

DEBT AND MORTGAGE DATA ANALYSIS & PREDICTION

Anwasha Tomar, Spencer Staub, Matteo Bucalossi

DATS 6103-Final Group Presentation
December 4, 2019

INTRODUCTION

- Performed analysis on the dynamics of mortgage and debt in US
- The following classification and regression models were used to predict which demographic groups are more affected by debt and mortgage:
 - K Nearest Neighbour
 - Support Vector Machines
 - Logistic Regression
 - Decision Tree
 - Hard-voting Ensembling
 - Random Forest

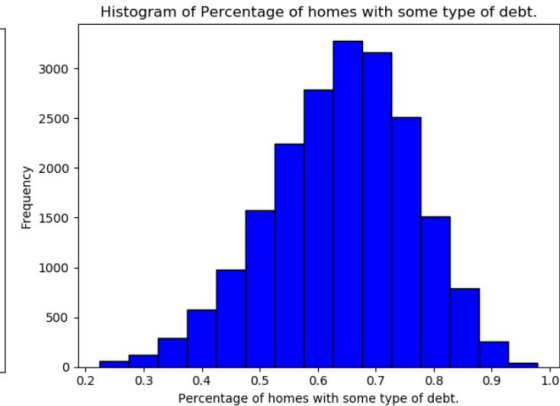
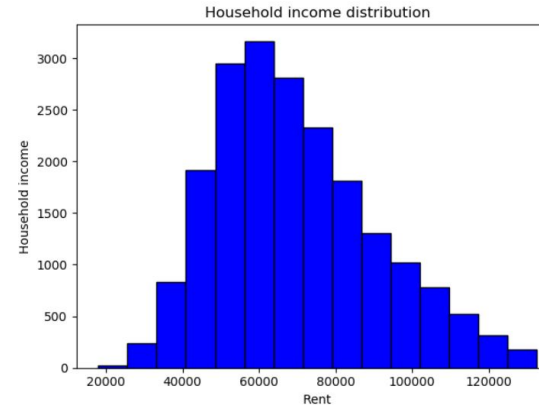
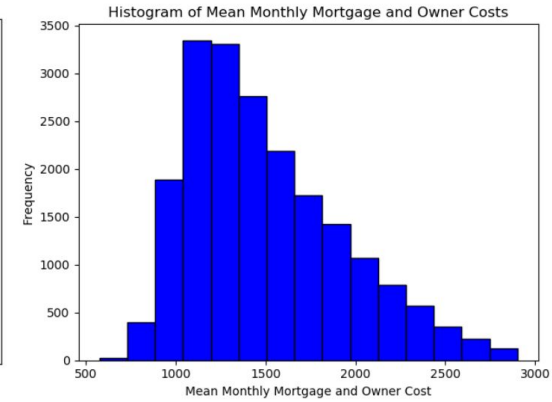
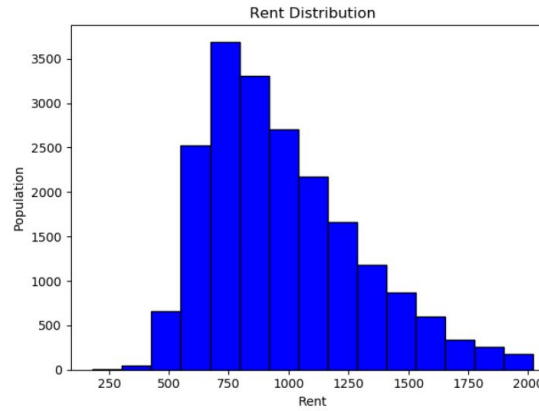
THE DATASET

- Collected by US Census in 2012-2016.
- 39,030 records with 80 features
- Geographic variables were converted to 'object'
- Numeric features as summary statistics:
 - Mortgage and Owner Costs
 - Owner Costs
 - Gross Rent
 - Household Income
 - Family Income
- Numeric features as proportions:
 - Second mortgage
 - Home equity loan
 - Debt
 - High school degree
 - Divorced, separated or married

PREPROCESSING

- Drop irrelevant/redundant columns of summary statistics
 - Removed outliers with function based on quartiles
 - Interpolation for null values with mean of column
-
- 38 features and 20,190 rows
 - Only mean and proportion features

VISUALIZATION OF TARGET VARIABLES RENT AND DEBT:



