PREDICTING CUSTOMER CHURN OF TELECOM DATA

November 29, 2017

Arindam Adak & Kalyanbrata Maity
Ramakrishna Mission Vivekanada University
Department of Data Sciences

Contents

0.1 INTRODUCTION

0.1.1 Data Science in Churn Analysis

Classification methods in Data Science have a wide range of applications for business problems. In this assignment, machine-learning techniques are applied to analyze real word data of an unknown telecommunications company in the USA to develop a model that predicts, as accurate as possible, whether a customer will churn or not. It is well known that in industries like that one where there is a lot of competition in the market, customer acquisition is very expensive and therefore trying to minimize the proportion of people that quit a contract is a very valuable thing to do. More accurate detection of people that will churn makes it possible to design more effective marketing campaigns or reward systems aimed at retaining these customers.

In the rest of this introduction the main characteristics of the dataset will be presented. Section 2 is dedicated to data preparation. The most important steps and decisions and their justification will be developed. The three different strategies used for variable selection will be explained in section 3. We tried a wide variety of models: Naive Bayes, Logistic regression, Support Vector Machine, Random Forests and Gradient Boosting. After getting the first round of results, we concentrated our effort into improving the results for the last two. The description of this process will be the subject of section 4. In section 5 we summarize the results and in section 6 we present a brief conclusion.

0.1.2 Description of the data set

The complete dataset consists of 100,000 "mature customer" i.e. who were in the company for at least six months that were sampled during 2001 and 2002. 171 variables capture customer's information about 1) socio-demographic, economic and geographic characteristics and 2) their usage and experience with the service provided by the telco company. Our target variable is "churn" only contained in the test dataset, which was calculated, according to the description provided, based on whether the customer left the company during the period 31-60 days after the customer was originally sampled. 49.56% of the customers in the training set churned, while the rest, 50.53% did not churn. **This means that, for this dataset there is no class-imbalance problem that had to be dealt with.**

0.2 DATA MANIPULATION

Every data analysis starts with importing the data. We used the option 'stringsAsFactors' = FALSE to deal with the data classes manually. The process of transforming the variables was not a one-step process. We first made some basic adjustments, **calculated correlations and analyzed the information value given by the WOE procedure and then made some extra manipulations**. Here just a summary of the changes will be presented.

To ensure that both the training and the test dataset contained the exact same variable transformations and factor levels, therefore avoiding problems for the prediction, the two datasets were merged into one using the command smartbind from the package gtools. After all the necessary transformations were made they were split again.

0.2.1 Variable transformations

- **age1 and age2**
  After transforming the large amount of zeros into NAs, we decided to bin this variable into the following categories: (0,30], (30,40], (40,50], (50-60], (60, max(age)] and relabel the NAs as "Missing". For each case, a dummy variable was created that indicates whether the value was missing or not.

- **Roaming-variables (rmcalls, rmmou, rmrev)**
  These all had a high number of NAs (48083), we calculated the correlation between the non-NAs and churn and since the correlation coefficient was too low, decided to drop them from the analysis. We however included a dummy variable that indicates whether the observation had a valid number or not.

- **retdays (48083 NASs)**
  NAs were transformed to missing and four categories according to the .25,.5 and .75 quantiles of the valid values were created. A dummy variable `retdays_bool` was also created with the levels "Called" and "Not called" (the retention team)

- **csa**
  This variable seemed to have potential explicative power according to the importance scores calculated in the **WOE** Method, however due to the fact that it consisted of over 700 levels, it turned out to be impossible to include it in our model. We extracted the city (contained in the first 3 characters) but this shortened version of csa (`csa_cities`)

lost its explicative power and did not make it into the pre-selected variable list for modeling.

- `last_swap`
  This variable gives the date of the last swap. To extract more meaning out of this information, we calculated the number of days that had passed between the last swap and an assumed date for the analysis of `2010-01-01`. Since this variable introduced too many NAs, we decided to bin it using the variable quantiles (.25, .5 and .75) of the valid observations and assigning "Missing" as an extra category for the rest.

