Individual Report

1. **Introduction. An overview of the project and an outline of the shared work**

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. 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.

Coding was split into three different tasks; EDA/Preprocessing, Model Building and GUI. Since EDA/Preprocessing was the easiest task of the three the individual with that responsibility would also write a majority of the report. Since we had three group members Matteo was tasked with EDA/Preprocessing, Spencer tasked with Model Building and Anwesha was tasked with the GUI. All members assisted with the report and presentation.

2. **Description of your individual work. Provide some background information on the development of the algorithm and include necessary equations and figures**

My individual task was to integrate the code generated from the models into a GUI. I used Tkinter for the integration of our models. Also to better understand the dataset and the models we can use, I performed preliminary data exploration in terms of visualization on all the numeric variables. I then performed KNN and SVM regression just to lay down the base for my team to start working on the models.

3. **Describe the portion of the work that you did on the project in detail. It can be figures, codes, explanation, pre-processing, training, etc.**

Our dataset contained 80 features and it was important to perform feature reduction in order to find the features that impacted our target the most. Once we found the most important features I performed visualization for them this also helped me better understand the dataset. Then using these features I performed SVM and KNN. I only used regression initially since most of our dataset was numeric in nature and since we wanted to find future values for debt and rent regression was needed. After the preliminary analysis we then performed some advanced modeling with involved both classification and regression.

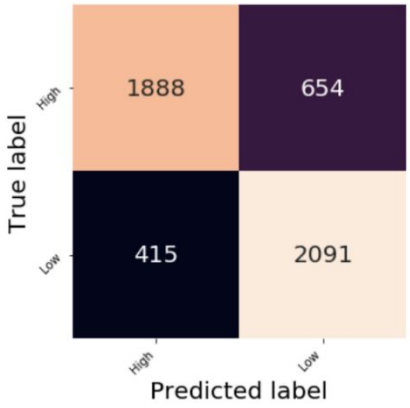
Once my teammates were done with the modeling I then used Tkinter to generate a GUI.

4. **Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.**

Analysis Performed on Rent:

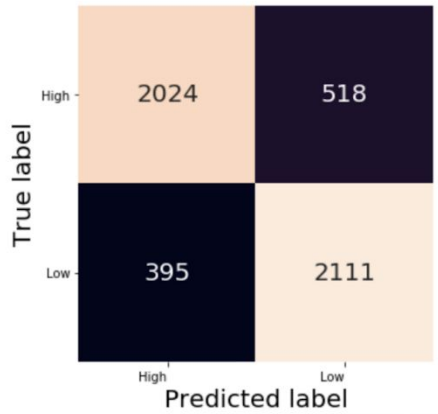
We continued our analysis by investigating the predictors for the average gross cost of rent within area interviewed. Thus, we used the numeric variable of rent to predict the mean of the rent with our models, and a categorization of rent costs by median and quartile splits to see if certain demographics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

KNN Confusion Matrix:



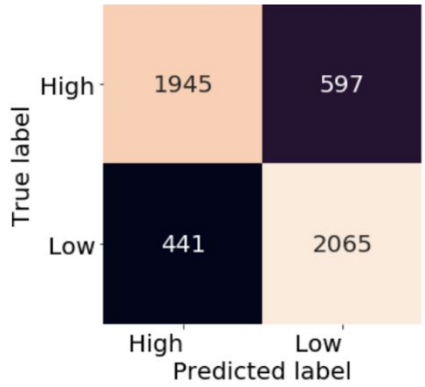
KNN model provided an accuracy score of 53% for its regressor and of 79% for its classifier. The confusion matrix for the latter is displayed below, showing that the KNN classification model predicted wrongly between 400 and 600 values for each low and high category.

SVM Confusion Matrix:



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.

