CKME 136 Data Analytics: Capstone Course
Ryerson University
Fall 2018

Richard Harman

Abstract

Major financial institutions do not typically service well the needs of lower income families. Lending to low-income or first-time borrowers is challenging when there is little or no credit history. This creates a segment of 'under-banked' individuals struggling to access formal, safe and fairly priced credit. For established banks, this represents a profit growth opportunity, while for new entrants offering financial services, an attractive underserved market. However, to be successful in this segment, a financial institution must find non-traditional and creative ways to identify and mitigate default risk.

The work that follows will explore a Kaggle competition dataset[1] provided by Home Credit B.V., a lender operating in this under-banked segment. The dataset contains details of loan applicants at the time of application plus information (where available) on their historical repayment performance. I will look for relationships between the characteristics of a loan applicant and their risk of default. For example, are people employed in a certain industry more likely to default, does length of employment, family status or size and location of residence impact default risk? Lastly, I will attempt to predict the likelihood of default for a new loan applicant.

Exploratory analysis will use Python data manipulation and visualisation packages such as numpy, pandas and matplotlib. Predictive analytics will use Python machine learning tools such as scikit-learn. The prediction task is a binary classification problem – likely to default: yes or no? Applicable models will be investigated such as logistic regression, decision trees and support vector machines.

[1] https://www.kaggle.com/c/home-credit-default-risk/data