CMPE 255 - Data Mining
Project Proposal

Student Name:

- Romil Viral Shah (SJSU-ID:-015250599)
- Pallav Kirankumar Parikh (SJSU-ID:-015275481)
- Ishan Hitesh Malkan (SJSU-ID:-015281968)

Project Title: Washington State Home Loan Analysis and Prediction

Project Description: Performing in depth analysis of the Washington home loans dataset available on Kaggle (link provided below). This dataset consists of a mixture of categorical, numerical and text data which requires pre-processing such as missing data interpolation, encoding data types etc. Later on we will also predict the approval decisions using the other columns of the dataset using various classification models such as tree based, SVM etc.

Proposed methodology: We will use multiple data pre processing techniques such as normalization, standardization, encoding, missing value interpolation, data transformation, data cleaning etc. and modelling techniques of classification such as logistic regression, support vector machines, neural networks etc. We will also use data visualization tools in python, to plot the data to get more insight.

Resources / Programming tools: Mainly Python programming language will be used for this project, (important libraries: pandas, numpy, sklearn, matplotlib, seaborn etc.). Anaconda will be used to manage python environments and packages.

Dataset: HDMA Washington State Home Loans, 2016
Number of columns: 47
Number of rows: 466,566

Link: https://www.kaggle.com/miker400/washington-state-home-mortgage-hdma2016