Final Project Proposal - Group 8

Team members

Erica Augustyniak
Karim Al Zeer Alhusaini
Kindeep Dhatt
Mathai Paul

Dataset

Home Credit Default Risk, source: Kaggle - Link

Division of work

Erica Augustyniak - Exploratory Data Analysis, build an LDA model
Karim Al Zeer Alhusaini - Data management and cleanup, build classification modes: Logistic regression, and KNN
Kindeep Dhatt - Write up and presentation, build a QDA model
Mathai Paul - Code implementation and write-up, build a Decision-Tree model

Problem Statement

As a lender, Home Credit gathers a considerable amount of demographic and financial data on its customers in addition to other credit bureau attributes. The problem in question is how to better predict the probability of default on a loan given the data available. In this project, we will build a variety of classification models and assess the predictive capability of each. The goal is to implement this model at the Point of Sale, so customers can be scored online, so the right loan can be offered.

Plan of Approach

- Data cleanup:
    - Identify and handle missing values
    - Rename columns so they are easily identified
    - Convert fields to the proper data type, e.g. TARGET field should be categorical

- Perform EDA and derive insights from the data: understand the various correlations between variables, and plot the distributions of relevant variables
- Build different classification models from Chapter 4:
    - Logistic Regression Model
    - LDA
    - QDA

    - KNN

- Assess the strengths and weaknesses of each model using the provided “test” dataset
    - Predictive capability
    - Error rates
    - Build an ROC graph for each model

- Build a decision tree model

- Explore Unsupervised Learning techniques such as Clustering, which may fit well with such dataset

Tools

We’re using R for the EDA and build all the models. We may use other tools, such as Python, for any unsupervised learning technique we may implement.