MODELING PROTOCOL

- Target variables used:
 - Rent Costs (average of gross rent costs)
 - Debt (percentage of houses with debt)
- To perform classification the target variables:
 - Categorize with median split (KNN, SVM, Logit) and quartile split (DT)
- Only the numeric features were used as predictors
- Split data in training (75%) and test (25%)
- The data was the scaled data with standardized based on the model used
- Predict target with test data using classification and regression
- Each model was then evaluated using the confusion matrix and accuracy scores

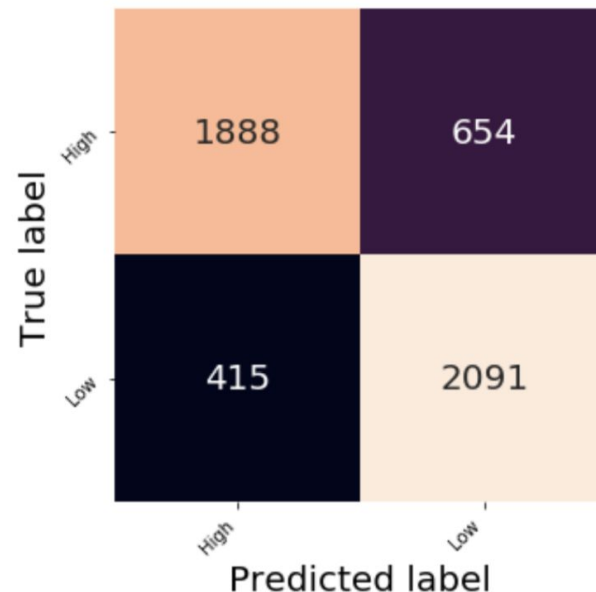
K NEAREST NEIGHBOUR

For KNN we used 9 as our K value since we have 20 features.

The classifier accuracy is found to be more than that of the regressor.

Classifier Accuracy = 78.82

Regressor Accuracy = 53.45



A confusion matrix for a K-Nearest Neighbour classifier. The vertical axis is labeled 'True label' with 'High' and 'Low' categories. The horizontal axis is labeled 'Predicted label' with 'High' and 'Low' categories. The matrix cells contain counts: True High/Predicted High is 1888 (orange), True High/Predicted Low is 654 (dark purple), True Low/Predicted High is 415 (dark blue), and True Low/Predicted Low is 2091 (light orange).

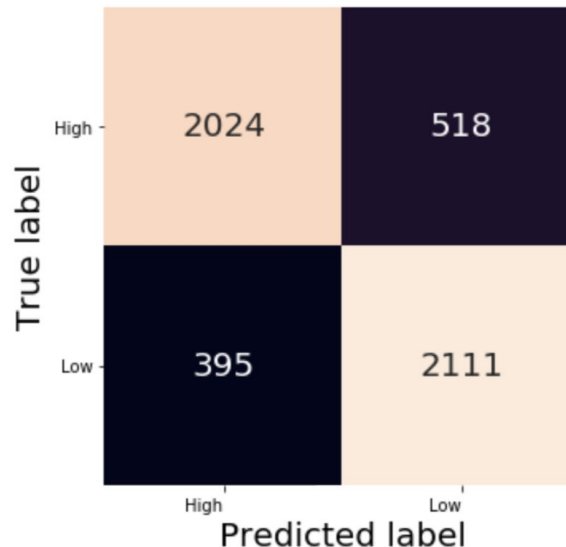
True label	High	Low
	1888	654
Low	415	2091

SUPPORT VECTOR MACHINES

In SVM we used two kernels, linear and radial basis function. We used the linear Kernel for classification and RBF for regression.

Classifier Accuracy = 81.91

Regressor Accuracy = 46.57



A confusion matrix for SVM classification results. The vertical axis is labeled 'True label' with 'High' and 'Low' categories. The horizontal axis is labeled 'Predicted label' with 'High' and 'Low' categories. The matrix cells contain counts: True High predicted High is 2024 (orange), True High predicted Low is 518 (dark purple), True Low predicted High is 395 (dark purple), and True Low predicted Low is 2111 (orange).

