**IS590 Final Project Report**

Team member: Lingsong Gao, Chujie Qin

**1. Approaches to solve the problem**

1) Tableau for KPI tracking: We build 6 different worksheets and 1 dashboard to visualize the data. The visualization focuses on finding the variation trend of the data.

2) SQL for extracting desired attributes: We use SQL to build a dataset for our model with predictors we think are important for donor prediction, including age, gender, married, events, and volunteer. Age and married are directly extracted, gender is changed to numerical binary attribute with 1 representing MALE and 0 for FEMALE. Events and volunteer are created attributes representing how many events a constituent attended and whether he or his relatives are volunteers. And we also generate the target binary attribute donor with 1 if he appears in Transaction table or 0 if not.

3) R for logistic regression model: Using the dataset derived from SQL, we then build a logistic regression model in R for donor prediction. We use the glm model in R package with several assumptions between our attributes. After assessment and modifications, we build our model based on 3 attributes: age, married, and volunteer. Our model is proved to have good performance for donor prediction.

**2. Assumptions at the beginning**

*Assumptions for attribute selection:*

We assume several attributes may influence whether a constituent is a donor or not, and we used the 5 attributes discussed above in our model analysis.

*Assumptions in model construction:*

We have three assumptions in our logistic regression. 1) Linearity 2) Independence of errors 3) Multicollinearity

**3. Assumptions Verification**

*Assumptions for attribute selection:*

Age, married and volunteer do influence the chance of a constituent becoming a donor. Events can make perfect predictions but it causes complete separations so we exclude it from our model. Gender proves to be an insignificant attribute.

*Assumptions in model construction:*

1) Linearity is satisfied for our only continuous attribute age in our model. 2) Independence of errors is satisfied. Cases of data are not related. 3) Multicollinearity is satisfied by checking VIF statistics.

**4. Project process**

1) Tableau: We first import the datasets. For each worksheet, we join two or three datasets to filter the data we need to show the the variation trend of the data. For dashboard, we use those worksheets and image designed in PowerPoint.

2) SQL: In general, SQL should be an easy task in our project, but it also has some difficulties especially for generated attributes: donor, events and volunteer. We need to use substantial group by, join, union and case clause to generate these attributes. Otherwise, extracting the age, married and gender attributes are simple tasks.

3) R: One of the difficulties in our project is the logistic regression model in R. Building the model is not difficult by simply using glm() function fed with data. However, the model assessment can be a little tricky when there are many things to be considered. We have met the complete separation and insignificant attribute to let us consider how to improve our model. We also encountered the situation of p-value = 1 and no outlier residuals, which should mean the perfection of our model, but we also have concerns about these perfections and we wonder if these are really good things or not.

**5. Tableau dashboard URL**

<https://tableau.admin.uillinois.edu/#/views/Gao_Qin_FinalProject/Gao_Qin_FinalProject>