Assignment 4: Random Forests by Mimi Trinh

Section 1: Summary and Problem Definition

The Boston housing study is a market response study including 506 census tracts in the Boston metropolitan area. The objective of the study is to advise a real estate brokerage firm in its attempt to employ machine learning methods to complement conventional methods for assessing the market value of residential real estate. The response variable is the median value of homes in thousands of 1970 dollars. The remaining variables in the dataset are predictors.

Section 2: Research Design, Measurement, and Statistical Methods

The study starts with 14 columns in the dataset, including the response variable and 13 predictors. However, the neighborhood column is dropped, so the dataset is narrowed down to 13 variables. There's no missing value, so the dataset is clean to be analyzed. First, standard scaler is implemented since it's best practice to standardize the variables before analysis. Second, an exploratory data analysis (EDA) is conducted to examine the response variable and correlation between the response variable and the predictors. Third, we build five models using linear regression, ridge regression, and three random forests with 7, 3, and 10 predictors included in individual trees. For each model, within a ten-fold cross validation design, we use the root mean squared error (RMSE) to evaluate the methods. In other words, the mean of 10 RMSE scores is an index of prediction error of each model.

Section 3: Programming Work

Multiple Python packages are utilized to do the programming: numpy, pandas, Scikit-Learn, and matplotlib. We start the project by feeding the csv raw data file into Python. Then we drop the neighborhood column from the dataset and start the standard scaler transformation process. Matplotlib is utilized to conduct the EDA to understand the data and correlation among

variables. Next we use LinearRegression(), Ridge(), and RandomForestRegressor() to build five models. We create a for loop to design a ten-fold cross validation for each method. We build all five models and let the RMSE metric determine the best model.

Section 4: Results and Recommendations

Exhibit 1 shows that the dataset has no null record, so we don't have to address missing value issue. Exhibit 2 gives the descriptive statistics of the dataset as part of the EDA result. Exhibit 3 shows that the response variable has outliers, but there's no extreme outlier since there's no observation outside the +/- 3 standard deviation range. The variable is skewed positive, but since there's no extreme outlier, we recommend not to remove any outlier and continue with the study. Exhibit 4 shows the correlation between response variable and each predictor. Rooms has the highest positive correlation, and lstat has the highest negative correlation with the dependent variable. In other words, the higher the number of rooms and the lower the percentage of lower socio-economic population, the higher the home value. This concludes the EDA part.

Random forest has the smallest RMSE, so we recommend using this approach. Among three random forest models, the method with 7 predictors has the lowest RMSE. Thus, random forest model with 7 predictors is the best model we recommend to management. Since this is the best method among all five models, we apply it on the full dataset and obtain results regarding the impact of explanatory variables. Specifically, exhibit 5 shows that the percentage of population of lower socio-economic status is the most important predictor in forecasting home prices. Then the average number of rooms per home is the second most important predictor, followed by crime rate and weighted distance to employment centers. The results match our findings in EDA regarding the number of rooms and percentage of lower socio-economic status.

Appendix

Exhibit 1

```
General description of the boston_input DataFrame:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
neighborhood    506 non-null object
crim            506 non-null float64
zn              506 non-null float64
indus           506 non-null float64
chas            506 non-null int64
nox             506 non-null float64
rooms           506 non-null float64
age             506 non-null float64
dis             506 non-null float64
rad             506 non-null int64
tax             506 non-null int64
ptratio         506 non-null float64
lstat           506 non-null float64
mv              506 non-null float64
dtypes: float64(10), int64(3), object(1)
memory usage: 55.4+ KB
None
```

Exhibit 2

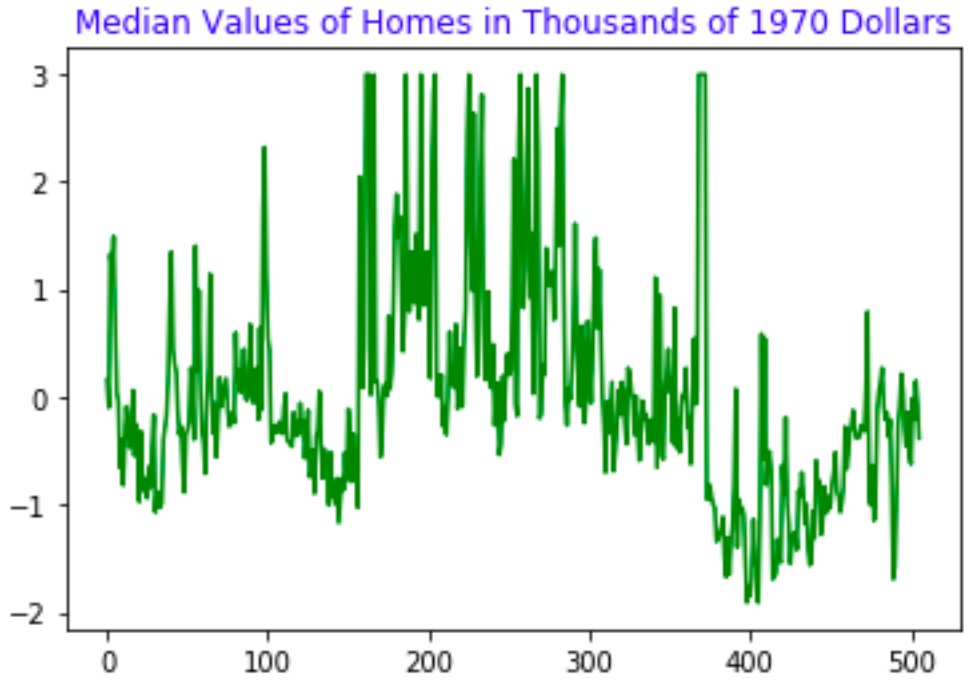
Descriptive statistics of the boston DataFrame:

	crim	zn	indus	chas	nox	rooms	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

	age	dis	rad	tax	ptratio	lstat	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063	
std	28.148861	2.105710	8.707259	168.537116	2.164946	7.141062	
min	2.900000	1.129600	1.000000	187.000000	12.600000	1.730000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	6.950000	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	11.360000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	16.955000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	37.970000	

	mv
count	506.000000
mean	22.528854
std	9.182176
min	5.000000
25%	17.025000
50%	21.200000
75%	25.000000
max	50.000000

Exhibit 3


Median Values of Homes in Thousands of 1970 Dollars
3
2
1
0
−1
−2
0
100
200
300
400
500

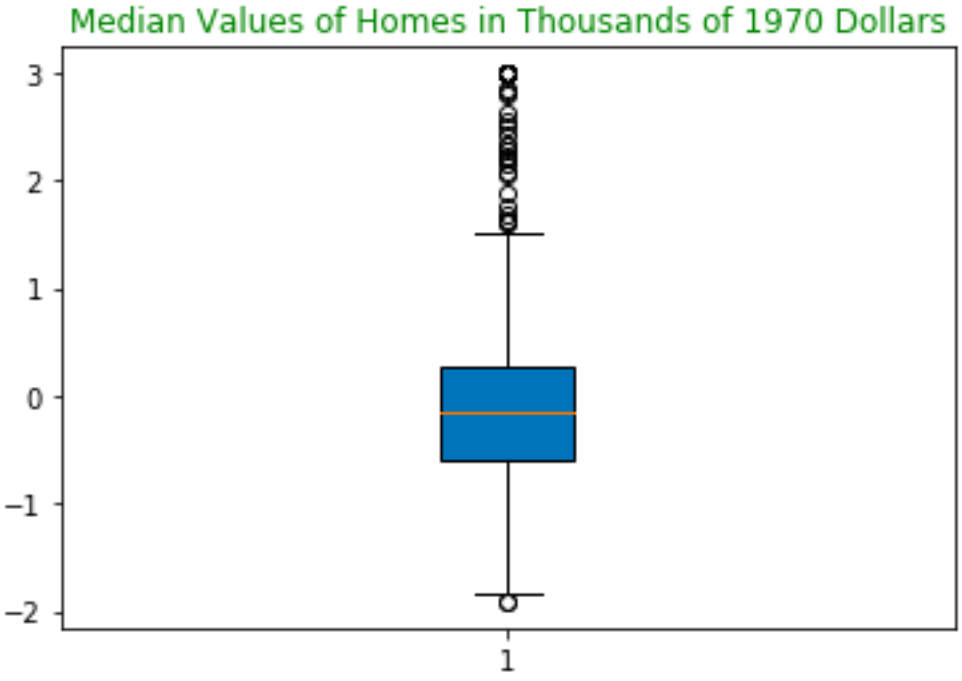
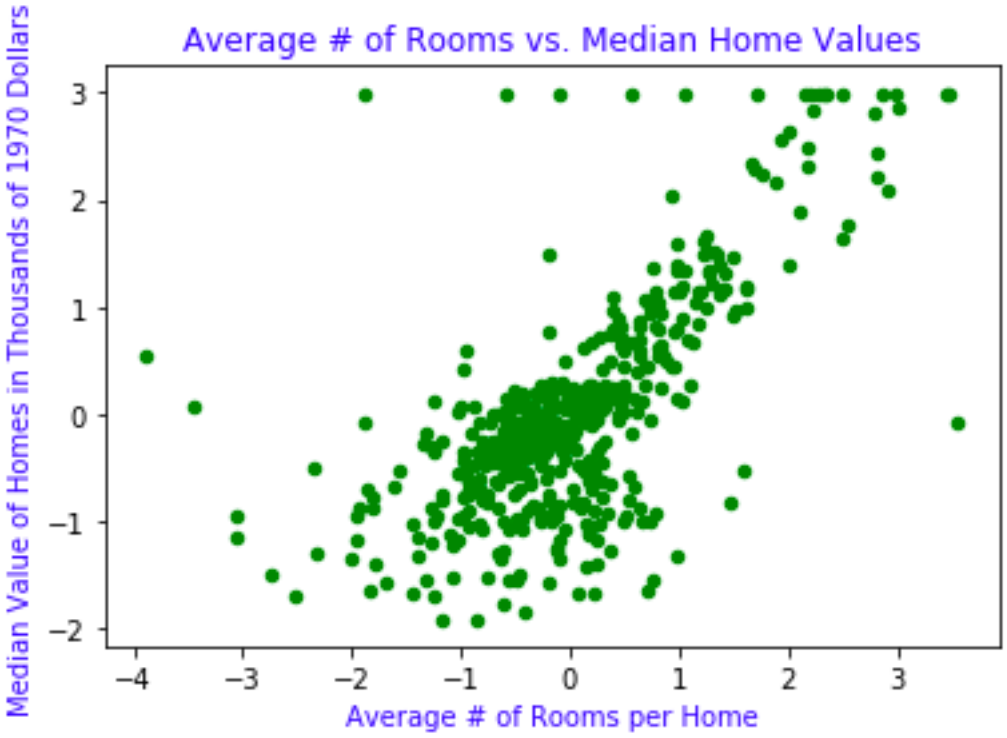

