DEBT AND MORTGAGE DATA ANALYSIS: Individual Report

Matteo Bucalossi

1. Introduction.

The objective of this project is to analyze the socio-economic dynamics of mortgage and debt in the country and build a classification model to understand the demographic factors that affect debt across American society. The ultimate goal is to determine if any disparity may exist within the population and which demographic groups would be more or less affected by higher or lower debt. The analysis of mortgage and debt data will lead to conclusions about how wealth inequality is distributed throughout the population. The project uses data collected by the U.S. Census during the period 2012-2016 as part of the ACS 5-Year Documentation, provided on Kaggle.com by the Golden Oak Research Group.

The team divided the work on the various components of the projects: preprocessing code, modeling code, visualization code, GUI code, presentation and report writing. We eventually all contributed equally to the final outcome throughout different tasks and everyone added their own code to some part of the final script.

2. Description of your individual work.

Personally, I mainly contributed to the preprocessing of the data, dealing with null values and outliers detection among other more basic tasks in this phase. Then, I wrote the group report, later edited and expanded by my colleagues at certain sections, as well as the presentation in ppt. Finally, I also integrated some code for ensembling and random forest models for prediction and visualization of features selected.

3. Describe the portion of the work that you did on the project in detail.

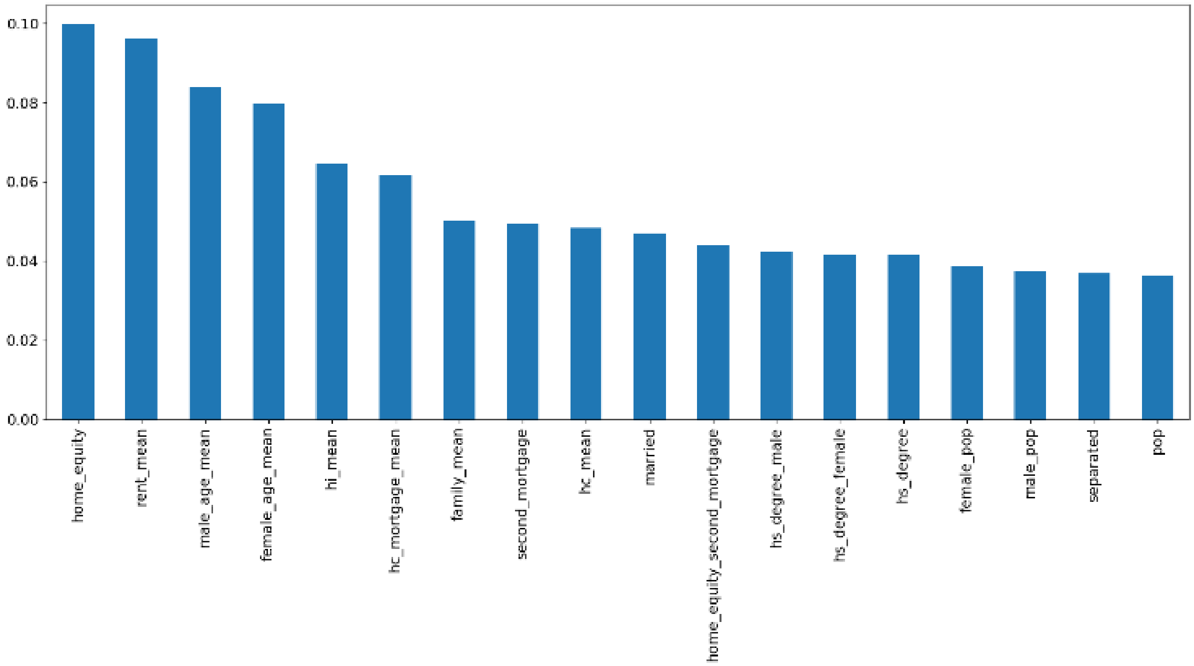
First of all, I proceeded with coding the preprocessing and dropped the columns that were not relevant for our analysis, including all the summary statistics features that were not mean. I then used a function to detect and remove outliers from every numeric column, and then checked for nulls and duplicates. I also built classes to call upon preprocessing and visualization as objects, but would not be later used team-wise. The code would then be integrated in the general script for GUI by Anwesha.

Second, I wrote the report in most of its entirety following the indicated steps as well as the entire presentation created with Google slides. While working on results, I had to integrate additional visualizations and results from the ensembling and random forests, which I then used principally for presentation purposes. The code can be found in the folder Code.

4. Results.

Given my work on the writing of the entire report, my contribution can be found on the Group report.

On the coding side, I tried different combinations of models' ensembling to find out the best accuracy score possible, eventually deciding for a SVM + Logit + DT ensemble. Then, I added classification random forest models for both target variables and added code for visualizing how often forests would use certain features in their reiterations. One example of the four graphs (2 for each target variable, and then each of those for regression and classification problems) can be found below. These bar graphs show indeed the use of each predictor variable by the random forest algorithm, and plot them against their importance in the splits operated by the model in the decision trees built.



5. Summary and conclusions.

Working on most of the report and presentation truly exposed me to the limitations and strengths of our models, and understanding them better by comparing their various accuracies and scores as well as investigating why each one was performing better or worse. This prompted me to reason about the data in the first phase and in the final phase where the results from the models were available, thus assessing their best uses and outcomes and arguing about the conclusions of our analysis.

On the other hand, coding the ensembling allowed me to explore which model would have increased the accuracy, while I was also able to assess which were the most important features used by the Random Forest algorithms. The preprocessing phase exposed me to the understanding of the dataset and how to use features and prepare them for modeling code performed by my colleagues. I was interested in also pursuing analysis with geographic variables and geocode the dataset per certain variables on a map, but time constraints limited my efforts on this less relevant direction.

6. Calculate the percentage of the code that you found or copied from the internet.

400-150/400+50 = 55%

7. References.

All modeling code was adapted from lectures and examples by Professor Amir Jafari at George Washington University. Background knowledge on the topic from Investopedia and domain expertise. The dataset was retrieved on Kaggle.com in .csv format.