Capstone project-2: Housing Affordability Prediction

Milestone Report

Gautham Gowda

Contents

Problem statement- why is it useful to answer the question

Housing dataset contains many important features about housing costs, income and burden of homeowners. This useful data is used to predict the housing affordability of consumers applying for mortgage to determine how much mortgage a consumer can afford. There are a lot of features in the data, but our aim is to build a model to use minimum features while keeping the prediction accuracy at the highest level.

Clients/ Intended audience

This is a very good prediction model for many users. The model can be used by individual home buyers wanting to know the house they can afford based on their income, costs, housing locations. It can be used by the lenders as well.

The clientele for this analysis could be banks or other mortgage lending institutions wanting to know the housing affordability of consumers based on the average home price in the area, consumer income levels and burden, interest rates etc.

The dataset has nearly 100 features to consider for making a decision. This model looks at the 10 most important features affecting housing affordability, thus reducing the time and effort in decision making.

.

Dataset used for the investigation

The dataset used for this analysis is from HADS database from the huduser.gov website and the data source is listed below:

https://www.huduser.gov/portal/datasets/hads/hads.html

The database consists of housing affordability data and contains many features such as income, burden, average housing cost, poverty income that are relevant for the analysis. The following table summarizes all the features available:

ABL30	Extremely Low Income Adjusted for # of Bedrooms, Only national, 2003 & later
ABL50	Very Low Income Adjusted for # of Bedrooms
ABL80	Low Income Adjusted for # of Bedrooms
ABLMED	Median Income Adjusted for # of Bedrooms
AGE	Age of head of household, 1985-1995
AGE1	Age of head of household, 1997& later
APLMED	Median Income Adjusted for # of Persons
ASSISTED	Assisted Housing
BEDRMS	# of bedrooms in unit
BUILT	Year unit was built
BURDEN	Housing cost as a fraction of income
CONTROL	AHS control number
COST06	Housing cost at 6 percent interest
COST06RELAMICAT	Cost06 Relative to Median Income (Category)
COST06RELAMIPCT	Cost06 Relative to Median Income (Percent)
COST06RELFMRCAT	Cost06 Relative to FMR (Category)
COST06RELFMRPCT	Cost06 Relative to FMR (Percent)
COST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)
COST06RELPOVPCT	Cost06 Relative to Poverty Income (Percent)
COST08	Housing cost at 8 percent interest
COST08RELAMICAT	Cost08 Relative to Median Income (Category)
COST08RELAMIPCT	Cost08 Relative to Median Income (Percent)
COST08RELFMRCAT	Cost08 Relative to FMR (Category)
COST08RELFMRPCT	Cost08 Relative to FMR (Percent)
COST08RELPOVCAT	Cost08 Relative to Poverty Income (Category)
COST08RELPOVPCT	Cost08 Relative to Poverty Income (Percent)
COST12	Housing cost at 12 percent interest
COST12RELAMICAT	Cost12 Relative to Median Income (Category)
COST12RELAMIPCT	Cost12 Relative to Median Income (Percent)
COST12RELFMRCAT	Cost12 Relative to FMR (Category)
COST12RELFMRPCT	Cost12 Relative to FMR (Percent)
COST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)
COST12RELPOVPCT	Cost12 Relative to Poverty Income (Percent)
COSTMED	Housing cost at Median interest
COSTMedRELAMICAT	CostMed Relative to Median Income (Category)
COSTMedRELAMIPCT	CostMed Relative to Median Income (Percent)
COSTMedRELFMRCAT	CostMed Relative to FMR (Category)
COSTMedRELFMRPCT	CostMed Relative to FMR (Percent)
COSTMedRELPOVCAT	CostMed Relative to Poverty Income (Category)
COSTMedRELPOVPCT	CostMed Relative to Poverty Income (Percent)
FMR	Fair market rent (average)
FMTASSISTED	Assisted Housing
FMTBEDRMS	# of bedrooms in unit
FMTBUILT	Year unit was built
FMTBURDEN	Cost Burden
FMTCOST06RELAMICAT	Cost06 Relative to Median Income (Category)
FMTCOST06RELFMRCAT	Cost06 Relative to FMR (Category)
FMTCOST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)
FMTCOST08RELAMICAT	Cost08 Relative to Median Income (Category)
FMTCOST08RELFMRCAT	Cost08 Relative to FMR (Category)
FMTCOST08RELPOVCAT	Cost08 Relative to Poverty Income (Category)
FMTCOST12RELAMICAT	Cost12 Relative to Median Income (Category)

