1 Theme: Big Data

2 Subject: Credit Risk Analysis with Big Data

3 Background and Motivation

Credit risk is an important and widely studied topic in the bank industry for lending decisions and profitability. Credit scoring has become one of the main analytical ways for financial institutions to assess credit risk. Traditionally, the score is calculated using statistical models (i.e., logistic regression), nonparametric statistical models (i.e., k-nearest neighbor), classification trees or neural networks.

With the fast growth of big data and internet, internet companies have started to step into the financial industry. For example, Sesame score from Alibaba collects customer basic information; real-time transaction data from Alibaba's shopping websites as well as banks and other companies. Sesame score ranges from 350 to 950. The higher the score, the more trustworthy a person is. Zest Finance targets Americans who do not have enough credit history and are unable to apply for loan from traditional banks. Zest Finance collects multi-dimensional datasets and then transforms the raw data to over 70,000 features. These features are then fed into over a dozen machine learning models to generate a credit score. These multi-dimensional models outperform traditional models for over 40% and provide better controls of the financial credit risks.

The success of these internet companies suggests that traditional methods are not sufficient to adapt the big data in the financial credit risk area. To leverage the potential of big data, new methods and tools need to be explored.

4 Scope

Possible research topics include:

- How to use big data technology to process and analyze the valuable information of structured data and unstructured data effectively. A person's social standing, online reputation and/or professional connections are factors that should be considered when extending credits, especially to someone who do not have enough credit history and are unable to apply loans from traditional banks. Many internet credit companies are started to collect social media data to assess consumer's credit risk, such as Lenddo,

Neo Finance and Affirm. There are many active research areas in this domain, including:

- Human identity validation.
- Missing data prediction. Big data does not necessarily lead to more information due to the inconsistent data and/or missing values.
- Social credit score prediction. These social credit score can help to answer questions like “Who has better credit? One with 50 FB friends or one with 500 FB friends?”
- Credit prediction. How to combine the structured data and unstructured data to assess the credits
- Future delinquencies prediction

- Data visualization. Machine learning is powerful but it is challenging to make the results interpretable for the human.
  - Rule extraction. How to extract the rules from the credit score predictive models?
  - Intelligent, interactive data visualization.

5 Expected Outcome and Deliverables

- New algorithm or system
- Demo system for an intelligent, interactive data interface
- One or more patents

6 Expected Project Duration

1 year