Data mining

PROJECT 2 CLASSIFICATION

Zhang, Shiyang

47319809

Contents

Charts and Table contents:

1. Execution Summary

In this project, we want to predict income according to given information like education, age, LOS, agency, etc. I use data 2001_03, 2005_03, 2009_03 and 2013_03 to do different prediction. Because 2001 is the first year of president Bush and 2009 is the first year of president Obama. Also, 2005 and 2013 are respectively 4 years later after their administration. I think analyzing these four years will be useful. In this project, I use decision tree, Ctree, PART, and KNN to do analysis.

I get some conclusion after analyzing data:

1. Supervisory Status is only important in year 2001. In year 2005 and later, it becomes trivial.
2. Decision tree has higher accuracy compared to other three models.
3. The most important factors are Education, LOS, Agency and State.
4. Under different administration, important factors are slightly different. Like Supervisory are not important at Obama government. But education, LOS and agencies are quite important.
5. If man want to predict incomes, using combination of prediction model may be useful.

2. Data Preparation

2.1 Basic cleaning for data[1]

At first, do basic cleaning for invalid data set. Using NA replace unknown characters like '########', 'UNSP', '*'. Then dealing with duplicate IDs:

1. Select rows with duplicate IDs
2. Order selection by ID, then Agency, then descending Pay
3. Select rows where the ID and Agency are duplicated (same employee at same agency)
4. Get row numbers for rows with the lowest pay for each grouping in the above selection
5. Reselect from data frame where rows are not required to removed

Then calculating percent of data saved.

2.2 Feature creation and feature selection

I add two features in original data set. One is agency name, the other one is state name result from substrings of station number. By using tool FSelector, I get attributes that are most important to salary prediction. Results are figure 1 and 2.

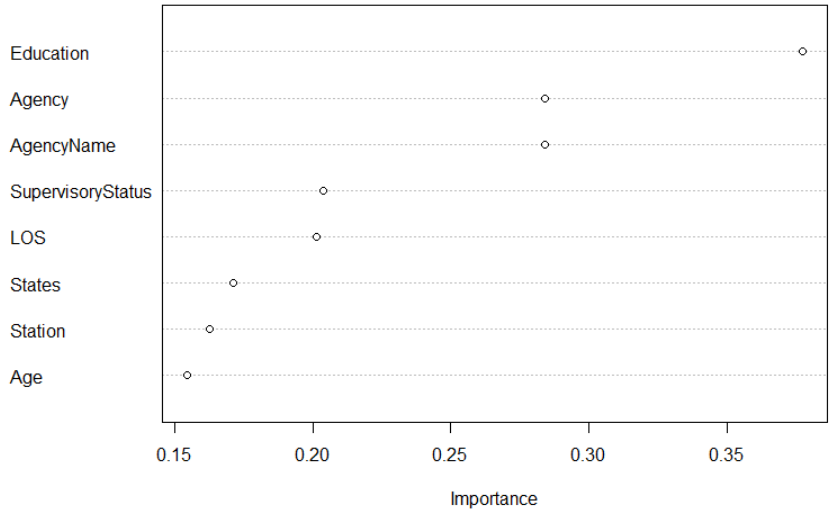

Fig 1 2001 dominant attributes

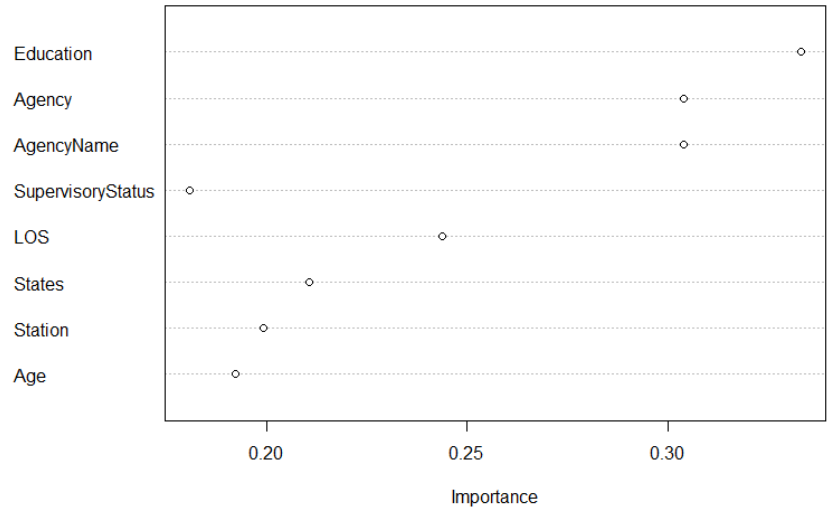

Fig 2 2005,2009 and 2013 dominant attributes

From figure1 and figure2, we can see most attributes are the same except supervisory. In 2001, supervisory status is a very import feature. But in 2005, 2009 and 2013, this feature is not so crucial. In addition, LOS becomes more important.

