ISOM 454/554 Final Project

The time has come for you to start working on your final project for ISOM 454/554. There are two data sets you can choose from. The first one is provided by a lender on loan repayment and the second one is provided by a telecommunications company on customer churning. Please choose **one of them** to work on.

You may work in a group of **up to 2** and please submit the following:

1. Slides presenting your analysis, findings, and recommendations for improved customer retention. Assume you are presenting to management, that has some understanding of the topics in this class. The slides must be well organized and self-contained, with clearly labelled tables and figures.
2. A video of the group presenting those slides. Presentation should be no more than 5 minutes, and all group members are expected to participate.
3. Clean, organized and clearly commented R code, so that I can quickly reproduce your analysis if needed. ALL results shown on your slides must be reproducible from your code.

The final project will be graded based on correctness and completeness of the analysis, and quality of the presentation including the slides and the video presentation. This project is similar to tasks you'll soon face in real life. Have fun!

(I) Loan repayment (LM)

A lender is interested in understanding how to best predict loan repayment. Your task is twofold with:

(T1) Inference: to help them understand the risk of non-repayment and propose a targeting plan for future loans;

(T2) Prediction: to build a predictive model with high prediction accuracy. For this task, you can create your own training and testing data split for prediction accuracy evaluation. You can try a deep learning model to take this project to the next level but this is not required.

The data include information about loan information such as Loan Type, FICO, Years In Business, ...

Your response variable of interest is **PRSM** (performance ratio at six months). Note that this column is not included in the file but can be computed as

$$PSRM = 2\frac{\text{Amount repaid at six months}}{\text{Total amount to be repaid}}.$$

PRSM should be approximately equal to 1 if the payments at 6 months are on track to fulfill the debt at the end of the year. Values of PRSM < 1 indicate a loan for which the payments are currently coming in slower than expected; PRSM > 1 indicates a loan that is being paid off faster than expected.

You can read in the data and compute PRSM as follows.

```
loans <- read.csv('https://zhang-datasets.s3.us-east-2.amazonaws.com/PRSM.csv')
loans$PRSM<-2*as.numeric(loans$Amt.Repaid.at.6.Months)/loans$Total.Amt.to.be.Repaid
```

Discussions with the lender have suggested several subtle issues that should be addressed when modeling these data. These comments from the lender may suggest variables that will be useful to you in modeling the performance of loans.

(a) Merchants that seek loans from your client often operate in distressed neighborhoods. The lender believes that improvements to the economic condition of the neighborhood (income, jobs, housing, etc) provide an environment in which the merchant will more easily be able to keep up with the target payment stream.

(b) The lender feels that some independent service organizations (ISOs) provide much better (or much worse) customers than others. The lender would very much like some evidence to either support or contradict this suspicion regarding differences among ISOs. Are there particularly good or bad ISOs among those represented in your data?

(c) There is a strong belief that overly aggressive commissions could be a red flag because the independent sales reps may have a propensity to push such loans, an example of the principal-agent problem.

(d) Lenders commonly make use of information from a credit bureau, such as the FICO score, as a means to judge the ability of a borrower to repay a loan. This information

is believed to be particularly useful when dealing with new loans but may not be so useful for repeat loans.

(e) There has been much discussion within the lender of the relevance of the tradelines variable. Some argue that having many tradelines is a good sign in that it proves that the merchant has been actively seeking credit and is a bona fide entity, while others suggest not needing tradelines is a positive sign because the merchant has not required credit to operate.

(f) Past performance is often a good indicator of future success (or failure). It is hard to believe that delinquencies against the merchant would not in some way be associated with loan performance.

(II) Customer churns (GLM)

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. Your task is twofold with

(T1) Inference: to help them understand what kind of customers are leaving and propose a retention plan to decrease churn and improve revenues;

(T2) Prediction: to build a predictive model with high prediction accuracy. For this task, you can create your own training and testing data split for prediction accuracy evaluation. You can try a deep learning model to take this project to the next level but this is not required.

The data can be read in here - 'https://zhang-datasets.s3.us-east-2.amazonaws.com/telcoChurn.csv' - and includes information about:

- If a customer left - the column is called Churn (i.e., Churn=1 means already left)
- How long they've been a customer - the column is called tenure
- Services that each customer has signed up for - phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information - contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers - gender, age range, and if they have partners and dependents

Your response variable of interest is **Churn**.

```
churn <- read.csv('https://zhang-datasets.s3.us-east-2.amazonaws.com/telcoChurn.csv')
```

Note: when you consider variables to include in the model, be careful of the nested ones. For example, **PhoneService** shows whether an account has phone service and **MultipleLines** shows whether an account has multiple lines if it has phone service. So it does not make sense to include both of them in the model because the information in **PhoneService** is nested in **MultipleLines**. Same thing with **InternetService** and **OnlineSecurity**, **OnlineBackup**, **DeviceProtection**, **TechSupport** and **StreamingMovies**, in that these additional add-on features are only relevant for accounts that have **InternetService**.