**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.

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

As the individual tasked with Model Building it was my job to decide which models could be run with our data and how our data needed to be manipulated to run good and accurate models. Once that was complete, I could then start testing our models for accuracy and modify them when needed. Initially we were only performing KNN and SVM regressions, but by the end of the project we were performing classification and regression as the pertained to KNN, SVM, Decision Tree, Logistic Regression and Ensembling.

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

Preprocessing

Before starting on the models, I realized that additional preprocessing was required to perform adequate and accurate models. First, we would have to perform feature reduction when we modeled certain target variables. For example debt and debt\_cdf had extremely high positive correlations, which provided great accuracy at the cost of reduced granularity. When performing a decision tree with the debt\_cdf feature it would be the only feature used to split the data. Therefor, this feature and other features with extremely high correlations with the target variable were removed. Additionally, we wanted to perform our analysis with as many models as possible, to do this we created categorical variables of our numerical targets. In this case all of our models could be classifications, as well as regressions. To do this we used lower quartile, median and upper quartile to split the data into four even subsets; “Low”, “MLow”, “MHigh” and “High”.

The KNN, SVM and Decision Tree models were all designed similarly. All models had the same numerical features and the same two target variables (Numerical and Object). All models performed both regression and classification, with the classification requiring encoding of the target variable. We decided on a 75/25 train/test split for all models. Individually for KNN we set K=9, which seemed to provide a greater level of accuracy. For SVM we decided on linear and radial basis function kernals for classification and regression respectively. Finally, for Decision Tree we set the max depth to five to prevent a significant loss in accuracy.

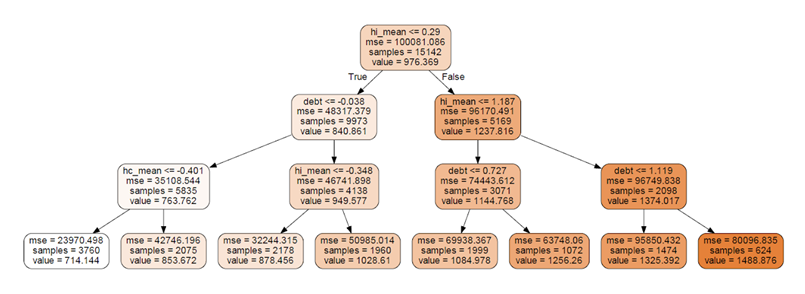
Logistic Regression was the only single model we used that had only classification, given that a binomial variable is required to be the target for the model. Like the other models we used a 75/25 train/test split and the same numerical features. In order to evaluate these models we also included ROC AUC accuracy scores.

In order to increase the accuracies of the models performed we decided to use ensembling and random forest techniques. In order to utilize all four models used in our analysis we performed classification ensembling with a ‘hard voting’ technique. Additionally, we used a Random Forest Regressor to individually increase the accuracy of our decision tree model. For this we used 100 estimators, increasing the accuracy significantly.

