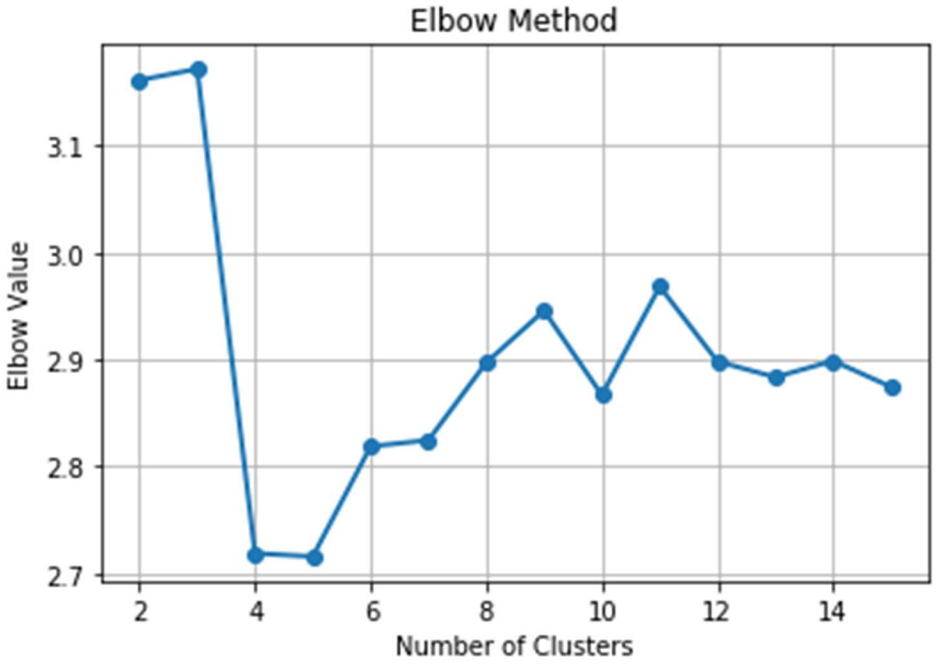
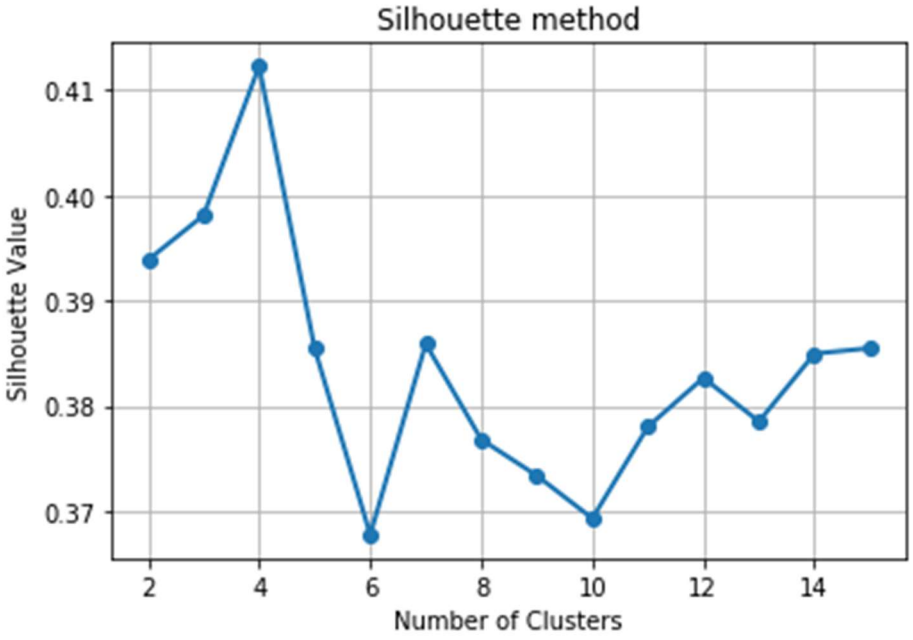
CS 584-04: Machine Learning

