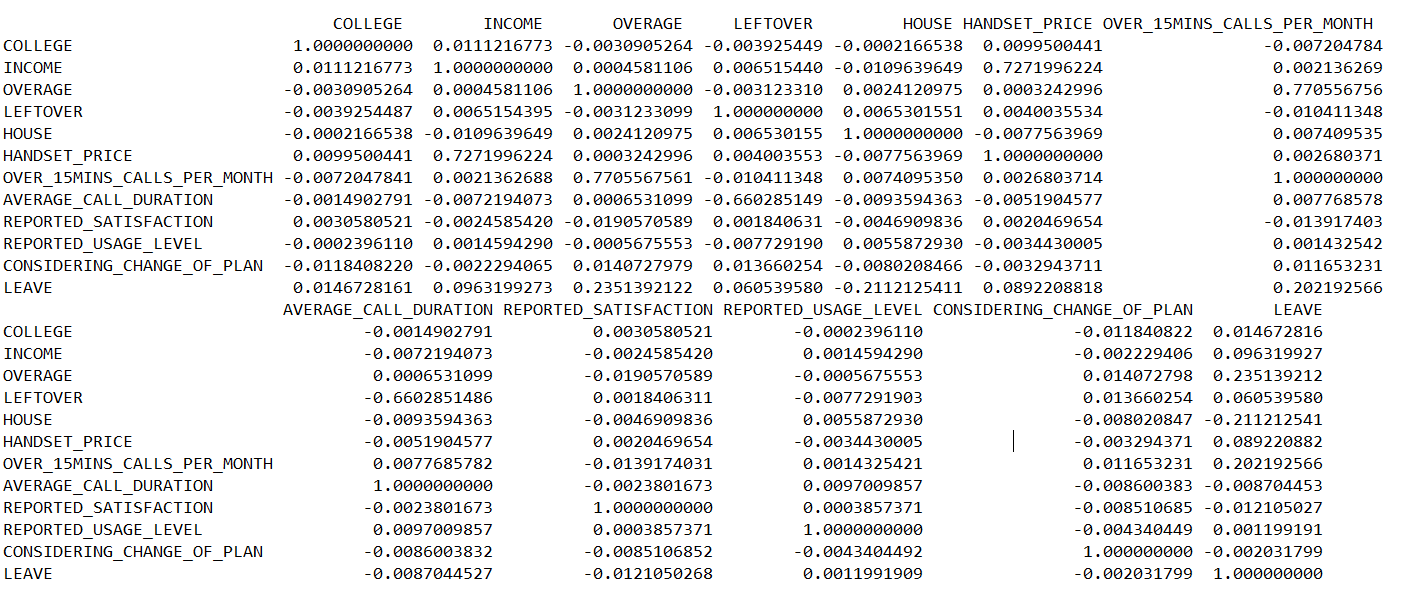
KSE521 Business Intelligence

Homework #7

### Data Understanding

1. Missing values inspection
   1. None of the variables contain missing values.
2. Outliers inspection
   1. None of the variables contain outliers.
3. Correlation checking
   1. This was performed after item 1 in Data Preprocessing.
   2. The correlation matrix is as below.
   3. A few important observations were obtained from the correlation matrix.
      1. OVERAGE and OVER\_15MINS\_CALLS\_PER\_MONTH are highly positively correlated by 0.77.
      2. HANDSET\_PRICE and INCOME are highly positively correlated by 0.73.
      3. LEFTOVER and AVERAGE\_CALL\_DURATION are moderately negatively correlated by 0.66.
      4. The variables with highest correlation with the target variable LEAVE are OVERAGE, HOUSE and OVER\_15MINS\_CALLS\_PER\_MONTH. However, since both OVERAGE and OVER\_15MINS\_CALLS\_PER\_MONTH are highly correlated, we might need to consider dropping one of them for building models later.

