BAN404 Assignment 2

a. Describe relevant features of the input and output variables with descriptive statistics.

The marketing dataset from *ElemStatLearn* package is composed of 14 variables, with 8993 observations. Those 14 demographic variables include *Sex*, *Age*, *Education*, and so on. First, we run some functions to have a brief understanding of the dataset to get information such as the class of individual variable, and whether the dataset contains any missing value, and we find 2694 missing values from the original dataset. Hence, we remove all the missing values in order to get a more accurate result.

```
summary(marketing)
```

```
  Sex     Marital  Age      Edu      Occupation   Lived   Dual_Income
  1:3067  1:2652   1:647    1:176    1     :2333  1:228   1:4114
  2:3809  2:536    2:1610   2:787    6     :1128  2:850   2:1742
          3:673    3:1768   3:1479   4     :834   3:594   3:1020
          4:202    4:1267   4:2407   2     :617   4:758
          5:2813   5:703    5:1207   3     :590   5:4446
                   6:488    6: 820   5     :504
                   7:393             (Other):870


  Household    Householdu18  Status  Home_Type  Ethnic      Language
  2     :2156  0     :4276   1:2584  1:4102     7     :4605 1:6277
  3     :1357  1     :1215   2:2882  2:521      5     :873  2:397
  1     :1244  2     :908    3:1410  3:1895     3     :631  3:202
  4     :1208  3     :321            4:102      2     :379
  5     :549   4     :91             5:256      8     :176
  6     :193   5     :37                        1     :112
 (Other):169  (Other):28                       (Other):100

  high
  0:5064
  1:1812
```

Table 1: Summary of cleaned marketing data

In the assignment, our task is to analyze the qualitive variables, therefore we need to transform our data from integer and numeric data type into factor. Then we create a new variable *high* by categorizing the variable *Income* into two groups: when the value of an *Income* variable is lower than $50,000, we set the dummy variable equal to 0; and when the value of an *Income* variable is equal or higher than $50,000, we set the dummy variable equal to 1. The summary above provides a view of preliminary information of cleaned data.

Next, we continue to study the distributions of each variable with the help of histograms (Figure 1). As the graphs show in Figure 1, the majority of variables show a sign of

skewness. The light grey color indicates the proportion of high-income, whereas the dark grey color indicates the low-income part.