FMTCOST12RELFMRCAT	Cost12 Relative to FMR (Category)
FMTCOST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)
FMTCOSTMEDRELAMICAT	CostMed Relative to Median Income (Category)
FMTCOSTMEDRELFMRCAT	CostMed Relative to FMR (Category)
FMTCOSTMEDRELPOVCAT	CostMed Relative to Poverty Income (Category)
FMTINCRELAMICAT	HH Income Relative to Median Income (Category)
FMTINCRELFMRCAT	HH Income Relative to FMR (Category)
FMTINCRELPOVCAT	HH Income Relative to Poverty Income (Category)
FMTMETRO	CENTRAL CITY / SUBURBAN STATUS, National 1985-1995, all metro
FMTMETRO3	CENTRAL CITY / SUBURBAN STATUS, National 1997 & later
FMTOWNRENT	Owner/Renter Status (adjusted)
FMTREGION	Census Region, National only
FMTSTATUS	Occupancy Status
FMTSTRUCTURETYPE	Structure Type
FMTZADEQ	ADEQUACY OF UNIT
GL30	Growth-adjusted extremely low income, National 2003 & later
GL50	Growth-adjusted very low income
GL80	Growth-adjusted low income
GLMED	Growth-adjusted median income
INCRELAMICAT	HH Income relative to AMI (category)
INCRELAMIPCT	HH Income relative to AMI (percent)
INCRELFMRCAT	HH Income Relative to FMR (Category)
INCRELFMRPCT	HH Income Relative to FMR (Percent)
INCRELPOVCAT	HH Income Relative to Poverty Income (Category)
INCRELPOVPCT	HH Income Relative to Poverty Income (Percent)
IPOV	Poverty Income
ISTATUS	Interview status, 1985-1995
L30	Extremely low income limit (average), National 2003 & later
L50	Very low income limit (average)
L80	Low income limit (average)
LMED	Area median income (average)
METRO3	CENTRAL CITY / SUBURBAN STATUS, National 1997 & later
NUNITS	# of units in building
OTHERCOST	Insurance, condo, land rent, other mobile home fees
OWNRENT	Tenure (adjusted)
PER	# of persons in household
REGION	Census Region
ROOMS	# of rooms in unit
SMSA	1980 design PMSA code, Metro only
STATUS	Interview status, 1997 & later
STRUCTURETYPE	Recoded structure type
TENURE	Owner/renter status of unit
TOTSAL	Total Wage Income
TYPE	Structure Type
UTILITY	Monthly utility cost
VACANCY	Vacancy status
VALUE	Current market value of unit
WEIGHT	Final weight
ZADEQ	Recoded adequacy of housing
ZINC2	Household Income
ZSMHC	Monthly housing costs

Data cleaning and wrangling

- Data cleaning is done to eliminate discrepancies in data formats. If we load the CSV file without formatting, many integer variables are loaded as objects containing data between single quotes representing string data. To load the data properly while importing the file, the parameter quotechar=" ' " is used in csv import to maintain the data type consistent between features.
- There are also formatted duplicate features that start with the characters 'FMT' such as 'FMTREGION', which is same as variable 'REGION'. All these duplicate features are eliminated and the total features reduced to 74 from 99
- Some of the features such as Vacancy, Status, Tenure, and Interview Status have only one value in the set and is not useful for the analysis. These features were deleted.
- Missing values: There are many variables with missing values. Some of these variables are continuous and some are discrete and categorical. Used the pandas data-frame method to fill the missing values in the columns:

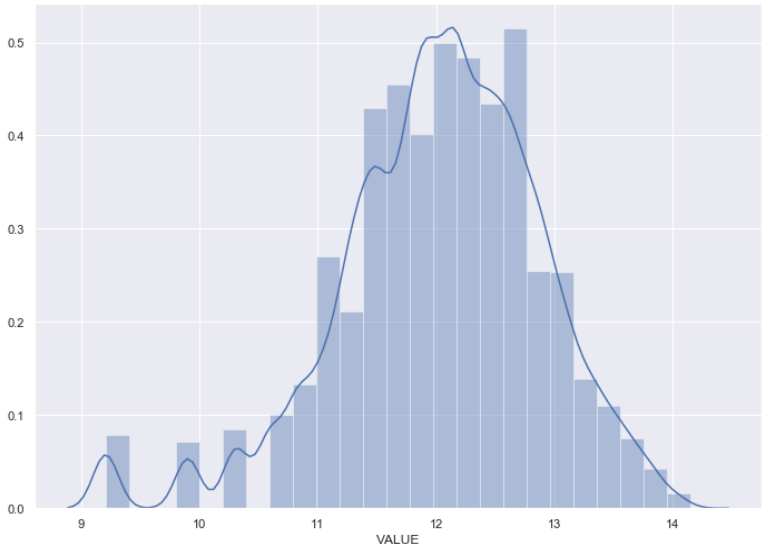
  data = df.fillna(method='ffill').fillna(method='bfill')

  This method to replace all missing values with forward fill and backward fill data. If the missing values are a lot for certain features, those rows were deleted entirely which resulted in data reduction from sixty six thousand variables to thirty five thousand variables.

