Report on Predictive Modeling of Education Levels Using a Random Forest Classifier

We undertook a study to develop a machine learning model capable of predicting individuals' education levels. The random forest classifier, selected for its robust performance in classification tasks, was trained on a dataset that was bootstrapped due to its initial small size. Our model demonstrated a high degree of accuracy, with further insights gained from an analysis of variable importance and ROC curves.

Data Preparation and Preprocessing

The Income2.csv dataset originally comprised 21 records, necessitating a bootstrap approach to create a sufficiently large dataset for our model. After resampling, we prepared 100 samples with replacement, which allowed us to build and validate the model using variables such as Experience, GPA, Age, and Income, with the categorical outcome, Education, as our target.

Model Training

We configured our random forest classifier with 500 trees (ntree=500) and determined that two variables at each split (mtry=2) provided a good balance between model complexity and performance. We enabled the calculation of variable importance to understand the contribution of each feature to the model's predictions.

Model Evaluation

Our model's Out-Of-Bag (OOB) error estimate was 4%, indicating strong predictive power. Upon validating against a separate dataset, we found high consistency between the observed and predicted education levels.

Results

ROC Curve Analysis: Our ROC curves revealed an exceptional ability of the model to classify education levels. The AUC scores for the levels BS, HS, and MS were nearly perfect, at 1, 1, and approximately 0.997, respectively.

Variable Importance Analysis: The variable importance plots revealed Experience as the primary predictor of education level. This was followed by Income, Age, and GPA, with the Mean Decrease in Accuracy indicating a tighter competition between Income and Age than what the Mean Decrease in Gini suggested.

Discussion

Despite the high accuracy and effective discrimination capabilities exhibited by our model, we express caution. The potential for overfitting is a concern due to the replicated instances in the bootstrapping process. Moreover, the primary reliance on the predictor 'Experience' could limit the model's generalizability across different demographics.

Conclusions

Our random forest model demonstrated a high accuracy rate in classifying individuals' education levels. However, we recognize the need for careful consideration of the model's capability to generalize beyond the scope of the bootstrapped data.

Model Insights

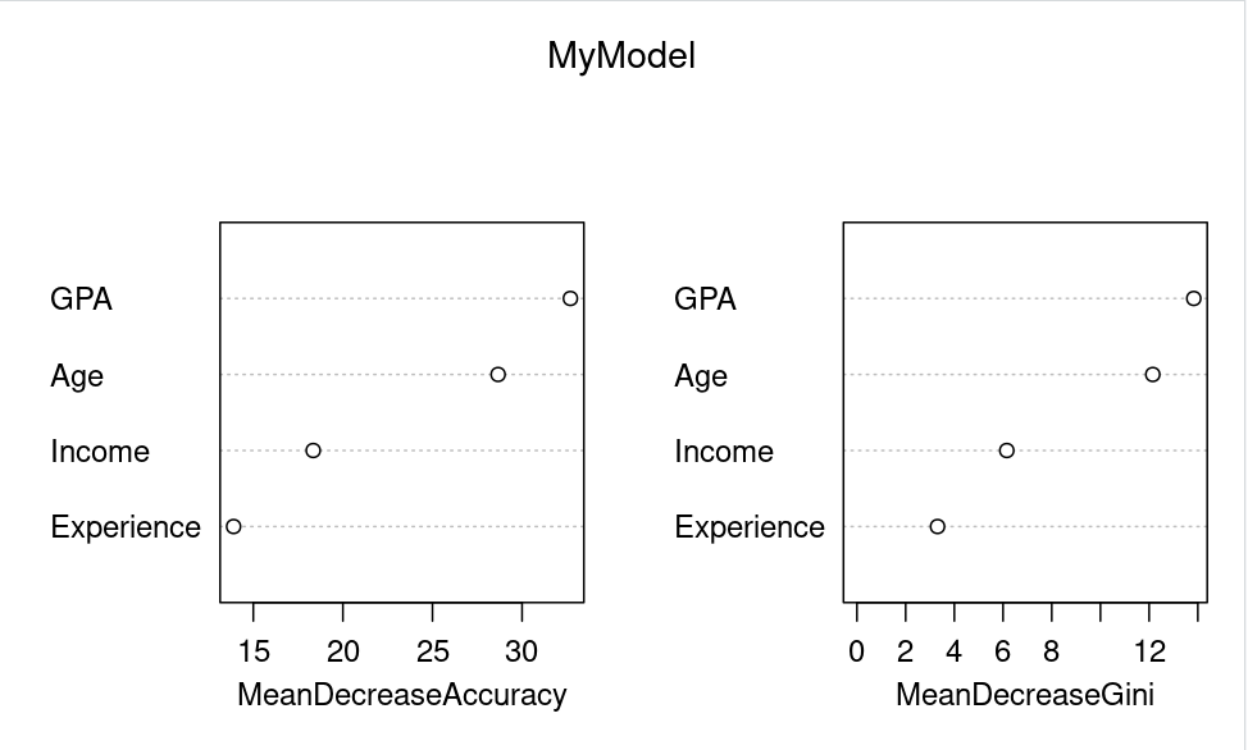
The model suggests that 'Experience' is the most crucial factor for predicting education levels, indicating that individuals with more experience tend to have higher education qualifications. This insight could have important implications for workforce development and educational policy.

