Business Analytics and Predictive Modelling

Predicting Churn Probabilities: Group Report

(submitted by Emma Hennig, Felicia Schneiderhan, Pierre Navarro and Christian Geike on 1/3/2015)

Introduction

Customer attrition, represented by the churn rate of a product, service or company, is the loss of a customer to a competitor. Churn rate is one of the most important indicators of the overall health of a venture – next to customer acquisition cost and customer life-time value. Especially business areas that rely on subscription services and recurring purchasing/using behaviour – such as mobile phone services – are vulnerable to high churn rates. Additionally, in the mobile telecommunications sector switching costs are low and the upcoming phone number portability will erase the last barriers for customers to easily change their providers. Thus, even more so when customer acquisition costs are high, it is much more economical for a business to invest in customer retention and thus achieve a state of low customer attrition. Thus, for the mobile telecommunications provider it is highly useful to predict churn probabilities of their customers, because on the one hand this will enable the company to engage in pre-emptive retention measures once a customer is identified as a high risk candidate and on the other hand it will save costly retention offers to customers who would have stayed regardless. Targeted marketing, offering of special incentives or fortified service support directed at customers most likely to defect, can be valuable tools to improve customer retention and thus improve overall profitability of the business.

Data pre-processing

The training database gathers data of 50000 customers described by 173 variables. As is almost always the case with such big databases, there was quite some work to be done in terms of data pre-processing. The first step in our data pre-processing was to give variables their proper class and format. Some of them were numeric, mainly variables related to phone use and consumption. Some others were categorical, mainly concerning socio-demographic features. As a result, we ended up with 124 numeric variables, 48 categorical variables and one date variable.

Missing values

The following problem we had to tackle was missing values. Indeed, the missing value rate for some variables was quite high. We took decisions according to how high this rate was and also according to whether we could interpret and replace these missing values or not.

For missing value rates higher than 30%, we simply dropped the variable. In total, we dropped 21 variables.

For lower rates, we found a way to replace missing values:

- For some of them, we interpreted the missing values and then replaced them by a value. For instance, regarding the total number of referrals (REF_QTY), a missing value can be interpreted as an absence of referrals. We then just replaced missing values by 0.
- However, for some variables, we did not manage to find a good way to interpret and replace missing values. For these variables, we took the median (or the mode, according to their class) and replaced them.

One variable was remaining: income (INCOME). The missing value rate was about 25% and we thought that this variable may be relevant for the suite of our analysis and that at least, it would be interesting to implement a more elaborate model to estimate the NAs. Moreover, the missing values could not be interpreted so that replacing 25% of the values by the mode would not have made sense. The income we considered to be an ordered categorical variable, with nine categories. The most accurate way to estimate missing values was therefore to settle a multinomial logistic regression with the non-missing values and then to predict the missing values based on this multinomial model. To build the model, we used four variables as possible explanatory variables for the income: the area (AREA), the marital status (MARITAL), the ethnics (ETHNICS) and finally the age (AGE1). We gathered together some modalities of both the area and the ethnics and then we transformed them into dummies.

In the annex, you can find in detail the class of each variable and how we treated missing values according to each variable.

Outliers

Regarding outliers, we chose the Tukey approach for numerical variables. This method requires to change the values higher (resp. lower) than the 3^{rd} quartile (resp. the 1^{st}) plus (resp. minus) three times the interquartile range. Nevertheless, we did not apply this method to all the variables since some of them had an interquartile range equal to zero: in particular this is the case for variables which take the same value for a very large majority of observations. The mean revenue of data overage (DATOVR_MEAN) for instance is mostly equal to zero since most of the people did not overpass their data limit.

For factors we identified two variables with outliers: number of active subscribers and number of unique subscribers (resp. ACTVSUBS and UNIQSUBS). For these variables, as well as for the number of models issued (MODELS) we grouped the highest categories into one.

Further transformations

A z-transformation was applied to all the numerical variables. This scaling procedure ensures similar values ranges and allow classifiers working with distances not to be biased because of different scales among variables.

For some categorical variables, we gathered together some levels in order to reduce the total number of levels of these variables. The affected variables are: number of active subscribers (ACTVSUBS), age of the first and second household members (AGE1 and AGE2), area (AREA), credit class code (CRCLSCOD), adjustments made to credit rating of individual (CRTCOUNT), ethnics (ETHNICS), marital status (MARITAL), number of models issued (MODELS), number of handsets issued (PHONES), total offers accepted from retention team (TOT_ACPT), number of unique subscribers (UNIQSUBS). Regarding the date of last phone swap (LAST_SWAP), we changed the variable into the number of days since swap (we considered the maximum date as the reference date to calculate it) and we created categories accordingly to how long ago the last swap was.

For non-ordered categorical variables we built dummies. The following variables were transformed: area (AREA), ethnics (ETHNICS), division type code (DIV_TYPE), dualband (DUALBAND), premier household status indicator (HHSTATIN), handset web capability (HND_WEBCAP), marital status (MARITAL), social group letter only (PRIZM_SOCIAL_ONE), total offers accepted from retention team (TOT_ACPT). In total, we got 27 dummies.

Variable Selection

For variable selection, we first did a very simple procedure to have a first idea whether this variable could influence the churn probability: we calculated the mean of each variable first on the clients that did not churn and then on the clients that churned. Nevertheless, we did not use this procedure to decide which variables to include in our model. Instead, we used WOE, correlations and Random Forest.

We calculated correlations between all the variables (separately) and the churn variable. Before that, we transformed every categorical variable into numerical ones so that the computation of the correlation was feasible. Here the 20 variables having the highest correlation with the churn variable (in absolute value) are displayed:

Variable	Correlation (absolute value)
eqpdays	0.11472894
hnd_price	0.10336092
hnd_webcap_3	0.08656398
retdays	0.07622461
tot_acpt_0	0.07247527
last_swap	0.06931283
tot_ret	0.06891415
asl_flag	0.06771693
totmrc_Mean	0.0677154
tot_acpt_1	0.06315947
hnd_webcap_1	0.06132199
mou_Mean	0.05532317
custcare_Mean	0.05332291
complete_Mean	0.05310789
hnd_webcap_2	0.05309482
comp_vce_Mean	0.05268276
mou_cvce_Mean	0.05243109
mou_opkv_Mean	0.05047662
attempt_Mean	0.05001695
plcd_vce_Mean	0.04964264

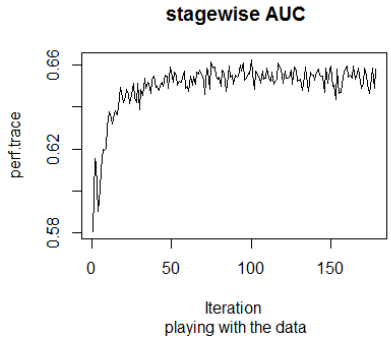
The WOE is designed for factors only and so, it concerned only 24 variables from our pre-processed dataset. None of the information values was strongly significant (greater than 0.3) or even in the 'medium' class (between 0.1 and 0.3). Instead, most of them are below the threshold of 0.02 which means that they are not predictive at all. Some of them nevertheless got an IV score between 0.02 and 0.3 which means that they have a low/medium predictive property.