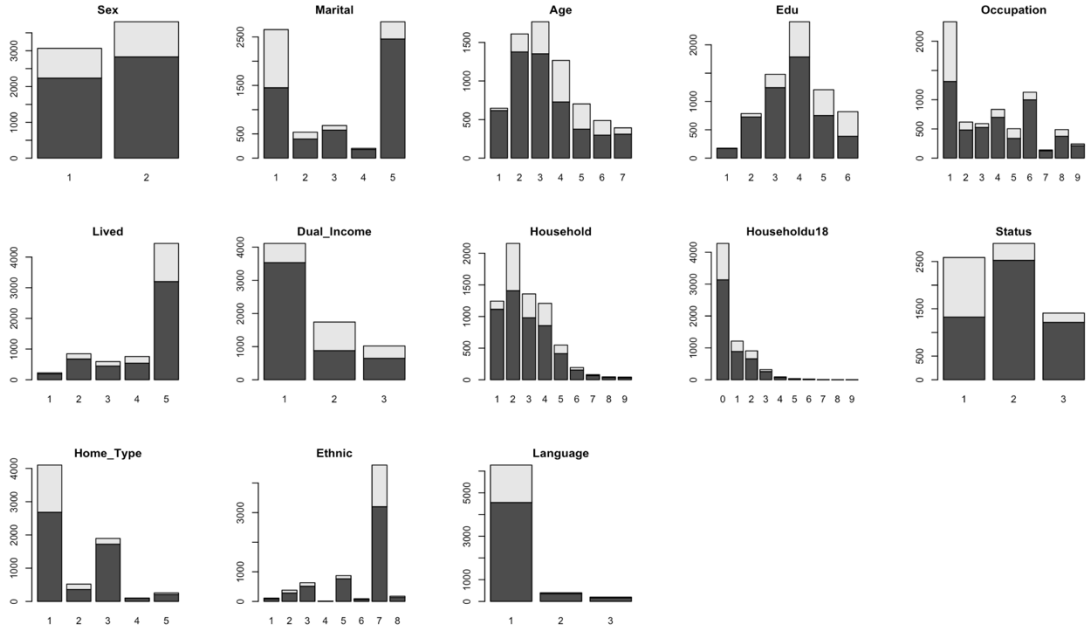

Figure 1: The distribution of variables

For categorical variables, it is necessary to measure the strength of association between variables to see whether the strength of relationship is strong or weak. For this measurement, we use *Goodman and Kruskal's tau* method since we deal with nominal categorical predicators. The Figure 2 is the association plot in which the K value indicates the number of unique levels for each variable. The off-diagonal elements contain the forward and backward tau measures for each variable pair. The numerical values appearing in each row represent the association measure from the variable indicated in the row name to the variable indicated in the column name. Most of the variables have weak association between each other, except the relationship between *Dual_Income* and *Marital* variables. Those two variables have relatively high forward and backward association (*Marital* forward association with *Dual_Income* is equal to 0.58, and *Dual_Income*'s forward association with *Marital* equals 0.45), which suggests that the variable *Marital* has higher prediction over *Dual_Income*. The association implies that if the *Marital* variable is known, it's easier to predict which level the *Dual_Income* variable belongs to.

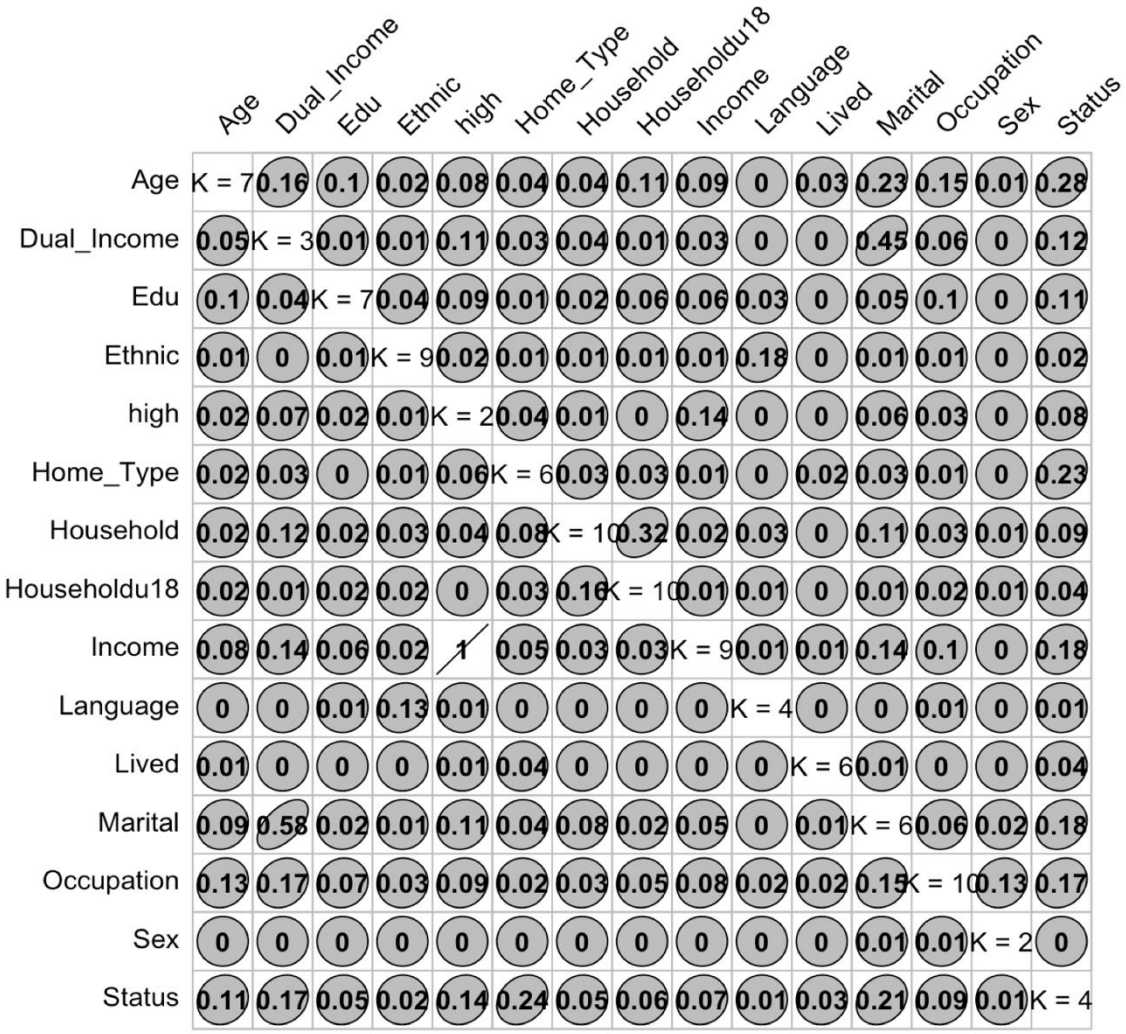

Figure 2: Strength of association between variables

b. Use different methods to predict high.