0.2.2 Missing Values

Two different strategies on how to treat missing values were applied depending on the type of the variable in question. For categorical variables and also for numeric that were binned into categories, we created an extra category labeled "missing" that enters normally to the models. **As for the numeric variables, after the variable transformations described in the previous section, we were left with a dataset for which none of the remaining numeric variables had more than 3% NAs. This meant that we got a data set in the end with 48,127 observations without NAs to build our models. In the case of the test set used for prediction, we tested our models both the mean and the median value of the variable in question.**

Table 1: Percentage of Missing Values

retdays	rmcalls	rmmou	rmrev
0.960	0.857	0.857	0.857

0.3 VARIABLE SELECTION

Our first approach to variable selection was splitting the dataset into factor and numeric variables to get a first idea of their relationships of these variables with our target variable churn using appropriate methods in each case.

0.3.1 Filter Approach

0.3.1.1 Factor Variables : Weights Of Evidence (WOE) and Information Value Score

The WOE Method is a useful start to analyze whether a level in one factor exhibits a considerable difference in the proportion of people that belong to one of the categories of the target variable relative to the other in this case whether a person churned or not churned.

$$WOE_{cat} = ln\left(\frac{p(churned)_{cat}}{p(notchurned)_{cat}}\right)$$

The Information Value (IV) is a measure derived from the WOE to assess not just a level of a variable but the whole variable and also to be able to compare among variables.

$$IV = \sum((p(churned)_{cat} - p(notchurned)_{cat}) \times WOE_{cat})$$

A useful rule of thumb is picking the variables with an **IV bigger than 0.02 (which are considered to be at least weakly explicative)**. In our case, out of the 63 variables that entered to the model, only 8 met this criterion.

Table 2: Variables' Information Value (WOE)

csa	`hnd_price`	`hnd_webcap`	**crclscod**
0.072	0.052	0.03	0.029
`retdays_fac`	`tot_acpt`	`retdays_bool`	`tot_ret`
0.027	0.029	0.021	0.021

Here goes the figure :

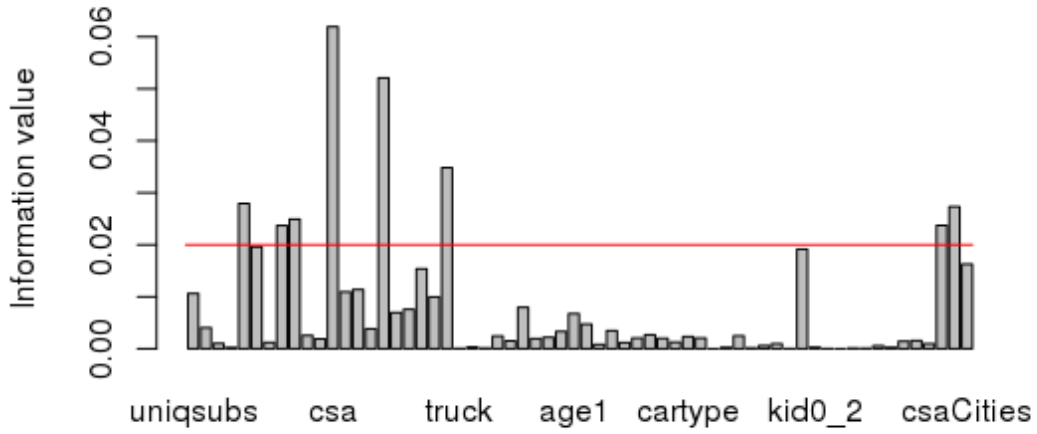

Figure 1: Plot of Information Value

0.3.1.2 Numerical variables: correlation coefficient

For the numerical variables we decided to use, as a first approach, the simple correlation coefficient, using the option that all the non-missing values for each pairwise set of variables are used. Normally we would have set a rule for which we would have kept all variables with at least a .5 correlation coefficient with the target variable. However out of the 111 variables, not even one had a correlation coefficient bigger than the established value. We therefore concluded that the relationship between the numerical variables, if any, could not be linear and therefore cannot be captured by correlation. Therefore this approach would not be right for variable selection.

