HOMEWORK 9

1. Read the article by Domingos: A few useful things to know about machine learning (communications of the ACM, Vol. 55 No. 10, Pages 78-87 doi: 10.1145/2347736.2347755 via ACM Digital library,https://courses.cs.ut.ee/MTAT.03.183/2012_fall/uploads/Main/domingos.pdf). Make a list of key messages with a supporting 1-2 sentence example or clari cation of that message (something like short summary of the article)

The article in the link above focuses mainly on classification. In order to understand any of the existing algorithms it is important to keep in mind the three major cardinal points of all of the existing algorithm. The cardinal points are as follows

1. Representation: This refers to the data being presented in a formal language.
2. Evaluation: This leans towards the possibility to separate good and bad results.
3. Optimization: optimization refers to the fact that the algorithm should be quick as much as possible.

It is also important to note that test dat is not equal to real data (This has also been highlighted in several other articles I have come across.

The article also talks about overfitting. This (overfitting) seems to be a common problem in machine learning. Overfitting usually occurs when we have what we can term the “ideal” test data. In such situation it is advisable to add some “noise data”.

Moving on from overfitting, another huge problem in data mining is known as the “curse of dimensionality”. Curse of dimensionality simply put is said to be the exponential growth of data in problems of large dimension however there is a counter effect which partly counteracts the curse and could be called “blessing of non-uniformity.

In regards to theoretical guarantees, I think every data scientist should take them with a grain of salt. The author was quite explanatory with his concluding statement that goes thus “The main role of theoretical guarantees in machine learning is not as a criterion for practical decisions, but as a source of understanding and driving force for algorithm design.”

Other Points worth noting

Machine learning is not one stop process but rather a continuous process with new data and the occurrence of minor changes in each cycle.

If you aren’t satisfied with the results of learning then try to process more data as this is cost effective when compare to towing the path of algorithm optimization.

Explore different models or group of models as this will help you to identify and choose the best one (similar to what we did in Homework 8 where we had to compare different classifiers).

Contrary to intuition, a more powerful learner is not necessarily better than a less powerful one as powerful learners can be unstable but still accurate.

2. Draw the ROC curves and calculate the ROC AUC for 4 classifiers based on the following data Attach:roc_data.zip. The data.class is the true class, and the roc1.txt etc are the orders in which different classifiers would classify examples as positive (so some are true positive, some false positive; after certain cutoff there remain false negatives and true negatives).

Roc1.txt

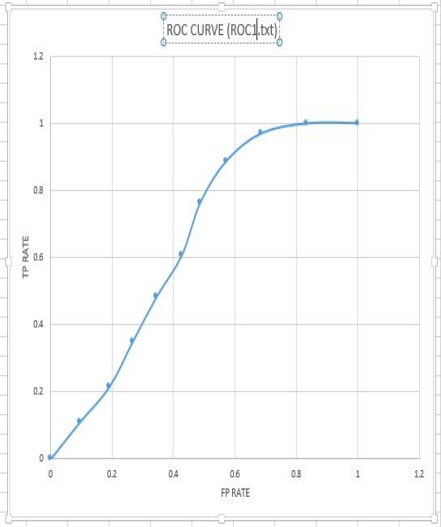
ACTUAL CLASS	
TRUE VALUES	1215
FALSE VALUES	1785
TOTAL	3000

		PREDICTED CLASS		
		TRUE	FALSE	TOTAL
ACTUAL CLASS	TRUE	1179	36	1215
	FALSE	1521	264	1785

PREDICTED CLASS			
		TRUE	FALSE
ACTUAL CLASS	TRUE	TP	FN
	FALSE	FP	TN

SENSITIVITY	TP/(TP+FN)
SPECIFICITY	TN/(TN+FP)
TN RATE	SPECIFICITY
TP RATE	SENSITIVITY
FP RATE	FP/(TN+FP)

	PREDICTION MODEL								
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX
0	1215	0	1785	0	0	1	1	0.099958848	0
0.1	1214	299	1486	1	0.167507003	0.832492997	0.999176955	0.098436214	0.166683958
0.2	1178	563	1222	37	0.315406162	0.684593838	0.969547325	0.092880658	0.284953488
0.3	1079	764	1021	136	0.428011204	0.571988796	0.888065844	0.082674897	0.316077048
0.4	930	915	870	285	0.512605042	0.487394958	0.765432099	0.068641975	0.278037141
0.5	738	1023	762	477	0.573109244	0.426890756	0.607407407	0.054485597	0.180516651
0.6	586	1171	614	629	0.656022409	0.343977591	0.482304527	0.041563786	0.138326936
0.7	424	1309	476	791	0.733333333	0.266666667	0.348971193	0.028148148	0.082304527
0.8	260	1445	340	955	0.80952381	0.19047619	0.21399177	0.016131687	0.023515579
0.9	132	1617	168	1083	0.905882353	0.094117647	0.108641975	0.005432099	0.014524328
1	0	1785	0	1215	1	0	0	0	0
								0.588353909	AUC