- Outliers: There is significant outliers present in some variables. Variables 'Age1', 'Value' and other continuous variables have some outliers such as value at $1 and age equal to -6, which are removed from the data frame.

Data visualization

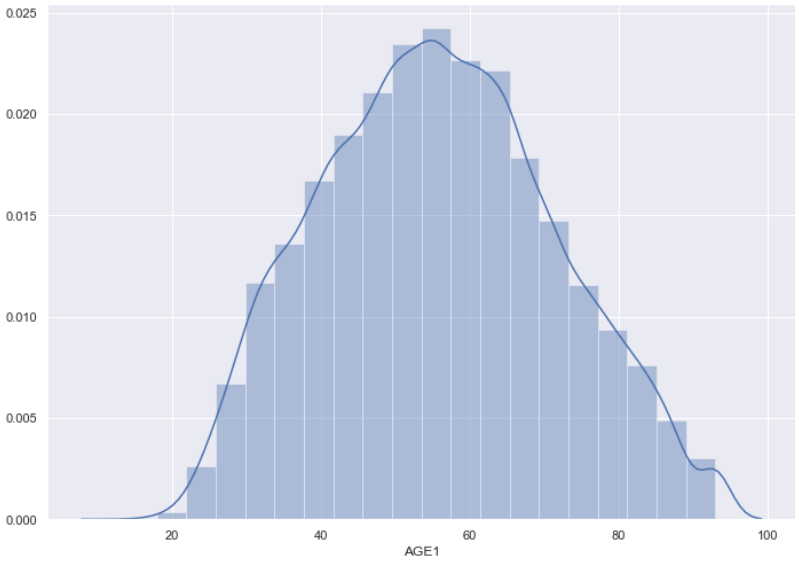
Following plots provide the visualization of the data for analysis. The continuous variables are plotted in histograms to check for normality.



Summary Stats	
count	34339
mean	231919.97
std	191595.55
min	10000
25%	100000
50%	180000
75%	300000
max	1400000

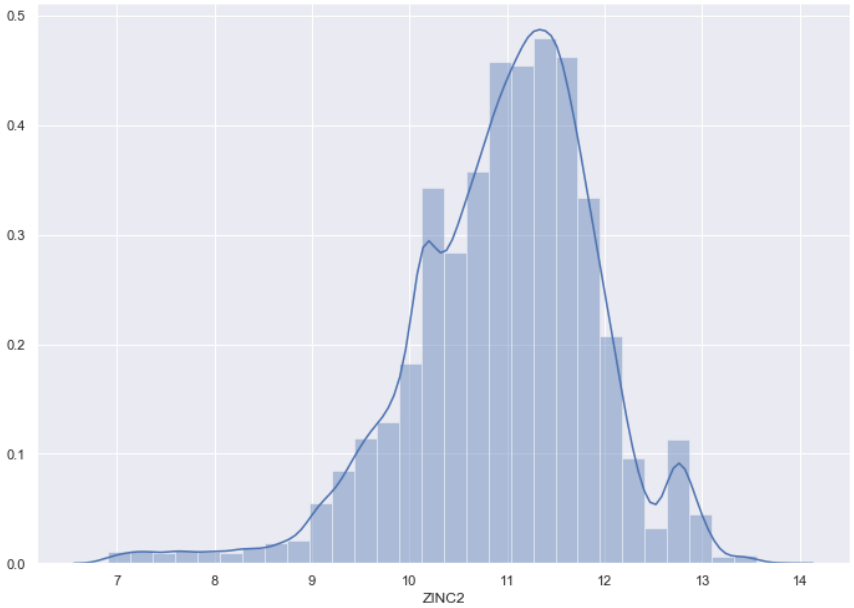
Fig-1: Normal distribution of home values

The above figure is histogram of the home values plotted in the log scale to demonstrate normal distribution.



Summary Stats	
count	34339
mean	55.74
std	15.7
min	14
25%	44
50%	55
75%	66
max	93

Fig-2: Age distribution of head of household and summary statistics



Summary Stats	
count	34339
mean	231920
std	191596
min	10000
25%	100000
50%	180000
75%	300000
max	1400000