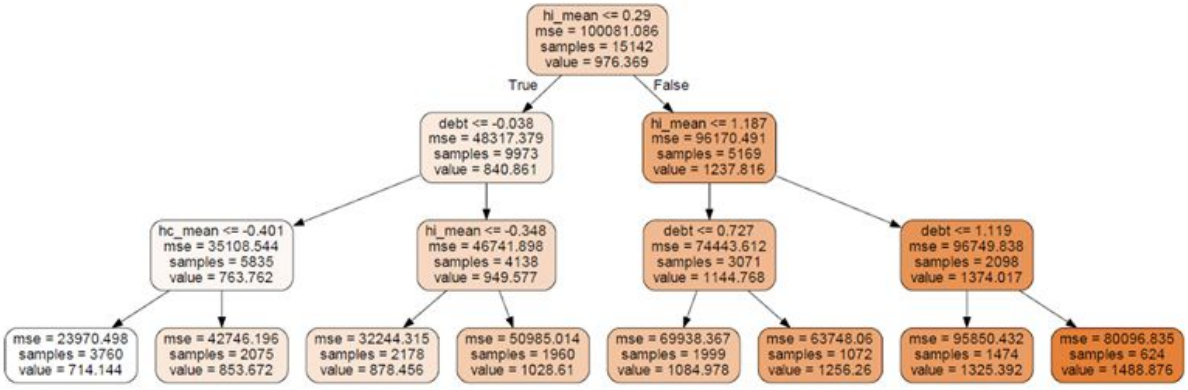
Logistic Regression Confusion Matrix:



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.

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.

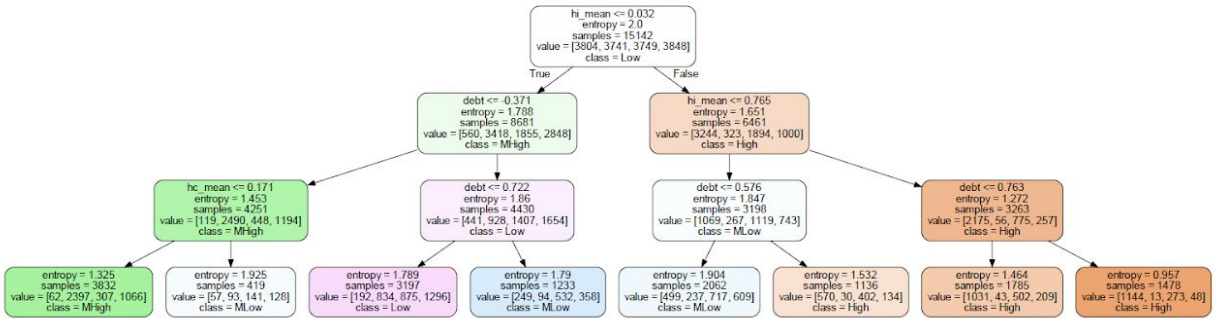
Decision Tree Regression:



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

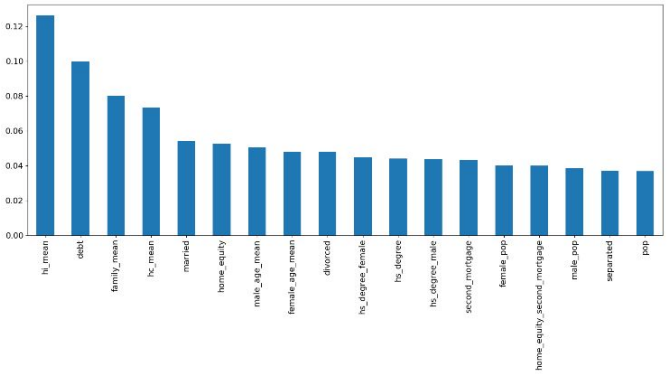
then goes on to use debt and hc_mean to split the dataset until it reaches our leaves after the 3rd split very similar to the classification model.

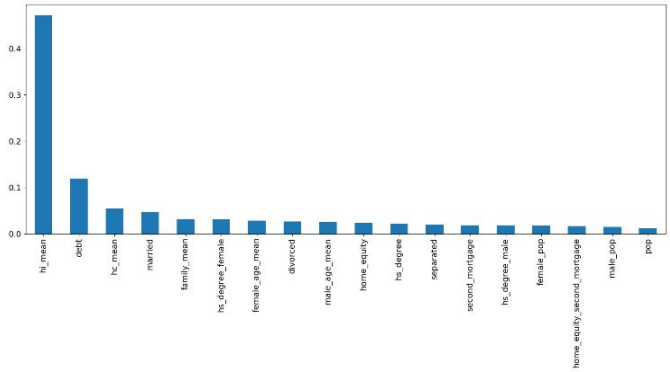
Decision Tree Classification:



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.





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.

