PREDICTING INCOME USING UNITED STATE ADULT CENSUS DATA

Humaira Asim

Final Report

https://github.com/HumairaAsim/CKME136

Introduction

Census is the process of systematically recording statistical information about the nation's population. The data is comprised of various people, their distribution, their living conditions, education and other key factors, which are critical for the development of the nation. This data helps the policy makers construct decisions for the future and betterment of the country. Income is one of the primary concerns for the standard of living and economic status of an individual and thus, has a significant impact on determining the nation's growth and prosperity.

In this project, the aim is to explore the U.S census income data set, relating the earnings/income with demographic information such as; age, gender, marital status, race, education level, employment type etc… and to build a classifier which can predict whether the income of an individual is greater than or less than $50K/year. The problem is a binary classification problem, since the target variable income has binary values.

To make these predictions, I will explore the census income data and determine the most remarkable features in the census income data, for predicting the income class of an individual and use these features to build/train models on training data, using different machine learning classification algorithms. The models will be tested on the test data, and the one with the highest accuracy will be selected as the best predictive model.

Research Questions:

1. Is the income of an individual greater than or less than $50K a year?

Literature Review

To obtain a better understanding and insight of the given problem, I reviewed several publications which focus on dealing with binary classification problems and adult census income data. Ron Kohavi in 1996, used income dataset to build a hybrid algorithm NBTree [1], which combines the features of decision-tree and naïve bayes classifiers. In this algorithm, the decision-tree nodes contain univariate splits (segmentation), like decision tree but at the leaves classification is done using naïve Bayesian classifiers. The author is able to prove through his experiments that NBTree works better on real-world datasets, like adult census data. It outperformed both decision tree and naïve bayes, but takes more time with regards to speed.

In another study, the author used random forest classifier as it gives better performance than decision tree and naïve Bayes. He achieved 85% accuracy with random forest considering marital status, capital gain, education, age and hours per week as the top features. In preprocessing author converted categorial variables into dummy variables, and merged capital gain and capital loss as one column. The author explores that all numerical attributes have a positive correlation with income. The groups with higher income are male, married people, self, or government employees, Whites and Asians, professionals, specialists, technology workers, and managers. The model has a good accuracy on low income but is weak when it comes to predicting high income [2].

In this paper [3], the author presented the effect of dimensionality reduction using; Principal Component analysis for dimensionality reduction on the performance of support vector machine on Adult income data set. The author also emphasized on the consequences in terms of accuracy and computation time, by reducing the input data vertically with regards to the number of training examples and horizontally in terms of the number of features and to find out the ideal ratio between them. They build four data sets: adult_13 dataset was built by removing education number, based on the strong correlation 0.8881 between education and education number. Next, adult_10 was built by removing four variables age, education number, marital status and sex. Third, adult_6 they keep components with eigenvalues greater than 1, and the last dataset adult_8, contains 13 principal components generated by the PCA method. As SVM method deals better with real numbers, they used the scaling method instead of discretization for continuous variables. They reported that the accuracy of adult_8, is the highest among all the datasets. This shows that PCA components capture maximum information and improved the performance of the Support Vector Machine.

In this study [5], the author compares the performance of different classification algorithms based on accuracy, ROC area and precision. They used eight supervised machine learning algorithms: SVMs, neural nets, decision trees, random forest, k-nearest neighbor, bagged trees, boosted trees and boosted stumps on different data sets including Adult income dataset and conducted empirical comparison. The Adult income data set is the only problem that has nominal attributes. The study concludes that bagged trees, random forest and neural nets have the best average performance over all the metrics and dataset. In terms of KNN, it gives phenomenal results, when attributes are weighted by gain ratio instead of unweighted Euclidean distance. Furthermore, for the selection of the best model they used the IK validation test, and then checked the performance on the test set.

After conducting a series of literature reviews, I have decided to use logistic regression, random forest and naïve bayes for this project. It is further supported by the fact that the project of predicting income from census data is a binary classification problem, as the target variable having binary output and data has mixed numerical and nominal attributes.

These are the benchmark algorithms, which are used for dealing with binary classification problems.

Logistic regression is also helpful as its gives the idea of important variables, which will be useful for predicting income, based on p values. Random forest gives better performance as compared to decision trees, in most of the studies. Feature selection will be done based on the correlation between attributes, exploratory analysis, literature review and in the model building step with forward/backward feature elimination techniques. Principle Component analysis (PCA) as used in [4] to reduce the dimensionality will not be used at this stage due to the scope and time limitation. The data used in this project will be the adult income census data extracted from complete census data, as used by most of the researchers.

