

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

$B = 1000(B_k - 0.63)^2$ where B_k 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)]

Inputer Parameters															
Variable	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT_MEDV
Reduction Type	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
# Records Treated	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missing Value Code															
# Output Records	506														
#Records Deleted	0														
NO MISSING VALUES															

A box plot showing the distribution of 12 variables: CRIM, ZN, INDUS, NOX, RM, AGE, DIS, TAX, PTRATIO, B, LSTAT, and MEDV. The y-axis is labeled 'Value' and ranges from 0 to 800. The x-axis is labeled 'All Variables'. The plot shows that TAX has the highest median value (around 330), followed by B (around 380). LSTAT has the lowest median value (around 15). MEDV has a median around 25. Most variables have a small number of outliers, particularly CRIM, ZN, RM, DIS, and B.

Correlation Matrix Analysis:

[illegible]

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.).

Component	Eigenvalue	Variance, %	Cumulative Variance, %
Component 1	6.126848826	47.12960636	47.12960636
Component 2	1.433275122	11.02519325	58.1547996
Component 3	1.242616673	9.558589793	67.7133894
Component 4	0.857575108	6.596731601	74.310121
Component 5	0.834815937	6.421661052	80.73178205
Component 6	0.657407175	5.056978272	85.78876032
Component 7	0.535356086	4.11812374	89.90688406
Component 8	0.396097314	3.046902419	92.95378648
Component 9	0.27694333	2.130333305	95.08411979
Component 10	0.220237825	1.694137115	96.7782569
Component 11	0.186014367	1.430879746	98.20913665
Component 12	0.169302975	1.302330579	99.51146723

Feature/Component	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6	Component 7	Component 8	Component 9
CRIM	0.250951397	-0.315252368	0.246566492	-0.061770707	-0.082156919	-0.219659612	0.777607207	0.153350477	-0.260390284
ZN	-0.256314541	-0.323312905	0.295857816	-0.128711591	-0.320616987	-0.323388102	-0.27499628	-0.402680309	-0.358137486
INDUS	0.346672065	0.112492908	-0.015945915	-0.017145714	0.007811194	-0.076137899	-0.339576454	0.173931716	-0.644416155
CHAS	0.005042434	0.454829136	0.289780815	-0.815941364	-0.086530945	0.167490141	0.074136208	-0.024662148	0.013727772
NOX	0.342852313	0.219115531	0.120964108	0.12822614	-0.136853557	-0.15298267	-0.19963484	0.08012056	0.018522012
RM	-0.18924257	0.149331541	0.593961167	0.280591838	0.423447195	0.059267074	0.063939924	-0.326752259	-0.047898035
AGE	0.313670596	0.311977778	-0.017674809	0.175206033	-0.016690847	-0.071709145	0.116010713	-0.600822917	0.067562182
DIS	-0.321543866	-0.349070004	-0.049736273	-0.215435854	-0.098592247	0.023438723	-0.10390044	-0.121811982	0.153291245
RAD	0.319792768	-0.271520937	0.287254835	-0.132349958	0.204131621	-0.143194012	-0.137942546	0.080358311	0.470890669
TAX	0.338469147	-0.239453645	0.220744471	-0.103335092	0.130460565	-0.192934282	-0.314886835	0.082774347	0.176563391
PTRATIO	0.204942258	-0.305896954	-0.323446272	-0.282621976	0.584002232	0.2731533	0.002323869	-0.317884202	-0.25442836
B	-0.202972612	0.238559443	-0.300145901	-0.168498497	0.345606947	-0.803454537	0.070294759	-0.004922915	0.044898024
LSTAT	0.30975984	-0.074322027	-0.267000248	-0.069414411	-0.394561129	-0.05321583	0.087011169	-0.424352926	0.19522139

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.

Record ID	Value
Record 1	1.78E-15
Record 2	1.78E-15
Record 3	3.55E-15
Record 4	0
Record 5	3.55E-15
Record 6	1.78E-15
Record 7	1.78E-15
Record 8	1.78E-15
Record 9	7.11E-15
Record 10	1.78E-15
Record 11	0
Record 12	1.78E-15
Record 13	4.44E-15
Record 14	2.66E-15
Record 15	1.33E-15
Record 16	3.55E-15
Record 17	4.44E-15
Record 18	3.55E-15
Record 19	3.55E-15
Record 20	3.55E-15
Record 21	5.33E-15
Record 22	3.55E-15
Record 23	3.55E-15
Record 24	4.44E-15

Record ID	Value
Record 1	7.548994
Record 2	4.612111
Record 3	4.972568
Record 4	6.928668
Record 5	7.13728
Record 6	6.483444
Record 7	6.058685
Record 8	14.24984
Record 9	25.70222
Record 10	13.67518
Record 11	16.84383
Record 12	10.93213
Record 13	7.847534
Record 14	5.962962
Record 15	6.360956
Record 16	6.063612
Record 17	10.29623
Record 18	4.947724
Record 19	10.95356
Record 20	5.200512
Record 21	8.239237
Record 22	5.706603
Record 23	7.001617
Record 24	8.268697

Multiple Linear Regression:

Regression Summary

Metric	Value
Residual DF	492
R2	0.740642664
Adjusted R2	0.733789726
Std. Error Es	4.745298182
RSS	11078.78458

Highest R-square using all variables compared to the variables

The R² and Adjusted R² show that the linear regression model predicted values fits 73-74% of the observed data.

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value	Column
Intercept	36.45948839	26.43222601	46.48675076	5.103458811	7.144074193	3.283E-12	
CRIM	-0.10801136	-0.172584412	-0.043438304	0.032864994	-3.28651687	0.0010868	
ZN	0.046420458	0.019448778	0.073392139	0.013727462	3.381576282	0.0007781	
INDUS	0.020558626	-0.100267941	0.141385193	0.061495689	0.334310042	0.7382881	INDUS
CHAS	2.686733819	0.993904193	4.379563446	0.861579756	3.118380858	0.001925	
NOX	-17.7666112	-25.27163356	-10.26158889	3.819743707	-4.65125741	4.246E-06	
RM	3.809865207	2.988726773	4.63100364	0.417925254	9.1161402	1.979E-18	
AGE	0.000692225	-0.02526232	0.026646769	0.013209782	0.052402427	0.9582293	AGE
DIS	-1.47556685	-1.867454981	-1.08367871	0.199454735	-7.3980036	6.013E-13	
RAD	0.306049479	0.175692169	0.436406789	0.06634644	4.612899768	5.071E-06	
TAX	-0.01233459	-0.019723286	-0.004945902	0.003760536	-3.28000914	0.0011116	
PTRATIO	-0.95274723	-1.209795296	-0.695699168	0.130826756	-7.28251056	1.309E-12	
B	0.009311683	0.004034306	0.01458906	0.002685965	3.466792558	0.0005729	
LSTAT	-0.52475838	-0.624403622	-0.425113133	0.050715278	-10.3471458	7.777E-23	

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.

Regression Summary

Metric	Value
Residual DF	494
R2	0.74058228
Adjusted R2	0.734805772
Std. Error Es	4.736233824
RSS	11081.36395

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

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	36.341145	26.38464913	46.29764088	5.067492203	7.171425935	2.727E-12
CRIM	-0.10841335	-0.17281767	-0.044009021	0.032779445	-3.30735754	0.0010104
ZN	0.045844929	0.019275889	0.07241397	0.01352267	3.390227667	0.0007543
CHAS	2.718716303	1.040324913	4.397107693	0.854239823	3.182614798	0.0015515
NOX	-17.3760234	-24.32199031	-10.43005655	3.535243066	-4.91508592	1.209E-06
RM	3.80157884	3.003258393	4.599899288	0.406315906	9.356214671	2.89E-19
DIS	-1.49271146	-1.857631161	-1.12779176	0.185730779	-8.03696333	6.837E-15
RAD	0.299608454	0.175037411	0.424179497	0.063402104	4.725528555	2.997E-06
TAX	-0.01177797	-0.018403857	-0.00515209	0.003372332	-3.49253054	0.0005214
PTRATIO	-0.94652457	-1.200109823	-0.692939318	0.129065618	-7.3336694	9.235E-13
B	0.009290845	0.004037216	0.014544473	0.002673905	3.474635544	0.0005566
LSTAT	-0.52255346	-0.615731781	-0.429375132	0.047424359	-11.018672	2.141E-25

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.

Regression Summary

Metric	Value
Residual DF	500
R2	0.633621077
Adjusted R2	0.629957288
Std. Error Es	5.594702908
RSS	15650.35032

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.

Regression Summary

Metric	Value
Residual DF	498
R ²	0.652510026
Adjusted R ²	0.647625629
Std. Error Est	5.459504882
RSS	14843.48439

The model fit didn't change much, the R² and Adjusted R² grew 2%. Checking the coefficients and their p-values might give us greater insight on the model.

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	13.70678783	3.219228295	24.19434737	5.337889474	2.567829083	0.0105244
CHAS	3.500181325	1.574569692	5.425792957	0.980085217	3.57130305	0.0003897
NOX	-25.211221	-32.980963	-17.44147899	3.954592481	-6.37517547	4.172E-10
RM	6.712133319	5.953859613	7.470407026	0.385941193	17.39159604	2.758E-53
DIS	-0.91401535	-1.27459233	-0.553438371	0.183524113	-4.98035563	8.766E-07
TAX	-0.0016671	-0.006096575	0.002762383	0.002254487	-0.73945675	0.4599781
PTRATIO	-1.11684049	-1.385197471	-0.848483501	0.136586584	-8.17679494	2.438E-15
B	0.01442867	0.008540389	0.02031695	0.002996978	4.814405742	1.961E-06

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.

Regression Summary

Metric	Value
Residual DF	499
R ²	0.652128487
Adjusted R ²	0.647945663
Std. Error Est	5.457025092
RSS	14859.78231

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

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	14.42589707	4.118588611	24.73320553	5.246172071	2.749794874	0.0061797
CHAS	3.542207846	1.620719159	5.463696532	0.977991522	3.621920811	0.0003223
NOX	-26.5726636	-33.44581507	-19.69951206	3.498268789	-7.59594679	1.519E-13
RM	6.705364729	5.947652342	7.463077117	0.385657379	17.38684412	2.766E-53
DIS	-0.91641061	-1.276765905	-0.55605531	0.183412178	-4.99645452	8.092E-07
PTRATIO	-1.15956426	-1.402617827	-0.916510688	0.12370842	-9.37336564	2.438E-19
B	0.014968987	0.009261019	0.020676954	0.002905218	5.152448597	3.71E-07

ANOVA

Source	DF	SS	MS	F-Statistic	P-Value
Regression	6	27856.51311	4642.752185	155.9062773	5.1622E-111
Error	499	14859.78231	29.77912286	#N/A	#N/A
Total	505	42716.29542	84.58672359	#N/A	#N/A

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.

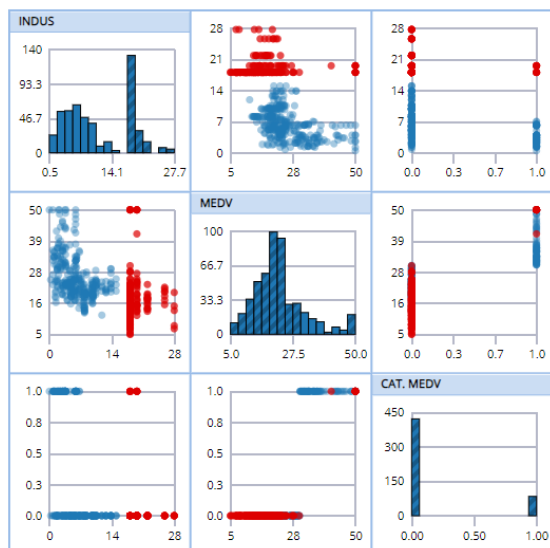
Linear regression: CHAS, NOX, RM, DIS, PTRATIO, and B

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

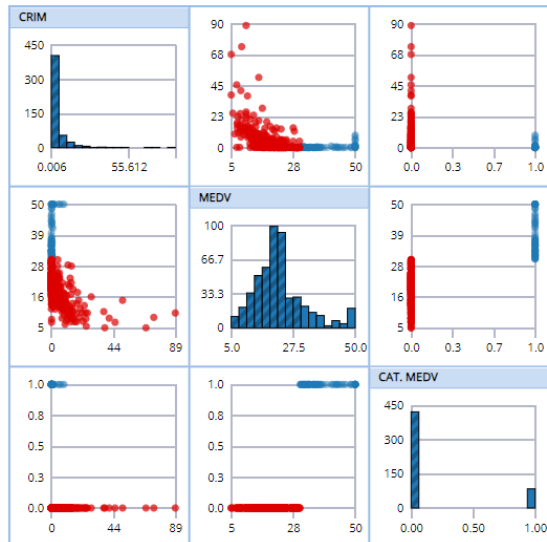
By comparing all the predictors in a scatter plot matrix, the scatter plots and histograms can be compared for relationship strength to confirm if the variables: CRIM, INDUS, NOX, RM, TAX, PTRATIO, LSTAT, AGE, B, DIS, CHAS. Excluding the variables RAD and ZN.