True label	High	Low
	2024	518
Low	395	2111

LOGISTIC REGRESSION

Classifier Accuracy = 79.44

Receiver Operating Characteristic
Curve = 86.67

True label	High	Low
	1945	597
Low	441	2065
Predicted label		

DECISION TREES

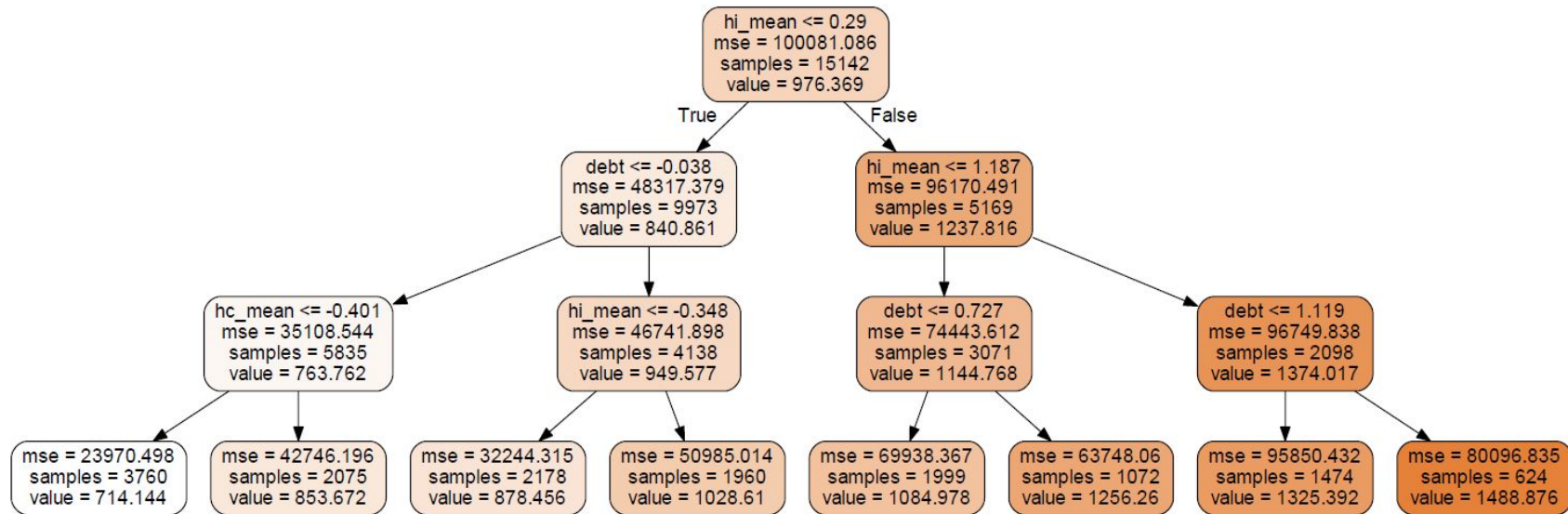
Entropy was used to calculate the homogeneity of a sample.