### 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 : Conversion of Variables

1. Equi-frequency Discretization
   1. This was only performed for logistic regression.
   2. Variables LEFTOVER, INCOME, AVERAGE\_CALL\_DURATION, HANDSET\_PRICE, OVERAGE, HOUSE and OVER\_15MINS\_CALLS\_PER\_MONTH were converted to 3 bins with almost same number of records respectively.
2. Negative Value Replacement
   1. There is one case with average overcharges (OVERCHARGE) per month = -2. The average overcharges is replaced by 0 for this observation.

### Model Building (using R)

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).

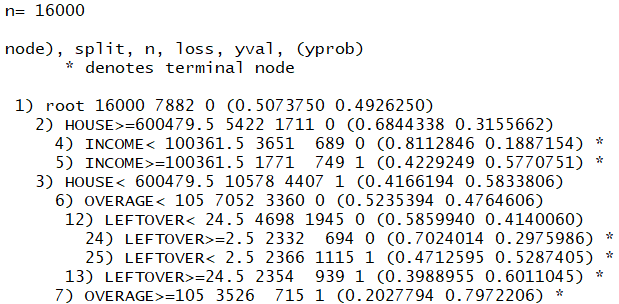
1. **Decision tree models** (Q1 AND Q2)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **CP** | **MINSPLIT** | **MAXDEPTH** | **ACCURACY (%)** | **NOTE** |
| 1 | 0.01 | 20 | 30 | 69.575 | Default settings |
| 2 | 0.05 | 20 | 30 | 67.525 | - |
| 3 | 0.005 | 20/500/1000 | 30 | 70.75 | - |
| 4 | 0.005 | 2000 | 30 | 69.575 | - |
| 5 | 0.001 | 20 | 30 | 70.5 | - |
| 6 | 0.001 | 20 | 5 | 70.15 | - |
| 7 | 0.001 | 500 | 30 | **70.8** | - |
| 8 | 0.001 | 1000 | 30 | 70.75 | - |
| 9 | 0.0005 | 20 | 30 | 70.5 | - |

Table : Decision Tree Model's Accuracy based on Different Parameter Settings

If we run the decision tree model with R using all default parameter settings, the accuracy is 69.575%.

The printed summary of model built based on the training data set from R is as below.



*\*You may skip the paragraphs below and go straight to plotted decision tree next page for visualization.*

From the printed result above, 1) root is the root node before decision tree starts splitting, where there are 16000 training observations with 50.74% of customers who stay (class ‘0’) and 49.26% of customers who churn (class ‘1’). The first level splits on variable HOUSE with splitting point 600479.5 into two children nodes 2) and 3).

Under node 2) HOUSE>=600479.5, the subtree further splits on variable INCOME with splitting point 100361.5 into two terminal nodes 4) and 5). At terminal node 4) INCOME<100361.5, there are 3651 observations with 81.13% of customers who stay and 18.87% of customers who churn. At terminal node 5) INCOME>=100361.5, there are 1771 observations with 42.29% of customers who stay and 57.71% of customers who churn.

Under node 3) HOUSE<600479.5, the subtree further splits on variable OVERAGE with splitting point 105 into two children nodes 6) and 7). At terminal node 7) OVERAGE>=105, there are 3526 observations with 20.28% of customers who stay and 79.72% of customers who churn. At children node 6) OVERAGE<105, the subtree further splits on variable LEFTOVER with splitting point 24.5 into two children nodes 12) and 13). At terminal node 13) LEFTOVER>=24.5, there are 2354 observations with 39.89% of customers who stay and 60.11% of customers who churn. At children node 14) LEFTOVER<24.5, the subtree further splits on variable LEFTOVER with splitting point 2.5 into two terminal nodes 24) and 25). At terminal node 24) LEFTOVE>=2.5, there are 2332 observations with 70.24% of customers who stay and 29.76% of customers who churn. At terminal node 25) LEFTOVE<2.5, there are 2354 observations with 47.13% of customers who stay and 52.87% of customers who churn.