Fall 2018 Midterm Test

Question 1

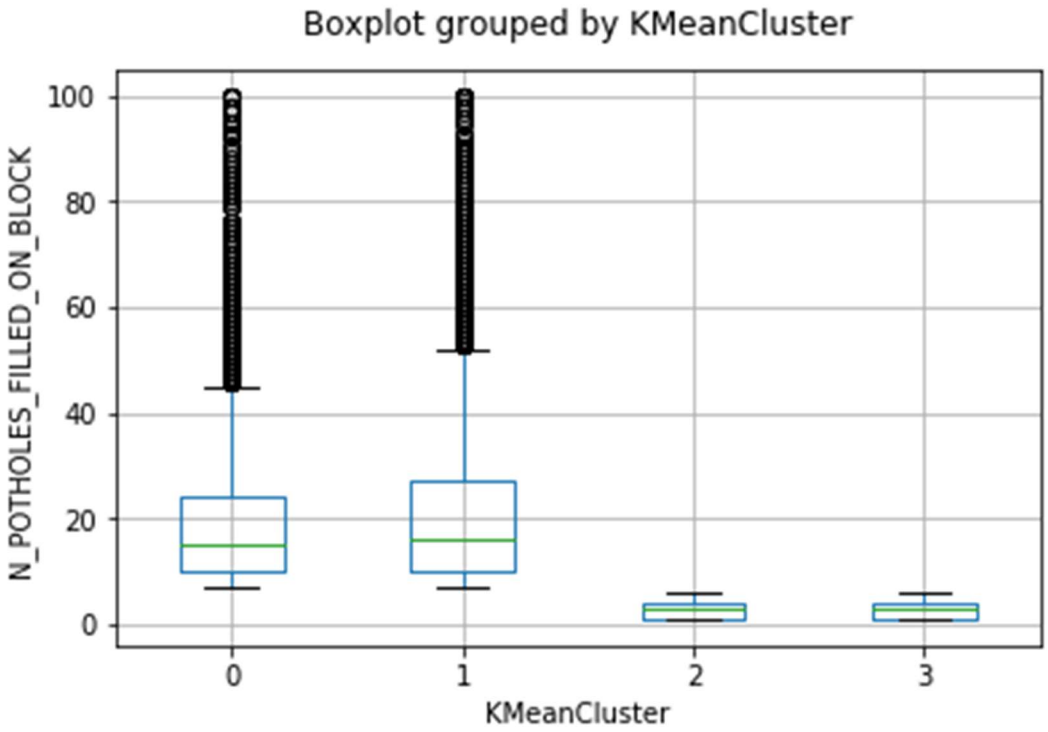
a) (10 points) How many clusters did you determine? Please provide the Elbow and the Silhouette charts and state your arguments. The charts must be properly labeled.

