Predicting Income Class with the Adult Dataset

By: Chris Beland and Bobby Mooney

Executive Summary

The main goal of this project was to develop models that would be able to accurately predict whether a person did or did not make over 50 thousand a year. We also examined what sorts of variables had strong predictability or influence in regards to whether or not a person makes over or below 50 thousand. The data set was the complete Adult dataset, which is found on the University of California Irvine machine learning repository website and provides basic demographic information as well as income class. Here are a few of our most insightful findings.

- The main business objective is to predict a person's income class to decide whether or not to potentially to increase available credit to the appropriate class (upper income).
- Our team concluded that because false positives resulted in giving credit extension to customers that were more risky, false positives were twice as costly missing out on a good customer.
- The most meaningful variables for predicting income class are education, relationship, and capital gains.
- The best models were the C50 and logistic regression due to their high specificity of 97.11% and 94.1% respectively at the cost of lower sensitivity of 36.54% and 52.51% respectively.
- To achieve a higher sensitivity going forward more data will need to be gather. More specifically, variables that have a meaningful impact on predicting income class.

Objective

As mentioned in the executive summary, the main business goal was to be able to classify whether a person makes over 50 thousand U.S. dollars or less. A few advantages of being able to predict this accurately might help decide whether or not to offer more credit to a person or to be used as a smaller predictor variable in a larger data mining set where information is given on a citizen but not income information. However, the company should look into rules and regulations about basing credit risk on demographic information Being able to predict this information might also lead to being able to help determine whether or not to target a person with a high income level advertisement. These models/information could be then sold to an advertising agency or used by our company. Overall, these are all possible useful bits of information that might assist in the decision making process of some upcoming business decisions.

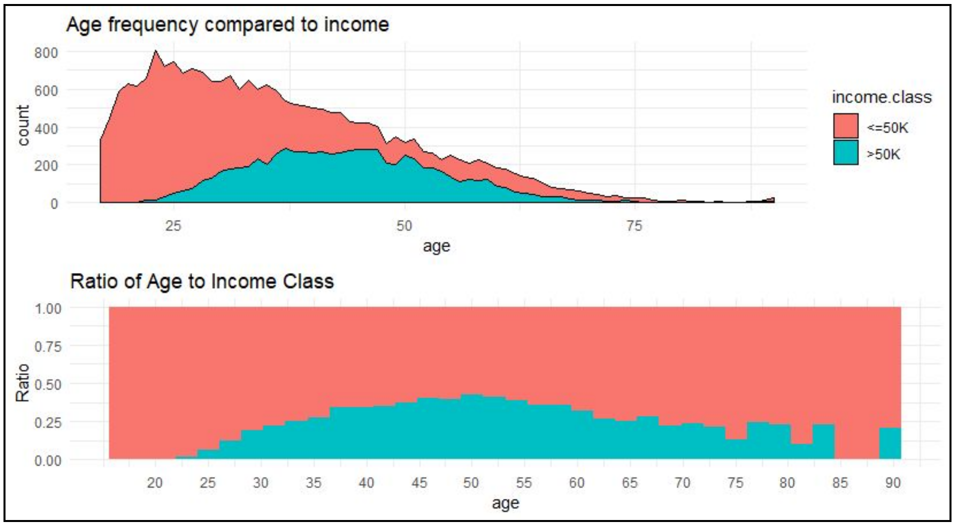
Since real world data may not be exactly like the data used in this dataset some variables with many different types of levels or categories were simplified which will be explained more in the EDA. The dataset itself contains 15 variables and 32,000 observations. Here is a list of variables and what they represent.

