KSE521 Business Intelligence

Homework #8

### **Data Preprocessing**

1. Convert ordinal and binary variables from string type to integer type.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Original Values** | **Converted Values** |
| COLLEGE | one, zero | 1, 0 |
| REPORTED\_SATISFACTION | very\_unsat, unsat, avg, sat, very\_sat | 1, 2, 3, 4, 5 |
| REPORTED\_USAGE\_LEVEL | very\_little, little, avg, high, very\_high | 1, 2, 3, 4, 5 |
| CONSIDERING\_CHANGE\_OF\_PLAN | never\_thought, no, perhaps, considering, actively\_looking\_into\_it | 1, 2, 3, 4, 5 |
| LEAVE | leave, stay | 1, 0 |

Table 1: Conversion of Variables

1. Negative Value Replacement
   1. There is one case with average overcharges (OVERCHARGE) per month = -2. The value of average overcharges is replaced by 0 for this observation.

### **Model Building using ALL Variables (Question 1a)**

Stratified sampling was performed with 80% of the data as train set and 20% of the data as test set. Train set was used to build the models while test set was used to measure the performance of the models in terms of accuracy (percentage of correctly classified instances).

Besides, baseline models were designed to evaluate the performance of each final model built.

1. **Classification Tree Model**

**C5.0 algorithm** is used to build the classification tree model.

|  |  |  |
| --- | --- | --- |
| **Model** | **Description** | **ACCURACY (%)** |
| 1 | Baseline model using majority classifier | 50.75 |
| 2 | Baseline model using only single node(attribute) - HOUSE | 61.05 |
| 3 | Model developed using all attributes | **70.55** |

Table 2: Baseline Models and Final Classification Tree Model's Accuracy

* 1. Baseline model using majority classifier (Model 1)
     1. Based on majority count in train set, 50.74% of the customers did not leave. If we predict all customers in the test set will stay (non-churn) eventually, then we would get an accuracy of 50.75% since 50.75% of customers in test set in fact did not churn.
  2. Baseline model using only single attribute (Model 2)
     1. The best attribute selected by C5.0 algorithm for the first split is HOUSE with splitting point at 600469. Customers with value of dwelling less than or equal to $600469 will be predicted as churn customers while customers with value of dwelling more than $600469 will be predicted as non-churn customers.
     2. By using the splitting rule as per b(i) to predict the customer churn status in test set, we would get accuracy of 61.05%.
  3. Building models using all variables (Model 3)
     1. We put all the attributes through the C5.0 algorithm and let it select the best sets of nodes to build the classification tree model.
     2. We set up the hyperparameters in the model such that global pruning (noGlobalPruning=FALSE) is allowed, and smallest number of samples that must be put in at least two of the splits is 30 (minCases=30).
     3. HOUSE variable is chosen as the first attribute to split the instances with splitting point at 600649.
     4. The tree built is as shown below (Figure 1 and Figure 2), with HOUSE variable as the first parent node, and each indentation represents children nodes of the parent node one level above. Further segmentation of customers based on this tree will be discussed in section later (Page 7).

