Machine Learning
Telecom Customer Churn Analysis
Rohit Jyotiba Chougule

Question. 1:
The phone company dataset consists of about 29 features, however, not all the features will be useful in predicting the customer churn from the dataset. Hence, we can reduce the dimension of the data using either of the feature selection strategies to reduce the dimension of the data and try to predict whether a customer will leave the service or not. We will use the Weka to use filter and wrapper feature selection strategy to take a subset of the features from the dataset.

a) Applying Filter feature selection strategy on the phone company dataset.

We use the Select attributes from Weka to choose 'InfoGainAttributeEval' and rank the attributes based on Information gain of the features and implement various methods to identify the subset of features that will give a good accuracy.
Selected attributes according to their rank: 6, 25, 13, 14, 7, 9, 20, 17, 26, 21, 27, 19, 24, 5, 8, 28, 2, 10, 11, 3, 29, 4, 15, 12, 16, 23, 22, 18, 1
As seen in the screenshot:

```
Ranked attributes:
 0.02195      6 handsetAge
 0.020515    25 lifeTime
 0.006154    13 avgMins
 0.00588     14 avgrecurringCharge
 0.005217     7 smartPhone
 0.003915     9 creditRating
 0.002787    20 avgOutCalls
 0.002739    17 callMinutesChangePct
 0.002676    26 lastMonthCustomerCareCalls
 0.002622    21 avgInCalls
 0.002313    27 numRetentionCalls
 0.002043    19 avgReceivedMins
 0.001515    24 avgDroppedCalls
 0.001078     5 numHandsets
 0.000997     8 currentHandsetPrice
 0.000747    28 numRetentionOffersAccepted
 0.000375     2 marriageStatus
 0.000127    10 homeOwner
 0.000115    11 creditCard
 0.000111     3 children
 0           29 newFrequentNumbers
 0            4 income
 0           15 avgOverBundleMins
 0           12 avgBill
 0           16 avgRoamCalls
```

i) First, we can remove the columns where the Information gain is 0 and hence their rank is low as well. We can remove the column number: 29,4,15,12,16,23,22,18,1 from the above feature list.

We now have 20 columns and we can test the accuracy of a classifier with these 20 columns.
The Naïve Bayes Classifier in Weka is used to predict the customer churn using k-fold validation on our dataset using the 20-feature subset, which gives accuracy of correctly classified instances as 52.5474%, and 47.4526% are incorrectly classified.

ii) From the ranking of features, we can choose top k features to identify whether which features gives high accuracy and then increase the number of features further to check if the accuracy can be improved.

Top 10 features are selected to be applied with Naïve Bayes Classifier in Weka, and it turns out we get an accuracy of correctly classified instances as 53.1579% and 46.8421% are incorrectly classified.

On subsequently performing the steps for further reduction of features we get the classification accuracy as follows:

Feature subset selected	**Correctly classified**	**Incorrectly classified**
Features with non-zero Information gain	52.5474%	47.4526%
Top 1 ranked features	53.8526%	46.1474%
Top 2 ranked features	54.5053%	45.4947%
Top 3 ranked features	55.8316%	44.1684%
Top 4 ranked features	55.7158%	44.2842%
Top 5 ranked features	56.4526%	43.5474%
Top 6 ranked features	56.1158%	43.8842%
Top 7 ranked features	56.1263%	43.8737%
Top 8 ranked features	54.7053%	45.2947%
Top 9 ranked features	53.8421%	46.1579%
Top 10 ranked features	53.1579%	46.8421%

It can be observed that when selecting Top 5 ranked features, the accuracy for Naïve Bayes classifier seems to be good as compared when selecting subset of other possibilities.

The subset of features using Filter technique for dimensionality reduction may suggest using Top 5 ranked features, which are: handsetAge, Lifetime, avgMins, avgrecurringCharge, smartphone.

a) Applying wrapper feature selection strategy on the phone company dataset:
- Unlike Filter selection strategy, the wrapper feature selection strategy evaluates the feature subsets based on their performance for a specific chosen classifier.
- The wrapper utilizes feature subset search that generates feature potential subsets for evaluation. Here, Sequential Search is implemented in Weka to generate potential subsets.
- Sequential Search uses two search methods, forward sequential search and backward sequential search.
- In Weka, a wrapper is now implemented with GreedyStepwise search method with forward search for Naïve Bayes Classifier and it turns out to give 3 feature subsets as below:

Selected attributes: 6,13,28 : 3
handsetAge
avgMins
numRetentionOffersAccepted

Whereas, if we set backwardSearch=True, i.e, use Backward GreedyStepwise search for wrapper attribute evaluator with Naïve Bayes classifier and it turns out to give about 21 feature subsets as below:
Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,14,16,17,18,20,21,23,25,27,28 : 21
age, marriageStatus, children, income, numHandsets, handsetAge, smartphone, currentHandsetPrice, creditRating, homeowner, creditCard, avgrecurringCharge, avgRoamCalls,

callMinutesChangePct, billAmountChangePct, avgOutCalls, avgInCalls, peakOffPeakRatioChangePct, lifetime, numRetentionCalls, numRetentionOffersAccepted

Now, as the list of features is shortlisted, we have about 3 features when using Backward Sequential Greedy search and 21 features when using forwards sequential GreedySearch.
We can verify the accuracy of the Naïve Bayes classifier for each of the feature subset:

For Naïve Bayes Classifier:

Feature subset selected	Correctly Classified	Incorrectly Classified
Forward Greedy Search (3 features)	57.3158%	42.6842%
Backward Greedy Search (21 features)	56.3053%	43.6947%

We can choose the subset generated by forward Greedy search feature evaluator using Naïve Bayes that gives 3 features in the subset, which are: handsetAge, avgMins, numRetentionOffersAccepted

Question 1.

b) Report and discuss differences between feature subsets produced by Filter and wrapper techniques from Task A.

Feature subset selected using Filter Technique	Feature subset selected using Wrapper Technique (Forward)
handsetAge	handsetAge
Lifetime	avgMins
avgMins	numRetentionOffersAccepted
avgrecurringCharge	
smartPhone	

From the above table of feature subsets for each of the dimension reduction technique, its seen that 2 of the features are common in both the Filter and Wrapper technique.

- Filter technique evaluates the feature depending on a statistical measure for the attributes in the dataset, Information gain is the measure implemented in this case.
- Whereas, the wrapper method uses and tries a subset of feature on a model to train in either forward or backward search approach to improve the classifier accuracy.
- In this case, the attributes selected in Filters seem to give highest classifier accuracy as compared to trying other combination hence we get the feature subset.
- Whereas, the wrapper worked in forward greedy search the handsetAge in both the subset is common because the Information gain is high(in case of filter)or the accuracy that the feature gives for the classifier is maximum (in case of wrapper), which can be inferred from the rank of this feature as well. Further features are evaluated in a similar manner in the wrapper in a greedy approach where we select the attributes which gives maximum accuracy when applied the classifier defined.
- In this wrapper, initially we have an empty subset and we add the features in the subset in a greedy manner which gives highest classifier accuracy for the defined classifier.

Question 1.

c) Evaluate and discuss the performance of both of the above feature selection techniques, when each one is combined with two different classifiers of your choice available in Weka (i.e. there will be four experimental combinations). Which combination do you believe is most suitable for this dataset?

- Using Decision Tree Classifier for each of the subsets:
  1. Features from Filter:
     handsetAge, lifeTime, avgMins, avgrecurringCharge, smartphone

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5628               59.2421 %
Incorrectly Classified Instances      3872               40.7579 %
Kappa statistic                          0.1843
Mean absolute error                      0.4779
Root mean squared error                  0.4906
Relative absolute error                 95.5713 %
Root relative squared error             98.1108 %
Total Number of Instances             9500
```

  2. Features from Wrapper:
     handsetAge, avgMins, numRetentionOffersAccepted

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5566               58.5895 %
Incorrectly Classified Instances      3934               41.4105 %
Kappa statistic                          0.1715
Mean absolute error                      0.4845
Root mean squared error                  0.4926
Relative absolute error                 96.9078 %
Root relative squared error             98.523  %
Total Number of Instances             9500
```

- Using K-NN Classifier for each of the subsets:

  1. Features from Filter:
     handsetAge, lifeTime, avgMins, avgrecurringCharge, smartphone

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5598               58.9263 %
Incorrectly Classified Instances      3902               41.0737 %
Kappa statistic                          0.1783
Mean absolute error                      0.4751
Root mean squared error                  0.4891
Relative absolute error                 95.0213 %
Root relative squared error             97.8186 %
Total Number of Instances             9500
```

  2. Features from Wrapper:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5447               57.3368 %
Incorrectly Classified Instances      4053               42.6632 %
Kappa statistic                          0.1467
Mean absolute error                      0.4821
Root mean squared error                  0.4931
Relative absolute error                 96.4291 %
Root relative squared error             98.6275 %
Total Number of Instances             9500
```

The below summary table shows the details about combination of each of the feature selection strategy for 2 classifiers:

Feature Selection Strategy	Correctly Classified Accuracy	
	Decision Tree Classifier	K-NN Classifier (k=97)
Filter	59.2421%	58.9263%
Wrapper	58.5895%	57.3368%

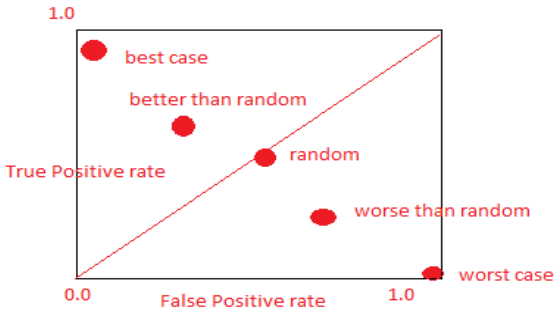
From the above summary table, I may conclude to use the feature subset obtained from Filter feature selection strategy to implement decision tree classifier as it gives the highest accuracy.

****

Question2:

a) What is a ROC Curve and what does it represent?

- The Receiver Operating Characteristic abbreviated as ROC curve is the graph of the Sensitivity vs Specificity of different thresholds for a given classifier. We draw a curve for each of the threshold value point.



- Y-axis: True positive rate (Sensitivity) is ratio of True Positives to sum of True Positives and False negatives.
- X-axis: False positive rate (Specificity) is ratio of False Positives to sum of False Positives and True Negatives.
- The reason we need to create a ROC is due to the need for comparison for various classifier performance at different threshold values which results in large number of confusion matrices.

- There is a reference line that passes through the 0,0 value for the graph, if our classifier lies on the line, it defines that there is an equal number of correctly and incorrectly classified set.
- The best-case scenario is the scenario where there are 0 false positives and 1 true positive, whereas the worst case scenario is the reverse, with 0 true positive and 1 false positive.
- Comparisons can be made based on the area that is covered under the graphs which is known as Area under the curve (AUC). Larger the Area under the curve better the classifier, provided the ROC curve is closer to the top left corner.

b) Difference in Lazy and Eager learning approach in Classification with example.
- Lazy Learning algorithm stores the training data and waits for a testing data to appear.
- The lazy learning algorithm classifies the data based on the most related data in the training set.
- Lazy learning algorithm have less training time but more predicting time.
- Examples of lazy learning: knn (k nearest neighbor)
- Eager learning algorithms construct a classification model based on the data, which is given as a training set, prior to getting the data for classification
- As the model construction takes place along with the training set, these algorithms take long time to train.
- Examples of Eager learning approach is Decision tree and Naïve Bayes.

c) Describe Conditional Independence assumption in Naïve Bayes.
- The Naïve Bayes classifier works on the principle of the Bayes Theorem, with an assumption that all the features in given dataset are conditionally independent.
- For example, if we have a dataset for predicting if there will be a cricket match game depending on weather conditions. We have features like- Outlook, Temperature, Humidity, Windy and Play Cricket in our dataset.
- The Play Cricket feature is the one which we predict whether there will be a game of cricket. Now we assume that each of the features in our dataset is independent of one another.
- We calculate the conditional probabilities for each of the feature independently and then use the test dataset to verify the prediction. The independent probabilities of each of the feature from the test data are considered along with the class probability.

d) Explain why classification accuracy may not be a good measure for classification problems with imbalanced data (e.g., fraud detection). Which evaluation measures are better suited for dealing with skewed class sizes?
- Imbalanced data is the data which is skewed towards one of the classes. Having imbalanced data to be applied on a model may lead to bias towards a class.
- There are various examples of imbalanced data in day to day life like telecom customer churning, fraud detection. Fraud detection using the credit card transactions can be termed as imbalanced as most of the credit card transactions are legitimate, very less number of transactions can be categorized as fraudulent.
- To evaluate classifiers when applying them to imbalanced data we can use two measures, which are: BAR(Balance Accuracy Rate) which is the mean of TP rate and TN rate and BER(Balance Error rate) which is mean of FP rate and FN rate.

e) Explain why is it not a good idea to select credit card number and name as features to split in a decision tree, even if these features result in the highest information gain?

- When we select Credit card number as a feature split, it will indeed result in highest information gain for the training data. However, this results in overfitting of the model, and it would work accurately only on the training data.
- The model may fail in case of unknown data or test data.
- Selecting such features which are unique for each of the training data values, it will result in n-1 split points where there are n unique values.
- It is important that feature split should not be applied on features having unique numeric values, rather using features that have categorical values are a better approach for splitting in a decision tree.