Scatter Plot Matrix:

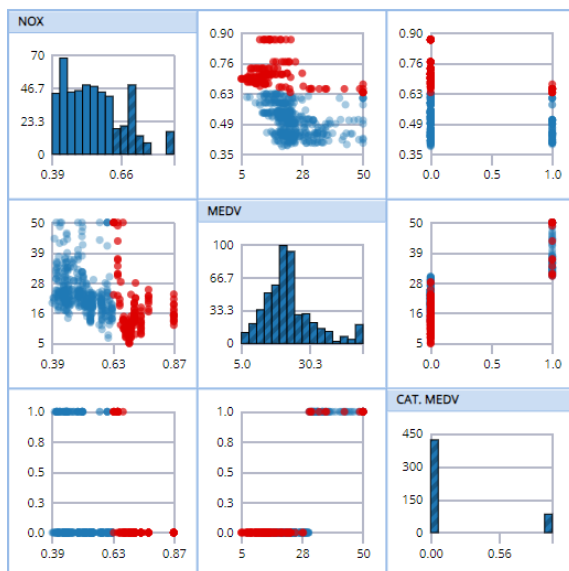
INDUS: Shows a slight correlation but there are some outliers



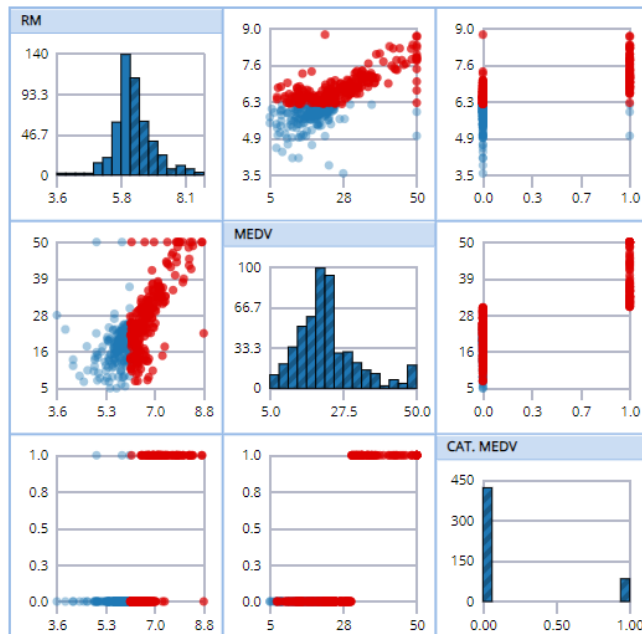
CRIM: Shows a clear negative relationship between CRIM and CAT.MEDV/MEDV with all values of high CRIM in the 0 CAT.MEDV category. Although there is a very large amount of data with low CRIM compared to higher amounts.



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



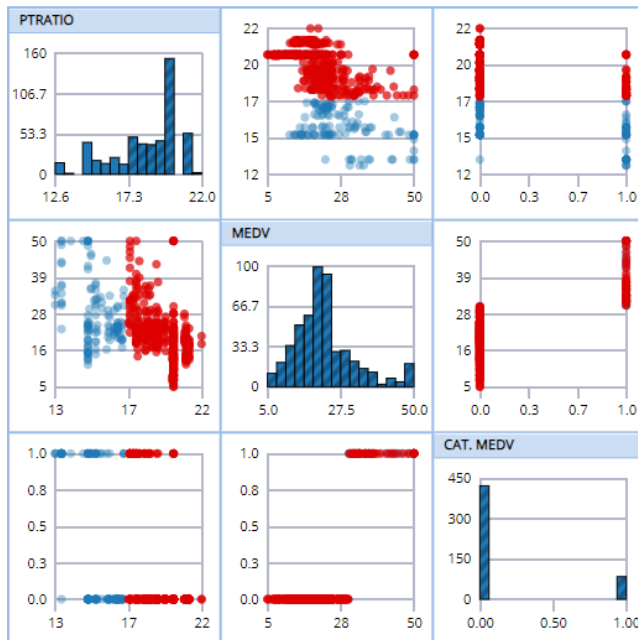
RM: There is a clear positive relationship between MEDV and RM where the higher amounts of RM correlate to higher values of MEDV.



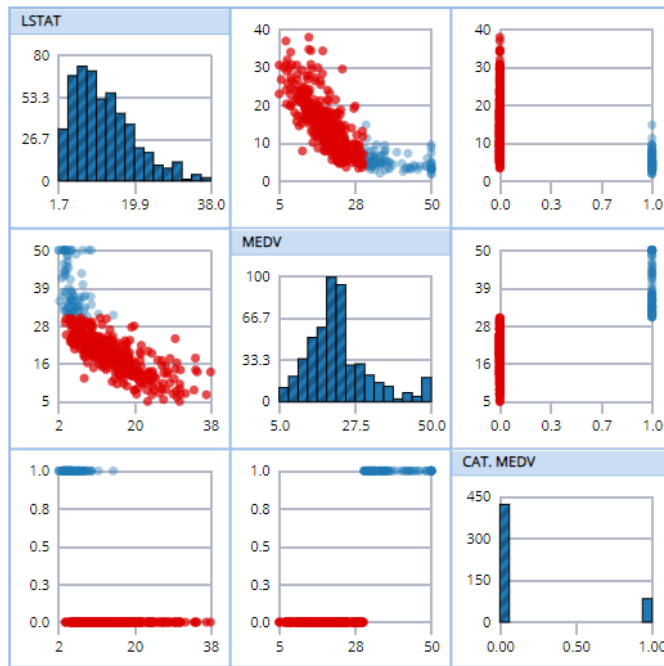
TAX: There is a slight negative relationship between TAX and MEDV and CAT.MEDV categories. There is more instances of there being higher TAX values when CAT.MEDV = 0, which could show a negative relationship, with one outlier being CAT.MEDV=1 with high TAX values.



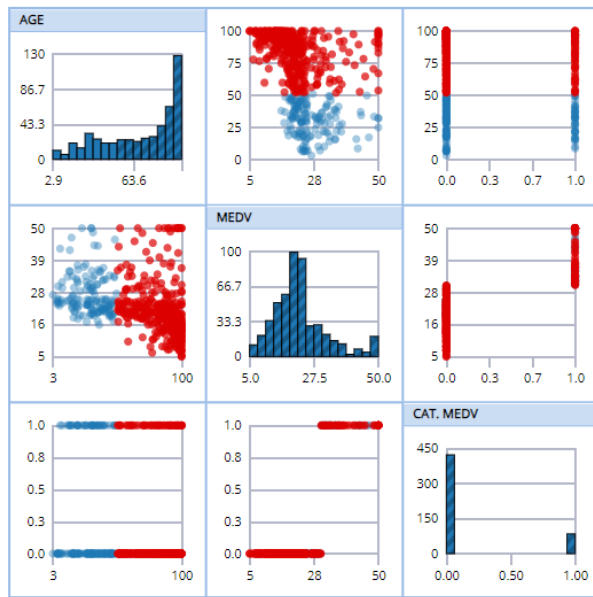
PTRATIO: There is no clear relationship between PTRATIO and MEDV and CAT.MEDV, showing it is actually not a clear predictor.



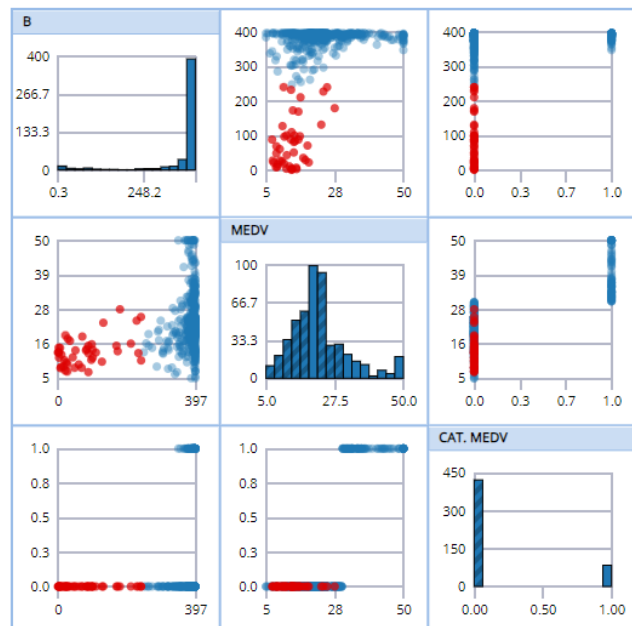
LSTAT: There is a very strong negative relationship between LSTAT and MEDV and CAT.MEDV



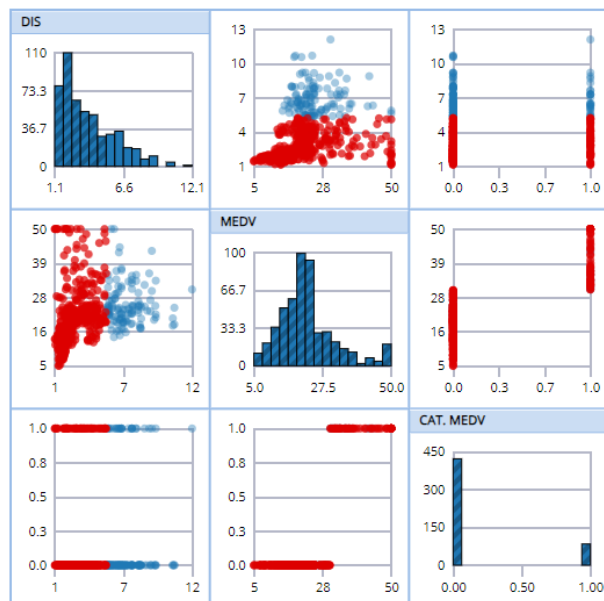
AGE: SHOWS there is no clear relationship between AGE and MEDV and CAT.MEDV.



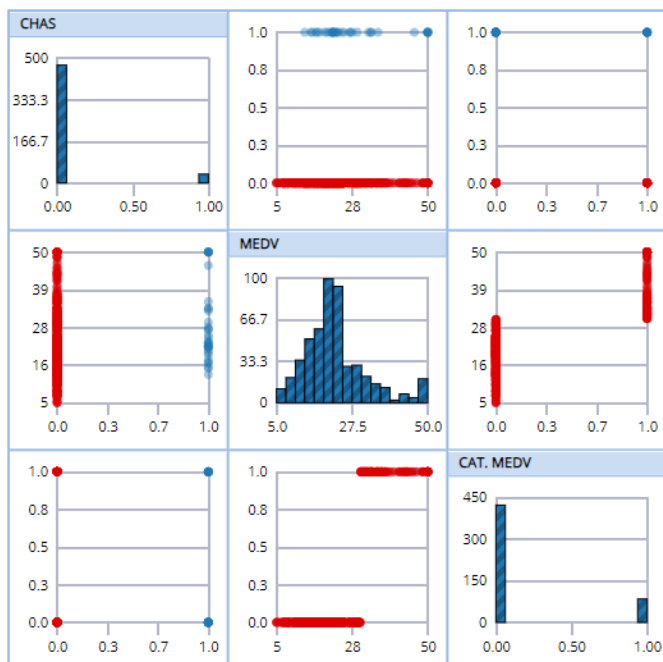
B: There shows a postive relationship between B and MEDV and CAT.MEDV. Although the distribution in B shows there might the relationship could be misleading by giving a higher average, because there is also a large amount that has lower MEDV values as well.