1. age: tells the age of the person [integer]
2. workclass: tells what sector a person works for (i.e government, private, ect.)[factor]
3. fnlwgt: estimate population that it represented by this observation [integer]
4. education: the education level (i.e.6th-grade, undergraduate, doctorate, ect..)[factor]
5. education-num: a number representing the education level [integer].
6. marital-status: the marital status (i.e.married, divorced, un-married)[factor]
7. occupation: The job type(i.e.office, labor, tech, sales, ect…)[factor]
8. relationship: explains that type of family member (i.e. husband, relative, own-child)
9. race: the race (i.e. black, white, asisa, ect..)[factor]
10. sex: male or female [factor]
11. capital-gain: whether or not their investments increased experience [integer]
12. capital-lose: whether or not their investments lost value for the year[integer]
13. hours per week: how many hours a week are worked [integer].
14. native-country: the native country (i.e. USA, Canada, France)[factor]
15. income.class: whether or not the person made over 50k or 50k or less.[factor]

EDA

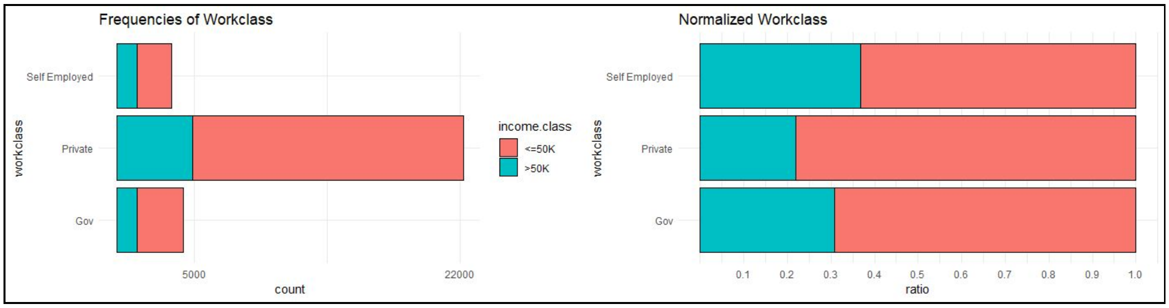
Once the data was introduced into R, we took a look at the structure of the dataframe by looking at type of data that each variable represented and the summary statistics, giving the mean and quartiles of numerical data and the distribution of the factor data. Note that the data was imported with each string of text being classified as a factor. These factor levels were also condensed into smaller groups based off of similarities. This was done manually through the R program by re-assigning the classes of each factors level. Making examination easier for the end-user to see relationships that each variable has with the target and its fellow predictors. The downside to this is that it might lead our models performing worse due to having less classes to choose from. Then the dataset was checked for duplicate records that were exact matches and discarded these duplicates. There was also a decent chunk of missing data points in the workclass, occupation, and native country variables. These missing data points will be dealt with later in the report in the pre-processing section. Once this was accomplished, each predictor variable and its relationship with the target variable was examined.

The first variable in the dataset is age. To get a better understanding of age refer to the density plot and a normalized histogram. The density plot is a type of histogram that connects the points with lines and works well with continuous numeric variables.



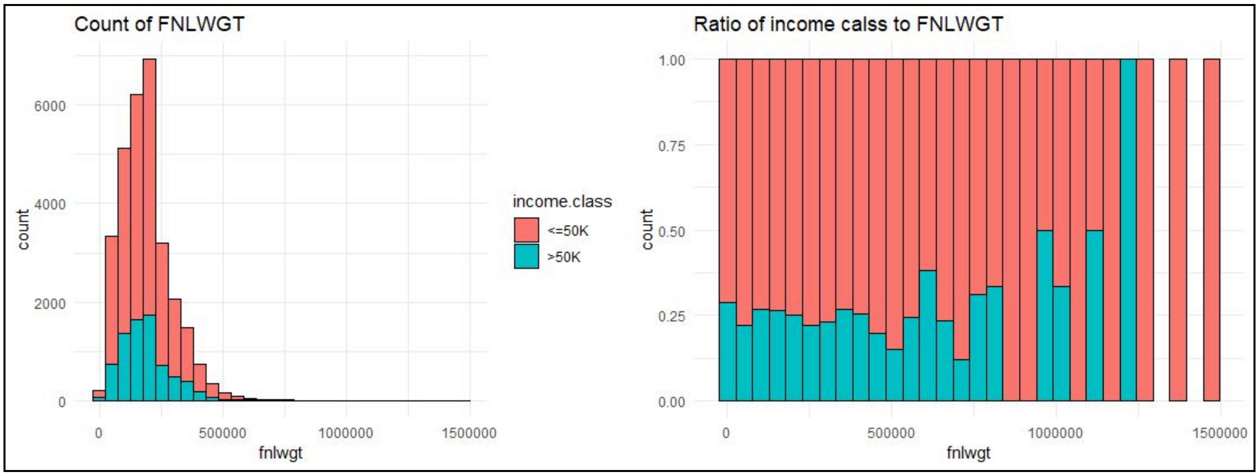
The graphs show age as a whole is right skewed. Because age is right skewed, making sure that its skewness will not violate any of the used models' assumptions will be important during modeling. Interestingly, age with income greater than 50k exhibits a distribution closer to normality. This makes sense as those who make more money tend to be older and in the later stage of their career due to the more experience they have to offer. Age will most likely be useful when determining the potential earning ability of an observation.

