Math-678 Statistical methods for data science "Project report"

- Tanmay Gupta (tg289)

Description:

Housing Prices Prediction using the Boston dataset, the purpose of this project will be to determine the prices of homes in the Boston areas using variables such as Crime rate per town, proportion of non-retail businesses acres, nitric oxide concentration, average amount of rooms and a variety of other variables in the data set. The object is to train a portion of the data to be able to predict the price of the test data set.

Background:

The project is based on the Boston dataset. The data to be analyzed were collected by Harrison and Rubinfeld in 1978 for the purpose of discovering whether or not clean air influenced the value of houses in Boston. Their results are documented in a paper titled Hedonic prices and the demand for clean air, published in J. Environ. Economics and Management 5, 81-102.
In this report, we will examine the several neighborhood attributes on the prices of housing, in an attempt to discover the most suitable explanatory variables. The specific neighborhood attributes to be considered are 'proximity to the Charles River, distance to the main employment centers, pupil-teacher ratio in schools, and levels of crime'. Whereas the original study focused on air pollution using nitrogen oxide concentrations as an explanatory variable, this report examines whether or not there are other, better explanatory variables for the median value of houses in Boston.

Research Question:

Based on the crime, age of a home, access to highways, and the 11 other attributes how well can our logistics regression model, SVM, and K-means predict the attributes and value of the home.

Potential Solution:

A potential solution is to use our static models to validate whether there is a correlation in the price of a house and the features of its city/town. Are residents willing to pay more for a home with better levels of oxygen (Nox variable) or for areas where the teacher pupil ratio is low. Using KNN, linear regression and SVM we'll dive into the data for attributes that are highly correlated to determine how well we can predict the attributes of a Boston home based on the housing price.

Method and procedure

The Boston Dataset contains the collected U.S Census Service concerning housing in the area of Boston Massachusetts. The data features 506 rows of housing information and features that include the crime rate, nitric oxides, avg amount of bedrooms, an index of accessibility to radial highways, and much more. The data set used is split into a training data set consisting of 404 homes, and a test set of 102 homes. The linear SVM model was built with a new differentiating column based on the mean price of homes for classification. Whiles the Radial and KNN predictive meansures….. All significant data attributes have been identified at the end of their name. W

Boston Housing Characteristics

Characteristics

- Age – the proportion of owner-occupied units built prior to 1940
- RM- the average number of homes*
- TAX – full value property tax rate per $10,000*

City Characteristics

- ZN – Proportion of residential land zoned for lost over 25,000 sqft *
- RAD – index of accessibility to radial highways*
- CRIM – per capita crime rate by town *
- DIS -weighted distances to five Boston employment centers*
- INDUS – the proportion of non-retail business acres per town
- PTRATIO – pupil-teacher ratio by town*