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



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.



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

Data	
Workbook	Boston_Housing.xlsx
Worksheet	Data
Range	\$A\$1:\$O\$507
# Records in the input data	506

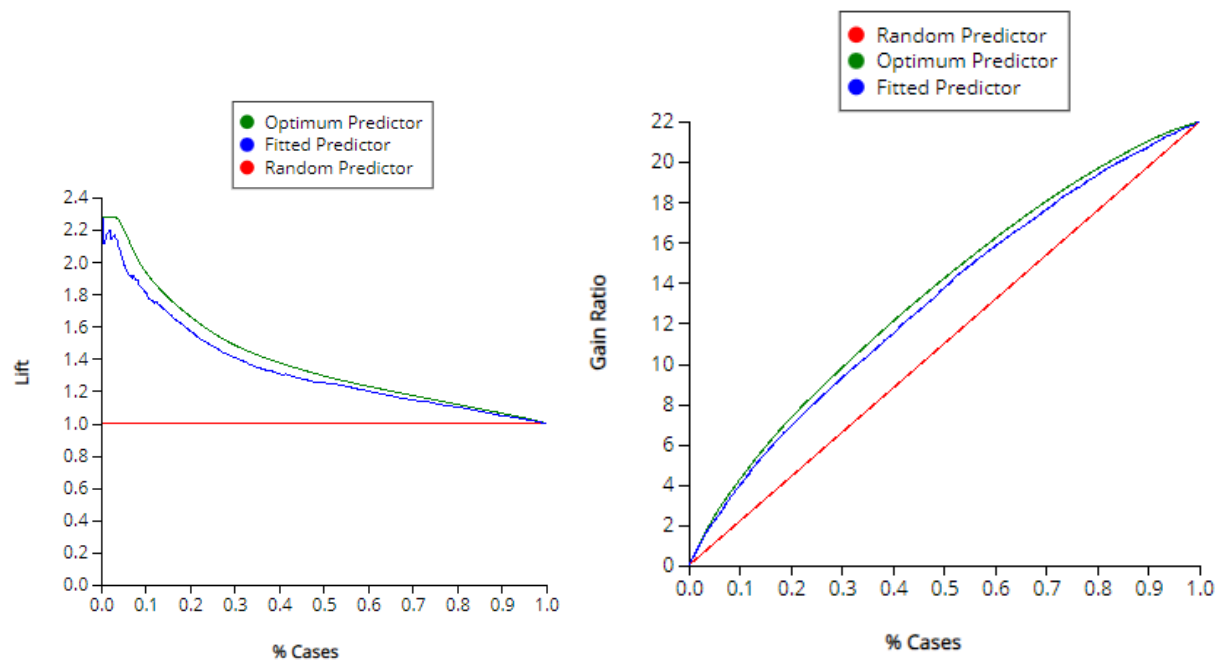
Variables								
# Selected Variables	8							
Selected Variables	CRIM	INDUS	RM	TAX	B	LSTAT	MEDV	CAT. MEDV

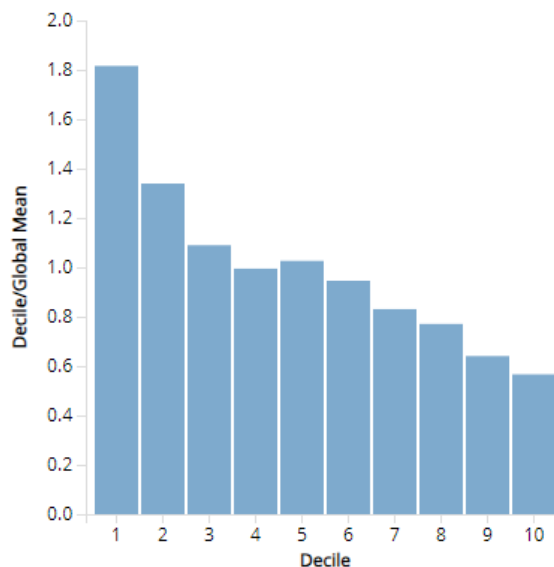
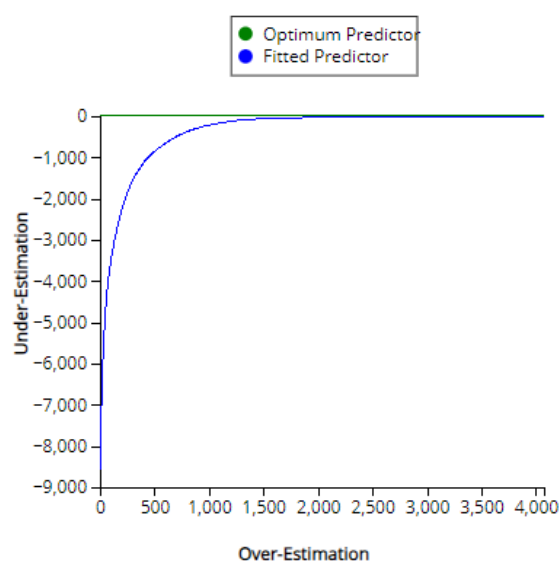
Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

tion Summary

Partition	# Records
Training	304
Validation	202

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



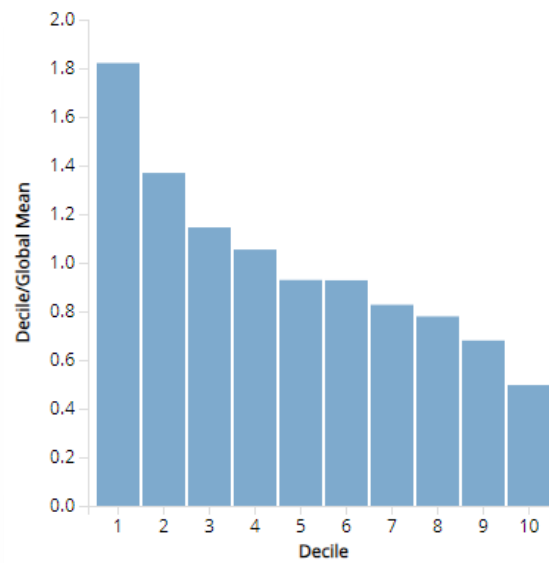
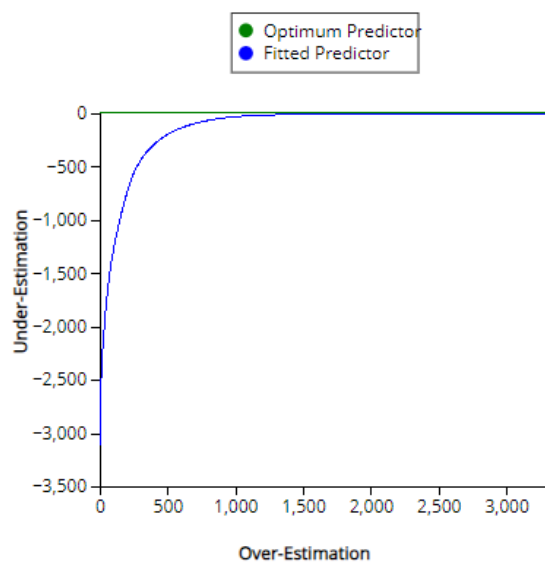
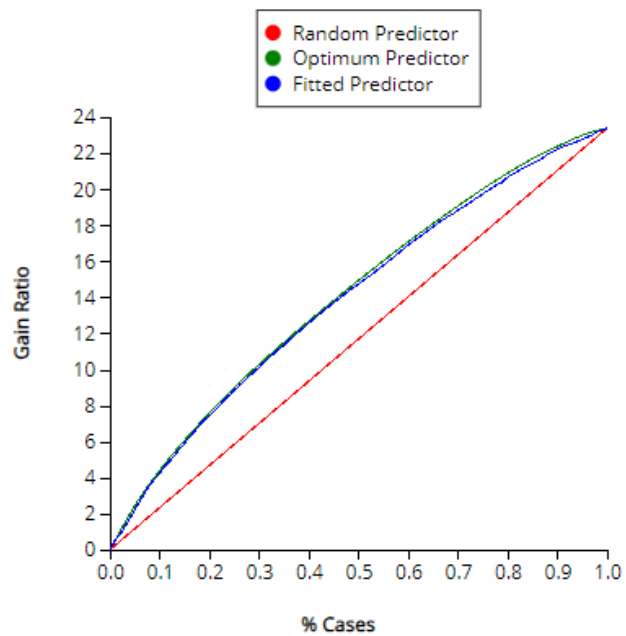
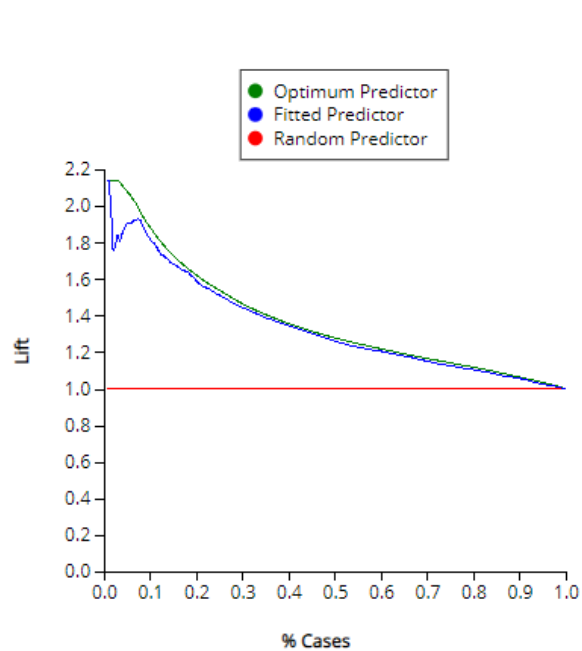


Training: Prediction Summary

Metric	Value
SSE	10189.13
MSE	33.51687
RMSE	5.789375
MAD	4.093294
R2	0.598919

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.



Validation: Prediction Summary

Metric	Value
SSE	4613.04
MSE	22.83683
RMSE	4.77879
MAD	3.503474
R2	0.729755

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

Feature Selection

☒ Perform Feature Selection

Maximum Subset Size: 6

Method

- ☐ Backward Elimination
- ☒ Forward Selection
- ☐ Sequential Replacement
- ☐ Stepwise Selection
- ☐ Best Subsets

Stepwise Selection Options

F-in: 4

F-out: 2.5

Best Subsets Options

Number of Subsets: 1

Help Done

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

Feature Selection

Best Subsets									
Subset ID	Intercept	CRIM	INDUS	RM	TAX	B	LSTAT		
Subset 1	1	0	0	0	0	0	0	0	0
Subset 2	1	0	0	0	0	0	0	1	
Subset 3	1	0	0	1	0	0	0	1	
Subset 4	1	0	0	1	0	1	1	1	

Best Subsets Details							
Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability	
Subset 1	1	25404.17	438.4990699	-1.1E-15	-1.11022E-15	4.89459E-56	
Subset 2	2	12165.82	54.61797944	0.521109	0.519523757	3.48922E-10	
Subset 3	3	10581.55	10.43849012	0.583472	0.580704517	0.02377885	
Subset 4	4	10293.73	4.048929714	0.594802	0.590749764	0.385730067	