The next variable looked at was workclass. Workclass is broken down into these classes '?', 'Gov", 'Private', and 'Self Employed'. The '?' represents a missing entry and will be excluded from the EDA graphs and looked at more indepthly later. The best way to see how each of these might affect a person's potential to earn over 50k is through a bar chart that displays the count and then the normalized version.

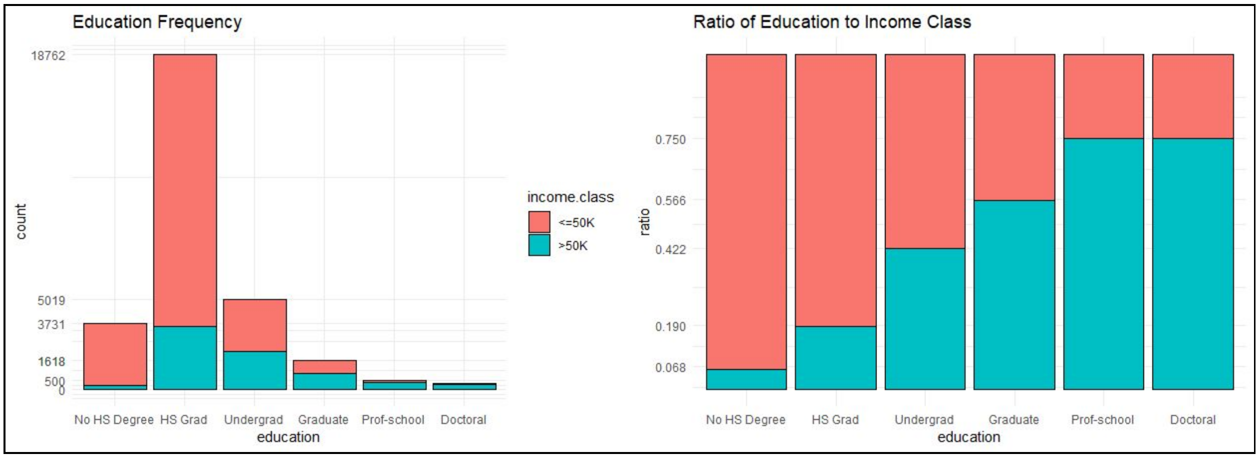


These charts show that the majority of the higher income respondents work in the private sector. Government and the self-employed workers are substantially smaller in number compared to private but tend to fall into the higher income class proportionally more often. A reason for this being the case is that the private sector employers employ more people and offers more lower paying jobs, however, this sector also employs over half of all the people that make over 50k. Overall, it seems that workclass does not carry much predictive capabilities due to its imbalanced and even ratio of high income class.

The fnlwgt variable is an interesting one. As mentioned earlier fnlwgt is basically giving the weight of each observation that represents how many people fall into this category in the United States. For the type of business objective that we are looking at, this variable does not carry any use and thus will be excluded from the models. Furthermore, it is unlikely that real data implement into our models would come with a fnlwgt which is another reason for its removal. Furthermore, looking at the graph below, there does not seem to be any solid trend or correlation to the type of fnlwgt in its relationship to income class.

