# **Predictive Modelling on Adult Dataset**

**Xinqi Gao, Abhisek Gupta, Xueyu Chen, Matthew A. Lanham**

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

gupta362@purdue.edu; gao131@purdue.edu; chen950@purdue.edu; lanhamm@purdue.edu

## **Abstract**

This paper attempts to build a binary classification model using decision tree, Linear Discriminant Analysis and Logistic regression. There were two primary reasons why we picked this dataset for the project. Adult dataset on UCI machine learning website is a binary classification problem where the target is to predict the whether the income of an individual will be greater than 50,000 or not, based on various socio economic indicators. First, our goal was to test the performance of various classification methods on the dataset post null value imputation. Second, given the debate on income inequality in our times we wanted to understand the role of various socio-economic factors on the level of income an individual command in different countries. We also felt that there could be implications in fraud identification justice department by matching the disclosed income by individuals against their predicted income, hence identifying if they are more likely to commit fraud. Some interesting insights that came out of the analysis were that Relationship, education, and capital gain/ loss are the three most important features which affect the Income class of a person. We also saw that Logistic Regression performed better than CART and Linear Discriminant analysis for our dataset.

## **Introduction**

The dataset which we will use for the project is Adult dataset provided by the UCI machine learning repository. Few important reasons why team identified this dataset are:

· Business Problem: Predicting if the income of a person is above aforementioned criteria by using socio-demographic variables as education, ethnicity, employment type etc. Team will also attempt to establish inferential relationship between the income group and demographic features

· Dataset Examination: Dataset is relatively clean and but does have null values in country, occupation and workforce. This gives us opportunity to test various classification techniques and also use some methodology to impute null values

· Availability of Performance Benchmarks: This will help team test recently developed modelling techniques or the dataset and improve upon the already tested methodologies

· Identification of Drivers: Identification of critical socio economic drivers in presence of each other and quantifying their impact on income group can help governing bodies design better policies delivering maximum impact. For example, the multiple logistic regression model in this case will assess the model while considering the impact of race on income in presence of age

· Lastly, this dataset gives us good opportunity to apply all important concepts taught in the class

While much extant economic research has studied the driving factors for individual income, one of our research focus is to classify individual income into brackets with different dollar values, and develop effective model to assign individuals into the proper brackets. Our study could contribute to helping the U.S. Internal Revenue Service (IRS) conduct preliminary examination on individual tax return and flag the individuals for further investigation. For instance, if an individual’s income falls into the bracket of $190,150 to $413,350 according to his or her actual tax filing, but the income bracket predictive model suggests that the individual should have income that fall into the bracket of greater than $415,050, the IRS could flag the individual as high-risk of potential fraud and conduct investigations later.

## **Literature Review**

[Paper 1:](http://www.cluteinstitute.com/ojs/index.php/IJMIS/article/viewFile/841/825)

One challenge we had at the beginning stage of the study is the missing values issue. Since our study was conducted majorly using R-Studio and SAS Enterprise Miner, we looked into literature about data cleansing techniques within both R Studio and SAS Enterprise Miner. According to the Ghoson’s 2010 study, SAS Enterprise Miner provides a couple of ways to address the missing values issue. SAS EM could exclude all subjects that contain missing values, or it could replace the missing values “using one of the imputation methods which are Seed of Nearest Cluster, Mean of Nearest Cluster, and Conditional Mean”. By applying the Filter Outlier node, SAS EM could also “remove categorical valued variable situations that do not happen” (Ghoson, 2010).

[Paper 2:](https://www.researchgate.net/publication/259519014_Algorithm_to_determine_e-distance_parameter_in_density_based_clustering)

The author Sunita Jahirabadkar and Parag Kulkarni in the journal Algorithm to determine e-distance parameter in density based clustering shown their research on the problem that DBSCAN cannot be used in various subspace clustering of high dimensional data. They did experiment on both UCI adult dataset with 14-dimensions where data contains 3 different densities, and Chameleon dataset. Their approach is check the clustering quality of each dimension to understand the data distribution based on the quality criterion of knn notion. For the adult dataset they checked the quality of the clusters and corresponding knn values. Through the experiment they observed that there exist four outliers in knn distances and observed the break points, and they successfully identify the cored e-value for each density cluster. Through this understanding using knn distances, the DBSCAN algorithm can be used in high dimensional data such as the UCI adult dataset and improved the scalability of the accuracy. Their future research direction is to strengthen the density based clustering and try to execute clustering technique on the large distributed dataset. Also try clustering algorithms in stream data, spatial data, text data and increase the scope of research.

[Paper 3:](http://aje.oxfordjournals.org/content/172/9/1070.long)

Missing data is inevitable to face during data collection stage. Multiple imputation always been used for handling missing data in large epidemiologic studies. Implementing multiple imputation by chained equations(MICE) is popular approach but with the deficiency of produce inconsistent imputation models and manually specify conditional models for all variables. Author Lane a Jerome present a MICE approach with classification and regression trees(CART) as conditional models approach to solve the current challenge of default MICE. The CART imputation requires manually minimal tuning needs and gives higher inference reliability than the naïve application of MICE. The author pointed out the problem of CART, where larger tree prone to over fit the data. However, they pruned tree weakly in order to minimize bias and preserve the complexity through the large imputation model. Through simulating data, they assess and compared two algorithms. They found out that CART-based MICE give lower mean-squared errors and biases and improved the prediction accuracy. They also found out that because of the imputation models are imperfect, both methods have intervals not covering truth. They also evaluate the algorithm in the data on adverse birth and examined the posterior predictive checks. In the future, they are look forward the nonparametric methods can be used and improved.

[Paper 4:](https://pdfs.semanticscholar.org/6db3/08779954a4006d87ce68148ff2aaa176dfb2.pdf)

When applying machine learning to extract true knowledge from data, the quality of data is vital. To solve the problem, there are three treatments are used for missing data. One is ignoring and discarding data when missing data are MCAR; second approach is through parameter estimation with the maximum likelihood algorithm third method is using Imputation. With the presence of missing data, the author Gustavo and Maria took approach through data imputation, especially compared three imputation methods performance: K-nearest neighbor, C4.5 and CN2 to treat missing data. The UCI dataset Bupa, Cmc and Pima are chosen for the experiment, where the dataset has no missing value. To have full control of the dataset, they only insert missing data into the training set completely at random (MCAR) and left the test set as the non-missing data. They used the 10-fold cross validation method and compared the performance of C4.5 and CN2 with 10-KNNI. The method 10-KNNI has the superior performance over C4.5 and CN2 for Bupa, Cmc and Pima. Hence the 10-NNI method did outperformed the other two methods. The researchers look forward to analyze the missing data are not in MCAR type and use more complex analysis to analyze the missing data.

[Paper 5:](https://www.cs.cornell.edu/~shmat/shmat_kdd08.pdf)

Microdata always contains private information about individuals which may be sensitive. The k-anonymity has limitation of the sensitive attributes that revealed to public and the limitation of been applied on high dimensional data, however, privacy has been researched and k-anonymity is the most popular method. How to reduce the privacy disclosure in public datasets and have gain in utility is the question raised from the researchers Justin and Vitaly. They compared the methodology of using k-anonymity and l-diversity with the trivial sanitization method and build a methodology for evaluating the tradeoff of the privacy versus gain of utility. The researchers use standard model and only the quasi-identifier t[Q] is applied generalization and suppression approach to keep “truth” of database. The author did experiment on the UCI adult dataset and contradict other researcher works. They observed the sanitization does not provide better compared to the trivial sanitization. The K-anonymity keep all the sensitive features but less privacy. The experimental result shown to keep the privacy and utility in one database is difficult and trivial sanitization can provide better privacy with equal utility with respect the other two methods. The future works is to design sanitization methodology with better privacy and utility.