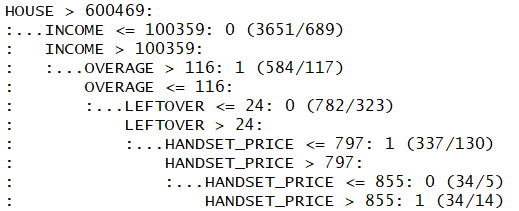


Figure 1: Branches and children nodes under Parent node HOUSE > 600469

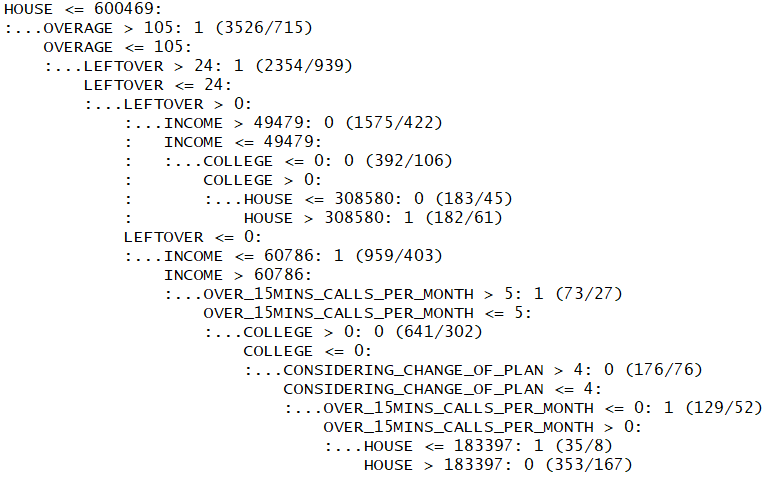


Figure 2: Branches and children nodes under Parent node HOUSE <= 600469

* + 1. The attributes selected to build the final model are HOUSE, OVERAGE, LEFTOVER, INCOME, COLLEGE, CONSIDERING\_CHANGE\_OF\_PLAN, HANDSET\_PRICE and OVER\_15MINS\_CALLS\_PER\_MONTH. The percentage of instances in the train set classified using the respective attributes is depicted in Figure 3.

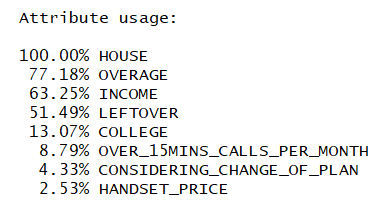


Figure 3: Attribute Usage of Classification Tree

* + 1. The generalization accuracy of this model is 70.55%, significantly higher than both the baseline models selected.

1. **Logistic regression model**

|  |  |  |
| --- | --- | --- |
| **Model** | **Description** | **ACCURACY (%)** |
| 1 | Baseline model using majority votes | 50.75 |
| 2 | Model developed using all attributes | **64.15** |

Table 3: Baseline Models and Final Logistic Regression Model's Accuracy

* 1. Baseline model using majority classifier (Model 1)
     1. Same as Model 1 used in decision tree induction.
  2. Building model using all variables (Model 2)
     1. All variables were used to build the logistic regression model. The estimated coefficient/slope of each variable is depicted in Figure 4.

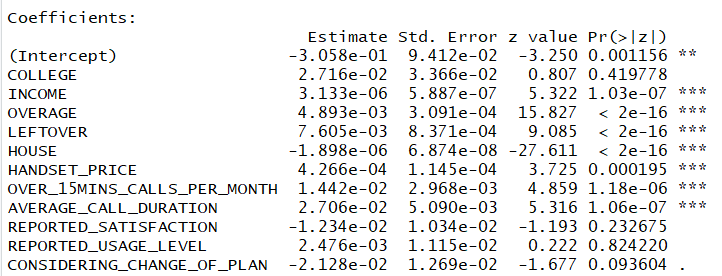


Figure 4: Logistic Regression Parameters and Estimated Coefficients

* + 1. By looking at the p-value (Pr(>|z|) of each regression coefficient (Estimate), we could see that the features INCOME, OVERAGE, LEFTOVER, HOUSE, HANDSET\_PRICE, OVER\_15MINS\_CALLS\_PER\_MONTH and AVERAGE\_CALL\_DURATION is statistically significant. Except HOUSE feature, they are all positively correlated with the odds ratio of customers churn, which means the customers are more likely to churn if the values of the features increase.
    2. The generalization accuracy of this model is 64.15%, 13.5% higher than the baseline model selected.

1. **Comparison between Tree Induction and Logistic Regression Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Description** | **Number of Correct Prediction** | **ACCURACY (%)** |
| 1 | Baseline model using majority votes | 2030 | 50.75 |
| 2 | Tree induction model | **2822** | **70.55** |
| 3 | Logistic regression model | 2566 | 64.15 |

Table 4: Comparison between baseline model, tree induction and logistic regression models

* 1. From the results above, we could see that both tree induction and logistic regression model performs better than the baseline model using majority votes.
  2. Tree induction model apparently has better performance than logistic regression model. The generalization accuracy of tree induction model is 6.4% higher than the logistic regression model, with 256 more customers’ churn status predicted correctly. This result is not surprising because in order to build a logistic regression model with good performance, careful feature engineering and tedious data preprocessing steps are required. Logistic regression model is also more sensitive towards outliers. On the other hand, tree induction using C5.0 algorithm post-prune the tree by first growing a large tree that overfits the training data, follow by removing the nodes and branches which have little effect on the classification errors. This could reduce the potential of overfitting.

### **Learning Curves Plots (Question 1b)**

The same train set which we used for building models in the section above was repeatedly cut into half to train and build the models again. Note that same hyperparameter settings was used here.

The outcome is depicted in Table 5, Figure 5 and Figure 6.