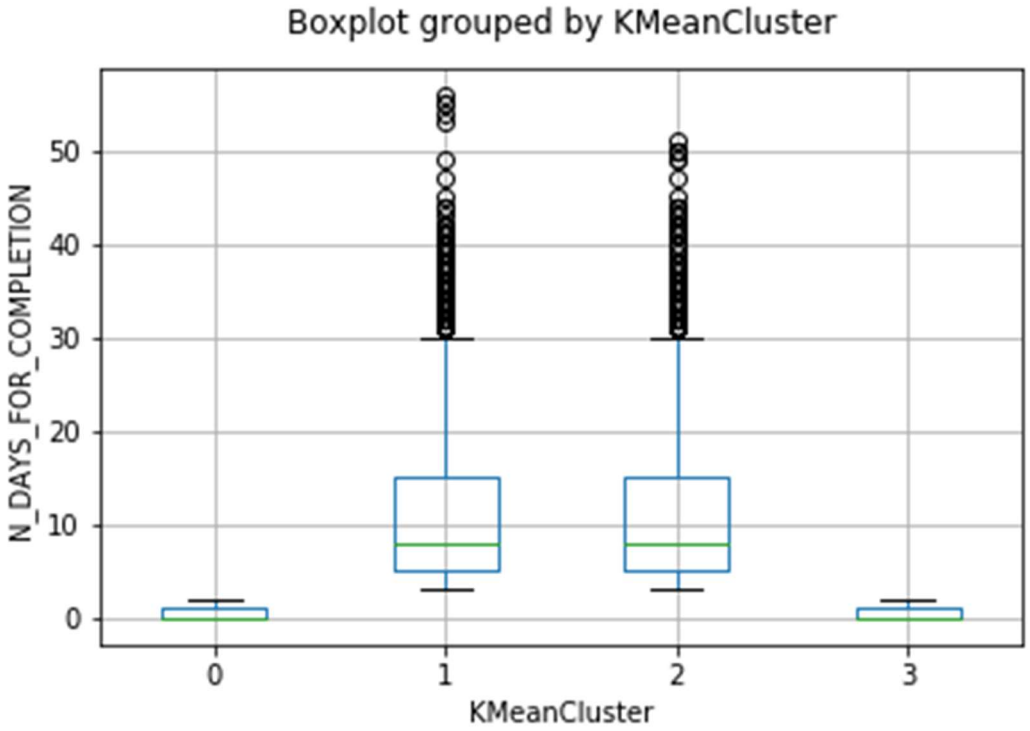


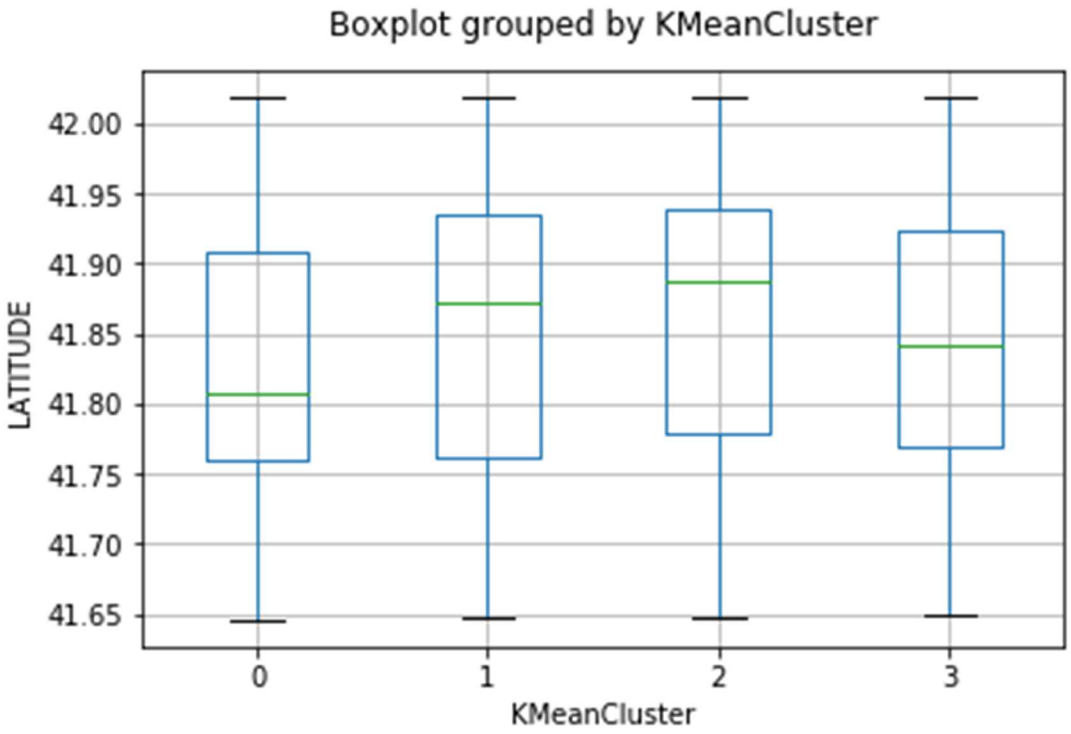


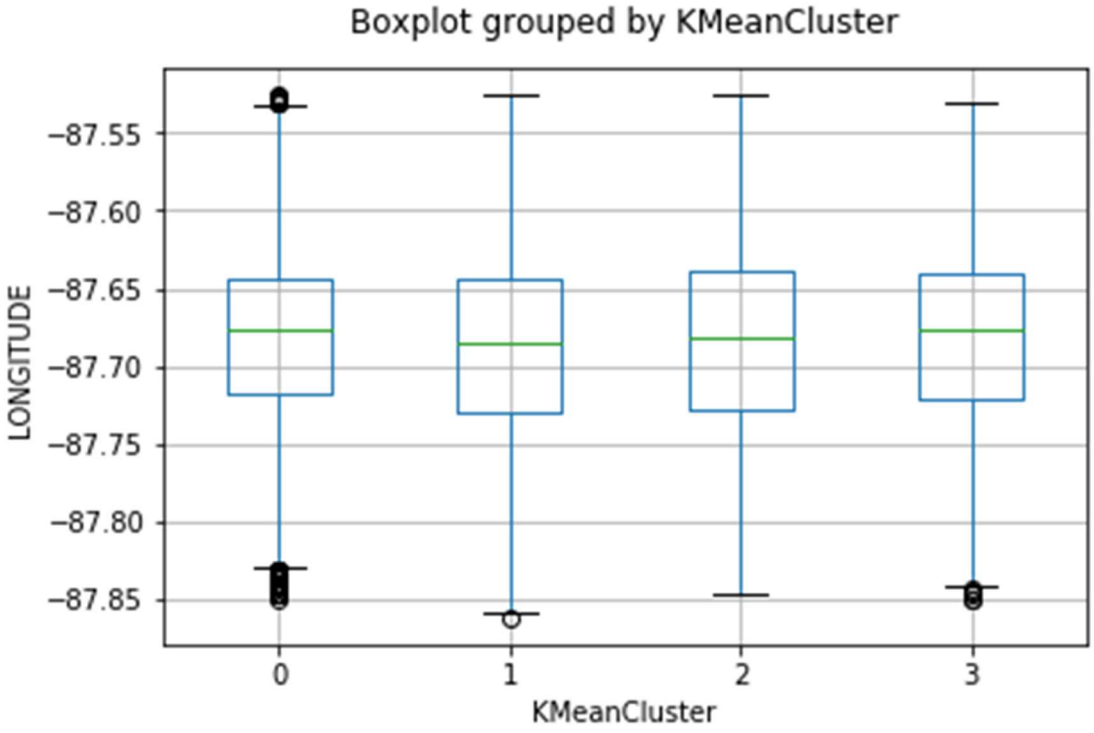
The elbow in the L- curve can be seen at 4. Hence, we select this as 'k', number of clusters. Also, the Silhouette value peaks at 4 in the graph. Therefore, I chose 4 to be the number of clusters.

b) (5 points) Create a box-plot for each of these four variables: N_POTHOLES_FILLED_ON_BLOCK, N_DAYS_FOR_COMPLETION, LATITUDE, and LONGITUDE, grouped by the Cluster ID.

