Project Proposal
I523/E534: Big Data Application
Fall 2018

Jatin Bhutka (jdbhutka)
Uma B Kota (umabkota)

October 5, 2018

1 Problem Statement

In 2006, Taiwan saw credit card debt crisis, the banks of Taiwan over issued credit cards in pursuit of increasing their market share. Also, many cardholders consumed more than their capacity thereby accumulating heavy debts. Utilizing this data, we aim to classify the defaulters by performing different statistical analyses, to help predict an individual customer's credit risk. We believe this analysis will help us better understand the route cause behind defaulting and also provide a better way to classify an applicant given his data.We'll be working on different modelling techniques like k means, logistic regression, naive Bayes and artificial neural networks to get provide our solution.

2 Data

The Default of credit card clients is multivariate Data Set from UCI machine learning Repository.It consists of 30,000 records with 23 features. The features includes amount of given credit, Sex, Education, Marital status, Age, History of past payments, amount of bill statements, amount paid.

3 Questions

- what is the best evaluation metric for the models and it's precision, recall or specificity and sensitivity.
- Which is the best model to predict the defaulters.
- Which features influence most to the default payments.
- How this can help to minimize bank's risk.
- We will analyze data to answer more questions.

4 Timeline

- Data Pre-processing – we will be using python for this

- Exploratory Data Analysis

- Data modelling

  1. k-means
  2. Logistic Regression
  3. Naive Bayes
  4. Artificial Neural Network

- Modelling Analysis

- Evaluation and Conclusion

References

[1] I-Cheng Yeh and Che-hui Lien, *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*