1. LOGISTIC REGRESSION

We fit a logistic regression model to predict *high* using all the 13 categorical variables. There are estimated coefficients for each level except the first level of the categorical predictors. We can tell if the coefficients are statistically significant from the z-values and the p-values, and find that for the variables *Lived, Ethnic* and *Language*, none of their levels' coefficients are significant.

We also find out that for the variable *Householdu18,* the coefficients of last six levels are insignificant. Furthermore, there are only a few observations in these levels, among which there is only one observation in the level "8", causing new level in test set when we apply k-fold cross validation or leave one out cross validation. Therefore, we merge the last seven levels, then the new classification of *Householdu18* is: 0-None, 1- One, 2-Two, 3-Three or more.

```
Call:
glm(formula = high ~ ., family = binomial, data = df)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
Lived2          0.031640   0.228468   0.138 0.889853
Lived3          0.035636   0.238569   0.149 0.881258
Lived4         -0.031881   0.231652  -0.138 0.890538
Lived5         -0.096930   0.215208  -0.450 0.652421
Householdu181  -0.188131   0.111129  -1.693 0.090473 .
Householdu182  -0.443198   0.141864  -3.124 0.001783 **
Householdu183  -0.613234   0.229769  -2.669 0.007610 **
Householdu184  -0.357682   0.385548  -0.928 0.353552
Householdu185   0.842244   0.536199   1.571 0.116236
Householdu186  -0.574737   0.842926  -0.682 0.495343
Householdu187  -0.240516   1.449681  -0.166 0.868228
Householdu188  -9.633241 324.744278  -0.030 0.976335
Householdu189  13.396484 210.212288   0.064 0.949187
Ethnic2        -0.185382   0.315697  -0.587 0.557060
Ethnic3        -0.032178   0.301965  -0.107 0.915136
Ethnic4        -1.513396   1.033913  -1.464 0.143261
Ethnic5        -0.529282   0.306392  -1.727 0.084084 .
Ethnic6        -0.196523   0.411739  -0.477 0.633149
Ethnic7         0.173228   0.278540   0.622 0.533998
Ethnic8        -0.007283   0.345980  -0.021 0.983206
Language2      -0.215295   0.214883  -1.002 0.316383
Language3       0.031409   0.220544   0.142 0.886753
```

Table 2: A logistic Regression using all the predictors (part of the results is shown)

Based on the above analysis, we fit a new logistic regression model after removing the three variables not helpful in predicting *high*. At the end of part b, we will apply cross validation on both models to see if removing these predictors can yield an improvement in prediction accuracy.

Regarding to the interpretation of the coefficients, a coefficient β of a predictor indicates that one unit increase in the quantitative predictor (or move from the first level to the current level of the qualitative predictor) is associated with the change in the log-odds of high income by β units, holding all other predictors constant. We take the exponential transformation e^{β} to interpret the estimated effects as relative odds ratios, which is easier to understand.

Furthermore, for the coefficients with p-value not tiny enough to reject null hypotheses that the coefficient equals to zero, we should not take them into account when interpreting the coefficients. Constant is also not of interest to interpret. Thus, we will interpret all statistically significant predictors and levels as below:

Sex: Female (level 2) is less likely to get high income than male (level 1). We expect a 0.3 decrease in the log-odds of high income for female versus male, which also means the odds of female getting high income is 0.7 times the odds of male getting high income, holding all other independent variables constant.

Marital: Divorced or separated person (level 3), windowed person (level 4) and single person (level 5) are less likely to get high income than married person (level 1), with a 0.86, 0.64, 0.48 decrease in the log-odds of high income respectively, holding all other independent variables constant. For example, the odds of single person getting high income is 0.6 times the odds of married person getting high income.

Age: Adults (level >1) are more likely to get high income than teenagers under 18 years old (level 1), among which the people between 35 to 44 years old (level 4) have the highest possibility with a 1.58 increase in the log-odds of high income, which also indicates the odds of the people between 35 to 44 years old getting high income are 4.9 times the odds of the people under 18 years old getting high income, holding all other independent variables constant.

Edu: The higher education people received, the higher the possibility of high income. For instance, the odds of the people accepted graduate school education (level 6) getting high income are 8.3 times the odds of the people with grade 8 or less (level 1) getting high income, holding all other independent variables constant.

Occupation: People with professional/managerial occupation (level 1) have the highest probability of high income. Among other levels of predictor *Occupation*, for example, the odds of sales workers (level 2) getting high income are 0.6 times the odds of the people with professional / managerial occupation getting high income, holding all other independent variables constant.