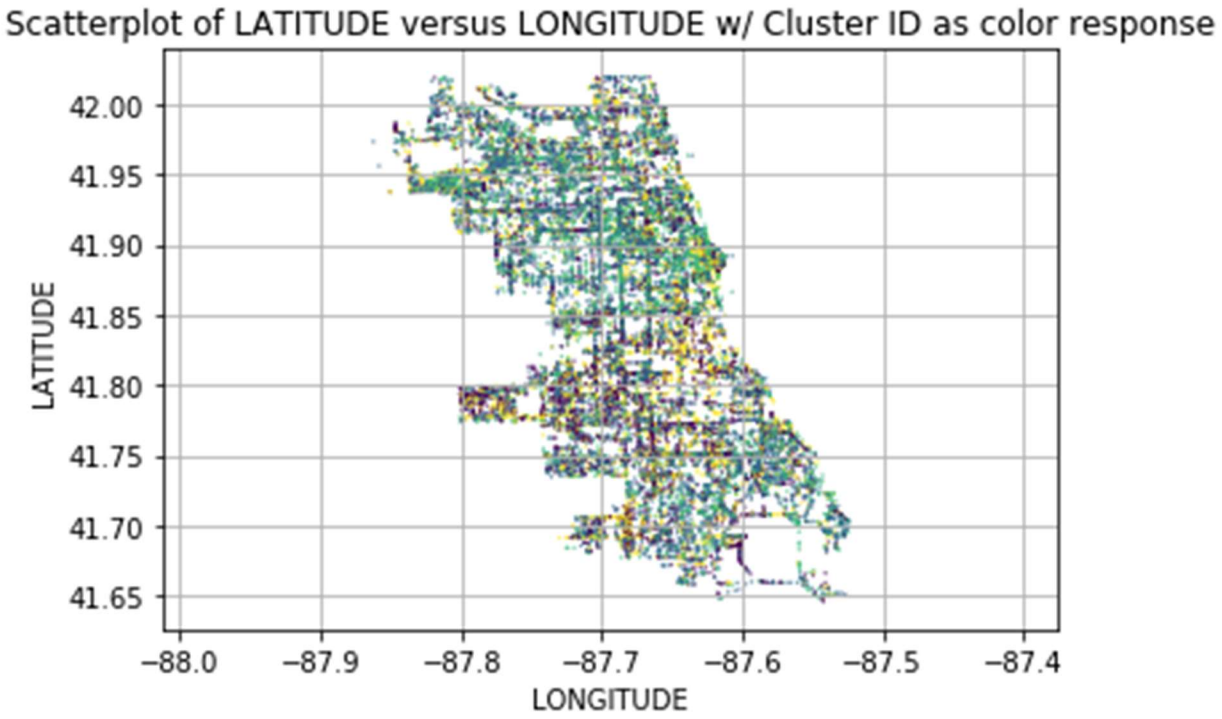





Boxplot grouped by KMeanCluster
LATITUDE
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
0
1
2
3
KMeanCluster


Boxplot grouped by KMeanCluster
LONGITUDE
−87.55
−87.60
−87.65
−87.70
−87.75
−87.80
−87.85
0
1
2
3
KMeanCluster

c) (5 points) Generate a scatterplot of LATITUDE (y-axis) versus LONGITUDE (x-axis) using the Cluster ID as the color response variable. You may need to adjust the marker size and set the aspect ratio to one in order to make the scatterplot more readable.

