Homework 1 : Statistical Data Mining II (EAS 507)

Shubham Sharma, Person No.: 50290293, Class No.: 43

March 2, 2019

Question 1: Consider the MovieLense data in the recommenderlab package.

Design and evaluate your own recommendation system based on the following principles:

For each user i and each movie j they did not see, find top k most similar users to i who have seen j and then use them to infer the user i s rating on movie. Handle all exceptions in a reasonable way and report your strategy if you did so; e.g., if you cannot find k users for some movie j, then take all users who have seen it.

Test the performance of your system using cross-validation. For each data set, the MovieLens database already provides a split of the initial data set into N = 5 folds. This means you will run your algorithm N times; in each step, use the training partition to make predictions for each user on all terms rated in the test partition (by that user). When you complete all N iterations, you will have a large number of user-movie pairs from the 5 test partitions on which you can evaluate the performance of your system. Measure the performance of your recommendation system.

A recommendation system is created for MovieLense data using the recommenderlab package. The predicted values is saved in the file q1ur.csv. Predictions have been made based on 50 nearest users using User based collaborative filtering. We don't get any exceptions in this case.

On examining the plot for Precision vs. Recall in figure 2 we find that k = 10 is ideal to get an accurate recommendation system. This maintains a balance between the precision and recall and for k = 10 we get the elbow.

In figure 1 we observe that at k = 10 we get the following order of performance for different methods: SVD > popular items > UBCF > random items > IBCF.