Demographic Characteristics

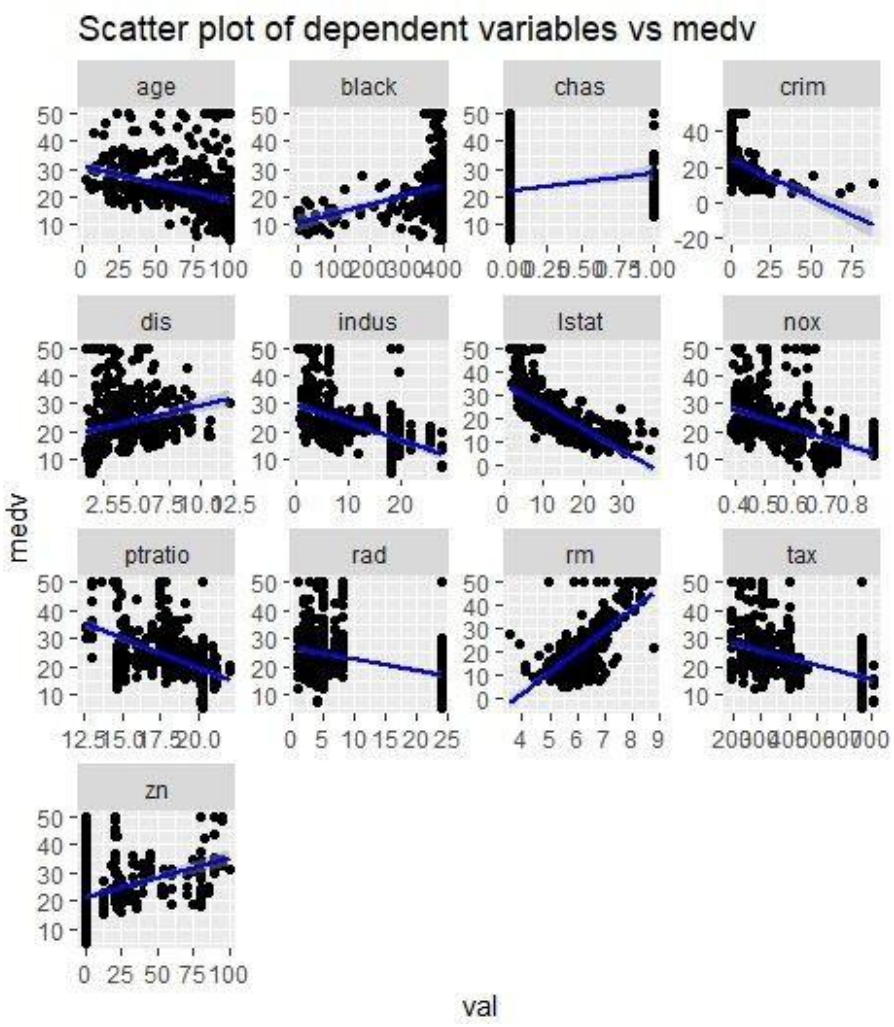
- B – 1000(Bk – 0.63)^2 where Bk is the proportion of blacks by town*
- LSTAT - % lower status of the population*
- CHAS – Charles river dummy variable*
- NOX – nitric oxides concentration (parts per million) *

Predictor characteristics

- Medv – Median value of owner-occupied homes in $10,000

Multiple predictive measures were applied to the above data variables In our first model we used Linear regression with the lm() function on the training set. The optimal variables were selected using stepwise selection of forward and backward selection The SVM radial and KNN models were evaluated classified the data using linear SVM,

