CDS Project

ZS Customer Modelling Challenge 2015

Group# 07

Ankit Katiyar (05)
Avinash Kumar (08)
Mohammed Tariq (22)
Mukul Kumar (26)

PGDBA - ISI Kolkata

Why this ?

In an environment where the only constant is innovation, the telecommunication sector is rapidly increasing its growth by understanding the customer behaviour and pattern.

As we are in an era where data leverage is a key feature to make smart business decisions, the telecommunication sector is rethinking all aspects to gain competitive advantage.

The problem

Company

ZS Associates Inc is a global leader in sales and marketing consulting, outsourcing, technology and software.

Context

Predictive Modeling:
To predict the customer behaviour as to whether the add-ons are preferred or not

Problem statement

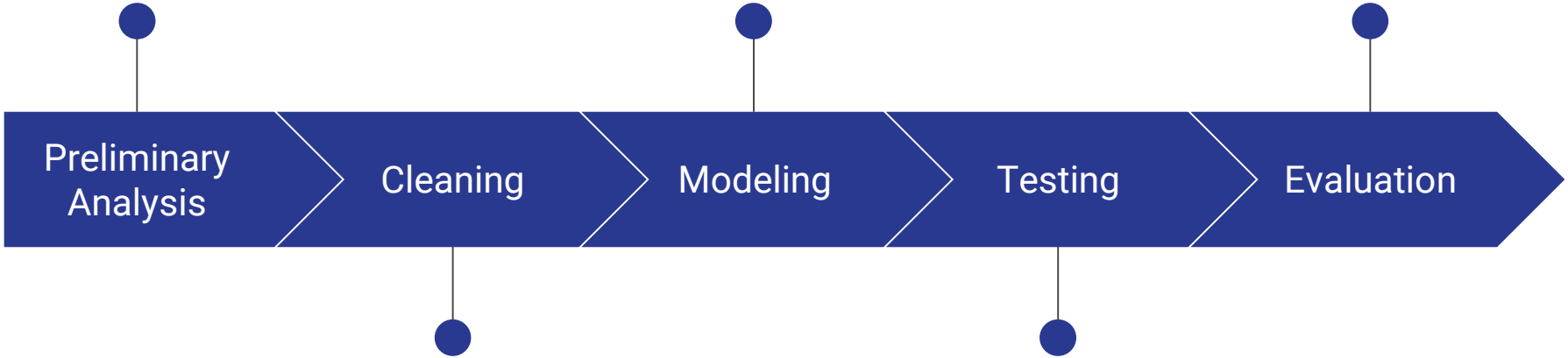
The challenge is to model customers of an telecom company and predict the propensity of them buying add-ons.