Since agency represents the same information with agency name, as well as states and station. Therefore, I choose Pay, Education, agency name, supervisory status, LOS, states, and age as features to do modelling analysis[2]. Before analyzing, I choose data that agencies have more than 10000 employees. Then sample 10000 points from those data. Otherwise, the amount of data is too large to build model quickly. For different models, the scale of features are different. I use sample data(10000 points) from data set 2001_03 as an example for data description. I make some attributes become ratio for better modelling.

Features	Scale	Description	Operation
Education	ratio	Range: 1-22 Median:10	Fill NA Education with median Education for employees at same Agency with same Age, otherwise fill with median education for Agency
Agency Name	nominal	Counting: e.g. VATA:2723 TR93:1534	Sometimes agency name becomes dummy values like 0-1 to represent if it belongs to this agency.
Supervisory Status	ratio	Range:2-8 Median: 8	Impute NA with median
LOS	ratio	Range:1-35 Median:12	Used to be discrete variable, use middle number in a range to replace for conveniently building model. Fill NA LOS with median LOS for employees of the same Age
States	nominal	Counting: e.g.Maryland:606 NY:477	Drop NA.
Age	ratio	Range:17-75 Median:47	Used to be discrete variable, use middle number in a range to replace for conveniently building model.
Pay	ordinal	Divide into 4 catigories: <50k, 50-75k, 75-100k,>100k	Drop NA pay

Table 1 Feature description for decision tree model

Building other models, I use subset from sample data comparing crosswise. Because Agency name and States are nominal variables, it is slower and harder for modelling. In addition, from figure 2 we can know, supervisory status is not important at year 2013. So I only use three most important features—Education, LOS and Pay to do prediction.

Features	Scale	Description
Education	ratio	Range: 1-22 Median: 12
LOS	ratio	Range: 1-35 Median: 7
Pay	ordinal	Count:<50k:2879, 50-75k:2834 75-100k:2026,>100k: 2210

Table 2 Feature description for C4.5 & PART model

Because KNN model needs features are ratio variables so that it can count distance from 1 point to other points. I convert pay from discrete into continuous variable.

Features	Scale	Description
Education	ratio	Range:1-22 Median:12
LOS	ratio	Range:1-35 Median:7
Pay	ratio	Range:17489-336521 Median: 67931

Table 3 Feature description for KNN model

I categorize LOS and Education this case because the ctree is too large if education and LOS are continuous number.

Features	Scale	Description
Education	ordinal	Count: <11:4063,11-22:4092
LOS	ordinal	Count: <10:3324, 10-20:2399 20-30:1698, >30: 734
Pay	ordinal	Count:<50k:1909, 50-75k:2481 75-100k: 1822, >100k: 1943

Table 4 Feature description for CTree model

3. Modeling

For this part, I first use decision tree model to do prediction for different years. Then I use 4 different models -- CTree, KNN, PART and C4.5 Decision tree to do model comparison for data 2013_03.