The Random Forest procedure, on the other hand, includes all the variables: factors and numeric ones. So, finally we used it. We chose to leave the number of trees at its default value (*ntree*=500) - it was taking a very long time - to ensure the accuracy of this procedure. We got the importance score for each variable. The highest ones are displayed in the following table:

Var	Importance score
eqpdays	38.9592976
months	33.9413254
mou_Mean	28.2467974
avg3mou	22.3882629
rev_Mean	21.544036
adjrev	21.3743268
totmrc_Mean	20.9226426
avg6mou	20.3557261
avgmou	20.1356259
avg3qty	19.9728789
mou_Range	19.7962141
totrev	19.6136664
avgqty	19.4548262
change_mou	19.308123
complete_Mean	18.971272
mou_peav_Mean	18.7969149
rev_Range	18.4936827
mou_rvce_Mean	18.4247529
retdays	18.1659521
Adjqty	18.0706282

We can notice that some of these variables already appeared in the top-20 correlation coefficients above. We based our variable selection, however, only on this Random Forest procedure, known to be more relevant than some filter approaches such as correlations.

To complete the variable selection, we had to define the number of variables to keep. Indeed, we knew how important every variable was for our analysis but we did not know where to stop the selection. For this purpose, we built a stage-wise feedforward loop that iteratively runs a model with the n most important variables, *n* varying between 1 and the total number of variables. The model was random forest and each time, the Area under the Curve (AUC) was calculated. Given the results, we chose to keep 18 variables as the following curve suggests.



Model Estimation and Evaluation

Data Partitioning

The so-called training set - consisting of 50000 observations for 173 variables including the target variable churn- represents our whole data set. The so-called test set, on the other hand, does not include the target variable. It can be seen as simulating a real world situation since the individual value of the target variable for each observation is unknown to us.

Our data partitioning strategy depends on the chosen classification algorithm. Above all, we used the logistic regression and neural network classifier. For the logistic regression classifier, we employed two different partitioning approaches. On the one hand, we conducted a non-repeated three-fold cross-validation. The aim was simply to see whether such a cross-validated sampling approach would yield better accuracy measures than the split-sample logistic regression approach. We preferred, however, the split-sample approach, which on the other hand was the second approach employed for the logistic regression classifier. The reason for this preference was that we wanted to compare the accuracy measures with those of the neural network classifier under conditions as equal as possible. A split-sample approach was necessary for our neural network

classifier because we wanted to conduct a grid-search to optimize the meta-parameters. This implies that we would need to hold back a fixed portion of our data for testing the optimized classifier. We split the data so that 70% of them, i.e. 35000 observations, were part of the training and 30%, i.e. 15000 observations, were part of the test set.

The logistic regression and neural network classifier directly estimate the posterior probability of churning. In contrast, naïve Bayes indirectly estimates the class-conditional probabilities first; above all, its assumption of independence is violated. When playing around with the naïve Bayes classifier, it did not perform very well. The same holds true for the decision tree classifier. Consequently, we focused on the logistic regression and neural network classifier described here.

Logistic Regression

In the logistic regression we conducted, we did not penalize for model complexity explicitly, i.e. implemented a LASSO or Ridge penalty. It may be, though, that the call function does already include such a penalization. The logistic regression of our cleaned, transformed and reduced data set is summarized in the following table:

```
Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = train.set)
Deviance Residuals:
   Min      1Q  Median    3Q     Max
-2.0885  -1.1417  -0.7801   1.1608   2.0046
Coefficients:
                 Estimate   Std. Error   z value      Pr(>|z|)
(Intercept)      -0.01274     0.01086     -1.173      0.240666
eqpdays           0.27902     0.01367      20.417     < 2e-16 ***
months           -0.10333     0.02085     -4.957      7.15e-07 ***
mou_Mean         -0.09930     0.10799     -0.920      0.357809
avg3mou          -0.24588     0.12115     -2.030      0.042406 *
rev_Mean          0.29776     0.02940      10.129     < 2e-16 ***
adjrev           -0.35238     0.17761     -1.984      0.047261 *
totmrc_Mean      -0.24259     0.01979     -12.257     < 2e-16 ***
avg6mou          -0.05814     0.05947     -0.978      0.328229
avgmou            0.27376     0.05416      5.054      4.32e-07 ***
avg3qty           0.04299     0.05309      0.810      0.418147
mou_Range         0.14800     0.01787      8.284      < 2e-16 ***
totrev            0.40596     0.17612      2.305      0.021165 *
avgqty            0.02054     0.04947      0.415      0.678050
change_mou       -0.07072     0.01910     -3.703      0.000213 ***
complete_Mean    -0.08974     0.02963     -3.028      0.002459 **
mou_peav_Mean    -0.03793     0.02531     -1.499      0.133902
rev_Range        -0.07751     0.02110     -3.673      0.000239 ***
mou_rvce_Mean     0.01455     0.02202      0.661      0.508763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48519  on 34999  degrees of freedom
Residual deviance: 47436  on 34981  degrees of freedom
AIC: 47474
Number of Fisher Scoring iterations: 4
```