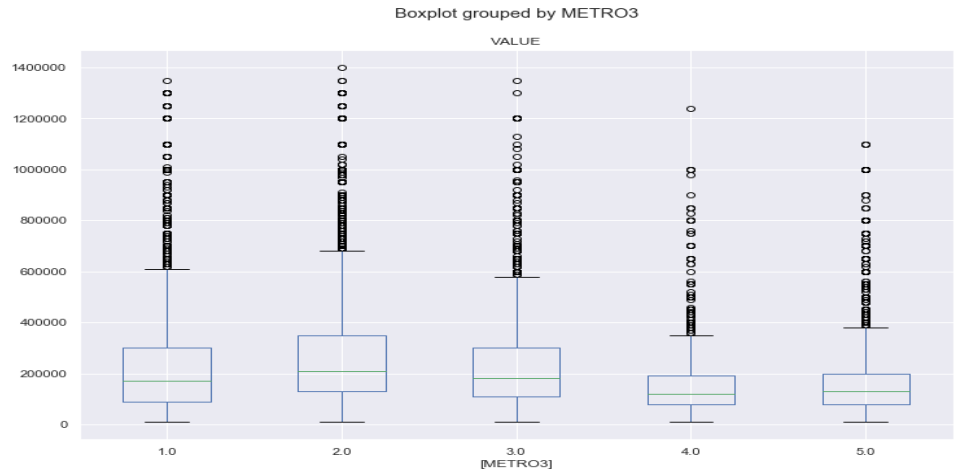
Fig-3: Household Income distribution plotted in a log scale

The above three figures are the examples of the continuous variables showing normal distribution.

Exploratory Data Analysis (EDA)

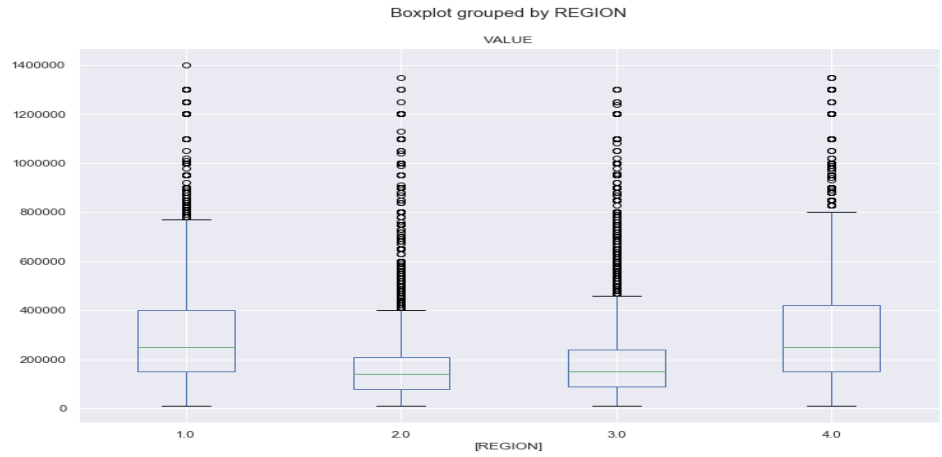
Boxplots to check for relationship between dependent and independent variables

Visual analysis using the box plots to see the response variable vs the independent variables give us information about the relationships. Below we have two plots with home prices grouped by Metro and Region.



	count	mean	std	min	25%	50%	75%	max
1.0	7811.0	233642.30	208241.55	10000.0	90000.0	170000.0	300000.0	1350000.0
2.0	14560.0	267257.55	202621.43	10000.0	130000.0	210000.0	350000.0	1400000.0
3.0	5123.0	227382.39	172433.49	10000.0	110000.0	180000.0	300000.0	1350000.0
4.0	2097.0	152594.18	123848.68	10000.0	80000.0	120000.0	190000.0	1240000.0
5.0	4748.0	160652.91	133056.45	10000.0	80000.0	130000.0	200000.0	1100000.0

Fig-4: Box plot & summary stats of home values for each metro region (1= central city, 2-5 =suburban zones)



	count	mean	std	min	25%	50%	75%	max
1.0	8395.0	298284.69	201480.20	10000.0	150000.0	250000.0	400000.0	1400000.0
2.0	10027.0	172948.04	141814.09	10000.0	80000.0	140000.0	210000.0	1350000.0
3.0	10392.0	188901.08	161205.12	10000.0	90000.0	150000.0	240000.0	1300000.0
4.0	5525.0	319020.81	238835.12	10000.0	150000.0	250000.0	420000.0	1350000.0

Fig-5: Box plot and summary statistics of home values for each region (1=Northeast, 2=Midwest, 3=South, 4=West)

The above two figures 4 & 5 represent examples of categorical variables metro and regions. The box plot shows the home values for these metros and regions. As can be seen from the data, the Northeast and West regions have higher home prices. Central city has higher home prices along with suburban zone-2. Suburban zones 3-5 has lower home prices.

