MGT 6203 Group Project Proposal Template

**Please edit the following template to record your responses and provide details on your project plan.**

**TEAM INFORMATION (1 point)**

**Team #:** **15**

**Team Members:**

1. Team Member 1 Grace Ye; GT Id: 903839619(gye37)

I’m currently working at an insurance company as an actuarial analyst, with a bachelor’s degree in mathematics.

1. Team Member 2 Yue Yuan; GT Id:903853476(yyuan375)

I am working at banking industry with master and bachelor in accountancy. I am working as financial system analyst now

1. Team Member 3 Juan Zhang; GT Id:903856394(jzhang3410)

I previously worked at bank financial department, with a master degree of finance. I had successfully completed programming courses online, using different tools such as Python and SQL step by step. And I have experience with R after I completed the courses ISYE6501 and ISYE6414.

1. Team Member 4 Liqing Jing; GT Id: 903844083(ljing32)

Senior quantitative analyst at Sun Life Financial model validation and analytics. I worked on different financial models including machine learning model, with Masters of software engineering, finance, mathematics and statistics. Experienced in different software languages including Python and R

1. Team Member 5 Qinghua Lin; 903851546(qlin81)

I’m currently working at a bank in a Fraud detection analytics role, with a BCom and Master’s degree both in Finance.

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:** Credit Card Default

**Background Information on chosen project topic:**

people are using credit card all over the world, the risk of default payments is high and predictive accuracy of probability of default would be helpful for banks. How can we prevent the risk of default payments and predict the probability of default? That would be the research we will focus on

**Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):** The dataset is called Default of credit card clients provided by UCI machine learning. We are going to research based on the variables we have from the dataset. Are default payments related to 23 variables (Amount of given credit, Gender, Education, Marital Status, Age and etc?

**State your Primary Research Question (RQ):**

The research topic will explore the likelihood of an individual defaulting on credit card payment.

**Add some possible Supporting Research Questions (2-4 RQs that support problem statement):**

1. Is there a correlation between credit limit and defaulting on credit card payment?
2. Which age group (young, middle, senior) is more likely to default on credit card payment?
3. Are married or single people more likely to default on credit card payment?
4. Is there a correlation between delayed payment history and defaulting on credit card payment? Are people who constantly delay on payment more likely to default on credit card payment?
5. Is there a correlation between education level and defaulting on credit card payment? Which education level (graduate school, university, high school, others) is more likely to default?
6. Is there a higher probability of defaulting if the individual delayed in payment in the previous month?

**Business Justification:** **(Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)**  
Banks can have serious impact from loan defaults. Based on the dataset, the total billed amounts from April 2005 to September 2005 is 8,095,850,136 (NT dollar), and the payment amount is 949,541,777 (NT dollar), only 11.73% of the total billed amount. Being able to identify which individual is more likely to default on their credit card bills can help banks better manage default risks and balance reserves.

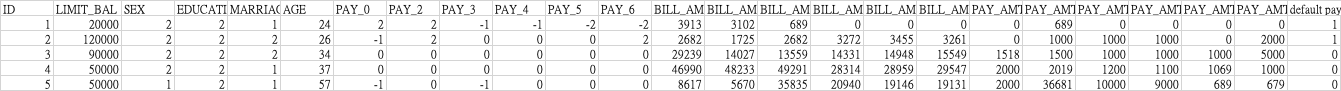
**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

Data source 1: <https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

Data source 1 includes characteristics of each individual such as gender, age, marital status and education level as well as the individual’s credit limit and past bill amounts and payments.



**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

Default payment next month in column Y will be considered dependent variable. Limit balance, sex, education, marriage, age, history of past payment, amount of bill statement and amount of previous payment will be considered independent variables. We are adding months delayed in payment which is the count of history of past payment and maximum months delayed which is the maximum months of past payment history. We are splitting age category into three binary variables consisting of young age(21-40), middle age(41-60) and senior age(61-79) and previous payment into binary variable of delay in each month. We hypothesize months delayed in payment and maximum month delayed being the most important predicting variables.

**APPROACH/METHODOLOGY (8 points)**

**Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))**

For data transformations, we convert the sex, education, marital status and age into the dummy variables. And we would calculate the difference between past payment and bill statement to create the new features “delay in month m\_n”, the dummy variables in which 1 means default.

We need to predict whether the clients would default based on given predictors. We try the models used for classification. we could try the KNN, Decision Trees, Random Forest, XGBoost model and Times Series model.

