**George Kacoyanis**

**Purpose of Analysis:**

The purpose of the Analysis is to begin data exploration and analysis for what variables predict the Median Value of Owner-Occupied Homes (MEDV) in the Boston Area. Using different aspects of the Boston area with different measurements it is best to use Correlation Analysis, Principle Component Analysis, and Multiple Linear Regression, for different analysis methods to find accurate predictors for MEDV. A scatter plot matrix will be used to evaluate the different procedures

**Description of Dataset:**

There are 506 observations and 15 attributes provided in the data set.

**Response Variables:**

CRIM per capita crime rate by town

ZN proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS proportion of non-retail business acres per town.

CHAS Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX nitric oxides concentration (parts per 10 million)

RM average number of rooms per dwelling

AGE proportion of owner-occupied units built prior to 1940

DIS weighted distances to five Boston employment centers

RAD index of accessibility to radial highways

TAX full-value property-tax rate per $10,000

PTRATIO pupil-teacher ratio by town

B 1000(Bk - 0.63)^2 where Bk is the proportion of African-Americans by town

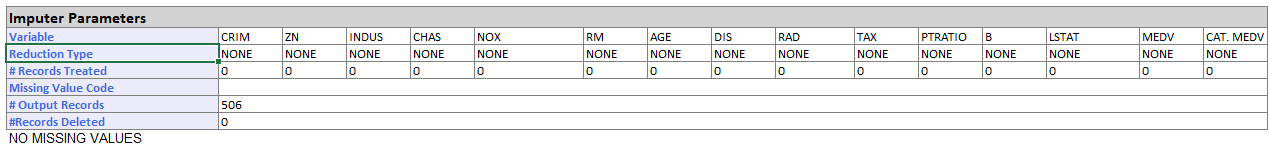
LSTAT % lower status of the population

**Indicator Variables:**

MEDV Median value of owner-occupied homes in $1000

CAT.MEDV [0 = MEDV<30($1000), 1 = MEDV >30($1000)]

The dataset comes from recordings in 1978 in the Boston area. The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.



From the Missing Data Handling there are no missing values in any of the variables.

A graph with a bar graph

Description automatically generated

Here it looks like there are a few of outliers so a Q-Statistics and Hotteling’s T-Squared Statistics in the PCA analysis might be required.

**Correlation Matrix Analysis:**

A graph with numbers and letters

Description automatically generated with medium confidence

Here is the correlation matrix between variables. For MEDV there looks to be a few variables with moderate to strong correlation. INDUS, NOX, RM, TAX, PTRATIO, and LSTAT. This calls for further analysis into which variables could have a high correlation value and those that could be excluded from the model.

**PCA Analysis:**

In the PCA analysis it looks as if 95% of the variance can be explained by the first 9 components. The data was Normalized/Standardized because many of the variables are measured differently (Dollars, Radius, CHAS is categorical, population, etc.).

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

For the 9 components it is visible that the dominant variables being used are CRIM, INDUS, CHAS, RM, AGE, PTRATIO, and B.

**Q-Statistics Test and Hotelling’s T-Squared Statistics Test:**

In the Q-Statistics Test and the Hotelling’s T-Squared Statistics Test, there were no values that were obscenely large except for one in the Hotelling’s T-Squared Statistics Test that has 153 at record 381. While the T-Squared Statistic was large, the actual record was not an outlier in any of the variables.

A screenshot of a computer

Description automatically generated

**Multiple Linear Regression:**

A white background with black text

Description automatically generatedThe R^2 and Adjusted R^2 show that the linear regression model predicted values fits 73-74% of the observed data.

A screenshot of a computer

Description automatically generated

This shows us the intercepts of the predictor variables. There are high absolute coefficients for CHAS, NOX, RM, DIS, and PTRATIO. There are 2 variables with very high p values showing that their relationship between the response can mostly be explained by random chance, that is INDUS and AGE. By removing those 2 there should be a better prediction with the regression model and can continue to further analysis on the fit of the variables.

A screenshot of a computer

Description automatically generated

The fit of the model has changed very little, almost unnoticeable from the removal of those variables.