Dual_Income: The odds of married people with dual incomes (level 2) getting high income are 1.8 times the odds of unmarried people (level 1) getting high income, holding all other independent variables constant.

Household: Holding all other independent variables constant, the odds of people with two persons (level 2) in household getting high income are 2.3 times the odds of people with one person in household (level 1), and the people with three persons in household (level 3) getting high income are 2.1 times the odds of people with one person in household, and so on.

Householdu18: In general, the more persons under 18 years old in household, the less likely to get high income, holding all other independent variables constant. The odds of people with two persons under 18 years old in household (level 2) getting high income are 0.64 times the odds of people without person under 18 years old in household (level 0). The odds of people with three or more persons under 18 years old in household (level 3) getting high income are 0.59 times the odds of people without person under 18 years old in household.

Status: The odds of people who rent a housing (level 2) getting high income are 0.3 times the odds of people who own a housing (level 1) getting high income, holding all other independent variables constant.

Home_type: The odds of people living in apartments (level 3) getting high income are 0.5 times the odds of people living in houses (level 1) getting high income. The odds of people living in mobile homes (level 4) getting high income are 0.2 times the odds of people living in houses getting high income.

```
 Call:
 glm(formula = high ~ . - Lived - Ethnic - Language, family = binomial)
              Coefficient P-value  Odds
(Intercept)        -2.903   0.000 0.055
Sex2               -0.305   0.000 0.737
Marital2            0.099   0.593 1.104
Marital3           -0.861   0.000 0.423
Marital4           -0.638   0.046 0.528
Marital5           -0.476   0.036 0.621
Age2                0.971   0.000 2.642
Age3                1.166   0.000 3.209
Age4                1.582   0.000 4.865
Age5                1.529   0.000 4.613
Age6                1.323   0.000 3.756
Age7                0.730   0.038 2.074
Edu2                0.728   0.088 2.071
Edu3                0.898   0.035 2.455
Edu4                1.288   0.002 3.625
Edu5                1.659   0.000 5.254
Edu6                2.121   0.000 8.339
Occupation2        -0.458   0.000 0.632
Occupation3        -1.456   0.000 0.233
Occupation4        -0.964   0.000 0.381
Occupation5        -0.317   0.028 0.728
Occupation6        -0.509   0.000 0.601
Occupation7        -1.298   0.000 0.273
Occupation8        -0.885   0.000 0.413
Occupation9        -0.856   0.000 0.425
Dual_Income2        0.599   0.002 1.821
Dual_Income3        0.113   0.584 1.119
Household2          0.816   0.000 2.262
Household3          0.722   0.000 2.059
Household4          0.839   0.000 2.315
Household5          0.845   0.000 2.327
Household6          0.808   0.004 2.243
Household7          0.673   0.072 1.960
Household8          1.618   0.000 5.042
Household9          1.561   0.001 4.764
Householdu181      -0.188   0.088 0.829
Householdu182      -0.443   0.001 0.642
Householdu183      -0.523   0.010 0.593
Status2            -1.170   0.000 0.311
Status3            -0.006   0.967 0.994
Home_Type2          0.021   0.866 1.021
Home_Type3         -0.705   0.000 0.494
Home_Type4         -1.624   0.000 0.197
Home_Type5         -0.163   0.405 0.850
```

Table 3: A logistic Regression using the statistically significant predictors (showing coefficients, p-values and relative odds ratio)

2. LINEAR DISCRIMINANT ANALYSIS (LDA)

We want to classify the observations into one of k =2 classes of the response variable *high*, with 13 predictors. Hence, we extract $\min(13, k-1) = 1$ discriminant LD. We fit an LDA model and interpret the results as below:

The first thing shown in the LDA output is the prior probabilities of two classes. The low-income group (level 0) accounts for 73.6% of the whole data set, while high income group (level 1) only accounts for 26.4%.

0	1
0.736	0.264

Table 4: Prior probabilities of groups

Then, the Group means are shown in the first two columns of the following table, which are the means of each predictor broken down by each class. Predictors having a large difference between the two classes are likely to be useful in predicting which group an observation should be classified.