Data Description

U.S Adult Census Income Data

The dataset used for the analysis is an extraction from the 1994 census data by Barry Becker and downloaded from UCI Machine Learning repository [6].

The data is comprised of 14 attributes and 32561 observations. 8 are categorical and 6 are numerical. Income is the dependence variable/class variable and rest are independent variables. The data attributes, their type, categories and description are explained as follow:

1. **Attribute Name** : Age
   Data Type : integer
   Description : Represents the age of an individual.

2. **Attribute Name** : Workclass
   Data Type : Categorical data (9 Distinct categories)
   Description : It describes the work class of an individual.
   Distinct Categories : 9 Distinct categories are as follow:
   Federal-gov, Local-gov, Never-worked, Private, Self-em, p-inc, Self-emp-not-inc, State-gov, Without-pay

3. **Attribute Name** : Fnlwgt
   Data Type : integer
   Description : Represents final weight. It is the number of units in the target population that the responding unit represents.

4. **Attribute Name** : Education

   Data Type : Categorical (16 Distinct categories)

   Distinct Categories : 16 Distinct categories are as follow:
   1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college.

   Description : It describes the education levels and consist of 16 different categories.

5. **Attribute Name** : Education number

   Data Type : integer

   Description : It specifies number of year of educations in total.

6. **Attribute Name** : Marital status

   Data Type : Categorical (7 Distinct categories)

   Distinct Categories : 7 Distinct categories are as follow:
   Divorced, Married-AF-spouse(AF – Armed forces), Married-civ-spouse(civ – civilian), Married-spouse-absent, Never-married, Separated, Widowed

   Description : Describes the marital status of an individual.

7. **Attribute Name** : Occupation

   Data Type : Categorical (15 Distinct categories)

   Distinct Categories : 15 Distinct categories are as follow:
   ?, Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving.

   Description : Describes the occupation of an individual.

8. **Attribute Name** : Relationship

   Data Type : Categorical with 5 distinct categories.

   Distinct Categories : 5 Distinct categories are as follow:
   Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife.

   Description : It represents the role of an individual in a family.

9. **Attribute Name :** race
Data Type : Categorical with 5 distinct levels.
Distinct Categories: Distinct categories are as follow:
White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

10. **Attribute Name :** Sex
Data Type : factor 2 levels Female, Male

11. **Attribute Name :** Capital gain
Data Type : integer
Description : It represents investment gain during a year.