A screenshot of a computer

Description automatically generated

The variables show that the predictors can explain the response very well, but some coefficients contribute very little to the Response variable; having very small coefficients. ZN, TAX, CRIM, RAD, B and LSTAT.

Removing those coefficients will see if there is a change in the model fit by trying to avoid overfitting.

A screenshot of a computer

Description automatically generated

Here R^2 and Adjusted R^2 model fitting has dropped 10%. Meaning there were too many variables. This could be explained by the fact that 2 of the variables’ coefficients values are relatively large in comparison to the other values: B, TAX. By adding these back in, there may be a better fit in the model.

A screenshot of a computer

Description automatically generated

The model fit didn’t change much, the R^2 and Adjusted R^2 grew 2%. Checking the coefficients and their p-values might give us greater insight on the model.

A screenshot of a computer

Description automatically generated

The TAX variable has a very high p-value showing that it has a very high amount of chance between its relationship with the response that we did not see before.

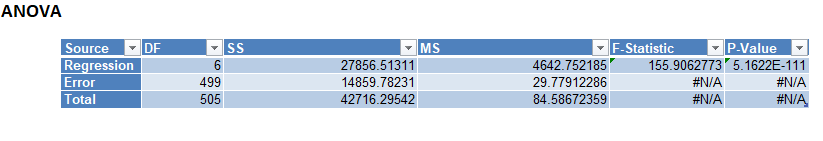
A screenshot of a computer

Description automatically generated

The model fit has barely changed, but there is a very low p-value with all predictors.

A screenshot of a computer

Description automatically generated



The P-value is very small still so the model can be greatly predicted with the chosen predictors, while avoiding overfitting and keeping the amount of predictor variables minimal. The predictors chosen are CHAS, NOX, RM, DIS, PTRATIO, and B.

From the three-dimensional reduction methods chosen we have 3 options.

**Correlation Analysis:** INDUS, NOX, RM, TAX, PTRATIO, and LSTAT

**PCA:** CRIM, INDUS, CHAS, RM, AGE, PTRATIO, and B.

**Linear regression:** CHAS, NOX, RM, DIS, PTRATIO, and B

There are 2 common variables throughout the 3 methods: PTRATIO and RM. Showing them to be significant predictors towards the response variable MEDV.

By comparing all the predictors in a scatter plot matrix, the scatter plots and histograms can be compared for relationship strength to confirm if the variables: CRIM, INDUS, NOX, RM, TAX, PTRATIO, LSTAT, AGE, B, DIS, CHAS. Excluding the variables RAD and ZN.

**Scatter Plot Matrix:**

INDUS: Shows a slight correlation but there are some outliers

A screenshot of a graph

Description automatically generated

CRIM: Shows a clear negative relationship between CRIM and CAT.MEDV/MEDV with all values of high CRIM in the 0 CAT.MEDV category. Although there is a very large amount of data with low CRIM compared to higher amounts.

A screenshot of a graph

Description automatically generated

NOX: The higher end of the NOX values indicate that there is no clear relationship between the CAT.MEDV category and MEDV value.

A screenshot of a graph

Description automatically generated

RM: There is a clear positive relationship between MEDV and RM where the higher amounts of RM correlate to higher values of MEDV.

A screenshot of a graph

Description automatically generated

TAX: There is a slight negative relationship between TAX and MEDV and CAT.MEDV categories. There is more instances of there being higher TAX values when CAT.MEDV = 0, which could show a negative relationship, with one outlier being CAT.MEDV=1 with high TAX values.

A screenshot of a graph

Description automatically generated

PTRATIO: There is no clear relationship between PTRATIO and MEDV and CAT.MEDV, showing it is actually not a clear predictor.

A screenshot of a graph

Description automatically generated

LSTAT:There is a very strong negative relationship between LSTAT and MEDV and CAT.MEDV

A screenshot of a graph

Description automatically generated

AGE: SHOWS there is no clear relationship between AGE and MEDV and CAT.MEDV.

A screenshot of a graph

Description automatically generated

B: There shows a postive relationship between B and MEDV and CAT.MEDV. Althought the distribution in B shows there might the relationship could be misleading by giving a higher average, because there is also a large amount that has lower MEDV values as well.

A screenshot of a graph

Description automatically generated

DIS: Shows there is not a clear relationship between DIS and MEDV and CAT.MEDV.