3.1 Decision Tree model[3]

3.1.1 2001 Education~Agencyname~Supervisory Tree

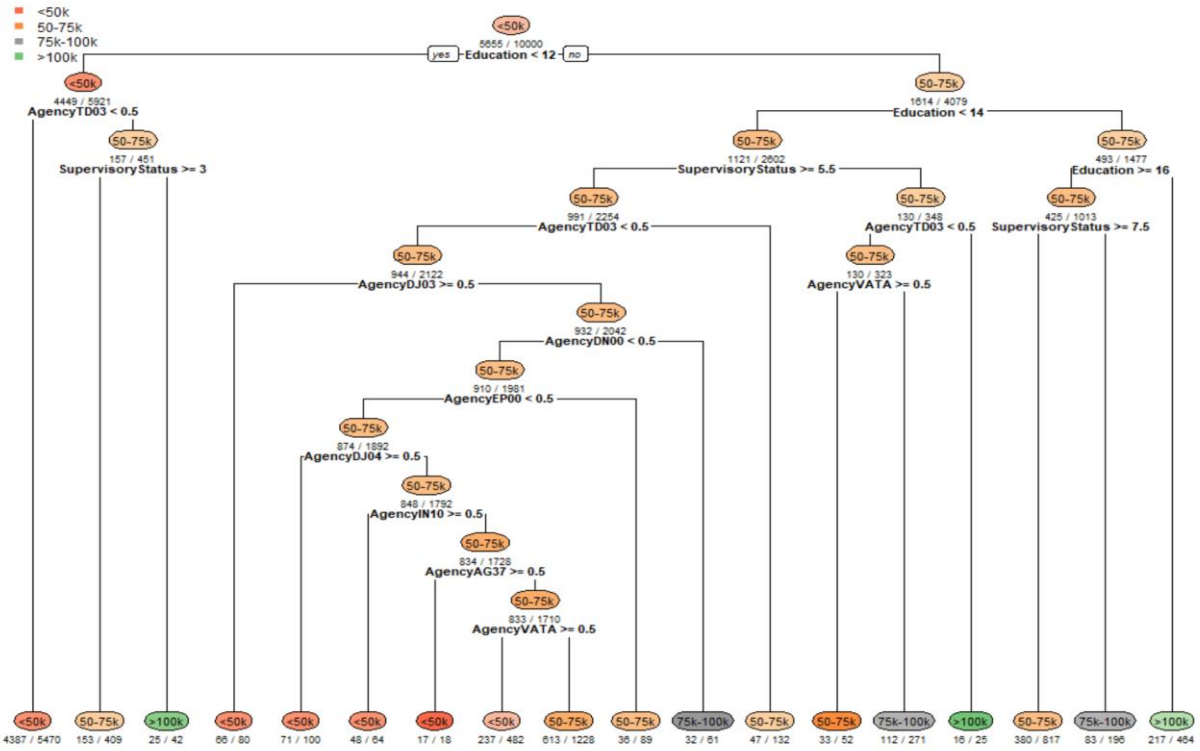

Fig 3. 2001 Education~Agencyname~Supervisory Tree

This model shows the prediction of Pay from Agency, Supervisory and Education aspect. From figure3, we can see that Education and supervisory status distribute broader than agency name. It means that in different agency, salary varies a lot. But the requirement for education level and supervisory status are similar. If education level is high, the people have more chance to get higher salary. The advantage of this model is that it uses the most important three features in 2001 to do the prediction. In addition, for each leaf node, the confidence estimation is different. We can predict very accurately if the result locates in high confidence node. For example, the 7th node from left leaf node has confidence 17/18. This is a pretty high confidence. So if people satisfy conditions listed above, the prediction will be accurate.

The most important variables in this model is

Education	100.000
AgencyTD03	32.845
SupervisoryStatus	10.865

Compared to education and agency, supervisory status dominants less.
By using caret tool, it use k-fold cross validation. Caret packages internally splits the data into training and testing sets and thus will provide us with generalization error estimates like table 5. Then we can choose optimal final model to generate tree.

cp	Accuracy	Kappa
0.002646720	0.6512994	0.3868126
0.003452244	0.6460991	0.3825003
0.003682394	0.6445991	0.3815442
0.003797468	0.6413991	0.3765566
0.004142693	0.6378992	0.3741391
0.004833142	0.6354001	0.3722018
0.008745685	0.6318003	0.3643760
0.017146145	0.6201006	0.3344917
0.021864212	0.6090006	0.3013422
0.093901036	0.5777963	0.1083626

Table 5 Accuracy and Kappa for data set

When we do prediction by using testing data, we get some statistics like table6 and figure 4. Table 6 represent statistics by class, figure 4 represents overall statistics.

	Class: <50k	**Class: 50-75k**	**Class: 75k-100k**	**Class: >100k**
Sensitivity	0.8439716	0.4832714	0.1818182	0.5087719
Specificity	0.6766055	0.7961696	0.9707865	0.9692471
Pos Pred Value	0.7714749	0.4659498	0.4347826	0.5000000
Neg Pred Value	0.7702350	0.8072122	0.9056604	0.9702760
Precision	0.7714749	0.4659498	0.4347826	0.5000000
Recall	0.8439716	0.4832714	0.1818182	0.5087719
F1	0.8060965	0.4744526	0.2564103	0.5043478
Prevalence	0.5640000	0.2690000	0.1100000	0.0570000
Detection Rate	0.4760000	0.1300000	0.0200000	0.0290000
Detection Prevalence	0.6170000	0.2790000	0.0460000	0.0580000
Balanced Accuracy	0.7602886	0.6397205	0.5763023	0.7390095

Table 6 Class statistics

```
                Accuracy : 0.655
                  95% CI : (0.6246, 0.6845)
    No Information Rate : 0.564
    P-Value [Acc > NIR] : 2.729e-09

                   Kappa : 0.3932
 Mcnemar's Test P-Value : 7.925e-11
```

Fig 4. Overall statistics

This model has testing accuracy 0.655 with kappa 0.3932. For class prediction, <50K and >100k are more accurate than other two classes.