The performance for different methods is summarised in figure 3

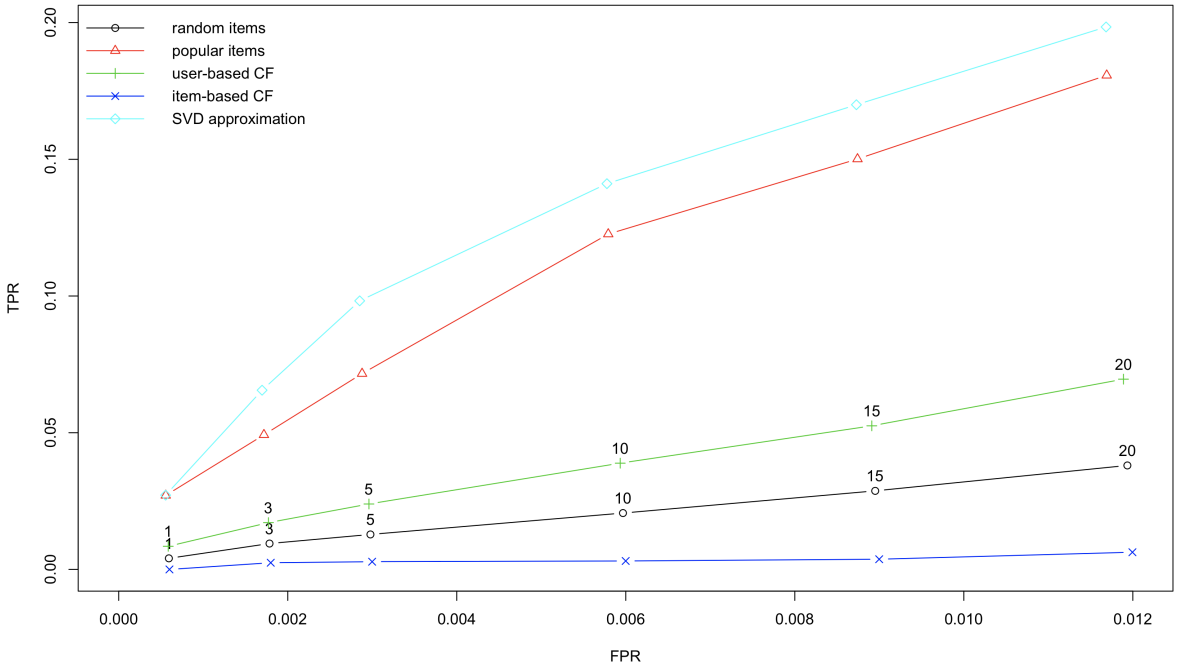

Figure 1: Plot for TPR vs. FPR

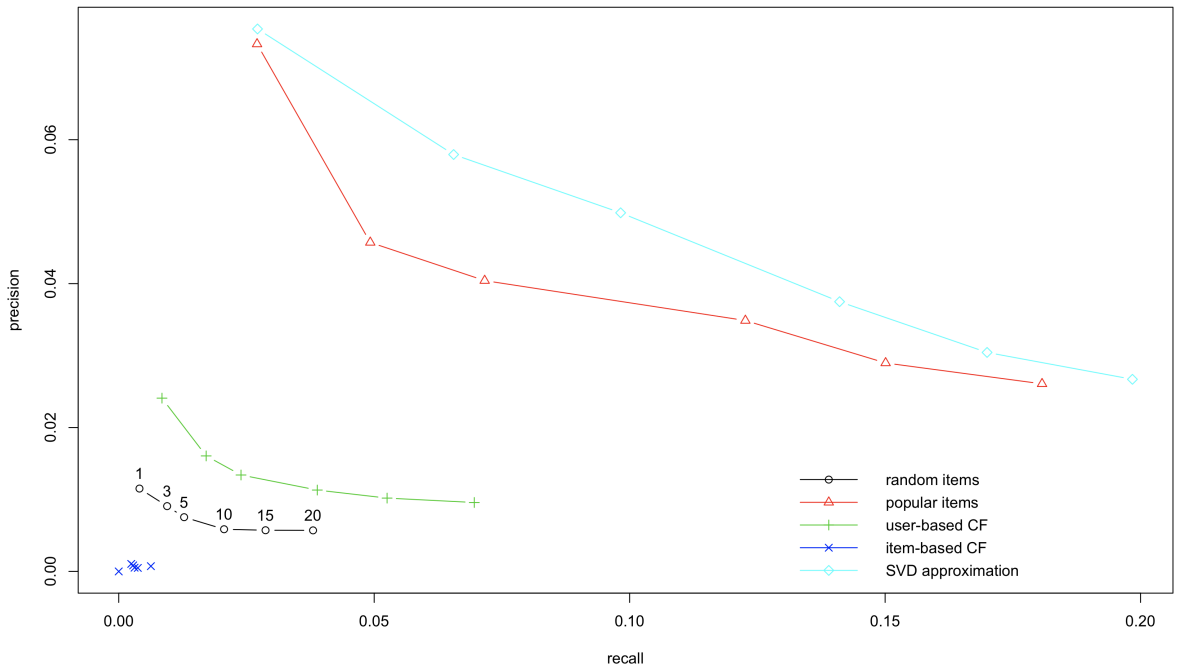

Figure 2: Plot for Precision vs. Recall

```
$`random items`
        RMSE      MSE      MAE
res 1.399989 1.961585 1.105544

$`popular items`
         RMSE       MSE       MAE
res 0.9861151 0.9735625 0.7701875

$`user-based CF`
        RMSE      MSE      MAE
res 1.036419 1.075268 0.820741

$`item-based CF`
        RMSE     MSE      MAE
res 1.408721 1.99436 1.059405

$`SVD approximation`
       RMSE     MSE       MAE
res 1.01069 1.02254 0.7962399
```

Figure 3: Perforamance of different methods

Question 2: Consider the given ratings table between five users and six items.

(a) Predict the values of unspecified ratings of user 2 using user-based collaborative filtering. Use the Pearson correlation with mean-centering.

(b) Predict the values of unspecified ratings of user 2 using item-based collaborative filtering algorithms. Use the adjusted cosine similarity.

The figure 4 shows the predicted values of unspecified ratings of user 2(as well as all the users) using user-based collaborative filtering using Pearson correlation with mean-centering

```
> getRatingMatrix(UR)
5 x 6 sparse Matrix of class "dgCMatrix"

[1,] .        .        . .        .        5.324102
[2,] .        4.080462 . 3.844854 .        .
[3,] 2.124195 .        . .        .        2.753220
[4,] .        .        . .        4.444854 .
[5,] .        2.875376 . .        .        .
```