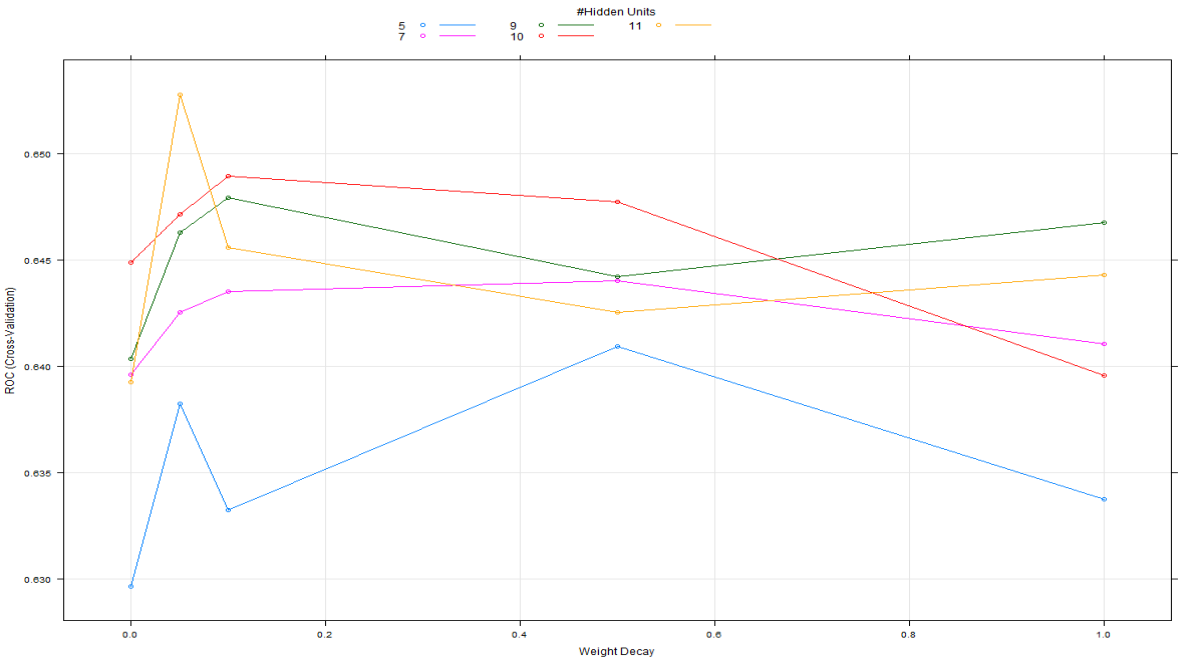
First of all, we notice that most of the coefficients are significant and even strongly significant. Some of them move positively with the log-odds of churning: number of equipment days, total revenue, range of number of minutes of use. In a sense, the customers that consume most are also more likely to churn. It is hard to interpret the coefficients in a numerical way since they are logarithmic odd-ratios. Moreover, they have been z-transformed. None of the coefficients is really high or low and significantly different from the others, which might confirm that none of the variables has a big effect on the churning probability.

Neural Network

When building the neural network classifier, we had to decide for two meta-parameters: the number of neurons and the strength of decay. In order to approach this decision systematically, we conducted a small grid search.

The grid search consisted of three rounds. In each round, several neural network models were built based on the combinations of three to five chosen values for the two meta-parameters number of neurons and decay. The classifiers were built using three-fold cross-validation on the training set only; each model was then evaluated taking the AUC accuracy measure. Based on the resulting AUC values, we chose finer values for the two meta-parameters. The results are as followed:

In the first round, a neural network classifier with 9 neurons and 0.1 decay performed best. The number of neurons were 3, 5 and 9 and the strength of decay was 0.1, 5 and 9. The corresponding plot depicting all the AUC values can be found in the annex. In the second round, other meta-parameters were chosen: the number of neurons were 5, 7, 9, 10 and 11 and the strength of decay was 0, 0.05 ,0.1 ,0.5 and 1. The plot on the next page shows the calculated AUC for the second round graphically:



When looking closely, one can see that a neural network with 11 neurons and a decay of 0.05 achieved the highest AUC value. In the third round, we updated the meta-parameters again: the number of neurons were 10, 11, 12, 13 and 14 and the strength of decay was 0.03, 0.05, 0.07, 0.1 and 0.2. The neural network classifier that performed best in the third round had 10 neurons and a decay of 0.2. The corresponding AUC-plot of the third round can again be found in the annex.

To decide for an optimal pair of meta-parameters, we took the winner neural networks of the three rounds, build the three final neural networks based on the training and validation set and calculated the AUC value for each one on the test set. The size-9-decay-0.1-neural network from the first round has an AUC value of 0.6477, the size-11-decay-0.05-neural network of the second round has an AUC of 0.6429 and the size-10-decay-0.2-neural network of the third round has an AUC of 0.6435. Interestingly, the initial neural network performed best. This is why we decided for this model.

A summary of this chosen model can be found in the annex. It shows the individual weights between the three layers.

Ensemble Algorithms

For model building, we did not apply any ensemble algorithms. During variable selection, however, we used a homogeneous ensemble algorithm, namely Random Forest, to calculate the variable importance scores and perform the stage-wise feedforward variable selection. The details have been described above.