Classifier Accuracy = 50.09

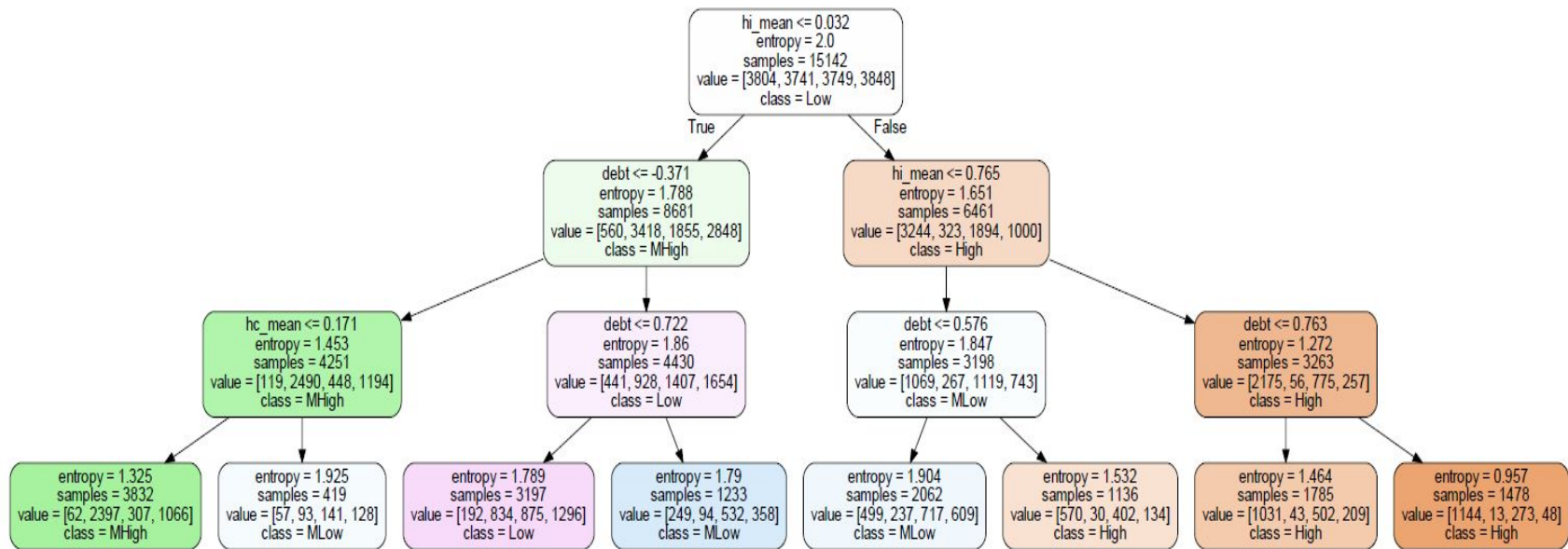
Regressor Accuracy = 49.34

True label	High	MHigh	MLow	Low
	888	17	259	80
	27	787	90	403
	414	108	440	336
Low	115	316	284	484
	High	MHigh	MLow	Low
Predicted label				

REGRESSION TREE



CLASSIFICATION TREE



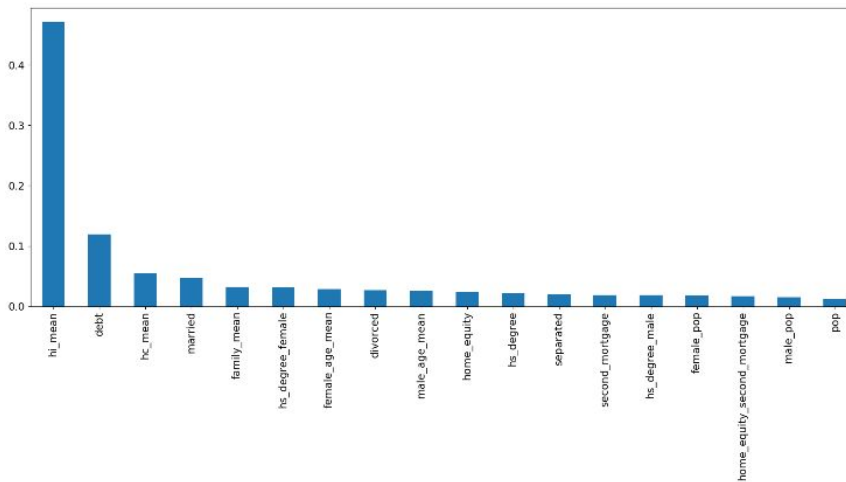
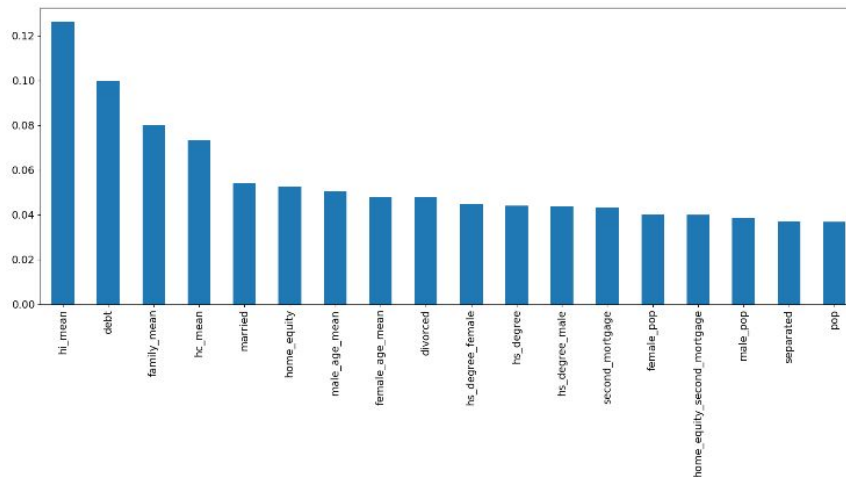
RENT