From the result, we can see that generalization accuracy of both decision tree and logistic regression models increases as the number of training instances increases. However, the performance of logistic regression models is relatively unstable and fluctuates throughout the varying number of training instances. Logistic regression model performs better when the size of training set is small (number of training instances less than 250), but when the size of training set gets bigger, decision tree model performs better. The performance of decision tree model surpasses logistic regression model starting from 250 training instances and its generalization accuracy increases gradually afterwards.

|  |  |  |
| --- | --- | --- |
| **Number of Training Instances** | **Accuracy of Decision Tree Model (%)** | **Accuracy of Logistic Regression Model (%)** |
| 0 | 0.00 | 0.00 |
| 2 | 50.75 | 50.75 |
| 4 | 50.75 | 48.93 |
| 8 | 50.75 | 54.38 |
| 16 | 50.75 | 55.53 |
| 32 | 50.75 | 52.25 |
| 63 | 55.00 | 62.23 |
| 125 | 59.50 | 62.98 |
| 250 | 61.23 | 60.68 |
| 500 | 67.85 | 62.93 |
| 1000 | 67.70 | 63.83 |
| 2000 | 68.95 | 63.43 |
| 4000 | 70.13 | 63.63 |
| 8000 | 69.90 | 63.93 |
| 16000 | **70.55** | **64.15** |

Table 5: Accuracy of Models Based on Different Number of Training Instances

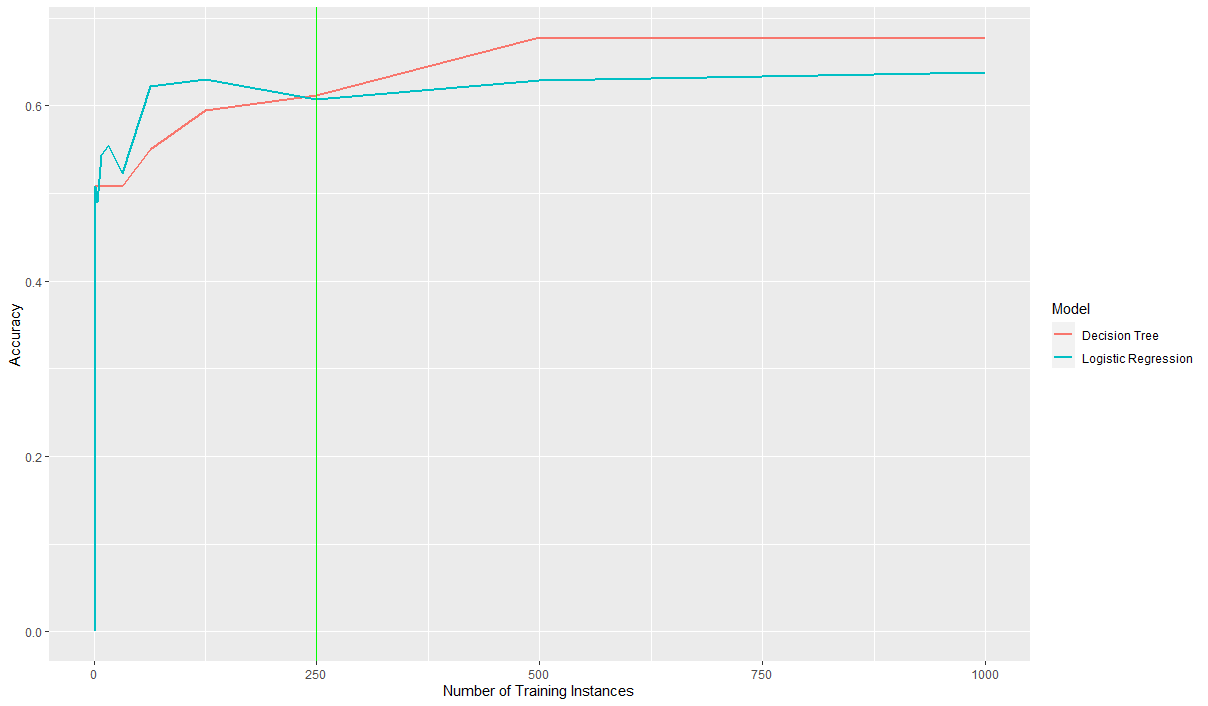
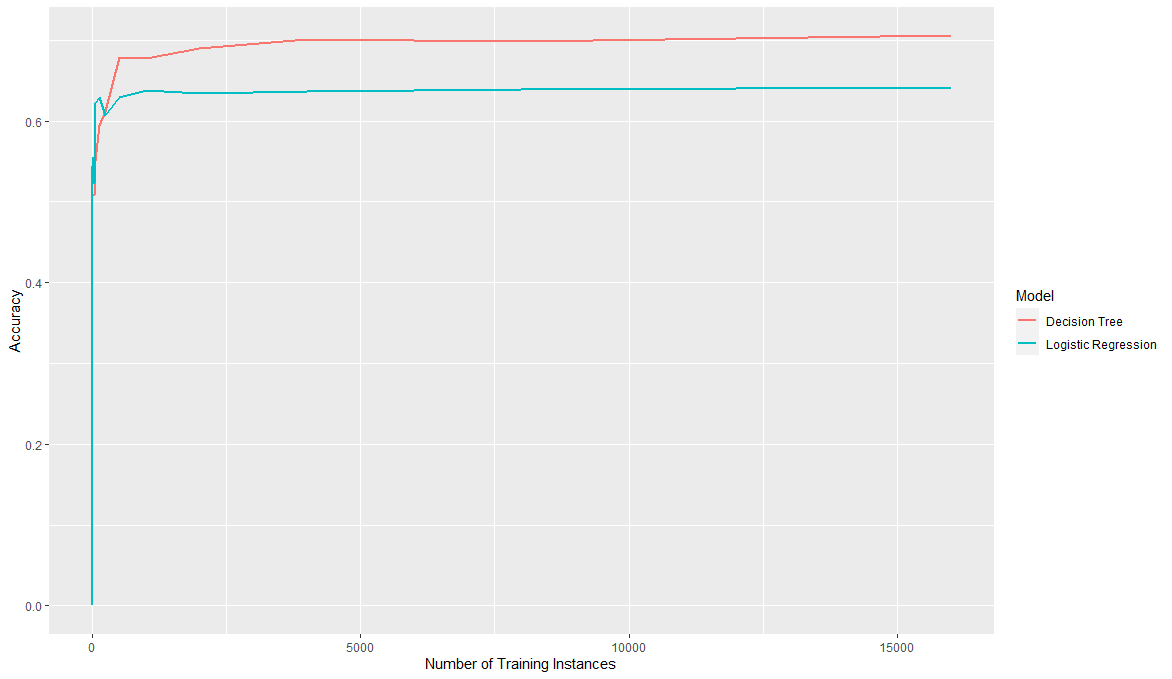
 