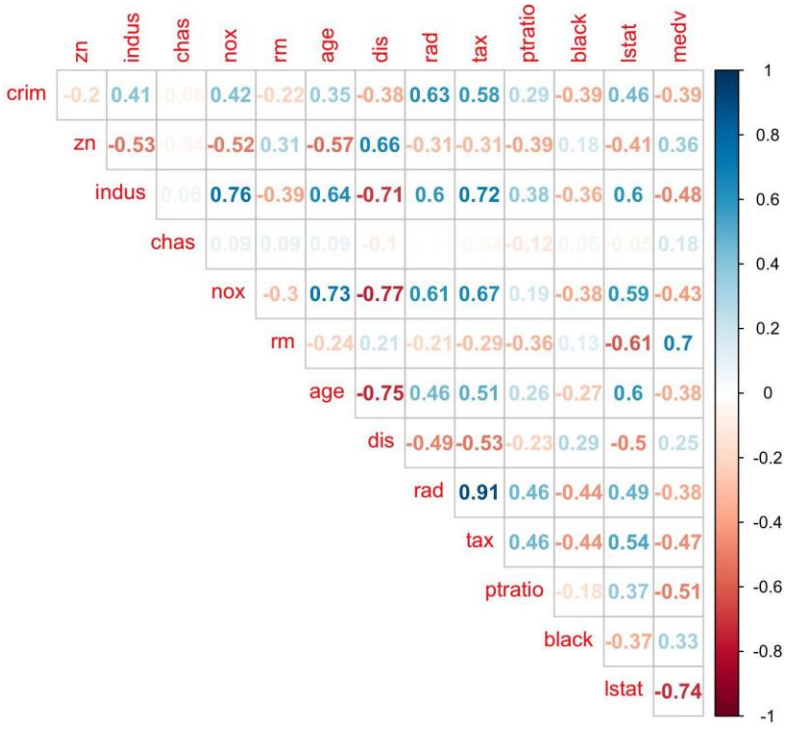
Understanding the data



About the plot

The scatterplot shows the relationship of all the variables with our predictor variable i.e median value of household (medv). The blue line shows the linear fit or relation between the variables. And the direction shows if there is a positive or negative relationship between the variables.
For example- “crim, nox and tax” show a negative relationship as the line is going ‘down from left to right’.
The variables “dis, rm and black” show a positive relationship as the line goes ‘up from left to right’.
For variables “chas, rad and zn” we are not sure about their relationship with medv as the ‘curve does not show any significant direction’.

About the plot “Correlation matrix”



The correlation matrix shows the correlation of every variable with a corresponding variable.. The last column shows the correlation of every variable with ‘medv’.The values are between -1 and 1 values greater than .5 generally show a strong relationship such as “indus” and tax” If the value is negative then the variables are negatively correlated and if positive then they are positively correlated.

- **Median value of owner-occupied homes (in 1000$) increases as average number of rooms per dwelling increases and it decreases if percent of lower status population in the area increases**
- **nox or nitrogen oxides concentration (ppm) increases with increase in proportion of non-retail business acres per town and proportion of owner-occupied units built prior to 1940.**
- **rad and tax have a strong positive correlation of 0.91 which implies that as accessibility of radial highways increases, the full value property-tax rate per $10,000 also increases.**
- **crim is strongly associated with variables rad and tax which implies as accessibility to radial highways increases, per capita crime rate increases.**
- **indus has strong positive correlation with nox, which supports the notion that nitrogen oxides concentration is high in industrial areas**

- ## Analysis and model fitting

We fit various models on our training set and used it to predict the median values of houses on our test set and to gain the importance of variables. Now we discuss the models.

1. Linear regression with subset selection

We performed the linear regression with the lm() function on the training set. Then we used the regsubsets() function to find the best subset on the 13 explanatory variables. To select the best variables, we went with the stepwise selection using “forward selection”, “backward selection” and a combination of both.

With the “lm” and “step” function with stepwise variable selection, a model was obtained with minimum AIC. The table below describes the characteristics of the model.

Variable	Estimate	Standard error	t-value
Lower status population	-0.650201	0.051231	-12.691
Average no. of homes	3.458593	0.458938	7.536
Pupil-teacher ratio	-0.960444	0.133653	7.86
Distance to employment center	-1.444988	0.193780	7.457
Charles river	3.624134	0.941359	-3.850
Population of blacks in town	0.008730	0.002709	3.222
Land zoned over 25000 sq. ft.	0.048326	0.013792	3.504
Nitric oxide concentration	-12.432692	3.662008	-3.395
Per capita crime rate	-0.127936	0.031170	-4.104
Accessibility to radial highway	0.268332	0.062466	4.296
Property tax rate	-0.008222	0.003342	-2.460

After the step selection model was evaluated, we identified that the variables “Indus” and “age” are statistically insignificant because their P-values are not below 0.05 and t-values are not <-2 or >2 and also in our subset selection these variables are left out. So now we make a new linear regression model (without the variables Indus and age) and predict the values on our test set.

To check how the model performed we used the Mean Squared Error (MSE) and Correlation accuracy. We made an actual to predicted table to see how good our predictions are as compared to the real data.