A screenshot of a graph

Description automatically generated

CHAS: Shows no clear relationship between CHAS and MEDV and CAT.MEDV. Although the distribution is heavily weighted for CHAS at 0 compared to those that are close to the Charles River.

A screenshot of a graph

Description automatically generated

From this we can conclude that Predictors we should use are: CRIM, RM, LSTAT, and B. Some other variables we could include are TAX and INDUS.

**What I have Learned:**

I have learned the process of PCA and the importance of data cleaning and exploration. I also learned the importance of Preprocessing data in the PCA analysis, since it is important to have the data standardized so that some variables don’t out weigh others based upon measurements and that the data being worked with is accurately represented. Using it with correlation analysis and multiple linear regression to find accurate predictors was very helpful for finding useful predictor variables. From the data we have learned that Important indicators for MEDV and CAT.MEDV are CRIM, RM, LSTAT, and B, through the different dimension reduction analysis procedures. Many of the different procedures had different outcomes and a couple similar outcomes when it comes to selecting predictor variables so comparing all the options provided ensured accurate results for which predictor variables were the best.

**Hands-On Exercise 2**

**George Kacoyanis**

**ISM 6136**

Multiple Linear Regression models are developed by using predictor variables to predict linear relationships with a response variable. This analysis will be using predictor variables chosen from the previous exercise using the Principle Component Analysis, the correlation matrix analysis, and the variable selection process using multiple linear regression to predict MEDV values. The variables chosen to be used in the feature selection process are CRIM, RM, LSTAT, B, TAX, and INDUS. These variables are to be used to find the best subset of variables to do our predictions using Multiple Linear Regression and Logistic Regression.

The data was partitioned into 60% training and 40% validation. The partitioned data was only done to variables being considered for the feature selection process and regression analysis as well as the response variables: MEDV and CAT.MEDV

A screenshot of a computer

Description automatically generated

**Training Score summary with lift charts, gain charts, decile charts, and RROC charts.**

A graph of a graph

Description automatically generated with medium confidence A graph of a graph with numbers and a red line

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generated A graph of a bar graph

Description automatically generated

A screenshot of a computer

Description automatically generated

The training data has an R^2 value of 0.598919 which is the same as the regression summary R^2. The RROC curve shows underestimation is occurring in the prediction with the training data.

**Validation Score Summary with lift charts, gain charts, decile charts and RROC charts.**

A graph of a graph

Description automatically generated with medium confidenceA graph with red and blue lines

Description automatically generated

A graph with numbers and lines

Description automatically generatedA graph of a bar chart

Description automatically generated

A screenshot of a graph

Description automatically generated

The Validation score R^2 is much better than the Training values, as well as the gain chart and lift chart showing much better predictions from the fitting with the optimal predictor. There is also less underestimation occurring. The lift chart, however, shows less response from the first 10% of the data compared to the training data. The decile chart shows around the same response from the first 10%. But the lift chart and decile chart for the validation data is better fitting overall.

**Forward Selection**

A screenshot of a computer

Description automatically generated

Choosing a value of 4 in the Forward Selection process for F-in.

A screenshot of a computer

Description automatically generated

Subset 4 for the forward selection shows the best model has Variables RM, B, and LSTAT with the highest R^2 and Adjusted R^2 values and lowest Mallows's CP value.

**Backward Elimination**

A screenshot of a computer

Description automatically generated

Choosing a value of 2.5 in the backward elimination process for F-out.

A screenshot of a computer screen

Description automatically generated

Subset 4 with variables RM, B, and LSTAT are shown to be the best subset with the lowest Mallows’s Cp but it has an R^2 close to the others as well.

**Stepwise Selection**

A screenshot of a computer

Description automatically generated

Used the values of 4 for F-in and 2.5 for F-out in the Stepwise Selection process.

A screenshot of a spreadsheet

Description automatically generated

Subset 5 shows that RM, B, and LSTAT make the best subset with the highest R^2 value and Adjusted R^2 as well as the lowest Mallows’s Cp value.