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

Feature Selection

☒ Perform Feature Selection

Maximum Subset Size: 6

Method

☒ Backward Elimination

☐ Forward Selection

☐ Sequential Replacement

☐ Stepwise Selection

☐ Best Subsets

Stepwise Selection Options

F-in: 4

F-out: 2.5

Best Subsets Options

Number of Subsets: 1

Help Done

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

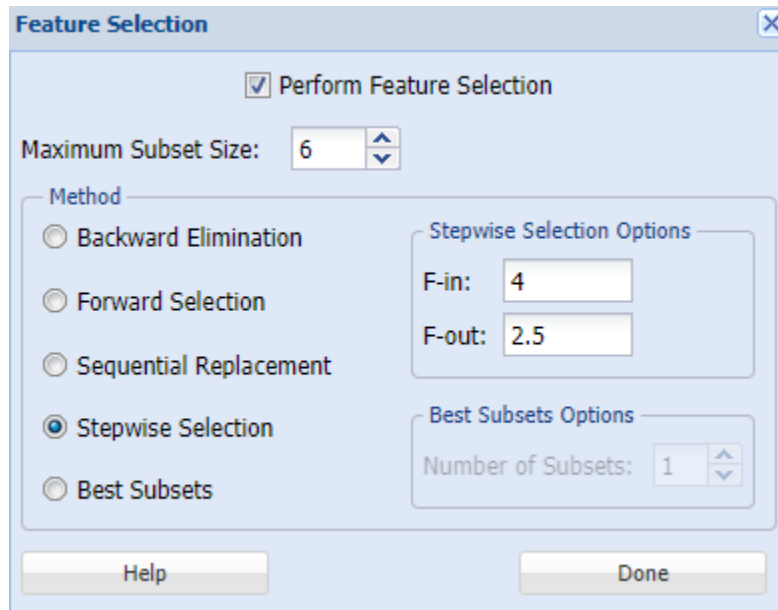
Feature Selection

Best Subsets								
Subset ID	Intercept	CRIM	INDUS	RM	TAX	B	LSTAT	
Subset 1	1	1	1	1	1	1	1	1
Subset 2	1	0	1	1	1	1	1	1
Subset 3	1	0	0	1	1	1	1	1
Subset 4	1	0	0	1	1	0	1	1

Best Subsets Details							
Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability	
Subset 1	7	10189.13	7	0.598919	0.590816502	#N/A	
Subset 2	6	10207.44	5.533726057	0.598198	0.591456742	0.465620446	
Subset 3	5	10258.81	5.031089802	0.596176	0.590773955	0.363459227	
Subset 4	4	10293.73	4.048929714	0.594802	0.590749764	0.385730067	

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



The image shows a 'Feature Selection' dialog box. At the top, there is a checkbox labeled 'Perform Feature Selection' which is checked. Below this, 'Maximum Subset Size' is set to 6. Under the 'Method' section, five radio buttons are present: 'Backward Elimination', 'Forward Selection', 'Sequential Replacement', 'Stepwise Selection' (which is selected), and 'Best Subsets'. To the right of these methods are two sections: 'Stepwise Selection Options' containing 'F-in' (set to 4) and 'F-out' (set to 2.5), and 'Best Subsets Options' containing 'Number of Subsets' (set to 1). At the bottom, there are 'Help' and 'Done' buttons.

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

Feature Selection

Best Subsets								
Subset ID	Intercept	CRIM	INDUS	RM	TAX	B	LSTAT	
Subset 1	1	0	0	0	0	0	0	0
Subset 2	1	0	0	0	0	0	0	1
Subset 3	1	0	0	1	0	0	0	1
Subset 4	1	0	0	1	0	1	1	1
Subset 5	0	0	0	1	0	1	1	1

Best Subsets Details							
Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability	
Subset 1	1	25404.17	438.4990699	-1.1E-15	-1.11022E-15	4.89459E-56	
Subset 2	2	12165.82	54.61797944	0.521109	0.519523757	3.48922E-10	
Subset 3	3	10581.55	10.43849012	0.583472	0.580704517	0.02377885	
Subset 4	4	10293.73	4.048929714	0.594802	0.590749764	0.385730067	
Subset 5	3	10302.05	2.291666396	0.594474	0.59177942	0.511383468	

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”

Feature Selection

Best Subsets								
Subset ID	Intercept	CRIM	INDUS	RM	TAX	B	LSTAT	
Subset 1	1	0	0	0	0	0	0	0
Subset 2	1	0	0	0	0	0	0	1
Subset 3	1	0	0	1	0	0	0	1
Subset 4	1	0	0	1	0	1	1	1
Subset 5	0	0	0	1	0	1	1	1

Best Subsets Details							
Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability	
Subset 1	1	25404.17	438.4990699	-1.1E-15	-1.11022E-15	4.89459E-56	
Subset 2	2	12165.82	54.61797944	0.521109	0.519523757	3.48922E-10	
Subset 3	3	10581.55	10.43849012	0.583472	0.580704517	0.02377885	
Subset 4	4	10293.73	4.048929714	0.594802	0.590749764	0.385730067	
Subset 5	3	10302.05	2.291666396	0.594474	0.59177942	0.511383468	

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.

Regression Summary

Metric	Value
Residual DF	300
R2	0.594801747
Adjusted R2	0.590749764
Std. Error Es	5.857680593
RSS	10293.72658

Coefficients

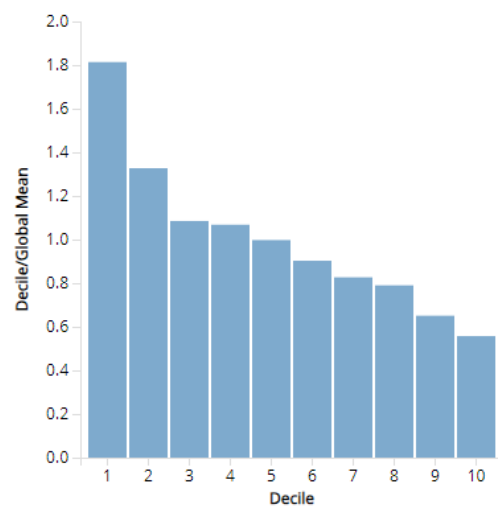
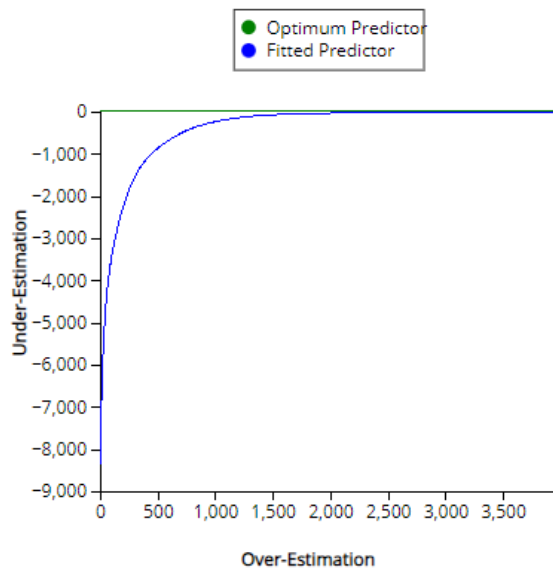
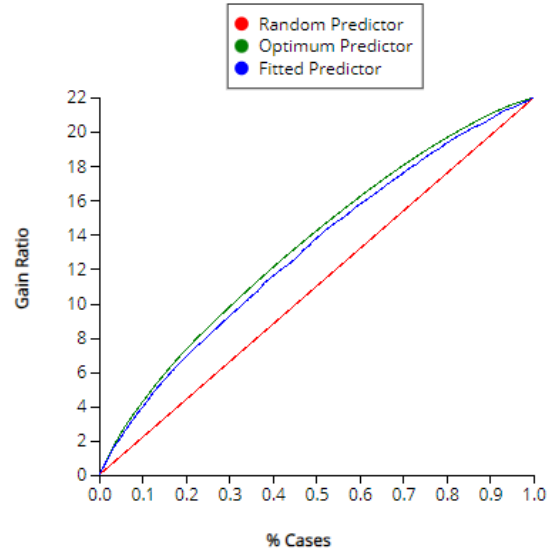
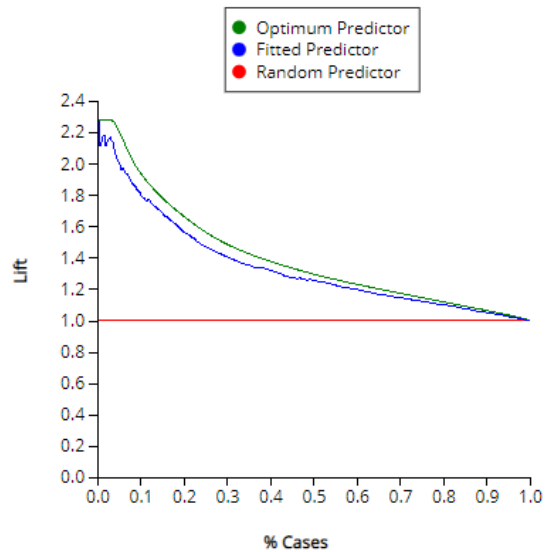
Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	-2.47190311	-12.34612407	7.40231786	5.017635985	-0.49264297	0.6226252
RM	4.543949067	3.299181577	5.788716558	0.632534979	7.183711916	5.389E-12
B	0.011620512	0.003724737	0.019516286	0.004012278	2.896237611	0.0040543
LSTAT	-0.60728866	-0.730825231	-0.483752079	0.062775744	-9.67393798	1.918E-19

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

Training: Prediction Summary

Metric	Value
SSE	10293.73
MSE	33.86094
RMSE	5.819016
MAD	4.132768
R2	0.594802

Lift Chart (Alternative)



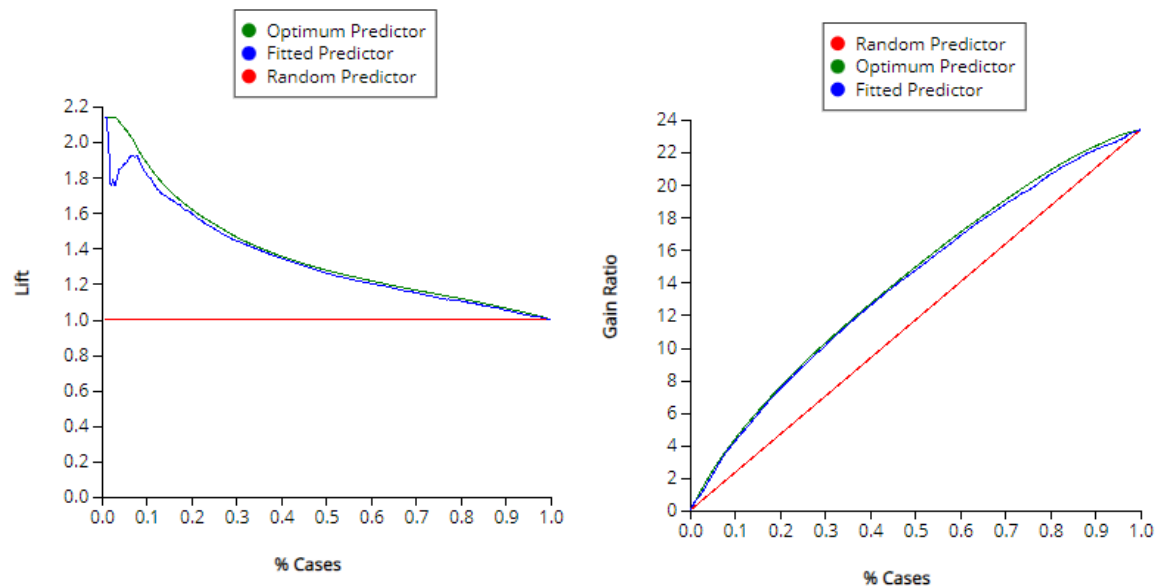
The fitting on the lift curve and gain curve are slightly better than the original but still show high amounts of under-estimation in the RROC curve.

