Credit Card Default Prediction

Adrian Sandoval-Vargas

Computer Science, Oakland University

**Abstract**

This study report was aimed to encompass *“The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”* by I-Cheng Yeh and Che-hui Lien [1] using statistical analysis, data mining techniques, and limiting the classification algorithm to Random Forest. The data used was 30,000 credit card holders from Taiwan in a period of 6 months [2]. In this study we will investigate which process yields the best accuracy.

1 Introduction

This report focuses of the use of the Random Forest [3] Classifier for the primary reason that our data is structured in such a way that it fits the supervised schema for the ensembled Random Forest Classifier. The data was obtained from the University of California – Irvine in the form of an excel spread sheet. The classifier should classify if users will default or not based on this data. I will first preprocess the data and clean the data with any outliers in specific area of interest and provide some visuals on this cleaning process. I will keep the original cleaned data and create segmentation of the data into different groups that best depicts a boundary. I will also show the effects of Data Normalization and Principle Component Analysis. I will discuss methods