**“Best Subsets”**

A screenshot of a spreadsheet

Description automatically generated

The “Best Subsets” selection shows subset 5 with the lowest Mallows’s Cp and highest R^2 and Adjusted R^2 values. The variables consistently in the best subset of all the feature selection processes are RM, B, and LSTAT.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

All coefficients are statistically significant, and the RM is the most influential on the response.

A screenshot of a graph

Description automatically generated

A graph of a graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generatedA graph of a bar chart

Description automatically generated

The fitting on the lift curve and gain curve are slightly better than the original but still show high amounts of under-estimation in the RROC curve.

Using the Predictors RM, B and LSTAT, the results in the multiple linear regression model are close to the original regression model that included the other 3 variables CRIM, INDUS and TAX for both Training and Validation.

A screenshot of a graph

Description automatically generated

Better validation prediction summary statistics such as better R^2 and lower error values compared to training.

A graph with numbers and lines

Description automatically generated A graph of a function

Description automatically generated with medium confidence

A graph with numbers and a line

Description automatically generated with medium confidence A graph of a bar chart

Description automatically generated

The validation data and training data results are similar to the original model’s charts and measurements. But removing the 3 other variables makes the model less complicated which is better for predictions.

A screenshot of a computer

Description automatically generated

A table with numbers and letters

Description automatically generated

By scoring the new data using the model built gives the predictions of the 10 records Median Housing values in Boston based off the three variables: RM, LSTAT and B.

There are 3 records that are predicted to be greater than or equal to 30 (1000$). Records 3, Records 5, and Records 8. In Logistic Regression we should see these 3 records as being classified as 1 for variable CAT.MEDV. The three records have the highest number of rooms and lower values of lower status and a distribution for B of both high and low values.

**Logistic Regression**

**Using logistic regression to classify whether a house is greater than or equal to 30 (1000$) in value or not.**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The cutoff value chosen was 0.5 and the variables chosen are RM, B and LSTAT. The response variable is CAT.MEDV where a success class is 1 (MEDV > 30k) and 0 (MEDV < 30k)

A screenshot of a computer

Description automatically generated

Training data shows an accuracy of 95% which is pretty good and a high specificity of .9844 but a sensitivity of .7659. There is more error in miss classifying those as 0 when they are actually 1 compared to those that are classified as actual being 0 and being predicted as 1.

A screenshot of a computer

Description automatically generated

The validation data has a better accuracy than the training with 96% and a specificity of 0.9576 and a sensitivity of 0.9730. The overall accuracy for both sensitivity and specificity is better. Using this model on the new data can give predictions on whether they are class 1 or class 0 for CAT.MEDV

A screenshot of a computer

Description automatically generated

The scores show that three of the records are class 1 out of the 10. The classification predicted matches the ones made using the multiple linear regression model.

Logistic Regression to create classification predictions, this can be helpful for classifying customers, target audiences, who is most likely to have an illness, etc. In this example logistic regression was used to predict Median Housing Values in Boston whether they were greater than 30 (1000$) equal or not.

The three predictor variables were chosen from the use of Principle Component Analysis, Correlation Matrix, and Multiple Linear Regression variable selection processes to choose variables TAX, INDUS, CRIM, RM, LSTAT, and B. The feature selection process then showed that variables RM, LSTAT, and B were the best subset of variables to use in the multiple linear regression analysis. These variables show an accurate prediction with 96% accuracy from the validation data set as well as high sensitivity and specificity (>.95). The logistic regression model and multiple linear regression models show high accuracy in their classifications and predictions.

**Hands-on Exercise 3**

**ISM 6136**

**Data Partition and Variable Selection:**

Neural Networks and Classification Trees are trained and used to classify output variables using several input variables. Neural Networks have several layers of connected neurons with the input layer receiving the data, the hidden layers to process the connections with weights to show the strength of the connections, and the output layer gives the predicted class. The network learns to adjust the weights during training to improve accuracy by comparing the target values to the output. Classification trees split data into branches to predict classifications. Each branch has nodes that split based on certain cutoff values, until it reaches a final node that would predict the class based on the previous node functions.

This analysis will be using variables chosen from the previous exercise using the Principle Component Analysis, the Correlation Matrix Analysis, and the variable selection process using Multiple Linear Regression to predict MEDV values. The variables chosen to be used in the Classification Tree and Neural Network Classification are CRIM, RM, LSTAT, B, TAX, and INDUS to classify the CAT.MEDV value. The data was partitioned into 60% training and 40% validation.