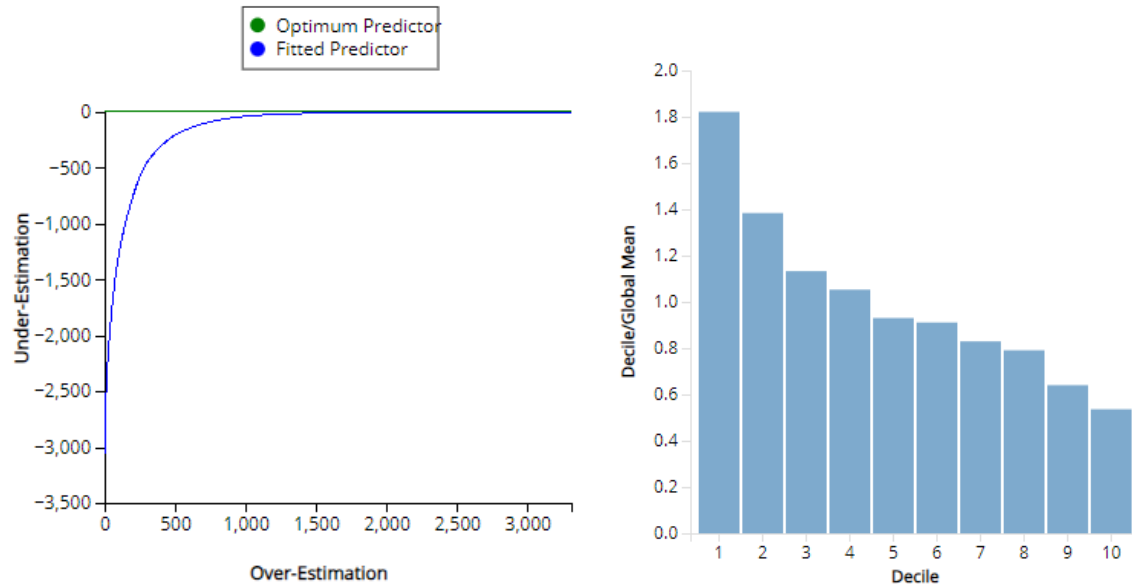
Using the Predictors RM, B and LSTAT, the results in the multiple linear regression model are close to the original regression model that included the other 3 variables CRIM, INDUS and TAX for both Training and Validation.

Validation: Prediction Summary

Metric	Value
SSE	4734.523
MSE	23.43823
RMSE	4.841305
MAD	3.562072
R2	0.722638

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





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.

Scoring

Record ID	Prediction: MEDV
Record 1	23.97054958
Record 2	14.51231691
Record 3	40.35272438
Record 4	11.18141032
Record 5	30.88549258
Record 6	26.96637006
Record 7	18.01923506
Record 8	44.16874487
Record 9	2.048879541
Record 10	10.01145071

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
2	0.00487	16	2.25	0	0.65	5.454	60	3.75	1	265	15.3	378	4.5
3	1.2456	1	8	0	0.522	5.5	95	3.65	4	298	20	366.57	20.2
4	0.03495	77	3	0	0.824	8.65	18.5	4.511	2	265	17	410	2.05
5	0.15	26	4.98	0	0.354	3.25	60	5.7422	9	295	23.7	591.31	13.15
6	0.87038	1	13.28	0	0.734	7.854	66.3	6.5402	5	389	17.8	369.06	10.9
7	0.0866	1	8.29	0	0.544	6.741	56.2	9.543	1	267	19	399.6	9.61
8	10.587	0	11	0	0.745	5.872	71.3	7.2457	7	423	18.7	383.36	17.53
9	3.22158	1	18.95	0	0.56	9.436	74.9	1.78773	6	340	17.4	634.33	5.94
10	0.06426	0	5.4	0	0.15	0.866	44.7	3.5921	4	629	15.3	379.12	6.29
11	0.086703	84	2.51	0	0.44	2.742	8.33	9.73	3	239	16.2	329.2	6.26

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.

Data Source
Worksheet: Workbook:
Data range: #Columns:
Rows In
Training Set: Validation Set: Test Set:

Variables
☒ First Row Contains Headers

Variables In Input Data	Selected Variables
Record ID	RM
CRIM	B
INDUS	LSTAT
TAX	
MEDV	

Weight Variable:
Output Variable:

Target
Classes
Number of Classes:
Binary Classification
Success Class:
Success Probability Cutoff:

Regression Summary

Metric	Value
# Iterations	6
Residual DF	300
Residual De	100.3086727
Multiple R2	0.616870273

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)

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	253	4	
1	11	36	

Error Report			
Class	# Cases	# Errors	% Error
0	257	4	1.55642
1	47	11	23.40426
Overall	304	15	4.934211

Metrics	
Metric	Value
Accuracy (#correct)	289
Accuracy (%correct)	95.065789
Specificity	0.9844358
Sensitivity (Recall)	0.7659574
Precision	0.9
F1 score	0.8275862
Success Class	1
Success Probability	0.5

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.

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	158	7	
1	1	36	

Error Report			
Class	# Cases	# Errors	% Error
0	165	7	4.242424
1	37	1	2.702703
Overall	202	8	3.960396

Metrics	
Metric	Value
Accuracy (#correct)	194
Accuracy (%correct)	96.039604
Specificity	0.9575758
Sensitivity (Recall)	0.972973
Precision	0.8372093
F1 score	0.9
Success Class	1
Success Probability	0.5

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

Scoring

Record ID	Prediction: CAT. MEDV	PostProb: 1	PostProb: 0
Record 1	0	0.014047638	0.985952362
Record 2	0	5.43671E-05	0.999945633
Record 3	1	0.999609393	0.000390607
Record 4	0	1.76194E-07	0.999999824
Record 5	1	0.866840388	0.133159612
Record 6	0	0.164209158	0.835790842
Record 7	0	0.000520274	0.999479726
Record 8	1	0.999849641	0.000150359
Record 9	0	7.08297E-10	0.999999999
Record 10	0	5.78288E-07	0.999999422

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.

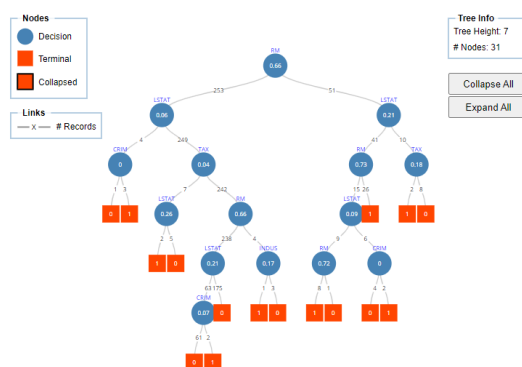
This analysis will be using variables chosen from the previous exercise using the Principle Component Analysis, the Correlation Matrix Analysis, and the variable selection process using Multiple Linear Regression to predict MEDV values. The variables chosen to be used in the Classification Tree and Neural 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-

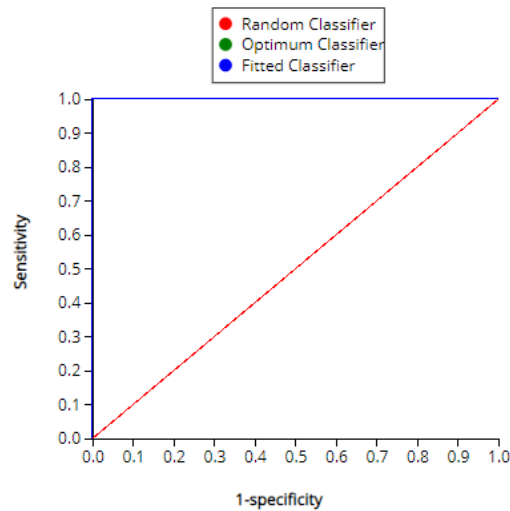
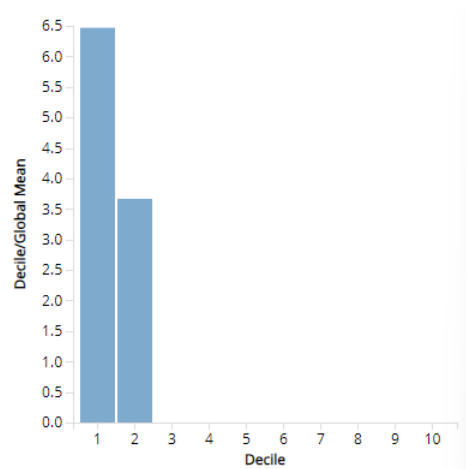
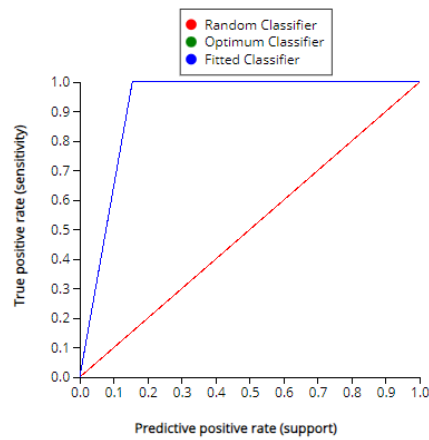


Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	257	0
1	0	47

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	0	0
Overall	304	0	0

Metrics	
Metric	Value
Accuracy (#correct)	304
Accuracy (%correct)	100
Specificity	1
Sensitivity (Recall)	1
Precision	1
F1 score	1
Success Class	1
Success Probability	0.5



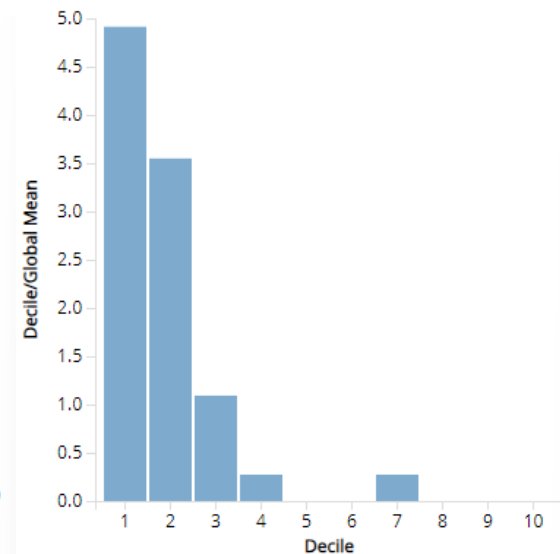
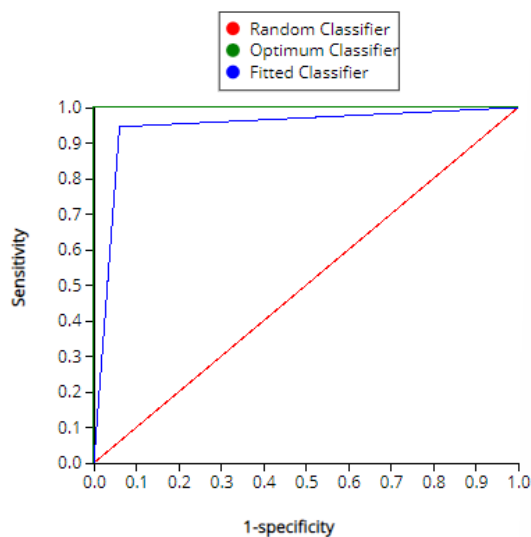
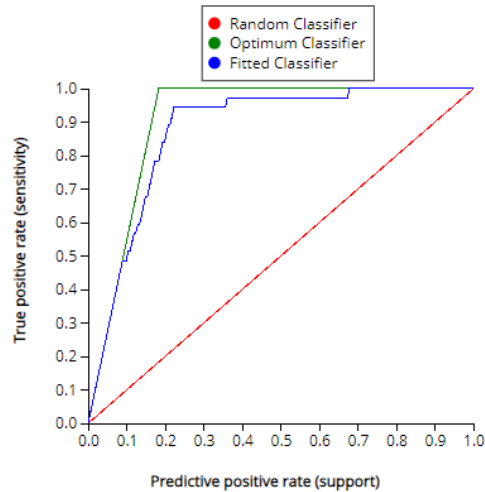
In the training portion, there are no errors when predicting. The ROC curve and Lift Chart have perfect fitting to the optimum classifier. The decile chart show that the model is highly accurate for classifying around 20% of the data but not as well with the rest of the data.

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	155	10	
1	2	35	

Error Report			
Class	# Cases	# Errors	% Error
0	165	10	6.060606
1	37	2	5.405405
Overall	202	12	5.940594

Metrics	
Metric	Value
Accuracy (#correct)	190
Accuracy (%correct)	94.059406
Specificity	0.9393939
Sensitivity (Recall)	0.9459459
Precision	0.7777778
F1 score	0.8536585
Success Class	1
Success Probability	0.5



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 classifies 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.

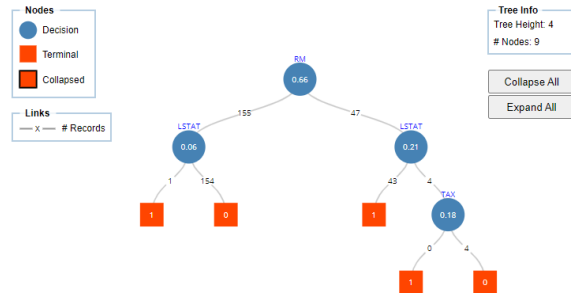
The RROC curve shows the sensitivity of detecting true positives across the range of specificities. The optimum has a 1 sensitivity across the range as it identifies all positive cases without any false negatives.