[Paper: 6](https://uta-ir.tdl.org/uta-ir/bitstream/handle/10106/25475/MAO-THESIS-2015.pdf?sequence=1)

Exploratory data analysis (EDA) refers to an iterative process through which analysts constantly ‘ask questions’ and extract knowledge from data. EDA is important exercise as it helps identify trends and relationships which can then be tested in a more rigorous manner using statistical tests. So in function, EDA is becoming important for modern data analysis, such as business analytics and business intelligence, given the need to explore the data to identify hypothesis which can tested through confirmatory data analysis, and involves analysts directly in the data mining process. EDA can be done through frequency tables, correlation matrix or visualization plots given the type of data and nature of relationship in enquiry. For example, correlation matrix is a good way to study relationship between two numerical variables. Mosaic plots and frequency matrix are quite useful while studying the relationship between two categorical variables. Similarly, relationship between a categorical and numerical variable can be analyses using boxplots.

[Paper: 7](https://arxiv.org/abs/1610.09075)

Missing data always add constraint social science research and modeling accuracy. There are different missing data imputation and mechanisms have been explored. In order to have higher prediction accuracy, researchers first understand the patterns of the missing data and then doing the imputation analysis. Most data is assumed as MAR where the missing data are dependent on the given observation. And Sliva-Ramirez did research comparison and found the ANN is the best imputation model for categorical data. The authors explored two datasets from UCI with one dataset is UCI Adult dataset. They first identify the pattern of Adult missing data are not MCAR and might be MAR since the” workclass” and “occupation” features are missing at the same time. They compared and trained three classifiers on the dataset with method ANN, decision trees and random forest and evaluate the performance by comparing the test error rate at different level of MCAR data. The researcher found out that the ANN classifier with knn data imputation beats the other classifiers with the lowest error rate and give the error with the best test error rate on the adult dataset. In the future, the regularization technique can be added to improve the accuracy of model.

**Summary:**

All the papers based on adult dataset do remove the null values from their analysis. In our attempt, we have tried to address this issue by doing multiple imputations using chain equations. We expect that this to improve the accuracy of our model over the models where null values are removed from analysis. One thing which is unknown and we would like to explore is performance of various imputation methods for multi class categorical variables for few datasets. Another things that’s missing from the research papers is commentary on rank deficient models encountered due to skewed ratio of number of observations to predictors. This can be an area to explore and methods to overcome this problem.

**Data**

· Data Set Description: Predict whether income exceeds $50K/year based on census data.

· Data Set Characteristics: Multivariate

· Attribute Characteristics: Categorical, Integer

· Associated Tasks: Classification

· # of Attributes: 14

· # of Instances: 32,652

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Age | Num | Age |
| Workclass | Char | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| Fnlwgt | Num | Final weight. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. |
| Education | Char | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. |
| Edu\_noy | Num | Number of years of education |
| marital-status | Char | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. |
| Occupation | Char | 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. |
| Relationship | Char | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| Race | Char | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. |
| Sex | Char | Female, Male. |
| Capital-gain | Num | Capital gain |
| Capital-loss | Num | Capital loss |
| Hours-per-week | Num | Hours per week |
| Native-country | Char | 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. |

**Methodology**

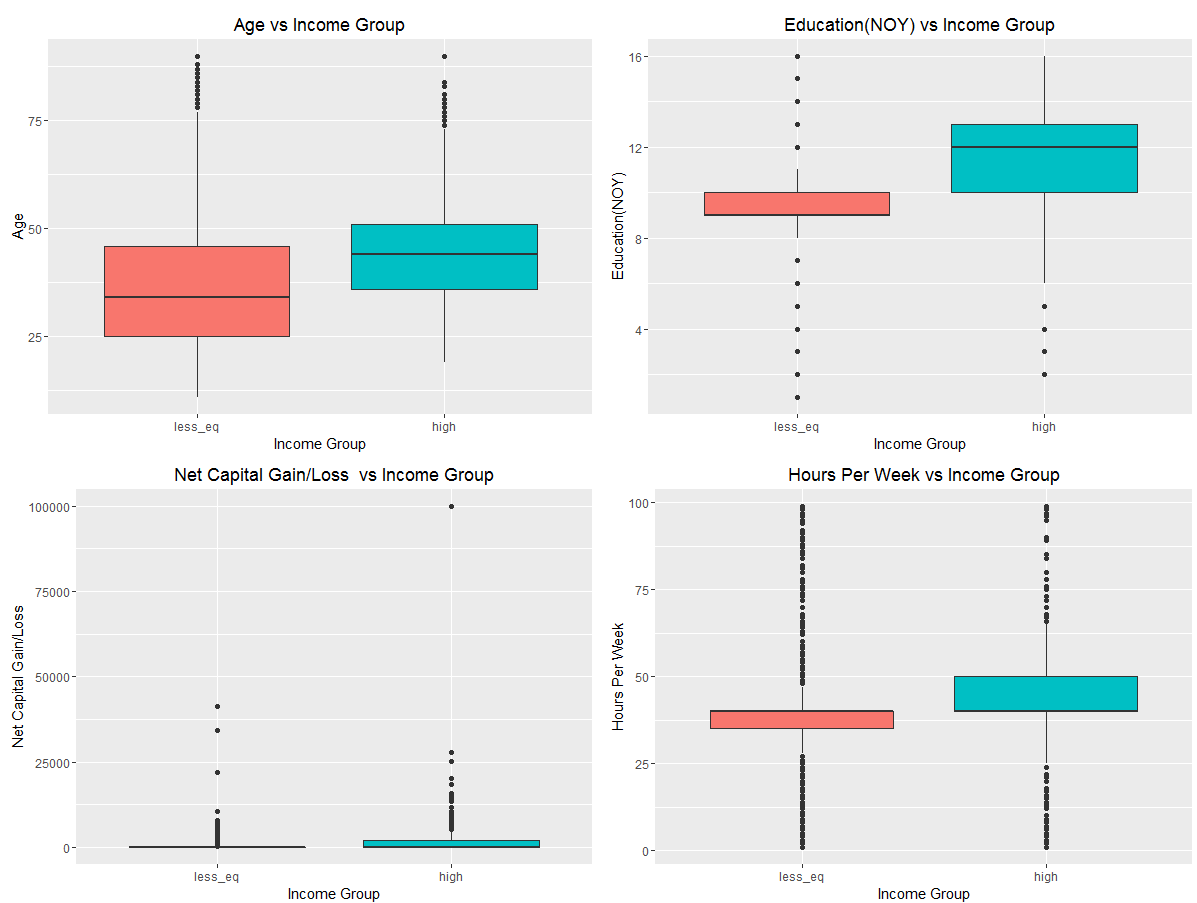
**SAS Enterprise Miner**:

Before conducting data analysis, our team examined and cleaned the dataset through two steps: (1) initial selection of variables, and (2) solution development for the missing values issue. For the first step, we started with identifying pairs of variables that are high correlated, and making a keep/ drop decision on those variables. For example, we decided to keep the variable “Education” and drop “Education\_Number of Years”, because we thought “Education” provides the highest degree received by each individual, which could be a valuable piece of information that the variable “Education\_Number of Years” might not necessarily provide. Another example is “Capital Gain” and “Capital Loss”. Because both are net values, which means no individuals would report a non-zero “Capital Gain” and a non-zero “Capital Loss” at the same time; In another word, if an individual has a non-zero “Capital Gain”, then his or her “Capital Loss” has to be zero. Therefore, our team combined the two variables into one variable “Capital Gain (Loss)”, which would show as positive (negative) if an individual has a capital gain (loss). The variables that we decided to keep from the initial selection step were: Age, Work class, CPS Final Weight, Education, Marital Status, Occupation, Relationship, Race, Capital Gain (Loss), Hours Per Week, Country, and Income Flag (target variable, whether annual income is greater than $50,000).