	0	1	LD1
Sex2	0.558	0.541	-0.206
Marital2	0.078	0.078	0.019
Marital3	0.114	0.052	-0.595
Marital4	0.035	0.013	-0.454
Marital5	0.486	0.195	-0.361
Age2	0.272	0.128	0.440
Age3	0.267	0.230	0.503
Age4	0.144	0.297	0.865
Age5	0.074	0.180	0.871
Age6	0.059	0.104	0.683
Age7	0.061	0.045	0.191
Edu2	0.144	0.033	0.092
Edu3	0.246	0.129	0.099
Edu4	0.352	0.343	0.326
Edu5	0.148	0.251	0.629
Edu6	0.076	0.240	1.046
Occupation2	0.095	0.075	-0.382
Occupation3	0.104	0.036	-0.896
Occupation4	0.138	0.075	-0.712
Occupation5	0.067	0.092	-0.302
Occupation6	0.197	0.072	-0.472
Occupation7	0.025	0.008	-0.886
Occupation8	0.074	0.063	-0.727
Occupation9	0.042	0.015	-0.590
Lived2	0.133	0.096	0.062
Lived3	0.089	0.080	0.109
Lived4	0.107	0.118	0.050
Lived5	0.632	0.687	0.008
Dual_Income2	0.174	0.475	0.570
Dual_Income3	0.128	0.206	0.061
Household2	0.278	0.412	0.444
Household3	0.194	0.208	0.380
Household4	0.169	0.194	0.456
Household5	0.082	0.074	0.462
Household6	0.031	0.021	0.495
Household7	0.013	0.008	0.434
Household8	0.007	0.006	1.052
Household9	0.006	0.006	0.932
Householdu181	0.175	0.180	-0.095
Householdu182	0.130	0.139	-0.248
Householdu183	0.076	0.052	-0.278
Status2	0.498	0.198	-1.073
Status3	0.240	0.108	-0.300
Home_Type2	0.071	0.090	-0.018
Home_Type3	0.340	0.095	-0.318
Home_Type4	0.018	0.005	-0.973
Home_Type5	0.041	0.026	-0.126
Ethnic2	0.056	0.051	-0.122
Ethnic3	0.101	0.065	-0.017
Ethnic4	0.002	0.001	-0.687
Ethnic5	0.151	0.059	-0.261
Ethnic6	0.014	0.010	-0.181
Ethnic7	0.632	0.776	0.131
Ethnic8	0.026	0.023	0.024
Language2	0.069	0.025	-0.122
Language3	0.031	0.024	0.009

Table 5: Group means (first two columns) and Coefficients of linear discriminants (the last column)

The last part is the coefficients of linear discriminants, which is shown is the last column of the above table. In general, predictors with large absolute values in the scaling are more likely to influence the process. The first discriminant function (LD1) is a linear combination of the predictors and the coefficients: $-0.2 * Sex2 + 0.02 * Marital2 - 0.6 * Marital3 - 0.5 * Marital4 - 0.4 * Marital5 + 0.4 * Age2 + \cdots$. If it is large, the LDA classifier would predict a high income, while predicting a low income if it is small.

3. CLASSIFICATION TREES WITH PRUNING

At first, we fit a classification tree without pruning, in which there are six terminal nodes and three variables used in tree construction, *Status, Occupation and Marital*. It is very easy to interpret with the nice graphical representation as below. Only the people who 1) own a housing, 2) with professional or managerial occupation and 3) married or living together with someone are expected to have high income. For the people who don't satisfy any one of the three conditions are estimated to have a low income.

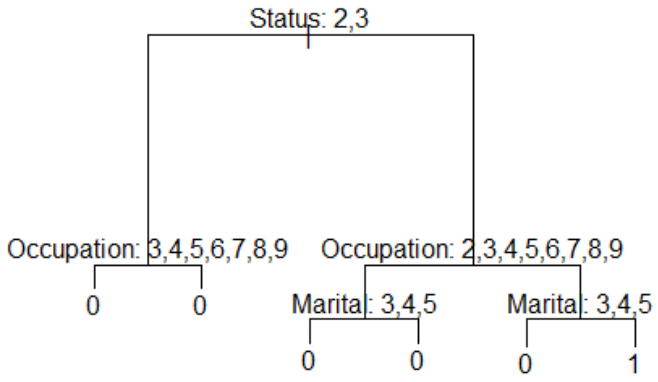

Figure 3: A classification tree without pruning

Then, we prune the tree to see if it would yield improvements. From the cross-validation results, the tree with four terminal nodes yields the same cross-validation error rate as the tree with six terminal nodes.

```
$size
[1] 6 4 3 1
$dev
[1] 1378 1378 1450 1812
$k
[1]  -Inf   0.0  69.0 182.5
$method
[1] "misclass"
attr(,"class")
[1] "prune"         "tree.sequence"
```

Table 6: Cross validation results of determining the optimal level of tree complexity

Finally, we prune the tree to obtain the four-node tree which is graphically shown in the below figure. The interpretation of the four-node tree is the same as the six-node tree that only the people who 1) own a housing, 2) with professional or managerial occupation and 3) married or living together with someone are expected to have high income. The only difference is that the four-node tree avoids splitting two terminal nodes that have the same predicted value.