3.1.2 2005 Education~Agency~LOS Tree

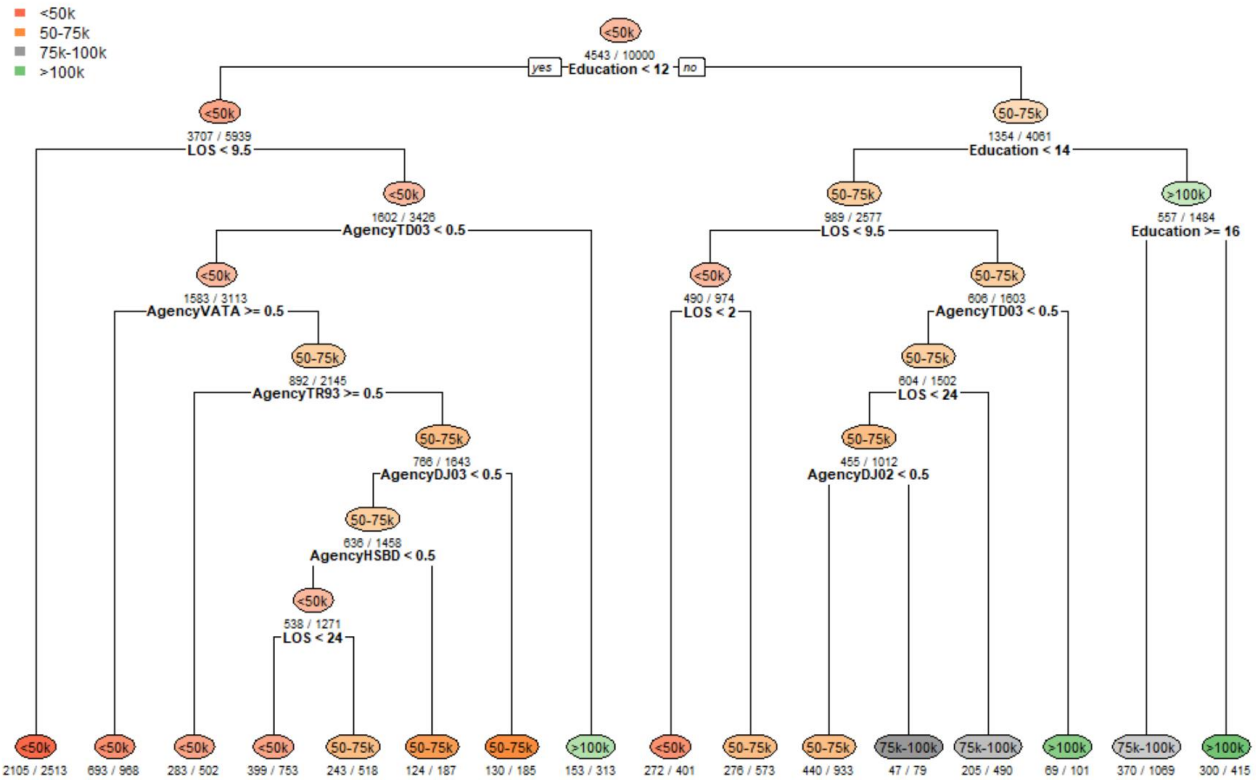

Fig 5. 2005 Education~Agency~LOS Tree

In 2005, supervisory status is not import as fselector indicates, so I use education, LOS and agency as features to predict pay. The importance of this model is

LOS	100.0000
Education	89.9119
AgencyTD03	61.6149
AgencyVATA	26.4044
AgencyHSBD	25.0916

Compare to last model, the agencyTD03 becomes more important. Also, LOS, agencyVATA and agencyHSBD becomes important as well. Education dominants decreasingly. It may testify that under bush administration, the education is important, but LOS is more crucial for fairly arrange salary for people. If someone works longer, he get more pay. This will encourage people go to work early once one have a okay diploma. In addition, TD03, VATA and HSBD agencies highly account for salary distribution. These three agencies are large department.

	Class: <50k	Class: 50-75k	Class: 75k-100k	Class: >100k
Sensitivity	0.8449438	0.4389439	0.3766234	0.5102041
Specificity	0.7693694	0.8364419	0.8711584	0.9645233
Pos Pred Value	0.7460317	0.5384615	0.3473054	0.6097561
Neg Pred Value	0.8608871	0.7742364	0.8847539	0.9477124
Precision	0.7460317	0.5384615	0.3473054	0.6097561
Recall	0.8449438	0.4389439	0.3766234	0.5102041
F1	0.7924131	0.4836364	0.3613707	0.5555556
Prevalence	0.4450000	0.3030000	0.1540000	0.0980000
Detection Rate	0.3760000	0.1330000	0.0580000	0.0500000
Detection Prevalence	0.5040000	0.2470000	0.1670000	0.0820000
Balanced Accuracy	0.8071566	0.6376929	0.6238909	0.7373637

Table 7 Class statistics

```
               Accuracy : 0.617
                 95% CI : (0.5861, 0.6472)
    No Information Rate : 0.445
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4259
 Mcnemar's Test P-Value : 0.001025
```

Fig 6. Overall statistics

In this model, overall accuracy is 0.617 with kappa 0.4259. For class prediction, <50k and >100k have Higher accuracy. The advantage for this model is it has relatively high accuracy.