AUC = 0.588353909

Roc2.txt

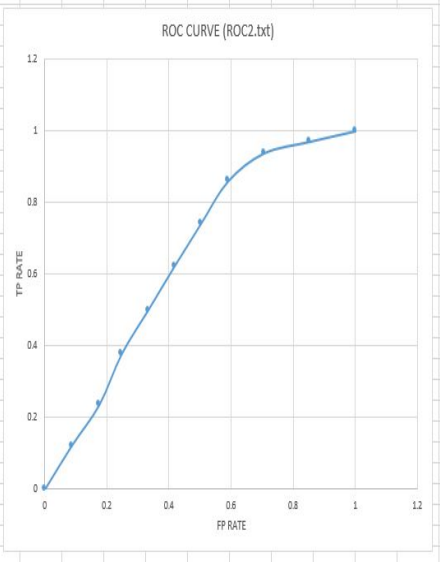
ACTUAL CLASS	
TRUE VALUES	1215
FALSE VALUES	1785
TOTAL	3000

		PREDICTED CLASS		
		TRUE	FALSE	TOTAL
ACTUAL CLASS	TRUE	1179	36	1215
	FALSE	1521	264	1785

PREDICTED CLASS			
		TRUE	FALSE
ACTUAL CLASS	TRUE	TP	FN
	FALSE	FP	TN

SENSITIVITY	TP/(TP+FN)
SPECIFICITY	TN/(TN+FP)
TN RATE	SPECIFICITY
TP RATE	SENSITIVITY
FP RATE	FP/(TN+FP)

	PREDICTION MODEL							
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES
0	1215	0	1785	0	0	1	1	0.098518519
0.1	1179	264	1521	36	0.14789916	0.85210084	0.97037037	0.095432099
0.2	1140	525	1260	75	0.294117647	0.705882353	0.938271605	0.089958848
0.3	1046	731	1054	169	0.40952381	0.59047619	0.86090535	0.080164609
0.4	902	887	898	313	0.496918768	0.503081232	0.742386831	0.068230453
0.5	756	1041	744	459	0.583193277	0.416806723	0.622222222	0.056090535
0.6	607	1192	593	608	0.667787115	0.332212885	0.499588477	0.043868313
0.7	459	1344	441	756	0.752941176	0.247058824	0.377777778	0.030699588
0.8	287	1472	313	928	0.82464986	0.17535014	0.236213992	0.017860082
0.9	147	1632	153	1068	0.914285714	0.085714286	0.120987654	0.006049383
1	0	1785	0	1215	1	0	0	0
								0.586872428 AUC



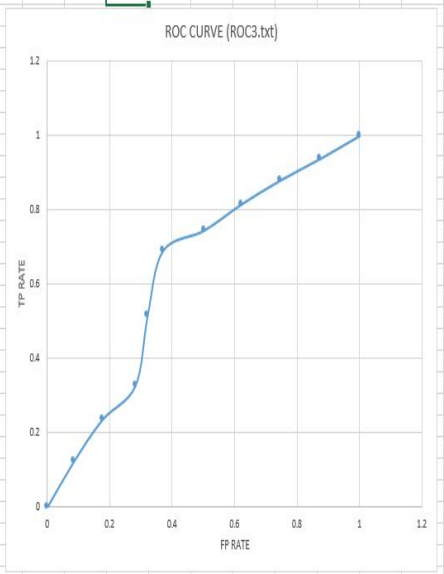
AUC = 0.586872428

Roc3.txt

ACTUAL CLASS		
	TRUE VALUES	1215
	FALSE VALUES	1785
	TOTAL	3000

	PREDICTED CLASS		
	TRUE	FALSE	TOTAL
TRUE	1179	36	1215
ACTUAL CLASS FALSE	1521	264	1785

PREDICTED CLASS				
		TRUE	FALSE	
	TRUE	TP	FN	SENSITIVITY TP/(TP+FN)
ACTUAL CLASS	FALSE	FP	TN	SPECIFICITY TN/(TN+FP)
				TN RATE SPECIFICTY
				TP RATE SENSITIVITY
				FP RATE FP/(TN+FP)



	PREDICTION MODEL							
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES
0	1215	0	1785	0	0	1	1	0.09691358
0.1	1140	225	1560	75	0.12605042	0.87394958	0.938271605	0.09090535
0.2	1069	454	1331	146	0.254341737	0.745658263	0.879835391	0.084773663
0.3	991	676	1109	224	0.378711485	0.621288515	0.81563786	0.078024691
0.4	905	890	895	310	0.49859944	0.50140056	0.744855967	0.071769547
0.5	839	1124	661	376	0.629691877	0.370308123	0.690534979	0.060329218
0.6	627	1212	573	588	0.678991597	0.321008403	0.516049383	0.042139918
0.7	397	1282	503	818	0.718207283	0.281792717	0.326748971	0.028106996
0.8	286	1471	314	929	0.824089636	0.175910364	0.235390947	0.017901235
0.9	149	1634	151	1066	0.915406162	0.084593838	0.122633745	0.006131687
1	0	1785	0	1215	1	0	0	0
								0.576995885 AUC

