PROJECT ASSIGNMENT 4

Customer Churn Prediction

Group - DA26

The given data set contains 21 predictor variables and outcome variable is a categorical one which consists of two outcomes “True ” and “False” .
Variable selection is an important aspect of model building , it helps in building predictive models free from correlated variables, biases and unwanted noise.
So, we applied Boruta algorithm for finding out feature selection.

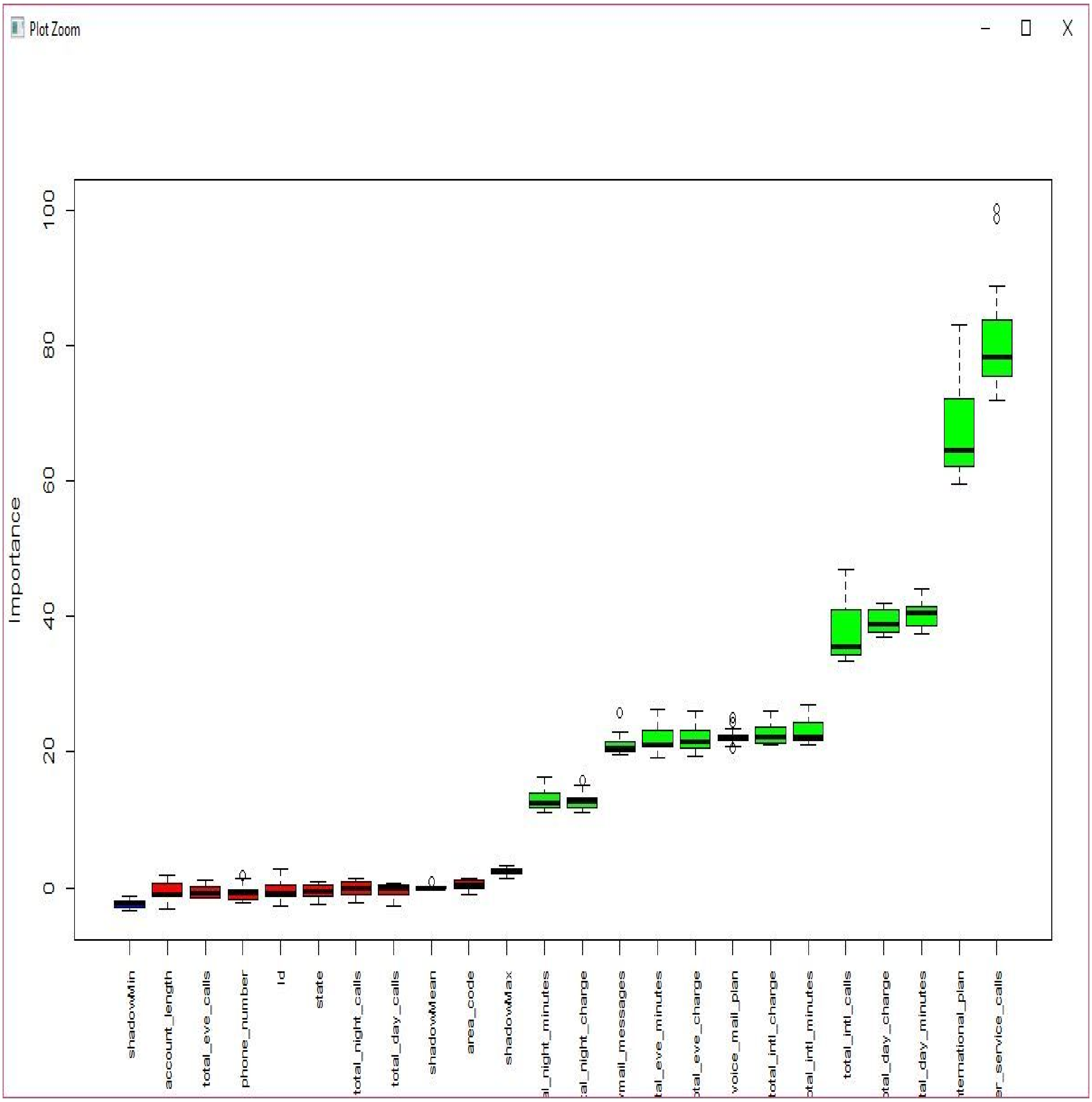
Precisely, it works as a wrapper algorithm around Random Forest.
This technique achieves supreme importance when a data set comprised of several variables is given for model building.