Problem Solving Approach

Summary of the data

Build a predictive model using random Forest

Validating the data to to ascertain model can predict future outcomes

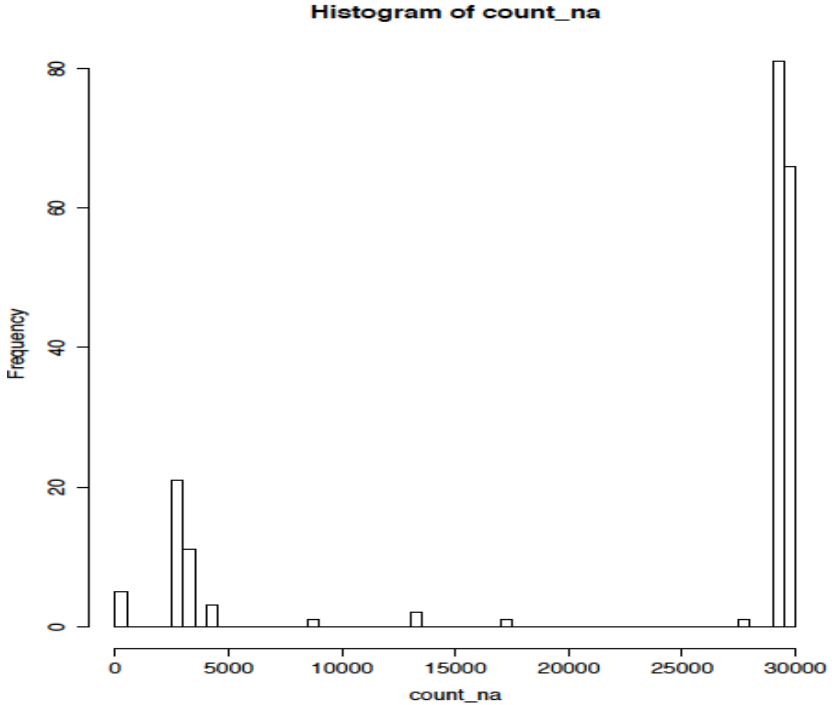


Removing NA values and performing multiple Imputation

Evaluate performance of the model using test data sets

Preliminary Analysis

Histogram for NA values



The data has been split into:
50% - Training
20% - Validation
30% - Testing

The data contains 190 explanatory variables out of which 38 has 10 percent NA values while the rest have 90 percent or more.

Data Cleaning

```
for (i in 1: 176)
  {
  if (sum(is.na(td2[,i]))< 5000 )
        {
        count[r] = i
        r = r + 1
        }
}
td_1<- subset( td2 , td$Labels== 1)
td_2 <- subset( td2 , td$Labels== 0)
summary(td_1)
summary(td_2)
```

```
library(mice)
tempData <- mice(td3,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)
```

DIMENSION REDUCTION (EXPLANATORY VARIABLES)

Remove the variables having NA values more than 90 percent and missing completely at random (MCAR).

MULTIPLE IMPUTATION

After selecting the explanatory variables ,we performed predictive mean matching with the help of MICE package.

Modeling

Selection of features

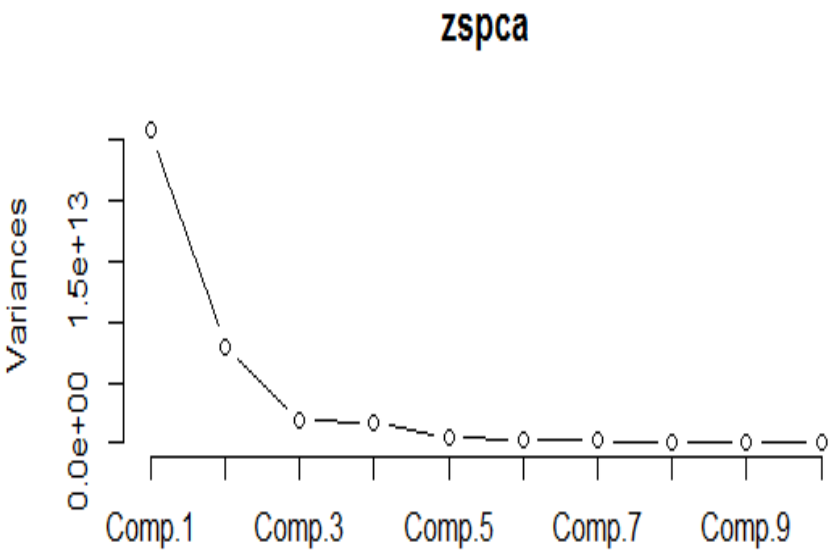
```
zspca <- princomp(td4[,-c(1,40)] , cor = "F")
summary(zspca)
zspca$loadings
biplot(zspca)
screeplot(zspca)
```

```
td4cov = cov(td4[,-c(1,40)])
td4cor <- cor(td4[,-c(1,40)])
heatmap(td4cor)
```

Correlation matrix

It gave the correlation between different explanatory variable to see the independent features

