DEBT AND MORTGAGE DATA ANALYSIS

Anwesha Tomar, Spencer Staub, Matteo Bucalossi

1. Introduction:

Mortgage balances and debt have been climbing in recent years, according to the Federal Reserve Bank of New York. Housing debt now totals $8.94 trillion, close to the $9.99 trillion peak of the third quarter of 2008. Mortgage debt is also the largest component of total household debt, making up 71% of total household debt.[1] Given the debt crisis of a decade ago, understanding the mortgage and debt trends across the U.S. population is extremely relevant and can help address areas and communities that may be more vulnerable to the housing market.

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.[2]

2. Description of the data set:

The dataset includes 39,030 records and 80 features. Some of these features are demographic variables, while some are summary statistics. 74 are numeric variables (either ‘integer’ or ‘float’ type), while the remaining six are string variables (coded as ‘object’) for geographic features such as state, city, and type of place.

Information about variables are presented through summary statistics which present mean, median, standard deviation, sample weight and samples of each indicator, and each of these statistics are then recorded as separate features in the dataset. This is the case for primary indicators as the following:

[1] https://www.investopedia.com/personal-finance/american-debt-mortgage-debt/

[2] https://www.kaggle.com/goldenoakresearch/us-acs-mortgage-equity-loans-rent-statistics

- Monthly Mortgage and Owner Costs, which constitutes the sum of payments for mortgages, deeds of trust, contracts to purchase and other debts on property, including taxes, insurance and utilities.
- Monthly Owner Costs, which constitutes the sum of all costs for the owner on property excluding mortgages and other loans.
- Gross Rent, as per the contracted rent plus estimated average monthly cost of utilities when paid by renter.
- Household Income, which is the sum of the income of the householder and everyone above 15 years old residing in it.
- Family Income, which differs from household income as it sums the incomes of all family members of the householders regardless of their residency, thus resulting often in larger amounts.
- Also variables of age of female samples and age of male samples are stored in summary statistics features.

Additional variables that are interesting for the purposes of this project are also recorded as proportion or mean values for the sample investigated. For instance, the dataset contains features of the percentage of houses with a second mortgage, with a home equity loan, with any type of debt, or features of the percentage of people with at least a high school degree and of the people married, divorced or separated.

3. Algorithms and Models Used:

The purpose of the project is to identify subgroups of the sample interviewed which are more or less likely to have mortgages, debts or a certain income. For this, we have developed a series of models to answer this classification problem. Then, we applied these models to solve this regression problem.

The following analysis is based on four different models: Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression. These models are then ensembled to increase their accuracy in classification with hard voting method, while a random forest regression is applied to increase prediction accuracy.

*Math equations and classic algorithm construction.

4. Experimental setup:

To perform the analysis on the dataset, a Python script was written in PyCharm and the data was then read as a Pandas dataframe.

Given the great amount of features present, the first step of pre-processing was to drop several features that were not needed, but they were indeed redundant given their recording as summary statistics. Thus, the dataset will only include the features containing the mean values for the mentioned indicators, excluding then features of median, standard deviation, samples, weight, and other statistics that are not the mean. Also, the features of 'BLOCKID' and 'SUMLEVEL' were also removed because empty or irrelevant.

Subsequently, we wrote a function to remove outliers from every feature, where outliers are defined as point values 1.5 times above or below the upper or lower bounds of the distributions of the data. The data type of features such as the unique identifier, or other geographic variables such as zip code, latitude and longitude, were temporarily converted to object so that the function could remove outliers only from numeric features that actually had outliers. The resulting dataset that will be used for the models presents 38 features and 20,190 observations.

Analysis Procedure

To accomplish our analytical purpose, we used all four models to classify and predict the following target variables: rent and debt, namely the average of rent costs and the percentage of houses with some sorts of debt. Consequently, we were able to obtain a compelling understanding of the data by looking at how predictors affect the two different target variables through four models and their ensembling. To do so, we wrote consistent code that could apply the models in both cases for the different target variables.

