**DSE 511 Final Project Report**

**Lancaster Barnstormers:**

Hairuilong Zhang

Inzamam Haque

Verónica G. Melesse Vergara

Tom Allemeier

## 1 Introduction

The objective of this project is to do a deeper dive into the data by performing some statistics and machine learning to try and extract more patterns from the data. The main idea of our group is to predict a loan status to see if it is a good or bad loan based on all the associated features with the loan, such as client characteristics, account statistics and so on. Generally speaking, this is a binary classification problem. In view of the fact that our original data has a severe class imbalance, with 606 good labels and 76 bad labels, we decided to implement the Synthetic Minority Oversampling Technique (SMOTE) to compare performance between original data and oversampled data.

## 2 Data Preparation

This section will briefly introduce the information of the dataset and the process of preparing the data used for fitting in different classifiers.

## 3 Methodology

Three commonly used machine learning classifiers are implemented respectively on the original and oversampled dataset.

### 3.1 k-Nearest Neighbors (kNN)

### 3.2 Support Vector Machine (SVM)

### 3.3 Random Forest (RF)

## 4 Performance Comparison

## References

## Appendix