Figure 4: User based Collaborative Filtering with mean centering

The figure 5 shows the predicted values of unspecified ratings of user 2(as well as all the users) using item-based collaborative filtering using the adjusted cosine similarity. The row numbers denote the user numbers.

```
5 x 6 sparse Matrix of class "dgCMatrix"
           1        2 3       4        5        6
[1,] .        .        . .       .        5.062993
[2,] .        3.999355 . 3.98319 .        .
[3,] 2.328387 .        . .       .        2.255409
[4,] .        .        . .       4.454013 .
[5,] .        2.734110 . .       .        .
>
```

Figure 5: Item Based Collaborative Filtering with adjusted cosine similarity

Question 3: Consider the Boston Housing Data. This data can be accessed in the ElemStatLearn package (available through CRAN).

a) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

b) Visualize the data using the itemFrequencyPlot in the arules package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).

c) A student is interested is a low crime area as close to the city as possible (as measured by dis). What can you advise on this matter through the mining of association rules?

d) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

e) Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?

The Boston Data Set has been taken from the MASS package. Histograms for all the variables have been plotted as shown in figures 6, 7, 8 and 9.

The individual features are grouped into different categories after looking at the distinction between frequencies for different values for a particular feature. In this way, we ensure adequate support for all the attributes.

The item frequency plot is shown in the figure 10. The support threshold for this plot is kept at 0.05 which means only those attributes are shown which have a support of greater than 5%, while the item frequencies are plotted relative to each other.

Association rules are mined using the apriori algorithm with support = 0.01 and confidence = 0.7. 1696494 rules are formed from 506 transactions in the binary incidence matrix.

Part c: On checking the rules for $crim = low$ and $dis = low$ we get the following: $nox = med$ and $ptratio = low$ with a support of 0.051, a confidence of 1 and a lift of 2.11.

Thus, these attributes have occurred together in 5.1% of the rules and we can say that if *crim* is low, *nox* is medium as well as *ptratio* is low then *dis* is also low each time. Thus, if a student is interested in a low crime area as close to the city as possible, he should look for places where nitrogen oxide concentrations are medium (0.55-0.75 ppm) and pupil-teacher ratio is low (12 -16).

Also, looking at some more rules we can say that he should also consider places where proportion of non-retail business acres per town (*indus*) is high (18 - 28) and the full-value property-tax rate per $10,000 (*tax*) is medium (300 - 500).

Part d: For the rules having $ptratio = low$, we get $nox = high$ and $black = low$ with support = 0.016, confidence = 1 and lift = 5.95. Thus, a low pupil-teacher ratio is available in a city when nitrogen oxide concentrations are high (0.75 - 0.9 ppm) and the proportion of blacks by town is also low (0-350). Also, the full-value property-tax rate per $10,000(*tax*) should be medium (300 - 500), the index of accessibility