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-

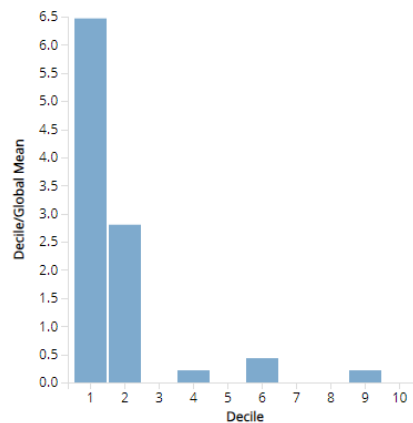
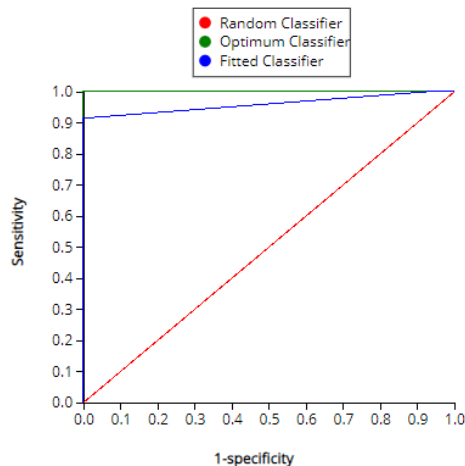
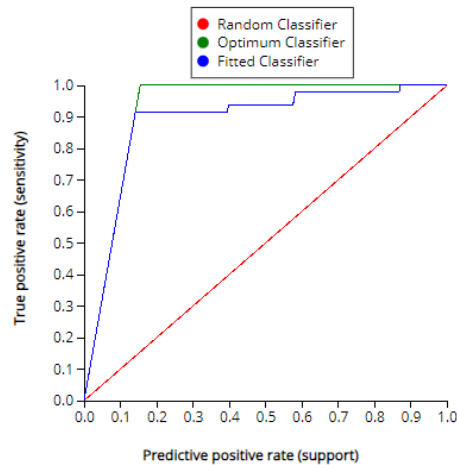


Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	257	0
1	4	43

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	4	8.510638
Overall	304	4	1.315789

Metrics	
Metric	Value
Accuracy (#correct)	300
Accuracy (%correct)	98.684211
Specificity	1
Sensitivity (Recall)	0.9148936
Precision	1
F1 score	0.9555556
Success Class	1
Success Probability	0.5



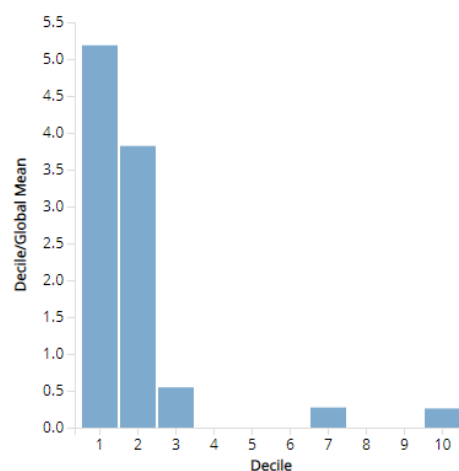
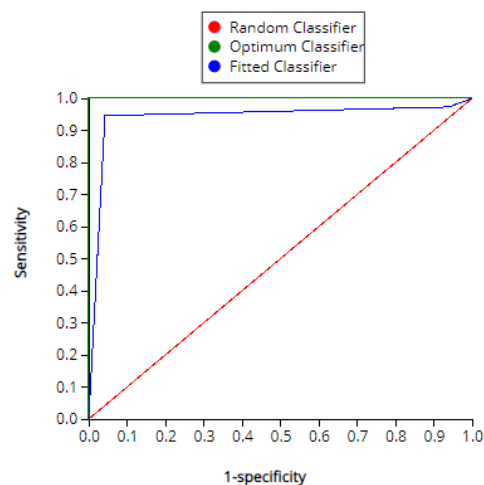
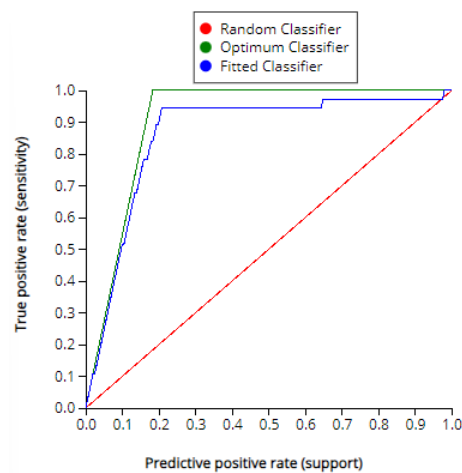
The Training data shows high accuracy (98.7%) and specificity at 1. The error comes from the sensitivity 0.915 where there are 4 error values predicted 0 when they were actually 1. The RROC curve and lift chart show the model to have a close fit to the optimum classifier. The Decile chart shows that 20% of the data is easy to classify compared to the other 80%.

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	158	7	
1	2	35	

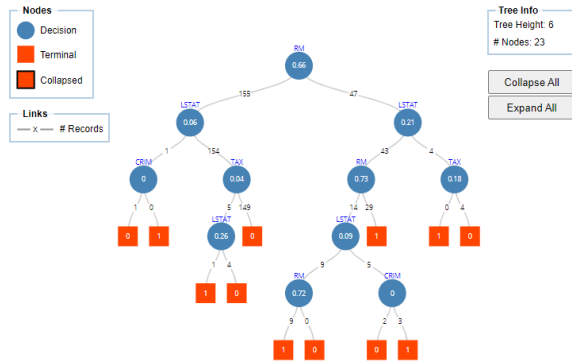
Error Report			
Class	# Cases	# Errors	% Error
0	165	7	4.242424
1	37	2	5.405405
Overall	202	9	4.455446

Metrics	
Metric	Value
Accuracy (#correct)	193
Accuracy (%correct)	95.544554
Specificity	0.9575758
Sensitivity (Recall)	0.9459459
Precision	0.8333333
F1 score	0.8860759
Success Class	1
Success Probability	0.5



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-

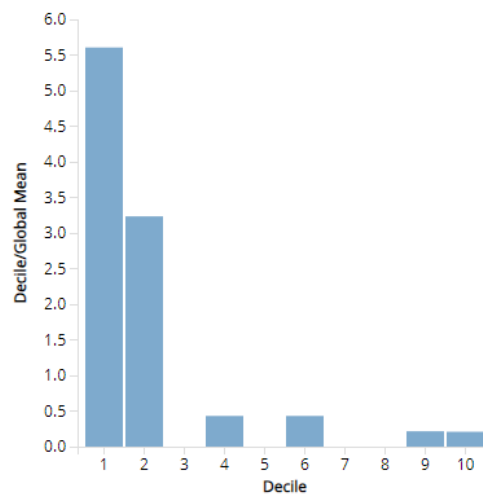
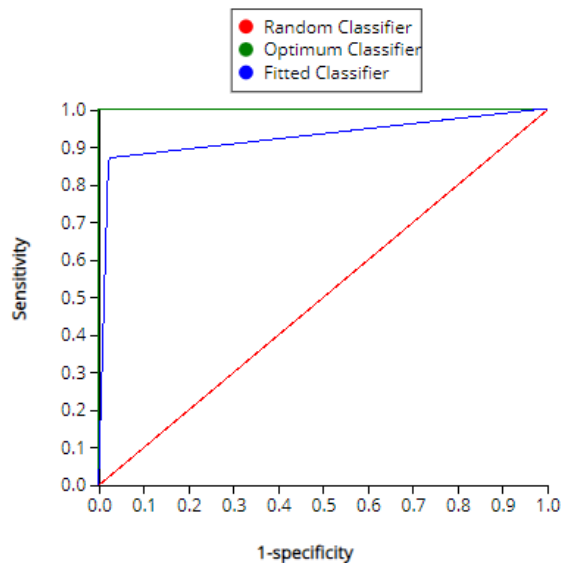
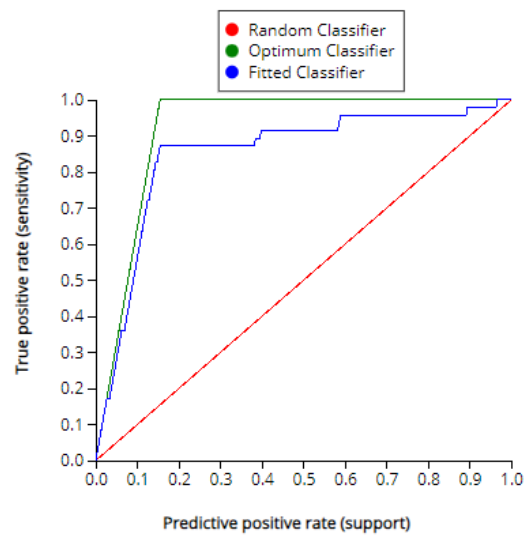


Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	251	6	
1	6	41	

Error Report			
Class	# Cases	# Errors	% Error
0	257	6	2.33463035
1	47	6	12.76595745
Overall	304	12	3.947368421

Metrics	
Metric	Value
Accuracy (#correct)	292
Accuracy (%correct)	96.05263158
Specificity	0.976653696
Sensitivity (Recall)	0.872340426
Precision	0.872340426
F1 score	0.872340426
Success Class	1
Success Probability	0.5



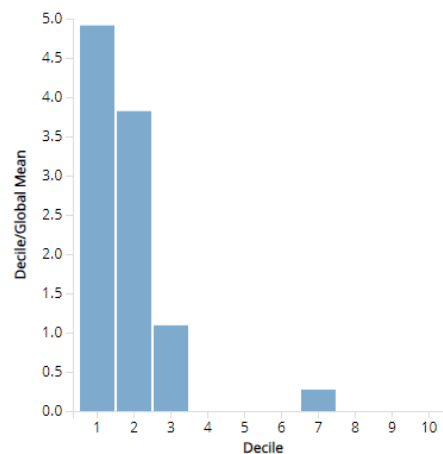
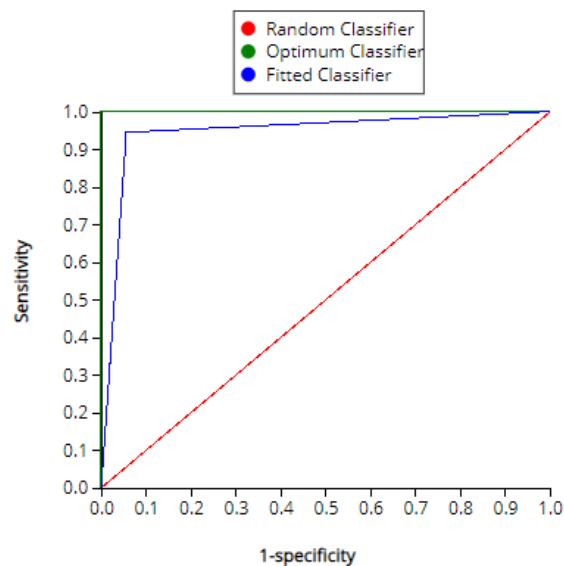
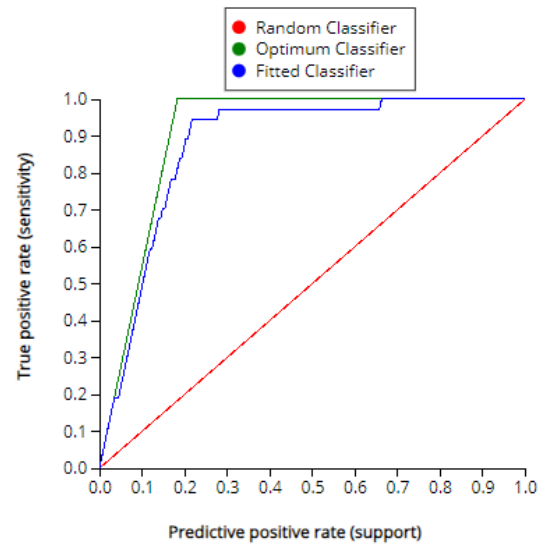
The minimum error tree training data has an accuracy of 96% with a specificity of 0.977 and a sensitivity of 0.872. The RROC curve and lift chart have ok fitting to the optimum and the decile chart shows 20% of the data is easy to classify.