**Classification trees**

**Model Settings-**

Data was normalized, using variables CRIM, RM, LSTAT, B, TAX, and INDUS to classify the CAT.MEDV class. The three model methods being used are Full Grown(Tries to have as much partitioning as possible), Best Pruned(focuses on removing branches and reducing overfitting) and Minimum error (focuses on minimizing error)

**Full Grown-**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated** **A diagram of a positive rate

Description automatically generated** A graph of a bar chart

Description automatically generated A diagram of a function

Description automatically generated

In the training portion, there are no errors when predicting. The RROC curve and Lift Chart have perfect fitting to the optimum classifier. The decile chart show that the model is highly accurate for classifying around 20% of the data but not as well with the rest of the data.

**A screenshot of a computer

Description automatically generated** **A graph of a positive rate

Description automatically generated with medium confidence**

**A diagram of a function

Description automatically generated** **A graph of a bar chart

Description automatically generated**

The validation data does not show perfect fitting in the RROC curve and the lift chart to the optimum predictor but it is close. There is ~6% error with .939 sensitivity and .946 specificity. The decile chart shows that it still classifys the top 20% of the data better than the rest but not as well as the training.

The close fits in the RROC curve and lift chart show how close the model’s predictions are with the outcomes.

The RROC curve shows the sensitivity of detecting true positives across the range of specificities. The optimum has a 1 sensitivity across the range as it identifies all positive cases without any false negatives.

The Lift Chart compares the sensitivity with proportion of true positive predictions. This shows how effective the model is at identifying true positive cases.

The Decile chart shows how well the model performs in 10 equal divisions of the data.

This chart shows that the validation has better classification across more data but has lower means in the first 20% compared to the training decile chart. Showing it has less overfitting.

**Best Pruned-**

**A diagram of a network

Description automatically generated**

**A screenshot of a computer

Description automatically generatedA graph of a positive rate

Description automatically generated with medium confidence** **A diagram of a function

Description automatically generatedA graph of a bar graph

Description automatically generated**

The Training data shows high accuracy (98.7%) and specificity at 1. The error comes from the sensitivity 0.915 where there are 4 error values predicted 0 when they were actually 1. The RROC curve and lift chart show the model to have a close fit to the optimum classifier.

The Decile chart shows that 20% of the data is easy to classify compared to the other 80%.

A screenshot of a computer

Description automatically generated A graph of a positive rate

Description automatically generated with medium confidence

A diagram of a function

Description automatically generated A graph of a bar graph

Description automatically generated

The validation data shows less accuracy but higher sensitivity, showing it correctly classifies more positive values than the training data. It did however incorrectly classified more negative cases, lowering the specificity. The RROC curve and the lift chart both show better fitting to the optimum, showing more correctly classified positive cases.

**Minimal Error-**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated** **A graph of a positive rate

Description automatically generated with medium confidence**

A diagram of a function

Description automatically generated with medium confidence A graph of a bar graph

Description automatically generated

The minimum error tree training data has an accuracy of 96% with a specificity of 0.977 and a sensitivity of 0.872. The RROC curve and lift chart have ok fitting to the optimum and the decile chart shows 20% of the data is easy to classify.

A screenshot of a data analysis

Description automatically generated A graph of a positive and negative rate

Description automatically generated with medium confidence

A diagram of a function

Description automatically generated with medium confidence A graph of a bar graph

Description automatically generated

The validation shows an accuracy of 94.6% and a specificity of 0.945 and sensitivity of 0.946. The RROC curve and Lift charts have better fitting on the optimum predictor and the decile chart shows it is better at classifying 30% of the data.

**Best Tree-**

The best tree that will be used for scoring the new data will be the Best Pruned tree. The Best Pruned Tree model is not prone to overfitting because of the minimal amount of branches. It correctly classified the most of the validation data compared to Minimum Error and Full Grown. Despite the Full Grown having perfect results in the Training it had the most incorrectly classified positive cases in the validation.

