Statistical Data Mining Project Proposal - Murali(mt788), Ranjini(rr735) & Subham(sb2564)

Objective:

- ***First Goal***: Our first goal is to build a model to predict the loan approval/denial process.
- ***Second Goal***: The predictions of approval/denial will then be used along with the data to build a model to predict the denial reason.
- ***Third Goal***: Explore bias and Fairness in our model, if any, in the models built.

Dataset:

- ***Act***: The Home Mortgage Disclosure Act (HMDA) requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages.
- ***Data Statistics***: We have 21 variables in total and over 14 million data points for the year 2017.
- ***Source***: Mortgage data from the Consumer Financial Protection Bureau website.

Data Analysis Plan:

- ***Exploratory Data Analysis***: To discern patterns underlying in the data, we will perform EDA by generating histograms, correlation plots, box plots etc and utilize that for feature engineering. Also, We plan to drop the missing values due to the abundance of data.
- ***Model Building***: We will build classification models like Logistic Regression, XGBoost, Random Forest, cross validate and tune the hyper-parameters to get to the best model.
- ***Unsupervised Learning***: We plan on performing clustering and dimensionality reduction algorithms to further understand the data better.

Conclusion:

- ***Pipeline Building***: The plan aids in building a pipeline to predict approval/denial process along with the potential denial reason. [*Outcome of Goal 1 and Goal 2*]
- ***Understanding Decision Making***: The methods we implement in our analysis will help us understand this decision-making process using various models built. [*Outcome of Goal 3*]