Median Values of Homes in Thousands of 1970 Dollars
3
2
1
0
−1
−2
1

Exhibit 4

```
mv          1.000000
rooms       0.696304
zn          0.360386
dis         0.249315
chas        0.175663
age        -0.377999
rad        -0.384766
crim       -0.389582
nox        -0.429300
tax        -0.471979
indus      -0.484754
ptratio    -0.505655
lstat      -0.740836
Name: mv, dtype: float64
```



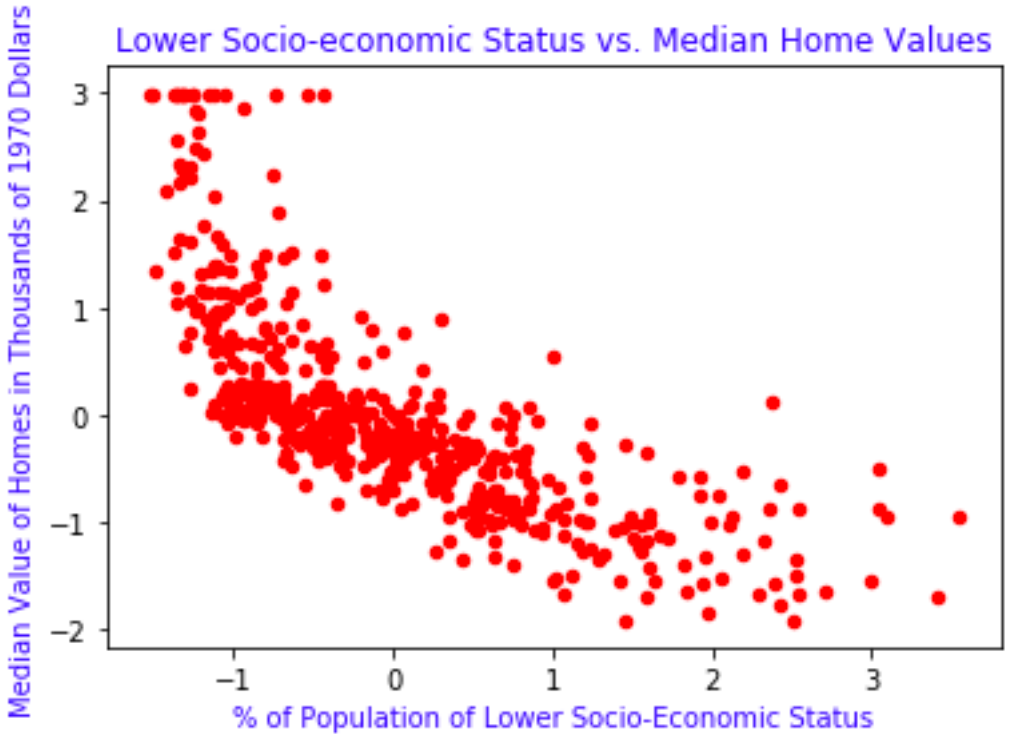


Exhibit 5

Random Forests Regression Model Explanatory Variable Importance Results

	name	description	importance
11	lstat	Percentage of population of lower socio-econom...	0.376417
5	rooms	Average number of rooms per home	0.365950
0	crim	Crime rate	0.050868
7	dis	Weighted distance to employment centers	0.050846
4	nox	Air pollution (nitrogen oxide concentration)	0.047793
10	ptratio	Pupil/teacher ratio in public schools	0.034480
2	indus	Percentage of business that is industrial or n...	0.022915
6	age	Percentage of homes built before 1940	0.022505
9	tax	Tax rate	0.018949
8	rad	Accessibility to radial highways	0.005754
3	chas	On the Charles River (1) or not (0)	0.001922
1	zn	Percentage of land zoned for lots	0.001602