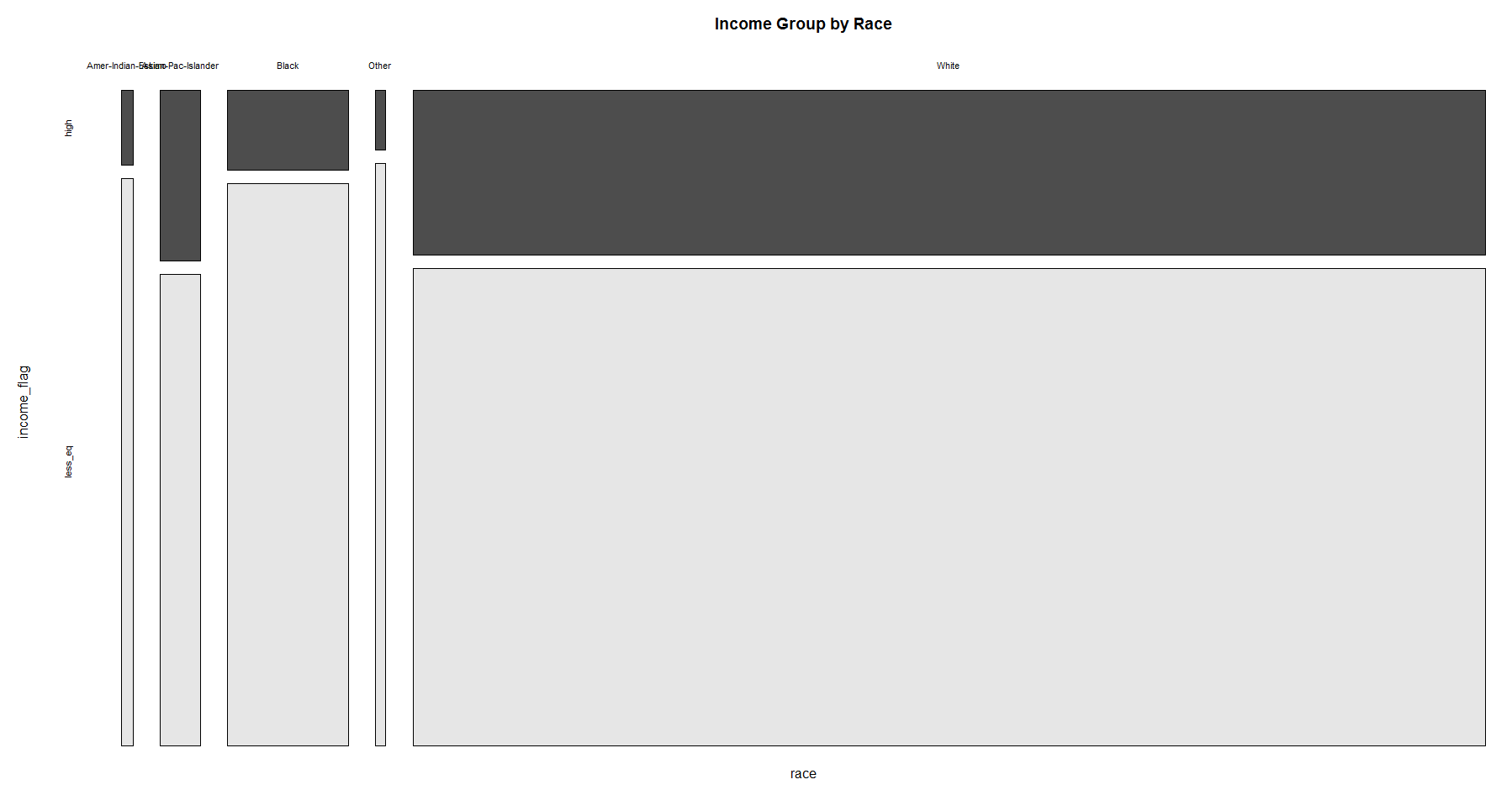
**R-Studio:**

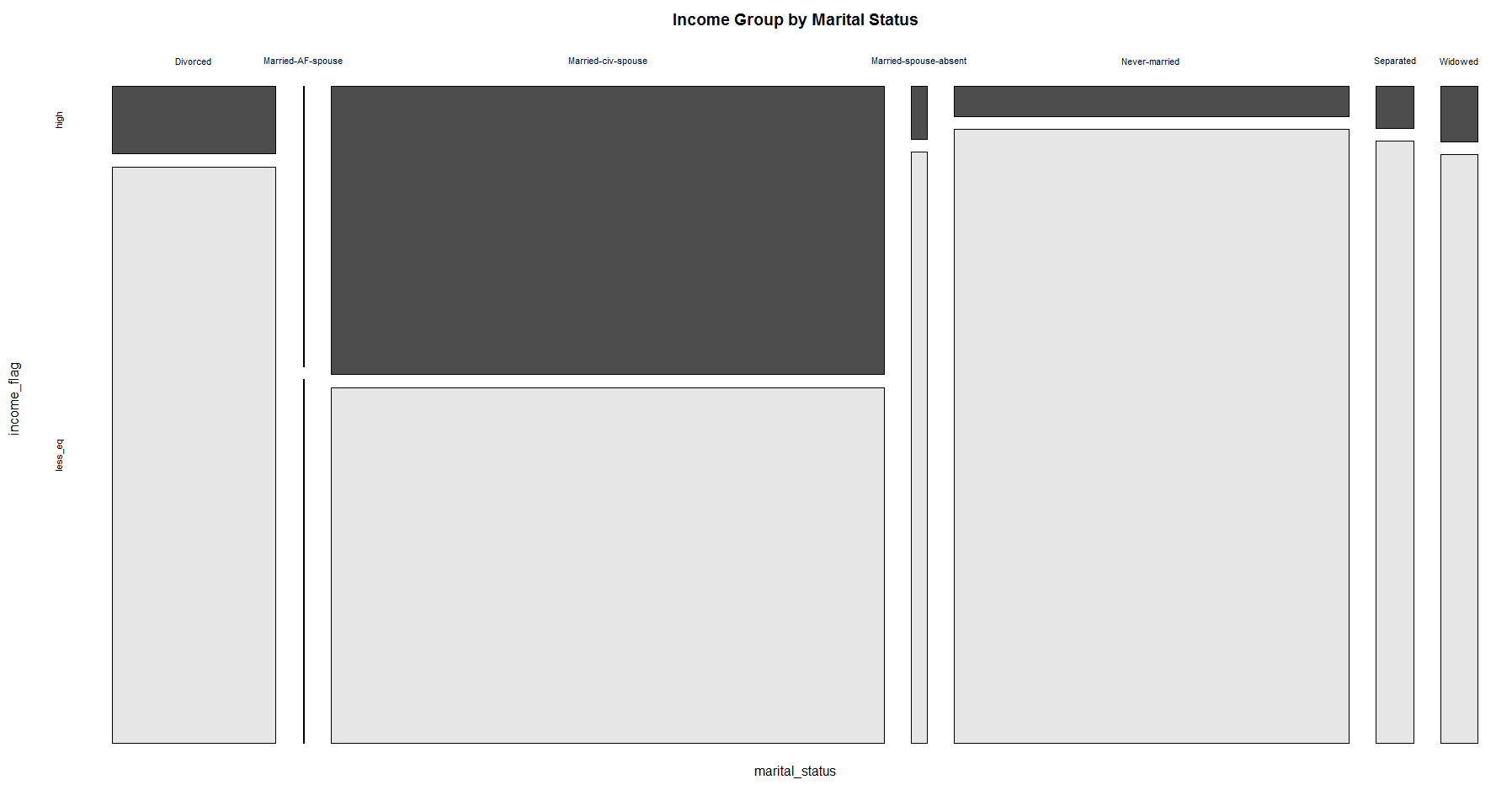
While encountering missing values, we tried to solve the problem with assigning the label to the missing data, instead of simply deleting the null data. One way to deal with them would be to assign them mean or mode of the column depending on whether the variable is categorical or numerical. Other way which can be done is to use some predictive technique to predict the missing values in the dataset for various columns. Through exploring the data, we identified that the features “work class”,” occupation”, and “country” has different levels of missing data, where “country” variable has 41 classes. First we used method 1 which is to use modal imputation for dealing with missing values. We did this by finding the most common value to identify and compare the proportion of imputed category with respect to the original category distribution. However, given that this method is not very robust, we used then focused on implementing multiple imputation by chained equations(MICE) to handle the missing data. We have a data matrix Y with Y=(Yp, Yc) with Yp is where Y is partially observed, and Yc is the columns that are completely observed. We need to specify a set of conditional distribution p(Yi|Y-i) where Yi is the ith column of the Yp, and Y-i is the matrix Y with ith column removed. (Burgette LF, Reiter JP. (2010).) To implement MICE on our Adult dataset, we use MICE package in R. Through multiple imputation, the three main step: imputation, analysis and pooling are all packaged in the mice package in R. Since our features have more than 2 classes and are unordered categorical type, the method polytomous regression imputation and linear discriminant analysis were used to achieve the same. With the imputed data, we expect the model building will perform better with higher accuracy.