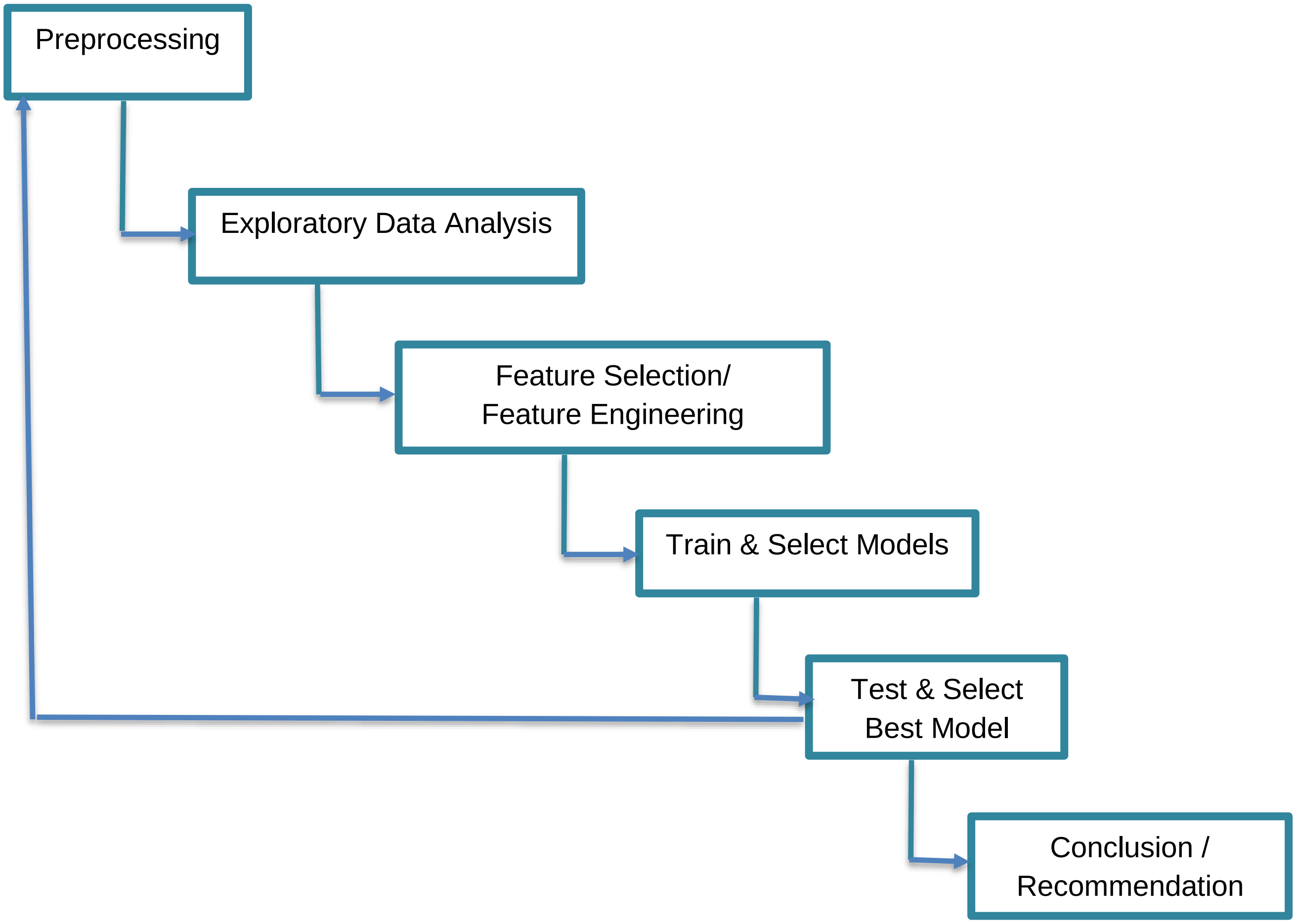
12. **Attribute Name :** Capital loss
Data Type : integer
Description : It represents investment losses during a year.

13. **Attribute Name :** Hours.per.week
Data Type : integer
Description : It represents the number of hours worked in a week.

14. **Attribute Name :** Native.country
Data Type : factor (42 levels) Cambodia
Description : It represents the native country of an individual.
Categories/Levels : Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad & Tobago, United-States, Vietnam, Yugoslavia.

15. **Attribute Name :** Income
Data Type : factor (2 levels)
Categories/Levels : <= 50K, > 50 K
Description : Class Label

Approach


Preprocessing
Exploratory Data Analysis
Feature Selection/
Feature Engineering
Train & Select Models
Test & Select
Best Model
Conclusion /
Recommendation

Step 1: Preprocessing

In this step, data is loaded. Following steps are conducted in this stage.

1.1 Divide the data into Training and Testing

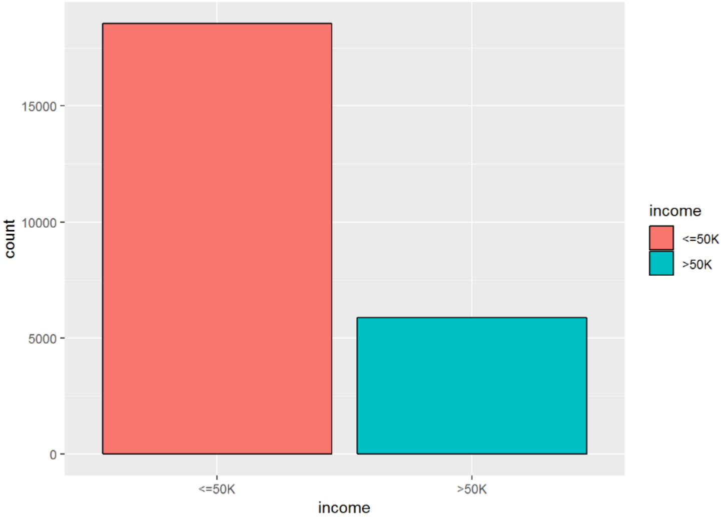
The data is divided into training and testing data in the ratio of 75% and 25%. Total data consist of 32561 rows and 15 attributes. Train data consist of 24421 rows and test data consist of 8140 rows.