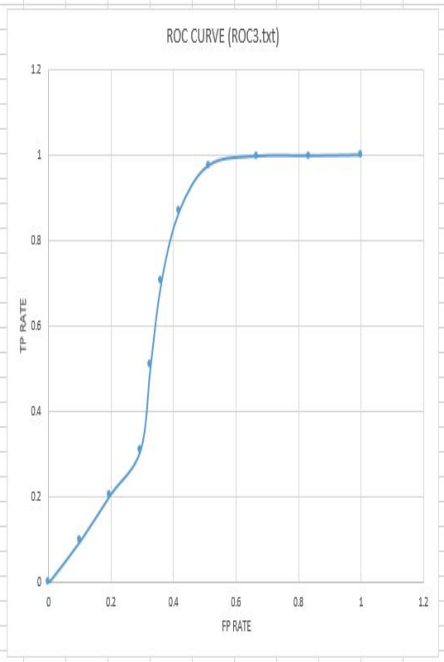
AUC = 0.576995885

Roc4.txt

ACTUAL CLASS		
	TRUE VALUES	1215
	FALSE VALUES	1785
	TOTAL	3000

	PREDICTED CLASS		
	TRUE	FALSE	TOTAL
TRUE	1179	36	1215
ACTUAL CLASS FALSE	1521	264	1785

PREDICTED CLASS				
		TRUE	FALSE	
	TRUE	TP	FN	SENSITIVITY TP/(TP+FN)
ACTUAL CLASS	FALSE	FP	TN	SPECIFICITY TN/(TN+FP)
				TN RATE SPECIFICTY
				TP RATE SENSITIVITY
				FP RATE FP/(TN+FP)



	PREDICTION MODEL							
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES
0	1215	0	1785	0	0	1	1	0.099917695
0.1	1213	298	1487	2	0.166946779	0.833053221	0.998353909	0.099794239
0.2	1212	597	1188	3	0.334453782	0.665546218	0.997530864	0.098683128
0.3	1186	871	914	29	0.487955182	0.512044818	0.976131687	0.092222222
0.4	1055	1040	745	160	0.582633053	0.417366947	0.868312757	0.078641975
0.5	856	1141	644	359	0.639215686	0.360784314	0.704526749	0.060658436
0.6	618	1203	582	597	0.67394958	0.32605042	0.508641975	0.04090535
0.7	376	1261	524	839	0.706442577	0.293557423	0.309465021	0.025720165
0.8	249	1434	351	966	0.803361345	0.196638655	0.204938272	0.015144033
0.9	119	1604	181	1096	0.89859944	0.10140056	0.097942387	0.004897119
1	0	1785	0	1215	1	0	0	0
								0.616584362 AUC

AUC = 0.616584362

3. Characterize the behavior of the 4 classifiers in task 2. Also, provide the "best" cutoff for each of the classifiers.

Roc1.txt

	PREDICTION MODEL								
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX
0	1215	0	1785	0	0	1	1	0.099958848	0
0.1	1214	299	1486	1	0.167507003	0.832492997	0.999176955	0.098436214	0.166683958
0.2	1178	563	1222	37	0.315406162	0.684593838	0.969547325	0.092880658	0.284953488
0.3	1079	764	1021	136	0.428011204	0.571988796	0.888065844	0.082674897	0.316077048
0.4	930	915	870	285	0.512605042	0.487394958	0.765432099	0.068641975	0.278037141
0.5	738	1023	762	477	0.573109244	0.426890756	0.607407407	0.054485597	0.180516651
0.6	586	1171	614	629	0.656022409	0.343977591	0.482304527	0.041563786	0.138326936
0.7	424	1309	476	791	0.733333333	0.266666667	0.348971193	0.028148148	0.082304527
0.8	260	1445	340	955	0.80952381	0.19047619	0.21399177	0.016131687	0.023515579
0.9	132	1617	168	1083	0.905882353	0.094117647	0.108641975	0.005432099	0.014524328
1	0	1785	0	1215	1	0	0	0	0
								0.588353909	AUC

The classifier that gave the output of roc1.txt had the second highest AUC rate hence ranked second best in terms of the performance of the four classifiers. From the Youden's index the best cut off point for the classifier used for roc1.txt data is;

Cutoff point: 0.3 (30%)

Roc2.txt

	PREDICTION MODEL							
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	YOUDEN INDEX
0	1215	0	1785	0	0	1	1	0
0.1	1179	264	1521	36	0.14789916	0.85210084	0.97037037	0.11826953
0.2	1140	525	1260	75	0.294117647	0.705882353	0.938271605	0.232389252
0.3	1046	731	1054	169	0.40952381	0.59047619	0.86090535	0.27042916
0.4	902	887	898	313	0.496918768	0.503081232	0.742386831	0.239305599
0.5	756	1041	744	459	0.583193277	0.416806723	0.622222222	0.205415499
0.6	607	1192	593	608	0.667787115	0.332212885	0.499588477	0.167375592
0.7	459	1344	441	756	0.752941176	0.247058824	0.377777778	0.130718954
0.8	287	1472	313	928	0.82464986	0.17535014	0.236213992	0.060863852
0.9	147	1632	153	1068	0.914285714	0.085714286	0.120987654	0.035273368
1	0	1785	0	1215	1	0	0	0

The classifier that gave the output of roc2.txt had the second lowest AUC rate hence ranked second from behind in terms of the performance of the four classifiers however the performance wasn't really bad when compared to that of roc1.txt as the difference in AUC was less than 0.01. From the Youden's index the best cut off point for the classifier used for roc2.txt data is;

Cutoff point: 0.3 (30%)