For the KNN, we can plot all the features on the graph where the horizontal axis shows education, the vertical axis shows the number of months had delays in prepayment and each previous client is either a blue data point if they finished the payment or a red point if they defaulted, assuming that each new client is similar to closest previous clients. Firstly, we need to choose the way to measure distance. Secondly some features might be more important than others to the classification, so we need to weight each dimension’s distance differently. The larger the weight, the greater the impact is on the distance. Thirdly, we fit a KNN model given the kernel and the value of parameter K using Leave-one-out Cross Validation. Find whether the model classifies the data point correctly. Apply the same method for all data points and find the misclassification rate.

For the decision trees, we partition the feature space into a set of rectangles and fit a simple model in each one. Each rectangle takes the value 0 or 1 with other specific rule. Then we could read the decision tree. The trees would have low bias but large variance. The Random Forest Model would be averaging. But we may try the Lasso Regression Model, Elastic Regression Model, Stepwise Regression Model as well. After that, we could compare the model, using accuracy, sensitivity and specificity.

For the Logistic regression models, we could start with exploratory data analysis, providing the significant differences in proportions for each group in the predicting variables like sex, education, marriage and age. Our responses are binary. Firstly, we test the linearity and independence assumptions. Then we explore the most significant factors on default and find the model goodness of fit and predictive power. To improve the goodness of fit, we may add predicting variables, transform predicting variables to improve linearity, identify unusual observations (outliers, leverage points) or try other S-shape functions.

For the XGBoost model, We will also use XGB algorithm to classify the next month default payment. XGBoost (Extreme Gradient Boosting) is an ensemble machine learning algorithm used for classification tasks efficiently. XGBoost is widely used in real-world applications due to its speed and accuracy. The features used in Xgb model are same as the above mentioned models.

For Times Series model, In this project, the next month default payment depends on the previous default payment history. We will use time series model(The exponentially weighted moving average (EWMA)) to see if the most recent history of past payment has more important role in the next default payment. The time Series model is a statistical technique used to analyze and predict patterns in sequential data points collected over time.

Then we could use the ANOVA to see differences in means across the age (21-40,41-60,61-70), education(graduate school, university, high school and others), gender(male and female) and marital status(married, single and others). Also we could compare the models using ANOVA, testing for subsets of coefficients. Does age alone significantly help in predicting default? Does the addition of X1(amount of given credit) significantly contribute to the prediction of default after controlling for the contribution of age?

For cross validate the models, we will compare the different models we choose, we will use K-Fold Cross Validation or Leave-One-Out Cross Validation to evaluate different models. Because this is a classification project, we could compare different models using ROC value based on the confusion matrix.

For optimizing the model hyper-parameters, we will use grid search algorithms in R to find the optimized hyper-parameters which with the higher ROC value of this model.

For training and testing model process, we will train different models by primarily splitting up the data sets into training/validation and test data sets(80/20). We would use the training/validation data set to train and cross validate (including K-Fold Cross Validation or Leave-One-Out Cross Validation) different models and ensure the selected model is performing effectively. Then we would use the test data set to evaluate the selected model.

**Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement**

Based on the research questions listed above, we anticipate that some factors would be potential strong patterns to help predict default. A smaller difference between the given credit amount and default payment may increase the likelihood of default. Middle age clients have higher chance of default than younger age and senior age because they tend to have more pressure and responsibilities and may have mortgages and family to support. Same reason for married people as they have more expense than single. Clients have delay history in the past and the more they delayed the higher chance they will delay and default in the upcoming month. Most recent history would have higher impact as it reflects client’s recent financial problem. Education level may affect too, clients with higher education may have more stable jobs and less likely to default.

**What business decisions will be impacted by the results of your analysis? What could be some benefits?**

We can adjust and improve our models to better predict default by selecting the right and most important features, for example if we know middle age married clients with bad history and lower education level are more likely to default, we can increase the risk ratings on these factors and so we would improve the model accuracy and lower false positives. And when our models perform well, it will further reduce loss from client default and save money. In addition, our analysis may also impact the risk ratings when new clients applying for new credit products. Clients will be classified as higher default risk if they meet those risk criteria we conclude from our analysis, this will help with the decision making whether or not to approve the application and prevent future loss.

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

Group Project Progress Report is due 11/4

Data gathering competed in Sep

Clean and manipulate Data and Explore Data

Create sampling methods and data separation into train/test/validate sets

Create exploratory models

Evaluation models

Agree with methodology for determine the best model

Group Project Final Report is due 12/3

**Appendix (any preliminary figures or charts that you would like to include):**