1.2 Null Values

The data has missing values in three attributes, as stated; work class, occupation and native country (90% values correspond to the United States), which are categorical and represent a very small fraction of whole data. The missing value are omitted from both training set (3173) and testing set (1089).

1.3 Income Attribute

Income is a target attribute, with values > 50K and <= 50K. The distribution of income attribute is shown below.



Income <= 50 K is 75.92% and income > 50 K is equal to 24.08.

The data consist of 24421 observations with total of 15 attributes, in which 6 are numerical. Their statistical information's minimum, maximum value, mean, median and standard deviation was calculated and shown in Table below.

1.4 Data Statistics of Numerical Variables:

Statistics of the numerical variables are as follow:

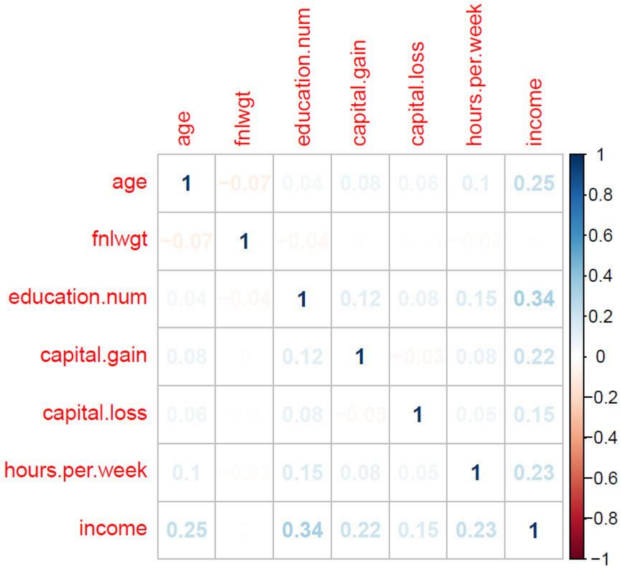
No.	Attribute Name	Minimum	Ist Qu.	Median	Mean	3rd Qu.	Max.	Sd
1.	Age	17.00	28.00	37.00	38.58	48.00	90.00	13.64
2.	Education	1.00	9.00	10.00	10.08	12.00	16.00	2.57
3.	Capital Gain	0	0	0	1078	0	999	7385.29
4.	Capital Loss	0.0	0.0	0.0	87.3	0.0	4356.0	402.96
5.	Hours per week	1.00	40.00	40.00	40.44	45.00	99.00	12.35

Step 2: Exploratory Data Analysis

This stage involves a detailed analysis of attributes and their relationships with each other, with target variable income. The following steps are conducted in this stage.

2.1 Correlation between Numerical Variables

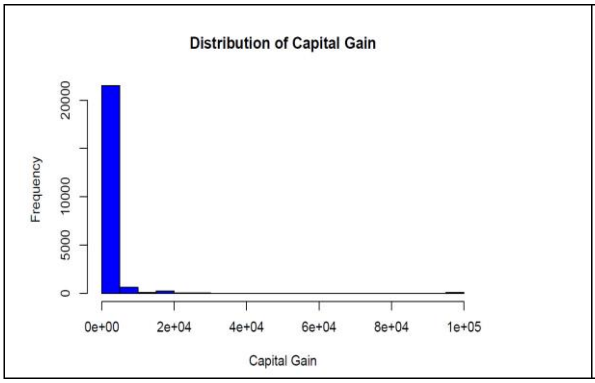
The dataset consists of five numerical attributes. i.e. age, fnlwgt, education number, capital gain, capital loss and hours per week. The correlation matrix is plotted below:

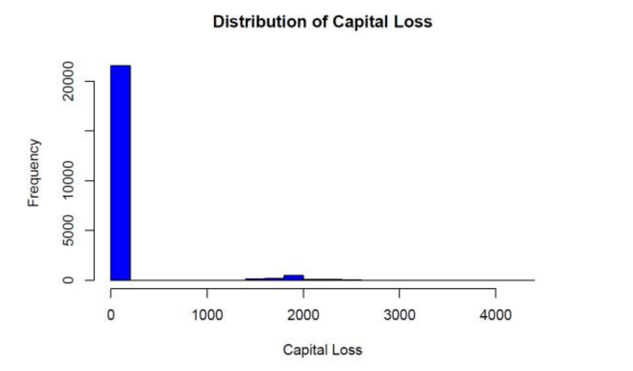


Correlations shows that numerical attributes are positively related but are not strongly correlated. Education has the highest correlation 0.33 with income, then capital gain (0.22), then age(0.24) and then hours per week (0.23). The variable fnlwgt has a value of -0.07, and it represent Represents final weight. It is the number of units in the target population that the responding unit represents. We will remove this variable.

