Hanan Adam

Statistical learning final project

Objective

The objective of this study is to analyze the Salary prediction datasets and fit at least three different statistical learning methods to make prediction of the individuals whose annual salary is less than equal to $50,000 or greater than $50,000 and also Use cross-validation (10-K fold CV) to compare the performance of the three different statistical learning methods. In addition, find out which predictors are significant in predicting the two salary group.

Explanation of the variables of the dataset

The Data was obtained form Kaggle,Which have 15 variables and 32,535 data points.

1. age : continuous.

2. workclass: a general term to represent the employment status of an individual

a. Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

3. fnlwgt: this is the number of people that census believes the entry represents :continuous.

4. education: Preschool , 1st-4th , 5th-6th , 7th-8th , 9th , 10th , 11th , 12th , HS-grad , Prof-school , Assoc-acdm , Assoc-voc , Some-college , Bachelors , Masters , Doctorate

5. education-num: a number that describe your education status from preschool to doctorate.

6. marital-status: marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.

Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

8. relationship: represents what this individual is relative to other

a. Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

10. sex: Female, Male.

11. capital-gain: continuous.

12. capital-loss: continuous.

13. hours-per-week: continuous.

14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland,

Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

15. salary: <=50K(Yes) or >50K(No)

Cleaning of Data

Rows with missing columns were delete from data set, while whiles columns with categorical responses were converted to factors. In addition, columns with numerical values were scaled. The data was finally divided into training and test data.

Methods

Since this is a classification problem, the statistical methods used in addition to the 10-fold cross validation as follows,

- k-nearest neighbors
- Random forests
- Gradient boosting machine
- support vector machine

The best statistical learning method was select based on the ROC of its 10-fold CV and the misclassification error based on the test data. Also the analysis was conducted using the Caret package for cross validation

Explanation of each statistical model and Results of Cross Validation

k-nearest neighbors

The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group, or another based on what group the data points nearest to it belong to. The KNN algorithm assumes that similar things exist in proximity. In other words, similar things are near to each other. The advantage of the k-nearest-neighbor is, it has no training period, hence example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data point's neighbors. However, it Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension. The results of the 10-fold CV for Knn is shown below with its prediction misclassification error rate .

```
k-Nearest Neighbors

22783 samples
   12 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 20504, 20505, 20504, 20505, 20504, 20506, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.8649799  0.6061319  0.9095307
   7  0.8770038  0.6065029  0.9166141
   9  0.8838101  0.6041013  0.9203566
  11  0.8873626  0.5985649  0.9207020
  13  0.8907059  0.5900713  0.9219693
  15  0.8921421  0.5948711  0.9229483
  17  0.8932213  0.5902565  0.9232361
  19  0.8945217  0.5922881  0.9239273
  21  0.8949381  0.5926557  0.9259428
  23  0.8959311  0.5911804  0.9264036

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 23.
     predicted.knn
        No  Yes
  No  1445  976
  Yes  565 6766
[1] 0.1580189
```

Random forests

Random forests aim to decorrelate the trees and hence improve the variance reduction of bagging. Unlike bagging, with random forest when building these decision trees, at each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. At each split only m < p predictors are allowed to use. Typically m is square root of p (classification) and m is equal to p divide by 3 (regression). By doing so, there are only at most m very correlated predictors across any two splits. Advantages of random forest is that Random Forest can automatically handle missing values and its comparatively less impacted by noise. Its disadvantage is Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.

```
Random Forest

22783 samples
   12 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 20504, 20505, 20504, 20505, 20504, 20506, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.8862092  0.1570602  0.9964872
  12    0.9027562  0.6120356  0.9435070
  23    0.9057510  0.6256952  0.9364242
  34    0.9039311  0.6266177  0.9296296
  44    0.9018270  0.6323366  0.9250225
  55    0.8997018  0.6292011  0.9210490
  66    0.8978558  0.6221852  0.9186302
  76    0.8964852  0.6238485  0.9172483
  87    0.8949262  0.6253269  0.9156359
  98    0.8941424  0.6251414  0.9141963

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 23.
     predicted.rf
        No  Yes
  No  1507  914
  Yes  479 6852
[1] 0.1428425
```