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	156	9
1	2	35

Error Report			
Class	# Cases	# Errors	% Error
0	165	9	5.454545455
1	37	2	5.405405405
Overall	202	11	5.445445454

Metrics	
Metric	Value
Accuracy (#correct)	191
Accuracy (%correct)	94.554454545
Specificity	0.945454545
Sensitivity (Recall)	0.945945946
Precision	0.795454545
F1 score	0.864197531
Success Class	1
Success Probability	0.5



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

NetID	# Hidden Layers	# Neurons (Layer)	# Neurons (Layer)	Training # Error	Training % Error	Training % Sensitivity	Training % Specificity	Training % Precision	Training % F1-Score	Validation # Error	Validation % Error	Validation % Sensitivity	Validation % Specificity
Net 1	1	1	0	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 2	1	2	0	193	63.48684211	95.74468	25.68093	19.0678	31.80212014	121	59.90099	100	26.66667

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

NetID	# Hidden Layers	# Neurons (Layer)	# Neurons (Layer)	Training # Error	Training % Error	Training % Sensitivity	Training % Specificity	Training % Precision	Training % F1-Score	Validation # Error	Validation % Error	Validation % Sensitivity	Validation % Specificity
Net 1	1	1	0	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 2	1	2	0	38	12.5	82.97872	88.32685	56.52174	67.24137931	27	13.36634	100	83.63636

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

NetID	# Hidden Layers	# Neurons (Layer)	# Neurons (Layer)	Training # Error	Training % Error	Training % Sensitivity	Training % Specificity	Training % Precision	Training % F1-Score	Validation # Error	Validation % Error	Validation % Sensitivity	Validation % Specificity
Net 1	1	1	0	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 2	1	2	0	66	21.71052632	89.3617	76.26459	40.7767	56	47	23.26733	100	71.51515
Net 3	1	3	0	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 4	1	4	0	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 5	1	5	0	146	48.02631579	70.21277	48.63813	20	31.13207547	89	44.05941	91.89189	47.87879
Net 6	2	1	1	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 7	2	1	2	47	15.46052632	0	100	#N/A	#N/A	37	18.31683	0	100
Net 8	2	1	3	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 9	2	1	4	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 10	2	2	1	257	84.53947368	100	0	15.46053	26.78062678	165	81.68317	100	0
Net 11	2	2	2	27	8.881578947	59.57447	96.88716	77.77778	67.46987952	14	6.930693	78.37838	96.36364

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.

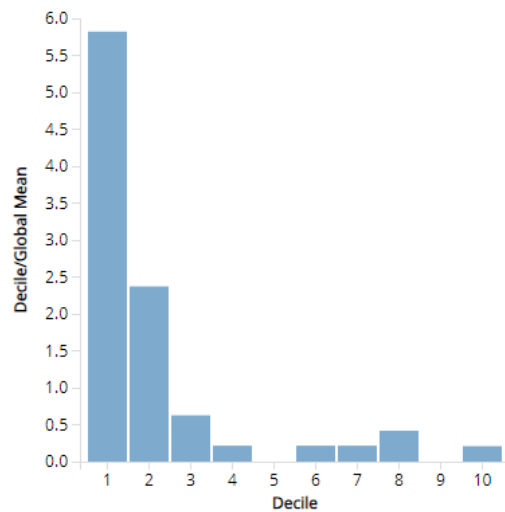
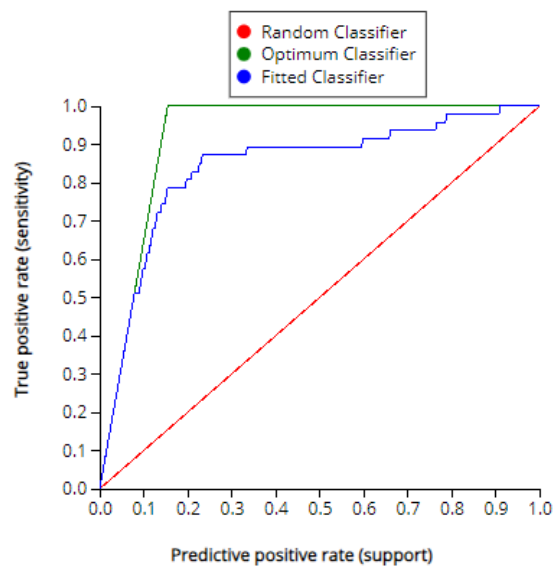
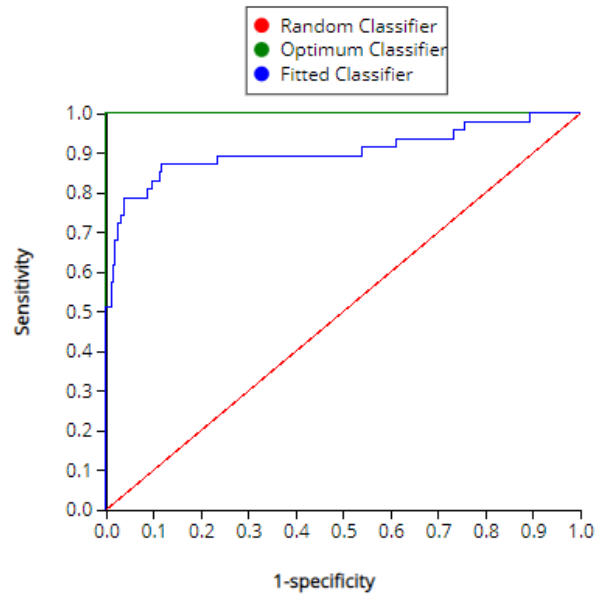
These models have shown that consistently a network with 1 layer with 2 neurons would be the best net to have.

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	248	9	
1	12	35	

Error Report			
Class	# Cases	# Errors	% Error
0	257	9	3.501946
1	47	12	25.53191
Overall	304	21	6.907895

Metrics	
Metric	Value
Accuracy (#correct)	283
Accuracy (%correct)	93.092105
Specificity	0.9649805
Sensitivity (Recall)	0.7446809
Precision	0.7954545
F1 score	0.7692308
Success Class	1
Success Probability	0.24



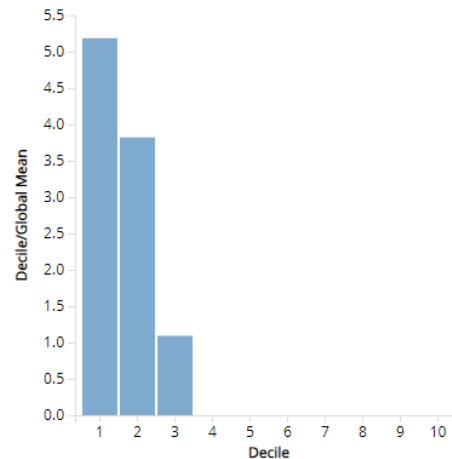
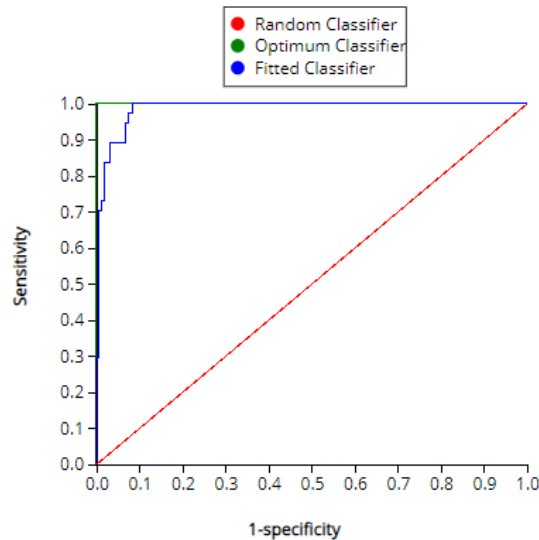
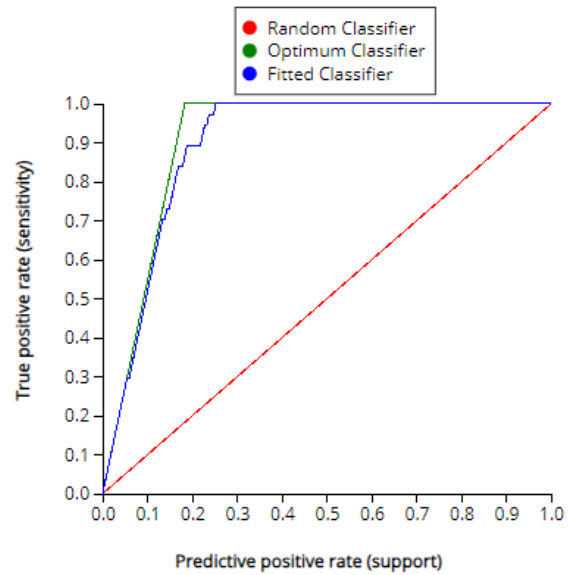
The training data shows an accuracy of 93% with a specificity of 0.965 and sensitivity of 0.745. The RROC curve and Lift chart show moderate fitting but not great.

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	153	12	
1	1	36	

Error Report			
Class	# Cases	# Errors	% Error
0	165	12	7.272727
1	37	1	2.702703
Overall	202	13	6.435644

Metrics	
Metric	Value
Accuracy (#correct)	189
Accuracy (%correct)	93.564356
Specificity	0.9272727
Sensitivity (Recall)	0.972973
Precision	0.75
F1 score	0.8470588
Success Class	1
Success Probability	0.24



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:

Record ID	Prediction: CAT. MEDV	PostProb: 0	PostProb: 1
Record 1	0	0.975903614	0.024096386
Record 2	0	0.975903614	0.024096386
Record 3	1	0.12195122	0.87804878
Record 4	0	0.975903614	0.024096386
Record 5	0	1	0
Record 6	0	0.975903614	0.024096386
Record 7	0	0.975903614	0.024096386
Record 8	1	0.12195122	0.87804878
Record 9	0	0.975903614	0.024096386
Record 10	0	0.975903614	0.024096386

Neural Network Score:

Record ID	Prediction: CAT. MEDV	PostProb: 0	PostProb: 1
Record 1	0	0.769931195	0.230068805
Record 2	0	0.786098025	0.213901975
Record 3	0	0.730997744	0.269002256
Record 4	0	0.81648522	0.18351478
Record 5	0	0.759257411	0.240742589
Record 6	0	0.764849923	0.235150077
Record 7	0	0.788331579	0.211668421
Record 8	0	0.760695243	0.239304757
Record 9	0	0.826552204	0.173447796
Record 10	0	0.796208084	0.203791916

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.

K	% Misclassification
1	5.940594059
2	9.405940594
3	6.435643564
4	6.435643564
5	5.445544554
6	5.940594059
7	6.435643564
8	4.95049505
9	5.940594059
10	5.445544554

Note: Scoring will be done using K=8

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	253	4
1	13	34

Error Report			
Class	# Cases	# Errors	% Error
0	257	4	1.55642
1	47	13	27.65957
Overall	304	17	5.592105

Metrics	
Metric	Value
Accuracy (#correct)	287
Accuracy (%correct)	94.407895
Specificity	0.9844358
Sensitivity (Recall)	0.7234043
Precision	0.8947368
F1 score	0.8
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	160	5
1	5	32

Error Report			
Class	# Cases	# Errors	% Error
0	165	5	3.030303
1	37	5	13.51351
Overall	202	10	4.950495

Metrics	
Metric	Value
Accuracy (#correct)	192
Accuracy (%correct)	95.049505
Specificity	0.969697
Sensitivity (Recall)	0.8648649
Precision	0.8648649
F1 score	0.8648649
Success Class	1
Success Probability	0.5

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.