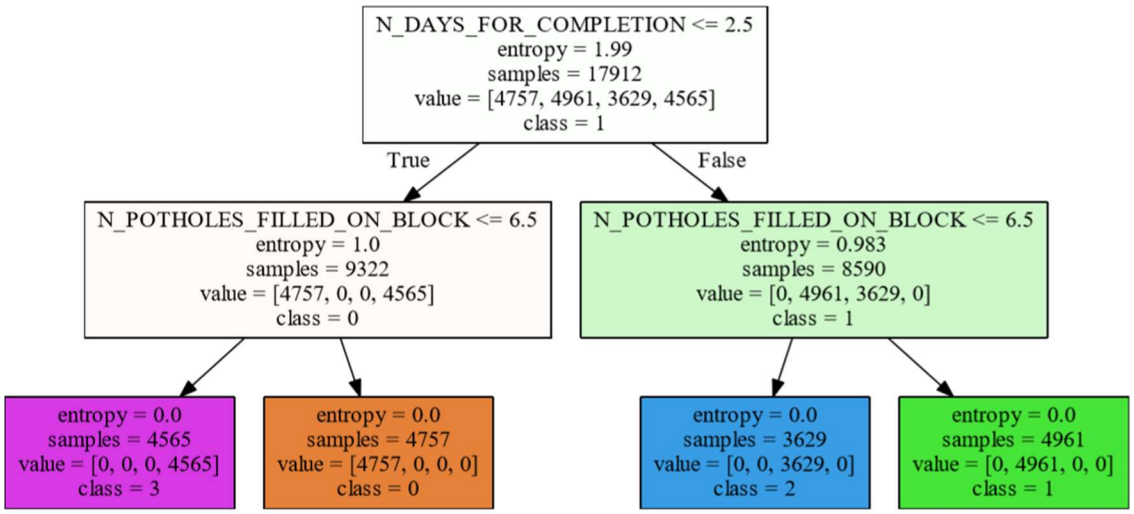


d) (5 points) Comment your scatterplot in (c). In particular, how effective or ineffective do you think the clustering analysis in dividing up the observations according to their geographical locations?

We can see that the data is not separated well enough for clustering analysis. Data from one cluster can be found amidst another and there is no clear demarcation. Therefore, it would be very ineffective to divide the observations according to their geographical locations(latitude and longitude) for clustering analysis.

e) (10 points) How many leaves did you have in your classification tree? Please attach the tree diagram which must be properly labeled.

I have 4 leaves in my classification tree.



f) (5 points) Calculate the Misclassification Rate and the Root Mean Squared Error of your classification tree.

The misclassification rate of the classification tree above is 0.

g) (10 points) Based on your classification tree, how would you describe the profiles of the clusters?
First of all, we can see that the clusters are not based on location at all as speculated in the earlier question. Rather, they are based on number of days taken for completion (N_DAYS_FOR_COMPLETION) and the number of potholes filled on the block (N_POTHOLES_FILLED_ON_BLOCK).

Cluster 3:
Contains requests which took less than 2.5 days to complete filling the pothole and less than 6.5 potholes were filled on the block.

Cluster 0:
Contains requests which took less than 2.5 days to complete filling the pothole but more than 6.5 potholes were filled on the block.

Cluster 2:
Contains requests which took more than 2.5 days to complete filling the pothole but less than 6.5 potholes were filled on the block.

Cluster 1:
Contains requests which took more than 2.5 days to complete filling the pothole and more than 6.5 potholes were filled on the block.

Question 2

a) (5 points) How many observations are in the Training and the Testing partitions?

Number of Observations in Training partition = 431
Number of Observations in Testing partition = 186

b) (5 points) What are the claim rates in the Training and the Testing partitions?