Gradient boosting machine

Boosting is primarily used to reduce the bias and variance in a supervised learning technique. In boosting each tree is fitted on a modified version of the original data set. Boosted trees are grown sequentially so that each tree is grown from using information from the previous trees. This method first involves building a decision tree with d splits (and d + 1 terminal notes) and next improving the model in areas where it under performed. This involves fitting a decision tree to the residuals of the model. This procedure is called learning slowly. The first decision tree is then updated based on the residual tree, but with a weight. The procedure is repeated until some stopping criterion is reached. Advantage Unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and overfitting, the boosting approach instead learns slowly and converts weak learners to strong learners. One disadvantage of boosting is that it is sensitive to outliers since every classifier is obliged to fix the errors in the predecessors. Thus, the method is too dependent on outliers.

```
Stochastic Gradient Boosting

22783 samples
   12 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 20504, 20505, 20504, 20505, 20504, 20506, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  ROC        Sens       Spec
   1                  50      0.8926404  0.3261317  0.9884829
   1                 100      0.9001971  0.4988933  0.9636044
   1                 150      0.9055499  0.5179038  0.9610707
   1                 200      0.9085294  0.5310093  0.9584217
  1                 250     0.9109729 0.5450372 0.9565793
  1                 300     0.9125418 0.5581409 0.9541029
  1                 350     0.9137718 0.5694003 0.9526055
  1                 400     0.9147086 0.5745674 0.9501867
  1                 450     0.9155026 0.5793665 0.9488621
  1                 500     0.9162805 0.5839801 0.9480559
  2                  50     0.9002811 0.5036938 0.9627407
  2                 100     0.9089233 0.5376561 0.9578462
  2                 150     0.9134916 0.5644167 0.9538149
  2                 200     0.9164008 0.5817657 0.9507627
  2                 250     0.9178717 0.5930237 0.9484017
  2                 300     0.9190020 0.5976380 0.9472501
  2                 350     0.9199623 0.6044676 0.9461558
  2                 400     0.9209573 0.6083452 0.9447162
  2                 450     0.9216281 0.6125888 0.9443707
  2                 500     0.9222784 0.6164650 0.9444283
  3                  50     0.9054853 0.5212286 0.9599763
  3                 100     0.9142153 0.5749364 0.9533540
  3                 150     0.9182419 0.5952405 0.9498988
  3                 200     0.9203729 0.6074230 0.9474227
  3                 250     0.9219363 0.6160977 0.9462709
  3                 300     0.9230821 0.6192336 0.9456375
  3                 350     0.9237032 0.6201574 0.9456375
  3                 400     0.9241438 0.6247679 0.9456950
  3                 450     0.9244900 0.6238454 0.9450616
  3                 500     0.9248105 0.6275368 0.9448890
  4                  50     0.9093929 0.5448517 0.9578457
  4                 100     0.9174449 0.5897027 0.9505321
  4                 150     0.9205176 0.6087139 0.9470773
  4                 200     0.9224797 0.6166512 0.9460411
  4                 250     0.9235643 0.6247737 0.9449468
  4                 300     0.9241560 0.6279082 0.9447738
  4                 350     0.9247617 0.6290176 0.9444858
  4                 400     0.9249529 0.6310451 0.9445434
  4                 450     0.9254021 0.6319679 0.9442554
  4                 500     0.9255931 0.6343678 0.9441402
  5                  50     0.9120601 0.5640450 0.9557728