1. **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.**

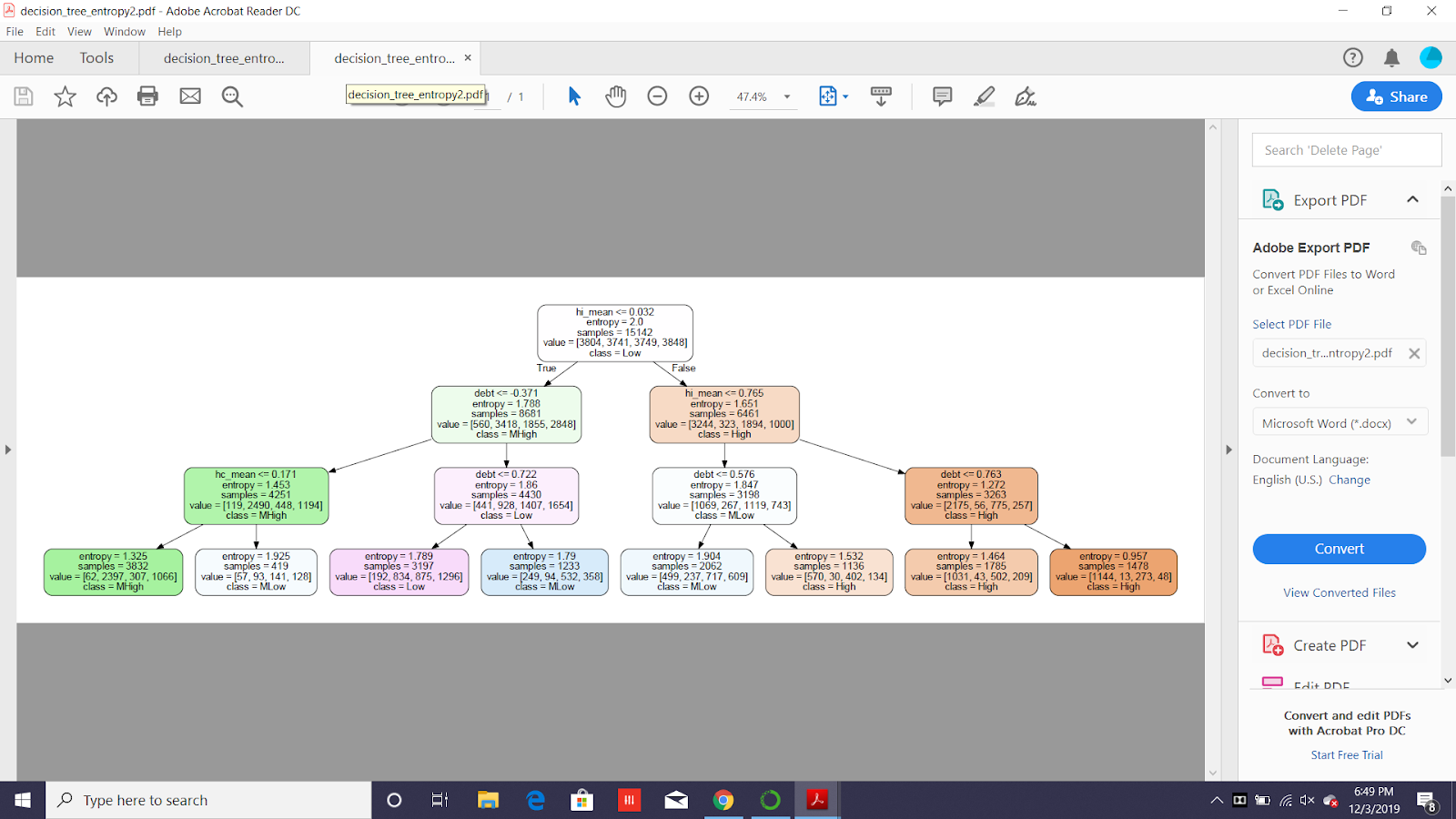
**Analysis Performed on Rent:**

**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:**



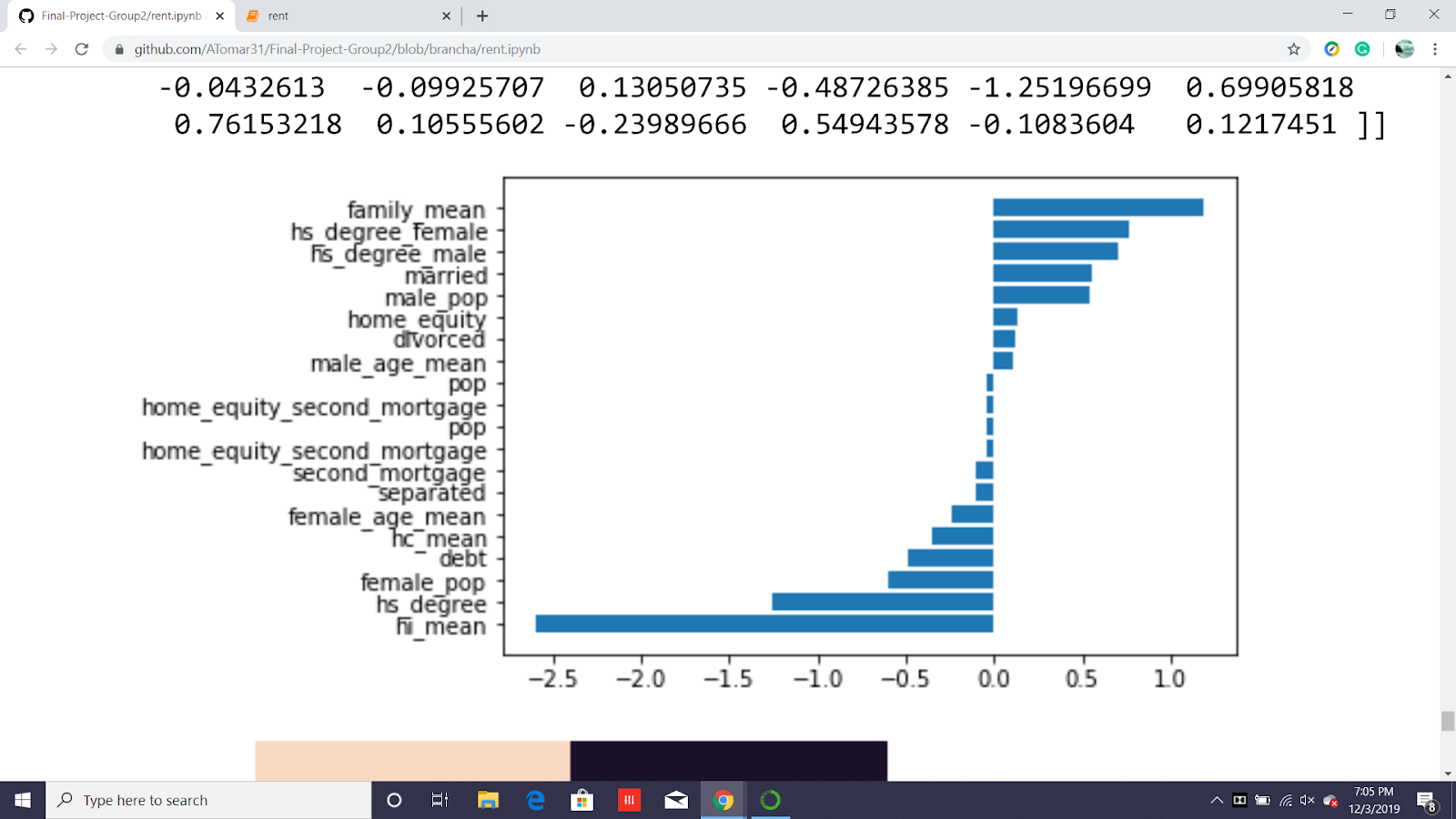
For the first root of our decision tree classifier our model uses hi\_mean or home income mean to divide our dataset – home income being the most inversely correlated to rent. It then goes on to use debt and hc\_mean to split the dataset until it reaches our leaves after the 3rd split. The leaf nodes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

**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 |

SVM Classifier is our most accurate model in this dataset for rent with logistic regression, ensembling and KNN Classifier all coming in above 78% accuracy. This makes sense as all of these models use classification and therefor 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.

**Top 20 Important Features and their coefficients:**



The Coefficients above show that hi\_mean, hs degree and family mean are all highly correlated

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

5

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

((180 – 72) / (180 + 9)) x 100 = 57.14%

\*Internet to include code provided by Amir Jafari GitHub

1. **References**

7