Model Assessment

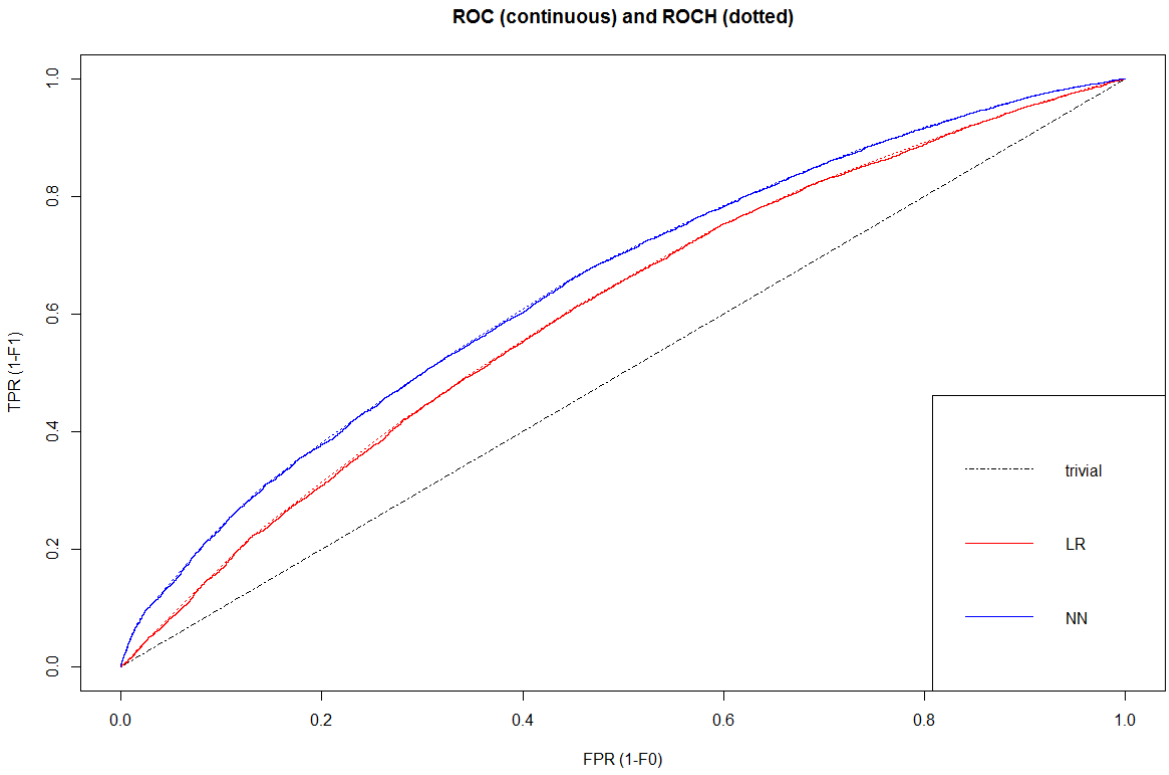
The split-sample logistic regressions and the final neural network model can be assessed according to the following six criteria.

Accuracy

To measure the accuracy of the models, we calculated the Brier score and the Area under the ROC Curve. The results are shown in the following table:

Model Accuracy	Brier	AUC
Logistic Regression	0.2422	0.6054
Neural Network (18-9-0.1)	0.2328	0.6477

The chosen (18-9-1)-neural network model outperforms the logistic regression model in both accuracy measures: the Brier score, representing a probability metric, is lower and the AUC, representing a ranking metric, is higher. It should be added that we think that the resulting AUC score for the neural network is still pretty low. We will discuss some possible reasons in the next and also final section. The Receiver Operating Characteristic (ROC) curve shown below graphically displays the better performance of the neural network model: the curve is closer to the upper left corner.



Scalability

When it comes to the consumption of time resources, the neural network classifier needs longer training time: the grid-search and final model building takes quite a while whereas the logistic regression is estimated within seconds. The computing time could be reduced if R would not only use CPU unit, which it does on default, but rather employ parallelization. We found some packages such as “Rmpi”, “Snow” or “Snowfall” which would allow for parallel computing but did not have time to get into the topic. Considering prediction time, both models are really fast so that predictions are generated quickly; apart from that, predicting churn probabilities does not have to happen in real time anyway. In this sense, it is not time sensitive.

When it comes to the consumption of memory resources, neural networks consume more memory resources. The calculations, however, can still be handled using a 32-bit computer. As a side-note, it can be said that the calculation of Random Forest with a default of *ntree* = 500 on the whole dataset can most likely not be handled by a 32-bit computer because – as some of us had to experience- R performs and saves its calculations in the computer working memory so that the memory limit is often reached. Once the models are calculated, the storage consumes comparably much less memory resources and is not a problem at all.

Robustness

After model training, both models were tested on an unused part of the data, the test set. This approach of simulating new data, i.e. a real world situation, probably increases the robustness of the models to some extend by design. This application simulation would have been further strengthened if we had managed to use cross-validation. Also penalizing for model complexity may reduce the variance of the models and therefore make them more robust in a real world setting. Apart from these little indications, it is hard to say anything about the models’ robustness.

Comprehensibility

The logistic regression model is definitely more comprehensible: even if not at all trivial how to interpret its log-odds coefficients, they can be interpreted. Looking at the effect size and sign of the coefficients, one can get to specific interpretations. That’s why the logistic regression model can be considered a white-box model. The neural network classifier, in contrast, is rather a black-box model: a straightforward interpretation of the input attributes on their effect on the probability of churning is hard, if not impossible, for larger neural networks.

Justifiability

Based on comprehensibility, it is well possible to find explanations for logistic regressions output that is in accordance with business rules and own beliefs. For a neural network model this is not really possible since -if at all - the transmission effect of its input variables is hard to understand.

Calibration

It is hard to say whether the models' probability predictions match the actual churning frequencies or not. Since we have no experience of how the models would perform in a real world setting over a longer period of time, we can assume that it is not unlikely for the models to be over- or underconfident, i.e. suboptimally calibrated.

Finally, we decided for the neural network model. The topmost reason for this decision was its better accuracy measures. Furthermore we could argue that although a black-box-model, we can get a rough idea of the variables' impact on churning simply by interpreting the output of the other model, the logistic regression.

Results – Limits - Conclusion

To sum up, we have to admit that we would have hoped to achieve better performance measures. Our aim was to get an AUC within the 0.75-0.80 range. Somehow it did not work out that way. At least the model is better than throwing a coin although it is not very good considering the time and energy we invested into this project. Now there is no more time left to improve the results. Our guess is that we could have done much better during variable selection. Perhaps we should have used partial dependence plots on the Random Forest importance-ranked individual variables and kicked out any one that does not show a nice plot. Maybe the number of trees (*ntree*) during the stage-wise feedforward Random Forest variable selection was too low. It was 50 only, as a result of the huge computation time involved in such a calculation. We also clearly underestimated the time necessary to lead a solid data pre-processing: we came to realise how important and determinant this part is for the rest of the analysis.

Be it as it is: we have to say that we learned quite a lot during this hands-on modelling assignment: oftentimes the hard way through trial and error. Above all, our command of the R-software has definitely increased – as well as our appreciation of machine learning and its possible benefits in the context of business analytics.