Correlation matrix to check for relationships between features

The correlation matrix to check dependency between variables is done to see the effect of some of the variables on the dependent variable. Many features are eliminated during the initial data cleaning methods. The remaining variables are checked to see dependencies between independent variables and their effect on the dependent variables. The following correlation matrix summarizes the results:

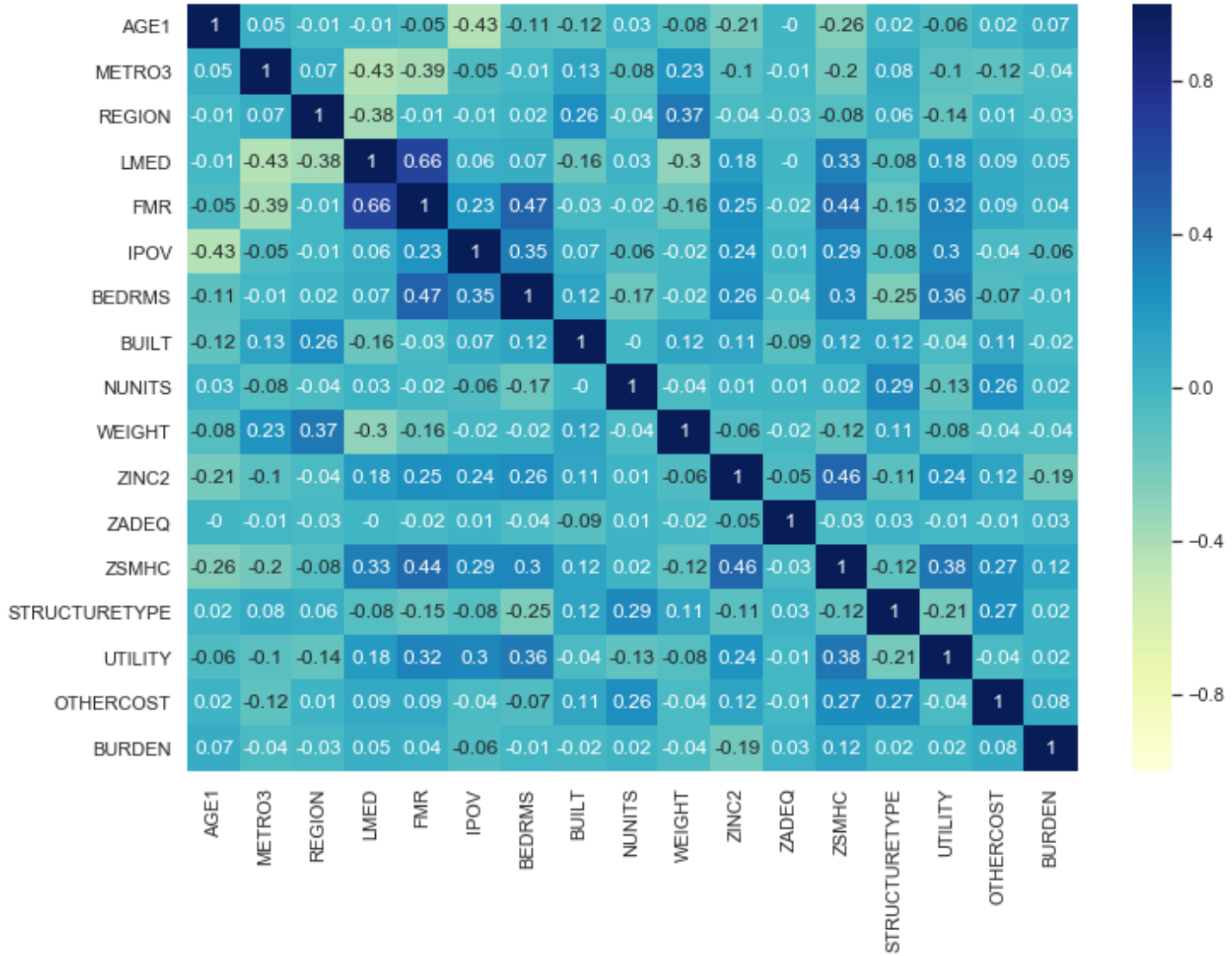


Figure-6: Correlation matrix of independent variables

Strong correlations between pairs of independent variables

Some of the independent variables have strong correlation between them. For example, Rooms and Bedrooms has strong correlation. There is also a strong correlation between many Cost features and they have all been dealt with taking only one significant variable. Figure-6 is a collection of independent variables that do not show strong correlation between them.

Significant variables:

Using features significance output of the Random Forest Regression, the top ten features affecting the home value feature is obtained. The following plot is the summary of the ten features and their significance.

