Springboard Data Science Career Track

**Capstone 1 Project Proposal**

*Bhaskar Das*

**Title: Home credit default risk prediction**

1. Business Problem:

Mainstream banks and financial institutions check traditional credit score models, which include demographic characteristics, historical payment data, credit bureau data and application data, to determine repayment success. However, many unbanked individuals do not have sufficient credit scores due to their past mistakes of unavoidable circumstances. Therefore, they have to deal with unconventional means such as loan sharks when borrowing money. Moreover, most of these individuals are hard-working and should get a chance to borrow money safely. It is important to identify these individuals from the financial records to provide a positive and safe borrowing experience.

2. Business Objective

The primary object of this project is to build a model from the financial data to predict the likelihood that an applicant will experience diﬃculty in repaying their loan. The output of the proposed model is the probability which determines an applicant in terms of having at least one late payment when repaying their loan.

3. Methodology

The dataset is provided by Home Credit Group’s data scientists which contains personal and ﬁnancial information belonging to 356,255 individuals who had previously been recipients of loans. The data is divided into training group, which contains 307,511 records, and test group, which contains 48,744 records.

An applicants’ personal background information is presented in the main data table which contains 120 features that comprise. An applicants’ previous loan and credit card balance payment histories are presented in the other six data tables contain.

Data Source: <https://www.kaggle.com/c/home-credit-default-risk/data>

4. Deliverables

The model will take inputs from the seven data tables and will be trained on the 307,511 records and computes delinquency probabilities. The delinquency probabilities will be in the range [0.0,1.0], where zero represents no delinquent repayment and 1.0 represents at least one delinquent repayment. The model will predict the highest and lowest probability of delinquency for the borrowers who had at least one late payment and zero late payments.