3.1.3 2005 Education~LOS~States

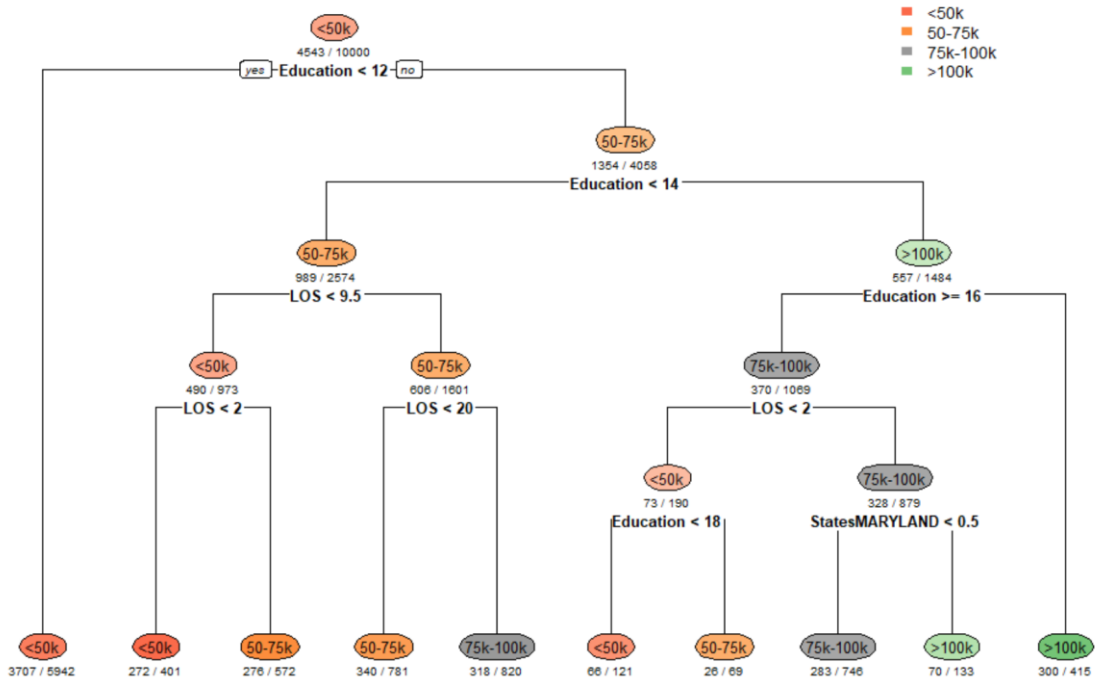

Fig 7. 2005 Education~LOS~States

In this model, Three features are LOS, States and Education. This tree is simple because there are two continuous variables. It is convenient for calculating. The most import four variables for this model are:

Education	100.0000
LOS	88.1754
StatesMARYLAND	15.4721
StatesDISTRICT OF COLUMBIA	14.0038

In this case, education accounts for more than LOS. The proportion is different from last case may result from states. In different states, education and LOS has different preference. Besides, Maryland and Colombia may have large people in work. Maybe their business is good.

	Class: <50k	Class: 50 - 75k	Class: 75k-100k	Class: >100k
Sensitivity	0.9084821	0.2230216	0.3795181	0.4629630
Specificity	0.5579710	0.9030471	0.8848921	0.9910314
Pos Pred Value	0.6251920	0.4696970	0.3962264	0.8620690
Neg Pred Value	0.8825215	0.7511521	0.8775268	0.9384289
Precision	0.6251920	0.4696970	0.3962264	0.8620690
Recall	0.9084821	0.2230216	0.3795181	0.4629630
F1	0.7406733	0.3024390	0.3876923	0.6024096
Prevalence	0.4480000	0.2780000	0.1660000	0.1080000
Detection Rate	0.4070000	0.0620000	0.0630000	0.0500000
Detection Prevalence	0.6510000	0.1320000	0.1590000	0.0580000
Balanced Accuracy	0.7332266	0.5630343	0.6322051	0.7269972

Table 8 Class statistics

```
                Accuracy : 0.582
                  95% CI : (0.5507, 0.6128)
    No Information Rate : 0.448
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.3459
 Mcnemar's Test P-Value : < 2.2e-16
```

Fig 8. Overall statistics

In this model, the overall accuracy is 0.582 with kappa 0.3459. For class prediction, <50k and >100k has more accurate prediction. Compared to last model, maybe last model is more useful for pay prediction. The advantage of this model is simple tree.