The variables considered are the following: population, male population, female population, rent, household income, family income, owner costs and mortgage, owner costs, second mortgage, home equity loan, second mortgage, debt, and home equity loan, education, male education, female education, male age, female age, married, separated. To be able to use the target variable for classification models, each considered target variable was categorized with median and quartile splits: the values were divided in two by median to run KNN, SVM and logistic regression, while they were divided into four equal quartiles for the Decision Tree.

First, we built the KNN model. We identified target and feature variables, also creating two different targets, one categorical and one numeric respectively for classification and regression. Then, we split the data twice for the two different problems with a 25% test split. After scaling the data with StandardScaler, we fit a KNN classifier and a KNN regressor to the training data, both with a K value equal to 9. We then used test data to predict values for the target variable selected and evaluated the model by confusion matrix, classification report, and accuracy scores.

Second, we built the SVM model. We followed the same practice as for KNN to split and scale the data, and then we fit the training data with SVM classifier and a linear kernel for the classification problem, and with SVM regressor and a radial basis function kernel. We eventually used the test data to predict values with the model, and evaluated it with confusion matrix and accuracy scores for both problems.

Third, we built a logistic regression for the binomial variable of the target. We split the data again with a 25% test split and fit the model to the training data, and use it to predict on the test values. Again, we evaluated this model with logistic and ROC AUC accuracy scores, and confusion matrix.

Finally, we built a Decision Tree for both regression and classification, using the numeric target in one case and its categorical transformation in the other. After splitting and scaling the data, we used a Decision Tree regressor and a Decision Tree classifier based on entropy to respectively train the training data for regression and classification. Evaluation was then based on accuracy scores and confusion matrix.

After having built and evaluated these models, we improved the accuracy of our analysis by ensemble. First, we used a hard voting method to ensemble linear regression, decision tree, and SVM classifiers, which improved the accuracy scores of the models. Then, we also improved the trees performance by using Random Forest regressor and classifier based on 100 estimators, which also increased the accuracy scores of the trees.

5. Results:

Analysis on debt:

We started our analysis by investigating the predictors for the percentage of households with some sort of debts per area interviewed. Thus, we used the numeric variable of debt to predict the percentage of such houses given our models, and a categorization of debt by median and quartile splits to see if certain demographics would have lower or higher debt.

KNN model provided an accuracy score of 57% for its regressor and of 78% for its classifier. The confusion matrix for the latter is displayed below, showing that the KNN classification model predicted wrongly around 550 values for each low and high category.

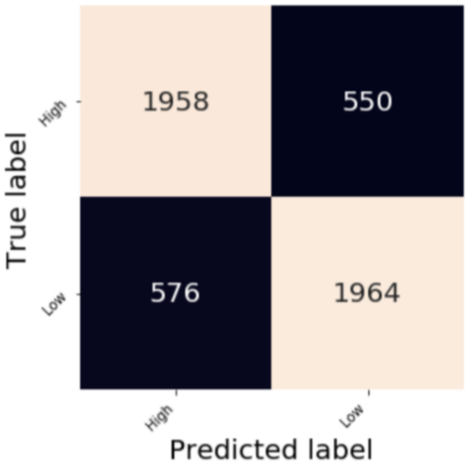

Figure 2. KNN Confusion Matrix: Debt

The SVM model provided a good accuracy of 80% for its classifier and one of 61% for its regressor. Below the confusion matrix of the SVM is indeed displayed, as well as the plot of the coefficients of the model per each predictor variable.

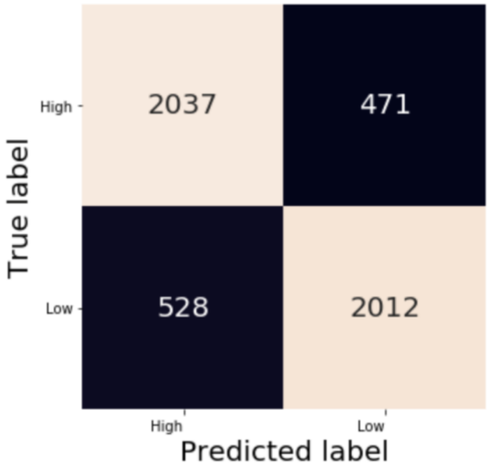

Figure 1. SVM Confusion Matrix: Debt

The Logistic Regression resulted in an accuracy of 75% and an Area under the ROC Curve of 84%. Below the confusion matrix for the classifier of the Logistic model.