Annex

I. Class Details and Missing Value Treatment of all Variables

Variable	Desc	% NA	Class	NA treatments
ADJMOU	Billing adjusted total minutes of use over the life of the customer	0	Numeric	no
ADJQTY	Billing adjusted total number of calls over the life of the customer	0	Numeric	no
ADJREV	Billing adjusted total revenue over the life of the customer	0	Numeric	no
ATTEMPT_MEAN	Mean number of attempted calls	0	Numeric	no
ATTEMPT_RANGE	Range of number of attempted calls	0	Numeric	no
AVG3MOU	Average monthly minutes of use over the previous three months	0	Numeric	no
AVG3QTY	Average monthly number of calls over the previous three months	0	Numeric	no
AVG3REV	Average monthly revenue over the previous three months	0	Numeric	no
AVG6MOU	Average monthly minutes of use over the previous six months	2.839	Numeric	replace with avgmou
AVG6QTY	Average monthly number of calls over the previous six months	2.839	Numeric	replace with avgqty
AVG6REV	Average monthly revenue over the previous six months	2.839	Numeric	replace with avgrev
AVGMOU	Average monthly minutes of use over the life of the customer	0	Numeric	no
AVGQTY	Average monthly number of calls over the life of the customer	0	Numeric	no
AVGREV	Average monthly revenue over the life of the customer	0	Numeric	no
BLCK_DAT_MEAN	Mean number of blocked (failed) data calls	0	Numeric	no
BLCK_DAT_RANGE	Range of number of blocked (failed) data calls	0	Numeric	no
BLCK_VCE_MEAN	Mean number of blocked (failed) voice calls	0	Numeric	no
BLCK_VCE_RANGE	Range of number of blocked (failed) voice calls	0	Numeric	no
CALLFWDV_MEAN	Mean number of call forwarding calls	0	Numeric	no
CALLFWDV_RANGE	Range of number of call forwarding calls	0	Numeric	no
CALLWAIT_MEAN	Mean number of call waiting calls	0	Numeric	no
CALLWAIT_RANGE	Range of number of call waiting calls	0	Numeric	no
CC_MOU_MEAN	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls	0	Numeric	no
CC_MOU_RANGE	Range of unrounded minutes of use of customer care calls	0	Numeric	no
CCRNDMOU_MEAN	Mean rounded minutes of use of customer care calls	0	Numeric	no
CCRNDMOU_RANGE	Range of rounded minutes of use of customer care calls	0	Numeric	no
CHANGE_MOU	Percentage change in monthly minutes of use vs previous three month average	0.891	Numeric	median
CHANGE_REV	Percentage change in monthly revenue vs previous three month average	0.891	Numeric	median
COMP_DAT_MEAN	Mean number of completed data calls	0	Numeric	no
COMP_DAT_RANGE	Range of number of completed data calls	0	Numeric	no
COMP_VCE_MEAN	Mean number of completed voice calls	0	Numeric	no
COMP_VCE_RANGE	Range of number of completed voice calls	0	Numeric	no
COMPLETE_MEAN	Mean number of completed calls	0	Numeric	no
COMPLETE_RANGE	Range of number of completed calls	0	Numeric	no
CUSTCARE_MEAN	Mean number of customer care calls	0	Numeric	no
CUSTCARE_RANGE	Range of number of customer care calls	0	Numeric	no
DA_MEAN	Mean number of directory assisted calls	0.357	Numeric	median
DA_RANGE	Range of number of directory assisted calls	0.357	Numeric	median
DATOVR_MEAN	Mean revenue of data overage	0.357	Numeric	median
DATOVR_RANGE	Range of revenue of data overage	0.357	Numeric	median
DROP_BLK_MEAN	Mean number of dropped or blocked calls	0	Numeric	no
DROP_BLK_RANGE	Range of number of dropped or blocked calls	0	Numeric	no
DROP_DAT_MEAN	Mean number of dropped (failed) data calls	0	Numeric	no
DROP_DAT_RANGE	Range of number of dropped (failed) data calls	0	Numeric	no
DROP_VCE_MEAN	Mean number of dropped (failed) voice calls	0	Numeric	no
DROP_VCE_RANGE	Range of number of dropped (failed) voice calls	0	Numeric	no
EQPDAYS	Number of days (age) of current equipment	0.001	Numeric	median
INONEMIN_MEAN	Mean number of inbound calls less than one minute	0	Numeric	no