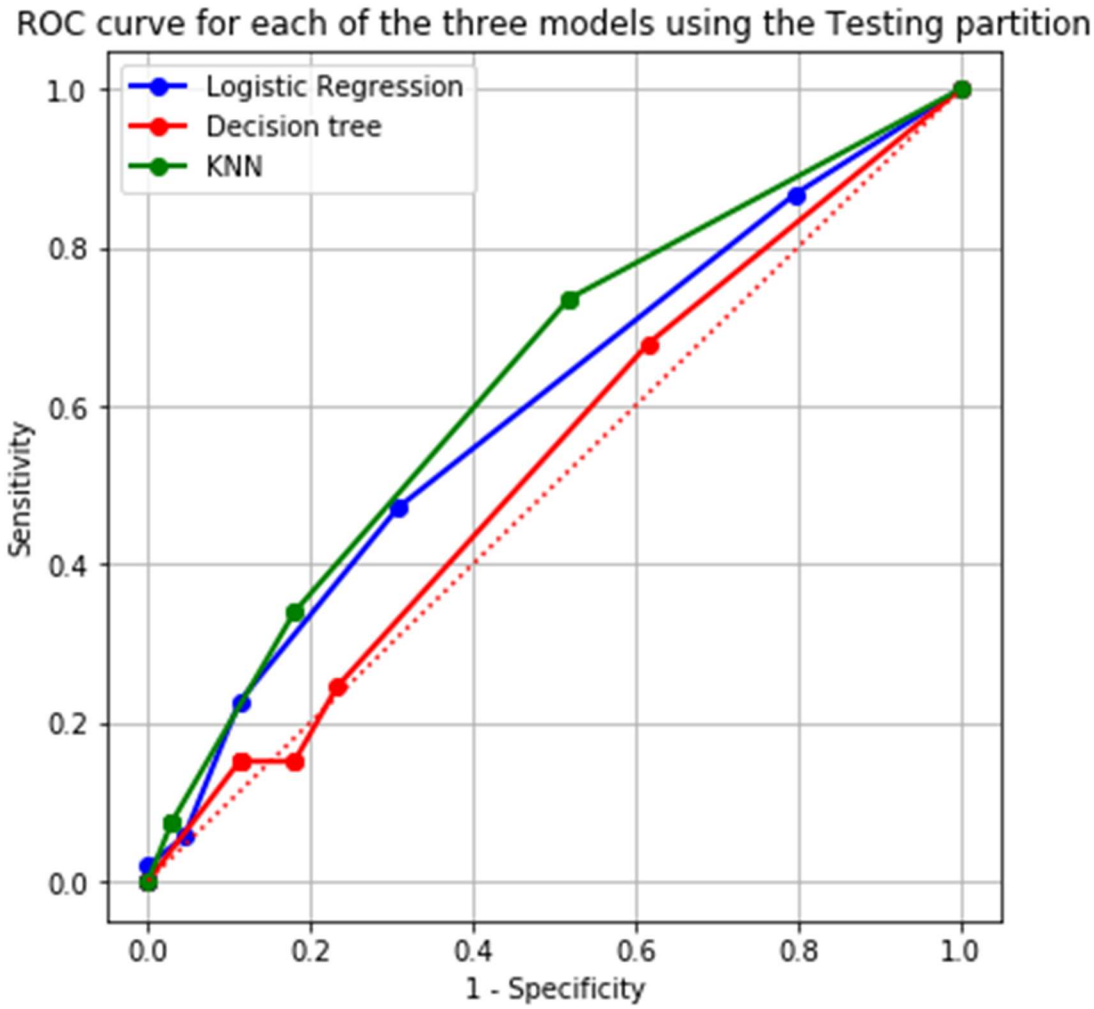
Claim rates in Training partition:
0 - 0.712297
1 - 0.287703