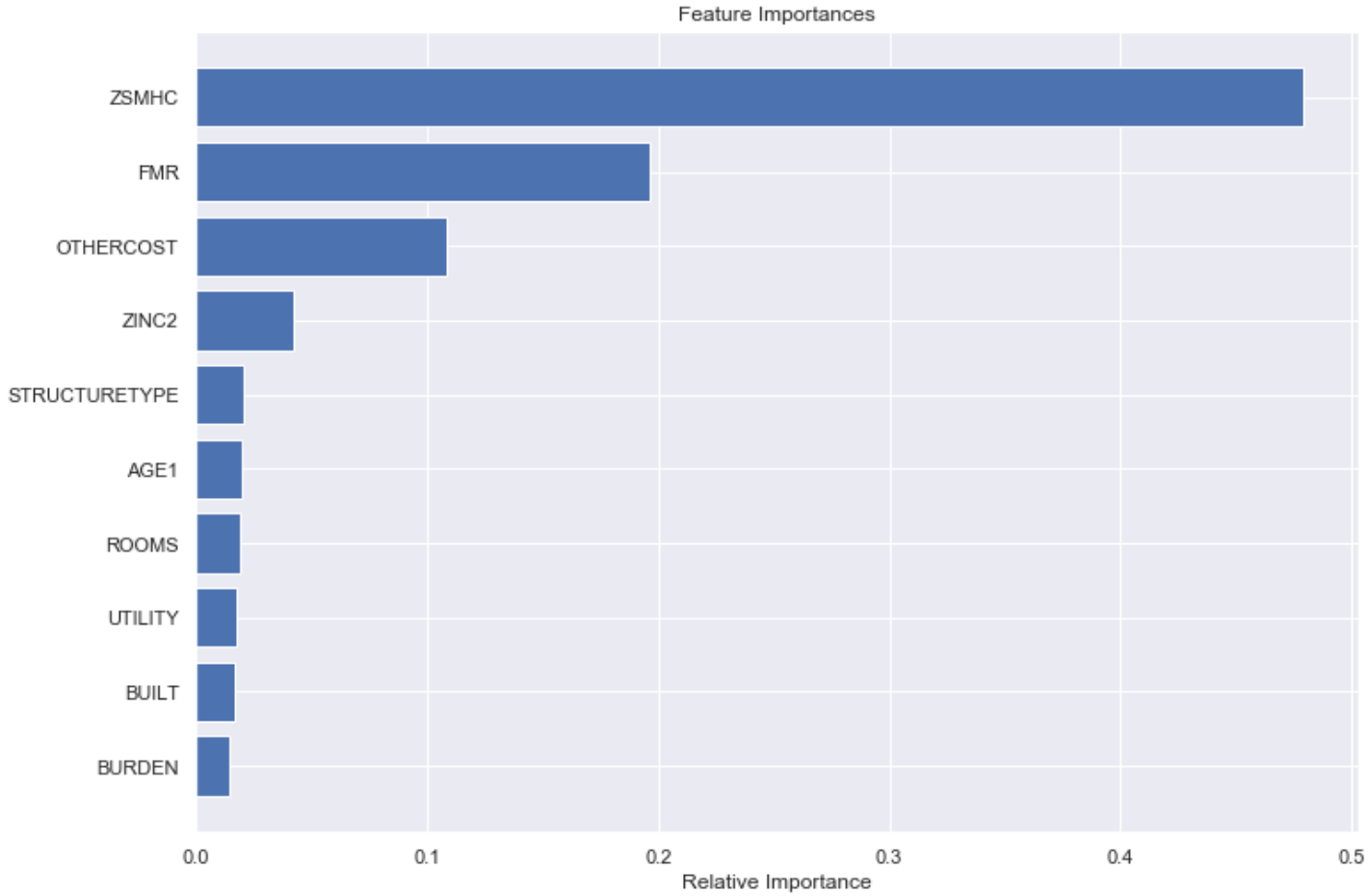


Figure-7: Top ten significant variables affecting the results

ZSMHC is the monthly housing costs and is the most significant, followed by Fair market rent. Other monthly costs and household income (ZINC2) are also significant in predicting house affordability along with Age of head of household, Rooms, Utility costs, year built, and burden.

Scatter plots

Scatter plots are useful in checking two features. Here we check scatter plots for linearity.