Post that we did some exploratory data analysis to identify trend in income buckets with numerical variables.



Some mosaic plots to study categorical variables with income flags:





Three methodologies we choose to analyze the data are:

1. Logistic Regression: Given that our target variable had two classes, we choose to start with Logistic Regression as our first modelling technique. In Logistic (or logit) regression, for any given X, the logit model provides the value for the observation that can be used with the logistic cumulative density function to find the probability that Y = 1 for that observation. This method uses maximum likelihood estimation to minimize the variance in predictions and estimate model parameters.
2. Linear Discriminant Analysis: LDA is a classification modelling technique, where two or more groups are known a priori and one or more new observations are classified into one of the known populations based on the measured characteristics. In this alternative approach, we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes’ theorem to flip these around into estimates for Pr(Y = k|X = x).
3. Classification and Regression Trees: Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Usually one of misclassification rate, Gini Index and Entropy is used to discriminate and do recursive portioning.

The reasons behind choosing these three methods were:

* There are times when both Logit and LDA not perform well at the same time. Some situations like this are:
  + When classes are far apart Logit tends not to be very stable
  + When the function does not have normal distribution and all classes don’t have common variance LDA tends to underperform Logit
  + LDA is quite sensitive to outliers while Logit is slightly more stable when in presence of outliers

We wanted to see if any of these problems do exist with the dataset. And if the test error for both logit and lda would have been different, we would easily identify the underlying issues with distributions in data which can then point us to correct methodology. Hence we tried using both of them

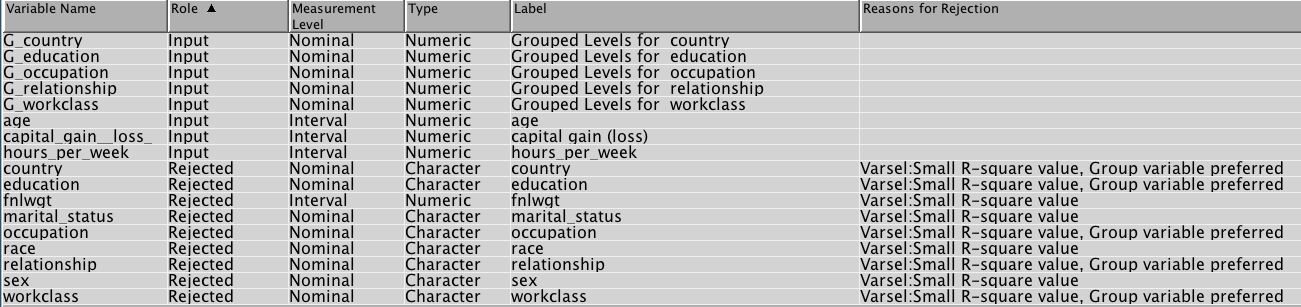
* CART gives a very interpretable tree which can help in identifying most important variables at each node. This helps in understanding relationships of target variable with various explanatory variables. The tree model sometime cannot be a very good model when thinking about accuracy if the decision boundary between classes is linear but it still does make the results more interpretable

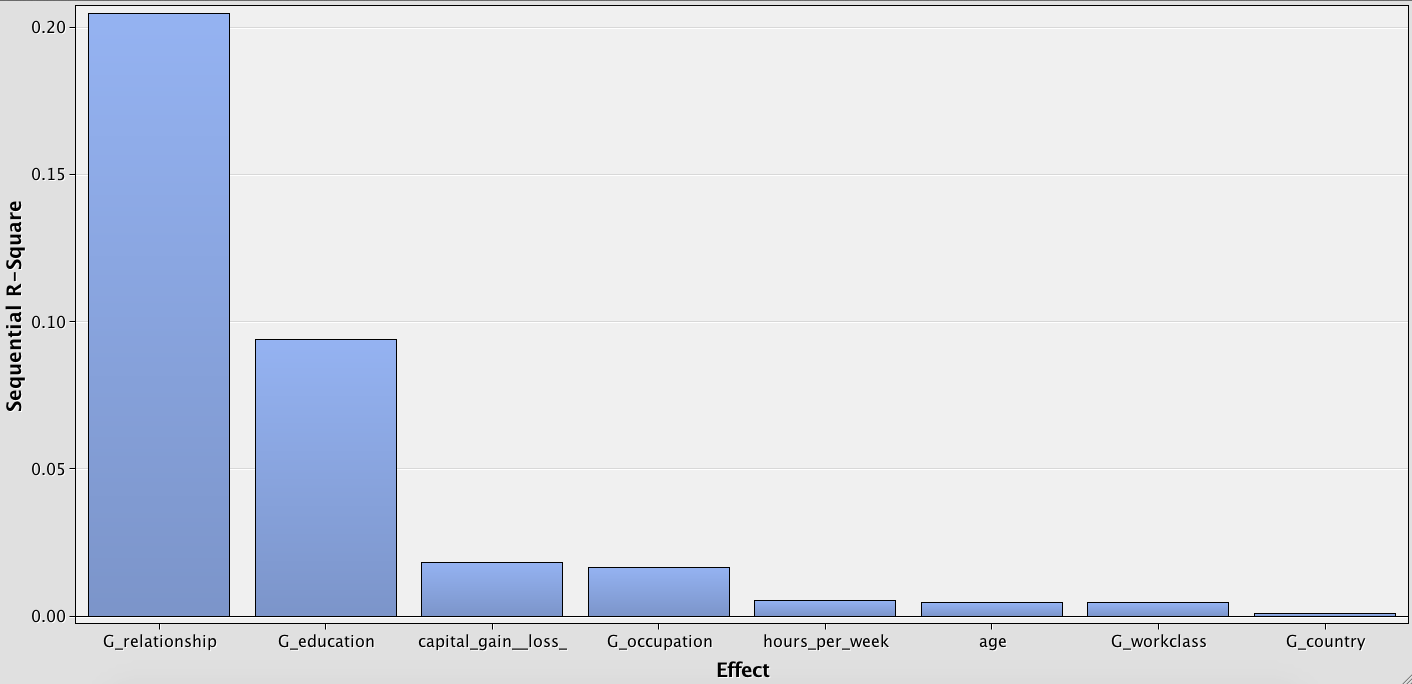
## **Model(s)**

**SAS Enterprise Miner:**

* Variable Selection

Using the “Variable Selection” node, which reduces the number of input variables using R-Square and Chi-Square, we have found that the variables selected are: Country, Education, Occupation, Relationship, Work Class, Age, Capital Gain (Loss), and Hours Per Week.