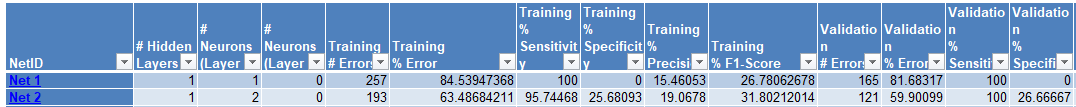
**Neural Networks**

Data was Normalized, the hidden layer functions are logistic sigmoid, the cutoff value will start at 0.5 and will be adjusted throughout to find the best Neural Network using the variables CRIM, RM, INDUS, LSTAT, B, and TAX on CAT.MEDV**.**

**NNC Output 1- Cutoff: 0.5**

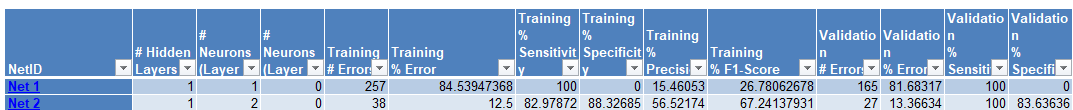
In the first output all nets had a sensitivity of 0 and a specificity of 100%. Where everything was predicted to be 0. Showing that we need a lower cutoff value to allow for higher sensitivity values.

**NNC Output 2- Cutoff: 0.2**

****

The lowest error in training and validation was net 2. It was also the only one to not have 100 sensitivity and 0 specificity where everything is classified as 1. This shows that the cutoff value needs to be adjusted higher to allow for a balance in sensitivity and specificity.

**NNC Output 3- Cutoff: 0.25**

****

Net 2 has a low error rate of 12.5% with training and 13.36% with validation. Specificity and sensitivity are balanced.

**NNC Output 4- Cutoff: 0.24**

**A screenshot of a computer screen

Description automatically generated**

Net 11 has the lowest error in training and validations with 8.88% and 6.93%. The specificity and sensitivity also seemed to be balanced, however net 2 seems to have a better balance between specificity and sensitivity.

These models have shown that consistently a network with 1 layer with 2 neurons would be the best net to have.

A screenshot of a computer

Description automatically generated A graph of a function

Description automatically generated

A graph of a positive and negative rate

Description automatically generated with medium confidence A graph of a bar graph

Description automatically generated

The training data shows an accuracy of 93% with a specificity of 0.965 and sensitivity of 0.745. The RROC curve and Lift chart show moderate fitting but not great.

A screenshot of a computer

Description automatically generated A graph of a positive rate

Description automatically generated with medium confidence

A diagram of a function

Description automatically generated A graph of a bar graph

Description automatically generated

The accuracy in the validation is 93.6% with a specificity of 0.927 and a sensitivity of 0.973. This is much better than the training. The RROC curve and lift chart also show much better fittings with the optimum classifier.

The Neural Network with 1 layer and 2 neurons shows great correct classification of positive cases, but still has a slightly low specificity. While it only misclassified 1 true positive it misclassified 10 true negative cases.

The decile chart shows that it can easily correctly classify 20% of the data.

**Best Models for Neural Networks and Trees**

The best model for the Classification Trees was the Best Pruned Tree showing 95.5% accuracy with high sensitivity (0.946) and specificity (0.958).

The best model for the Neural Networks was 1 hidden layer with 2 neurons having a cutoff value of 0.24. The accuracy was 93.6% in the validation with higher sensitivity (0.973 than the classification tree but lower specificity (0.927).

**Scoring For Trees and Neural Network**

**Best Pruned Tree Score:**

**A screenshot of a computer

Description automatically generated**

**Neural Network Score:**

A screenshot of a computer

Description automatically generated

The classification tree had classified records 3 and 8 as class 1 for CAT.MEDV while the Neural Network had not classified any as 1.

Compared to the Logistic Regression done in exercise 2. The records 3,5, and 8 were predicted as class 1. I think shows that the neural network is not the best model to use compared to the Classification Tree and Logistic Regression. This could be due to overfitting since there is more generalization in the Classification Tree (the Best Pruned Tree method removes complexities by removing branches) and Logistic Regression (using the best subsets the variables used was reduced to 3). The Neural Network had kept all 6 variables.

**Hands-on exercise 4**

Data is partitioned into 60% training and 40% validation.