Accuracies of Models:

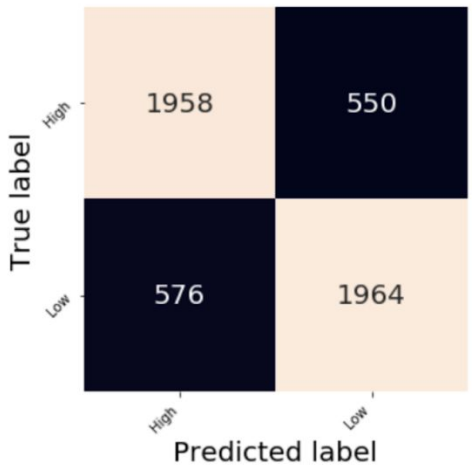
MODEL USED	ACCURACY
SVM Classifier	81.91
Logistic Regression	79.43
Ensembling Accuracy	78.88
KNN Classifier	78.82
Random Forest Regression	61.58
KNN Regressor	53.44
Decision Tree Classifier	50.09
Decision Tree Regression	49.34
SVM Regression	46.57

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

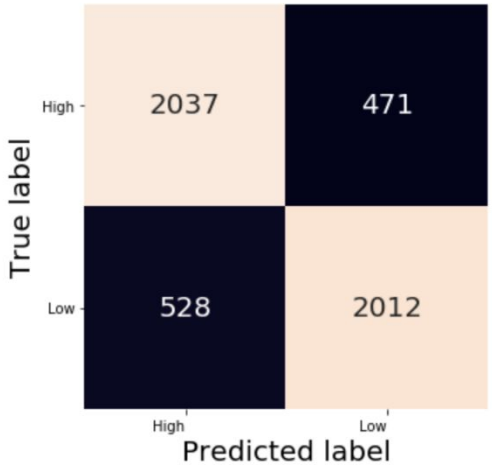
percentage of such houses given our models, and a categorization of debt by median and quartile splits to see if certain demographics would have lower or higher debt.

KNN Confusion Matrix:



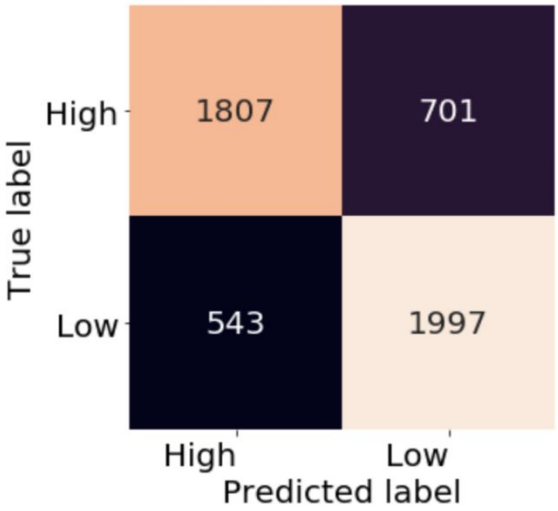
KNN model provided an accuracy score of 57% for its regressor and of 78% for its classifier. The confusion matrix for the latter is displayed below, showing that the KNN classification model predicted wrongly around 550 values for each low and high category.

SVM Confusion Matrix:



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.

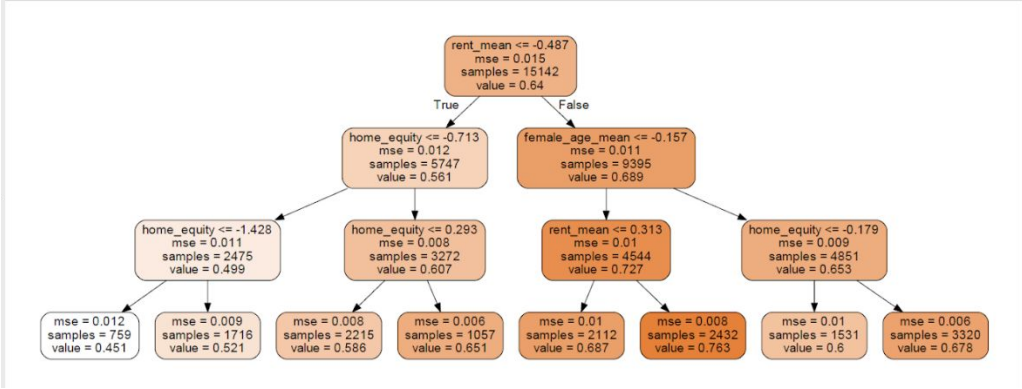
Logistic Regression Confusion Matrix:



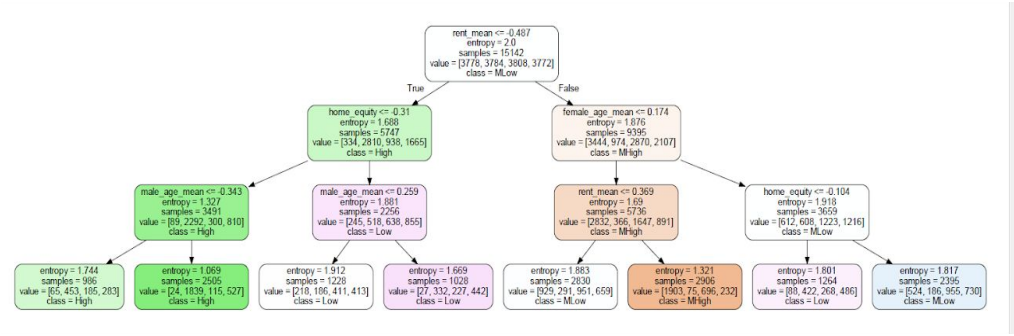
The Logistic Regression resulted in an accuracy of 75% and an Area under the ROC Curve of 84%. Below the confusion matrix for the classifier of the Logistic model.

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.

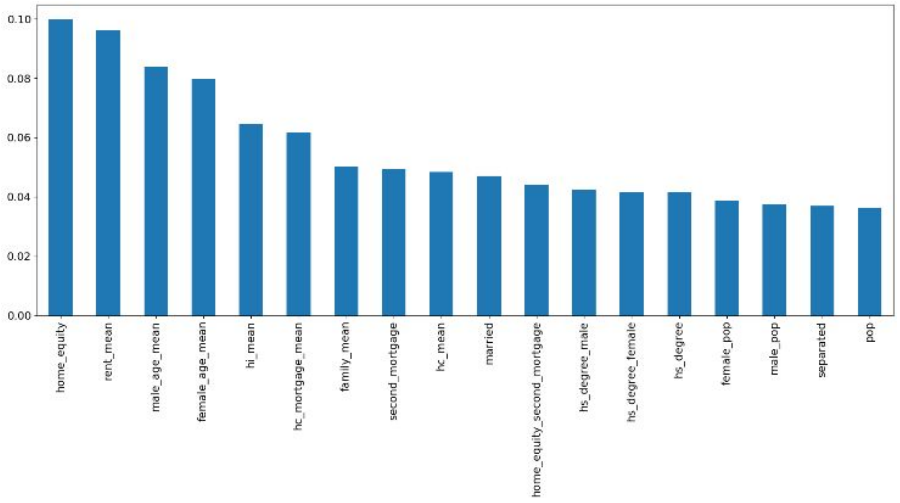
Decision Tree Regression:

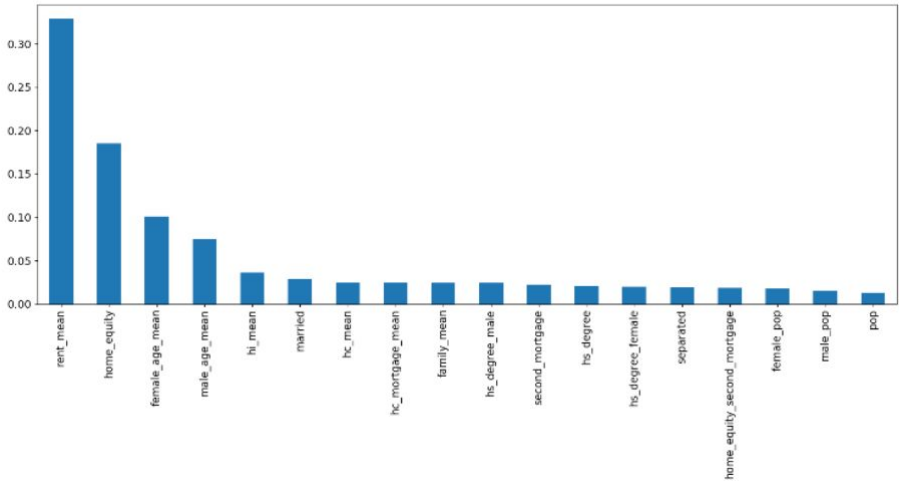


Decision Tree Classification:



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.