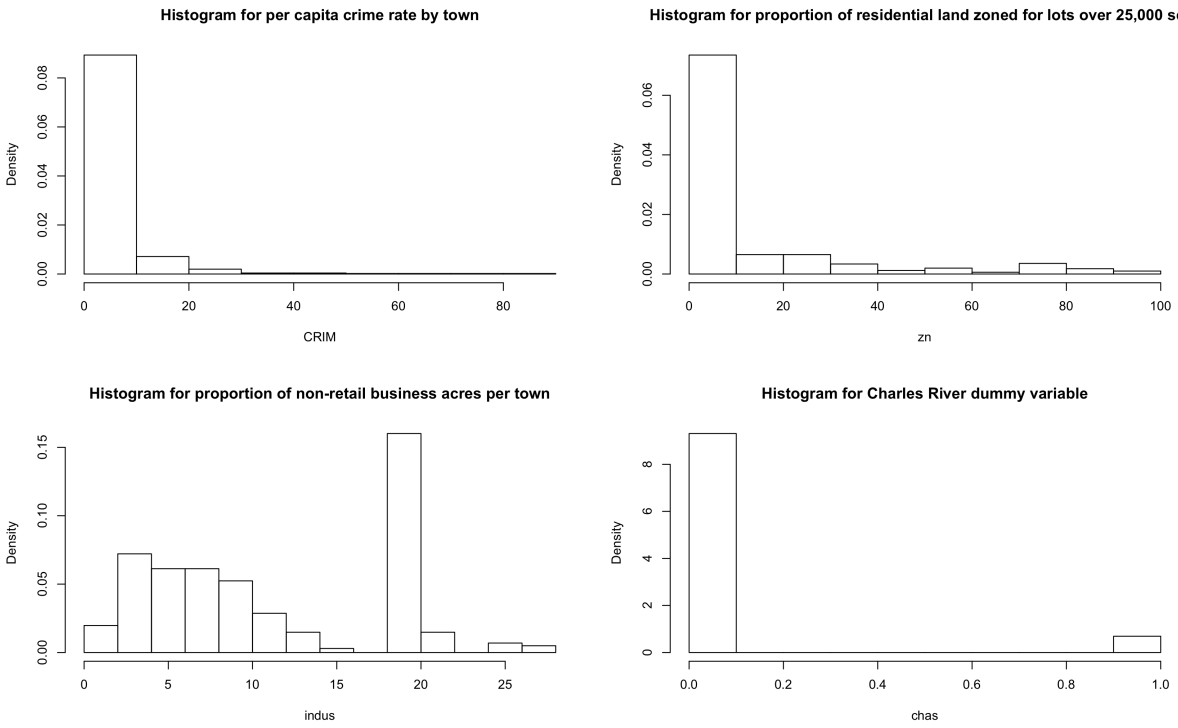

Figure 6: Histograms for *crim*, *zn*, *indus* and *chas*

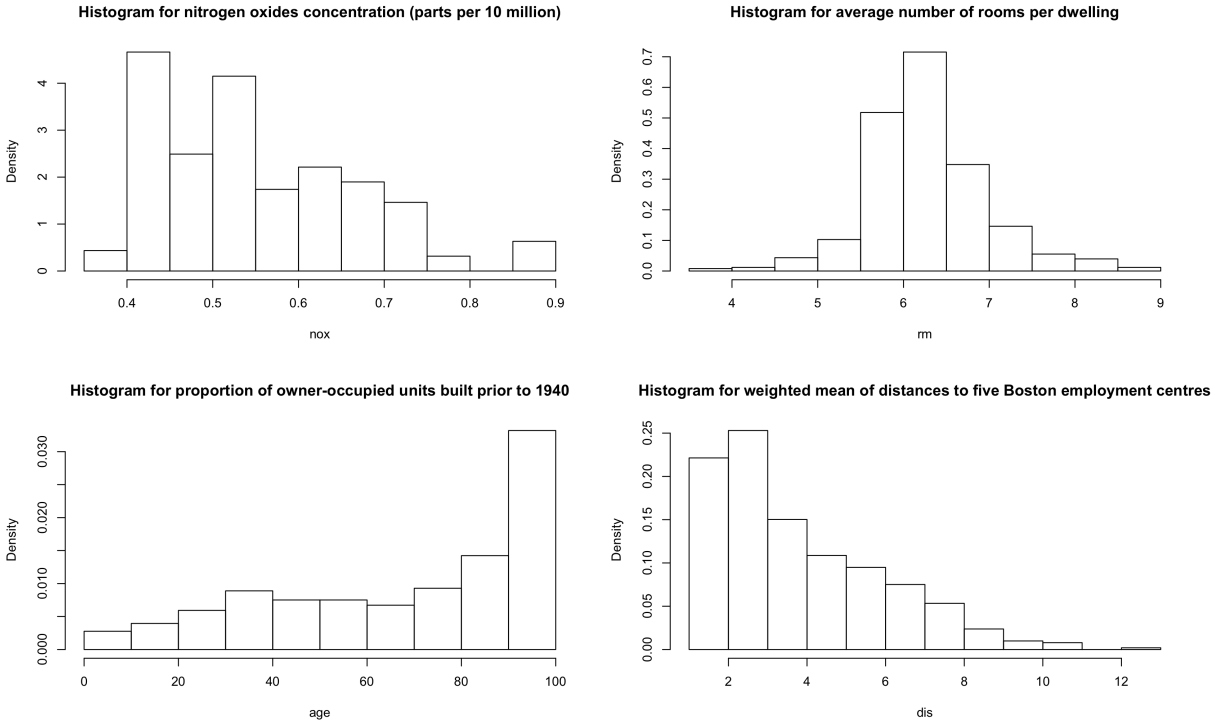

Figure 7: Histograms for *nox*, *rm*, *age* and *dis*

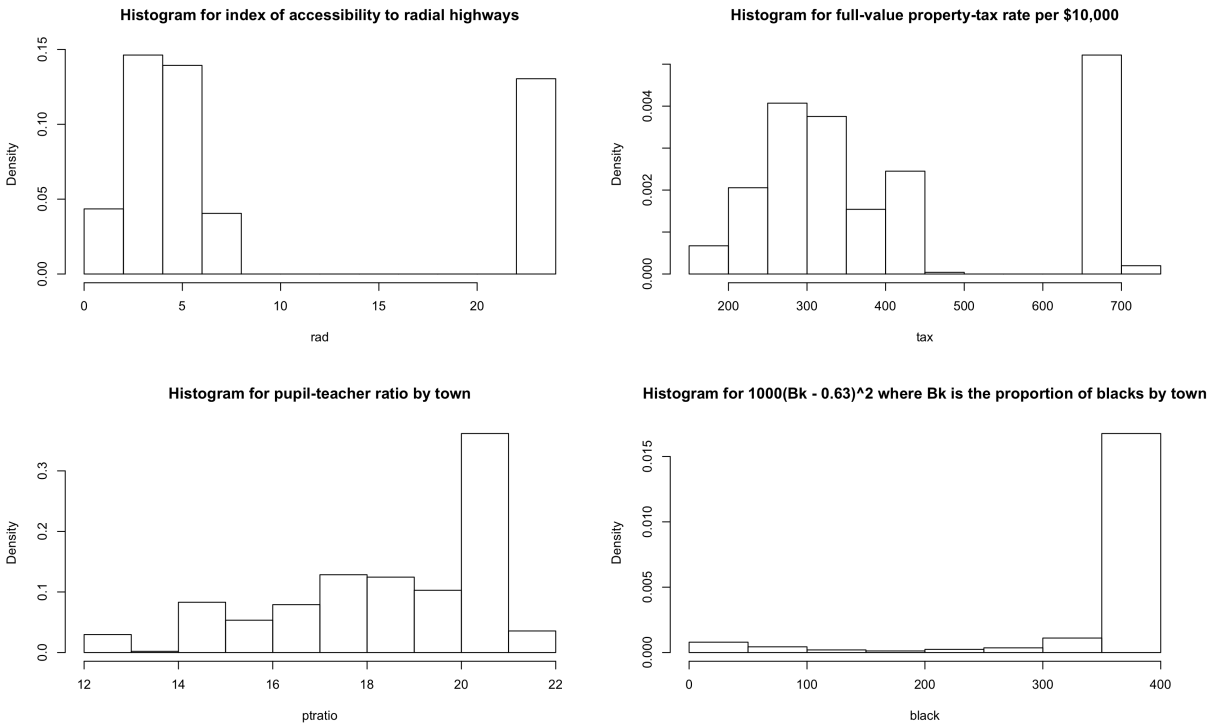

Figure 8: Histograms for *rad*, *tax*, *ptratio* and *black*

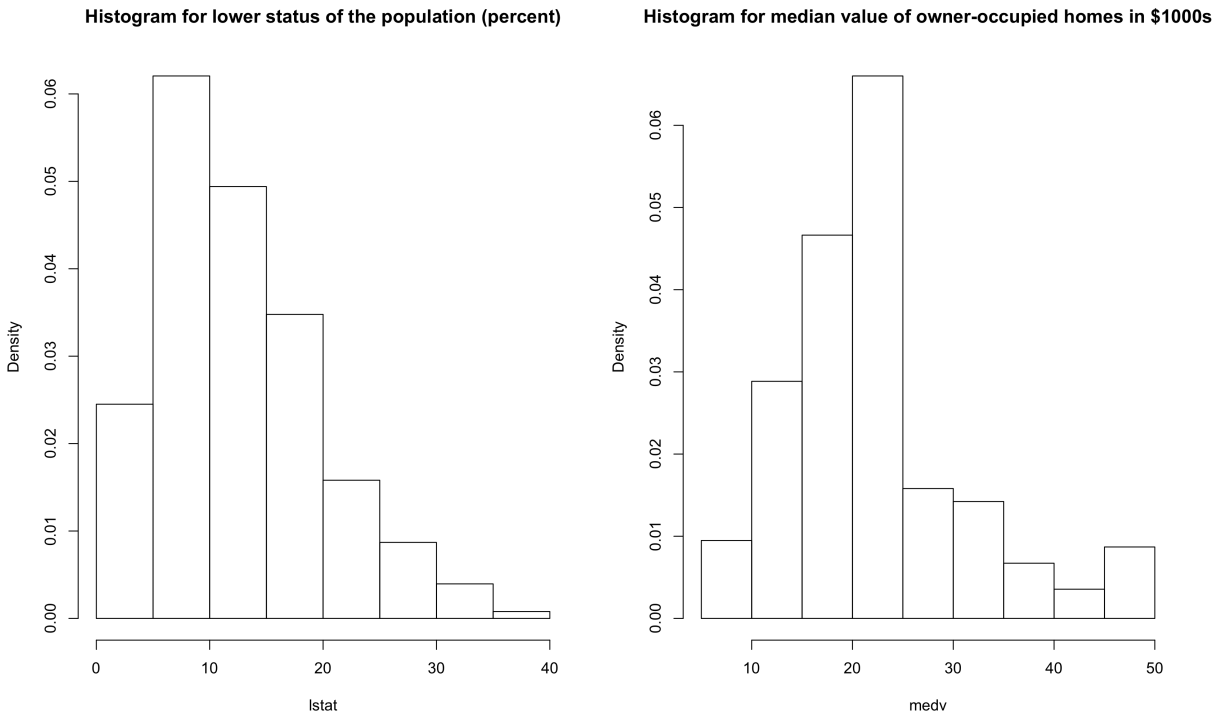

Figure 9: Histograms for *lstat* and *medv*

to radial highways(rad) should be low (0-10) as per other associations rules with comparable support, confidence and lift.

Part e: In Linear regression we get nox and rad as statistically significant variables from p-values which is comparable with the results obtained from mining association rules where both of them occur in the same rule with high support and confidence as seen in the previous part. However, Linear regression is a supervised learning technique whereas Association rules are unsupervised learning