Machine Learning: Adult Census Income Project

Xiaoqing Wu

6/18/2020

1 Introduction

This document describes a machine learning project as apart of Harvard edX Data Science course. The project goal is to use Adult Census dataset hosted on kaggle.com to develop a machine learning model to predict if an adult's income is equal or less than $50K based on 14 predictors

Machine Learning has two common approaches: supervised learning and unsupervised learning. Supervised learning involves data visualization, data analysis, manual training and tuning data models, whereas unsupervised learning focuses on finding predictors to group observations as clusters and the prediction would be to identify the right clusters for new data entries. This document reports the exploration and findings from a supervised study of Adult Census income

Supervised learning generally has the following phases and each phase will be discussed in a corresponding section:

- Visualizing and analysing data in order to identify predictors that determine rating outcomes
- Wrangling data
- Developing training model for the machine to learn behaviors from observations
- Applying the training model to test data set to obtain predictions
- Measuring the accuracy of the predictions against true results and evaluating the performance of the recommendation system

2 Methods and Analysis

2.1 Data downloading

The income census data was downloaded from the author's github site and imported to a data set called adult.csv with 32,561 entries:

1994 Adult Census Income

age	workclass	fnlwgt	education	education.num	marital.status	occupation
Min. :17.0	Private :22696	Min. : 12285	HS-grad :10501	Min. : 1.0	Divorced : 4443	Prof-specialty :4140
1st Qu.:28.0	Self-emp-not-inc: 2541	1st Qu.: 117827	Some-college: 7291	1st Qu.: 9.0	Married-AF-spouse : 23	Craft-repair :4099
Median :37.0	Local-gov : 2093	Median : 178356	Bachelors : 5355	Median :10.0	Married-civ-spouse :14976	Exec-managerial:4066
Mean :38.6	Unknown : 1836	Mean : 189778	Masters : 1723	Mean :10.1	Married-spouse-absent: 418	Adm-clerical :3770
3rd Qu.:48.0	State-gov : 1298	3rd Qu.: 237051	Assoc-voc : 1382	3rd Qu.:12.0	Never-married :10683	Sales :3650
Max. :90.0	Self-emp-inc : 1116	Max. :1484705	11th : 1175	Max. :16.0	Separated : 1025	Other-service :3295
	(Other) : 981		(Other) : 5134		Widowed : 993	(Other) :9541

relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
Husband :13193	Amer-Indian-Eskimo: 311	Female:10771	Min. : 0	Min. : 0.0	Min. : 1.0	United-States:29170	<=50K:24720
Not-in-family : 8305	Asian-Pac-Islander: 1039	Male :21790	1st Qu.: 0	1st Qu.: 0.0	1st Qu.:40.0	Mexico : 643	>50K : 7841
Other-relative: 981	Black : 3124		Median : 0	Median : 0.0	Median :40.0	Unknown : 583	
Own-child : 5068	Other : 271		Mean : 1078	Mean : 87.3	Mean :40.4	Philippines : 198	
Unmarried : 3446	White :27816		3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.:45.0	Germany : 137	
Wife : 1568			Max. :99999	Max. :4356.0	Max. :99.0	Canada : 121	
						(Other) : 1709	

