DS GA 1017 Home Credit Default Risk Proposal

Joby George (jg6615), Bruno Stefoni (bas9893)

Competition & Data Overview

Problem Statement: Granting a home mortgage is a predictive problem for home lenders. The lender has to ensure that the loan seeker is qualified for the loan they are asking for, otherwise the lender could lose the entire loan amount, known as **default risk**.

Competition Background: Home Credit Group, incentivized to better address this problem, posed a challenge on Kaggle four years ago with a grand prize of $70 K. 7,176 teams participated in this challenge, with a lively discussion and numerous leaderboard competitors, posting their solution to the challenge.

Data background: There are 307,512 sample data points with 120 predictors and a target variable, whether or not a consumer missed a payment. The feature space can broadly be described as the following:

1. Demographic attributes, including Sex, Age, Marital Status
2. Credit Risk attributes: including, Previous Home Loan data, Credit Bureau Data, Previous Loan Applications, Payment history, and Credit Card Data
3. Home Attributes: data about the house, average statistics about houses in that region, etc

Interpretability Criteria: Our Nutritional Label strives to understand algorithmic lending fairness on the protected classes of age, gender and marital status, with the goal of understanding whether bias exists for any of these groups.

Code Overview

In choosing a specific submission to analyze, we had four priorities: performance of the solution on the problem, documentation level, modeling complexity and execution complexity. We preferred a top performing submission, with a high level of documentation and moderate modeling and execution complexity. If the model used was outside our domain of expertise, and the execution complexity was high with minimal documentation, we would not be able to meaningfully create a nutritional label. The submission selected to analyze the fairness was a top-20 solution by *NoxMoon*, which we were able to successfully run.

Why did we chose this project

The Home Credit Default Risk competition has many of the interesting properties of an ADS which can benefit from nutritional labeling. The models behind it use sensitive features such as Gender, Age and marital status. Also, the database is likely to contain all three forms of bias: pre-existing, technical and emergent. In addition, there exists a potential irreparable impact in people's lives caused by the decisions made from this ADS. Finally, career wise we think it is a very relevant and interesting case as it is related to our future professional endeavors.