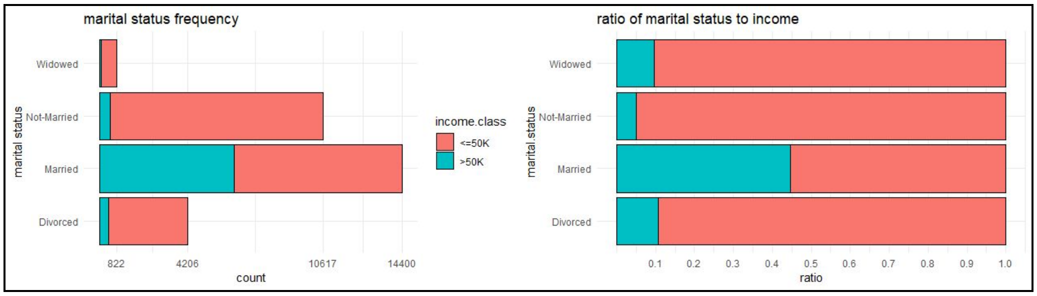


Next comes the education and education number. Education represents the level of schooling achieved by the individual and education num was originally represented the number of years education completed but after education was condensed it lost its meaning and was removed. A similar version could be implemented for the new education levels however, note that the distance between a highschool graduate and an undergraduate compared to an undergraduate and graduate student is subjective. As with most categorical data types the best way to examine it is to look at table or barchart.



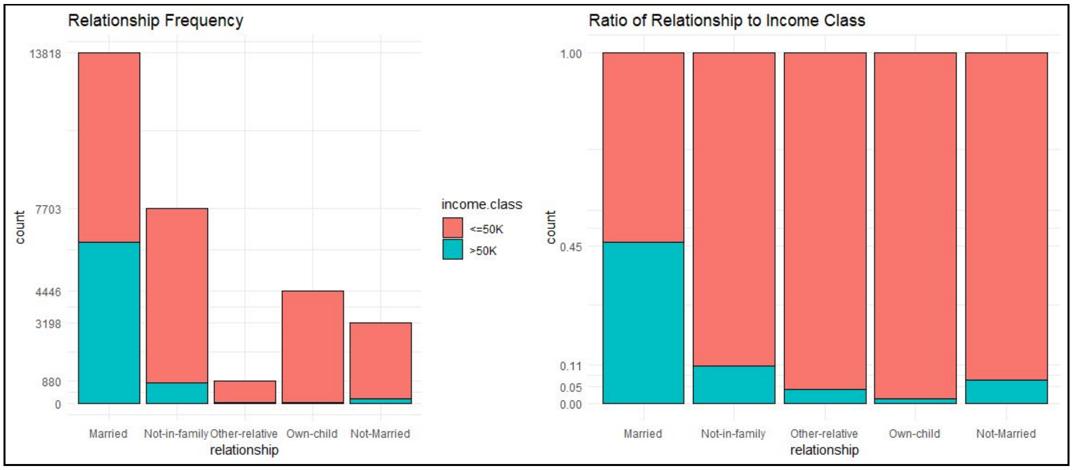
The count or frequency of education levels clearly show that the majority of people in this dataset have a highschool level of education. Looking at the normalized version it is also clear that as the level of education increase so does a person's chance of making over 50K. This seems intuitive and is what would be expected in the real world. Furthermore this graphs shows that the level of education seems to have a large effect on weather or not a person makes over 50 thousand or not.

Here is a bar chart that shows the normalized and frequency of marital status with respect to income class.