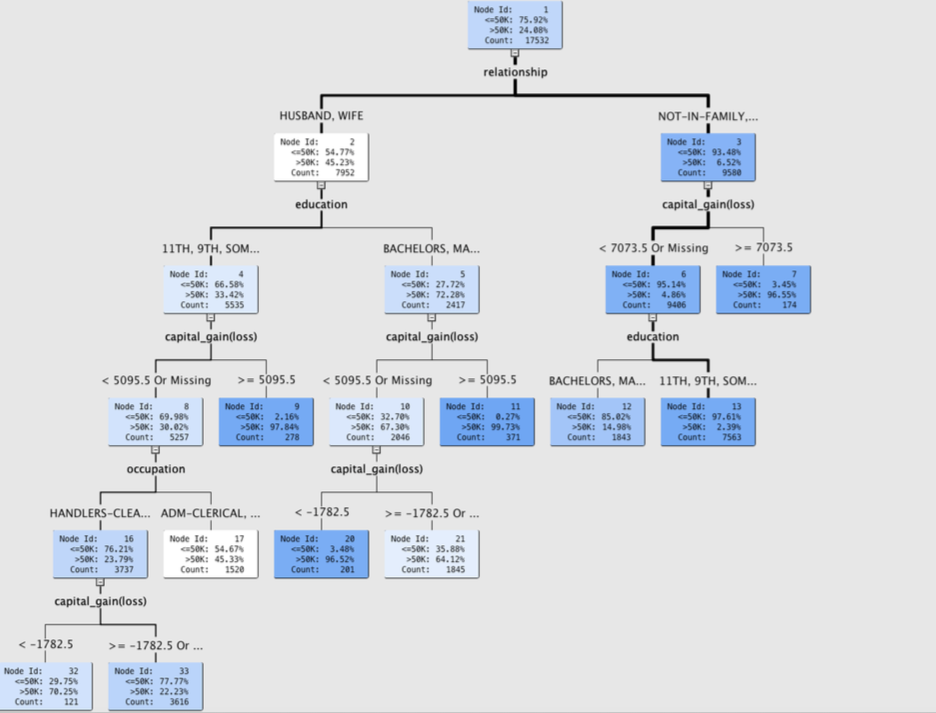


* Logistic Regression Model

The challenge of building the logistic regression model in the SAS EM comes from categorical variables that have many different values (i.e. country, work class, and occupation, etc.). In another word, the challenge is the selection of variables. In this case, we consider assigning dummy variables to different values of each categorical variable, and our goal is to identify the variables that are significant to our target (whether income is greater than $500K). To achieve the goal, besides the “Variable Selection” performed above, we can also take advantage of the “Stepwise selection” “Forward selection” and “Backward selection methods available within the SAS EM.

* Decision Tree Model

The decision tree is a hierarchical model composed of decision rules that recursively splits independent variables into homogenous zones (Myles et al., 2004; Cho and Kurup, 2011; Pradhan, 2013). The easiness of interpretation and the applicability to both classification and regression-type problem make decision tree model popular in predictive analytics. In our study, we have applied decision tree method to identify ways of splitting the training dataset into branch-like segments. We firstly started building a “large” tree using SAS Enterprise Miner. To do this, we imported the Decision Tree node in SAS Enterprise Miner, selecting “Yes” to allow the SAS EM perform cross validation, setting the “Method” under “Subtree” to “Largest” (to build a full tree), selecting “Average Square Error” as “Assessment Measure”, and keeping other settings default. Then we’ve got the decision tree with more than 100 of terminal nodes. This is due to the variables (i.e. country, education, and occupation, etc.) with many different values. Because “a decision tree combines the features in a hierarchical fashion such that the most important feature is located at the root of the tree” (Lee and Park, 2013), we know that that our initial attempt of building a “large” tree over fitted the data. So we further improved the model by “pruning” the tree and setting the number of leaves to “9”. Our goal was to identify the most important factors that contribute to income.



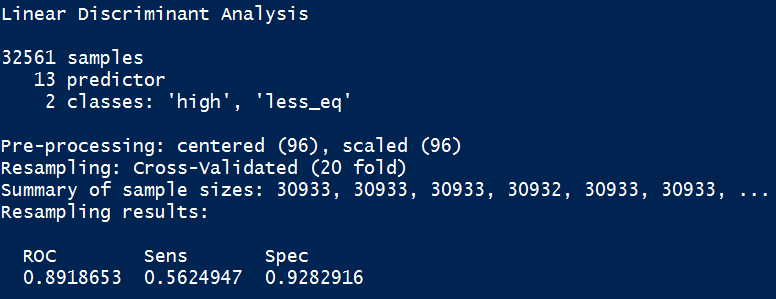
R-Studio:

**Sampling and Partitioning the data:**

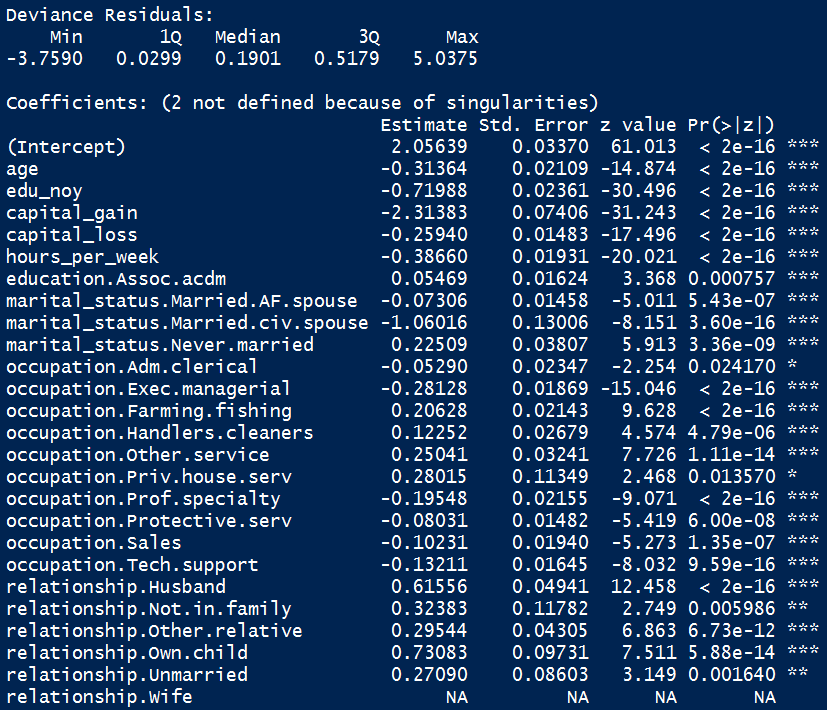
The UCI adult dataset already comes with two datasets separately for training and testing purposes. For this reason, we choose not to partition our data. However, we used a functionality of caret package to choose k-fold cross validation to train our model. We choose 20 folds as the standard for all the three models we developed.