Roc3.txt

	PREDICTION MODEL								
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	
									YOUDEN INDEX
0	1215	0	1785	0	0	1	1	0.09691358	0
0.1	1140	225	1560	75	0.12605	0.87395	0.93827161	0.09090535	0.064322025
0.2	1069	454	1331	146	0.254342	0.745658	0.87983539	0.084773663	0.134177128
0.3	991	676	1109	224	0.378711	0.621289	0.81563786	0.078024691	0.194349345
0.4	905	890	895	310	0.498599	0.501401	0.74485597	0.071769547	0.243455407
0.5	839	1124	661	376	0.629692	0.370308	0.69053498	0.060329218	0.320226856
0.6	627	1212	573	588	0.678992	0.321008	0.51604938	0.042139918	0.19504098
0.7	397	1282	503	818	0.718207	0.281793	0.32674897	0.028106996	0.044956254
0.8	286	1471	314	929	0.82409	0.17591	0.23539095	0.017901235	0.059480583
0.9	149	1634	151	1066	0.915406	0.084594	0.12263375	0.006131687	0.038039907
1	0	1785	0	1215	1	0	0	0	0
								0.576995885	AUC

The classifier that gave the output of roc3.txt had the worst performance amongst the four classifiers as can be seen from the AUC.This is also visible in the difference in cut off points between it and the rest classifier. From the Youden's index the best cut off point for the classifier used for roc4.txt data is;

Cutoff point: 0.5 (50%)

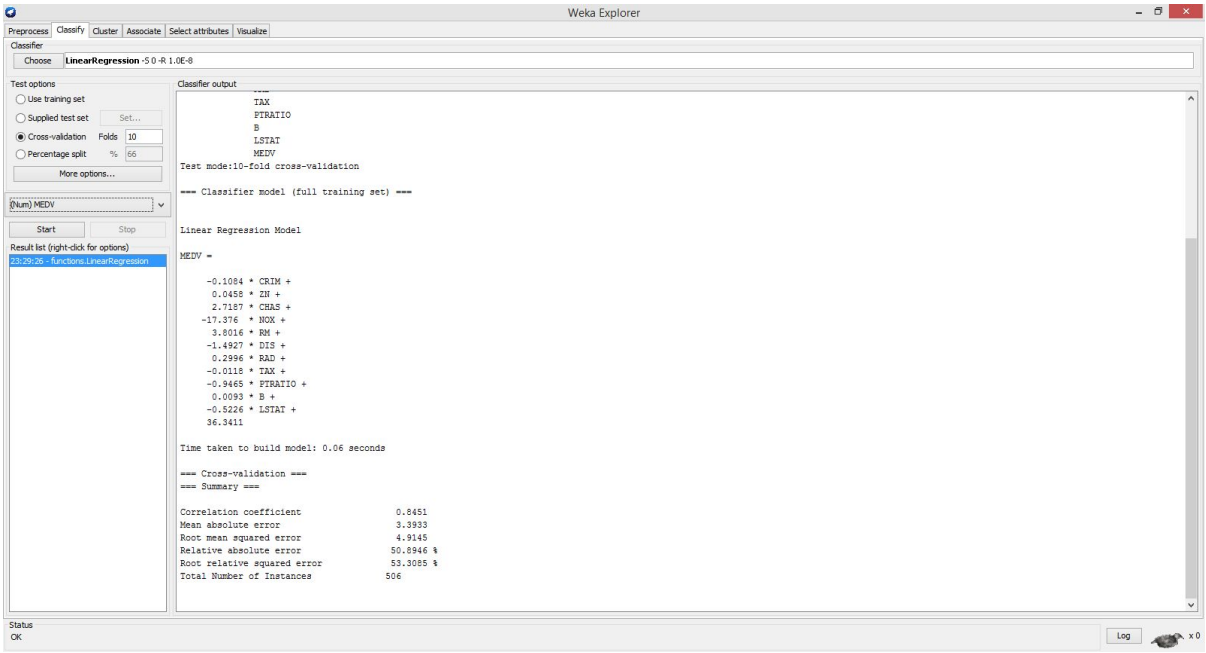
Roc4.txt

	PREDICTION MODEL								
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX
0	1215	0	1785	0	0	1	1	0.099917695	0
0.1	1213	298	1487	2	0.166946779	0.833053221	0.998353909	0.099794239	0.165300688
0.2	1212	597	1188	3	0.334453782	0.665546218	0.997530864	0.098683128	0.331984646
0.3	1186	871	914	29	0.487955182	0.512044818	0.976131687	0.092222222	0.464086869
0.4	1055	1040	745	160	0.582633053	0.417366947	0.868312757	0.078641975	0.45094581
0.5	856	1141	644	359	0.639215686	0.360784314	0.704526749	0.060658436	0.343742435
0.6	618	1203	582	597	0.67394958	0.32605042	0.508641975	0.04090535	0.182591555
0.7	376	1261	524	839	0.706442577	0.293557423	0.309465021	0.025720165	0.015907598
0.8	249	1434	351	966	0.803361345	0.196638655	0.204938272	0.015144033	0.008299616
0.9	119	1604	181	1096	0.89859944	0.10140056	0.097942387	0.004897119	-0.003458173
1	0	1785	0	1215	1	0	0	0	0
								0.616584362	AUC

The classifier that gave the output of roc4.txt had the highest AUC rate hence is the first and best in terms of the performance of the four classifiers. From the Youden's index the best cut off point for the classifier used for roc4.txt data is;

Cutoff point: 0.3 (30%)

4. Use the data about housing (http://archive.ics.uci.edu/ml/datasets/Housing) and estimate by regression analysis the last column - report RMSE score.