Claim rates in Testing partition:
0 - 0.715054
1 - 0.284946

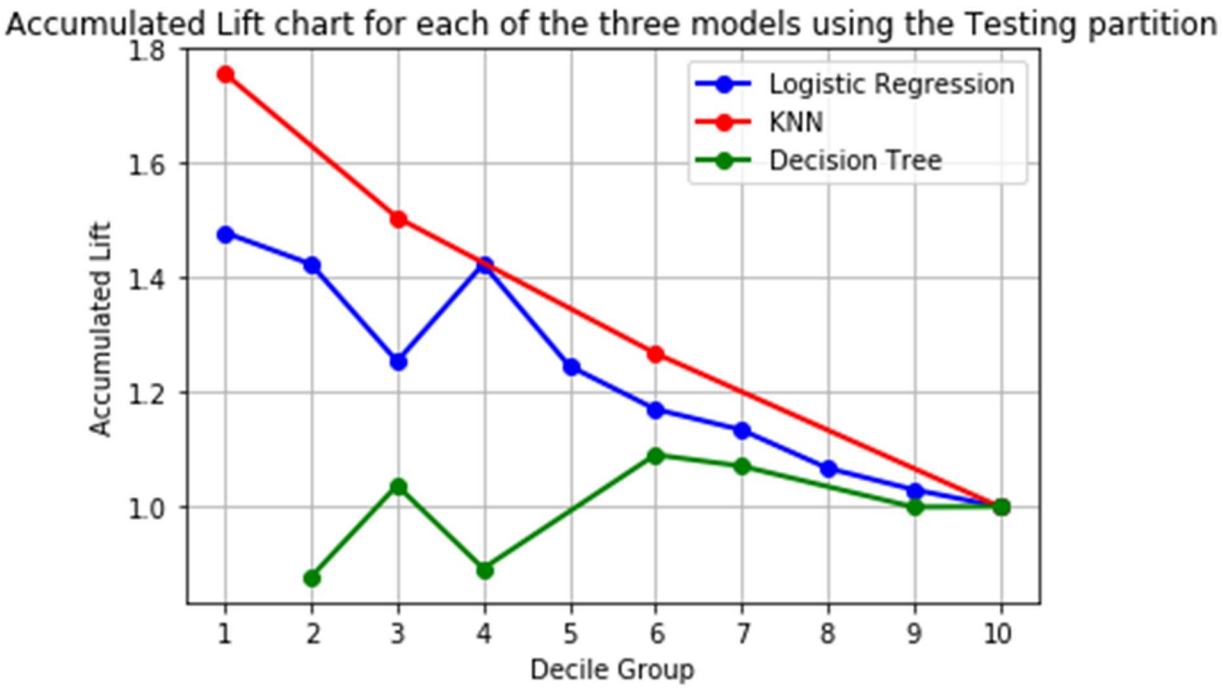
c) (10 points) Use **the claim rate in the Training partition** as the probability threshold in the misclassification rate calculation. A claim is predicted if the predicted probability of filing a claim is greater than or equal to the probability threshold. Calculate the Area Under Curve metric, the Root Mean Squared Error metric, and the Misclassification Rate for all three models using the Testing partition. List the metrics as the rows and the models as the columns in a table.

	Decision Tree	Logistic Regression	KNN
Misclassification Rate	0.3817204301075269	0.3602150537634409	0.446236559139785
Area under curve	0.4852461342034331	0.5057454958150092	0.5795857568449425

d) (10 points) Calculate (but no need to display) the coordinates of the Receiver Operating Characteristic curve for each of the three models using the Testing partition. Plot all three curves in the same chart but use a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.



e) (10 points) Calculate (but no need to display) the coordinates of the Accumulated Lift chart for each of the three models using the Testing partition. Plot all three accumulated lift curves in the same chart but use a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.



f) (10 points) Based on the evaluation and the comparison results in (c), (d), and (e), which single model will you recommend? Please state your reasons for your recommendation.

As we have a binary target, we consider the Area Under Curve metric and Lift in the first few deciles, both of which should be high. A model with an AUC > 0.5 is deemed to be acceptable and higher the AUC above 0.5, better the model. Decision tree, with an AUC of 0.48 right away becomes unacceptable. The logistic regression and the KNN model have and AUC of 0.505 and 0.57 each; KNN is the better model according to AUC. Even the lift is higher for KNN in the first few deciles. However, the misclassification rate for KNN is the highest with 0.44, with the logistic regression model faring much better at 0.36 (Even lower than misclassification for decision tree,0.38). This indicates that the logistic regression model has learnt the entire data better. Since there seems to be no apparent reason to learn only a portion of the dataset better, I would prefer a model that fits the entire dataset well. Therefore, I will consider metrics like AUC and misclassification rather than lift. I would recommend the logistic regression model as it has an acceptable AUC with a lower misclassification rate than KNN and hence it will generalize better.