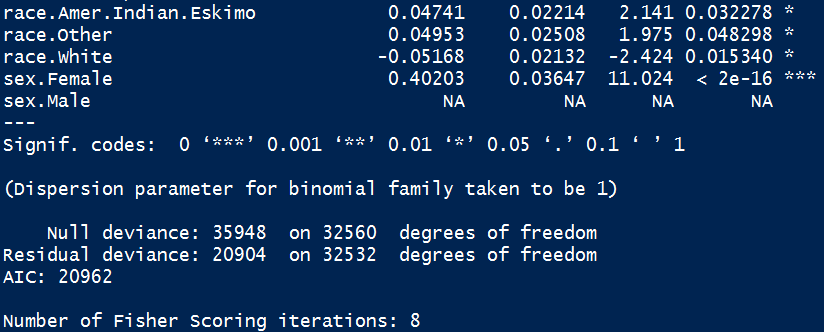
Caret package was used to run all the three models and we choose to center and scale the dataset for each of them. Income greater than 50K was selected as the target class for all the three.

**Linear discriminant analysis:** We ran LDA on our dataset using caret package on the entire dataset. The result is given below.

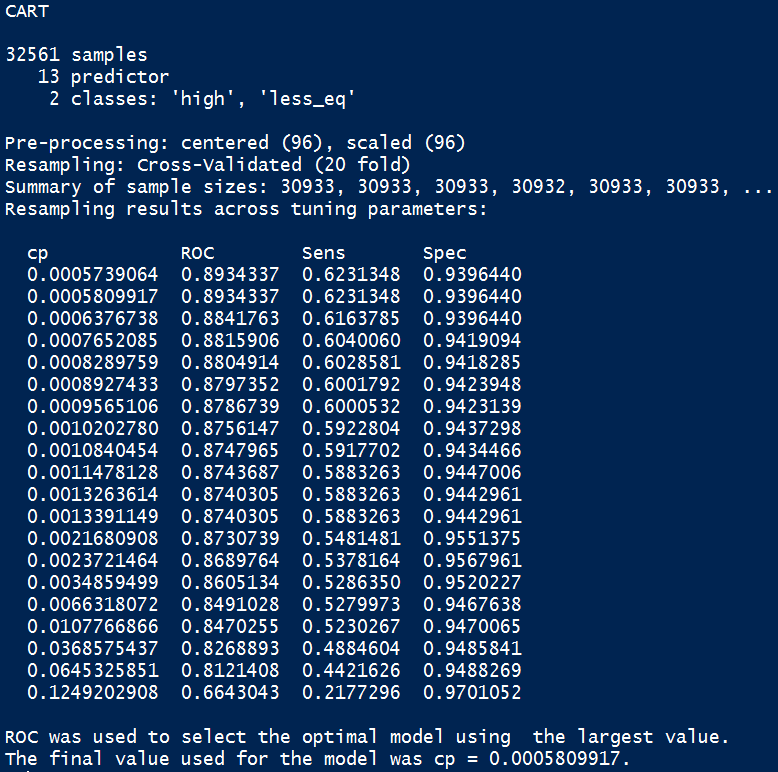


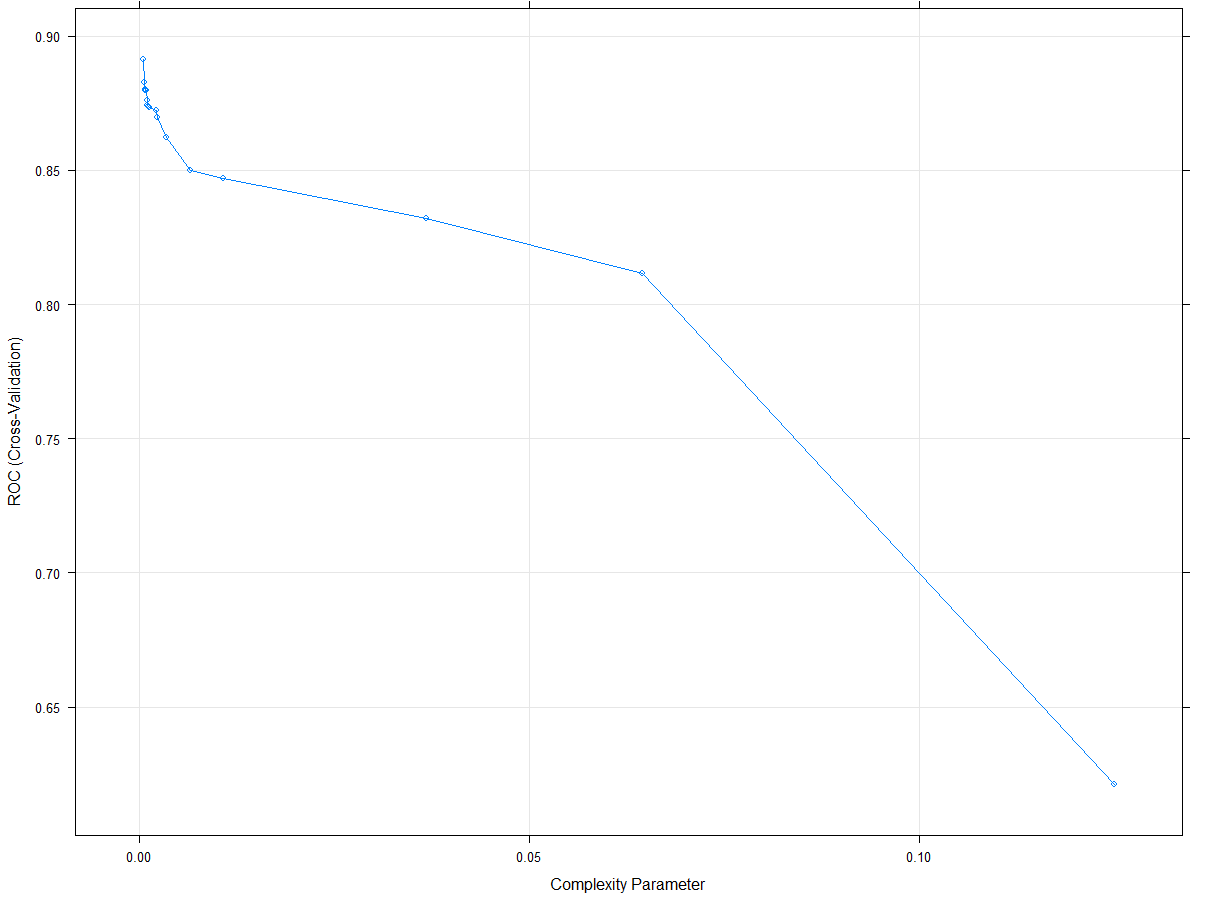
**Logistic Regression:** We created a function to convert all categorical variables into dummy variables using “Ade4” package. We first tried to use all the variables in our model but received error of rank-deficient model which could been result of two reasons. First, the units of various input variables might be very different and on different scale which makes it difficult for R to do matrix calculations in backend. To counter this, we pre-processed the data using center and scale functionality in caret before running the second iteration. Second, this error happens when there are not enough observations in the dataset and we removed country with 41 levels to counter it. Even after this the model kept giving the warning for rank deficient model and we iteratively started removing variables with highest p values to reduce dimensions. The final model was resolved of this error and all estimates where significant at 95 % confidence Interval.