2.2 Capital Gain and Capital Loss

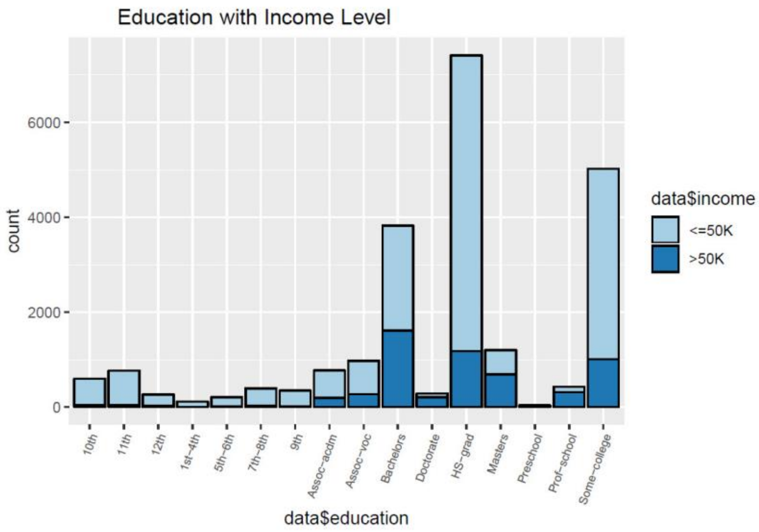
Majority of the values in capital gain and capital loss are zero and there are highly skewed. We will exclude both the attributes.





2.3 Correlation between Education, Education number and Income Level

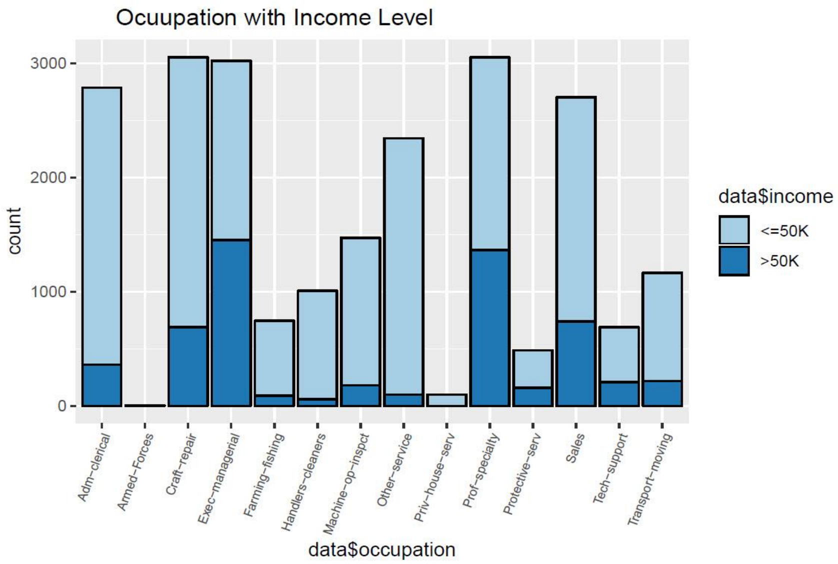
Education and education number are linearly correlated so we will consider only education. The plot between education level and income level is as follow:



The plot shows that the maximum number of adults earning income greater than 50K have bachelor's degree. In doctorate and masters also, the largest proportion is earning greater than 50 K. In lower education levels the largest proportion have income less than 50K. We can conclude that higher education results in higher income. Education attribute is important for predicting income of an individual.