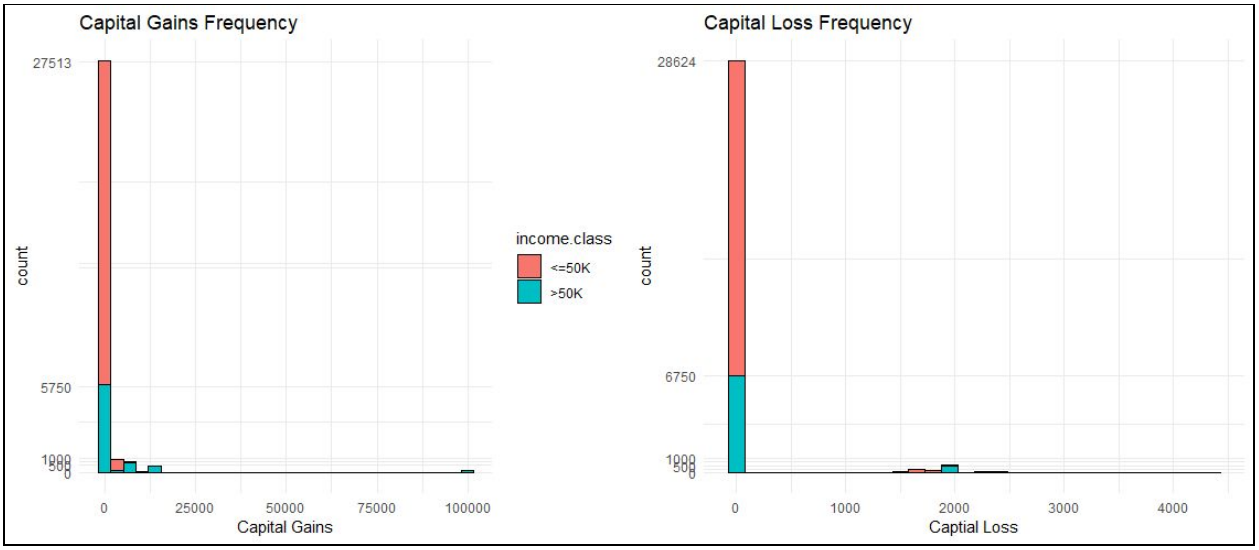
These graph shows that the majority of the dataset falls within the married and unmarried with a smaller portion being divorced or widowed. It also shows that being married has a big influence on whether or not a person makes over 50k. This might be the case due to the sharing of household income. Basically, it is much more feasible for two people to earn 25K each compared to one person earning 50k. However, the dataset does not actual say if married couples income are combined so there's no clear way to make this a definitive claim.

After marital status comes relationship status and these two variables have similarities with respect to the information that they provide. Nevertheless let's look at relationship status with income class and marital.status. It is worth noting that relationship includes the levels of 'husband' and 'wife which tells us if the observation is married and their gender. Due to gender already disclosed it is best to re-name both to married.

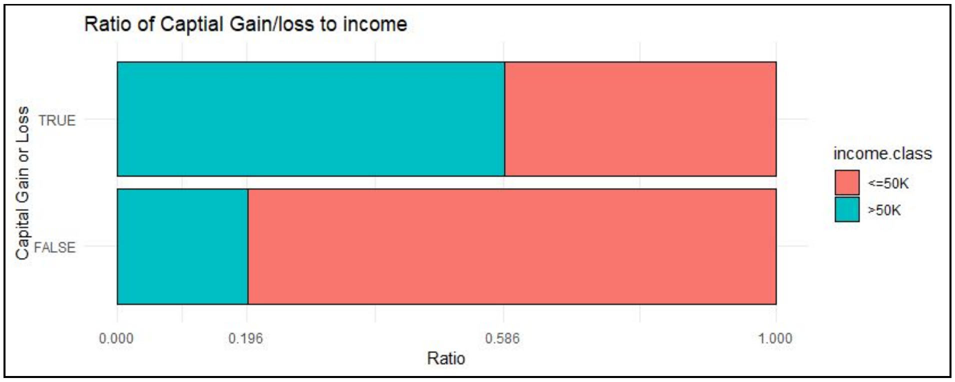


Here the graph shows that relationship is extremely similar to marital status. The only difference is the way it classifies non-married couples with relationship breaking down into relative and own child. This makes sense as it was originally established that married people earn over 50k more often than non-married people by the marital status variable. Looking at marital status and relationship this chart clearly shows some overlap as these two variables capture similar information. Either way, one of these two variables should be dropped as they are similar in the information they provide. To help determine which variable should be dropped the results of a logistic regression using both variables as predictors to predict income class was looked at. The regression showed the relationship classes 'Not-in family', 'Other-relative' and 'Own-child' were more significant (lower p-value) than the classes "Divorced' and 'Widowed'. Also their beta values were larger which indicates that they influence the decision more than marital status. Furthermore, looking at the charts of relationship and marital status, the relationship variables do seem to break down the non-married sections into larger and more meaningful splits relating to income class. This was the reasoning behind dropping marital status and retaining relationship. The implication for dropping marital status resulted in a reduction of .02% in overall accuracy in the logistic regression model ability to predict the test dataset. Because this value was small our team felt that was further evidence for omitting marital status

Here are the capital gains and losses histograms which give a visual representation of their respective frequencies.