Figure 5: Learning Curve Plot (16000 training instances and below)

Figure 6: Learning Curve Plot (1000 training instances and below)

### **Segmentation of Customers based on Decision Tree Model (Question 2)**

Based on the classification tree model built in 1(a), there are 19 segments of customers. Based on the number of customers in train set, he biggest two main groups are 5422 customers with value of dwelling more than $600469 and 10578 customers with value of dwelling less than or equal to $600469. Under these two huge groups, the customers are further divided into more segments. 4 most noticeable segments (with number of customers more than 1000) are tabulated in Table 6.

|  |  |  |  |
| --- | --- | --- | --- |
| **Segment** | **Characteristics** | **Number of Customers** | **Ratio of Churn Customer within Segment (%)** |
| 1 | * Value of dwelling more than $600469 * Annual income less than or equal to $100359 | 3651 | 18.87 |
| 2 | * Value of dwelling less than or equal to $600469 * Average overcharges per month more than $105 | 3526 | 79.72 |
| 3 | * Value of dwelling less than or equal to $600469 * Average overcharges per month less than or equal to $105 * Average % leftover minutes per month more than 24 | 2354 | 60.11 |
| 4 | * Value of dwelling less than or equal to $600469 * Average overcharges per month less than or equal to $105 * Average % leftover minutes per month less than 24 but NOT 0 * Annual income more than $49479 | 1575 | 26.79 |

Table 6: Representative Customer Segments based on Classification Tree Model

Based on 4 segments above, we could say that customers with value of dwelling less than or equal to $600469 are more likely to churn. For the customers with lower overcharges per month, those with more leftover minutes are more likely to stay than those with less leftover minutes. We hypothesize that customers with expensive house but relatively low income (could be possibly due to the inaccurate information provided by customers) are more likely to stay. Moreover, customers who use the service less frequently but are overcharged are more likely to churn.