3.1.4 2013 Education~LOS~States

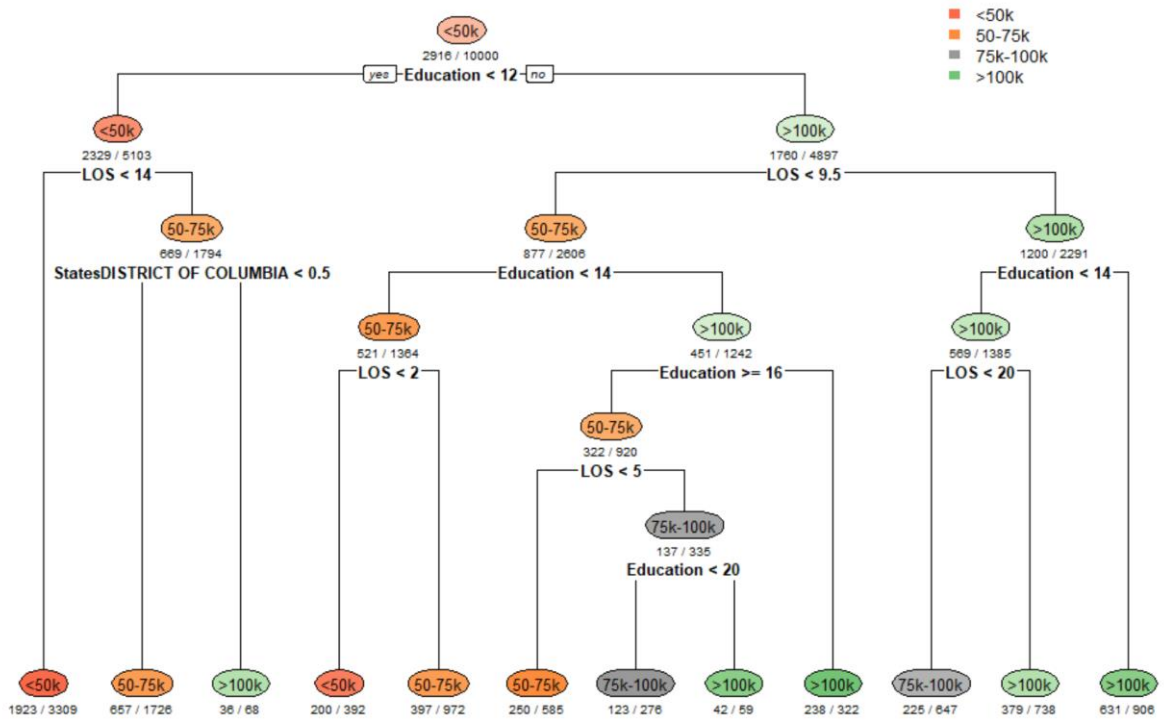

Fig 9. 2013 Education~LOS~States

In this case, I still use education, LOS and states as features to do prediction. The most important variables are:

Education	100.0000
LOS	83.7787
StatesDISTRICT OF COLUMBIA	23.5078
StatesMARYLAND	21.0937

The LOS becomes less important compared to last case. Also, columbia and maryland exchange ranking. Maybe because columbia becomes poorer or richer.

	Class: <50k	Class: 50-75k	Class: 75k-100k	Class: >100k
Sensitivity	0.5354610	0.3457249	0.1545455	0.7719298
Specificity	0.8830275	0.6703146	0.9089888	0.8186638
Pos Pred Value	0.8555241	0.2784431	0.1734694	0.2046512
Neg Pred Value	0.5950541	0.7357357	0.8968958	0.9834395
Precision	0.8555241	0.2784431	0.1734694	0.2046512
Recall	0.5354610	0.3457249	0.1545455	0.7719298
F1	0.6586696	0.3084577	0.1634615	0.3235294

Prevalence	0.5640000	0.2690000	0.1100000	0.0570000
Detection Rate	0.3020000	0.0930000	0.0170000	0.0440000
Detection Prevalence	0.3530000	0.3340000	0.0980000	0.2150000
Balanced Accuracy	0.7092443	0.5080198	0.5317671	0.7952968

Table 9 Class statistics

```
                Accuracy : 0.456
                  95% CI : (0.4248, 0.4875)
    No Information Rate : 0.564
    P-Value [Acc > NIR] : 1

                   Kappa : 0.2093
 Mcnemar's Test P-Value : <2e-16
```

Fig 10. Overall statistics

In this model, overall accuracy is 0.456 with kappa 0.2093. Also, <50k and >100k has higher accuracy.

