**A Real-Time Credit Risk Assessment Using Machine Learning and Apache Spark Streaming**

Chunhua Wu

[chunhua.wu@ryerson.ca](mailto:chunhua.wu@ryerson.ca)

**Abstract**

This project aimed at predicting credit risk in advance.

Credit risk is the possibility of a loss resulting from a borrower failing to make required payments under agreed terms. For lenders, it means not receiving the owed principal and interest, so to result in the cash flow interruption, and both collection costs and borrowing costs increase.

To address this issue, financial institutions and lenders try to assess the risk probability and to predict potential default, by using credit data such as payment and bill statement history, income level, so on and on. Traditional statistical techniques, such as manually analyzing these data, is time consuming, error prone, and resource intensive.

With the development of blockchain and the internet of things (IoT), financial institutions could process real-time assessment within shorter time and with fewer costs. Nowadays, big data technology, such as machine learning and data streaming tools, offers the platform for advanced automated process.

This project will build a real-time and automatic credit risk assess system by using big data techniques. A public dataset, Default of Credit Card Clients Dataset from Kaggle (<https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>), provides 30000 observations and will be used. Data clean and machine learning algorithms, such as supervised-classification and regression, will be introduced. Python, Apache Spark, and data-related libs or software will be applied. This project is also updated in the GitHub repository (<https://sherylgit.github.io/>).