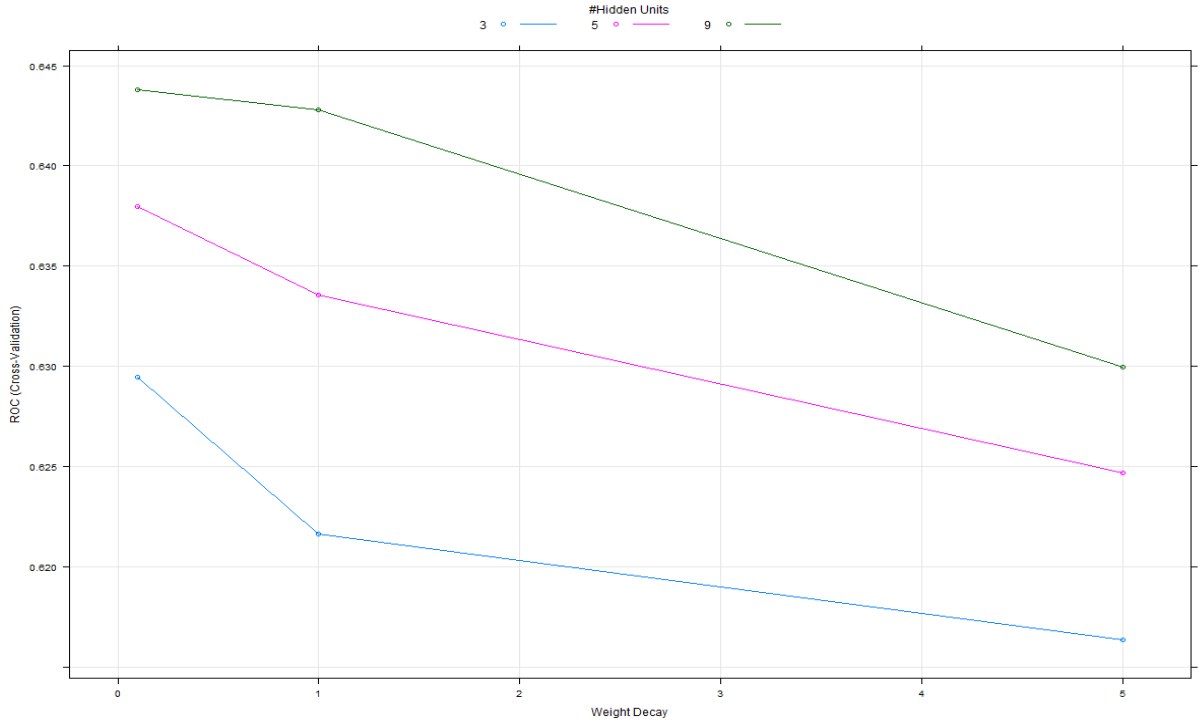
INONEMIN_RANGE	Range of number of inbound calls less than one minute	0	Numeric	no
IWYLIS_VCE_MEAN	Mean number of inbound wireless to wireless voice calls	0	Numeric	no
IWYLIS_VCE_RANGE	Range of number of inbound wireless to wireless voice calls	0	Numeric	no
MONTHS	Total number of months in service	0	Numeric	no
MOU_CDAT_MEAN	Mean unrounded minutes of use of completed data calls	0	Numeric	no
MOU_CDAT_RANGE	Range of unrounded minutes of use of completed data calls	0	Numeric	no
MOU_CVCE_MEAN	Mean unrounded minutes of use of completed voice calls	0	Numeric	no
MOU_CVCE_RANGE	Range of unrounded minutes of use of completed voice calls	0	Numeric	no
MOU_MEAN	Mean number of monthly minutes of use	0.357	Numeric	replace with avgmou
MOU_OPKD_MEAN	Mean unrounded minutes of use of off-peak data calls	0	Numeric	no
MOU_OPKD_RANGE	Range of unrounded minutes of use of off-peak data calls	0	Numeric	no
MOU_OPKV_MEAN	Mean unrounded minutes of use of off-peak voice calls	0	Numeric	no
MOU_OPKV_RANGE	Range of unrounded minutes of use of off-peak voice calls	0	Numeric	no
MOU_PEAD_MEAN	Mean unrounded minutes of use of peak data calls	0	Numeric	no
MOU_PEAD_RANGE	Range of unrounded minutes of use of peak data calls	0	Numeric	no
MOU_PEAV_MEAN	Mean unrounded minutes of use of peak voice calls	0	Numeric	no
MOU_PEAV_RANGE	Range of unrounded minutes of use of peak voice calls	0	Numeric	no
MOU_RANGE	Range of number of minutes of use	0.357	Numeric	median
MOU_RVCE_MEAN	Mean unrounded minutes of use of received voice calls	0	Numeric	no
MOU_RVCE_RANGE	Range of unrounded minutes of use of received voice calls	0	Numeric	no
MOUIWYLISV_MEAN	Mean unrounded minutes of use of inbound wireless to wireless voice calls	0	Numeric	no
MOUIWYLISV_RANGE	Range of unrounded minutes of use of inbound wireless to wireless voice calls	0	Numeric	no
MOUOWYLISV_MEAN	Mean unrounded minutes of use of outbound wireless to wireless voice calls	0	Numeric	no
MOUOWYLISV_RANGE	Range of unrounded minutes of use of outbound wireless to wireless voice calls	0	Numeric	no
OWYLIS_VCE_MEAN	Mean number of outbound wireless to wireless voice calls	0	Numeric	no
OWYLIS_VCE_RANGE	Range of number of outbound wireless to wireless voice calls	0	Numeric	no
OPK_DAT_MEAN	Mean number of off-peak data calls	0	Numeric	no
OPK_DAT_RANGE	Range of number of off-peak data calls	0	Numeric	no
OPK_VCE_MEAN	Mean number of off-peak voice calls	0	Numeric	no
OPK_VCE_RANGE	Range of number of off-peak voice calls	0	Numeric	no
OVRMOU_MEAN	Mean overage minutes of use	0.357	Numeric	median
OVRMOU_RANGE	Range of overage minutes of use	0.357	Numeric	median
OVRREV_MEAN	Mean overage revenue	0.357	Numeric	median
OVRREV_RANGE	Range of overage revenue	0.357	Numeric	median
PEAK_DAT_MEAN	Mean number of peak data calls	0	Numeric	no
PEAK_DAT_RANGE	Range of number of peak data calls	0	Numeric	no
PEAK_VCE_MEAN	Mean number of inbound and outbound peak voice calls	0	Numeric	no
PEAK_VCE_RANGE	Range of number of inbound and outbound peak voice calls	0	Numeric	no
PLCD_DAT_MEAN	Mean number of attempted data calls placed	0	Numeric	no
PLCD_DAT_RANGE	Range of number of attempted data calls placed	0	Numeric	no
PLCD_VCE_MEAN	Mean number of attempted voice calls placed	0	Numeric	no
PLCD_VCE_RANGE	Range of number of attempted voice calls placed	0	Numeric	no
RECV_SMS_MEAN	Mean number of received SMS calls	0	Numeric	no
RECV_SMS_RANGE	Range of number of received SMS calls	0	Numeric	no
RECV_VCE_MEAN	Mean number of received voice calls	0	Numeric	no
RECV_VCE_RANGE	Range of number of received voice calls	0	Numeric	no
RETDAYS	Number of days since last retention call	96.017	Numeric	replace with 1000
REV_MEAN	Mean monthly revenue (charge amount)	0.357	Numeric	median
REV_RANGE	Range of revenue (charge amount)	0.357	Numeric	median
RMCALLS	Total number of roaming calls	85.777	Numeric	replace with 0
RMMOU	Total minutes of use of roaming calls	85.777	Numeric	replace with 0
RMREV	Total revenue of roaming calls	85.777	Numeric	replace with 0
ROAM_MEAN	Mean number of roaming calls	0.357	Numeric	median
ROAM_RANGE	Range of number of roaming calls	0.357	Numeric	median