R Code:

```
> data = read.csv("Income2 (1).csv")
> data$Education = as.factor(data$Education)
>
> install.packages("boot",dep=TRUE)
Error in install.packages : Updating loaded packages
```

```
> library(boot)
> install.packages("randomForest")
Error in install.packages : Updating loaded packages
> library(randomForest)
> install.packages("ROCR")
Error in install.packages : Updating loaded packages
> library(ROCR)
>
> totalS = sample(1:21, 100, replace = T)
> validateS = totalS[1:50]
> buildS = totalS[51:100]
>
> build = data[buildS,]
> validate = data[validateS,]
>
> summary(build)
   Experience        GPA             Age             Income      Education
 Min.   : 6.0   Min.   :238.0   Min.   :24.00   Min.   :51176   BS : 6
 1st Qu.: 9.0   1st Qu.:250.8   1st Qu.:28.00   1st Qu.:54318   HS :15
 Median :11.0   Median :308.5   Median :35.00   Median :67616   MS :17
 Mean   :12.5   Mean   :304.7   Mean   :34.42   Mean   :68540   PHD:12
 3rd Qu.:15.0   3rd Qu.:331.0   3rd Qu.:38.00   3rd Qu.:78063
 Max.   :23.0   Max.   :393.0   Max.   :48.00   Max.   :92108
> summary(validate)
   Experience         GPA             Age             Income      Education
 Min.   : 6.00   Min.   :238.0   Min.   :24.0   Min.   :51176   BS :10
 1st Qu.: 9.25   1st Qu.:260.0   1st Qu.:28.0   1st Qu.:56835   HS :11
 Median :12.00   Median :302.0   Median :33.0   Median :66267   MS :19
 Mean   :12.04   Mean   :302.5   Mean   :33.9   Mean   :66590   PHD:10
 3rd Qu.:14.75   3rd Qu.:327.0   3rd Qu.:37.0   3rd Qu.:70322
 Max.   :23.00   Max.   :393.0   Max.   :48.0   Max.   :92108
> MyModel = randomForest(Education ~ ., data=build, ntree=500, mtry=2,
importance=TRUE)
> MyModel

Call:
 randomForest(formula = Education ~ ., data = build, ntree = 500,      mtry = 2,
importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 4%
Confusion matrix:
    BS HS MS PHD class.error
BS   4  2  0   0   0.3333333
HS   0 15  0   0   0.0000000
```

```
MS   0  0 17   0   0.0000000
PHD  0  0  0  12   0.0000000
>
> varImpPlot(MyModel)
>
> MyPredictions = predict(MyModel, validate[,-5])
> table(observed=validate[,5],predicted=MyPredictions)
        predicted
observed BS HS MS PHD
     BS  10  0  0   0
     HS   0 11  0   0
     MS   0  2 17   0
     PHD  0  0  0  10
>
> ROC_Predictions= predict(MyModel,validate[,-5],type="prob")
> Colors = c("Green","Blue","Red","Black")
> Class = levels(validate$Education)
> for (i in 1:3)
+ {
+    true_values = ifelse(validate[,5]==Class[i],1,0)
+    pred = prediction(ROC_Predictions[,i],true_values)
+    perf = performance(pred, "tpr", "fpr")
+    if (i==1)
+    {
+        plot(perf, col=Colors[i], main="ROC Curve for Each type of Education
+ (Green=BS) (Blue=HS) (Red = MS) (Black = PHD)")
+    }
+    else
+    {
+        plot(perf, main="ROC Curve", col=Colors[i], add=TRUE)
+    }
+    AUC = performance(pred, measure = "auc")
+    print(AUC@y.values)
+ }
[[1]]
[1] 1

[[1]]
[1] 1

[[1]]
[1] 0.9966044
```



'Experience' is the most critical predictor in determining education levels according to both criteria of variable importance. This indicates that as the number of years of experience increases, the model becomes better at predicting the education level, possibly implying that higher education levels might correlate with more years of experience.

'Income' follows as the second most important variable but with a notable gap from 'Experience'. This could suggest that individuals with higher income levels might be associated with higher education levels, but the strength of this relationship is not as strong as with experience.

'Age' and 'GPA' are less important in this model, which could be due to a weaker or more complex relationship with the education level that is not captured by the model as directly as 'Experience' and 'Income'.

The curves for BS (Green), HS (Blue), and MS (Red) are starting at the top-left corner, which suggests that the model has an excellent classification performance for these categories. The perfect starting point (0,1) for these curves indicates that the model has a high sensitivity (true positive rate) and specificity (1 - false positive rate) for these education levels.

There is a lack of visibility for the PHD (Black) curve, which could mean one of two things: either the PHD curve overlaps perfectly with the other curves, indicating equally strong performance, or it was not included in the plot. Since PHD is a unique category with likely fewer samples, if it's not visible due to overlap, it suggests that the model is very accurate across all levels of education.