RANDOM FOREST

A hundred trees were used as estimators

Classifier Accuracy = 54.42

Regressor Accuracy = 61.74



ENSEMBLING

A hard voting classifier was used with the following models:

- Logistic Regression,
- Decision Trees

Accuracy = 78.88

MODEL USED	ACCURACY
SVM Classifier	81.91
Logistic Regression	79.43
Ensembling	78.88
KNN Classifier	78.82
Random Forest Regression	61.74
Random Forest Classifier	54.42
KNN Regressor	53.44
Decision Tree Classifier	50.09
Decision Tree Regression	49.34
SVM Regression	46.57

K NEAREST NEIGHBOR

As we saw in Rent analysis we used 9 neighbours for classification and we see that the classification accuracy is still higher than that of regressor.

Classifier Accuracy = 77.69

Regressor Accuracy = 56.72

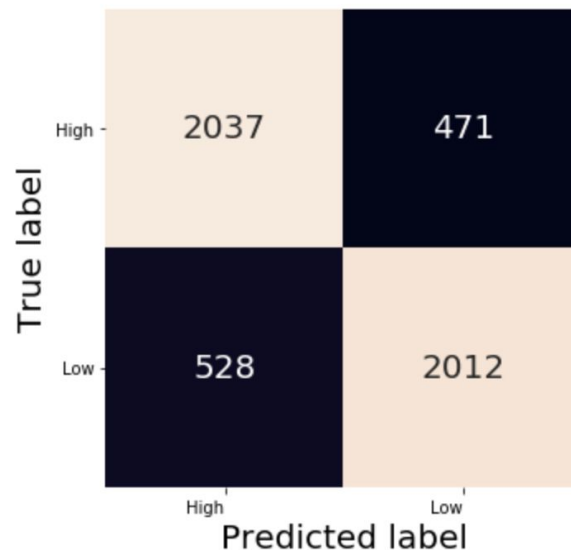
True label	High	Low
	1958	550
Low	576	1964
Predicted label		

SUPPORT VECTOR MACHINES

In SVM we used two kernels, linear and radial basis function again for analysis.

Classifier Accuracy = 80.21

Regressor Accuracy = 60.71



A confusion matrix for a binary classifier. The vertical axis is labeled 'True label' with 'High' and 'Low' categories. The horizontal axis is labeled 'Predicted label' with 'High' and 'Low' categories. The matrix cells contain counts: True High & Predicted High is 2037 (light orange), True High & Predicted Low is 471 (dark blue), True Low & Predicted High is 528 (dark blue), and True Low & Predicted Low is 2012 (light orange).

