Data Analytics
CS40003

Churn Prediction

Assignment #3

Submitted by
Manish Agrawal
15IM10032

Customer Churn Prediction

Variable Name Type

Account_length : number of months active user
Total_eve_charge : total charge of evening calls
area_code : area code of customer
total_night_minutes : total minutes of night calls
international_plan : local/international call
total_night_calls : total number of night calls
voice_mail_plan : voice mail or normal
total_night_charge : total charge of night calls
Number_vmail_messages : number of voice-mail messages
total_intl_minutes : total minutes of international calls
total_day_minutes : total minutes of day calls
total_intl_calls : total number of international calls
total_day_calls : total number of day calls
total_intl_charge : total charge of international calls
total_day_charge : total charge of day calls

Derivable variables

Using the above features following derivable features were derived:
total_minutes = total_day_minutes + total_eve_minutes + total_night_minutes
total_charge = total_day charge + total_eve_charge + total_night_charge
day_rate = total_day_charge / total_day_minutes
eve_rate = total_eve_charge / total_eve_minutes
night_rate = total_night_charge / total_night_minutes
intl_rate = total_intl_charge / total_intl_minutes

Given dataset was split into training(80%) and testing set(20%) using **createDataPartition()** of R Caret package. createDataPartition() does a stratified split of the data. **churnTrain** and **churnTest** were the new training and testing dataset which were then used to perform further tests.

Naive Bayes model was developed using the library : **e1071**
Decision Trees model was developed using the library : **rpart**
Support Vector Machine model was developed using the library **: e1071**

Confusion Matrix was also tabulated with predicted values in the columns and true values in the rows.

For Naive Bayes, confusion matrix is as:

		True Values	
		0	1

Predicted Values	0	803	72
	1	52	73

For Decision Trees, confusion matrix is as:

		True Values	
		0	1
Predicted Values	0	835	41
	1	20	104

For SVM, confusion matrix is as:

		True Values	
		0	1
Predicted Values	0	843	67
	1	12	78

Classification Report on used models:

Models	**Precision**	**Recall**	**Accuracy**
Naive Bayes	91.8%	93.9%	91.67%
Decision Trees	97.3%	97.66%	88.9%
SVM	92.6%	98.60 %	92.1%

The accuracy of the given classification models are improved by selecting the variables having highest importance among them. With the following step, we can compare and conclude with the classification models with highest accuracy.