DEBT AND MORTGAGE DATA ANALYSIS & PREDICTION

Anwesha 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

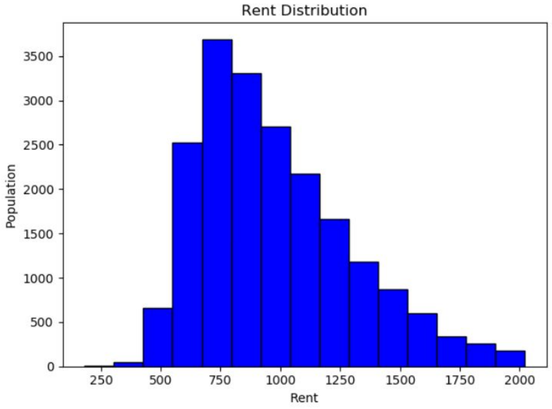
- Numeric features as proportions:
  - Second mortgage
  - Home equity loan
  - Debt
  - High school degree
  - Divorced, separated or married

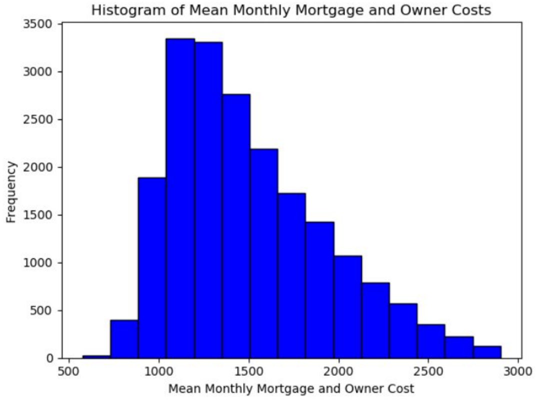
PREPROCESSING

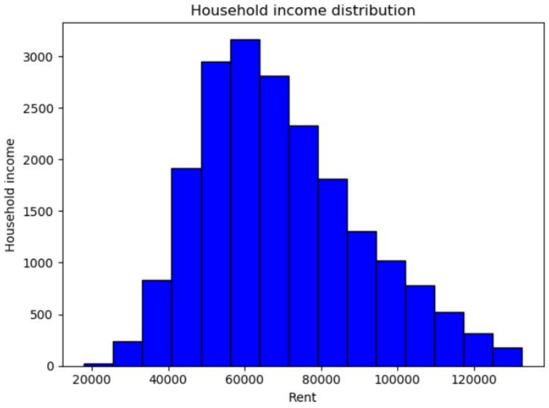
- Drop irrelevant/redundant columns of summary statistics
- Removed outliers with function based on quartiles
- Interpolation for null values with mean of column

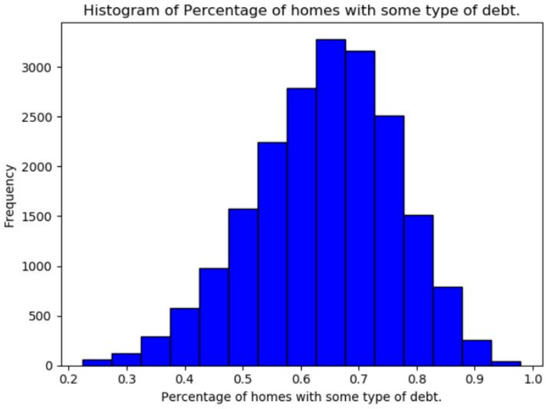
- 38 features and 20,190 rows
- Only mean and proportion features

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)
- Only the numeric features were used as predictors
  - Features that were CDF (Cumulative Distribution Functions) were removed from the analysis due to extremely high correlation
- Both Classification and Regression Models were used
  - Categorize with median split (KNN, SVM, Logit) and quartile split (DT)
- Split data in training (75%) and test (25%)
- The features were scaled with the standardized function
- Classification models were evaluated using the predict target, confusion matrix and accuracy score functions
- Regression models were tested for accuracy using the score function