To visualize the decision tree model better, please refer to the following diagram.

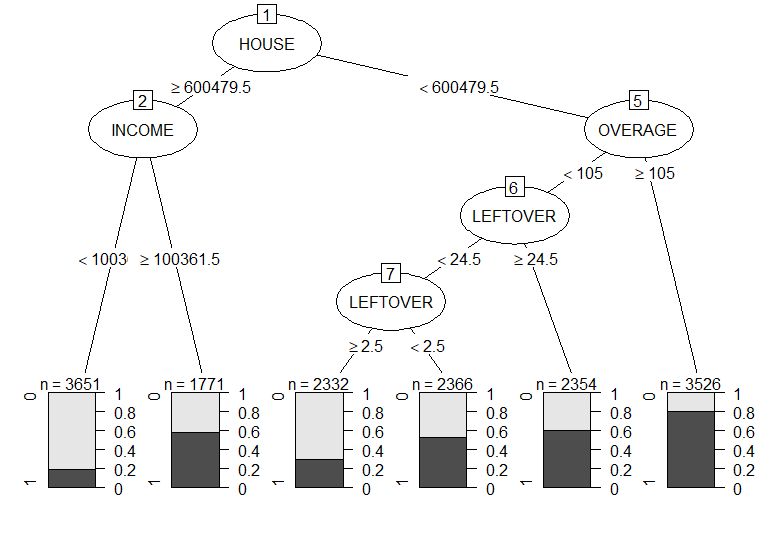


Figure : Decision Tree Model 1 (CP=0.01)

A few decision tree models with alternative parameters settings were built as shown in Table 2. The model with higher complexity parameter (CP) performs poorer (lower accuracy) than default decision tree model (CP=0.01) while models with lower CP generally perform better (higher accuracy). However, for CP=0.005, the model performance remains the same until we increase the minimum number of instances of a node for further splitting (MINSPLIT) to 2000. The model shows the same performance at MINSPLIT=2000 as compared to default model due to underfitting most possibly. For CP=0.001, all the models perform better than default model, but it is interesting to note that if we hold the MINSPLIT constant and decrease the maximum number of levels the tree can grow (MAXDEPTH), the accuracy drops; and if we hold the MAXDEPTH constant and increase MINSPLIT, the accuracy increases first but decreases later as the value of MINSPLIT increases.

Overall, Model 7 produces the highest accuracy. All decision tree models were built based on at least 4 variables (except the model 2), namely, HOUSE, LEFTOVER, OVERAGE and INCOME. However, we can see that different splitting point and different splitting level can affect the accuracy. For example, if we compare Model 1 and Model 8, both of them use exactly same set of variables for building the decision tree, but Model 8 produces higher accuracy than Model 1.