2.2 Data Wrangling

Processing question marks and adding a new factor level

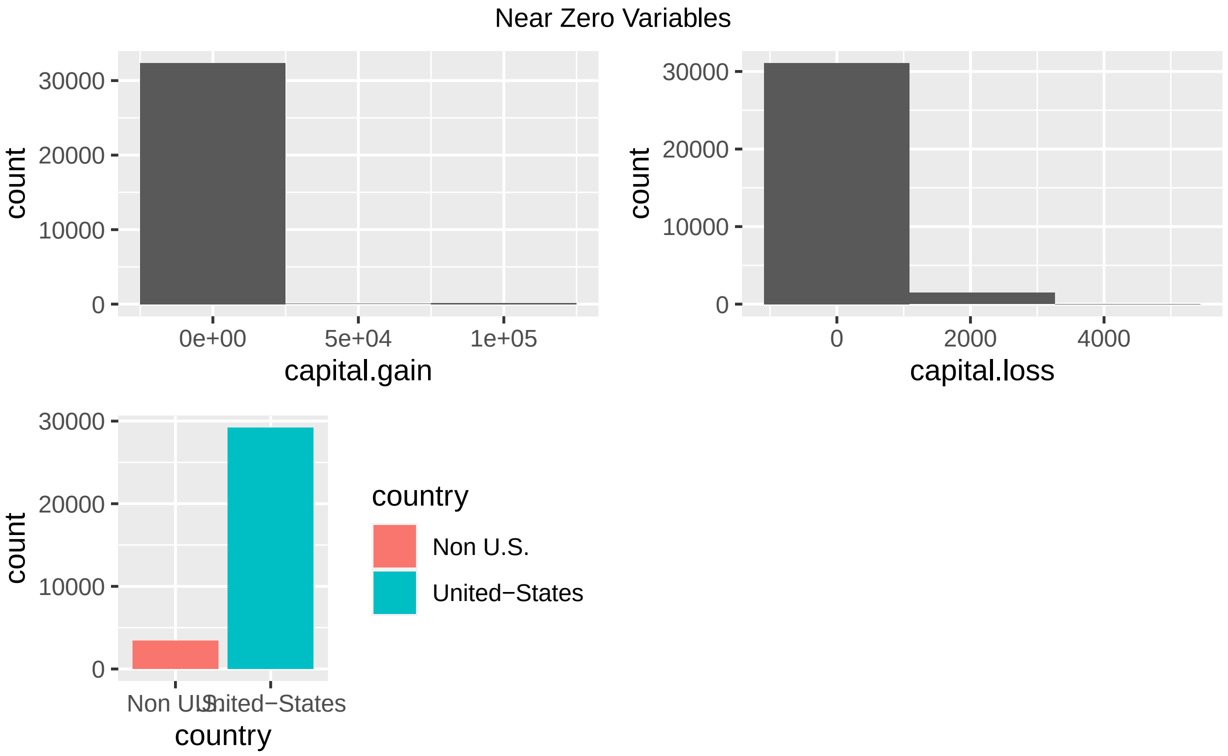
The adult income data has 4,262 question marks distributed in workclass, occupation, and native.country columns. To facilitate downstream processes, all question marks were replaced with “Unknown” and a new factor level “Unknown” was added to these three components.

The next step was to check if there are a large number of entries that have unknown values in all three components as they can skew prediction outcomes; there are only 27 such entries and hence were left in the data set.

2.3 Dimension Reduction

Near Zero Variables

The nearZeroVar function identified three nzv variables, they are capital.gain, capital.loss and native.country; the capital.gain has 91% zeros and capital.loss 95% zeros, 95% of native.country is the United-States. Near zero variables aren’t informative and hence were removed from the adult data set and the resultant data was stored in a new object called adultr to denote it’s a reduced dataset.