THREEWAY_MEAN	Mean number of three way calls	0	Numeric	no
THREEWAY_RANGE	Range of number of three way calls	0	Numeric	no
TOTCALLS	Total number of calls over the life of the customer	0	Numeric	no
TOTMOU	Total minutes of use over the life of the customer	0	Numeric	no
TOTMRC_MEAN	Mean total monthly recurring charge	0.357	Numeric	median
TOTMRC_RANGE	Range of total monthly recurring charge	0.357	Numeric	median
TOTREV	Total revenue	0	Numeric	no
UNAN_DAT_MEAN	Mean number of unanswered data calls	0	Numeric	no
UNAN_DAT_RANGE	Range of number of unanswered data calls	0	Numeric	no
UNAN_VCE_MEAN	Mean number of unanswered voice calls	0	Numeric	no
UNAN_VCE_RANGE	Range of number of unanswered voice calls	0	Numeric	no
VCEOVR_MEAN	Mean revenue of voice overage	0.357	Numeric	median
VCEOVR_RANGE	Range of revenue of voice overage	0.357	Numeric	median
ACTVSUBS	Number of active subscribers in household	0	numeric	no
ADULTS	Number of adults in household	0.23019	numeric	deleted
AGE1	Age of first household member	0.01732	numeric	put default value '00'
AGE2	Age of second household member	0.01732	numeric	put default value '00'
AREA	Geographic area	0.0004	factor	put value 'Unknown'
ASL_FLAG	Account spending limit	0	factor	no
CAR_BUY	New or used car buyer	0.01732	factor	put value 'Unknown'
CARTYPE	Dominant vehicle lifestyle	0.68412	factor	deleted
CHILDREN	Children present in household	0.65928	numeric	deleted
CHURN	Instance of churn between 31-60 days after observation date	0	factor	no
CRCLSCOD	Credit class code	0	factor	no
CREDITCD	Credit card indicator	0.01732	factor	median
CRTCOUNT	Adjustments made to credit rating of individual	0.965	numeric	Replacement with value 0
CSA	Communications local service area	0	string	no
CUSTOMER_ID	Unique tournament specific customer ID for scoring purposes	0	string	no
DIV_TYPE	Division type code	0.81459	factor	Replacement with value 0
DUALBAND	Dualband	0.00001	factor	Replacement with value 'U'
DWLLSIZE	Dwelling size	0.38308	factor	deleted
DWLLTYPE	Dwelling unit type	0.31909	factor	deleted
EDUC1	Education of first household member	0.86478	factor	deleted
ETHNIC	Ethnicity roll-up code	0.01732	factor	Replacement with value 'U'
FORGNTVL	Foreign travel dummy variable	0.01732	factor	median
HND_PRICE	Current handset price	0.00847	numeric	median
HHSTATIN	Premier household status indicator	0.37923	factor	deleted
HND_WEBCAP	Handset web capability	0.00001	factor	Replacement with value 'UNKW'
INCOME	Estimated income	0.25436	factor	Multinomial logit
INFOBASE	InfoBase match	0.22079	factor	deleted
KID0_2	Child 0 - 2 years of age in household	0.01732	factor	Replacement with value 'U'
KID3_5	Child 3 - 5 years of age in household	0.01732	factor	Replacement with value 'U'
KID6_10	Child 6 - 10 years of age in household	0.01732	factor	Replacement with value 'U'
KID11_15	Child 11 - 15 years of age in household	0.01732	factor	Replacement with value 'U'
KID16_17	Child 16 - 17 years of age in household	0.01732	factor	Replacement with value 'U'
LAST_SWAP	Date of last phone swap	0.58	date	Replacement with '01/01/1960'
LOR	Length of residence	0.3019	numeric	deleted
MAILFLAG	DMA: Do not mail flag	0.98523	factor	deleted
MAILORDR	Mail order buyer	0.64363	factor	deleted
MAILRESP	Mail responder	0.62889	factor	deleted
MARITAL	Marital status	0.01732	factor	Replacement with value 'U'
MODELS	Number of models issued	0.00001	numeric	median
MTRCYCLE	Motorcycle indicator	0.01732	factor	median

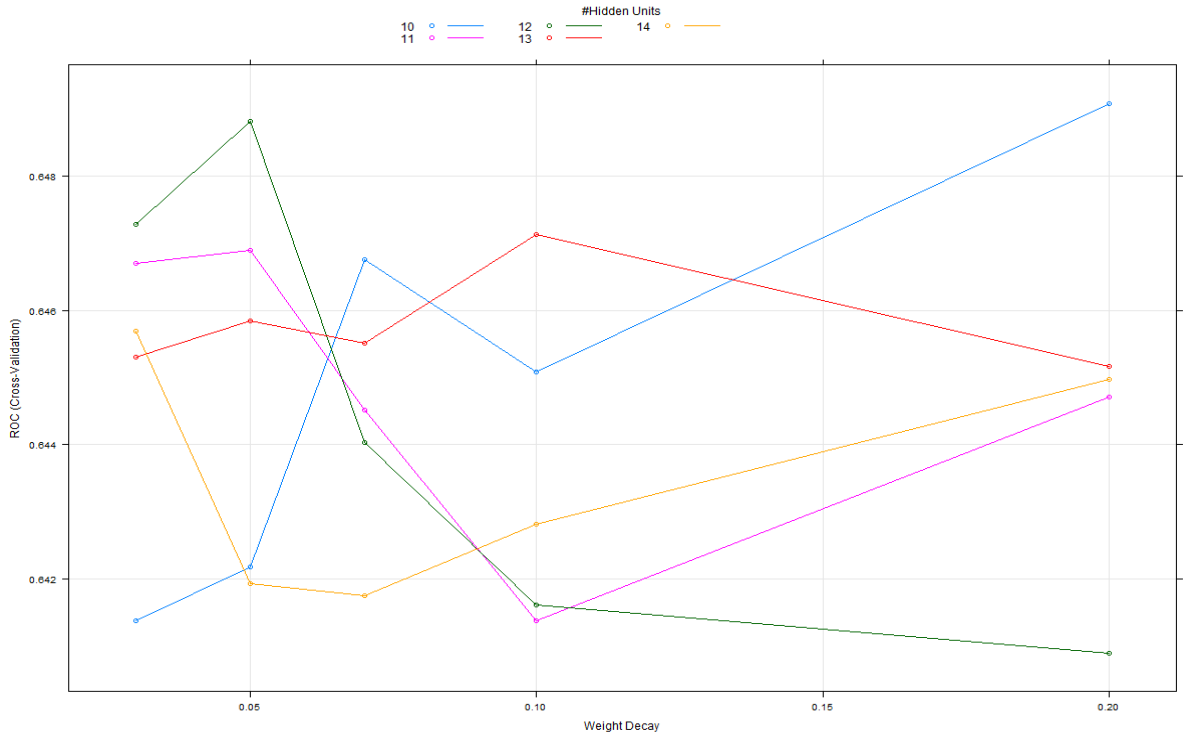
NEW_CELL	New cell phone user	0	factor	no
NUMBCARS	Known number of vehicles	0.49366	numeric	deleted
OCCU1	Occupation of first household member	0.73353	factor	deleted
OWNRENT	Home owner/renter status	0.33706	factor	deleted
PCOWNER	PC owner dummy variable	0.81534	factor	deleted
PHONES	Number of handsets issued	0.00001	numeric	median
PRE_HND_PRICE	Previous handset price	0.57515	numeric	deleted
PRIZM_SOCIAL_ONE	Social group letter only	0.07388	factor	Replacement with value 'Z'
PROPTYPE	Property type detail	0.71788	factor	deleted
REF_QTY	Total number of referrals	0.95545	numeric	Replacement with value 0
REFURB_NEW	Handset: refurbished or new	0.00001	factor	Replacement with value 'N'
RV	RV indicator	0.01732	factor	median
SOLFLAG	Infobase no phone solicitation flag	0.98039	factor	deleted
TOT_ACPT	Total offers accepted from retention team	0.96017	numeric	Replacement with value -1
TOT_RET	Total calls into retention team	0.96017	numeric	Replacement with value 0
TRUCK	Truck indicator	0.01732	factor	median
UNIQSUBS	Number of unique subscribers in the household	0	numeric	no
WRKWOMAN	Working woman in household	0.87491	factor	deleted