Variables selected for the k-Nearest Neighbors and Ensemble analysis are: CRIM, RM, b, TAX, LSTAT, and INDUS. These were chosen through the variable selection analysis in the Correlation Matrix Analysis, Principal Component Analysis and Multiple Linear Regression Variable Selection from exercise 1.

The kNN models were done using a 1…k search where k=10. This compares 10 different models with k values 1-10 using the RMSE values for predictions and misclassification percents for classification and chooses the lowest value for scoring the Training and Validation data.

**kNN Model Classification**

The classification model using the 6 predictor variables and CAT.MEDV response variable was found to have the lowest misclassification rate with k = 8.A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The model shows high accuracy in both the training and validation summary. The model sensitivity is low misclassifying many positives with 13% error in the positive class for the validation summary.

**kNN Model Prediction**

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph with red and blue lines

Description automatically generatedA graph of a bar graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A graph of a graph

Description automatically generated with medium confidenceA graph of a bar graph

Description automatically generated

The training summary shows that the model is doing optimum predictions, but in the validation, it has a lower R^2 but still shows good model fitting. The RMSE and MAD values are low.

**Ensembles**

When comparing the three ensemble approaches: Boosting, Bagging and Random Trees, in ASDM Classification Trees was chosen as the Weakest Learner to compare the 3 approaches

**Boosting model**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The training summary is perfect, and the validation shows minimal error. There is higher sensitivity than there is specificity, but the precision is low.

**Bagging model**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The bagging model shows a training summary with low error and high sensitivity and specificity. The Validation summary has a higher error and lower precision than boosting, showing that it misclassifies more positives and negatives.

**Random Trees**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The random tree had the same training summary errors, showing low error and high sensitivity and specificity values. But showing better accuracy with classification on the negatives but similar misclassification on the positives compared to bagging.

Boosting shows the lowest misclassification on positive cases and negative cases compared to Bagging and Random Trees.

**Boosting Model Comparisons with different Weakest Learners: NNC, KNN and Discriminant Analysis**

**Neural Network**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The neural network weakest learner has all cases classified as negative showing an unsuitable model.

**k-Nearest Neighbors**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The k-Nearest Neighbors Training Summary has no error but higher error with the validation. Compared to the decision tree weakest learner, the kNN model has higher error and misclassifies positives much more.

**Discriminant Analysis**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The Discriminant Analysis shows low error in the training summary showing high misclassifications with positive cases at 9. The validation summary shows much better results with only 1 positive misclassified. Compared to the decision tree, this model has a lower precision value but the same recall and specificity. This model is more likely to misclassify a positive than the decision tree boosting model making it the most suitable ensemble.

**Classification Model Comparisons:**

**kNN**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**Boosting**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

The boosting model shows better training and validation summary metrics with less positive cases misclassified. The training summary shows perfect model prediction with the training portion of the data and the validation summary shows only 6 cases misclassified. The boosting model is the better classifier compared to the kNN classifier for this data.

**Prediction Model Comparison:**

**kNN**

**A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated**

**Boosting**

**A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated**

The kNN predicition model has better R^2 values with lower RMSE and MAD than the boosting decision tree model. Showing that kNN is better for predictions than Ensembles for this data.

**Discuss what would be the recommended model and its performance**

**kNN where k=10 with variables CRIM, INDUS, RM, TAX, B, and LSTAT to predict the MEDV value score.**

A screenshot of a computer

Description automatically generated

The prediction from the kNN model shows that records 3,5, and 8 would be >= 30 MEDV. They would be classified as 1 with the CAT.MEDV variable.

**Boosting with the decision tree weak learner with variables CRIM, INDUS, RM, TAX, B, and LSTAT to classify the CAT.MEDV value score.**

A screenshot of a computer

Description automatically generated

The Boosting Classification model agrees that records 3 and 8 should be class 1, but on record 5 classifies as 0. While all the other class 0s have 100% probability of being class 0 record 5 shows a near split probability, while still favoring class 0. Record 8 also doesn’t have a 100% probability of being class 1.

The most common class 1 records in the new data from exercises 2, 3, and 3, show records 3, 5, and 8 to be class 1, with records 3 and 8 being classified as class 1 every time.