Positive predictive value is the probability that subjects predicted to be dissatisfied are dissatisfied.
2. True positive rate: Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (i.e. the percentage of sick people who are correctly identified as having the condition).
3. Log loss: Log loss is the average of log of probability of misclassification.
4. Accuracy: Accuracy is the proportion of correctly classified examples.
5. Kappa: Kappa statistic is a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy.
6. Confusion matrix: Confusion matrix is the contingency table of prediction vs observed outcome.
7. Area under curve: ROC curve plots true positive rate vs false positive rate. The area under ROC curve can be used to compare classifiers.