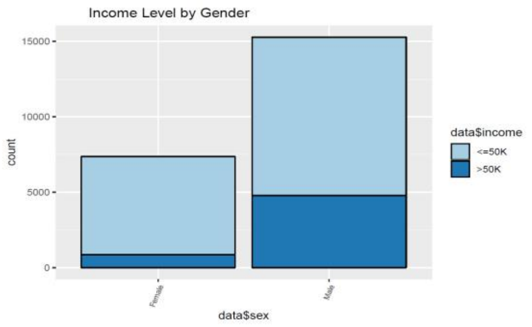
2.4 Occupation and Income Level

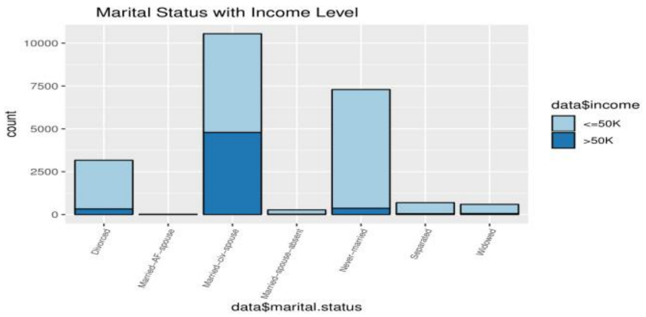
The plot between income of people in different education levels shows that people in managerial jobs and professors are earning more than 50 K in the highest ratio. It shows that occupation also effects at the level of income of an individual.



2.5 Gender, Marital Status and Income Level

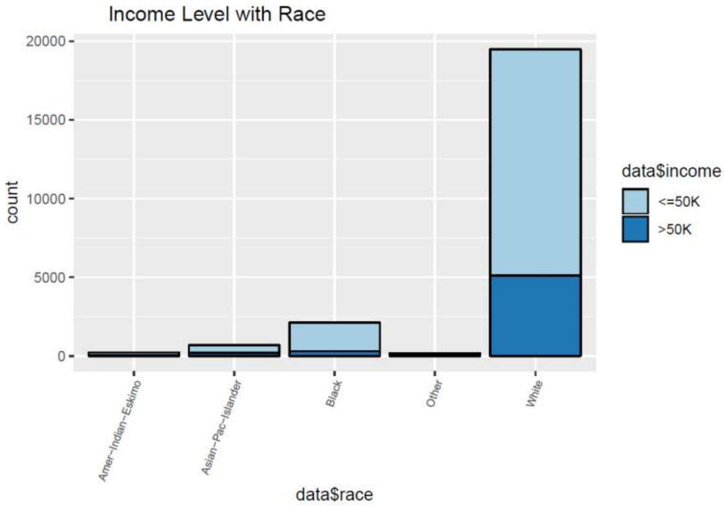
The plots show that Male and married people are earning more than 50K, as compared to female and unmarried people. These attributes are also important for predicting the income level of an individual.





2.6 Race and Income Level

The plot shows that majority of the individual belongs to white race and then black.



Step 3: Feature Selection/ Feature Engineering

3.1. Reducing and Combining Native Country

The native country attribute consists of 42 levels, with 90% data corresponds to united states. So, the levels are reduced into following 5 levels.

i. asia <- c("Cambodia", "China", "Hong", "India", "Iran", "Japan", "Laos","Philippines", "Taiwan", "Thailand", "Vietnam")

ii. northAmerica <- c("Canada", "Cuba", "Dominican-Republic", "El-Salvador", "Guatemala","Haiti", "Honduras", "Jamaica", "Mexico", "Nicaragua", "Outlying-US(Guam-USVI-etc)", "Puerto-Rico", "Trinadad&Tobago","United-States")

iii. southAmerica <- c("Columbia", "Ecuador", "Peru")

iv. europe <- c("England", "France", "Germany", "Greece", "Holand-Netherlands", "Hungary", "Ireland", "Italy", "Poland", "Portugal", "Scotland", "Yugoslavia")