**CART:** To build classification tree, we pre-processed the data and chose the tuning length of 20 to build the model.



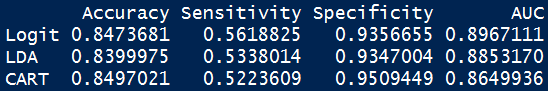


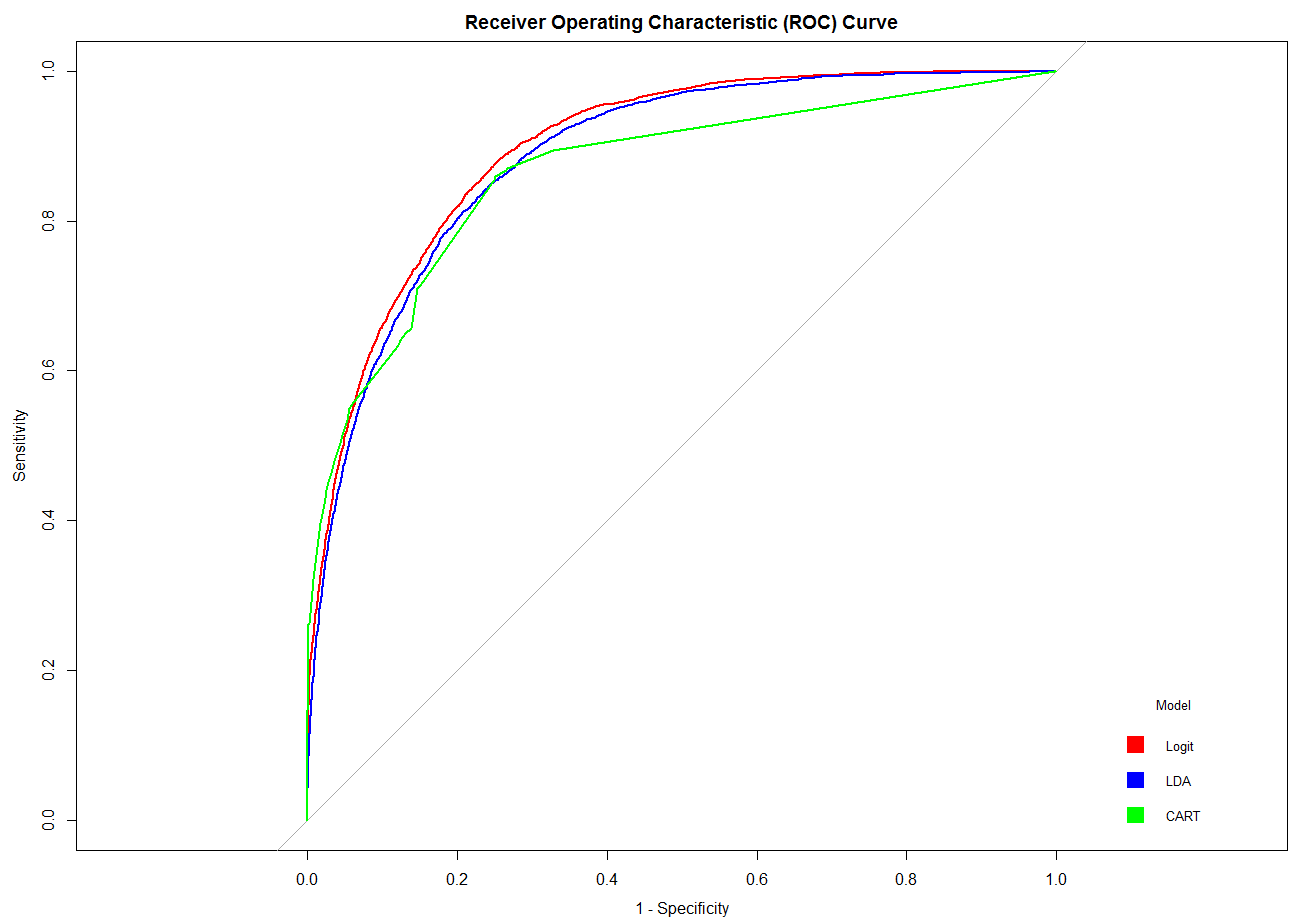
As seen by the output, the optimal model with highest ROC (Receiver Operating Characteristics) is one with complexity parameter of 0.0005809917. The fact that complexity parameter chosen over repeated cross validation is so low tells that highly complex model in this case might result in high variance and overfitting on the test dataset. We will see how the model performs when we get the test misclassification error and area under the curve for all the models and compare them to pick our choice.

## **Results**

Given that the test dataset was present to benchmark out performance, we didn’t use traditional criteria like Akaike Information Criteria(AIC), Bayesian Information Criteria(BIC) to see model performance on the training sets. We looked to compare the performance of our model using Receiver Operating Curve, Misclassification Rate, Specificity and Sensitivity on the test set to pick the best model for our needs. We also resampled 20 times from our training dataset to see the variation in performance of the three models based on the above mentioned parameters. This resampling also gave insights on how similar is the performance of three models for a particular. Let’s look at each of the above mentioned part in detail:

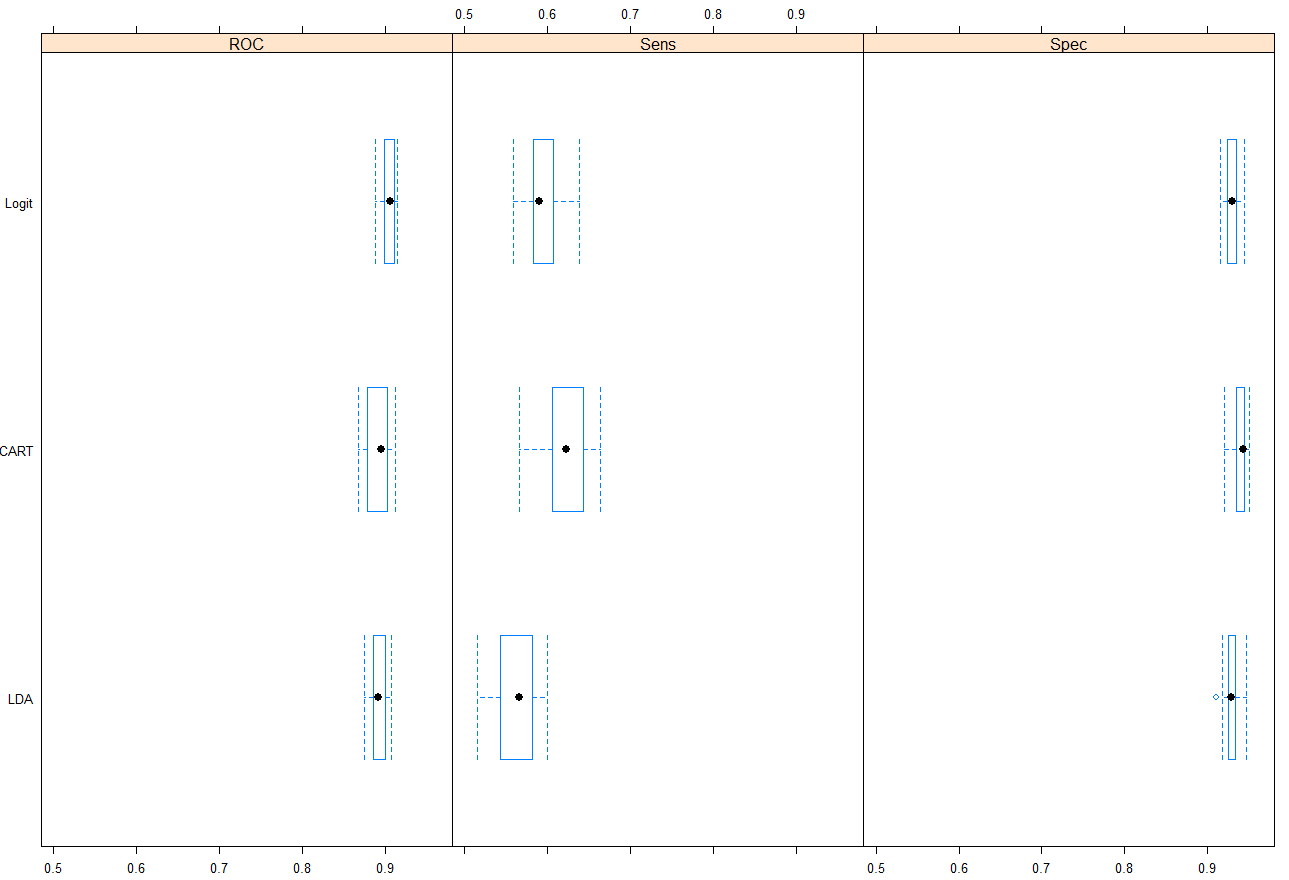
* **Performance on Test set**:





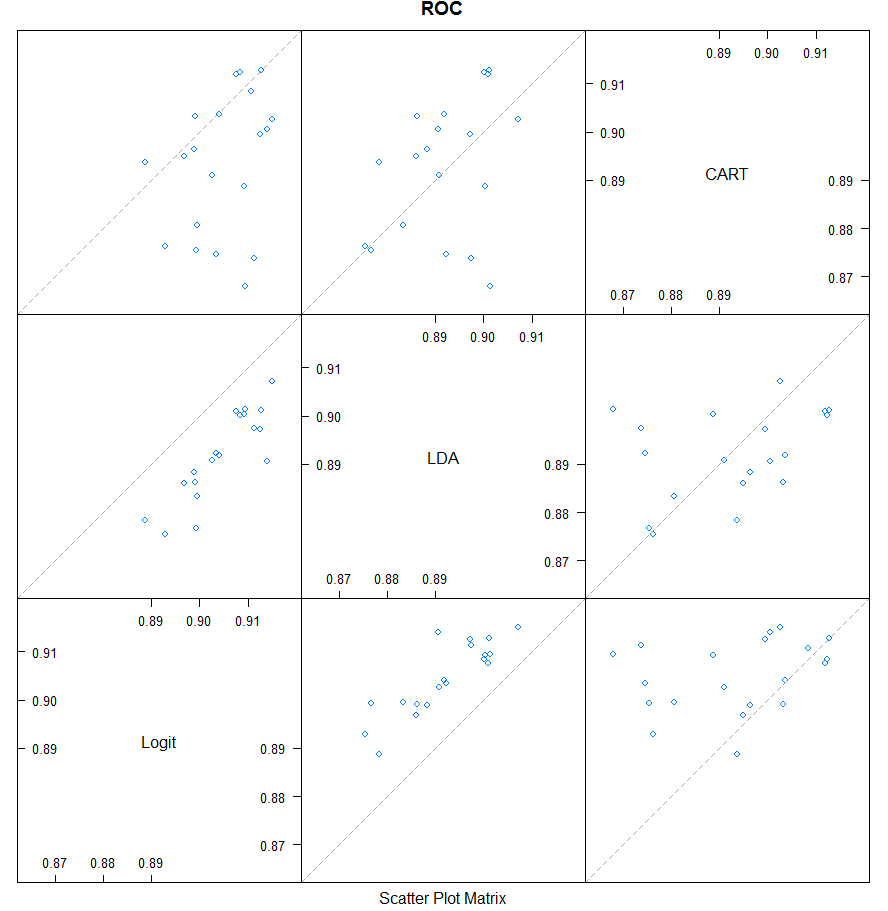
Firstly, all the three models perform better than the random predictor which would AUC of 0.5. The above ROC curve show that Logit model has the highest AUC closely followed by LDA and CART. However, the parameter statistics show that CART has the least misclassification rate out of the three models, while LDA does perform the worst out of the above in terms of misclassification rate. The logistic regression does best job of predicting the target class given by highest sensitivity while CART does a great job of predicting true negatives for the usual cutoff of 0.50.

* **Boxplot of performance metrics for 20 resamples of training set:**



The box plots show that Logit and LDA are very stable for various samples in terms of performance metrics of ROC, sensitivity and specificity. The Logit does perform slightly better in terms of having the least variation out of three models on all three metrics. CART shows striking variation while correctly predicting our target class(Income greater than 50K) which might be off concern to the business stakeholders. In the previous section we saw that the performance of CART on test set was worst out the three models in terms of sensitivity. The fact that the median sensitivity in training sample sets for CART is higher than Logit while for the test set it does a worse job than logit. The CART might be overfitting the given that its high variance in its performance on training resamples and inferior performance on test. LDA also does inferior job of predicting the target class and show high variance for various resamples.

* **Scatter plot of ROC for each resample for the three models:**



There are very few points close to 45 degree line which point that various models give different performance for different resamples. Few trends which are consistent are that for each resample Logit outperformed the LDA model while it also outperformed CART for majority of the samples. Another interesting thing to note is there is high variation in performance of CART and Logit model which means that for few sample when CART gives the highest ROC, logit tends to give the perform in the lower ranges and vice versa.

Lastly, it’s important to note that even though there are clear trends in terms of variation in performance, the difference on the actual range is never greater than 10 percentage points. This means that although some models conclusively perform better in some scenarios, all the three models do achieve high standards in terms of AUC.

**Best Model and Business Implications:**

Based the AUC and the business objective of correctly predicting the positive class of people having income greater than 50K, we would choose the logit model out of the three. They can expect this model of having high overall accuracy but will perform significantly better in predicting the negative classes (High Specificity) than positive class (Low sensitivity). Logit nonetheless was the best modeling predicting the positive classes and it still adds value in terms of overall accuracy. However, decision tree also helped us in identifying some important drivers for these relationships and we found that Relationship, education, and capital gain/ loss are the three most important features. When it comes to family relationship, the group “husband and wife” and the group “not -in-family, own-child, or other” have significant difference in the income classification (whether income is greater than $50K); the population in the latter group have higher probability of having income less than or equal to $50K). Within the group of “not-in-family, own-child, or other”, whether an individual has a capital gain (loss) greater than $7,073.5 is significant to whether an individual has income greater than $50K. Within the group of “husband and wife”, individuals with a degree of bachelor's or higher tend to achieve income of $50K or higher, compared to individuals with a degree of high school or lower.

**Conclusions**

Predicting if the income of a person is above aforementioned criteria by using socio-demographic variables as education, ethnicity, employment type etc. Team will also attempt to establish inferential relationship between the income group and demographic features. While much extant economic research has studied the driving factors for individual income, one of our research focus is to classify individual income into brackets with different dollar values, and develop effective model to assign individuals into the proper brackets. Our study could contribute to helping the U.S. Internal Revenue Service (IRS) conduct preliminary examination on individual tax return and flag the individuals for further investigation.

Our solution achieves high prediction accuracy through one methodology while also identifies the important drivers through another. This will help business understand not only possible causations but also predict likelihood of a person being in high or low income bracket for unseen observations.

Few ways in which the performance of our solution of are:

* Check for performance improvement of some other imputation methods on the null data and if any other method improves the accuracy on the test set
* Collect more data points so that categorical variable like country can be analyzed and problem of rank deficient models can be solved
* Test some other methods to see if they can improve the accuracy in predicting target class

## **References**

Ghoson, A. M. (2010). Decision Tree Induction & Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, And IBM Intelligent Miner - A Comparative Analysis. *International Journal of Management and Information Systems, 14*(3), 57-70.

Burgette LF, Reiter JP. (2010). Multiple imputation for missing data via sequential regression trees. *Am J Epidemiol*. 172(9): 1070–1076.

Sunita Jahirabadkar, Parag Kulkarni,. Algorithm to determine ε-distance parameter in density based clustering, Expert Systems with Applications, Volume 41, Issue 6, May 2014, Pages 2939-2946, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2013.10.025.

Gustavo EAPA Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. HIS, 87:251–260, 2002.

J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD, 2008.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemometrics 18, 275e285.

Data Visualization In Exploratory Data Analysis: An overview of methods and technology by Yingsen Mao

Missing Data Imputation for Supervised Learning, October 2016, Jason Poulos, Rafael Valle