**STAT 430 Project Proposal**

**A. Group Members:**

* Huanhuan Yue (hyue4)
* Bian Yuan (ybian3),
* Taer Du (taerdu2),
* Xiuqi Jiang(xjiang40)

**B. Title of the project:**  Analysis on Default of Credit Card Clients

**C. Basic description**

a.default of credit card clients Data Set

1) Data source:<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#>

2) Number of variables in the data sets: 23 variables

3) Information: This data set contains binary and explanatory variables about customers default payments in Taiwan in 2016. It shows six data mining methods to predict the accuracy of probability of default payment (Yes = 1, No = 0), as the response variable. This data reviewed the literature and used the following 23 variables as explanatory variables:

* X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
* X2: Gender (1 = male; 2 = female).
* X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
* X4: Marital status (1 = married; 2 = single; 3 = others).
* X5: Age (year).
* X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
  + X6 = the repayment status in September, 2005;
  + X7 = the repayment status in August, 2005; . . .;
  + X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
* X12-X17: Amount of bill statement (NT dollar).
  + X12 = amount of bill statement in September, 2005;
  + X13 = amount of bill statement in August, 2005; . . .;
  + X17 = amount of bill statement in April, 2005.
* X18-X23: Amount of previous payment (NT dollar).
  + X18 = amount paid in September, 2005;
  + X19 = amount paid in August, 2005; . . .;
  + X23 = amount paid in April, 2005.

**D. Background Information:**

At present, the default rate of credit card is increasing around the world. People always spend the money on unexpected fields, which causes them don’t have ability to repay at the maturity date. Therefore, we are curious about which factors would let people spend money than expected and what kind of people are more likely to be holder of default credit card. In this data, we could predict the default rate of a specific credit card holder by using five mining methods in the data.

**E. Statistical Learning Task:**

* Aim:
  + **Classification** of the default of credit card clients using all predictors in the dataset.
  + **Compare** their performance of various algorithms on the prediction
  + Know which predictors are the most **influential** among all the predictors
* Steps to achieve the aim:
  + Data Cleaning: clean the data as needed in order to further the analysis
  + Apply Statistical Methods: apply several statistical methods
  + Method Validation: validate the models and select the one performs the best
  + Interpret Result: understand the result and offer interpretation of the data and visualization

We want to do using several statistical learning methods and subsequently compare their performance on the prediction. We also want to know which predictors are the most influential among all the predictors.

**F. Method:**

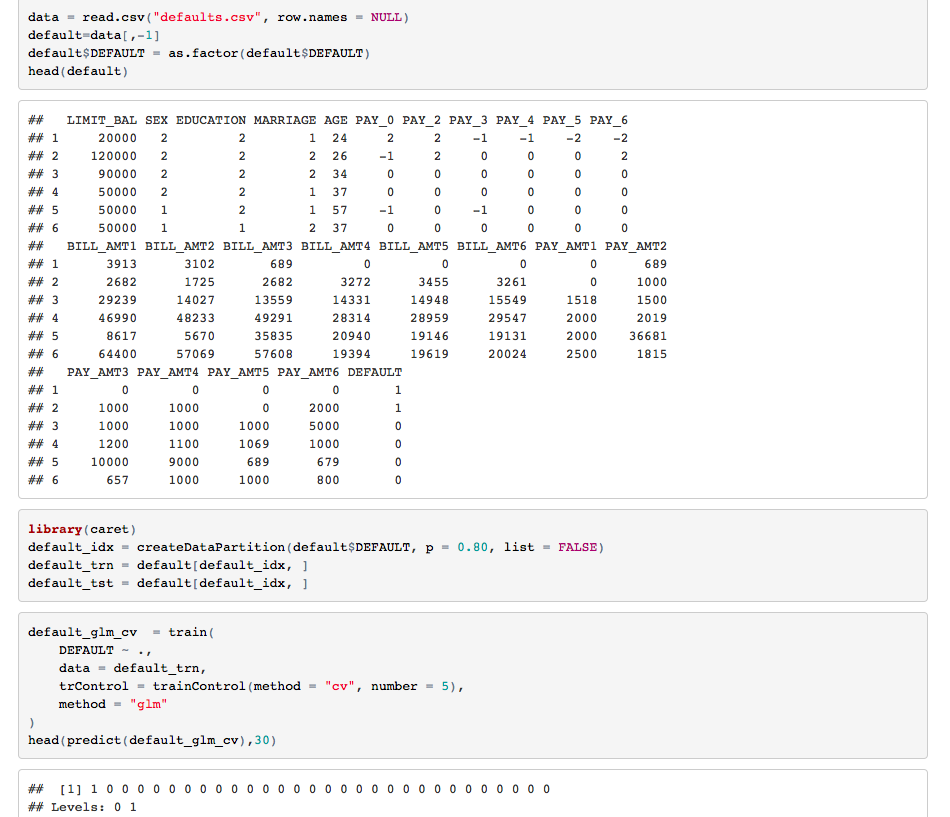
Generalized Linear Model (GLM), K-nearest neighbor, Decision Tree, Random Forest, Boosted Tree

**G. Reasons for the Chosen Methods:**

* Generalized Linear Model (GLM): We use GLM instead of linear model (lm) because glm is a more general version of a linear model, it can do both regression and classification. But lm can only do regression. In our data set, the associated tasks is classification, so GLM is suitable for our task.
* K-nearest neighbor: KNN is very powerful classification method used in pattern recognition. It stores all available cases and classifies new cases based on a similarity measure. Unlike LDA and QDA, KNN is nonparametric so it does not anything about the distribution of the dataset.
* Decision Tree Model: A decision tree model is easy to interpret and can be easily converted to a set of rules. They can classify categorical data and require no prior assumptions about the data. Because of the advantages listed above, the decision tree approach is extensively utilized for classification purpose.
* Random Forest and Boosted Tree: Random forest is an ensemble method that combines two ideas in order to improve the performance of decision tree. The first idea is bagging, or bootstrap aggregation. Instead of learning a single decision tree, bagging resamples the training dataset with replacement T times. The second idea is to further reduce the correlation among each of the trees. Thus, it would be of interest to use the model.

**H. Data cleaning regarding the original data set**

* Firstly, we deleted the first row of the original data set. This allows us to convert all values to numeric..
* We then deleted ID column because ID value does not provide any useful information.
* Besides, we changed the response variable name “default payment for the next month” to “DEFAULT”. Now all the variables have same format, and is ready for next step.

**I. Evidence that the data can be loaded and modeled in R:**