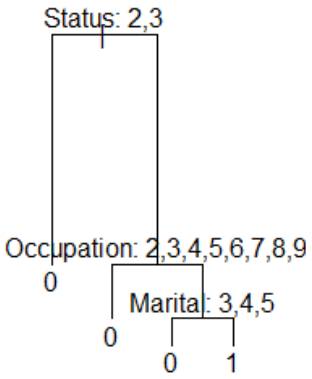

Figure 4: A classification tree with pruning

4. COMPARE THE PREDICITONS WITH CROSS-VALIDATION

So far, we have used logistic regression (with all predictors and selected predictors), linear discriminant analysis and classification trees with pruning to predict *high*. We will compare the four models with cross-validation.

Regarding to the decision of choosing among the three validation methods, validation set approach, LOOCV and k-fold cross-validation, LOOCV is the first one to be weeded out, because it is computational expensive to implement on more than 6000 observations. Then we weed out validation set approach due to its worst performance in bias reduction. Therefore, we choose k-fold cross-validation which can generate accurate estimates of test error rate and suffers neither high bias nor high variance.

With respect to the choice of fitting function used to measure the quality of the predictions, we compute the fraction of observations for which the prediction was correct on test set of each fold, then calculate the mean of the k folds. Except the total accuracy rate, we are also interested in the accuracy rate on two different classes separately, which are the fraction of low-income examples that are classified as low income (level 0 accuracy rate), and the fraction of high-income examples that are classified as high income (level 1 accuracy rate).

```
                          glm1     glm2      lda     tree
Total   accuracy rate   0.805532 0.805677 0.804512 0.797089
Level 0 accuracy rate   0.836661 0.834932 0.841861 0.804623
Level 1 accuracy rate   0.665650 0.669237 0.652302 0.732002
```

Table 7: Mean of the fraction of correct classification on test set

We draw the conclusion from above tables that the three classifiers yield similar total accuracy rate, which are around 80%. The logistic regression has the best total accuracy rate, which is 0.8% higher than the classification tree. Removing predictors not statistically significant can have a little bit improvement in the prediction accuracy of the logistic regression.

However, the level 0 accuracy rate and level 1 accuracy rate splits dramatically. For the logistic regression, the level 0 accuracy rate is 17% higher than level 1 accuracy rate. The classification tree has the narrowest gap between the level 0 accuracy rate and level 1 accuracy rate, which is 7%.

In summary, among the three classifiers, the logistic regression has the highest total accuracy rate, the classification tree has the narrowest gap between the level 0 accuracy rate and level 1 accuracy rate, while the linear discriminant analysis is quite moderate. Furthermore, although the classification tree doesn't have the same high accuracy as other classifiers, it is quite easy to explain and can be interpreted by people without any statistical background.

c. Choose two additional prediction methods from the chapters 7-10 in James et al. (2015) and compare their predictions with the other methods.

1. SUPPORT VECTOR MACHINE (SVM)

In order to fit a Support Vector Machine using a non-linear kernel, we use the svm() function. Then, we perform the cross-validation using function tune() to select the best choice of gamma and cost for an SVM with a radial kernel. We decide to apply radial kernel because of its local behavior that only nearby training observations influence the class label of a test observation. Although we do not specify levels of gamma and cost since the function tune() perfectly fit our data by its arbitrary complex ability, it is possible for the user to create a list of exact values to assign to the two variables. However, we decide not to show the results in our report, because those lists certainly lead to outperform out-of-sample or dominate our in-sample predictions, which means the accuracy of the confusion Matrix would decrease.

```
 Call:
 best.tune(method = svm, train.x = high ~ ., data = train, kernel = "ra
dial")
  Parameters:
    SVM-Type:  C-classification
  SVM-Kernel:  radial
        cost:  1
       gamma:  0.01754386
 Number of Support Vectors:  1626
```

Table 8: Summary of cross-validation for selecting gamma and cost

The model that reduces the most errors in the training data uses a cost of 1 and gamma value of 0.0175.