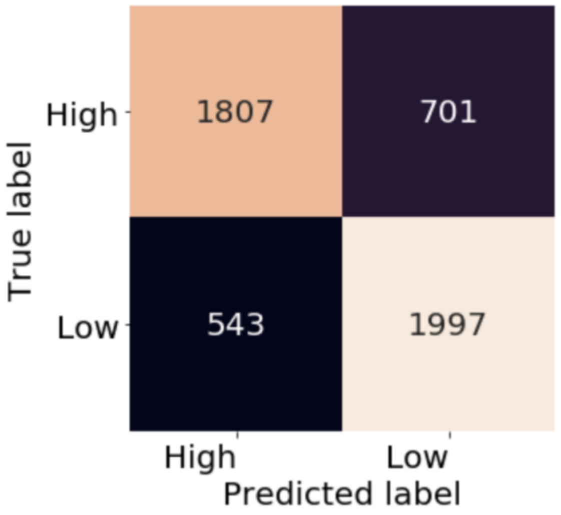

Figure 3. Logit Confusion Matrix: Debt

Finally, the decision trees built are displayed below, respectively regression and classification trees. These models resulted in accuracy scores of 54% and 52% respectively. We can see that the 3 splits operated to predict an average proportion of debt are based on the main predictors of rent, home equity loan and average age, in both problems.

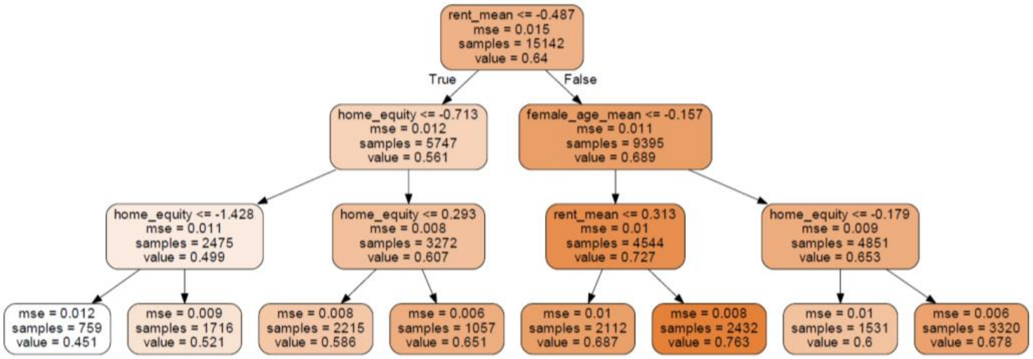

Figure 4. Regression Tree: Debt

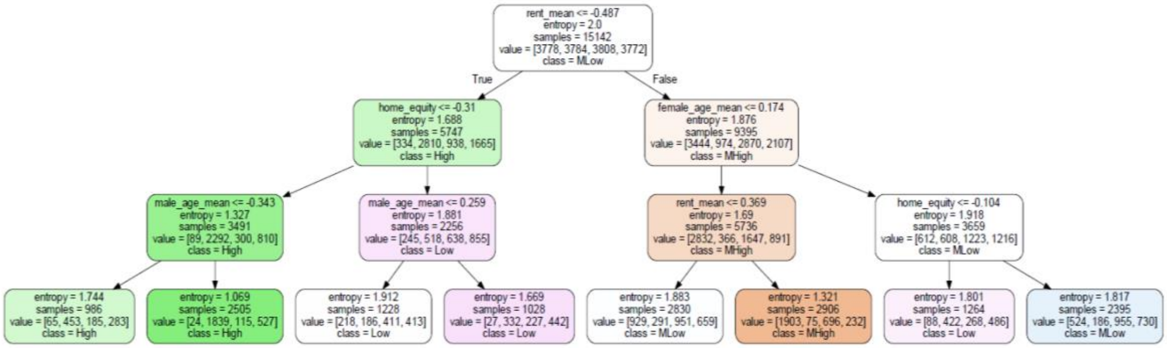

Figure 5. Classification Tree: Debt

Decision trees were then ensembled with Random Forest algorithms, which increased their accuracy only partially as the regressor scored 62% and the classifier 54%. It also resulted that both Random Forest models used mainly the same features to split the data for their estimator trees, as shown by the following bar graphs.