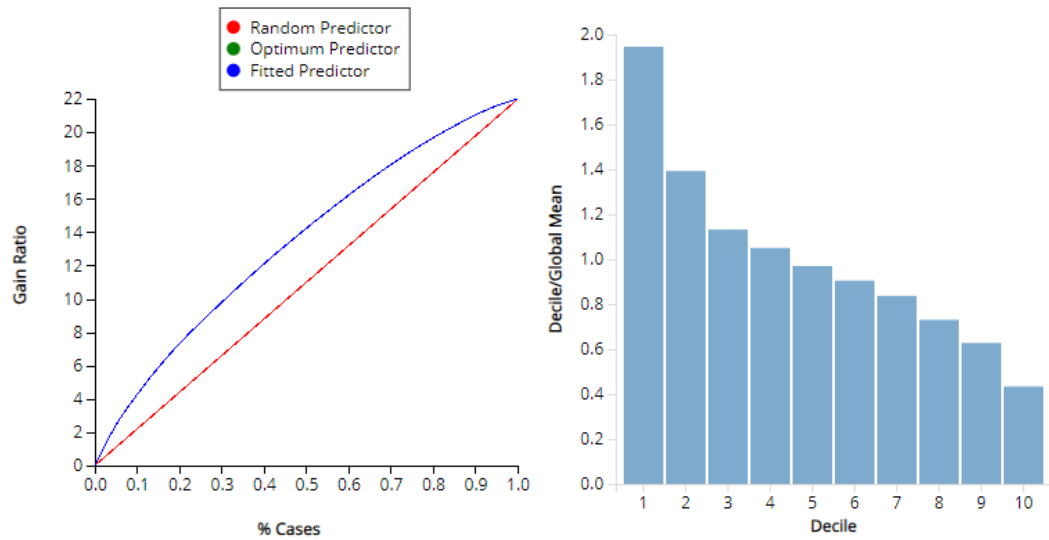
kNN Model Prediction

K	RMSE
1	4.597815
2	3.652861
3	3.30803
4	3.157657
5	3.10972
6	3.130371
7	3.101573
8	3.106233
9	3.084421
10	3.08158

Note: Scoring will be done using K=10

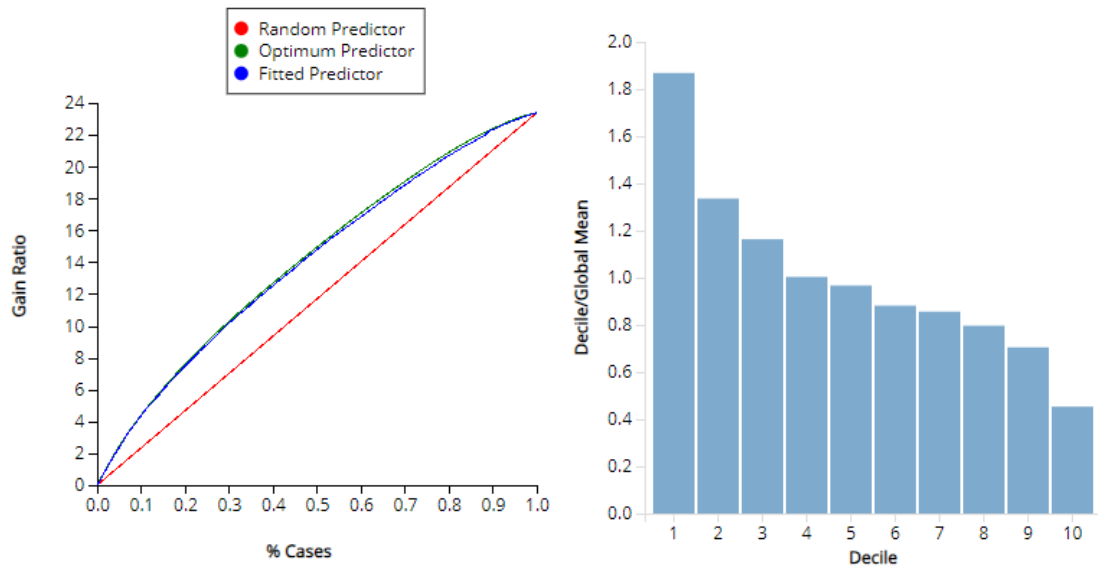
Training: Prediction Summary

Metric	Value
SSE	0
MSE	0
RMSE	0
MAD	0
R2	1



Validation: Prediction Summary

Metric	Value
SSE	1918.219
MSE	9.496136
RMSE	3.08158
MAD	2.326975
R2	0.887625



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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	257	0
1	0	47

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	0	0
Overall	304	0	0

Metrics	
Metric	Value
Accuracy (#correct)	304
Accuracy (%correct)	100
Specificity	1
Sensitivity (Recall)	1
Precision	1
F1 score	1
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	160	5
1	1	36

Error Report			
Class	# Cases	# Errors	% Error
0	165	5	3.030303
1	37	1	2.702703
Overall	202	6	2.970297

Metrics	
Metric	Value
Accuracy (#correct)	196
Accuracy (%correct)	97.029703
Specificity	0.969697
Sensitivity (Recall)	0.972973
Precision	0.8780488
F1 score	0.9230769
Success Class	1
Success Probability	0.5

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

Validation: Classification Summary

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	255	2	
1	2	45	

Error Report			
Class	# Cases	# Errors	% Error
0	257	2	0.77821
1	47	2	4.255319
Overall	304	4	1.315789

Metrics	
Metric	Value
Accuracy (#correct)	300
Accuracy (%correct)	98.684211
Specificity	0.9922179
Sensitivity (Recall)	0.9574468
Precision	0.9574468
F1 score	0.9574468
Success Class	1
Success Probability	0.5

Confusion Matrix			
Actual\Predicted	0	1	
0	159	6	
1	2	35	

Error Report			
Class	# Cases	# Errors	% Error
0	165	6	3.636364
1	37	2	5.405405
Overall	202	8	3.960396

Metrics	
Metric	Value
Accuracy (#correct)	194
Accuracy (%correct)	96.039604
Specificity	0.9636364
Sensitivity (Recall)	0.9459459
Precision	0.8536585
F1 score	0.8974359
Success Class	1
Success Probability	0.5

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

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	255	2	
1	2	45	

Error Report			
Class	# Cases	# Errors	% Error
0	257	2	0.77821
1	47	2	4.255319
Overall	304	4	1.315789

Metrics	
Metric	Value
Accuracy (#correct)	300
Accuracy (%correct)	98.684211
Specificity	0.9922179
Sensitivity (Recall)	0.9574468
Precision	0.9574468
F1 score	0.9574468
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	160	5	
1	2	35	

Error Report			
Class	# Cases	# Errors	% Error
0	165	5	3.030303
1	37	2	5.405405
Overall	202	7	3.465347

Metrics	
Metric	Value
Accuracy (#correct)	195
Accuracy (%correct)	96.534653
Specificity	0.969697
Sensitivity (Recall)	0.9459459
Precision	0.875
F1 score	0.9090909
Success Class	1
Success Probability	0.5

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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	257	0
1	47	0

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	47	100
Overall	304	47	15.46053

Metrics	
Metric	Value
Accuracy (#correct)	257
Accuracy (%correct)	84.539474
Specificity	1
Sensitivity (Recall)	0
Precision	#N/A
F1 score	#N/A
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	165	0
1	37	0

Error Report			
Class	# Cases	# Errors	% Error
0	165	0	0
1	37	37	100
Overall	202	37	18.31683

Metrics	
Metric	Value
Accuracy (#correct)	165
Accuracy (%correct)	81.683168
Specificity	1
Sensitivity (Recall)	0
Precision	#N/A
F1 score	#N/A
Success Class	1
Success Probability	0.5

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

k-Nearest Neighbors

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	257	0
1	0	47

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	0	0
Overall	304	0	0

Metrics	
Metric	Value
Accuracy (#correct)	304
Accuracy (%correct)	100
Specificity	1
Sensitivity (Recall)	1
Precision	1
F1 score	1
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	158	7
1	6	31

Error Report			
Class	# Cases	# Errors	% Error
0	165	7	4.242424
1	37	6	16.21622
Overall	202	13	6.435644

Metrics	
Metric	Value
Accuracy (#correct)	189
Accuracy (%correct)	93.564356
Specificity	0.9575758
Sensitivity (Recall)	0.8378378
Precision	0.8157895
F1 score	0.8266667
Success Class	1
Success Probability	0.5

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

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	250	7	
1	9	38	

Error Report			
Class	# Cases	# Errors	% Error
0	257	7	2.723735
1	47	9	19.14894
Overall	304	16	5.263158

Metrics	
Metric	Value
Accuracy (#correct)	288
Accuracy (%correct)	94.736842
Specificity	0.9727626
Sensitivity (Recall)	0.8085106
Precision	0.8444444
F1 score	0.826087
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	159	6	
1	1	36	

Error Report			
Class	# Cases	# Errors	% Error
0	165	6	3.636364
1	37	1	2.702703
Overall	202	7	3.465347

Metrics	
Metric	Value
Accuracy (#correct)	195
Accuracy (%correct)	96.534653
Specificity	0.9636364
Sensitivity (Recall)	0.972973
Precision	0.8571429
F1 score	0.9113924
Success Class	1
Success Probability	0.5

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

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	253	4	
1	13	34	

Error Report			
Class	# Cases	# Errors	% Error
0	257	4	1.55642
1	47	13	27.65957
Overall	304	17	5.592105

Metrics	
Metric	Value
Accuracy (#correct)	287
Accuracy (%correct)	94.407895
Specificity	0.9844358
Sensitivity (Recall)	0.7234043
Precision	0.8947368
F1 score	0.8
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	160	5	
1	5	32	

Error Report			
Class	# Cases	# Errors	% Error
0	165	5	3.030303
1	37	5	13.51351
Overall	202	10	4.950495

Metrics	
Metric	Value
Accuracy (#correct)	192
Accuracy (%correct)	95.049505
Specificity	0.969697
Sensitivity (Recall)	0.8648649
Precision	0.8648649
F1 score	0.8648649
Success Class	1
Success Probability	0.5

Boosting

Training: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	257	0	
1	0	47	

Error Report			
Class	# Cases	# Errors	% Error
0	257	0	0
1	47	0	0
Overall	304	0	0

Metrics	
Metric	Value
Accuracy (#correct)	304
Accuracy (%correct)	100
Specificity	1
Sensitivity (Recall)	1
Precision	1
F1 score	1
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix			
Actual\Predicted	0	1	
0	160	5	
1	1	36	

Error Report			
Class	# Cases	# Errors	% Error
0	165	5	3.030303
1	37	1	2.702703
Overall	202	6	2.970297

Metrics	
Metric	Value
Accuracy (#correct)	196
Accuracy (%correct)	97.029703
Specificity	0.969697
Sensitivity (Recall)	0.972973
Precision	0.8780488
F1 score	0.9230769
Success Class	1
Success Probability	0.5

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

Training: Prediction Summary Validation: Prediction Summary

Metric ▾	Value ▾
SSE	0
MSE	0
RMSE	0
MAD	0
R2	1.

Metric ▾	Value ▾
SSE	1918.219
MSE	9.496136
RMSE	3.08158
MAD	2.326975
R2	0.887625

Boosting

Training: Prediction Summary Validation: Prediction Summary

Metric ▾	Value ▾
SSE	301.6263
MSE	0.992192
RMSE	0.996088
MAD	0.769735
R2	0.988127

Metric ▾	Value ▾
SSE	2846.274
MSE	14.09046
RMSE	3.753727
MAD	2.470048
R2	0.833257

The kNN prediction 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.

Scoring

Record ID	Prediction: MEDV
Record 1	21.37964311
Record 2	15.65530338
Record 3	45.78970677
Record 4	18.58325332
Record 5	30.1792256
Record 6	24.24235996
Record 7	19.8178018
Record 8	40.55951807
Record 9	24.01396957
Record 10	19.46511759

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.

Scoring

Record ID	Prediction: CAT. MEDV	PostProb: 0	PostProb: 1
Record 1	0	1	0
Record 2	0	1	0
Record 3	1	0	1
Record 4	0	1	0
Record 5	0	0.647219535	0.352780465
Record 6	0	1	0
Record 7	0	1	0
Record 8	1	0.169089623	0.830910377
Record 9	0	1	0
Record 10	0	1	0

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.