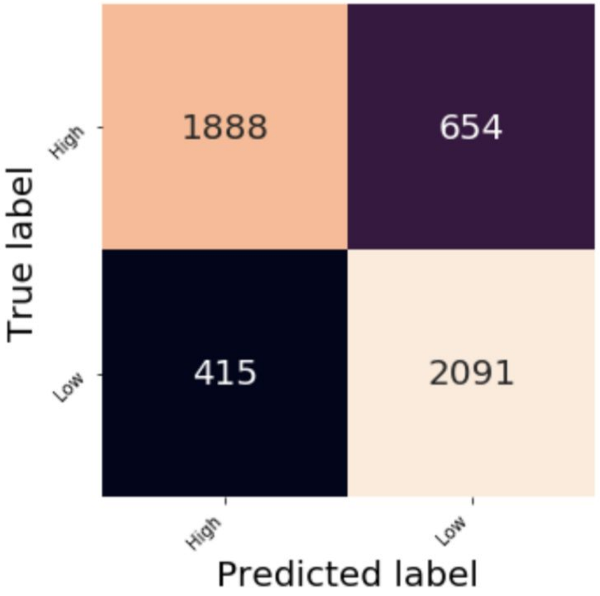
K NEAREST NEIGHBOUR

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

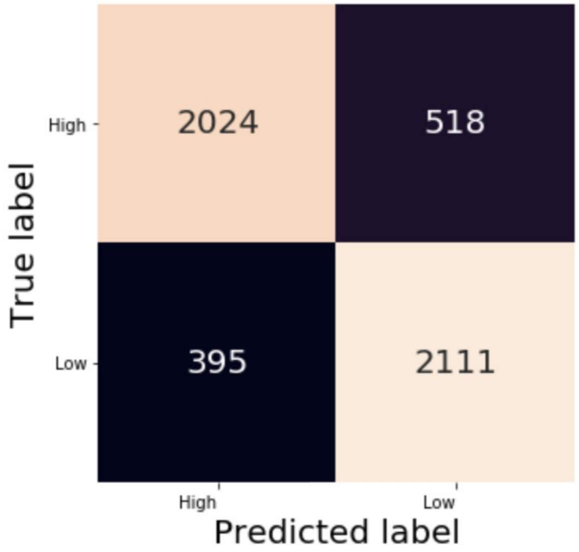


SUPPORT VECTOR MACHINES

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

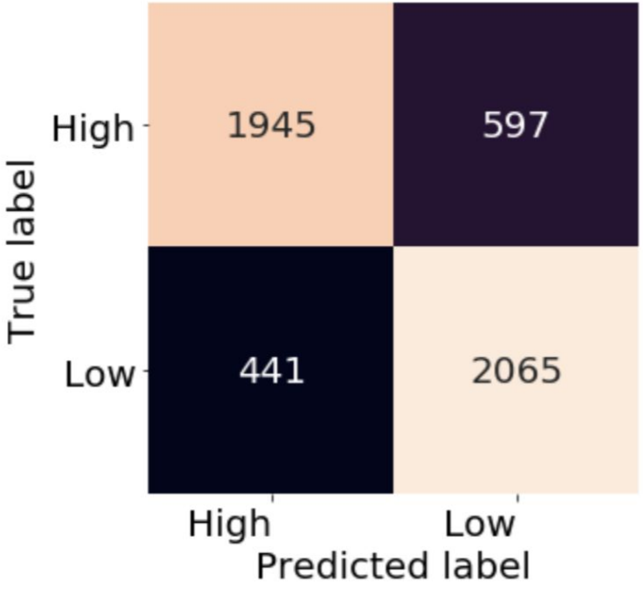


RENT

LOGISTIC REGRESSION

Classifier Accuracy = 79.44

Receiver Operating Characteristic Curve = 86.67

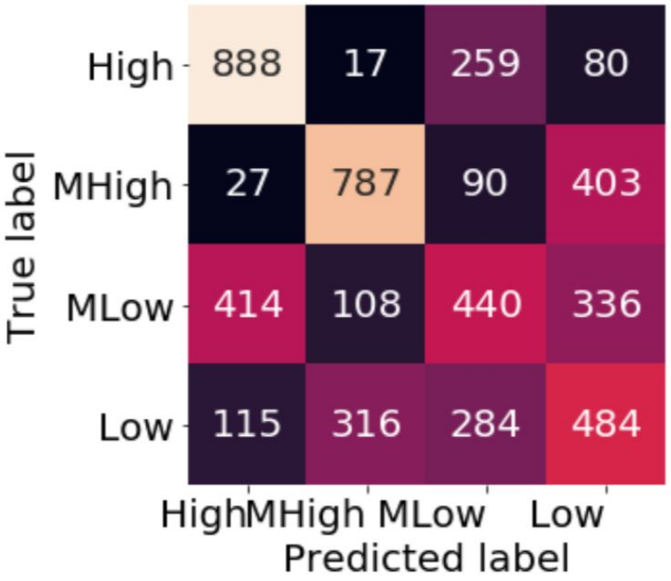


DECISION TREES

Entropy was used to calculate the homogeneity of a sample.