=== Classifier model (full training set) ===

Linear Regression Model

MEDV =

-0.1084 * CRIM +
0.0458 * ZN +
2.7187 * CHAS +
-17.376 * NOX +
3.8016 * RM +
-1.4927 * DIS +
0.2996 * RAD +

-0.0118 * TAX +
-0.9465 * PTRATIO +
0.0093 * B +
-0.5226 * LSTAT +
36.3411
Time taken to build model: 0.06 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8451
Mean absolute error	3.3933
Root mean squared error	**4.9145**
Relative absolute error	50.8946 %
Root relative squared error	53.3085 %
Total Number of Instances	506

For this task I converted the data I converted the data into csv and loaded into WEKA via the CSV loader. I used the Linear regression model from the functions branch in WEKA over SimpleLinearRegression model because in this situation we have more than one Variable and I also used cross-validation. The final **RMSE** score for the last column (MEDV) is **4.9145** as can be seen above.

5. Estimate every variable one by one using all other attributes in this data set - report RMSE scores for each. What are the most important predictors and what are the most correlated ones? Which variables are "easier" to predict than others? If so, then why?

Linear Regression Model for CRIM variable

CRIM =
0.0479 * ZN +
-11.6094 * NOX +
-0.961 * DIS +
0.6121 * RAD +
-0.0055 * TAX +
-0.2884 * PTRATIO +
-0.008 * B +
0.1075 * LSTAT +
-0.1871 * MEDV +
20.5742
Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.6415
Mean absolute error	2.9332
Root mean squared error	**6.598**
Relative absolute error	61.234 %
Root relative squared error	76.6856 %
Total Number of Instances	506

Linear Regression Model for ZN variable

ZN =

0.2608 * CRIM +
-0.4812 * INDUS +
2.8395 * RM +
-0.1158 * AGE +
6.4508 * DIS +
-0.6495 * RAD +
0.066 * TAX +
-2.4257 * PTRATIO +
0.4525 * LSTAT +
0.4825 * MEDV +
-11.1639

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.7465
Mean absolute error	10.9076
Root mean squared error	**15.5063**
Relative absolute error	65.1671 %
Root relative squared error	66.4325 %
Total Number of Instances	506

Linear Regression Model for INDUS variable

INDUS =

-0.0245 * ZN +
1.4482 * CHAS +
16.6801 * NOX +
-0.6004 * RM +
-0.6705 * DIS +
-0.3084 * RAD +
0.0269 * TAX +
0.2762 * PTRATIO +
0.0622 * LSTAT +
-5.5469

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8581
Mean absolute error	2.4645
Root mean squared error	**3.5197**
Relative absolute error	39.6015 %
Root relative squared error	51.1579 %
Total Number of Instances	506

Linear Regression Model for CHAS variable

CHAS =

0.0058 * INDUS +
0.3974 * NOX +
-0.0002 * TAX +
0.0072 * MEDV +
-0.2859

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.1997
Mean absolute error	0.1273
Root mean squared error	**0.2496**
Relative absolute error	98.6476 %
Root relative squared error	98.1077 %
Total Number of Instances	506

Linear Regression Model for NOX variable

NOX =

-0.0008 * CRIM +
0.0045 * INDUS +
0.0009 * AGE +
-0.0175 * DIS +
0.0037 * RAD +
-0.0133 * PTRATIO +
-0.0023 * MEDV +
0.775

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8772
Mean absolute error	0.0424
Root mean squared error	**0.0556**
Relative absolute error	44.2188 %
Root relative squared error	47.9324 %
Total Number of Instances	506

Linear Regression Model for RM variable

RM =

```
  0.0028 * ZN +
 -0.014  * INDUS +
  0.0055 * AGE +
  0.0088 * RAD +
 -0.001  * B +
 -0.035  * LSTAT +
  0.0381 * MEDV +
  5.8765
```

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.7213
Mean absolute error	0.3165
Root mean squared error	**0.4866**
Relative absolute error	61.5145 %
Root relative squared error	69.1042 %
Total Number of Instances	506

Linear Regression Model for Age variable

AGE =

```
 -0.1256 * ZN +
 77.848  * NOX +
  6.5907 * RM +
 -4.4436 * DIS +
 -0.3621 * RAD +
  0.0076 * TAX +
  0.7367 * PTRATIO +
  0.0135 * B +
  1.3035 * LSTAT +
-32.2627
```

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8117
Mean absolute error	12.7975
Root mean squared error	**16.4323**
Relative absolute error	51.8417 %
Root relative squared error	58.2427 %
Total Number of Instances	506

Linear Regression Model for DIS variable

DIS =

-0.0221 * CRIM +
0.0279 * ZN +
-0.0592 * INDUS +
-5.8401 * NOX +
-0.0174 * AGE +
-0.0259 * LSTAT +
-0.0641 * MEDV +
10.4241

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8749
Mean absolute error	0.7668
Root mean squared error	**1.0188**
Relative absolute error	44.5095 %
Root relative squared error	48.2965 %
Total Number of Instances	506

Linear Regression Model for RAD variable

RAD =

0.1414 * CRIM +
-0.0273 * ZN +
-0.2501 * INDUS +
0.9754 * CHAS +
10.2517 * NOX +
0.4384 * RM +
-0.0146 * AGE +
0.1542 * DIS +
0.0445 * TAX +
0.4861 * PTRATIO +
-0.0041 * B +
0.0972 * LSTAT +
0.1355 * MEDV +
-25.9234

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9278
Mean absolute error	2.5071
Root mean squared error	**3.2457**
Relative absolute error	33.1666 %
Root relative squared error	37.2151 %
Total Number of Instances	506