II. Neural Network: Grid Search Results Round 1 and 3

Grid search round 1



Grid search round 3



III. Summary of the final (18-9-0.1) Neural Network

a 18-9-1 network with 181 weights
options were - entropy fitting decay=0.1

```
 b->h1 i1->h1 i2->h1 i3->h1 i4->h1 i5->h1 i6->h1 i7->h1 i8->h1 i9->h1 i10->h1 i11->h1 i12->h1 i13->h1
  0.43  0.77 -0.03 -1.42 -0.73 -2.25 -1.57  0.01  1.48  0.73  0.43  1.33  1.30 -0.29
i14->h1 i15->h1 i16->h1 i17->h1 i18->h1
  0.07  0.10  0.30  2.26 -0.78
 b->h2 i1->h2 i2->h2 i3->h2 i4->h2 i5->h2 i6->h2 i7->h2 i8->h2 i9->h2 i10->h2 i11->h2 i12->h2 i13->h2
 -2.16  0.71 -0.73 -0.58  0.29  3.54 -1.61 -2.08 -3.56  1.06  1.18 -0.11  1.31 -0.62
i14->h2 i15->h2 i16->h2 i17->h2 i18->h2
  0.45 -0.73  0.08 -1.11  0.16
 b->h3 i1->h3 i2->h3 i3->h3 i4->h3 i5->h3 i6->h3 i7->h3 i8->h3 i9->h3 i10->h3 i11->h3 i12->h3 i13->h3
  6.50 13.88  2.21  1.48 -0.70 -0.24  0.53  0.24 -0.92  0.35 -0.33  0.04 -0.62  0.27
i14->h3 i15->h3 i16->h3 i17->h3 i18->h3
 -0.18 -0.39  0.29 -0.32 -0.12
 b->h4 i1->h4 i2->h4 i3->h4 i4->h4 i5->h4 i6->h4 i7->h4 i8->h4 i9->h4 i10->h4 i11->h4 i12->h4 i13->h4
  2.20  9.23  3.05 -0.56  1.62 -0.66  0.39  0.08 -1.17 -0.13 -0.72  0.28  0.35  0.97
i14->h4 i15->h4 i16->h4 i17->h4 i18->h4
  0.21 -0.28  0.48  0.45  0.24
 b->h5 i1->h5 i2->h5 i3->h5 i4->h5 i5->h5 i6->h5 i7->h5 i8->h5 i9->h5 i10->h5 i11->h5 i12->h5 i13->h5
  2.87  1.12 -2.41 -0.15 -0.22  0.01 -1.13  0.41  1.69  0.72  0.98  0.71 -1.94  0.24
i14->h5 i15->h5 i16->h5 i17->h5 i18->h5
  0.18 -1.22  0.40 -0.70  1.72
 b->h6 i1->h6 i2->h6 i3->h6 i4->h6 i5->h6 i6->h6 i7->h6 i8->h6 i9->h6 i10->h6 i11->h6 i12->h6 i13->h6
  2.77 -2.76  4.09  0.55 -0.53  1.01 -0.10 -0.43 -0.66  0.85 -0.31  0.10  0.41 -0.03
i14->h6 i15->h6 i16->h6 i17->h6 i18->h6
 -0.30 -0.25 -0.02 -0.70  0.44
 b->h7 i1->h7 i2->h7 i3->h7 i4->h7 i5->h7 i6->h7 i7->h7 i8->h7 i9->h7 i10->h7 i11->h7 i12->h7 i13->h7
  7.65  0.36 -0.08  7.49  1.09 -0.44 -1.01 -0.17 -0.83 -0.54 -1.24 -0.12  0.93  0.37
i14->h7 i15->h7 i16->h7 i17->h7 i18->h7
  3.15  1.44 -1.25  1.85  0.68
 b->h8 i1->h8 i2->h8 i3->h8 i4->h8 i5->h8 i6->h8 i7->h8 i8->h8 i9->h8 i10->h8 i11->h8 i12->h8 i13->h8
  2.26  4.11  4.26  0.00  0.22 -2.02 -0.99  3.65  1.10  0.10  1.46 -1.62 -0.38  2.27
i14->h8 i15->h8 i16->h8 i17->h8 i18->h8
 -0.99  0.88  0.03  2.01 -0.10
 b->h9 i1->h9 i2->h9 i3->h9 i4->h9 i5->h9 i6->h9 i7->h9 i8->h9 i9->h9 i10->h9 i11->h9 i12->h9 i13->h9
  5.02  0.23 -0.40  0.59 -0.28  0.10 -0.04 -0.25 -1.61  1.57 -0.86  0.77  0.48  2.82
i14->h9 i15->h9 i16->h9 i17->h9 i18->h9
  0.69  0.90 -0.55  1.72  0.48
b->o h1->o h2->o h3->o h4->o h5->o h6->o h7->o h8->o h9->o
-2.16  0.78  1.02  1.34 -1.04  0.20  1.07 -2.19  0.26  2.28
```