The results are shown here -

Mean Squared Error	Root mean squared error
23.44	4.84

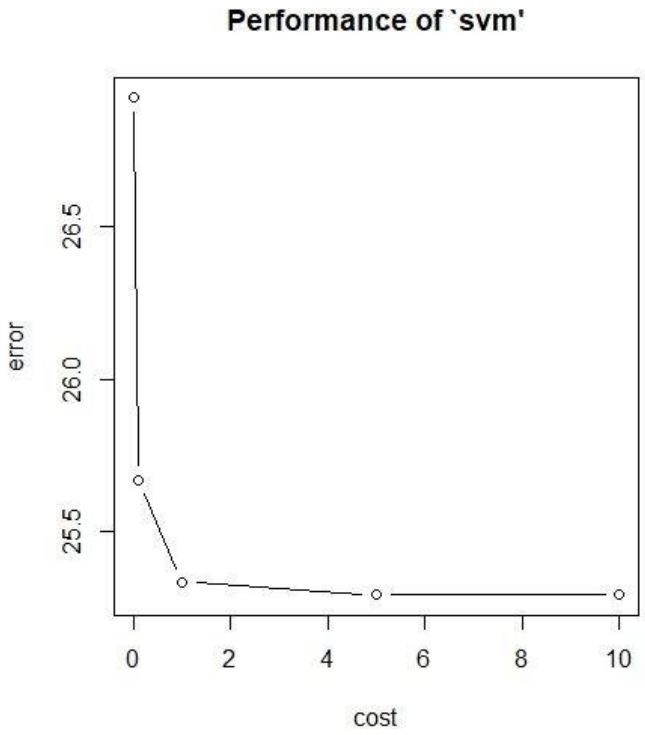
	Actual	Predictors
Actual	**1.00**	**.80**
Predictors	**.80**	**1.00**

The lower the MSE the better the model and **the higher the accuracy the better the model.** Our R^2 was around **74% for our linear regression model.**

Linear Support Vector machine

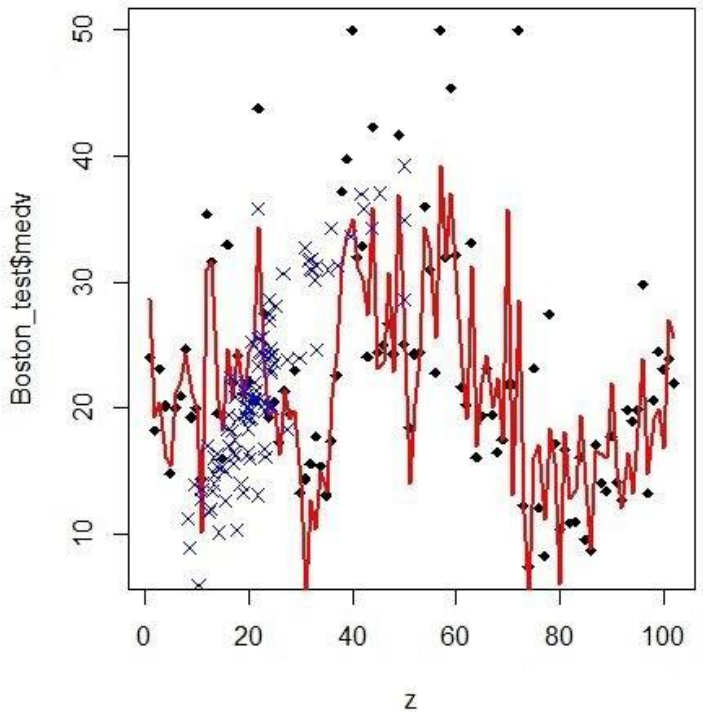
We fit a linear SVM model to our training data with the defined parameters (C=1, sigma(gamma) = 0.07 and epsilon= 0.1). Now we tune our model to get the best parameters, we used the “tune” function from the “e1071” library (library for SVM in R). After tuning the model we got the best parameters which were “C= 10”.

The plot below shows the performance of ‘svm’ with ‘error’ on Y axis and the value of Cost(c) on X axis. The error is minimum at C= 10



Now we fit another model with the tuned parameters and used this new model to predict the values of medv on our test set. The results are somewhat better than the linear regression model because the SVM fits a margin around the hyperplane and then predicts the values.