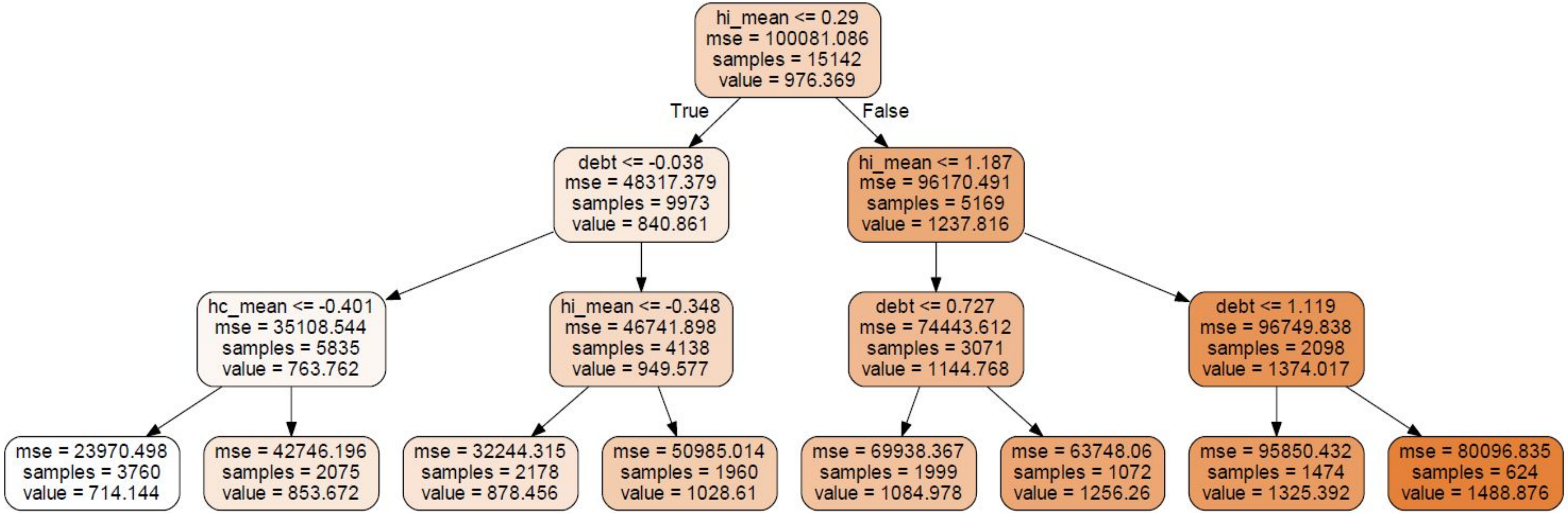
Classifier Accuracy = 50.09

Regressor Accuracy = 49.34



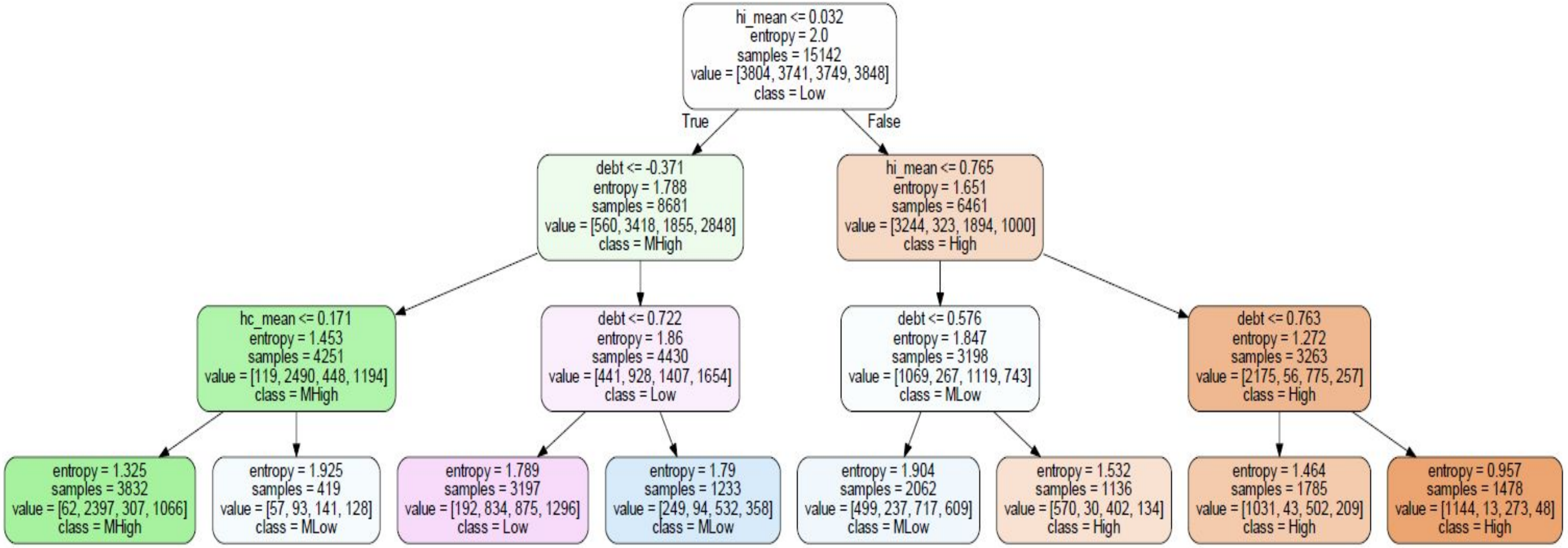
RENT

REGRESSION TREE



RENT

CLASSIFICATION TREE



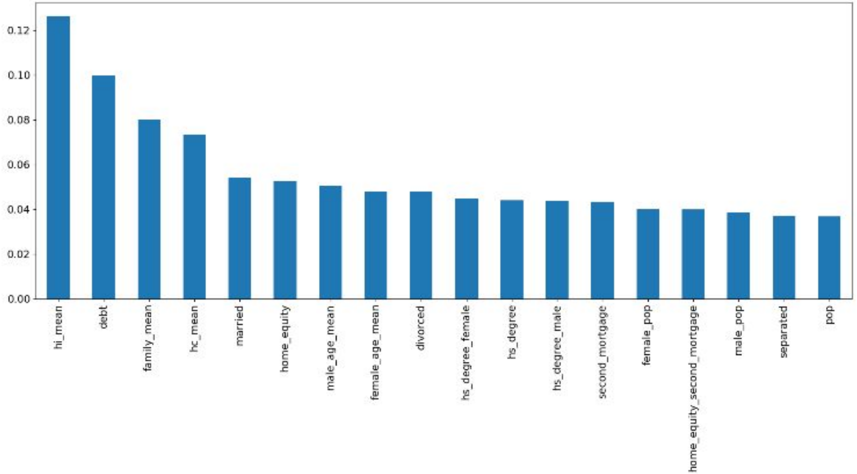
RENT

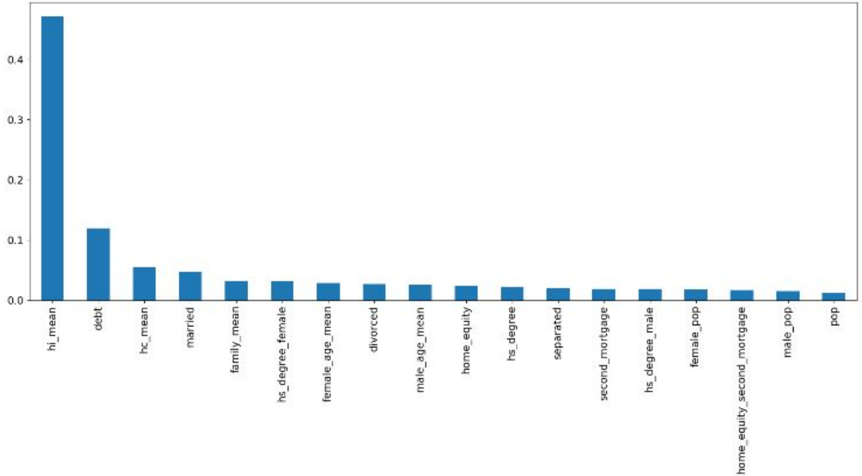
RANDOM FOREST

A hundred trees were used as estimators

Classifier Accuracy = 54.42

Regressor Accuracy = 61.74





RENT

ENSEMBLING