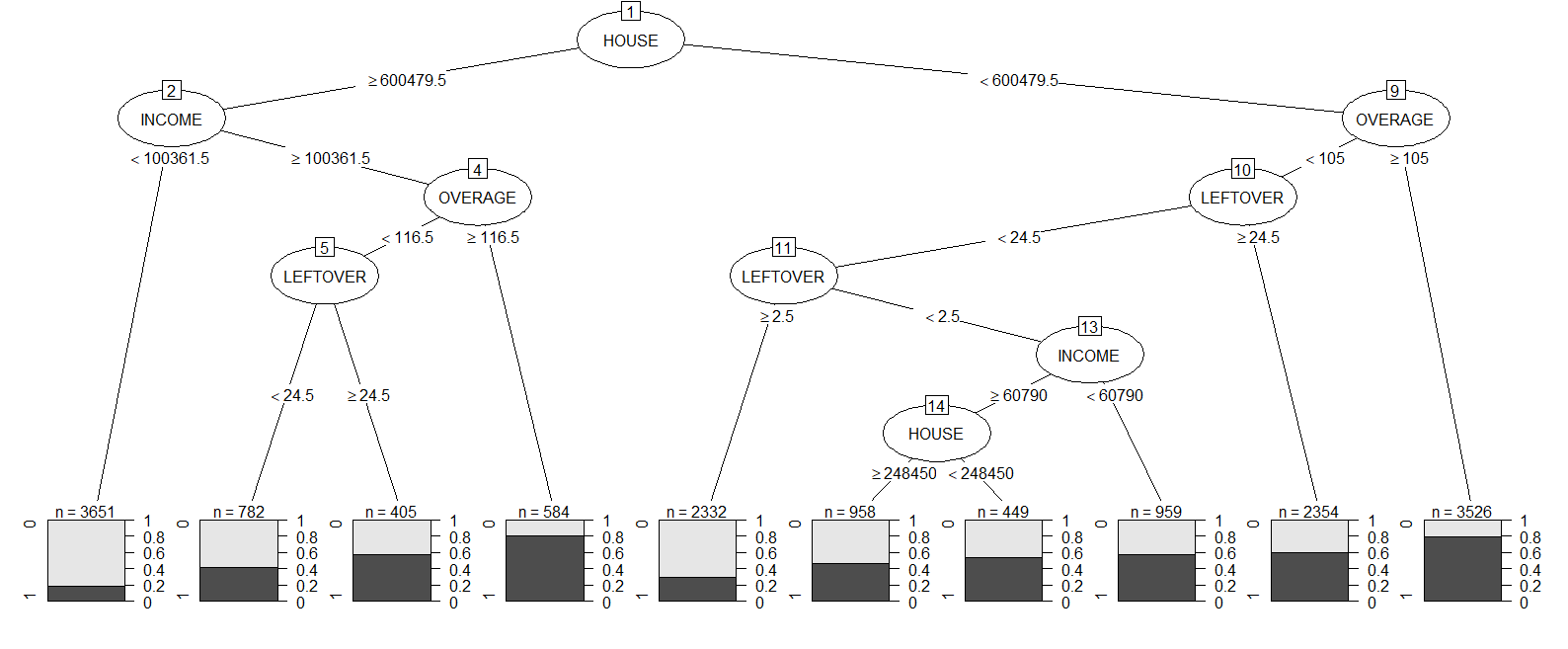


Figure : Decision Tree Model 8 (CP=0.001, MINSPLIT=1000)

1. **Logistic regression model** (Q3)
   1. Models with different combination of variables and interaction effects were built and compared as shown in table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | | | | | |
| **Variables** | **1** | **2** | **3** | **4** | **5** | **6** |
| COLLEGE | √ | X | X | X | X | X |
| INCOME | √ | √ | √ | X | X | √ |
| OVERAGE | √ | √ | √ | √ | √ | √ |
| LEFTOVER | √ | √ | √ | √ | √ | √ |
| HOUSE | √ | √ | √ | √ | √ | √ |
| HANDSET\_PRICE | √ | √ | √ | X | X | X |
| OVER\_15MINS\_CALLS\_PER\_MONTH | √ | √ | √ | √ | X | X |
| AVERAGE\_CALL\_DURATION | √ | √ | √ | √ | X | X |
| REPORTED\_SATISFACTION | √ | X | X | X | X | X |
| REPORTED\_USAGE\_LEVEL | √ | X | X | X | X | X |
| CONSIDERING\_CHANGE\_OF\_PLAN | √ | X | X | X | X | X |
| INCOME:HANDSET\_PRICE | X | X | √ | √ | √ | X |
| OVERAGE: OVER\_15MINS\_CALLS\_PER\_MONTH | X | X | √ | √ | √ | X |
| LEFTOVER: AVERAGE\_CALL\_DURATION | X | X | √ | √ | √ | X |
| **Accuracy (%)** | 64.3 | 63.8 | 64.2 | **64.4** | 63.8 | 63.5 |

Table : Logistic Regression Model's Accuracy based on Different Combination of Variables

* 1. From the results above, it can be seen that the performance between the models only varies between 0.1 to 0.8%, which is very little. Hence, if we were to use logistic regression on this dataset, we could pick the one with least variables and without interaction effects for simpler interpretation of the model.
  2. Interesting result: By using same train and test set, but using only 4 variables INCOME, OVERAGE, LEFTOVER and HOUSE, and **applying discretization (page 2) to numerical variables**, the accuracy of logistic regression model increases to 66.675%.

