# GBA 6210 – Data Mining for Business Analytics

# Dr. Shuo Zeng

# Case Study

# Customer Retention – Telco Customer Churn Dataset

**Description**

A telecommunication company (known as Telco company) provides subscription-based telecommunication service, which is its major revenue source. In order to grow their revenue generating customer base, it is important for a Telco company to attract new customers as well as avoid termination of existing contracts, which is known as churn. Customer turnover, or churn rate, is the percentage of a company's customer base lost during a given period of time, usually on monthly or annual basis. A high churn rate may hurt revenue and profit badly. Many different reasons may trigger customers to terminate their contracts, such as better price offers and/or more interesting packages from competitors, bad service experiences, or change of customers' personal situations.

In order to reduce churn rate, many Telco companies adopt a ***reactive*** approach: if a customer called with a request to cancel his or her contract, then the customer service representative would try to convince the customer to extend the contract, most often by offering free services or discounts on existing services. However, it would be more effective to estimate the probability that a given customer would churn in the near future, identify the factors that contributed most to that customer's decision, and then ***actively*** reach out to the customer to enhance his or her service experience and divert churn without giving up costly discounts. The goal of this project is to build a ***classification model*** using R to predict customer churn (probability and classification) for a Telco company.

**Dataset**

Dataset used in this project comes from IBM sample datasets from Kaggle (*Telco-Customer-Churn.csv*). It contains 7043 rows and 21 columns. Each row represents a customer, and each column contains one of customer's attributes, described below.

*Customers demographic information*

customerID — Customer ID
gender — Whether the customer is a Male or a Female
SeniorCitizen — Whether the customer is a senior citizen or not (1, 0)
Partner — Whether the customer has a partner or not (Yes, No)
Dependents — Whether the customer has dependents or not (Yes, No)

*Customer account information*

tenure — Number of months the customer has stayed with the company
Contract — The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling — Whether the customer has paperless billing (Yes, No)
PaymentMethod — The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges — The amount charged to the customer monthly
TotalCharges — The total amount charged to the customer

*Customer services booked*

PhoneService — Whether the customer has a phone service (Yes, No)
MultipleLines — Whether the customer has multiple lines (Yes, No, No phone service)
InternetService — Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity — Whether the customer has online security (Yes, No, No internet service)
OnlineBackup — Whether the customer has online backup (Yes, No, No internet service)
DeviceProtection — Whether the customer has device protection (Yes, No, No internet service)
TechSupport — Whether the customer has tech support (Yes, No, No internet service)
StreamingTV — Whether the customer has streaming TV (Yes, No, No internet service)
StreamingMovies — Whether the customer has streaming movies (Yes, No, No internet service)

*Classification labels*

Churn — Whether the customer churned or not (Yes or No)

**Tasks**

Please complete the following tasks for this data mining project ***with R only***.

1. *Data exploration (required)*: use proper summary statistics and visualization tools to understand the distribution of ***ALL*** variables. Please also explore the relationship between two or more variables that support your model building in step 5.
2. *Data preprocessing (optional)*: detect and remove outliers, compute new variables from existing variables if necessary.
3. *Data and dimension reduction (optional)*: choose a subset of variables for model building (dimension reduction), partition the records into homogeneous groups (using unsupervised learning techniques such as cluster analysis) and build separate models (data reduction).
4. *Partition the data (required)*: use 10-fold cross validation to evaluate model performance. Please refer to the description of *k*-fold cross validation in my slides on data mining process (Chapter 2) and implement 10-fold cross validation using "for" loop, the "sample" function, the "setdiff" function, the "union" function, and the "rbind" function from scratch (Hint: "for" loop and the "sample" function can be used together to partition the dataset, and "for" loop and the "rbind" function can be used to pool training/validation results). Please be

noted that you are ***NOT*** allowed to use existing functions that has already implemented the *k*-folds cross validation (including but not limited to: createFolds, createMultiFolds, train, and trainControl functions from the "caret" package, etc.).