A hard voting classifier was used with the following models:

- Logistic Regression
- Decision Trees

Accuracy = 78.88

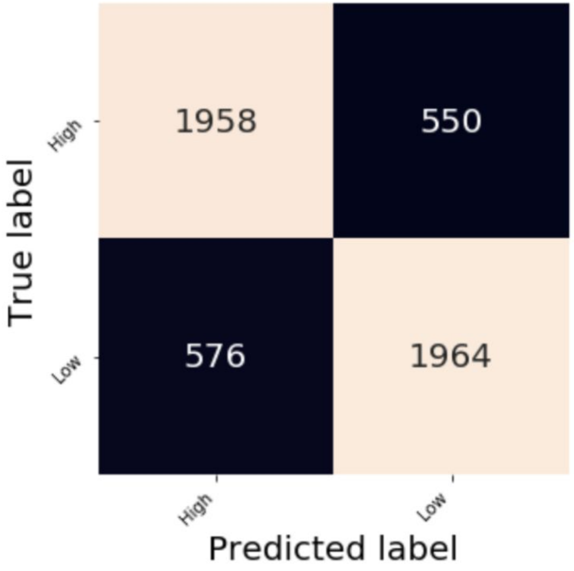
MODEL USED	ACCURACY
SVM Classifier	81.91
Logistic Regression	79.43
Ensembling	78.88
KNN Classifier	78.82
Random Forest Regression	61.74
Random Forest Classifier	54.42
KNN Regressor	53.44
Decision Tree Classifier	50.09
Decision Tree Regression	49.34
SVM Regression	46.57