Next, we predict the variable *high* on the test data and we compute the confusion matrix to return the accuracy of the predictions shown as below:

```
Confusion Matrix and Statistics
          Reference
Prediction    0    1
         0 2306  468
         1  207  457
               Accuracy : 0.8037
                 95% CI : (0.79, 0.8168)
    No Information Rate : 0.7309
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.452
 Mcnemar's Test P-Value : < 2.2e-16
            Sensitivity : 0.9176
            Specificity : 0.4941
         Pos Pred Value : 0.8313
         Neg Pred Value : 0.6883
             Prevalence : 0.7309
         Detection Rate : 0.6707
   Detection Prevalence : 0.8069
      Balanced Accuracy : 0.7058
       'Positive' Class : 0
```

Table 9: Confusion matrix for SVM

The Support Vector Machine successfully predict with an accuracy of 80.37%. Although the prediction fraction is quite high, the SVM has a major disadvantage, the interpretation of the model. The complexity of the algorithm does not allow users to draw a clear conclusion about the dummy variable *high*.

2. BAGGING (BOOTSTRAP AGGREGATION)

Bagging, or bootstrap aggregation, is a general-purpose procedure for reducing the variance of a statistical learning method. The general idea behind is that bagging average a number of trees based on same amount of training set, which certainly returns a less bias model than pruning back the trees. However, bagging improves prediction accuracy at the expense of interpretability.

Since we are facing a classification problem where Y is qualitative, the bagging method has got a different approach. For a given test observation, we can record the class predicted by each of the tree and take a majority vote. Moreover, the following method allow the user to overcome the computational issue to estimate the test error with the need to perform cross-validation or the validation set approach. It utilizes the Out of Bag observation to

compute the prediction - again by apply a majority vote due to the classification problem. We perform bagging as follow: first we use the function randomForest() on the training data - by specify that all the 13 predictors should be considered for each split of the tree (mtry=13) - then plot the importance of each variable into the selection process - we look at the Gini index for bagging classification trees.

```
Gini Index
          Sex       Marital           Age           Edu    Occupation
     51.13619      86.50318     120.93133     131.76809     163.62930
        Lived  Dual_Income     Household Householdu18        Status
     89.58291      49.42489     103.90855      57.34737     205.72158
    Home_Type         Ethnic      Language
     63.73268      93.22472      22.52092
```

Table 10: Gini index for importance of predictors

The results from the importance chart indicate that across all the trees considered in the bagging, the householder status (Status), the type of industry where the householder work (Occupation) and the level of its education (Education) are the three most important variables to predict variable *high*.

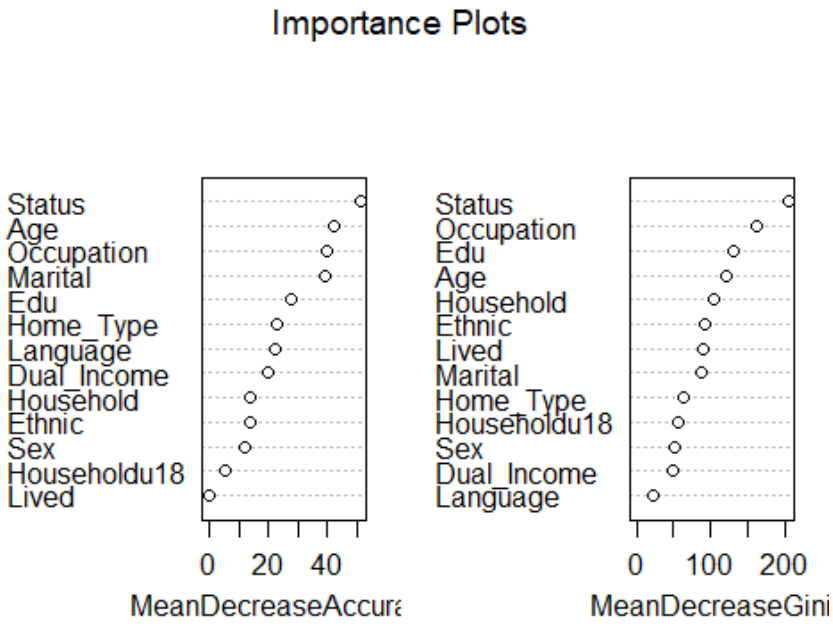

Figure 5: The importance plot returned with bagging approach

Once again, we predict the variable *high* on the test data and we compute the confusion matrix (shown below) to return the accuracy of the prediction:

```
Confusion Matrix and Statistics
          Reference
Prediction    0    1
         0 2230  283
         1  462  463
               Accuracy : 0.7833
                 95% CI : (0.7691, 0.797)
    No Information Rate : 0.783
    P-Value [Acc > NIR] : 0.4933
                  Kappa : 0.4132
 Mcnemar's Test P-Value : 6.965e-11
            Sensitivity : 0.8284
            Specificity : 0.6206
         Pos Pred Value : 0.8874
         Neg Pred Value : 0.5005
             Prevalence : 0.7830
         Detection Rate : 0.6486
   Detection Prevalence : 0.7309
      Balanced Accuracy : 0.7245
       'Positive' Class : 0
```

Table 11: Confusion matrix for bagging

The bagging method successfully predicts *high* with an accuracy of 78.33%. However, it lacks of interpretability. When we bag many trees, it is no longer possible to represent the resulting statistical learning procedure using a single tree, and it is no longer clear which variables are most important to the procedure.