The plot below shows our SVM model. The red line shows the variation of our predicted values to the actual values in our data.



The statistics and results are shown below-

Mean Squared Error	Root mean squared error
23	4.80

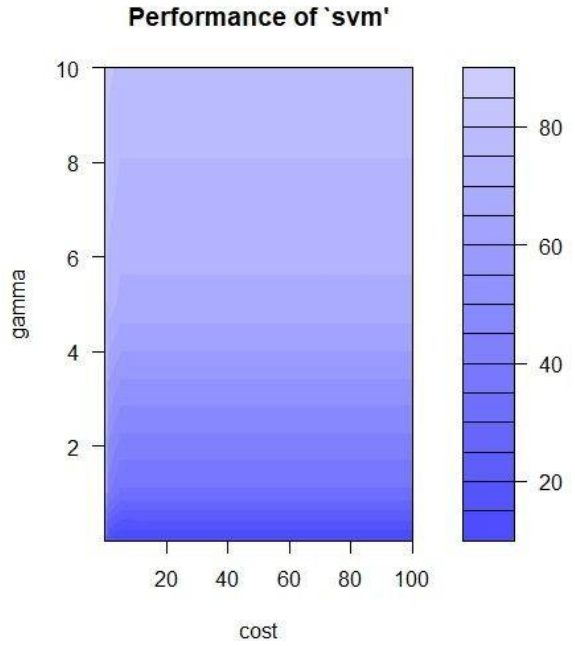
	Actual	Predictors
Actual	**1.00**	**.863**
Predictors	**.863**	**1.00**

The correlation accuracy increased from 80 % in linear regression to 86% in linear SVM.

Radial Support Vector Machine

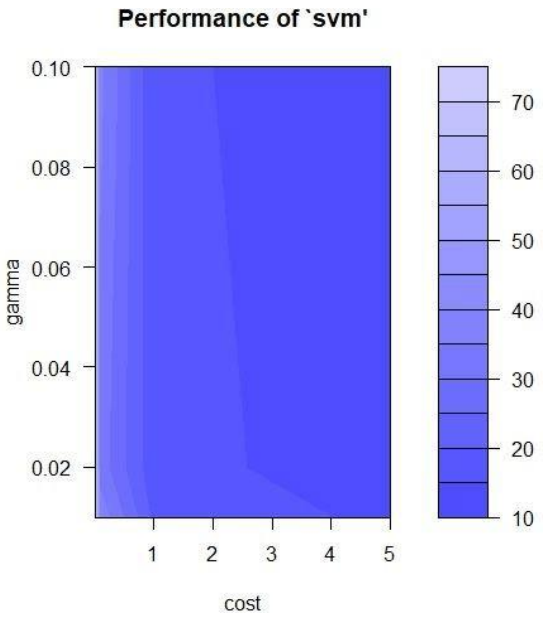
Now we fit a radial SVM model to our training data with the defined parameters (C=1, sigma(gamma) = 0.07 and epsilon= 0.1). Now we tune our model to get the best parameters, we use the “tune” function from the “e1071” library . After tuning the model we got the best parameters which were “C= 10” and “gamma= 0.1” with the best performance of error being 11.3672. Now we make a plot of our tuned svm to check the performance.

The plot below shows cost and gamma on X and Y axes (it shows the RMSE in the dark part) , so the darker the region is, the better the model. This model is poor and has a higher RMSE, so we tune the model again with narrower values of the parameters.