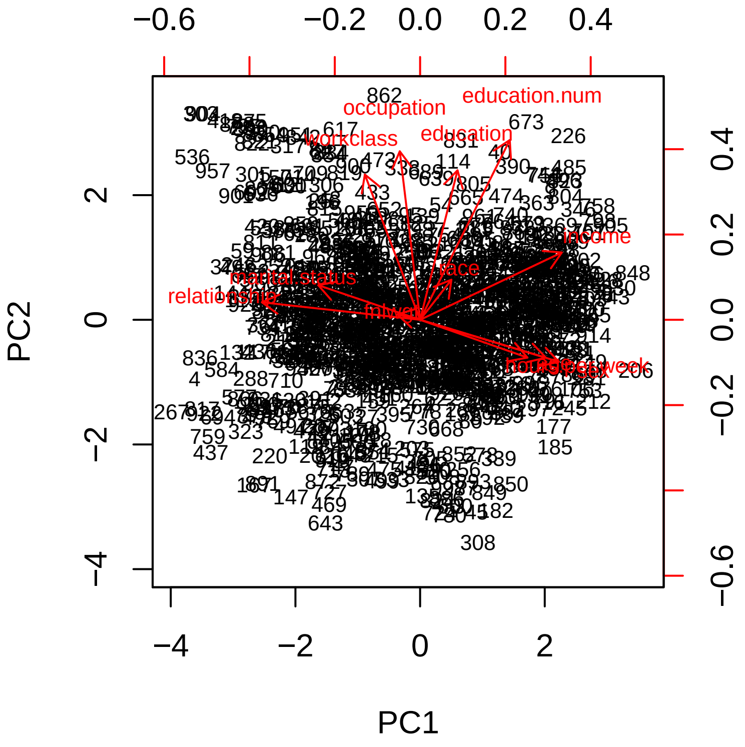


Principal Components

After three near zero variables were removed, prncomp function was run to check if the components could be trimmed further, but there wasn't more to be trimmed. The variability of sex's cumulative proportion is 0.9247 and hours.per.week's cumulative proportion is 0.9684. In order to maintain 95% variance all variables were retained.

```
## Importance of components:
##                         age workclass fnlwgt education education.num
## Standard deviation     1.51      1.20   1.15    1.0559        0.9971
## Proportion of Variance 0.19      0.12   0.11    0.0929        0.0829
## Cumulative Proportion  0.19      0.31   0.42    0.5126        0.5955
##                        marital.status occupation relationship   race    sex
## Standard deviation             0.9761     0.9201       0.9047 0.8321 0.8003
## Proportion of Variance         0.0794     0.0706       0.0682 0.0577 0.0534
## Cumulative Proportion          0.6749     0.7454       0.8136 0.8713 0.9247
##                        hours.per.week income
## Standard deviation             0.7242 0.6156
## Proportion of Variance         0.0437 0.0316
## Cumulative Proportion          0.9684 1.0000
```

Below is a principal components biplot created from 500 random samples:



2.4 Training and Test Data Separation

In a supervised machine learning process, observation data is typically split into a training set and a validation set. The training set is used to learn predictor behaviors and create recommendation algorithms, the validation set is used to produce final prediction results. The census income was split into a 26,371 records training set, called dev, and a 2,932 records validation set, called validation.

Overfitting Avoidance

In order to avoid overfitting and losing effectiveness for future data, the training dataset, dev, was further split into dev_train and dev_test; dev_test was used to select the final prediction model.

2.5 Model Training

Initial training used five methods - Generalized Linear Models (glm), Linear Discriminant Analysis (lda), k-Nearest Neighbour (knn ks = seq(34, 41, 1)), gamLoess, and Random Forest (rf mtry = c(3, 5)).

A model training function called run_models was written to provide the flexibility of testing different model combinations, datasets and tuning parameters.

Training Model Overall Accuracies

```
##      glm      lda      knn gamLoess       rf
##  0.79843  0.80355  0.75409  0.80593  0.83186
```

3 Results

In the multiple model trainings, random forest, rf, scored the highest overall accuracy and was chosen to be the final prediction model. Below is the final prediction accuracy from rf using the validation set:

Final Results

```
##      rf
## 0.82873
```

```
## Accuracy    Kappa
##  0.82873  0.51526
```

```
## Sensitivity Specificity
##     0.91411     0.57196
```

```
## rf variable importance
##
##                Overall
## age            100.000
## fnlwgt          98.354
## relationship    87.168
## education.num   84.769
## hours.per.week  58.448
## marital.status  56.913
## occupation      41.603
## workclass       20.480
## education       19.640
## race             0.846
## sex              0.000
```

3.1 Analyzing The Final Prediction Results

The table below reveals that low specificity is a major source of miss classification, over 40% of >50K cases were predicted to be <=50k.

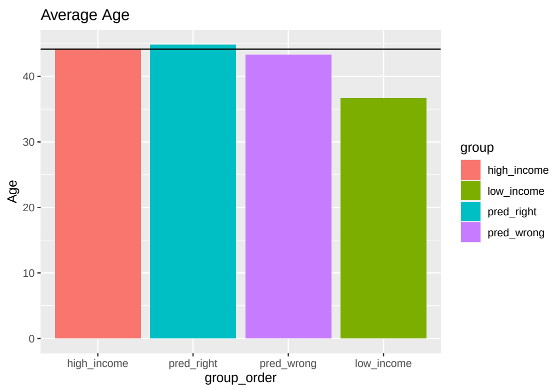
```
##           Reference
## Prediction <=50K >50K
##      <=50K  2235  348
##      >50K    210  465
```