techniques. Thus, for prediction of a response variable linear regression would be preferred, e.g. prediction the value of $ptratio$ based on other variables. However, in this case Association rules provide a better interpretability since we get an approximate range for different features to get $ptratio = low$. Thus, regression tells us the relation as well as the extent to which a particular variable affects the response variable however association rules provide a better interpretabillity in cases where relations need to be made for different ranges of values of particular features.

Question 4: Cluster the demographic data of Table 14.1(ESL) using a classification tree. Specifically, generate a reference sample the same size as the training set. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability.

The given dataset has 8993 observations and 15 variables. However, 2694 observations are missing values which is replaced using median for each column. A reference data set is generated randomly using the unique values for each columns and the two data sets are combined giving a response variable Y a value of 1 for the original data set and a value of 0 for the reference data set.

The tree is grown with 10 number of cross-validations and cp = 0. The tree is pruned back at index = 10 as we can see from the plot of the cv error in figure 10.

The pruned tree is shown in figure 11. The highest probability for class = 1 comes out to be 0.956 for node 31 for which the path is as follows:

Householdu18 $< 2.5 -- >$ Language$< 1.5 -- >$ Household$< 5.5 -- >$ Home_Type< 3.5

Thus, we infer that we obtain the highest estimated class probability for class 1 (i.e. a sample from the original data set) when the Householdu18 is less than 2.5, Language < 1.5, Household < 5.5 and Home Type < 3.5.