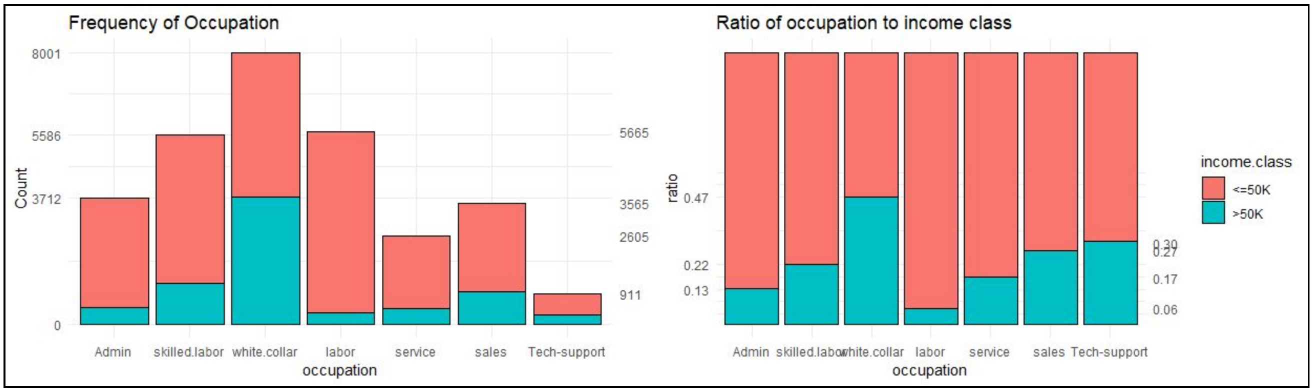


Notice that the majority of the data set did not have any capital gains or losses. This results in the distribution of both being right skewed and thus not normally distributed. Notice how as capital gains and losses increase the proportion of higher income earners increases as well. This relationship can best be seen by utilizing the following chart on the next page which shows the observations that had capital gains or losses and their respective ratio of income class.



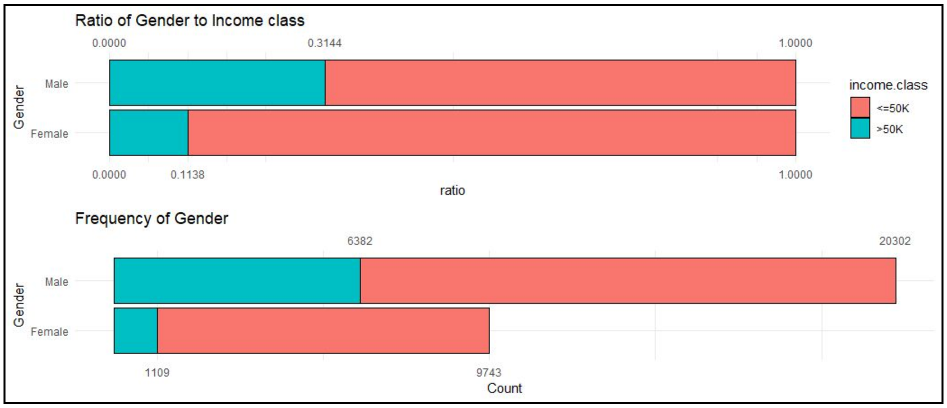
Here the continuous capital gains and losses are being transformed and the general trend is being captured by creating a categorical variable based on whether or not a person has capital gains or losses.This process is called binning. By binning we probably lose some prediction accuracy but the problems of the skewed data and distribution are now gone. However, the new categorical version is imbalanced. The capital gains variable also had some observations with values of 99,999. This might be due to a system entry error or not as it is impossible to tell. However, all of these capital gain entries fall in the >50K class. This points to the possibility that when the data was originally recorded the variable had max entry of 5 digits and any gain over that was recorded at 99,999 as there were no capital gains data point above one hundred thousand. It is impossible to know for sure but due to the fact that these observations all fell in the right income class they were left in the dataset.