Linear Regression Model for TAX variable

TAX =

```
  0.7667 * ZN +
  7.2742 * INDUS +
-22.425  * CHAS +
 56.8028 * NOX +
 14.0697 * RAD +
 -0.9296 * LSTAT +
 -1.6967 * MEDV +
204.1929
```

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9401
Mean absolute error	35.7655
Root mean squared error	**57.3815**
Relative absolute error	24.8313 %
Root relative squared error	33.9598 %
Total Number of Instances	506

Linear Regression Model for PTRATIO variable

PTRATIO =

```
 -0.0253 * ZN +
  0.0645 * INDUS +
-10.3959 * NOX +
  0.0069 * AGE +
  0.1195 * RAD +
  0.0017 * B +
 -0.0381 * LSTAT +
 -0.1056 * MEDV +
 24.4216
```

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.6802
Mean absolute error	1.2521
Root mean squared error	**1.5862**
Relative absolute error	69.8116 %
Root relative squared error	73.0791 %
Total Number of Instances	506

Linear Regression Model for B variable

B =
-1.1198 * CRIM +
-23.6043 * RM +
-3.2089 * RAD +
4.7817 * PTRATIO +
-1.6018 * LSTAT +
2.6521 * MEDV +
411.9698

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.4452
Mean absolute error	47.8831
Root mean squared error	**81.9674**
Relative absolute error	87.4073 %
Root relative squared error	89.6547 %
Total Number of Instances	506

Linear Regression Model for LSTAT variable

LSTAT =
0.0458 * CRIM +
0.0277 * ZN +
0.076 * INDUS +
-2.3238 * RM +
0.0717 * AGE +
-0.3468 * DIS +
0.1375 * RAD +
-0.0052 * TAX +
-0.2056 * PTRATIO +
-0.0035 * B +
-0.3361 * MEDV +
35.776

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.8355
Mean absolute error	2.8952
Root mean squared error	**3.9206**
Relative absolute error	50.5054 %
Root relative squared error	54.7932 %
Total Number of Instances	506

6. (Bonus 2p) Continue the task from ROC examples. Assume there is different cost assigned for different types of mistakes. E.g. cost 20€ for missing a case (false negative) and 15€ for false classification (false positive). Or vice versa. Calculate for each of the 4 classifiers with ROC curve, what would be the optimal cutoff to minimize cost. Provide yourself examples of four such costs based on which you can say for each of the 4 classifiers that exactly that provides the best classification.

For this task first I defined the misclassification error costs based on the given values as the following

FN cost = 20
FP cost = 15
Total = 3000 (Number of classified data)

Misclassification cost = ((FN * FN cost) + (FP x FP cost)) / Total

Roc1.txt

							FN COST	20			
							FP COST	15			
	PREDICTION MODEL										
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS	
0	1215	0	1785	0	0	1	1	0.099958848	0	8.925	
0.1	1214	299	1486	1	0.167507003	0.832492997	0.999176955	0.098436214	0.166683958	7.436666667	
0.2	1178	563	1222	37	0.315406162	0.684593838	0.969547325	0.092880658	0.284953488	6.356666667	
0.3	1079	764	1021	136	0.428011204	0.571988796	0.888065844	0.082674897	0.316077048	6.011666667	
0.4	930	915	870	285	0.512605042	0.487394958	0.765432099	0.068641975	0.278037141	6.25	
0.5	738	1023	762	477	0.573109244	0.426890756	0.607407407	0.054485597	0.180516651	6.99	
0.6	586	1171	614	629	0.656022409	0.343977591	0.482304527	0.041563786	0.138326936	7.263333333	
0.7	424	1309	476	791	0.733333333	0.266666667	0.348971193	0.028148148	0.082304527	7.653333333	
0.8	260	1445	340	955	0.80952381	0.19047619	0.21399177	0.016131687	0.023515579	8.066666667	
0.9	132	1617	168	1083	0.905882353	0.094117647	0.108641975	0.005432099	0.014524328	8.06	
1	0	1785	0	1215	1	0	0	0	0	8.1	
								0.588353909	AUC		

Minimal misclassification cost = 6.011666667
The optimal cutoff to minimize cost for the Roc1.txt is 0.3. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **0.3**

Roc2.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.098518519	0	8.925
0.1	1179	264	1521	36	0.14789916	0.85210084	0.97037037	0.095432099	0.11826953	7.845
0.2	1140	525	1260	75	0.294117647	0.705882353	0.938271605	0.089958848	0.232389252	6.8
0.3	1046	731	1054	169	0.40952381	0.59047619	0.86090535	0.080164609	0.270429159	6.396666667
0.4	902	887	898	313	0.496918768	0.503081232	0.742386831	0.068230453	0.239305599	6.576666667
0.5	756	1041	744	459	0.583193277	0.416806723	0.622222222	0.056090535	0.2054155	6.78
0.6	607	1192	593	608	0.667787115	0.332212885	0.499588477	0.043868313	0.167375592	7.018333333
0.7	459	1344	441	756	0.752941176	0.247058824	0.377777778	0.030699588	0.130718954	7.245
0.8	287	1472	313	928	0.82464986	0.17535014	0.236213992	0.017860082	0.060863852	7.751666667
0.9	147	1632	153	1068	0.914285714	0.085714286	0.120987654	0.006049383	0.035273369	7.885
1	0	1785	0	1215	1	0	0	0	0	8.1
								0.586872428	AUC	

Minimal misclassification cost = 6.396666667
The optimal cutoff to minimize cost for the Roc2.txt is 0.3. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **0.3** as seen in the image above