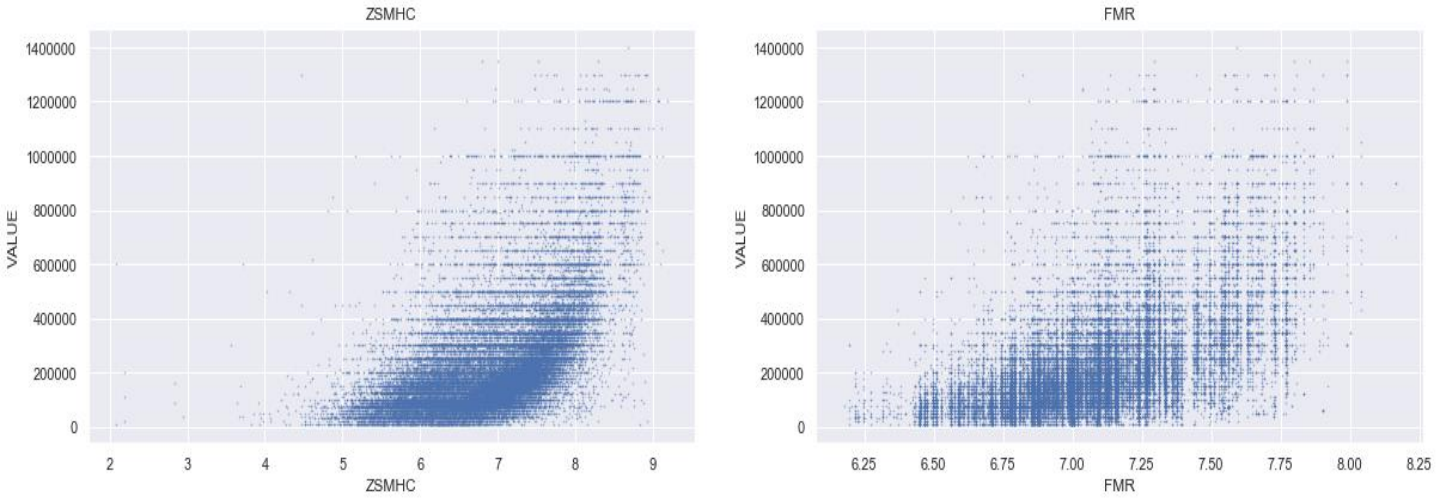


Figure-8: Scatter plots of Value v. housing costs (ZSMHC) & Value v. Fair market rent (FMR)

The above scatter plots show the relationship is not linear. In order have a linear relationship, square root is taken of the house values and log scale for the housing costs and fair market rent as seen below. This is helpful should we choose a linear regression model to predict affordability.

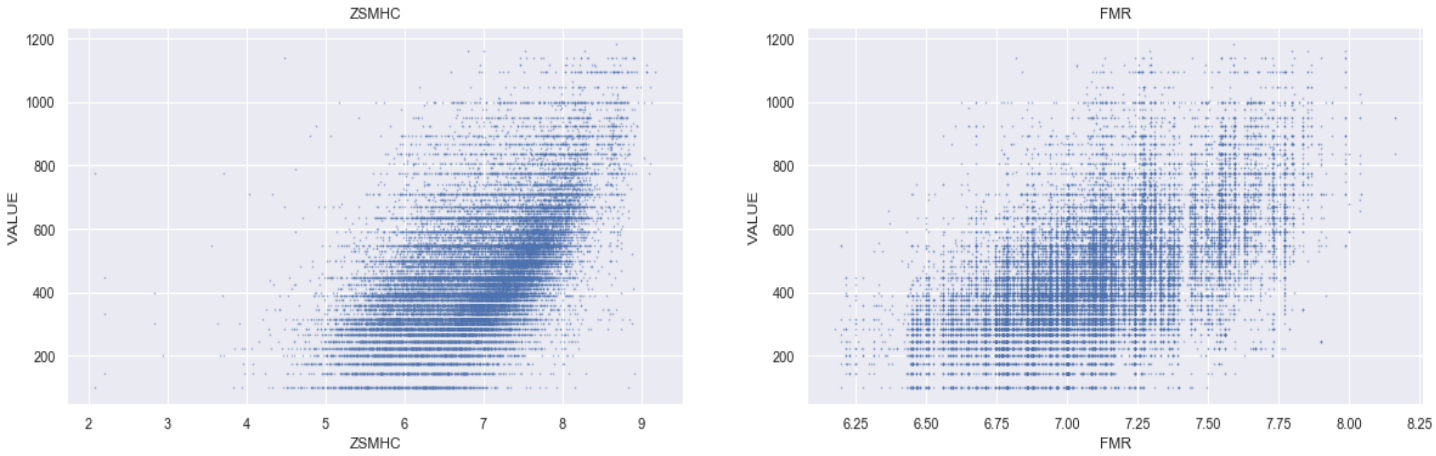


Figure-9: Similar to Fig-8 with square root of house values and log scale for costs and fair market rent

Machine learning models

The target variable we are trying to predict is the house values based on information of the consumers and is a continuous variable. We will use regression to solve this problem.

Three different machine learning models were used to provide the solution. Linear regression, KNN method and Random Forest Regression were used. We can try many different methods but these models are widely used for similar problems at hand.

