**CKME 136 Data Analytics: Capstone Course**

**Ryerson University**

**Fall 2018**

**Using predictive analytics to detect defaults in bank loan applications**

Constantin Romanescu

https://github.com/cromanes/Ryerson-Capstone

Detecting possible defaults during a loan application processing, is a sought-after goal of any institution in the banking system.

The goal of this project is to showcase a full data analytics process aimed at estimating the default risk of bank loan applications, and, therefore, classifying them as default or fully paid loans. The dataset will be provided by LendingClub1, a fintech peer-to-peer lending company headquartered in San Francisco, US. The set contains anonymized data from loan applications along with the outcome at each loan.

The analytics pipeline will follow closely the pathway outlined in the Cross Industry Standard Process for Data Mining (CRISP-DM) model and it will be performed using R/ Python/ Spark programming languages. Following the data ingestion and the exploratory data analysis, we will use Feature Engineering to create new features that will enhance the predictive power of the model.

For the classification task we will be using both standalone techniques (logistic regression, trees, random forests, gradient techniques, SVM, neural networks, deep learning, etc.), and ensembles of classification techniques.

As the last step, the models featuring high performance metrics will be deployed as web applications, such as Shinny, and/ or using cloud services, such as Microsoft Azure. This will enable the prescreening of applications using simple tools such as a web browser or spreadsheet software (Excel).

The approach outlined above can be easily extended to similar problems that confront a number of industries: credit card defaults, customer retention.