Roc3.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.09691358	0	8.925
0.1	1140	225	1560	75	0.12605042	0.87394958	0.938271605	0.09090535	0.064322025	8.3
0.2	1069	454	1331	146	0.254341737	0.745658263	0.879835391	0.084773663	0.134177128	7.628333333
0.3	991	676	1109	224	0.378711485	0.621288515	0.81563786	0.078024691	0.194349345	7.038333333
0.4	905	890	895	310	0.49859944	0.50140056	0.744855967	0.071769547	0.243455407	6.541666667
0.5	839	1124	661	376	0.629691877	0.370308123	0.690534979	0.060329218	0.320226856	5.811666667
0.6	627	1212	573	588	0.678991597	0.321008403	0.516049383	0.042139918	0.195040979	6.785
0.7	397	1282	503	818	0.718207283	0.281792717	0.326748971	0.028106996	0.044956254	7.968333333
0.8	286	1471	314	929	0.824089636	0.175910364	0.235390947	0.017901235	0.059480582	7.763333333
0.9	149	1634	151	1066	0.915406162	0.084593838	0.122633745	0.006131687	0.038039907	7.861666667
1	0	1785	0	1215	1	0	0	0	0	8.1
								0.576995885	AUC	

Minimal misclassification cost = 5.811666667

The optimal cutoff to minimize cost for the Roc3.txt is 0.5. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **0.5** as seen in the image above

Roc4.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.099917695	0	8.925
0.1	1213	298	1487	2	0.166946779	0.833053221	0.998353909	0.099794239	0.165300688	7.448333333
0.2	1212	597	1188	3	0.334453782	0.665546218	0.997530864	0.098683128	0.331984646	5.96
0.3	1186	871	914	29	0.487955182	0.512044818	0.976131687	0.092222222	0.464086869	4.763333333
0.4	1055	1040	745	160	0.582633053	0.417366947	0.868312757	0.078641975	0.45094581	4.791666667
0.5	856	1141	644	359	0.639215686	0.360784314	0.704526749	0.060658436	0.343742435	5.613333333
0.6	618	1203	582	597	0.67394958	0.32605042	0.508641975	0.04090535	0.182591555	6.89
0.7	376	1261	524	839	0.706442577	0.293557423	0.309465021	0.025720165	0.015907598	8.213333333
0.8	249	1434	351	966	0.803361345	0.196638655	0.204938272	0.015144033	0.008299616	8.195
0.9	119	1604	181	1096	0.89859944	0.10140056	0.097942387	0.004897119	-0.003458173	8.211666667
1	0	1785	0	1215	1	0	0	0	0	8.1
								0.616584362	AUC	

Minimal misclassification cost = 4.763333333

The optimal cutoff to minimize cost for the Roc4.txt is 0.3. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **0.3** as seen in the image above

Switching costs

In the first part of the analysis we assumed that the cost for missing a case (false negative) was higher than the cost for false classification (false positive) however in this second part of the analysis we will assume vice versa in order to get more insight into the analysis. I believe this will give us some more insight into the influence of the values in calculating the cost of misclassification for all the four classifiers provided.

FN cost = 15 FP cost = 20 Total = 3000 (Number of classified data)

*Misclassification cost = ((FN * FN cost) + (FP x FP cost)) / Total*

Roc1.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.099958848	0	11.9
0.1	1214	299	1486	1	0.167507003	0.832492997	0.999176955	0.098436214	0.166683958	9.911666667
0.2	1178	563	1222	37	0.315406162	0.684593838	0.969547325	0.092880658	0.284953488	8.331666667
0.3	1079	764	1021	136	0.428011204	0.571988796	0.888065844	0.082674897	0.316077048	7.486666667
0.4	930	915	870	285	0.512605042	0.487394958	0.765432099	0.068641975	0.278037141	7.225
0.5	738	1023	762	477	0.573109244	0.426890756	0.607407407	0.054485597	0.180516651	7.465
0.6	586	1171	614	629	0.656022409	0.343977591	0.482304527	0.041563786	0.138326936	7.238333333
0.7	424	1309	476	791	0.733333333	0.266666667	0.348971193	0.028148148	0.082304527	7.128333333
0.8	260	1445	340	955	0.80952381	0.19047619	0.21399177	0.016131687	0.023515579	7.041666667
0.9	132	1617	168	1083	0.905882353	0.094117647	0.108641975	0.005432099	0.014524328	6.535
1	0	1785	0	1215	1	0	0	0	0	6.075
								0.588353909	AUC	

Minimal misclassification cost = 6.075

As can be seen from above the image, the optimal cutoff to minimize cost for the Roc1.txt is now at 1. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **1**

Roc2.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.098518519	0	11.9
0.1	1179	264	1521	36	0.14789916	0.85210084	0.97037037	0.095432099	0.11826953	10.32
0.2	1140	525	1260	75	0.294117647	0.705882353	0.938271605	0.089958848	0.232389252	8.775
0.3	1046	731	1054	169	0.40952381	0.59047619	0.86090535	0.080164609	0.270429159	7.871666667
0.4	902	887	898	313	0.496918768	0.503081232	0.742386831	0.068230453	0.239305599	7.551666667
0.5	756	1041	744	459	0.583193277	0.416806723	0.622222222	0.056090535	0.2054155	7.255
0.6	607	1192	593	608	0.667787115	0.332212885	0.499588477	0.043868313	0.167375592	6.993333333
0.7	459	1344	441	756	0.752941176	0.247058824	0.377777778	0.030699588	0.130718954	6.72
0.8	287	1472	313	928	0.82464986	0.17535014	0.236213992	0.017860082	0.060863852	6.726666667
0.9	147	1632	153	1068	0.914285714	0.085714286	0.120987654	0.006049383	0.035273369	6.36
1	0	1785	0	1215	1	0	0	0	0	6.075
								0.586872428	AUC	