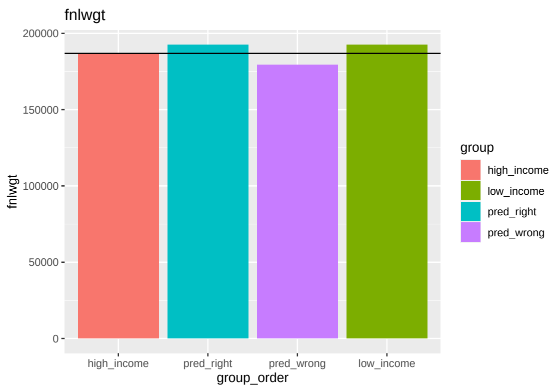
In an effort to root cause the low specificity that missclassified >50K to <=50k, four of the top five important predictors were investigated: age, education.num, fnlwgt and hours per week and their respective averages were compared in four groups:

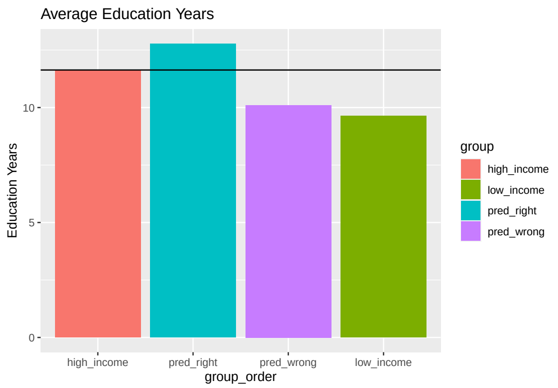
- high income: true >50K entries
- low income: true <=50K entries
- high income predicted correctly by rf: >50K predicted as >50K
- high income predicted wrongly by rf: >50K predicted as <=50K

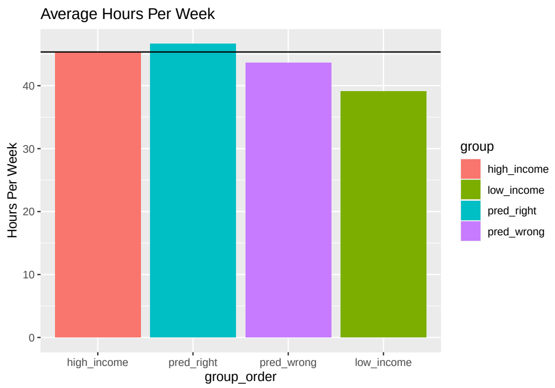
Averages of rf Important Predictors

group	edunum_avgs	age_avgs	fnlwgt_avgs	hwk_avgs
high_income	11.6322	44.164	186869	45.358
pred_right	12.7763	44.839	192519	46.671
pred_wrong	10.1035	43.261	179321	43.603
low_income	9.6348	36.677	192576	39.090









As it can be observed from above histgrams, the pred_wrong group's average education year is lower than the average of true high income group, similarly age, work hours per week and fnlwgt are all lower than the true high income group. So it appears that as far as random forest is concerned, the pred_wrong group doesn't fit in the "typical" profile of the high income group and as a result this group's income was predicted to be <=$50K.

3.2 Final Result

Prediction model: random forest 'rf', mtry=3

```
##      rf
## 0.82873
```

3.3 Performance

Run Time

The R script ran on a power lacking 2013 MacBook Pro with 8 GB Memory and 2.4 GHz Dual-Core Intel Core i5. The entire script took about 50 minutes to run including downloading and wrangling data, creating dev and validation summaries, running five evaluation models and then the final rf model.

Random Forest rf Model Run Time (in seconds)

```
## $everything
##    user  system elapsed 
## 622.352  19.375 657.857 
## 
## $final
##    user  system elapsed 
##  24.642   0.625  26.014 
## 
## $prediction
## [1] NA NA NA
```

R Script Run Time (in seconds)

```
##     user   system  elapsed 
## 3050.128   92.206 6923.991
```

4 Conclusion

This project created an adult income prediction model by learning behaviors of eleven predictors from the 1994 adult census, the algorithm is simple but takes time to run.

The data set is small with not many predictors, and hence random forest worked well; however due to it's long run time random forest may not be a suitable algorithm for larger datasets.

This is a preliminary step in machine learning; much more can be done in future. A next major step would be to study unsupervised machine learning using similar data.

References

I. Rafael and Harvard dev. Introduction to Data Science 2020