  5                 100     0.9194321 0.6028057 0.9488048
  5                 150     0.9225510 0.6184935 0.9466166
  5                 200     0.9239141 0.6221839 0.9456954
  5                 250     0.9246811 0.6267968 0.9459257
  5                 300     0.9252761 0.6291987 0.9451770
  5                 350     0.9256255 0.6336261 0.9451193
  5                 400     0.9258004 0.6345489 0.9446009
  5                 450     0.9260320 0.6375026 0.9433339
  5                 500     0.9260889 0.6402702 0.9435643
  6                  50     0.9141568 0.5758603 0.9516268
  6                 100     0.9211559 0.6122204 0.9471926
  6                 150     0.9235129 0.6240320 0.9462711
  6                 200     0.9246567 0.6299360 0.9446014
  6                 250     0.9254867 0.6319649 0.9436797
  6                 300     0.9255854 0.6347351 0.9433920
  6                 350     0.9258628 0.6365805 0.9426432
  6                 400     0.9259308 0.6367636 0.9436223
  6                 450     0.9258158 0.6356563 0.9429312
  6                 500     0.9258511 0.6367643 0.9428736
  7                  50     0.9156860 0.5845366 0.9514539
  7                 100     0.9223063 0.6157270 0.9468472
  7                 150     0.9239855 0.6253259 0.9445437
  7                 200     0.9250309 0.6295708 0.9445438
  7                 250     0.9256039 0.6360287 0.9435648
  7                 300     0.9259538 0.6356610 0.9431042
  7                 350     0.9259205 0.6424890 0.9427006
  7                 400     0.9258101 0.6393504 0.9411460
  7                 450     0.9257422 0.6411954 0.9416068
  7                 500     0.9256224 0.6432266 0.9402821
  8                  50     0.9176609 0.5961643 0.9495534
  8                 100     0.9226875 0.6227411 0.9453498
  8                 150     0.9246979 0.6310461 0.9438525
  8                 200     0.9255288 0.6349196 0.9433917
  8                 250     0.9259624 0.6406402 0.9425279
  8                 300     0.9261746 0.6421169 0.9419520
  8                 350     0.9262927 0.6448882 0.9407427
  8                 400     0.9261140 0.6469174 0.9404551
  8                 450     0.9260120 0.6476567 0.9401672
  8                 500     0.9260199 0.6493189 0.9387276
  9                  50     0.9183556 0.5994850 0.9498991
  9                 100     0.9235359 0.6319700 0.9455802
  9                 150     0.9249386 0.6380579 0.9441977
  9                 200     0.9258509 0.6384275 0.9442555
  9                 250     0.9261629 0.6411957 0.9425282
  9                 300     0.9259504 0.6406416 0.9420100
  9                 350     0.9257582 0.6432249 0.9409733
  9                 400     0.9256479 0.6454430 0.9403399
  9                 450     0.9254247 0.6472887 0.9396486
  9                 500     0.9251324 0.6474739 0.9390729
 10                  50     0.9196010 0.6040993 0.9503021
 10                 100     0.9241835 0.6297529 0.9460410
 10                 150     0.9253377 0.6375047 0.9432767
 10                 200     0.9253968 0.6380575 0.9425855
 10                 250     0.9255917 0.6413803 0.9422402
 10                 300     0.9254076 0.6447047 0.9413188
 10                 350     0.9253429 0.6459989 0.9405704
 10                 400     0.9250433 0.6458158 0.9391884
 10                 450     0.9246746 0.6434145 0.9390154
 10                 500     0.9244053 0.6432290 0.9382669

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 10
ROC was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 350, interaction.depth
 = 8, shrinkage = 0.1 and n.minobsinnode = 10.
     predicted.gbm
        No  Yes
  No  1539  882
  Yes  438 6893
[1] 0.1353568
```