True label	High	Low
	2037	471
Low	528	2012

LOGISTIC REGRESSION

Classifier Accuracy = 75.36

Receiver Operating Characteristic
Curve = 83.71

True label	High	Low
	1807	701
Low	543	1997
Predicted label		

DECISION TREES

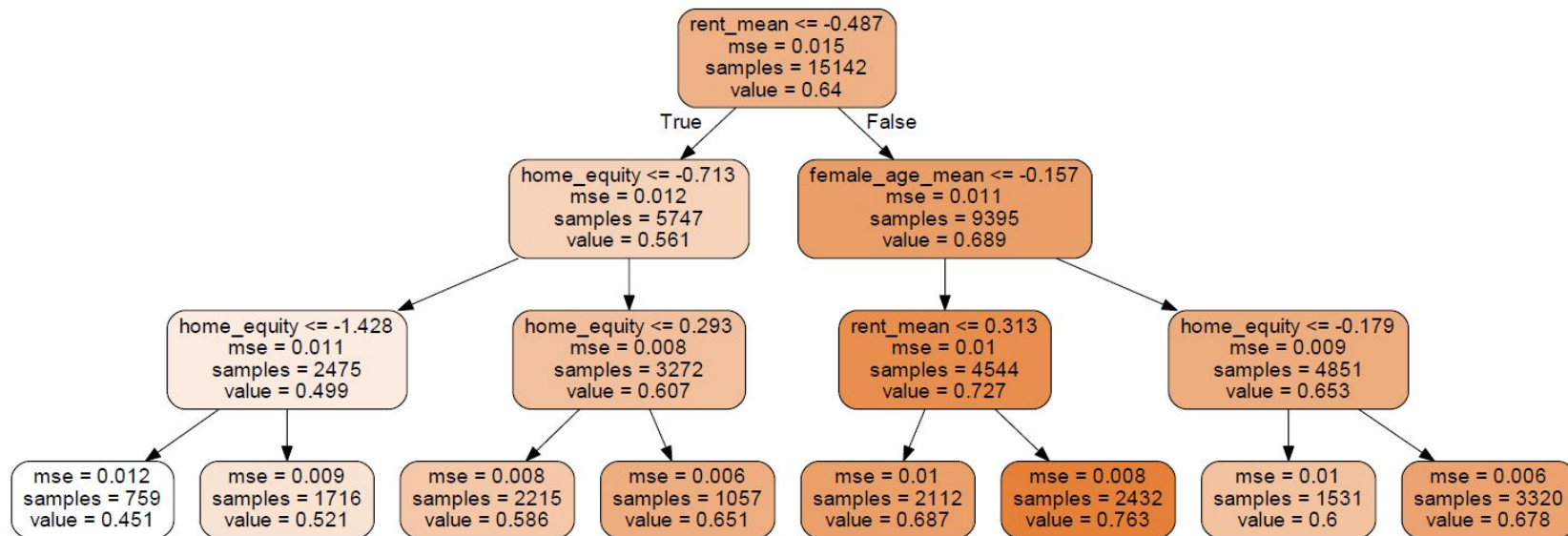
Entropy was used to calculate the homogeneity of a sample. Here the regressor accuracy is slightly higher than that of the classifier.

Classifier Accuracy = 51.58

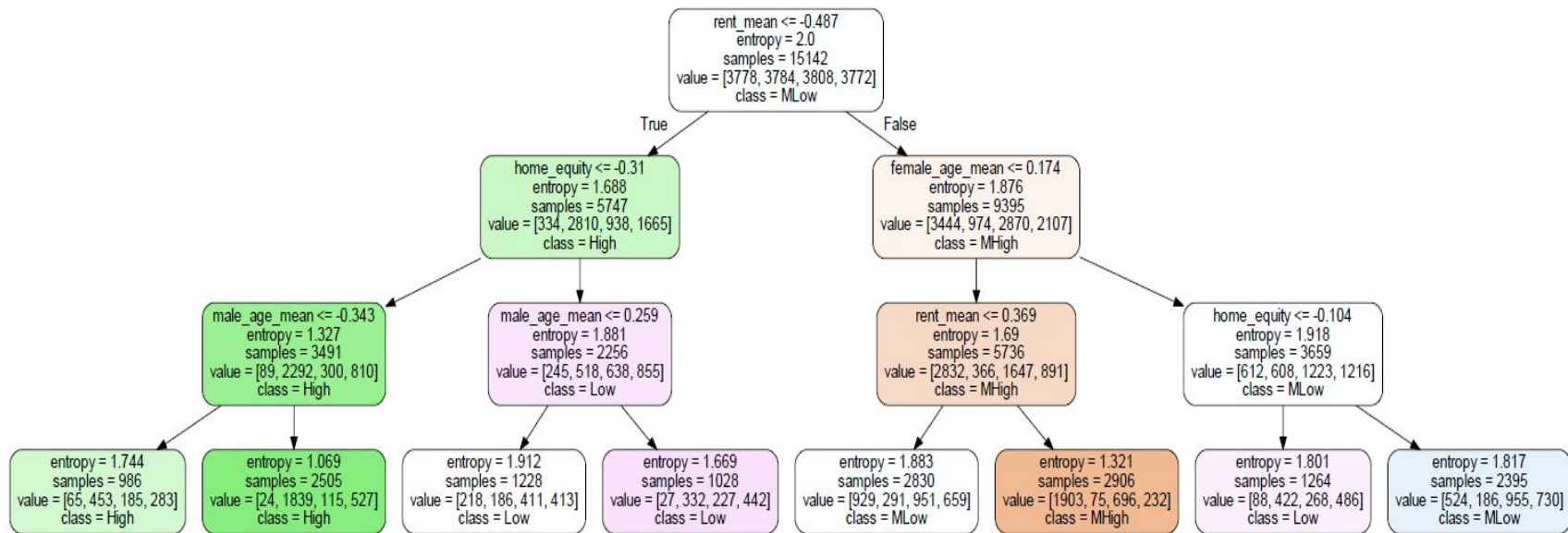
Regressor Accuracy = 54.16

True label	MHigh	High	MLow	Low
	669	21	480	100
	32	900	123	209
	242	104	604	288
Low	105	387	353	431
	MHigh	High	MLow	Low
Predicted label				