3.2 PART-rule based model[4]

For following models, model evaluation uses 10-fold cross validation method.

Part of rule based model result:

```
LOS > 17 AND
Education > 9 AND
Education <= 10 AND
LOS <= 32 AND
LOS <= 27: 50-75k (117.0/77.0)

LOS > 17 AND
Education <= 10 AND
LOS <= 32 AND
Education <= 9: 50-75k (390.0/241.0)

LOS <= 17: 75k-100k (18.0/9.0)

Education > 11: 75k-100k (16.0/10.0)

LOS > 32 AND
Education <= 10: 50-75k (12.0/5.0)

LOS <= 32 AND
Education <= 10: >100k (26.0/16.0)

LOS <= 32 AND
LOS <= 27: >100k (21.0/12.0)

: 75k-100k (15.0/10.0)

Number of Rules  :      45
```

Fig 11. PART model

The advantage of this model is that it list every entry very clearly. We only need to compare conditions and trace the condition, then we can get our class. The most important features for this model is

Education 100
LOS 0

threshold	pruned	Accuracy	Kappa
0.0100	yes	0.5149253	0.3399471
0.0100	no	0.5168350	0.3427132
0.1325	yes	0.5178402	0.3440378
0.1325	no	0.5168350	0.3427132
0.2550	yes	0.5182416	0.3447774
0.2550	no	0.5168350	0.3427132
0.3775	yes	0.5181411	0.3447013
0.3775	no	0.5168350	0.3427132
0.5000	yes	0.5175383	0.3440411
0.5000	no	0.5168350	0.3427132

Table 10 Accuracy and Kappa of PART model

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were threshold = 0.255 and pruned = yes.

3.3 K-Nearest Neighbors model

```
k   RMSE      Rsquared   MAE
 1  27830.89  0.4885852  20544.17
 2  27867.28  0.4871538  20585.07
 3  27891.43  0.4863443  20596.07
 4  27904.69  0.4858654  20610.08
 5  27919.44  0.4853362  20625.33
 6  27952.90  0.4840571  20634.34
 7  27956.33  0.4839432  20631.92
 8  27967.61  0.4835084  20642.52
 9  27978.94  0.4831161  20654.14
10  27985.72  0.4828362  20657.43

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was k = 1.
```

Fig 12. KNN model

For this model, KNN uses Euclidean distance, so data should be standardized (scaled) first. Here LOS, Education and Pay are ratio variables. The advantage for this model is that when drawing the image

of this model, it will directly shows distances between nodes. So we can classify classes by dividing points. The most important variable is:

