Statistical Learning and Bankruptcy Prediction

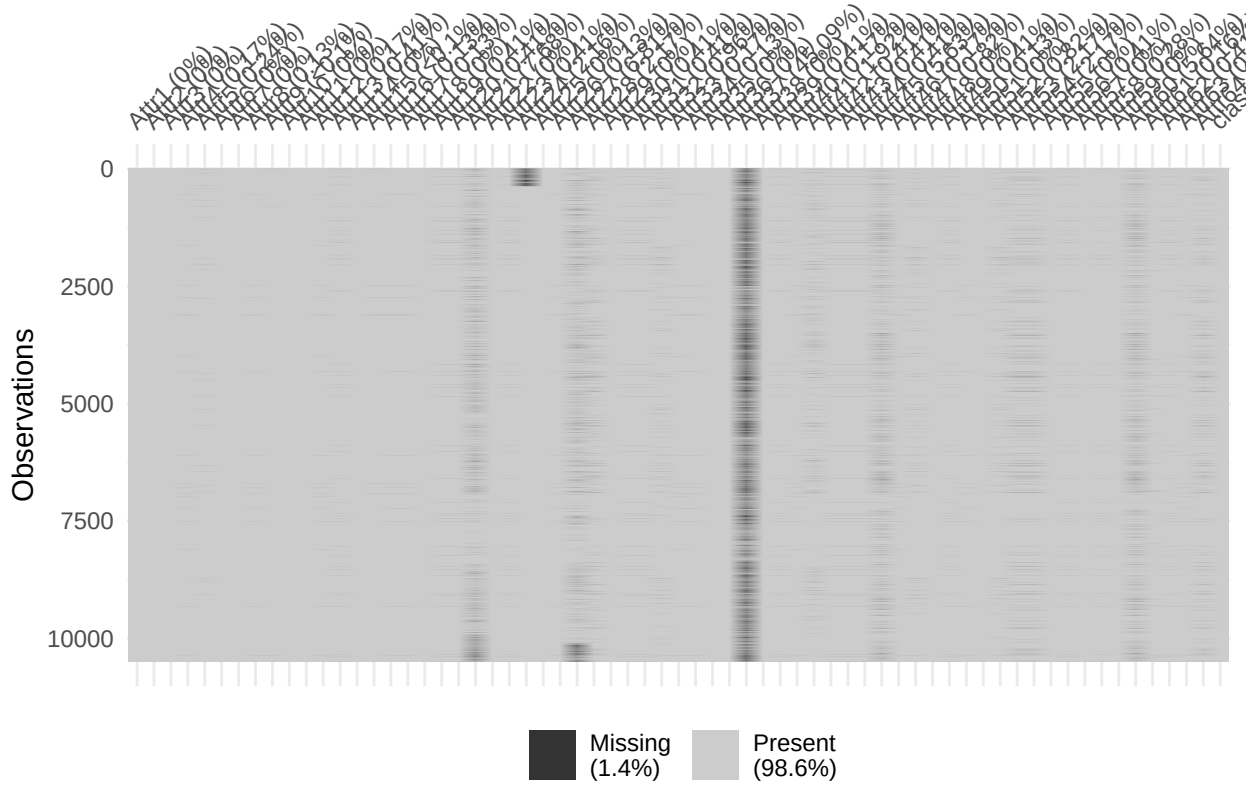
STAT432 Final Project

Group Stepanov

4/1/2019

Prepare packages

Missing Values

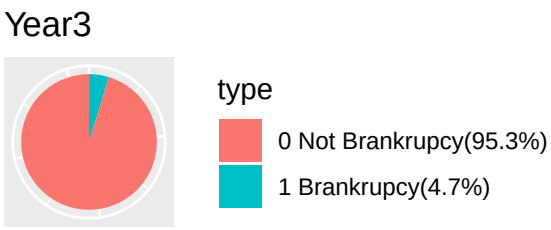



```
## [1] 5618
```

We first conduct basic data preprocessing. Missing values for each dataset are shown in the graph below. Due to the large number of missing values in each dataset, completely delete missing values will result to a large amount of data loss. Thus, we use variable means to replace missing values. We also drop the first variable `id` and factorize variable `class`.

Imbalance Data

Pie Charts to show the imbalance in response variable



The pie charts above show that the data is imbalanced. It has 0 with above 95.3%. The we use the SMOTE method to oversample the minority group and achieve a more balanced dataset.

SMOTE Algorithm For Unbalanced Classification

```
##
##     0     1
## 10008   495
```

```
##
##    0    1
## 7425 5445
```

By applying SMOTE, the new data set is more balanced.

Finally, we test 1. NA values, 2. Data Imbalance

```
## [1] TRUE
```

```
##
##    0    1
## 7425 5445
```

Data Modeling

Logistic Regression

Full model logistic regression

Gini Coefficient