Next, we compute the fraction of observations for which the prediction was correct on test set of each fold, then we calculate the mean of the k folds for the two selected models:

```
          SVM bagging
 [1,] 0.79330 0.76419
 [2,] 0.82678 0.79913
 [3,] 0.83261 0.80204
 [4,] 0.79767 0.78020
 [5,] 0.81951 0.80349
 [6,] 0.78311 0.77001
 [7,] 0.81223 0.78020
 [8,] 0.82096 0.79622
 [9,] 0.80495 0.79622
 [10,] 0.79185 0.75400
     SVM  bagging
0.808297 0.784570
```

Table 12: Mean of the fraction of correct classification on test on SVM and bagging methods

Although the Support Vector Machine is higher than the Bagging of about 2.37%, both models obtain relatively high accuracy rates.

Finally, we can conclude that Support Vector Machine (kernel = "radial") returns the best accuracy, compared with logistic regression, linear discriminant analysis, classification trees, and bagging methods. The mean accuracy in the 10 K fold-cross validation is about 81%, 0.03% higher than the accuracy produced by the logistic regression model in part b. But, the SVM's output is difficult to interpret. Hence, the classification tree method remains the best option for audiences who do not hold a strong analytics skill and statistical background.

d. Elaborate on what your analysis has to say about the different predictors' association with high? Does a high-income household have any features?

Since the Support Vector Machine model sacrifice its ability of the interpretability of the variables to obtain better prediction accuracy, we limited our analysis on *high* to the models elaborated in part b and Bagging.

According to the logistic regression, high income individuals are those who are middle age, most likely men – the log-odds ratio is negative (-0.3) for female – which have got a managerial position and have completed a high education degree – higher levels of education are most significant and with positive odds-ratio. Moreover, those in which household has two or more people according the coefficients significance have got a higher probability to have a high Income.

Further, the classification tree with pruning confirm the higher chances for a married individual which own a house and have got a professional or managerial occupation to have a high income.

Finally, by look at table 5 – Linear Discrimination Analysis – we can identify which predictors' levels have higher chances to obtain high household Income. For example, be single (*Martial 5*) has a high chance to do not earn high Income, as well factory workers (*Occupation 3*), clerical or social service employees (*Occupation 4*) and student (*Occupation 6*) most likely fall in the category 0 of *high,* low income. Those who rent accommodations are more likely to be categorized as low earnings group.

Opposite who have finished at least a bachelor's degree have got better chance to be categorize into the class 1 of *high,* earn more than $50,000. Same logic applies to dual income family (*Dual_Income 2*) and in which the household has got 2 people *(Household 2).*

Although, the Bagging process does not return the significance of each levels for every predictor, it is still helpful to confirm which variables returns the highest importance versus *high* by using the Gini Index: *Status, Occupation*, *Education*, *Age* and *Household*.

e. Summarize your results in a conclusion.

In this assignment, we analyze the qualitative dataset *marketing* from package *ElemStatLearn*. We practice the three methods: logistic regression, linear discriminant analysis (LDA), classification trees, and compare the predictions that four models (we build two models: with all predictors and with selected predictors, with logistic regression method) produce with the k fold cross-validation approach, considering the bias-variance tradeoff. We discover that people with demographic characteristics, such as: 1) own accommodation; 2) managerial occupation; 3) married or live together are more likely to be grouped into high-income class, and that all three methods obtain the similar accuracy. Furthermore, we select additional two methods: support vector machine (SVM) and bagging (bootstrap aggregation) to see which model from the five methods we use generate the best predication. We noticed that although the SVM provides the highest accuracy results, we still prefer to choose logistic or classification tree method when dealing with qualitative variables for their simplicity in interpretation and relatively high accuracy.

References

Dutt, Anish. *To eat or not to eat! That's the question? Measuring the association between categorical variables*. https://www.r-bloggers.com/to-eat-or-not-to-eat-thats-the-question-measuring-the-association-between-categorical-variables/. Accessed on 25th March 2019

Halvorsen, Kjetil B. 2019. *ElemStatLearn: Data Sets, Functions and Examples from the Book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman*. Material from the book's webpage and R port and packaging by the author, https://CRAN.R-project.org/package=ElemStatLearn.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning*. Springer.