Education	100
LOS	0

3.4 CTree model

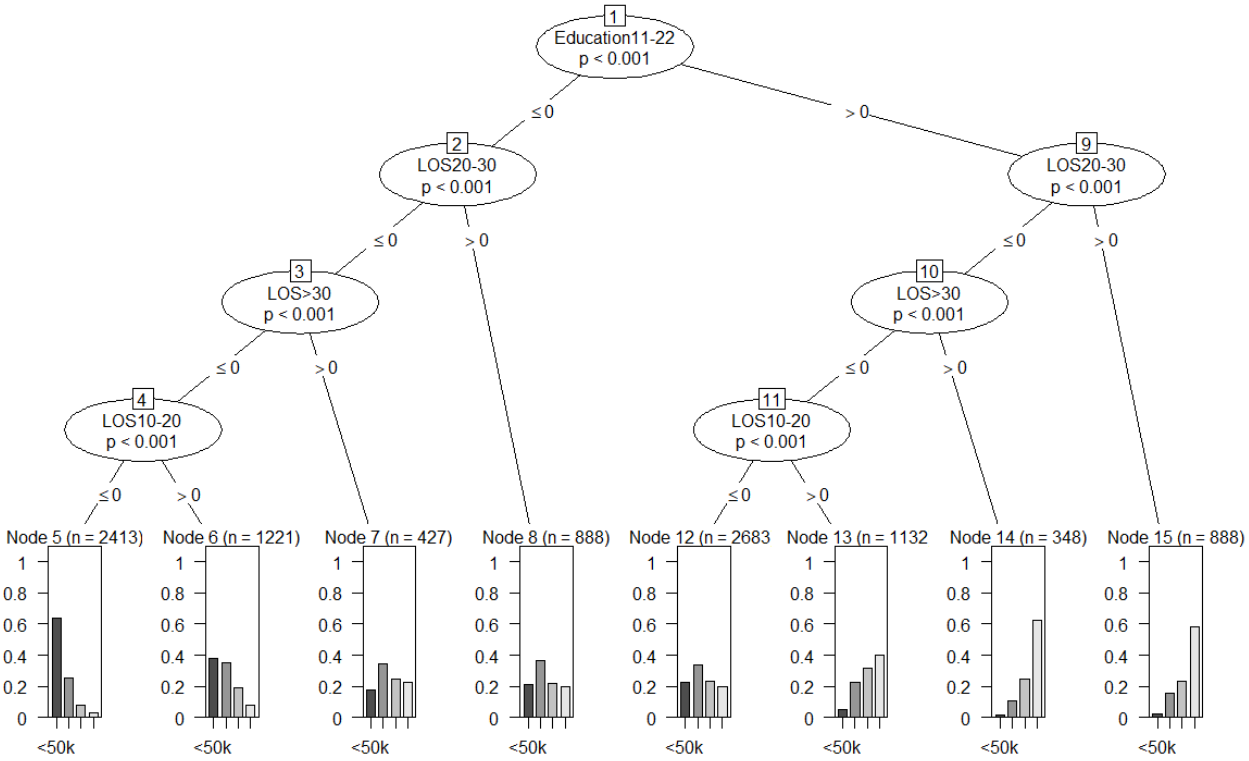

Fig 13. CTree model

In this model, all three features are categorial variables. From the leaf node of figure 13, we can see from left to right, the histogram varies a lot. At the most left one, <50k counts most while at the rightest side, >100k becomes the most. There is a transformation between those nodes. Comparing conditions above, the left side has lower education and lower LOS. In the middle leaf node, they have fairly similar distribution with 4 pay classes. Also, most of them do not have large LOS or Education.

The advantage of this model is that it shows very clearly. It combines tree and histogram together to better represent information for users. Users do not need to read statistics if they are not familiar with handling data. The most important variables for this model:

	X.50k	X50.75k	X75k.100k	X.100k
Education	65.74	100.00	19.67	100.00
LOS	20.14	59.88	0.00	59.88

For different classes, education and LOS counts differently.

mincriterion	Accuracy	Kappa
0.010	0.4566563	0.2525541
0.255	0.4566563	0.2525541
0.500	0.4566563	0.2525541
0.745	0.4566563	0.2525541
0.990	0.4566563	0.2525541

Table 11 CTree accuracy and kappa

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mincriterion = 0.99.

3.5 Model comparison

The comparison is used for decision tree, ctree and part model. It seems like that decision tree has higher accuracy. Also, the difference for different models are small.

```
summary.resamples(object = resamps)

Models: rpart, ctree, rules
Number of resamples: 10

Accuracy
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rpart 0.4860000 0.5060060 0.5090000 0.5088999 0.5147500 0.5310000    0
ctree 0.4162621 0.4275758 0.4336760 0.4406079 0.4570848 0.4690909    0
rules 0.5055276 0.5100503 0.5180905 0.5188479 0.5271357 0.5362173    0

Kappa
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rpart 0.2991967 0.3283333 0.3306076 0.3312212 0.3400077 0.3607419    0
ctree 0.2120742 0.2273914 0.2341074 0.2439469 0.2647628 0.2836881    0
rules 0.3271455 0.3343632 0.3461096 0.3458549 0.3573531 0.3697179    0
```

Fig 14. Resample figure

```
summary.diff.resamples(object = difs)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
      rpart     ctree     rules
rpart            0.068292 -0.009948
ctree 7.195e-06           -0.078240
rules 0.2992    5.300e-08

Kappa
      rpart     ctree     rules
rpart            0.08727  -0.01463
ctree 1.012e-05           -0.10191
rules 0.2307    7.431e-08
```

Fig 15. Diff summary

4. Evaluation and Deployment

For stake holders, he can combine those models together to do the prediction.
For employers, he can use KNN to get rough point information about company's salary distribution. Then make some adjustments for it.
For employees, if they want to know how much they can get according to their condition, they can combine different models. For example, they can use Ctree model first, get rough categorial information. Because histogram can help people directly get information. Then they will know which feature is important. Then use decision tree or rule-based model doing prediction so exactly ensure agencies' treatment. Decision tree may lose unimportant information. But rule-based model is divided fine.

5. Refences

1. https://github.com/jakecarlson1/data-mining-projects/blob/master/project-2/project2.R
2. https://github.com/jicee13/Obush/blob/master/rCode/idk.r
3. https://rawgit.com/mhahsler/Introduction_to_Data_Mining_R_Examples/master/chap4.html
4. https://rawgit.com/mhahsl