Boruta follows an all-relevant feature selection method where it captures all features which are in some circumstances relevant to the outcome variable. In contrast, most of the traditional feature selection algorithms follow a minimal optimal method where they rely on a small subset of features which yields a minimal error on a chosen classifier.
Boruta find all features which are either strongly or weakly relevant to the decision variable.

So, after applying the Boruta Algorithm , a plot between Importance and features is drawn.

Summary of Boruta Results :-
13 attributes confirmed important:
International_plan, number_customer_service_calls, number_vmail_messages, total_day_charge, total_day_minutes and 8 more.

8 attributes confirmed unimportant:
account_length, area_code, Id, phone_number, state and 3 more.



Boruata Plot : Importance Vs features

We removed the least important feature variables (Id, state, area_code, phone_number), referring to the Boruda plot and trained our algorithms on remaining features.

Splitting of Data into training and test set :
In **k-fold cross-validation**, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter.
When k = n (the number of observations), the k-fold cross-validation is exactly the leave-one-out cross-validation.
In stratified k-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. In the case of a dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels.

<u>Algorithms:</u>

1. **Naive Bayes**

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model.
In simple terms, a Naive Bayes (NB) classifier assumes that the presence (or absence) of a particular feature of a class (i.e., customer churn) is unrelated to the presence (or absence) of any other feature.

10-fold cross validation applied on data set and the results are as follows :

Confusion Matrix :

Reference

Prediction	**False**	**True**
False	**4282**	**593**
True	**11**	**114**

Accuracy - 87.92%
Recall - 99.74%
Precision - 87.84%
F-Measure - 93.41%

2. Decision Tree

Decision Trees (DT) are tree-shaped structures representing sets of decisions capable to generate classification rules for a specific data set, a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. More descriptive names for such tree models are Classification Trees or Regression Trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. DT have no great performance on capturing complex and non-linear relationships between the attributes.

10-fold cross validation applied on data set and the results are as follows:

Confusion Matrix :

Reference

Prediction	**False**	**True**
False	**4217**	**399**
True	**76**	**308**

Accuracy - 90.05%
Recall - 98.23%
Precision - 91.36%
F-Measure - 94.66%

3. Support Vector Machines

Support Vector Machines (SVM), also known as Support Vector Networks, are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM is a machine learning technique based on structural risk minimization. Kernel functions have been employed for improving performance.

Results for :
i) **Gaussian Radial Basis kernel function :**

10-fold cross validation applied on data set and the results are as follows:

Confusion Matrix :

Reference

Prediction	**False**	**True**
False	**4271**	**258**
True	**22**	**449**

Accuracy - 94.40%
Recall - 99.49%
Precision - 94.30%
F-Measure - 96.82%

ii) Polynomial kernel (SVM POLY)

10-fold cross validation applied on data set and the results are as follows:

Confusion Matrix :

Reference

Prediction	**False**	**True**
False	**4276**	**161**
True	**17**	**546**

Accuracy - 96.44%
Recall - 99.60%
Precision - 96.37%
F-Measure - 97.96%

Conclusion:

Precision or recall alone cannot describe the efficiency of a classifier since good performance in one of those indices does not necessarily imply good performance on the other. For this reason, **F-measure**, a popular combination is commonly used as a single metric for evaluating classifier performance. **F-measure** is defined as the the harmonic mean of precision and recall.

F - measure = (2 × Precision × Recall) / (Precision + Recall) .

We observed that, SVM-Poly is best classifier as it’s **F-measure** is highest.