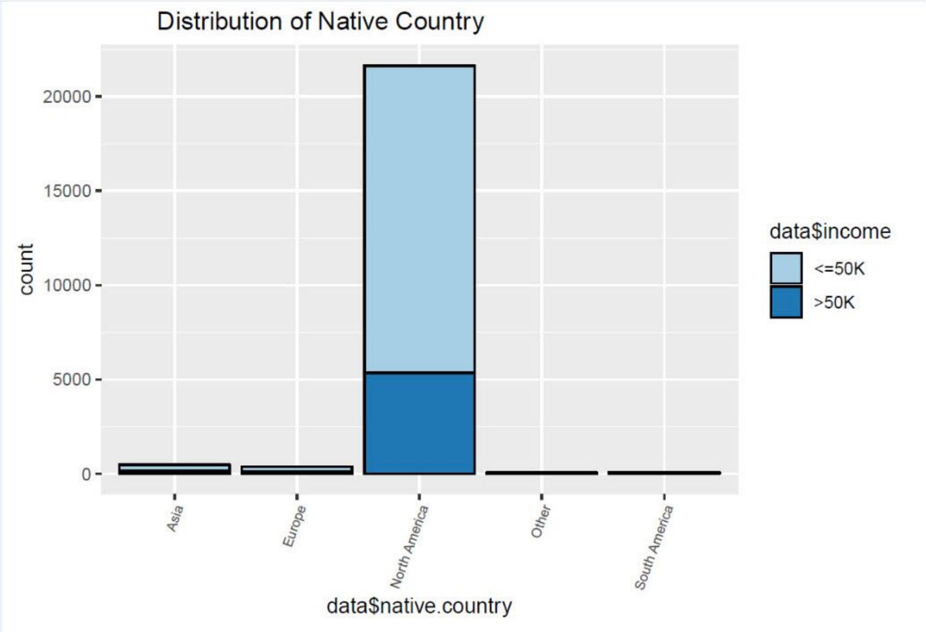
v. other <- c("South")

```
data$native.country[data$native.country %in% northAmerica] <- "North America"
data$native.country[data$native.country %in% asia] <- "Asia"
data$native.country[data$native.country %in% southAmerica] <- "South America"
```

```
data$native.country[data$native.country %in% europe] <- "Europe"
data$native.country[data$native.country %in% other] <- "Other"
```

so, we reduced and combined native countries as follow:

Distribution of population with reference to the income and native country is shown below:

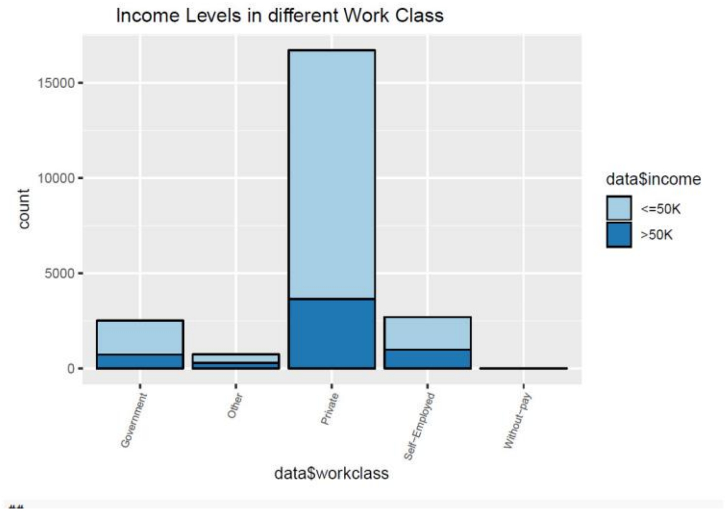


Native country of most of the population is north America, so we will remove native country as there’s not much variation in this attribute.

3.2. Reducing and Combining Work Class

Levels of the work class attribute are also reduced into four levels Government, self-employed, other and unknown as below:

```
data$workclass <- gsub('^Federal-gov', 'Government', data$workclass)
data$workclass <- gsub('^Local-gov', 'Government', data$workclass)
data$workclass <- gsub('^State-gov', 'Government', data$workclass)
data$workclass <- gsub('^Self-emp-inc', 'Self-Employed', data$workclass)
data$workclass <- gsub('^Self-emp-not-inc', 'Self-Employed', data$workclass)
data$workclass <- gsub('^Other', 'Other', data$workclass)
data$workclass <- gsub('^Unknown', 'Other', data$workclass)
```



The plot shows that most of the people earning more than 50K are in private sector, then self-employment and then Government. Thus, work class attribute also seems to contribute to predicting the income level.

https://github.com/HumairaAsim/CKME136/blob/master/EDA.Rmd

Step 4: Model Fit

We selected eight features age, workclass, education, marital status, occupation, race, sex and hours.per.week to predict income. We used three classification algorithms regression, decision tree and random forest first to train the model and later for prediction.