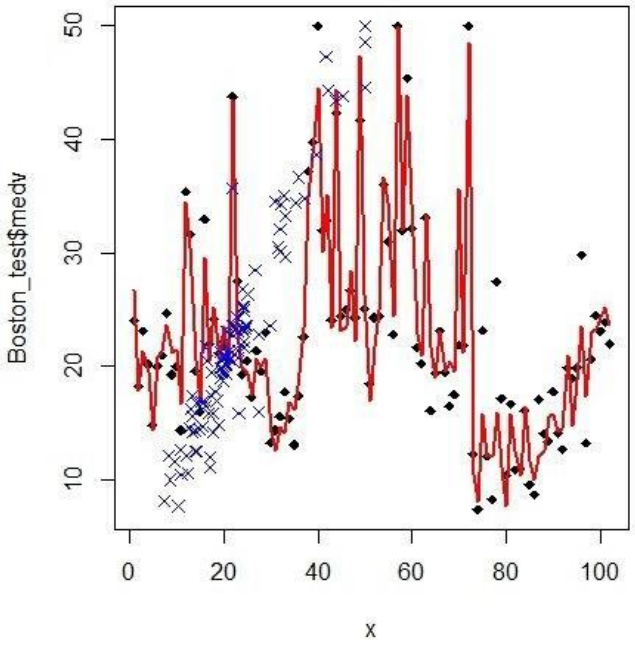


Initially while tuning we took the **range of parameters** to be between 0.1 and 100, now we take the range of cost between 0.1 and 5 and for gamma 0.01,0.1,0.2 and tune our model again.

The plot below cost and gamma on X and Y axes. This model is better than the last one because of the amount of darker region in the plot. It shows that the RMSE is less as compared to the last one. These values are considerable and we can fit our model by taking these parameter values.



We fit a radial SVM model with the tuned parameters and use it to predict the test values on our test set. It performs very well as compared to Linear regression and Linear SVM which shows that our data performs better with non-linear technique. This could happen because of the non-linear relationship of some variables with our predictor variables. We predicted our values and the plot shows a better result as our model covers almost all the values with the line.



The statistics and results of our Radial SVM model are shown below

Mean Squared Error	Root mean squared error
8.4	2.9

	Actual	Predictors
Actual	**1.00**	**.9256**
Predictors	**.9256**	**1.00**

The model performed really well with our new tuned parameters.

K-Nearest Neighbors (KNN)

We fit a K-nearest neighbor regression on our data set by selecting the best value of k by using cross validation . The best value for k was ‘5’ , so we predicted by choosing 5 nearest points are giving the prediction on that basis. We used the ‘train’ function in “caret” library to train our model by selecting the best value of K over many values. (RMSE was used to select the best K)

The statistics and results are shown below

Mean Squared Error	Root mean squared error
19.7	4.4

	Actual	Predictors
Actual	**1.00**	**.88**
Predictors	**.88**	**1.00**

The MSE again raised to 19.7 for our KNN model

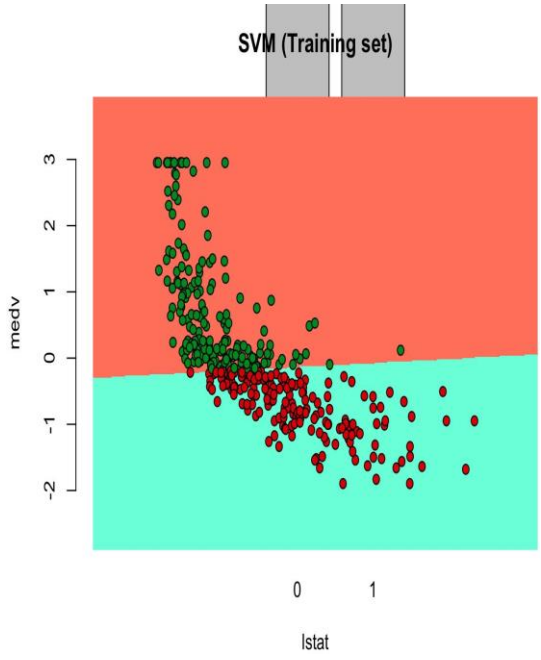
Now we try model for “classification”