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



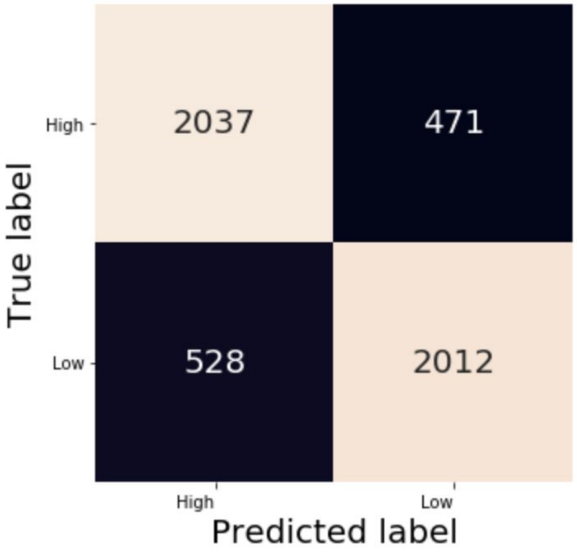
DEBT

SUPPORT VECTOR MACHINES

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

Classifier Accuracy = 80.21

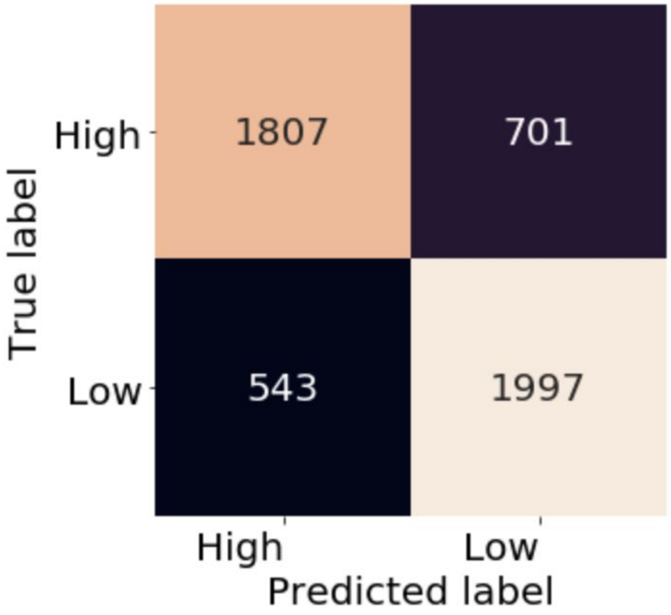
Regressor Accuracy = 60.71



LOGISTIC REGRESSION

Classifier Accuracy = 75.36

Receiver Operating Characteristic Curve = 83.71

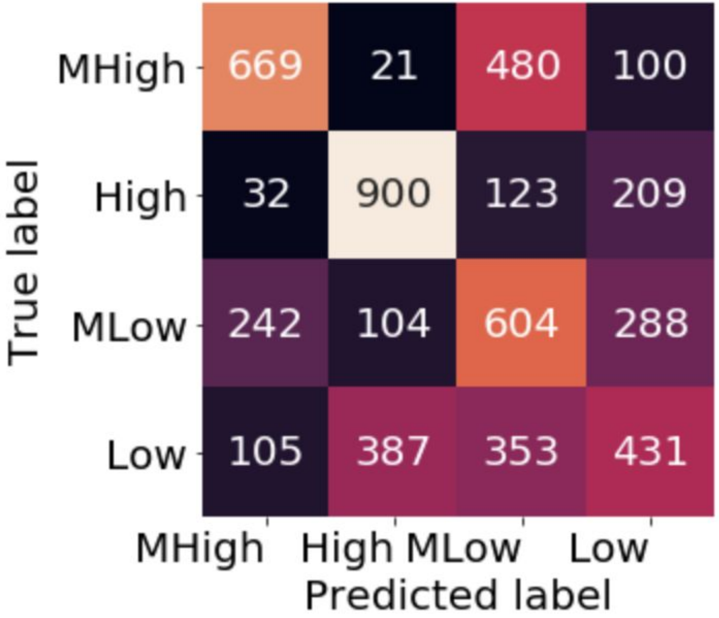


DECISION TREES