1. **SVM Model** (Radial Basis Function Kernel) (Q4)
   1. Models with different combination of variables were built and compared as shown in table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Models** | | | | | |
| **Variables** | **1** | **2** | **3** | **4\*** | **5** | **6** |
| COLLEGE | √ | X | X | X | X | X |
| INCOME | √ | √ | √ | √ | √ | √ |
| OVERAGE | √ | √ | √ | √ | √ | √ |
| LEFTOVER | √ | √ | √ | √ | √ | √ |
| HOUSE | √ | √ | √ | √ | √ | √ |
| HANDSET\_PRICE | √ | √ | X | X | X | X |
| OVER\_15MINS\_CALLS\_PER\_MONTH | √ | √ | X | X | √ | √ |
| AVERAGE\_CALL\_DURATION | √ | √ | X | X | X | √ |
| REPORTED\_SATISFACTION | √ | X | X | X | X | X |
| REPORTED\_USAGE\_LEVEL | √ | X | X | X | X | X |
| CONSIDERING\_CHANGE\_OF\_PLAN | √ | X | X | X | X | X |
| **Accuracy (%)** | 67.6 | 68.8 | 69.4 | **69.9** | 68.4 | 69.0 |

Table : SVM Model's Accuracy based on Different Parameter Settings and Combination of Variables (Model 4’s hyperparameters (C=1.0, Gamma=2.0) are tuned based on Model 3)

* 1. From the results above, we could see that Model 4 (tuned based on Model 3) with only 4 variables INCOME, OVERAGE, LEFTOVER, and HOUSE performs the best (achieves highest accuracy).
  2. SVM model achieves lower accuracy than decision tree model, but higher accuracy then logistic regression model. When redundant variables are removed, as compared to logistic regression model, SVM model archives slightly higher accuracy.

### Discussion (Q5)

For decision tree models, overfitting occurs when the value of CP is too low (the tree is too complex) without adjustment of other hyperparameters; underfitting occurs when the value of CP is high (the tree is too simple). One obvious weakness of decision tree induction here is the model performance could be unstable or even much worse if there are variations in training data (setting different seed).

For logistic regression models, adding more variables could increase the prediction accuracy but the model would become too complicated. Besides, removing redundant variables (highly correlated variables) somehow did not help in achieving higher accuracy.

As compared to both decision tree and logistic regression models, it is more difficult to explain to business users about the rules or prediction criteria of SVM model. Moreover, adding more variables (disregard interaction effect) to SVM model does not guarantee higher accuracy like logistic regression. The computational time and cost took to build SVM model is much higher than building decision tree and logistic regression models.

In terms of feature selection, we could see that all the models with highest accuracy do not favor the categorical variables such as customers’ personal information (COLLEGE), customers’ opinion (CONSIDERING\_CHANGE\_OF\_PLAN, REPORTED\_SATISFACTION), and customers’ self-reported usage level (REPORTED\_USAGE\_LEVEL). This suggests that a better metric needs to be designed in the future for collecting customers’ opinion and feedback, so that the data collected could be made useful to prediction of churning.

In my opinion, based on this dataset, variables LEFTOVER, HOUSE, OVERAGE and INCOME are the most important variables for the prediction of churning. From 3 types of models built above, I think that decision tree is a better model for prediction of churning here. The reason is decision tree requires the least effort to build, computationally inexpensive, best interpretability and could achieve high accuracy easier with proper pruning.