Cell2Cell – The Churn Game

The intention of this problem is to develop skills to perform logistic regression, conduct feature scaling, interpret results and understand the intuition behind logistic regressions. Also, the intent is to learn from this exercise and apply the learning in your functional area.

The case and problem originates from a case study developed by Duke University (attached below; read to understand the context). This provides you context of Wireless Telecommunication Industry in 2000’s. Customer churn was (in some cases is) a big problem in the industry where customer chose to change their carriers in record number. Those of you who are new to this industry should understand that Churn is big driver of the revenue and profit in the industry – (CLV customer Lifetime Value).



You job is to work as a consultant/s to this carrier named, Cell2Cell, in developing predictive Churn model that identifies variables which are driver of churn. Then propose a retention plan for Cell2Cell to adopt to reduce churn.

**Data Set:**

Attached is complete definition of the data fields and description of it (data dictionary).



The data set is in csv format. The data is already segregated into training and test category – field name “traintest”.



Note: This data is not a tidy data. There is N/A in the data and you would need to clean as you analyze. Be careful how you do it. As a good analytics engineer you should understand the reason you are removing N/As or null values.

Logistic Regression in R (Reference: Chapter 4 – Classification; Book An Introduction to Statistical Learning with Applications in R – Authors (Gareth Hanes\*Daniela Witten\*Trevor Hastie) ISBN 978-1-4614-7137-0 <http://www-bcf.usc.edu/~gareth/ISL/>

GLM to run the logistic regression format:

*>model <- glm (categorical\_variable~., data=dataset\_name, family = binomial)*

family=binomial : is the option that tells R to run a logistic model

~. : run the regression with all parameters

For example of running with specific 3 independent variables with revenue, mou and roam

*>model <- glm(churndep~revenue+mou+roam, data=dataset\_name, family=binomial)*

Predict() function can be used to predict the probability using the mdoel. The type=”response” option tells R to output probability for P(categorical variable=1|idependent variables).

*>predict <- predict(model, test\_dataset, type = 'response')*

if you do not specify the test\_data\_set then predict() computes the probability for training data.

BONUS Credit: Evaluating Model Generalization

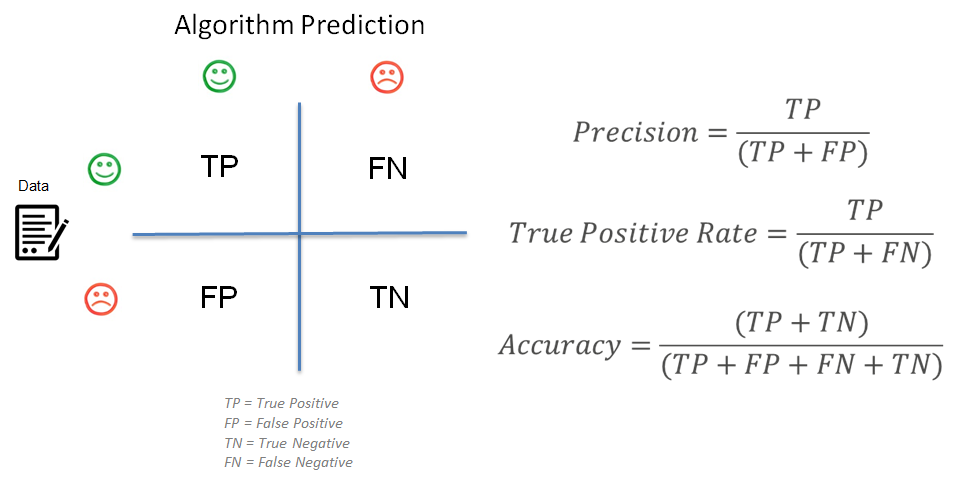
Example: K-Fold Cross Validation (use any approach in Chapter 5; Book An Introduction to Statistical Learning with Applications in R – Authors (Gareth Hanes\*Daniela Witten\*Trevor Hastie) ISBN 978-1-4614-7137-0 <http://www-bcf.usc.edu/~gareth/ISL/>)

When evaluating models, we often want to assess how well the model performance generalize to unseen new data. This is also called “model generalization”. . One technique for evaluating the level of generalization is k-fold cross-validation, which partitions the data into k equally sized segments (called ‘folds’). One fold is held out for validation while the other k-1 folds are used to train the model and then used to predict the target variable in our testing data. The distribution of errors obtained using cross-validation is an estimate of the error distribution that we could observe when using the model with brand new data.

You can use the “caret” package in R to do this or you write a loop to do it.

Model Evaluation Criteria:

Binary Classification models usually use the confusion matrix to understand different aspects of the model performance including accuracy, true positives/negatives, and false positives/negatives.



**Use confusion matrix to evaluate the model** –

>table (dataset\_name$churndep, predict > 0.5)

You can calculate model accuracy using the output of = (True Positive + True Negative)/(True positive + True Negative+ false positive + false negative)

FYI: The ultimate goal in a modeling exercise is to have good performance models (i.e. accurate) with good generalization (i.e. training error ~ testing error).

BONUS Credit if you do the following -

ROC Chart to model evaluation (use ROCR package) – Create ROC chart using the ROCR package. Typically ROC charts are used for binary classification to measure the efficacy of logistic model. Calculate Area under the curve (AUC)

<https://cran.r-project.org/web/packages/ROCR/ROCR.pdf>

**Extra Resource:**

If you have not done Logistic regression and really want to walk through an example step by step then use the following: This example works on a Kaggle dataset “Titanic” that I thought was very interesting that you could walk though once you have read the Chapter 4 on Classification:

<https://datascienceplus.com/perform-logistic-regression-in-r/>