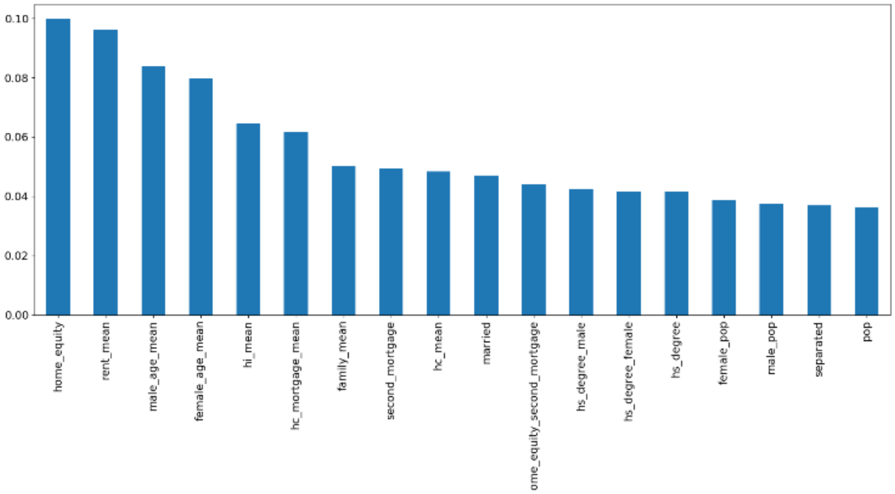

Figure 6. Classifier Forest: Debt

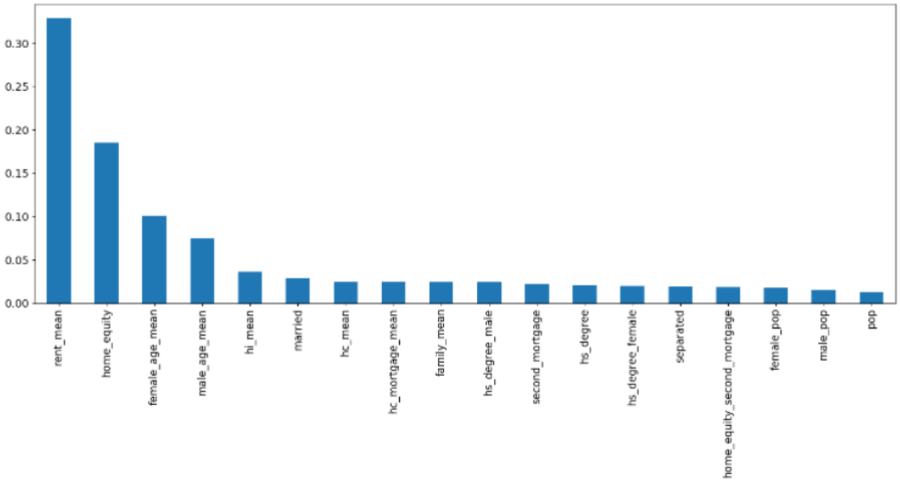

Figure 7. Regressor Forest: Debt

Eventually, an ensembling by hard-voting of Decision Tree, and Logistic Regression models resulted in higher accuracy than other possible combinations, i.e. 76%. Below a table summarizing the accuracies for every model attempted to predict debt proportion.

MODEL USED	ACCURACY
SVM Classifier	80.20
KNN Classifier	77.69
Ensembling	75.65
Logistic Regression	75.35
Random Forest Regression	62.32
SVM Regression	60.71
KNN Regressor	56.71
Decision Tree Regression	54.16
Decision Tree Classifier	51.58

Analysis performed on rent:

We continued our analysis by investigating the predictors for the average gross cost of rent within area interviewed. Thus, we used the numeric variable of rent to predict the mean of the rent with our models, and a categorization of rent costs by median and quartile splits to see if certain demographics would pay more or less rent.