support vector machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. The advantages of the SVM method are the better accuracy in classification and the best performance in the analysis. SVM is effective in cases where the number of dimensions is greater than the number of samples and works relatively well when there is a clear margin of separation between classes. However, As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.

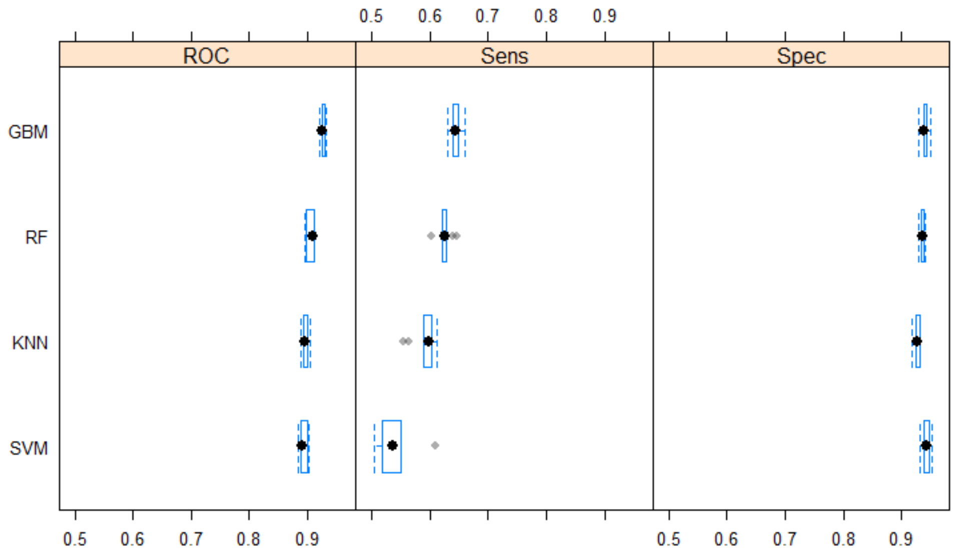
```
Support Vector Machines with Radial Basis Function Kernel

22783 samples
   12 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 20504, 20505, 20504, 20505, 20504, 20506, ...
Resampling results across tuning parameters:

  C       ROC        Sens       Spec
    0.25  0.8923413  0.5394981  0.9391883
    0.50  0.8926668  0.5387597  0.9428157
    1.00  0.8921972  0.5418993  0.9437369
    2.00  0.8904089  0.5463274  0.9433338
    4.00  0.8882398  0.5430022  0.9421242
    8.00  0.8853657  0.5418973  0.9395905
   16.00  0.8823578  0.5426349  0.9383235
   32.00  0.8747186  0.5342463  0.9358841
   64.00  0.8693975  0.5319864  0.9343485
  128.00  0.8715065  0.5442804  0.9366544

Tuning parameter 'sigma' was held constant at a value of 0.01177868
ROC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.01177868 and C = 0.5.
     predicted.svm
        No  Yes
  No  1311 1110
  Yes  421 6910
[1] 0.1569934
```

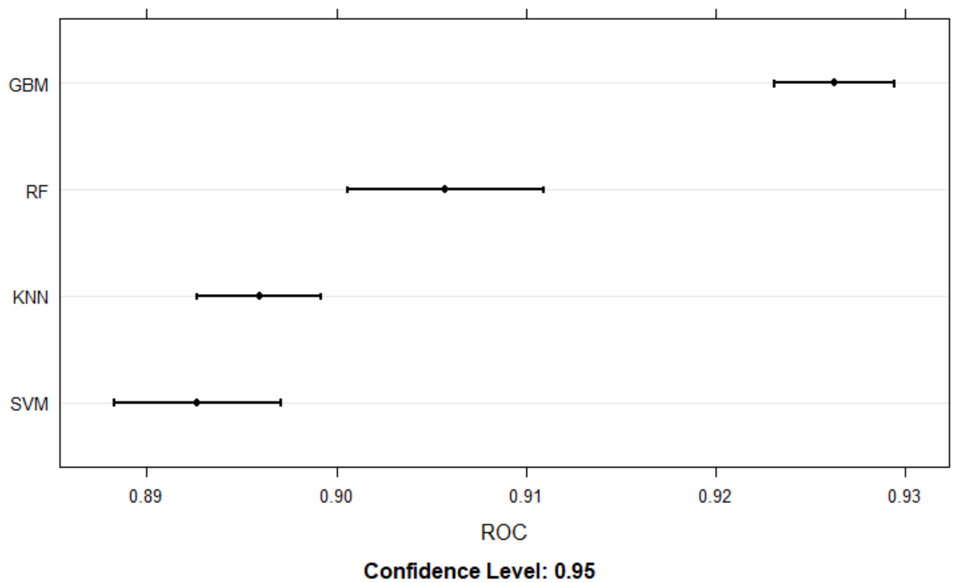
<u>Performance based on Misclassification error by using test data</u>