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

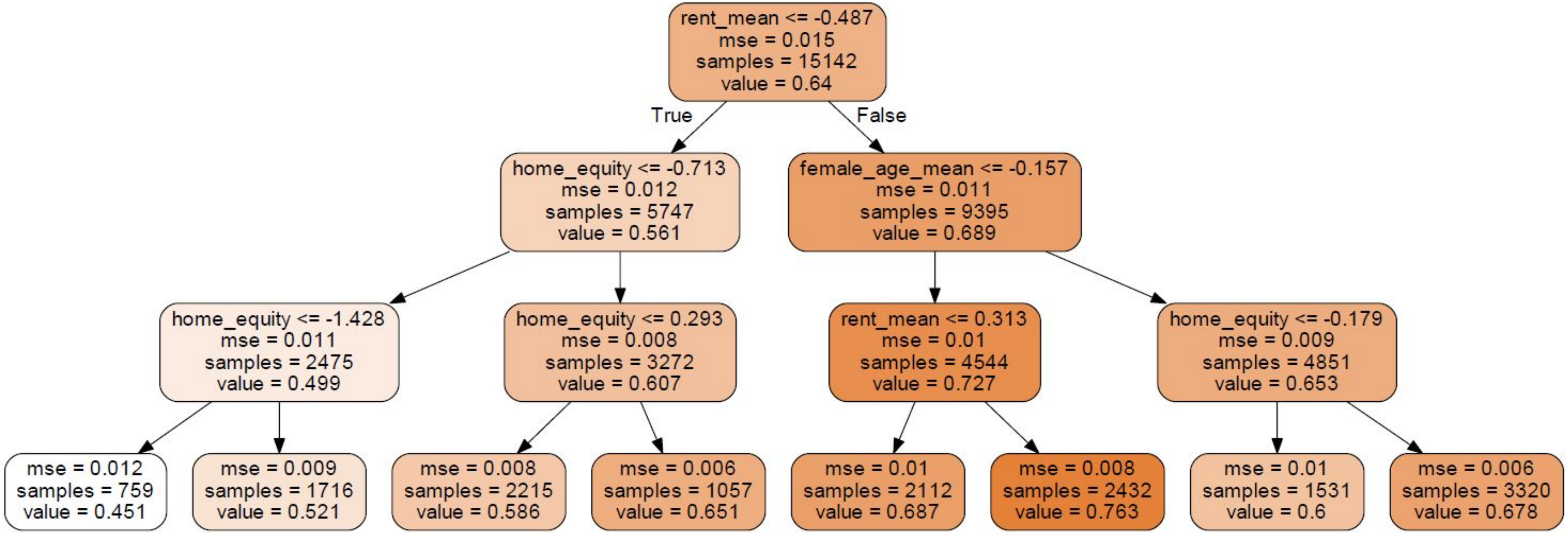
Classifier Accuracy = 51.58

Regressor Accuracy = 54.16



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

REGRESSION TREE



rent_mean <= -0.487
mse = 0.015
samples = 15142
value = 0.64

True

False

home_equity <= -0.713
mse = 0.012
samples = 5747
value = 0.561

female_age_mean <= -0.157
mse = 0.011
samples = 9395
value = 0.689

home_equity <= -1.428