KNN model provided an accuracy score of 53% for its regressor and of 79% for its classifier. The confusion matrix for the latter is displayed below, showing that the KNN classification model predicted wrongly between 400 and 600 values for each low and high category.

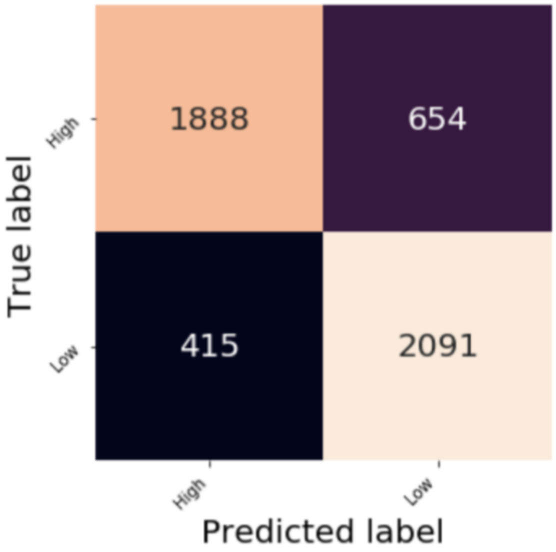

Figure 8. KNN Confusion Matrix: Rent

The SVM model provided an accuracy of 82% for its classifier and one of 47% for its regressor. Below the confusion matrix of the SVM is indeed displayed, as well as the plot of the coefficients of the model per each predictor variable.

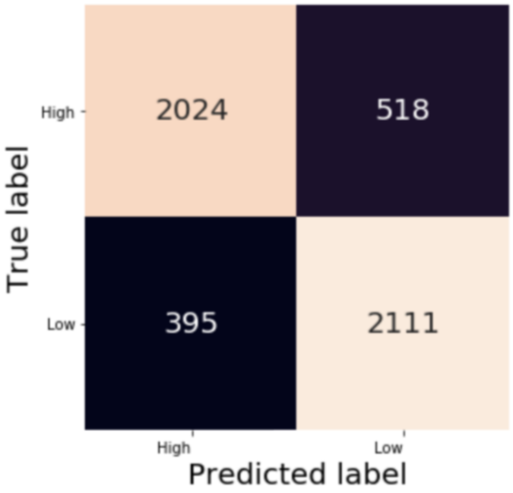

Figure 9. SVM Confusion Matrix: Rent

The Logistic Regression resulted in an accuracy of 79% and an Area under the ROC Curve of 87%. Below the confusion matrix for the classifier of the Logistic model.

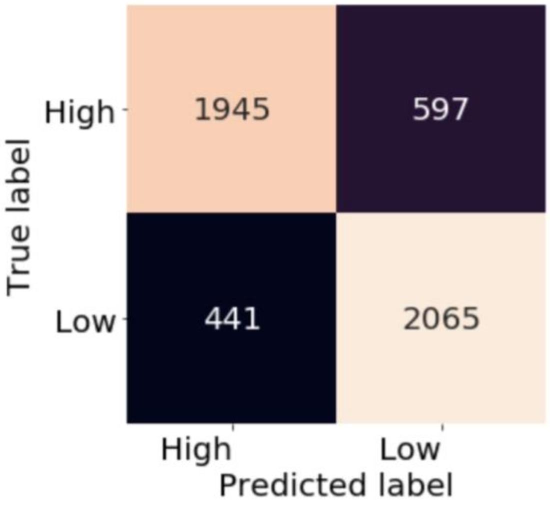

Figure 10. Logit Confusion Matrix: Rent

Finally, the decision trees built are displayed below, respectively regression and classification trees. These models resulted in accuracy scores of 49% and 50% respectively. We can see that the 3 splits operated to predict the average rent cost in the tree are based on the main predictors of household income, debt proportion and homeowner costs, in both problems.