Occupation was a variable that required a bit of work to condense its levels down in a logical manner to a smaller size. It originally had 15 different levels. Some of these levels shared similarities in the type of occupation/job and ratio of income class. These levels were broken down based on these factors resulting in these new condensed levels.



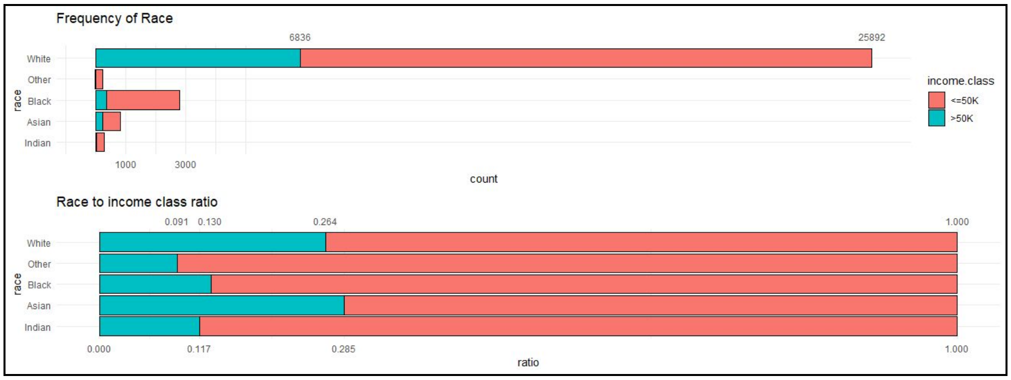
These graphs show that occupation looks to have a significant impact on whether or not a person makes over 50K which makes sense. The only issue that might arise when using these classes is determining what job falls into which category or class. This is important to get right due to the predictive power that occupation appears to have. Thus, whoever is using the models should have a basic understanding of the un-edited occupation classes.

Now let's look at the distribution of gender in the dataset. As usual here are barcharts based on frequency and normalization.



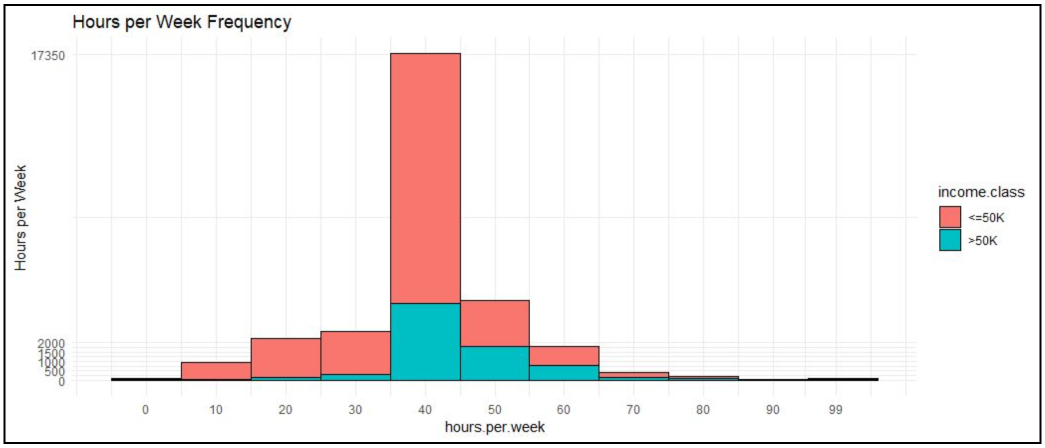
One note worthy piece of information that can be gained from looking at these charts is that men are represented more than women by a significant margin. This is not indicative of the entire US population as the split between males and females are pretty even according the CIA world factbook (CIA). The reason for this imbalance could be due to more males being polled than females or that the working population consists of more males than females. Regardless, the charts do seem to show gender having some predictive power in regards to income class with 30% of males make over 50K compared to just 11% of females.

Race on the other hand, does not seem to be as useful due to the majority of Americans being white as these charts depict.