0.3.2 Random Forest for variable selection

Given the above results, we decided to use the Random Forest Method directly to select our variables. This had also the extra benefit that we use a single method that can handle both numeric and categorical variables. To do this we used the option "importance = TRUE" to get the scores on Mean Decrease of Accuracy for each variable.

The big challenge in this data set is that there is not one or a small set of variables that really help predict churn. We have to include several variables in model assuming that each

will help to explain little parts. This, of course, with the risk of over fitting our data. We decided to reduce our data set to contain the top 20 variables and apply the models and for each one decide the best amount of variables.

Here we present the top 20 variables and their scores:

Table 3: Variables' Importance Scores (Random Forest)

eqpdays	**months**	mou_Mean	totmrc_Mean	last_swap
0,0068	0,0041	0,0020	0,0016	0,0013
hnd_price	**adjrev**	change_mou	mou_cvce_Mean	**avg3mou**
0,0013	0,0012	0,0012	0,0012	0,0011
totrev	mou_Range	mou_opkv_Mean	**totcalls**	**phones**
0,0011	0,0010	0,0010	0,0009	0,0009
avg3qty	complete_Mean	peak_vce_Mean		
0,0008	0,0008	0,0008		

0.3.3 Gradient Boosting for variable selection

We also tried to find out importance variables through GBM with 1000 trees :

Table 4: Variables' Importance Scores (GBM)

eqpdays	**months**	mou_Mean	**crclscod**	hnd_price	change_mou
23.28	17.32	8.75	7.36	6.63	6.24
totmrc_Mean	**avgqty**	rev_Range	mou_Range	mou_cvce_Mean	retdays_factor
3.92	3.44	3.07	2.99	2.20	2.03
ovrmou_Mean	tot_ret	tot_acpt	**totcalls**	overmou_Range	**totrev**
1.93	1.18	1.08	0.97	0.90	0.72

From these two method we got almost the same set of variables , so we can assume that these set of variables are more all less consistent to predict churn. But the real problem here is that the variables have very low correlation with the churn variable i.e. why we can expect that 20 variables are not enough to predict churn good.

0.4 MODELLING

One of the things that had the most impact in our modeling decisions was the computational capacity of running our models.

We decided to try several different models by dividing our training dataset into 70% of the observations to real training and the rest to predict and assess the predictions. Finally we fitted three competing models Random Forest and Gradient Boosting, Naive bayes . **For these three models we performed 10-fold Cross-Validation and averaged their performance metrics lift score and area under the ROC-Curve.**

0.5 RESULTS : ON BASIS OF AUC AND LIFT SCORE

Table 5: Results

Classifier	AUC	Lift Score
Naive Bayes	0.58	0.79
Random Forest (100 trees)	0.63	0.59
Gradient Boosting (500 trees)	0.67	0.49

0.6 CONCLUSION

With respect to the AUC (Area under the ROC curve) value we got **Gradient Boodting Model as the winner**. The predicted churn probablity corresponding to the customers that are selected in the validation set are written in the "Customer churn probability " . From this churn probability we can predict that which customer can churn or not, of course the cut-off of probability should be given.

0.7 APPENDIX

List of packages that are used in this analysis are :

1. leaps
2. klaR
3. randomForest

4. Hmisc
5. ROCR
6. AUC
7. GBM
8. gModels
9. caret
10. verification
11. gtools
12. dplyr]

REFERENCES

<http://www.souravsengupta.com/ml2017/index.html>