### **Model Building by Varying Model Complexity (Question 3)**

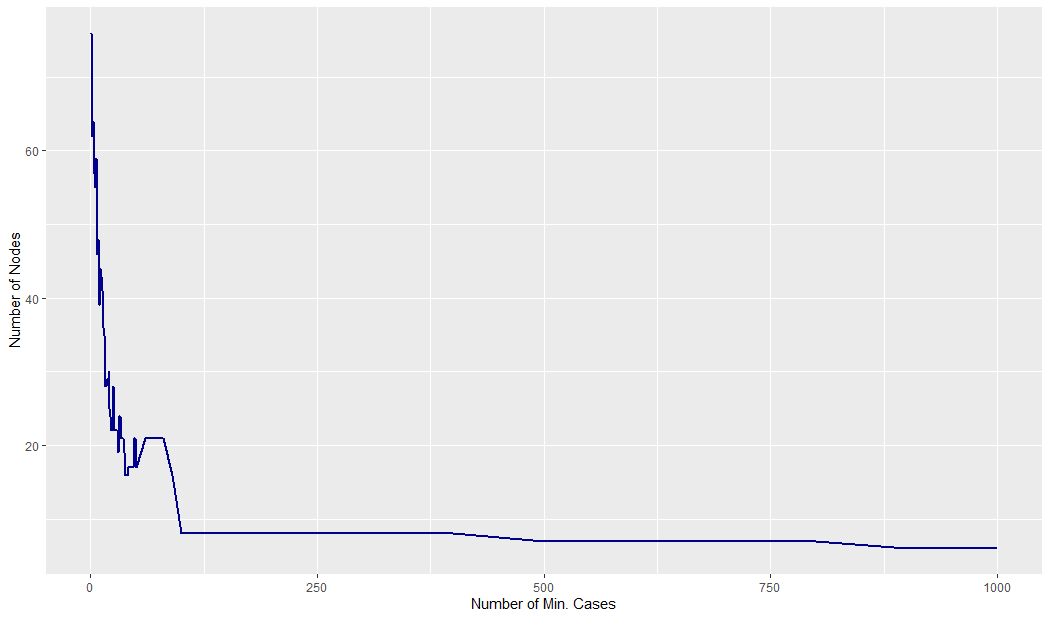
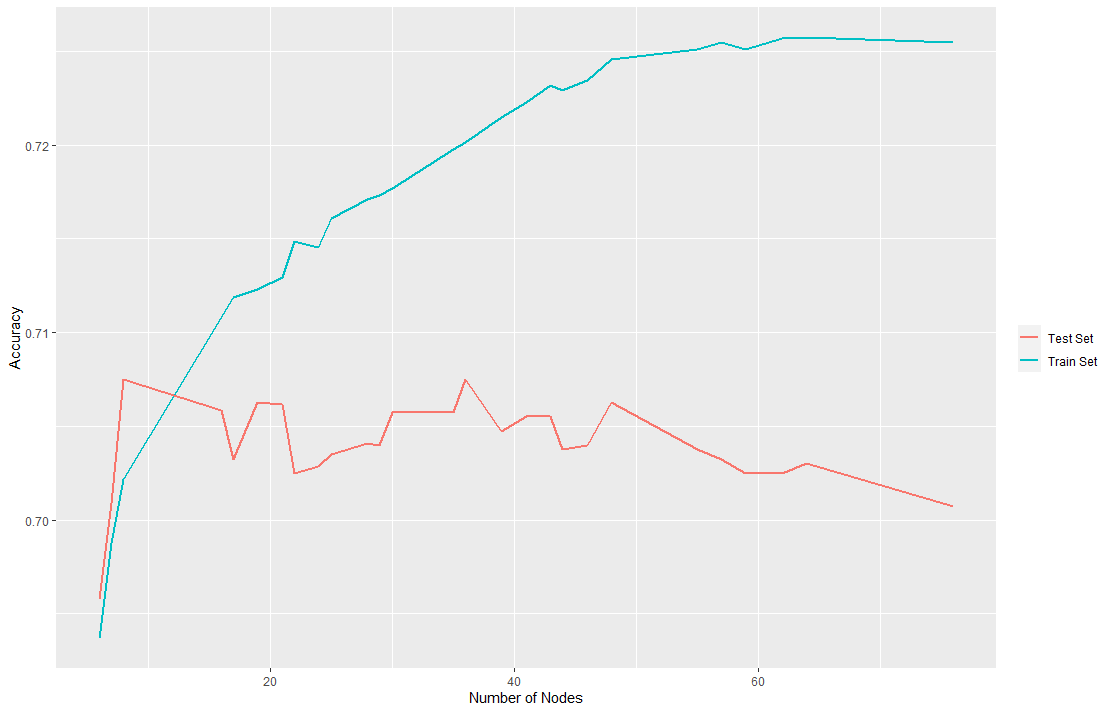
1. **Classification Tree Model**
   1. Minimum number of samples (hyperparameter ‘minCases’) that must be put in at least two of the splits was varied to control the classification tree model’s complexity. In general, models with lower value of minCases produce higher number of nodes (Figure 7). Higher value of minCases might be able to prevent overfitting but could also cause underfitting when the value is too high.
   2. For classification tree models which produced same number of nodes but having different value of minCases, the final accuracy is calculated as the average of these models’ accuracy.

