**[Using different method rather than the ensemble selection]**

For preprocessing data and tackling its problems, we cleaned it up as the following:

1. Remove all attributes that has > 50% missing values
2. Use unsupervised filters to:
   1. Replace the missing values for nominal and numeric attributes with the modes and means
   2. Change numeric data type to nominal
   3. Apply discretization
3. Select features using InfoGainAttributeEval filter

Which evaluates the worth of an attribute by measuring the information gain with respect to the class, ranks attributes by their individual evaluations. Based on the ranking, we remove the least significant attributes that has < 0.001

We carried out three classifiers (Naïve Bayes, Lazy IBK, and Decision Table) on after preprocessing the data using 3-fold cross validation. Naïve Bayes Classifier (NBC) is simple in structure and it is based on the assumption that predictors are conditionally independent given the target variable. Because of its simplicity, NBC was attractive choice since we have a large set of variables. The second classifier is Lazy IBK (Instance-Based K) which is one of the Nearest Neighbor algorithms. Thirdly, we used Decision Table Classifier (DTC) which uses a simple decision table majority algorithm; it makes decisions on attributes for each instance. Table-1 shows the AUC results of each classifier. Among the three classifiers, we can see that NBC did better than Lazy IBK and DTC while classifying for churn and appetency, and for up selling; DTC was the best (highest AUC). The highest overall score was for DTC with slight difference with NBC, while Lazy IBK was the worst. Moreover, comparing the results we got with the KDD results 1; up selling always get the highest AUC value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | AUC | | | |
| Churn | Appetency | Up Selling | Score |
| NBC | 0.568 | 0.523 | 0.682 | 0.591 |
| Lazy IBK | 0.533 | 0.5 | 0.564 | 0.532 |
| Decision Table | 0.53 | 0.5 | 0.756 | 0.595 |

Table-1

We also, carried the same three classifiers but this time we got rid of the categorical attributes to see how that affects the AUC results. Table-2 shows that DTC still got the highest overall score and this time there is a noticeable difference between DTC and NBC. Looking at each label, Lazy IBK did well while classifying churn and appetency though for up selling, DTC was the best. Generally, DTC is the best among those three classifiers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | AUC | | | |
| Churn | Appetency | Up Selling | Score |
| NBC | 0.529 | 0.497 | 0.658 | 0.561 |
| Lazy IBK | 0.532 | 0.525 | 0.619 | 0.558 |
| Decision Table | 0.53 | 0.5 | 0.757 | 0.595 |

Table-2

1: http://www.kdd.org/kdd-cup-2009-customer-relationship-prediction