mse = 0.011
samples = 2475
value = 0.499

home_equity <= 0.293
mse = 0.008
samples = 3272
value = 0.607

rent_mean <= 0.313
mse = 0.01
samples = 4544
value = 0.727

home_equity <= -0.179
mse = 0.009
samples = 4851
value = 0.653

mse = 0.012
samples = 759
value = 0.451

mse = 0.009
samples = 1716
value = 0.521

mse = 0.008
samples = 2215
value = 0.586

mse = 0.006
samples = 1057
value = 0.651

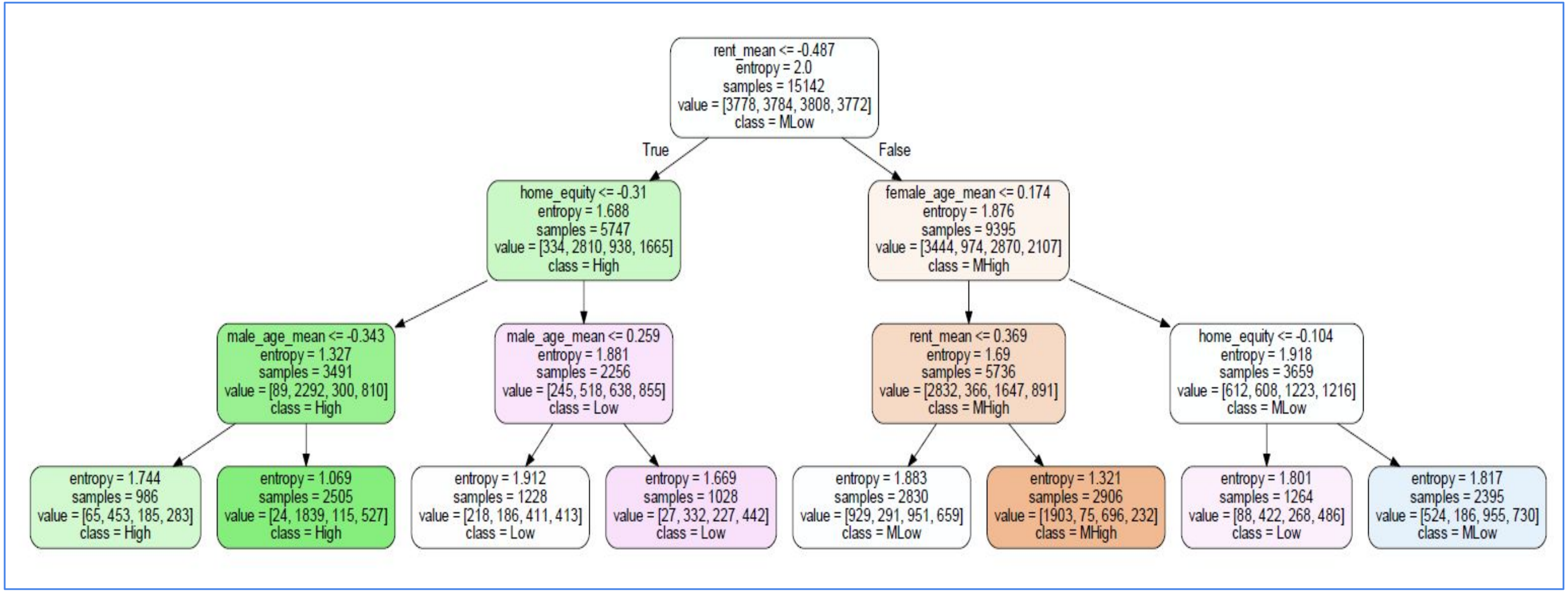