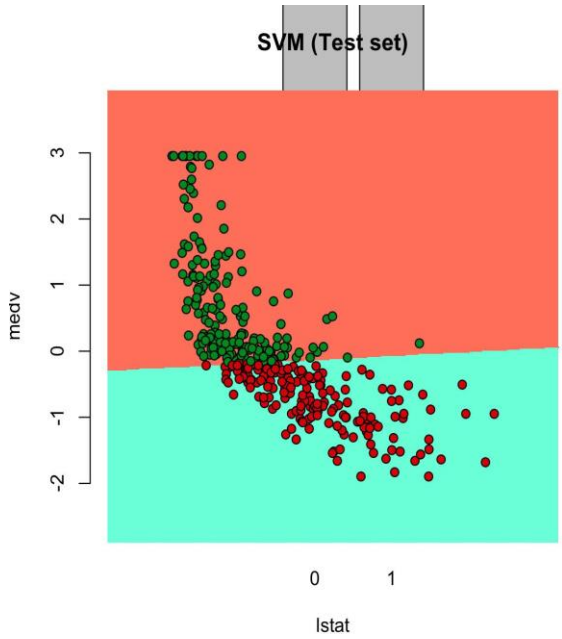
Linear SVM for classification:

The SVM algorithm works on finding a hyperplane which will maximize the distance between the hyperplane and the support vectors. The formulation of our hyperplane here was non-linear and our testing error is still lesser than training error with an underlying fit that appears to be non-linear

Now we fit a linear SVM for classification by taking a dummy variable for medv by taking medv < than the median as ‘0’ and medv>median as ‘1’. The dummy variable is added at the end as a 15th column. We use the variables to classify the value of medv as 0 or 1.

The plot below shows the classification on the basis of the value of lstat. The red dots are above the median value and the green dots are above the median value.





The confusion matrix is shown below-

	0	1
0	50	1
1	0	53

The results are not so good and classification does not fulfill our goal as the medv variable is itself a mean value and taking its mean and classifying the values is not a wise option.

Model Comparison

As we found our results in the form of MSE and correlation accuracy so we will compare them on the basis of that only.

Model	MSE	RMSE	Correlation accuracy
Linear Regression	**23.44**	**4.84**	**80%**
LInear SVM	**23**	**4.8**	**86.3%**
Radial SVM	**8.4**	**2.9**	**92.56%**
KNN	**19.7**	**4.4**	**88%**

Almost all the models performed adequately well on our dataset. The radial SVM model performed best in terms of MSE and accuracy . This tells us about the nature of the dataset and of the model , how well the model handels the dataset and gives the prediction results.

Limitations:

The data set could have used more varieties in its variables to tell a better story or provide more detail about the boston area. Information relating other demographics would be helpful in determining the buying patterns in the boston area. The data set had a bias by only including the proportion of blacks by town. Also the data set did not include information describing the demographics which may have been helpful in predicting the likelihood of purchased homes.

Conclusion:

The goal of this report was to determine the neighborhood attributes that best explained variation in house pricing. Various statistical techniques were used to eliminate predictors and extraneous observations. In examining the final model, one finds – quite reasonably – that house prices are higher in areas with lower crime and lower pupil-teacher ratios. House prices also tend to be higher closer to the Charles River, and houses with more rooms are pricier. This report is interested in the neighborhood attributes of houses, so the number of rooms is not an important predictor. The most interesting factors to consider are nitrogen oxide levels and distance to the main employment centers. On the one hand, people would want to live close to their place of employment. Yet it is reasonable to suggest that pollution levels are higher as one moves closer to these main employment centers. Most importantly, when talking of pollution, it is not just nitrogen oxide levels that are higher, but also noise pollution levels. The regression model that was fitted shows that higher levels of pollution decrease house prices to a greater extent than distance to employment centers. This suggests that people would prefer to live further away from their place of employment if it meant lower levels of pollution, which is an interesting point to consider. On a concluding note, it is important to note that the data for this report was collected several decades ago. In the years since, there is no doubt that pollution levels have risen and it would be interesting to examine the ways in which that affects house pricing in Boston today.