5. *Model building (required)*: build at least two classification models using only the algorithms we have introduced in class, at least one of the built models must be interpretable (in terms of model structure and parameters, which essentially reflects the factors that contributed to customer's decision to churn).
6. *Model evaluation (required)*: evaluate performance of each model built. Specifically, record and pool all training outcomes and validation outcomes from all folds separately, and evaluate model performance using the following metrics and tools: accuracy, sensitivity, specificity, precision, FDR, FOR, ROC chart, AUC of ROC curve, and Lift chart. Compare the performance between training and validation and discuss whether there is any overfitting issue or not.
7. *Model deployment (required)*: choose one of the ***interpretable*** algorithms from step 5 and build a model using the whole dataset. Interpret the model from both technical and managerial perspectives.

**Deliverables**

1. A single report created using Microsoft Word (Times New Roman font, 12-point font size, 1.5-line spaced) that includes:
    a. A cover page that includes the project title and the list of team members.
    ***b.*** A ***1-page*** Executive Summary of the project. It should include the goal of the project, a description of the modeling process, and the most important, your findings and managerial insights. It should be precise and concise and ***should not*** include too much technical details.
    c. ***Detailed*** and ***organized*** documentation of description, key results (including tables, charts, etc.), and discussions for each of the tasks you have completed. ***R code must NOT be included in the report.***
    d. Convert your report to *pdf* format for submission.
2. A single R file that includes code for all tasks you have completed, in their corresponding order.

**Deadline: the Friday of week 15.**

| Telco Group Project Rubric | | | | |
|---|---|---|---|---|
| **Item (100% Total)** | **90 – 100 points (Exemplary)** | **85 – 90 points (Satisfactory)** | **60 – 85 points (Need Improvement)** | **< 60 points (Not Acceptable)** |
| **Report (20%)** | Report is well organized and contains all required components. Discussions are thorough. Language is formal, precise, and concise. | Report contains all required components. Some of the discussions are brief. Language is formal, but not sufficiently precise and/or concise. | Report is poorly organized. Some required components are missing. Most of the discussions are brief or missing. Language is not formal, precise, and concise. | A significant number of required components are missing. No discussions at all. |
| **Data Exploration (20%)** | All variables are properly summarized numerically and graphically. Correlations, especially between predictors and the outcome, are summarized. Existing variables are transformed, and/or new variables are created, with proper justification. Dimension reduction/data reduction are explored with proper justification, possibly with the help of unsupervised learning techniques. Discussion is thorough. | All variables are summarized numerically and graphically, with a few errors in visualization. Some of the correlations are summarized. No existing variables are transformed, and no new variables are created. No dimension reduction/data reduction is explored. Discussion is brief. | Only some of the variables are summarized. There are many errors in numerical and/or graphical summaries. Correlations are not summarized. No existing variables are transformed, and no new variables are created. No dimension reduction/data reduction is explored. Discussion is missing. | Missing numerical and graphical summary of a significant number of variables. Correlations are not summarized. No existing variables are transformed, and no new variables are created. No dimension reduction/data reduction is explored. Discussion is missing. |
| **Modeling (20%)** | Three or more models are built, with careful selection of predictors. Models are implemented correctly. Discussion on the selection of machine learning techniques is thorough. Choice of the final model deployed is properly justified and thoroughly discussed. | Two models are built. Models are implemented correctly. Discussion on selection of machine learning techniques is proper but brief. Choice of the final model deployed is properly justified but the discussion is brief. | Only one model is built. Choice of machine learning technique is not proper and/or is not justified. Model implementation is on the right track, with some errors. No discussion about the choice of the final model deployed. | Modeling part is missing or is significantly incomplete. |
| **Model Evaluation (20%)** | 10-folds cross validation is correctly implemented without using existing function. Each model is evaluated with the required metrics and plots. Discussion about evaluation results is proper and thorough. Advanced topics are explored and/or discussed, such as oversampling, etc. | 10-folds cross validation is mostly correctly implemented without using existing function. Each model is evaluated with the required metrics and plots, with some errors. Discussion about evaluation results is brief. Some metrics are not discussed. | 10-folds cross validation is not correctly implemented. Some of the required metrics and plots are missing or not correctly implemented. There is no discussion about the evaluation results. | Model evaluation part is missing or is significantly incomplete. |
| **R Code (20%)** | All tasks are implemented in R. Code is well organized with no error. Necessary comments are added. | Most of the tasks are implemented in R. Code is organized, but without any comment. There are less than 3 errors. | A significant number of tasks are not implemented in R. There are more than 3 errors. | The project is not implemented using R, or the code is missing, or the code does not run. |