# Final Project - Predictive Modeling of Credit Risk

Jacqueline Liu, Steven Chen, Zhongling Jiang

December 4, 2016

## 1 Methods

### 1.1 Logistic Regression

*Logistic regression* is one of the most commonly used tools for applied statistics and discrete data analysis. In this model, we have a binary output variable Y, and we want to model the conditional probability

$$Pr\,(Y = 1 \,|\, X = x) \tag{1}$$

as a function of $x$; any unknown parameters in the function are to be estimated by maximum likelihood.

Formally, the model logistic regression model is that

$$log \frac{p(x)}{1 - p(x)} = \beta_0 + x \cdot \beta \tag{2}$$

Solving for p(x), this gives

$$p(x) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} \tag{3}$$

To minimize the mis-classification rate, we should predict Y = 1 when p > 0.5 and Y = 0 when p < 0.5. This means guessing 1 whenever $\beta_0 + x \cdot \beta$ is non-negative, and 0 otherwise. Therefore, the decision boundary separating the two predicted classes is the solution of $\beta_0 + x \cdot \beta = 0$.

### 1.2 Gradient Boosting Machines

We also approached the problem from a regression point of view. The idea behind *gradient boosting* is to combine weak learners in an iterative fashion in order to create a stronger model. We used decision trees as our base model, as they are the most popular for this method of learning. Our goal is to find a model **M** that predicts the label (in this case, our default rate)

# 2 Analysis

## 2.1 Logistic Regression

Since Logistic Regression is a classification method, we'll have to transform our labels, *CDR3*, into binary labels. Based on other papers, we set our threshold at 0.15, which default rates above that considered 'high' risk and those below 'low' risk. We then performed cross-validation to gauge the accuracy of logistic regression on this particular problem. We calculated three statistics for each iteration of cross validation: precision, recall, and F1-score (see **Results** for more information on these values).

## 2.2 Gradient Boosting Machines

# 3 Results

## 3.1 Logistic Regression

```
> stats.df = read.csv("../data/logistic-result.csv")
> xtable(stats.df, caption = "Logistic Regression Statistics of 5-Fold Cross Validation")
```

| | X | Precision | Recall | F1 |
|---|---|---|---|---|
| 1 | 1 | 0.71 | 0.81 | 0.76 |
| 2 | 2 | 0.70 | 0.81 | 0.75 |
| 3 | 3 | 0.72 | 0.84 | 0.78 |
| 4 | 4 | 0.72 | 0.82 | 0.77 |
| 5 | 5 | 0.74 | 0.82 | 0.78 |

Table 1: Logistic Regression Statistics of 5-Fold Cross Validation

Precision is the percentage of the positive predictions were correct. Recall is the percentage of positive cases that were correctly classified. Lastly, F1-score considers both precision and recall. The 5-fold cross validation results are displayed in the below table, and on average of the 5-fold, the F1-score is around 76 percent, which is not bad.