Figure 7: Number of Nodes vs Number of minimum number of samples that must be put in at least two of the splits

* 1. The fitting graph of the model’s generalization accuracy with varying number of nodes is depicted in Figure 8. From the fitting graph, we could see that the training accuracy increases from lowest number of nodes to highest number of nodes with slight fluctuation. As for generalization accuracy, it starts to decrease gradually from 36 nodes (**sweet spot** as shown in the rectangular box). The generalization accuracy at the sweet spot is 70.75%. The gap between generalization and training accuracy becomes wider after the sweet spot. However, it is interesting to note that when the number of nodes is less than 10, the generalization accuracy is higher than the training accuracy. Although the difference in accuracy is not too big, it could be due to slight difference between the distribution of train and test dataset.
  2. The classification tree model built at 1(a) contains 19 nodes, and its generalization accuracy is only 0.2% less than the model built at sweet spot. Hence, we can say that the model built at 1(a) performs fairly well as it does not overfit and at the same time does not really underfit.

Figure 8: Fitting Graph of Classification Tree Model



1. **Logistic Regression Model**
   1. The complexity of logistic regression model is varied by the number of features used to build the model. In general, just like a linear mathematical function, a regression model is more complicated when the number of features included increases. For each number of feature(s), the model with combination of variable(s) which performs best is selected and presented in this report (Table 6). For example, for a combination of 2 features, only logistic regression model which uses OVERAGE and HOUSE features is selected for comparison.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Predictors** | | | | | | | | | | | |
| **Variables** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **14** |
| COLLEGE | X | X | X | X | X | X | X | X | X | √ | √ | √ |
| INCOME | X | X | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| OVERAGE | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| LEFTOVER | X | X | X | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| HOUSE | X | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| HANDSET\_PRICE | X | X | X | X | X | X | √ | √ | √ | √ | √ | √ |
| OVER\_15MINS\_CALLS\_PER\_MONTH | X | X | X | X | X | √ | √ | √ | √ | √ | √ | √ |
| AVERAGE\_CALL\_DURATION | X | X | X | X | √ | √ | √ | √ | √ | √ | √ | √ |
| REPORTED\_SATISFACTION | X | X | X | X | X | X | X | X | √ | √ | √ | √ |
| REPORTED\_USAGE\_LEVEL | X | X | X | X | X | X | X | X | X | √ | √ | √ |
| CONSIDERING\_CHANGE\_OF\_PLAN | X | X | X | X | X | X | X | √ | √ | X | √ | √ |
| INCOME:HANDSET\_PRICE | X | X | X | X | X | X | X | X | X | X | X | √ |
| OVERAGE: OVER\_15MINS\_CALLS\_PER\_MONTH | X | X | X | X | X | X | X | X | X | X | X | √ |
| LEFTOVER: AVERAGE\_CALL\_DURATION | X | X | X | X | X | X | X | X | X | X | X | √ |
| **Training Accuracy (%)** | 61.16 | 62.36 | 63.19 | 63.82 | 64.01 | 64.08 | 64.26 | 64.16 | 64.29 | 64.19 | 64.19 | **64.66** |
| **Generalization Accuracy (%)** | 60.83 | 62.13 | 62.93 | 63.50 | 63.65 | 63.60 | 63.80 | 63.68 | 63.75 | 63.85 | 64.15 | **64.33** |

Table 7: Logistic Regression Models with Different Combination of Variable(s) and Accuracy