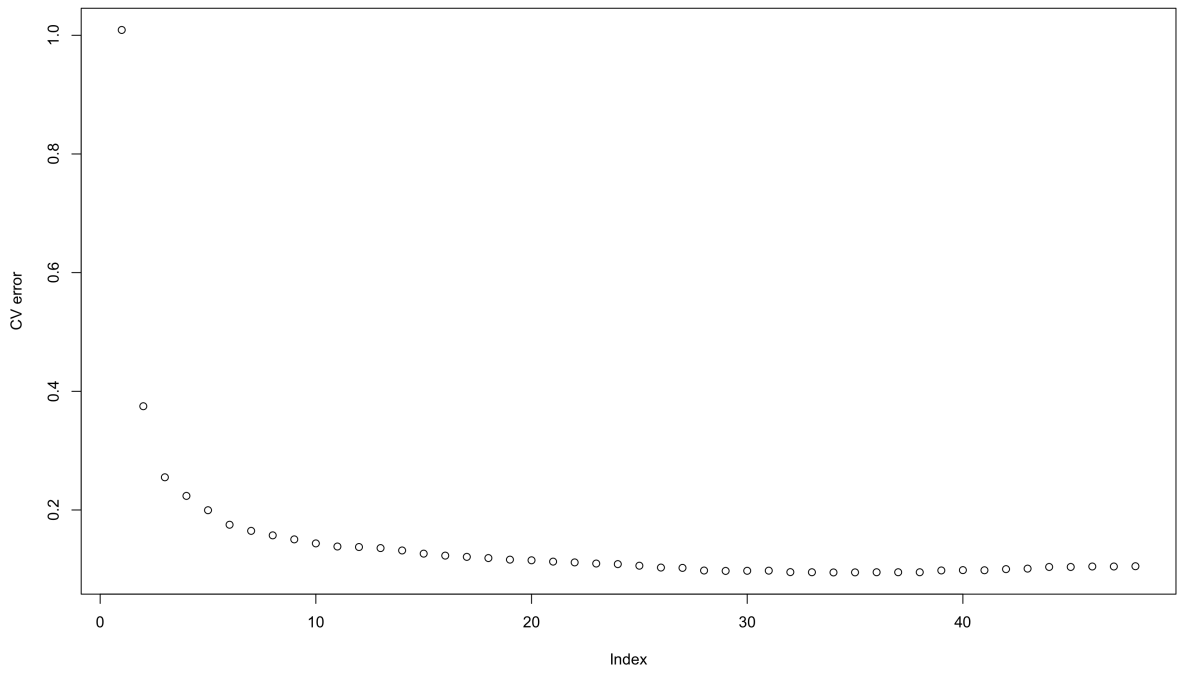

Figure 10: CV error vs. Index

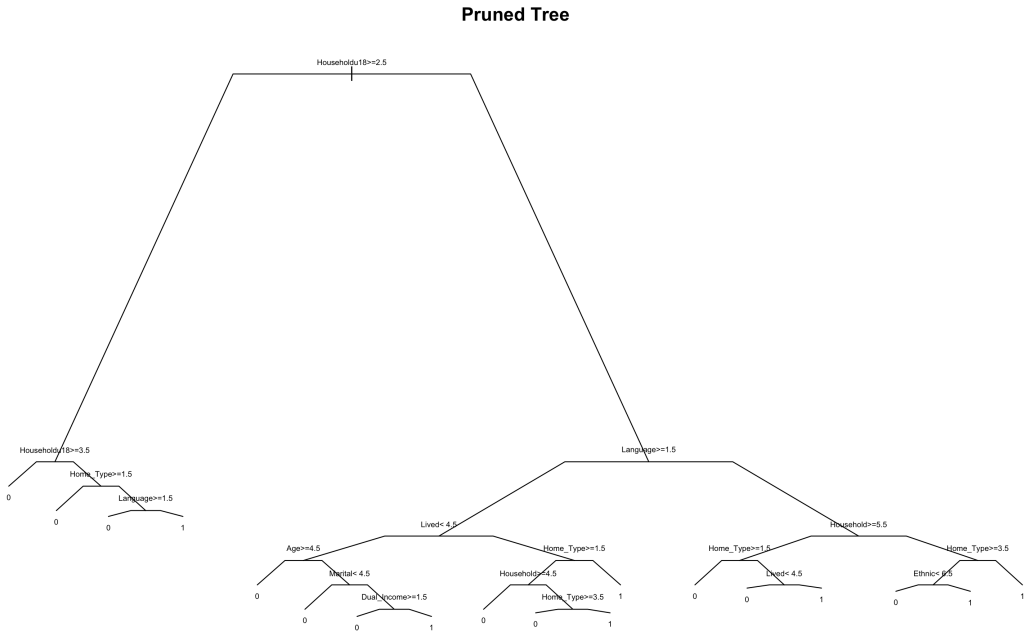

Figure 11: Pruned Tree