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.

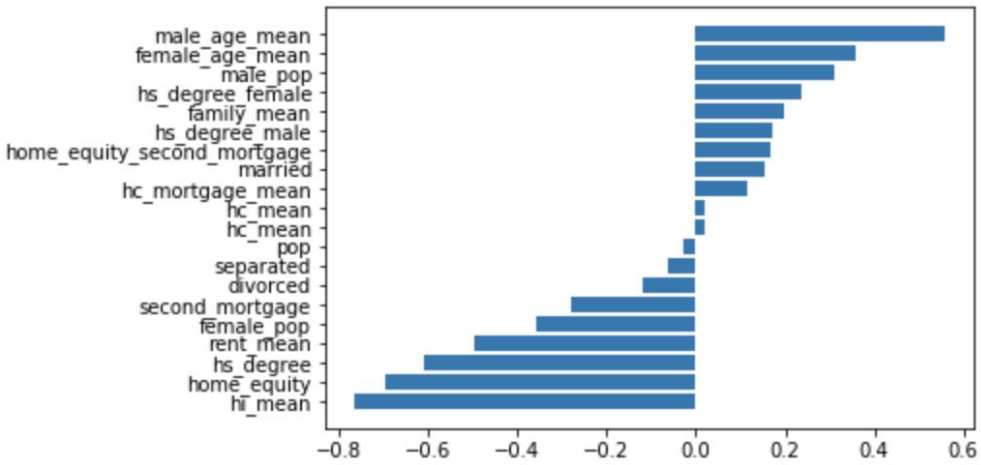
MODEL USED	ACCURACY
SVM Classifier	80.20
KNN Classifier	77.69
Ensembling	75.65
Logistic Regression	75.35
Random Forest Regression	62.32
SVM Regression	60.71
KNN Regressor	56.71
Decision Tree Regression	54.16
Decision Tree Classifier	51.58

5. **Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.**

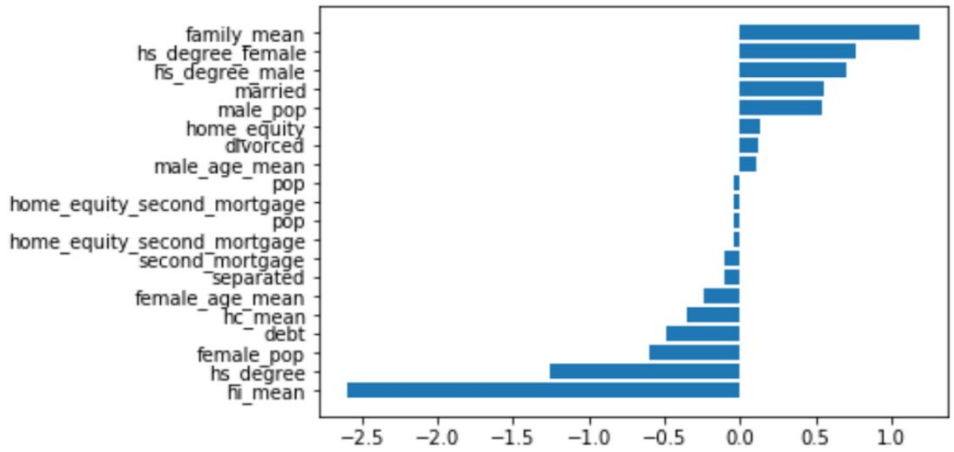
In both cases of debt and rent as target variables, SVM Classifier is our most accurate model in this dataset along with logistic regression, ensembling and KNN Classifier all resulting in above 75% accuracy. This makes sense as all of these models use classification and therefore only have a set amount 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.

Rent Coefficients:



Debt Coefficients:



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.

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

((80 – 30) / (80+40)) x 100 = 41.16%

(Since I was responsible for the GUI I wrote most of that code, it quite a lot and I thought it would be redundant to add it here since it is in the main file.)

*Internet to include code provided by Amir Jafari GitHub

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.