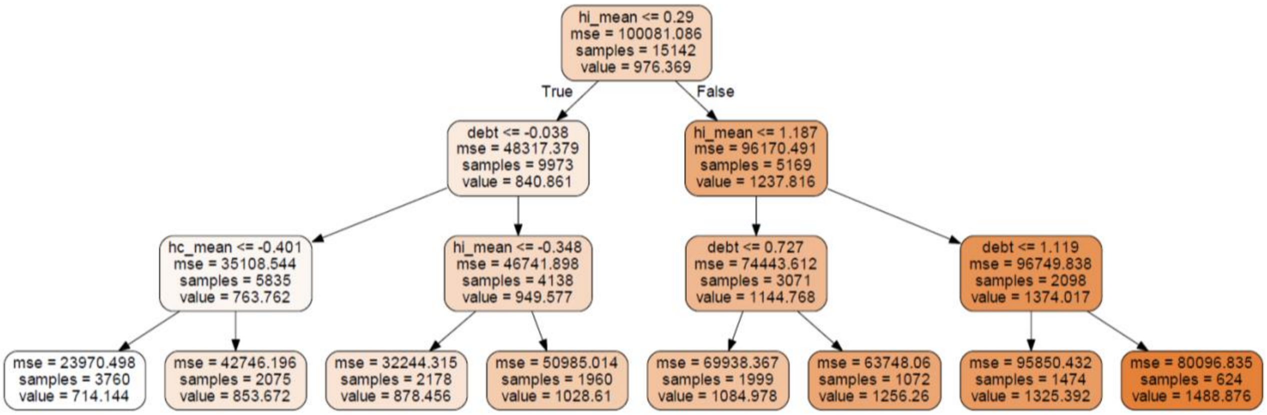

Figure 11. Regression Tree: Rent

For the first root of our decision tree regression our model uses hi_mean or home income mean to divide our dataset just like our classifier– home income being the most inversely correlated to rent. It then goes on to use debt and hc_mean to split the dataset until it reaches our leaves after the 3rd split very similar to the classification model.

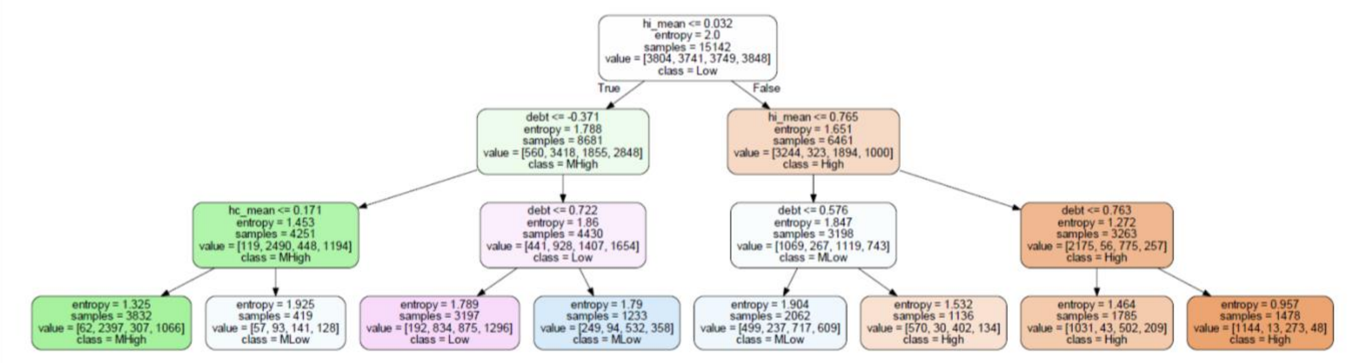

Figure 12. Classification Tree: Rent

For the first root of our decision tree classifier our model uses hi_mean or home income mean to divide our dataset – home income being the most inversely correlated to rent. It then goes on to use debt and hc_mean to split the dataset until it reaches our leaves after the 3rd split. The leaf nodes are divided into one "Low", two "MLow", one "Mhigh" and three "High" rent classes.

Decision trees were then ensembled with Random Forest algorithms, which increased their accuracy only partially as the regressor scored 62% and the classifier 54%. It also resulted that both Random Forest models used mainly the same features to split the data for their estimator trees, as expected and showed by the following bar graphs.