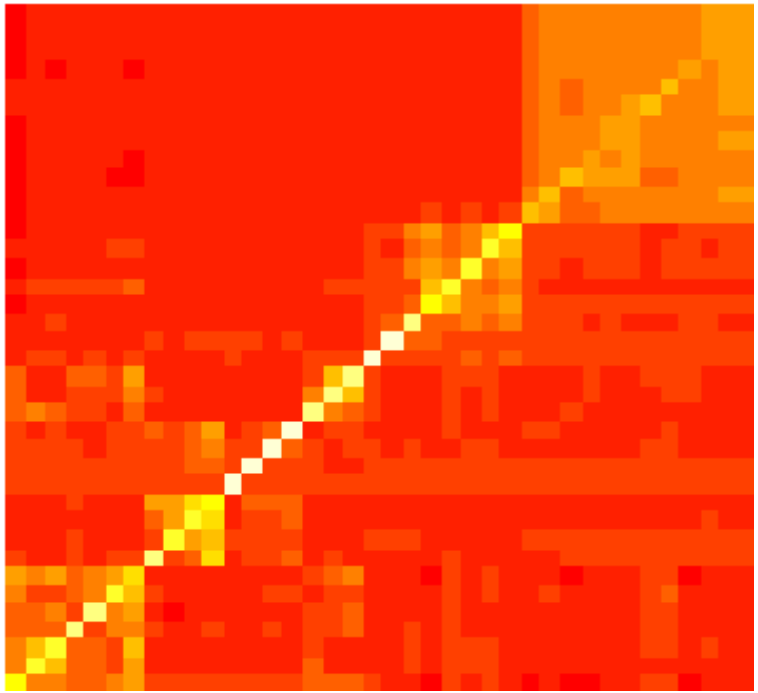
Scree Plot



PCA

It gave the Major features which cover the maximum variances of the explanatory variable

Heat Map



It provides the graphical representation of the correlation matrix.

Data modeling for Random forest

```
rndf <-randomForest(Labels ~ Variable_12 + Variable_38 +
                    Variable_83 + Variable_4 +
                    Variable_34 + variable_107 ,
                    data=train,ntree=500,mtry=5,
                    importance=TRUE, na.action=na.omit )
model_pred_probs = predict(rndf, test , type="response")
```

We modeled our data for the prediction using random forest model. The features chosen were features making principal components and other independent feature.

Testing and evaluation

Contingency Table

		Actual	
		Not Buy '0'	Buy '1'
Test	Not Buy '0'	True Negative	False Negative
	Buy '1'	False Positive	True Positive

The Evaluation was based on Mean F1 Score

$$F1 = \frac{2PR}{P + R}$$

Precision P is the ratio of true positives (T_P) to all predicted positives ($T_P + F_P$)

$$P = \frac{T_P}{T_P + F_P}$$

R is Recall or Sensitivity which is equal to the ratio of true positives (T_P) to all actual positives ($T_P + F_N$)

$$R = \frac{T_P}{T_P + F_N}$$

Improvements

- Further tuning was done in Random Forest for better results.
  - increasing the number of variables to be used for making the tree
  - Limiting the maximum number of nodes at the bottom level
  - changing the weightage of 1's and 0's to control the precision (because the proportion of 1 is very low approx 7 %)

p value= 0.26		Actual	
		0	1
Test	0	8261	125
	1	518	97

The F1 Score = 0.23178

RANKING

Rank	Name	F1 Score	Number of attempts
1.	Anonymized User	0.261393	28
2.	Anonymized User	0.257895	6
3.	Anonymized User	0.257375	17
26.	Anonymized User	0.049421	10
27.	*You*	*0.026094*	*8*
28.	Anonymized User	0.019688	2

The Highest Score was 0.26 and now we have 0.23

Software and Packages Used

- Excel
- R (Mice, Random Forest, Miss Forest)

Special Thanks

- Robin Singh
- Manaswi

Thank You!

তোমাকে ধন্যবাদ