Minimal misclassification cost = 6.075

As can be seen from above the image, the optimal cutoff to minimize cost for the Roc2.txt is now at 1. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **1**

Roc3.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.09691358	0	11.9
0.1	1140	225	1560	75	0.12605042	0.87394958	0.938271605	0.09090535	0.064322025	10.775
0.2	1069	454	1331	146	0.254341737	0.745658263	0.879835391	0.084773663	0.134177128	9.603333333
0.3	991	676	1109	224	0.378711485	0.621288515	0.81563786	0.078024691	0.194349345	8.513333333
0.4	905	890	895	310	0.49859944	0.50140056	0.744855967	0.071769547	0.243455407	7.516666667
0.5	839	1124	661	376	0.629691877	0.370308123	0.690534979	0.060329218	0.320226856	6.286666667
0.6	627	1212	573	588	0.678991597	0.321008403	0.516049383	0.042139918	0.195040979	6.76
0.7	397	1282	503	818	0.718207283	0.281792717	0.326748971	0.028106996	0.044956254	7.443333333
0.8	286	1471	314	929	0.824089636	0.175910364	0.235390947	0.017901235	0.059480582	6.738333333
0.9	149	1634	151	1066	0.915406162	0.084593838	0.122633745	0.006131687	0.038039907	6.336666667
1	0	1785	0	1215	1	0	0	0	0	6.075
								0.576995885	AUC	

Minimal misclassification cost = 6.075

As can be seen from above the image, the optimal cutoff to minimize cost for the Roc3.txt is now at 1. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **1**

Roc4.txt

							FN COST	20		
							FP COST	15		
	PREDICTION MODEL									
PROBABILITY CUTOFF	TP	TN	FP	FN	TN RATE	FP RATE	TP RATE	AOC VALUES	YOUDEN INDEX	MISCLASSIFICATION COSTS
0	1215	0	1785	0	0	1	1	0.099917695	0	11.9
0.1	1213	298	1487	2	0.166946779	0.833053221	0.998353909	0.099794239	0.165300688	9.923333333
0.2	1212	597	1188	3	0.334453782	0.665546218	0.997530864	0.098683128	0.331984646	7.935
0.3	1186	871	914	29	0.487955182	0.512044818	0.976131687	0.092222222	0.464086869	6.238333333
0.4	1055	1040	745	160	0.582633053	0.417366947	0.868312757	0.078641975	0.45094581	5.766666667
0.5	856	1141	644	359	0.639215686	0.360784314	0.704526749	0.060658436	0.343742435	6.088333333
0.6	618	1203	582	597	0.67394958	0.32605042	0.508641975	0.04090535	0.182591555	6.865
0.7	376	1261	524	839	0.706442577	0.293557423	0.309465021	0.025720165	0.015907598	7.688333333
0.8	249	1434	351	966	0.803361345	0.196638655	0.204938272	0.015144033	0.008299616	7.17
0.9	119	1604	181	1096	0.89859944	0.10140056	0.097942387	0.004897119	-0.003458173	6.686666667
1	0	1785	0	1215	1	0	0	0	0	6.075
								0.616584362	AUC	

Minimal misclassification cost = 5.766666667
As can be seen from above the image, the optimal cutoff to minimize cost for the Roc4.txt is now 0.4. The reason is based on the fact that the we get the minimal misclassification cost for this classifier at the **cutoff** for **0.4**

Summary of bonus Task
For the analysis I can deduce that the optimal cutoff for any of the classifiers in regards to the misclassification when taking cost into account is highly dependant on cost for missing a case (false negative) vs the cost for false classification (false positive) hence when the values are swapped the optimal cutoff also changes. Roc4.txt in this case had the best observation as the difference between the optimal cutoff based on Youdens index and optimal cut off based on misclassification cost is just one step down as opposed to the rest of the classifiers. Below are summary tables for each of the classifiers

Roc1.txt Conclusion

roc1.txt	FN cost	FP cost	Optimal Cutoff
	20	15	0.3
	15	20	1

Roc2.txt Conclusion

roc2.txt	FN cost	FP cost	Optimal Cutoff
	20	15	0.3
	15	20	1

Roc3.txt Conclusion

roc3.txt	FN cost	FP cost	Optimal Cutoff
	20	15	0.5
	15	20	1

Roc4.txt Conclusion

roc4.txt	FN cost	FP cost	Optimal Cutoff
	20	15	0.3
	15	20	0.4

Task 5 contd.
Which variables are "easier" to predict than others? If so, then why?

From my observation I believe the variables CHAS, NOX and RM are the easiest to predict. My reason for saying this is because they have the least RMSE score as well as the use the fewest variables for calculations. **NOX** comes first in my ranking with an **RMSE score of 0.0556** based on the calculation with **7 variables** and is followed by **CHAS** which has an **RMSE score of 0.2496** based on the calculations with **4 variables**. In third place comes **RM** with an **RMSE score of 0.4866** based on calculations from **7 variables**.