Statistical model	**Misclassification error**
Gradient boosting machine	0.1353568
Random forests	0.1428425
support vector machine	0.1569934
k-nearest neighbors	0.1580189

Gradient boosting machine is the best model based on its small misclassification error rate.

Visualization of model performance base on ROC





In general, you want the model with the higher median AUC, as well as a smaller range between min and max AUC. Hence, we can observe that the gradient boosting machine has the highest median ROC value with the lowest range between the minimum and Maximum ROC value.

T-test for model selection

Q[1]-> k-nearest neighbors

Q[2]-> Random forests

Q[3]-> Gradient boosting machine

Q[4]-> support vector machine

```
        Paired t-test

data:  Q[, 1] and Q[, 2]
t = -5.2037, df = 9, p-value = 0.0005613
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01408873 -0.00555092
sample estimates:
mean of the differences
           -0.009819824


        Paired t-test

data:  Q[, 1] and Q[, 3]
t = -31.77, df = 9, p-value = 1.489e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03252349 -0.02819968
sample estimates:
mean of the differences
            -0.03036158


        Paired t-test

data:  Q[, 2] and Q[, 3]
t = -13.226, df = 9, p-value = 3.348e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02405509 -0.01702843
sample estimates:
mean of the differences
            -0.02054176


        Paired t-test

data:  Q[, 1] and Q[, 4]
t = 1.7638, df = 9, p-value = 0.1116
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0009223348  0.0074509368
sample estimates:
mean of the differences
            0.003264301
```

```
        Paired t-test

data:  Q[, 2] and Q[, 4]
t = 7.4522, df = 9, p-value = 3.883e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.00911238 0.01705587
sample estimates:
mean of the differences
             0.01308413


        Paired t-test

data:  Q[, 3] and Q[, 4]
t = 23.084, df = 9, p-value = 2.555e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03033061 0.03692115
sample estimates:
mean of the differences
             0.03362588
```

<u>Summary of t-test</u>

T-test	**Significance**
KNN vs Random forest	Statistically the different
KNN vs Gradient boosting machine	Statistically the different
KNN vs support vector machine	Statistically the same
Random forest vs Gradient boosting machine	Statistically the different
Random forest vs support vector machine	Statistically the different
Random forest vs Gradient boosting machine	Statistically the different

From our analysis of t-test we realized that it was only k-nearest neighbors and support vector machine that has its p-value to be greater than 0.05(Hence there is no statistical difference between the two models and they both performed bad compared to both Random Forest and Gradient boosting machine). Whiles the t-test for Random Forest and Gradient boosting machine shows significant difference among them. Also based on the ROC and misclassification error rate we consider the t-test on Gradient boosting machine and Random forest(statistically different model).Moreover Gradient boosting machine has the highest ROC and lowest misclassification error rate, hence its chosen as our final model..

Fitting final model based on the final chosen model

A model was fitted based on the Gradient boosting machine using the best parameters that was obtain from performing 10-fold CV. This model shows how each independent factors contributed (variable relevance) in making prediction of the dependent variable.

	var <chr>	rel.inf <dbl>
relationship	relationship	25.2238133
education	education	16.0018593
capital.gain	capital.gain	14.8610026
occupation	occupation	12.4367574
age	age	7.0496723
native.country	native.country	5.6926476
capital.loss	capital.loss	5.2056371
marital.status	marital.status	5.1194243
hours.per.week	hours.per.week	4.7615525
workclass	workclass	2.9002142