The following Figure-10 summarizes the training and test accuracies from the models:

model	Training Accuracy	Test set accuracy	Delta RSME (Test-Train) *
Linear Regression (8 features)	0.47	0.45	2566
Linear Regression (4 features)	0.45	0.45	117
K-nearest neighbor (KNN)	0.54	0.36	16824
KNN hyper tuned with Random Search CV	0.50	0.38	11021
Random Forest	0.67	0.54	14809

Higher delta RSME and delta R2 score between test and train data is an indication of model overfitting. Ideally, we should see minimal overfitting to have good prediction of new data by the model.

Linear regression has the lowest delta RSME compared to the other two models, but the R2 score is lower. Linear regression with just the top 4 features were tried. The model accuracy did not change, but the delta RSME reduced significantly.

K nearest neighbor model (KNN) is decent model with better accuracy than linear regression but with high delta RSME, suggesting overfitting. Even with fine tuning the model with Random Search CV, the results do not improve much.

Finally, a Random Forest Regressor model gives better accuracy and but the RSME is not great. But with fine tuning the model with hyper parameter optimization, we can improve this model with good accuracy and RSME.

Random Forest model optimization using cross validation:

Randomized Search CV:

Randomized search CV is method where we tune the hyper parameters using a combination of different parameters which are picked using a random function. This method of hyper tuning is faster and gives a good improvement of the model over the default parameter model. After running the randomized search CV, the following parameters were tuned:

- Max_depth:9
- Max_features: 4
- Minimum_sample_leaf: 6
- N_estimators:100

Grid Search CV:

In grid search CV, the hyper parameters are tuned such that each combination of parameters are tried out and the best parameters chosen. In a random forest model the most important parameters are the tree depth (max_depth) and the number of trees (n_estimators) used in the model. Minimum samples in a leaf and max_features can also be optimized, but their effect on the model is minimal. Also, using default values for max_features and samples reduces the number of iterations and time in training the model.

The Figure-11 below shows the plot of accuracy score with max depth and n_estimators as parameters. From the plot we can see that we get good results with 50 trees (n_estimators) and a depth of 6-7 layers. After depth of 7, the model tends to over fit.

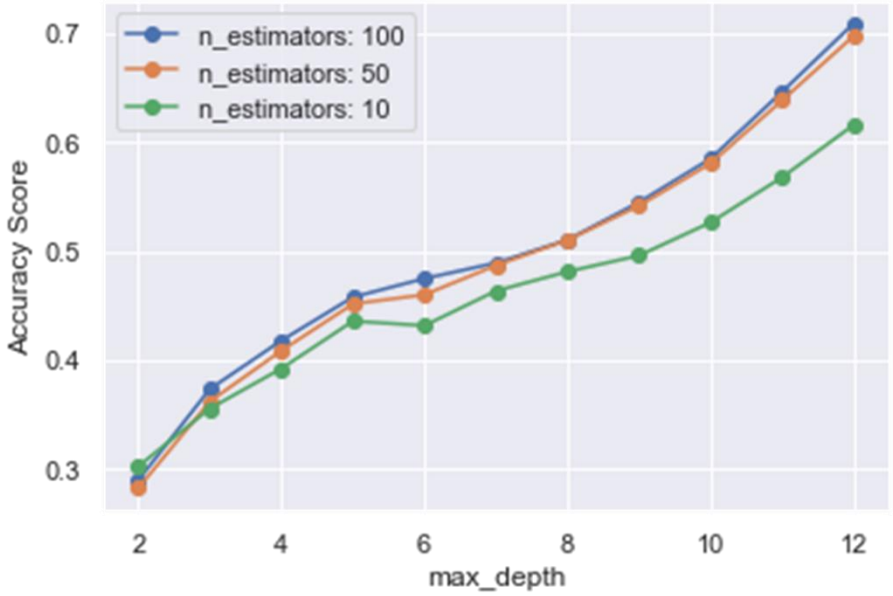


Figure-11: grid search CV results.

Random Forest model comparison:

model	Training Accuracy	Test set accuracy	Delta R2 (Test-Train)	Delta RSME (Test-Train)
Random Forest Default	0.67	0.54	0.13	14809
Random Forest Randomized CV	0.62	0.54	0.08	9103
Random Forest Grid Search CV	0.58	0.53	0.05	5954

Figure-12: Random forest models R2 and RSME comparison

Conclusion:

1. The dataset has 100 features. Using dimension reduction, only 10 most important features are selected for decision making.
2. ML algorithms considered: Linear Regression, KNN, and Random Forest.
3. Model evaluation is based on both R2 score and RSME.
4. Based on the accuracy scores obtained, Random Forest model is chosen to train the data.
5. Hyper parameter tuning is done to minimize overfitting using Randomized Search CV and Grid Search CV.

6. Random Forest Model with hyper parameter tuned model with grid search CV is chosen as the final model based on model performance metrics R2 and RSME.