mse = 0.01
samples = 2112
value = 0.687

mse = 0.008
samples = 2432
value = 0.763

mse = 0.01
samples = 1531
value = 0.6

mse = 0.006
samples = 3320
value = 0.678

CLASSIFICATION TREE



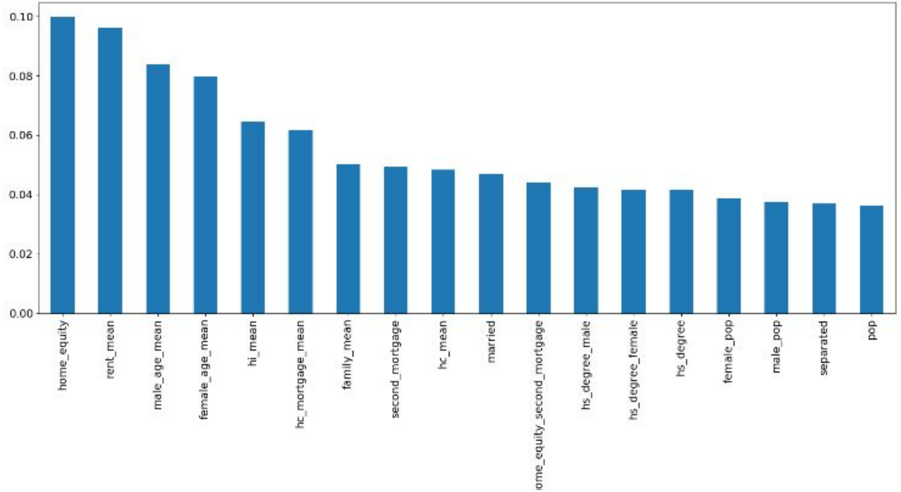
DEBT

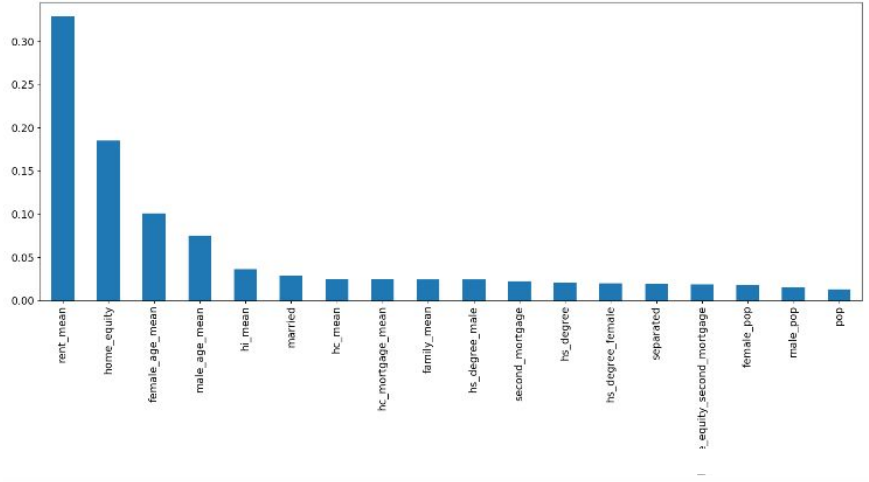
RANDOM FOREST

A hundred trees were used as estimators

Classifier Accuracy = 54.46

Regressor Accuracy = 62.33





DEBT

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.