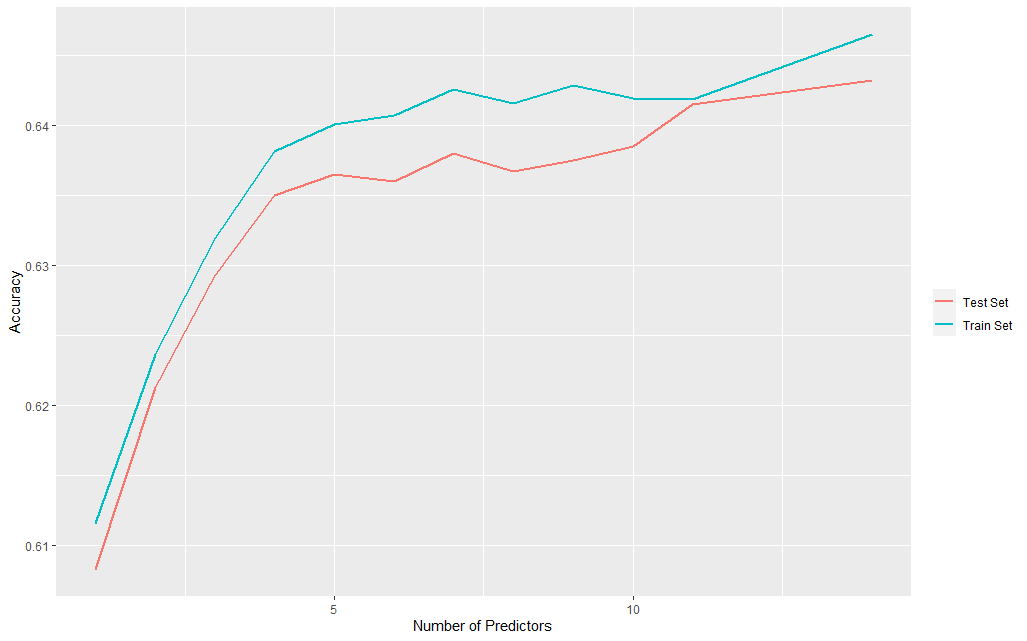
* 1. A model which includes interaction effects INCOME:HANDSET\_PRICE, OVERAGE: OVER\_15MINS\_CALLS\_PER\_MONTH, and LEFTOVER: AVERAGE\_CALL\_DURATION was also built to check if a more complicated model would outperform the model we built in 1(a).
  2. The fitting graph of logistic regression model with different number of predictors used is depicted in Figure 9. Overall, both the training accuracy and generalization accuracy increases as the number of predictors increases. No sweet spot is discovered on this fitting graph. It could be due to the nature of regression model in which the more the variables/predictors added, the more the variation of dataset is explained by the model.

Figure 9: Fitting Graph of Logistic Regression Model

### **Cross-Validation (Question 4)**

10-fold cross-validation (without repetition) is applied during process of building classification tree model and logistic regression tree model, using all features. The average score of all folds’ accuracies of each model after using cross validation is depicted in Table 8.

|  |  |
| --- | --- |
| **Model** | **Accuracy (%)** |
| Classification Tree | **70.12** |
| Logistic Regression | 64.12 |

Table 8: Average Scores of the Accuracies from the Cross-Validation Procedure

By comparing the average score of the accuracies, classification tree model performs better (higher accuracy) than logistic regression model. In fact, if we look at each fold’s accuracy (Figure 10 and Figure 11), classification tree model performs better than logistic regression model in all folds.