REGRESSION TREE



CLASSIFICATION TREE



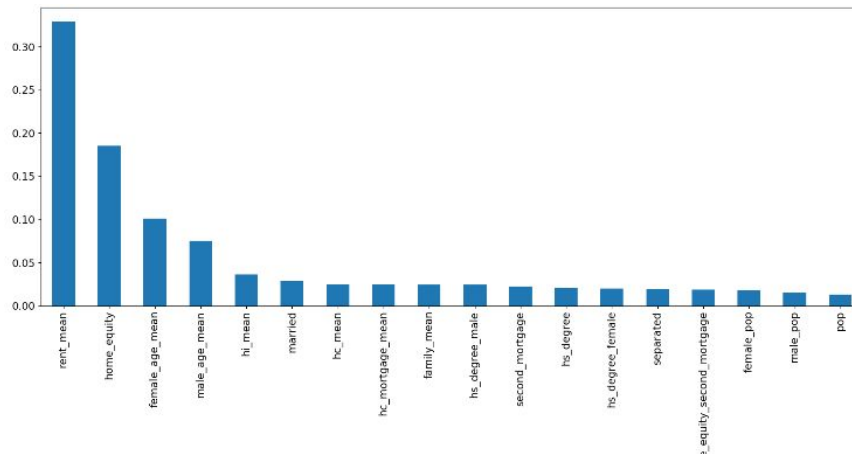
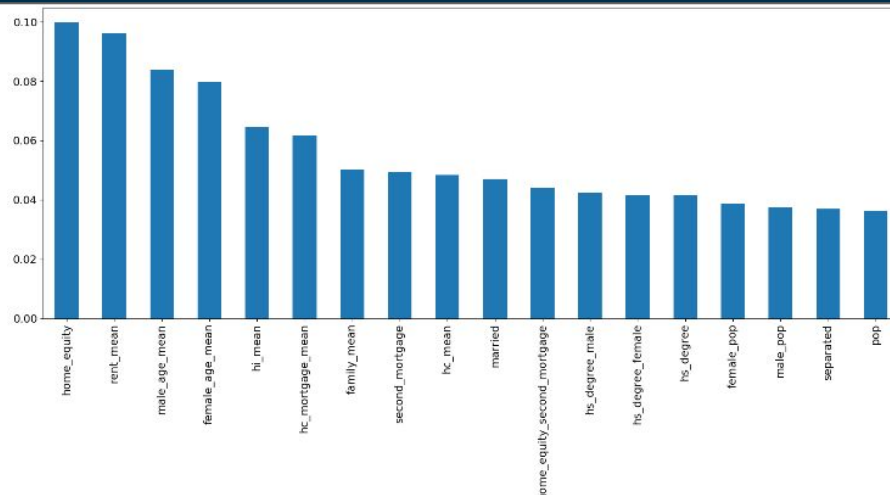
DEBT

RANDOM FOREST

A hundred trees were used as estimators

Classifier Accuracy = 54.46

Regressor Accuracy = 62.33



ENSEMBLING

A hard voting classifier was used with the following models:

- Logistic Regression,
- Decision Trees

Accuracy = 76.76

MODEL USED	ACCURACY
SVM Classifier	80.21
KNN Classifier	77.69
Ensembling	76.76
Logistic Regression	75.36
Random Forest Regression	62.33
SVM Regression	60.71
KNN Regressor	56.71
Random Forest Classifier	54.46
Decision Tree Regression	54.16
Decision Tree Classifier	51.58

CONCLUSIONS & FUTURE WORK

CONCLUSION:

- Classification in most cases performs better than regression. This may be due the nature of our data which is in summary format.
- Support Vector Machines Classifier performed the best for fitting and classification of the data

FUTURE WORK:

- If the data was record based we could have found more accurate information.
- Record based data would also enable us to make conclusions based on the geographic area.