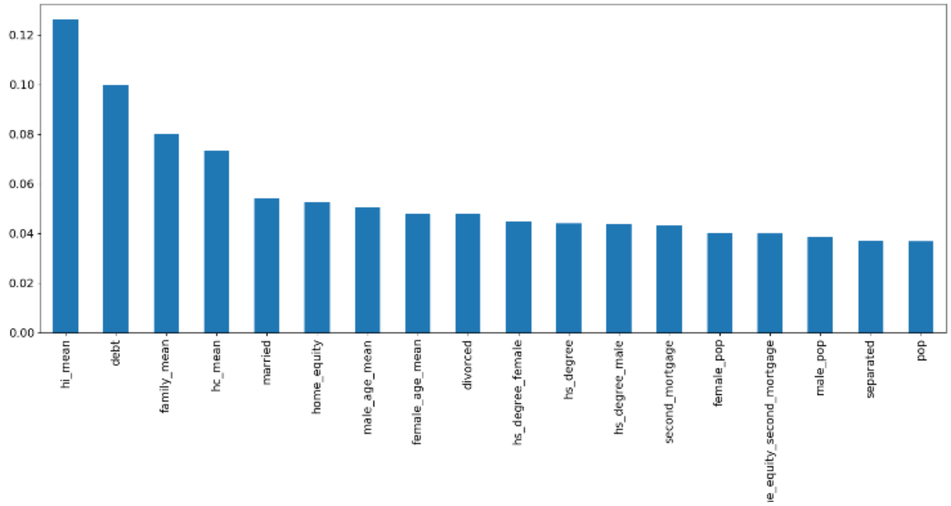

Figure 14. Classifier Forest: Rent

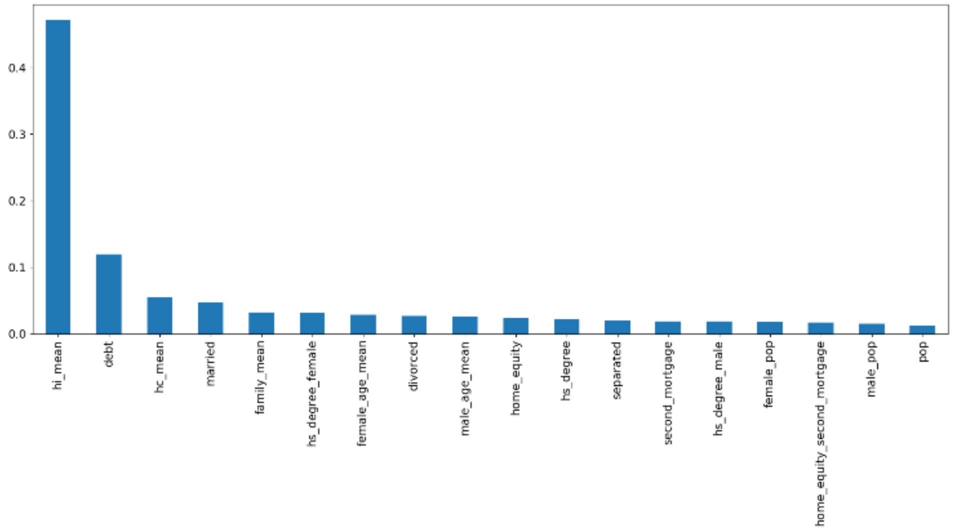

Figure 13. Regressor Forest: Rent

Eventually, an ensembling by hard-voting of Decision Tree, and Logistic Regression models resulted in higher accuracy than other possible combinations of multiple models, i.e. 80%. Below a table summarizing the accuracies for every model attempted to predict average rent cost.

MODEL USED	ACCURACY
SVM Classifier	81.91
Logistic Regression	79.43

Ensembling	78.88
KNN Classifier	78.82
Random Forest Regression	61.58
KNN Regressor	53.44
Decision Tree Classifier	50.09
Decision Tree Regression	49.34
SVM Regression	46.57

6. Summary and conclusions:

In both cases of debt and rent as target variables, SVM Classifier is our most accurate model in this dataset along with logistic regression, ensembling and KNN Classifier all resulting in above 75% accuracy. This makes sense as all of these models use classification and therefore only have a set amount of classes to test against, whether it be two or four classes. Most of our regressions are lower in accuracy since the model is required to predict a number from the min to max of a target, increasing the amount of error exponentially. This lower regression accuracy is then due to the fact that the values of features are in most cases summary or proportion statistics.