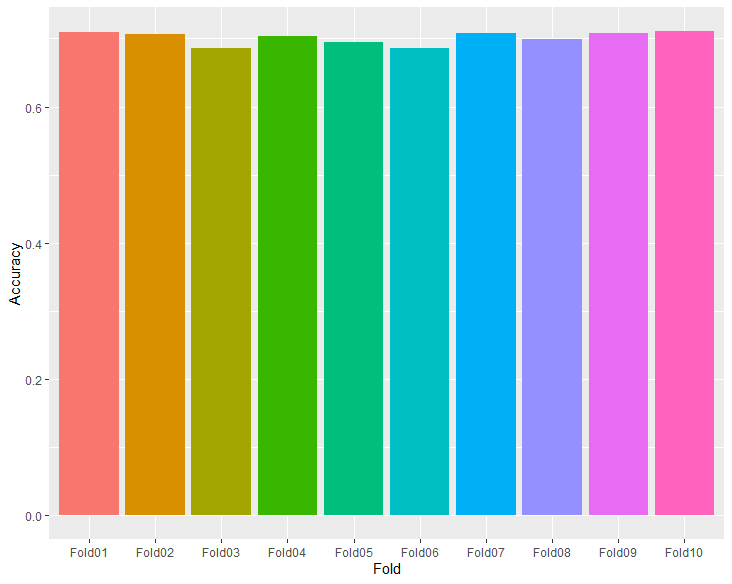


Figure 10: Accuracy of Each Fold for Classification Tree Model

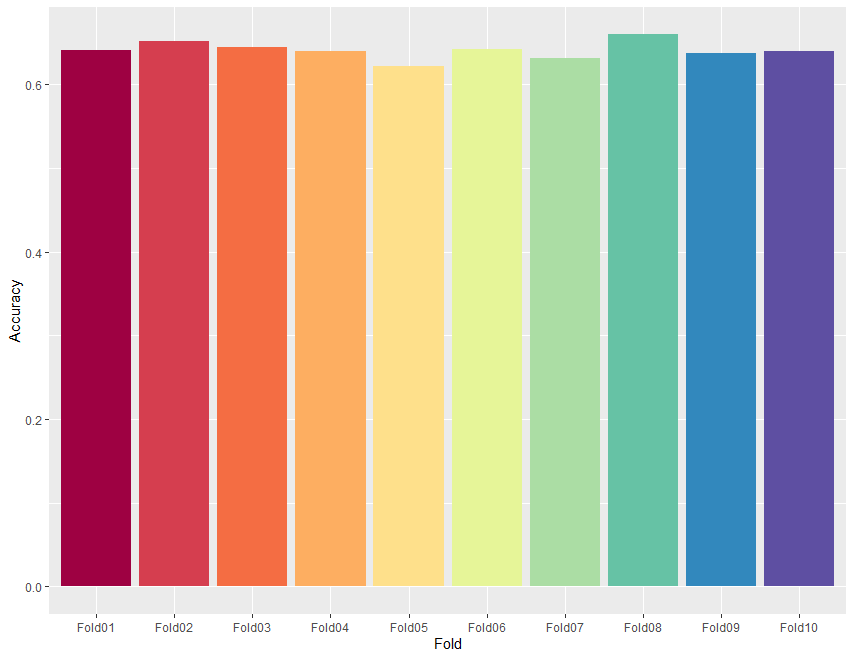


Figure 11: Accuracy of Each Fold for Logistic Regression Model

### **Ensemble Modeling Approach (Question 5)**

Two ensemble modeling approach (using holdout sampling approach) were performed for classification tree model: Random Forest (3 variables selected at each split) and Boosting (3 iterations) of C5.0 algorithm. In addition, a simple ensemble method which makes prediction based on the majority votes of top 3 models (Random Forest, C5.0 without boosting, C5.0 with boosting) with highest accuracy was used too. The result is depicted in Table 9.

|  |  |
| --- | --- |
| **Ensemble Method** | **Accuracy (%)** |
| Random Forest | 69.90 |
| Boosting (C5.0) | **70.63** |
| Simple Ensemble | 70.53 |

Table 9: Ensemble Modeling Approach and Respective Accuracy

From the result above, by using suitable number of boosting iterations on C5.0 algorithm, the performance of the classification model could be improved.

Cross-validation procedure and different number of randomly selected variables were also attempted on Random Forest, and the performance is worse than the C5.0 algorithm. The result is depicted in Table 10.

|  |  |
| --- | --- |
| **Number of Variables Selected at Each Split** | **Accuracy (%)** |
| 2 | 69.68 |
| 6 | 69.58 |
| 11 | 69.35 |

Table 10: Performance of Random Forest Approach with Different Hyperparameter

### **Discussion (Question 6)**

For classification tree model using C5.0 algorithm, boosting using suitable hyperparameters improves the performance. Overfitting occurred when minimum number of samples that must be put in at least two of the splits decreased to certain threshold while other hyperparameters were hold fixed. By using boosting, the model performance was improved but it depends on the optimal settings of hyperparameters. Random Forest performed worse than C5.0 most probably because the number of features selected are not optimal and selection of features in some trees were not significant for the prediction.

For logistic regression models, adding more variables could increase the prediction accuracy but the model would become too complicated to analyze. Due to the nature of logistic regression, no overfitting was observed when more variables were added to build the model.

One obvious advantage of cross-validation procedure over holdout sampling approach, we could fully utilize all available data for training and testing. This provides better model evaluation as the model is exposed to different distribution of data at each fold.

In terms of feature selection, I think based on this dataset, variables LEFTOVER, HOUSE, OVERAGE and INCOME are the most important variables for the prediction of churning. By just looking the attributes usage percentage in classification tree and p-value of the variables in logistic regression, these 4 variables are ranked at the top. This is also in accordance with the customer segmentation which we explained in Segmentation of Customers based on Decision Tree Model (Question 2). Among all the models built above, I think that classification tree using C5.0 algorithm is a better model for prediction of churning here. The reason is classification tree built using C5.0 algorithm is post-pruned which can avoid overfitting, and could achieve high accuracy without loss of high interpretability and explainability.