LOGISTIC REGRESION

```
glm(formula = income ~ age + workclass + education + marital.status + occupation + race +
sex + hours.per.week, family = binomial("logit"), data = traindata)
```

Results :

```
pred     <=50K   >50K
<=50K     5247    846
>50K       453   1029
```

DECISION TREE

```
Dtree<- rpart(income~ age+ workclass+ education+marital.status+ occupation+ sex
+hours.per.week, data = traindata, method='class',cp =1e-3)
rforest.pred.prob <- predict(rforest, newdata = testdata1, type = 'prob')
rforest.pred <- predict(rforest, newdata = testdata1, type = 'class')
```

Results:
Confusion Matrix

```
Prediction <=50K >50K
<=50K 5263 437
>50K 833 1042

Accuracy : 0.8323
Sensitivity : 0.8634
Specificity : 0.7045
Pos Pred Value : 0.9233
Neg Pred Value : 0.5557
```

RANDOM FOREST

```
rforest <- randomForest(income ~ age+ workclass+ education+marital.status+occupation+
sex+hours.per.week, data = traindata, ntree = 500)
```

Results:

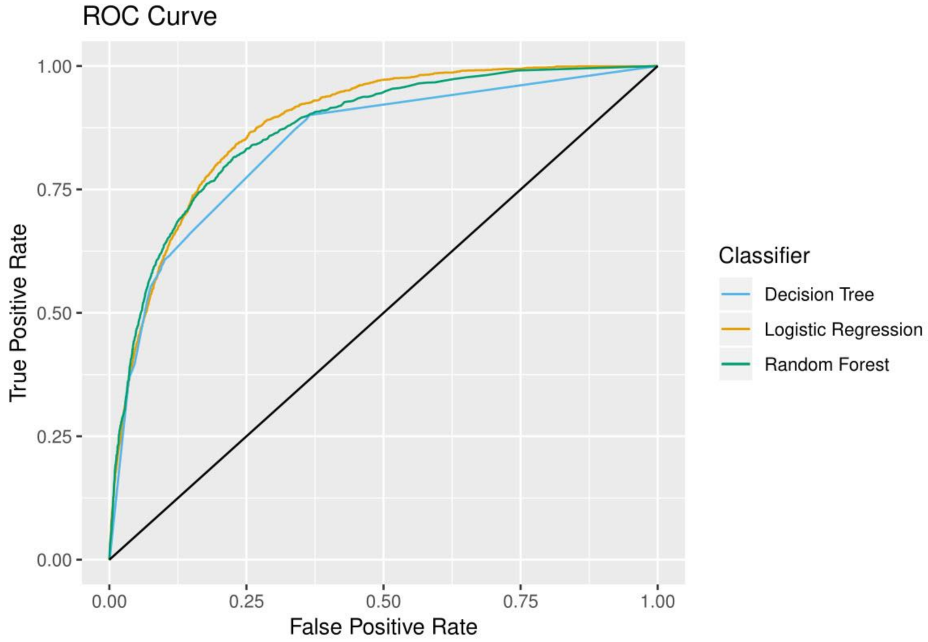
```
Prediction
              <=50K   >50K
<=50K        5220    480
 >50K          755   1120

Accuracy : 0.837
Sensitivity : 0.8736
Specificity : 0.7000
Pos Pred Value : 0.9158
Neg Pred Value : 0.5973
```

Step 5: Prediction/Conclusion

The final model is selected based on the Area under curve, which is plotted against true positive rate and true negative rate

	Area Under Curve
Logistic Regression	0.8815
Decision Tree	0.8421
Random Forest	0.8712



Logistic regression has the highest area under curve (AUC) value, then random forest and lowest is with the decision tree. So, we selected logistic regression as our final model for predicting income of an individual as it gives the largest area under the curve.

Github Link:

https://github.com/HumairaAsim/CKME136/blob/master/modelFit.rmd

Bibliography

[1] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996.

[2] S. M. Bekena, "Using decision tree classifier to predict income levels," MPRA Archive, 2017.

[3] A. Lazar, "Income prediction via support vector machine," in *2004 International Conference on Machine Learning and Applications, 2004. Proceedings.*, Louisville, Kentucky, USA, 2004.

[4] N. K. A. Y. Rich Caruana, "An Empirical Evaluation of Supervised Learning for ROC Area," in *Proceedings of the 25th international conference on Machine learning*, Helsinki, Finland, 2008.

[5] B. Becker, "UCI Machine Learning Repository," [Online]. Available: http://mlr.cs.umass.edu/ml/datasets/Census+Income. [Accessed 2018].