Considering the best model among these ones, we can identify the features that contributed the most in the algorithms. Below we can see these features as plotted by their correlation coefficients for debt and rent as targets.

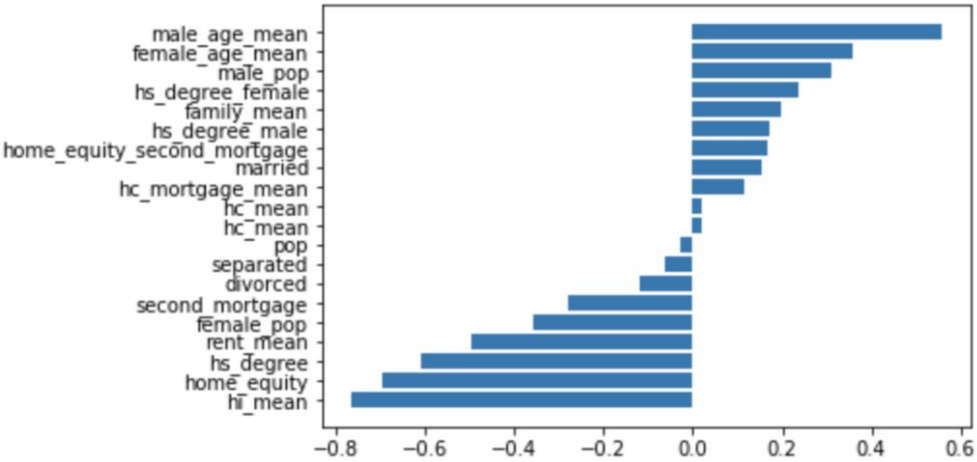

Figure 15. Correlation Coefficients: Debt

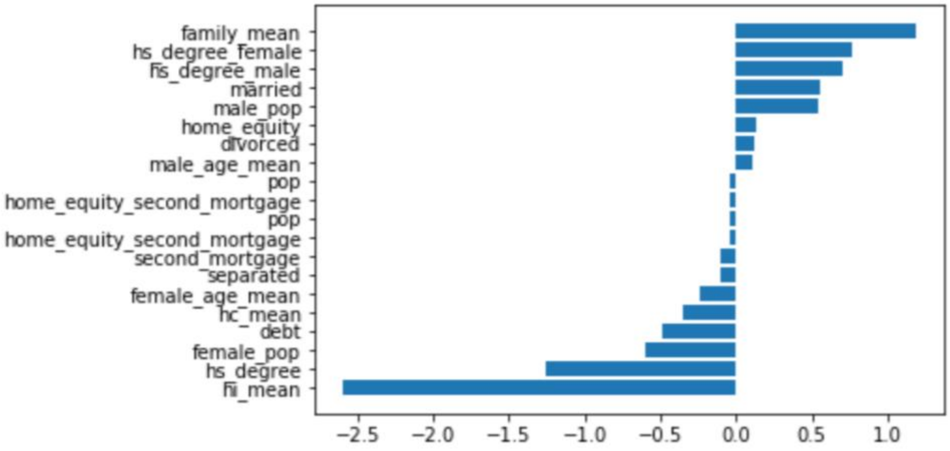

Figure 16. Correlation Coefficients: Rent

The coefficients in these bar graphs show a similar trend from the ones of the decision trees and random forests as well. Indeed, in the case of debt prediction, age seems to be positively correlated to the percentage of debt (as more age allowed for more time to accumulate debt we could assume) , while higher household income, home equity and education seem to be negatively correlated to debt. In the case of average rent, family income and education seem to be the most correlated for predicting higher rent for these subgroups, while household income seems to be inversely significant.

Given the extension and depth of these dataset, multiple models based on different ideas could have been developed to explore socio-economic dynamics in the country as related to housing and mortgage. Indeed, possibilities for future work within the scope of the project would include investigating more than numeric features by considering geographic variables and assess which area, state, city could have more widespread debt or higher rent and mortgage for instance.

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.

8. Computer listings.

Software

- Pycharm

Packages

- Tkinter
- Pandas

- Matplotlib
- Seaborn
- Numpy
- Sk-learn
- Pydotplus