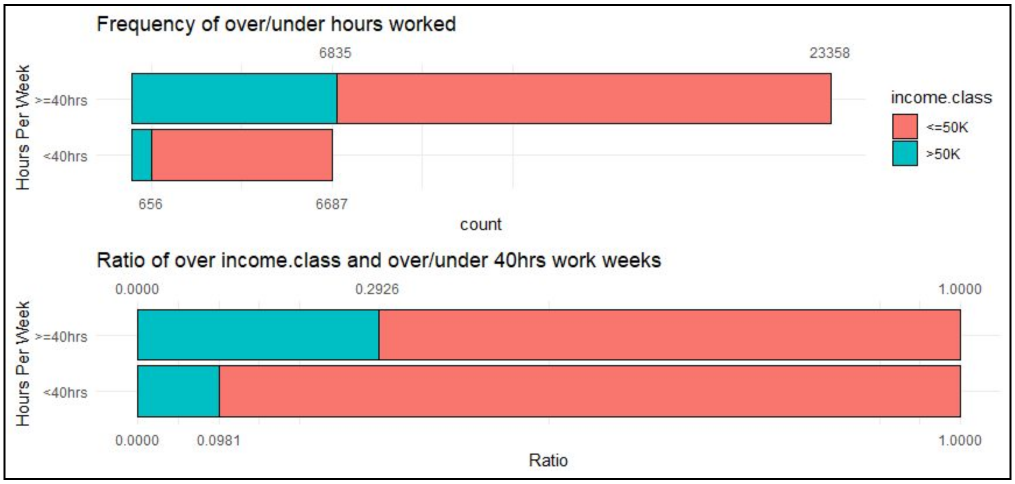


It is apparent when looking at these charts that the white and asian races have a higher percentage of over 50K earners. However, this should have little effect on predictability due to the balance of the classes. No matter how good a variable is as a predictor if the majority of the data falls within one class it loses its ability to separate into useful splits.

The last numeric variable in this dataset is hours per week. To help get a better understanding to the distribution of this data here is a histogram.

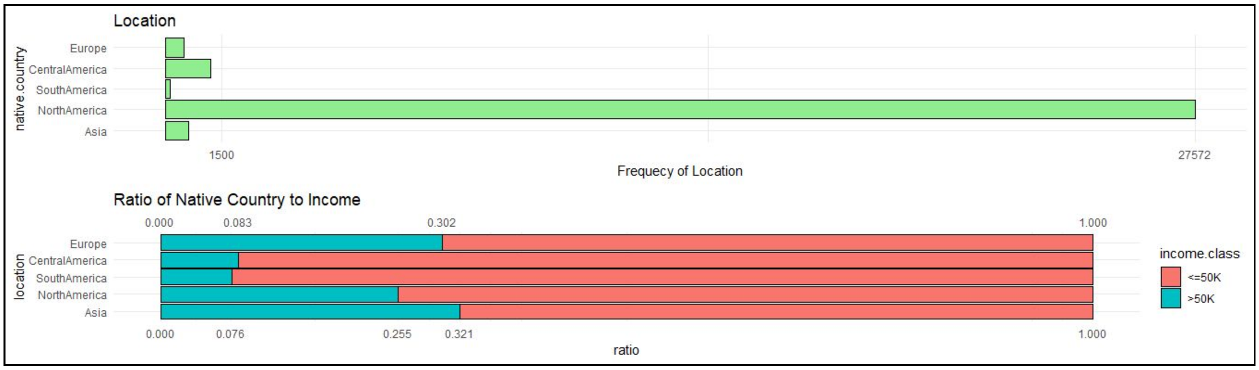


Here we can see that the majority of people in this dataset work around 35 to 45 hours. There is also a significant drop off when looking at the ratio of income class when split below and at or above 40 hours per week. This bar chart does a good job in capturing this trend. Due to the uneven distribution of the work hours, the variable was turned into a categorical type telling whether or not a person works 40 hrs or more. This also generalizes making it more useful when dealing without a specific amount of working hours.



Here it is clearly depicted that working at least 40 hrs has an increase in the likelihood of a person making over 50K. However, only about ⅕ of the dataset works less than 40 hours so this will limit the predictive power.

Finally the last variable in this dataset is native country. There was almost 50 different countries on the list so they were put together based on their similar location. However, as this is an American census the most common area or country will be the United States or North America as this chart depicts.

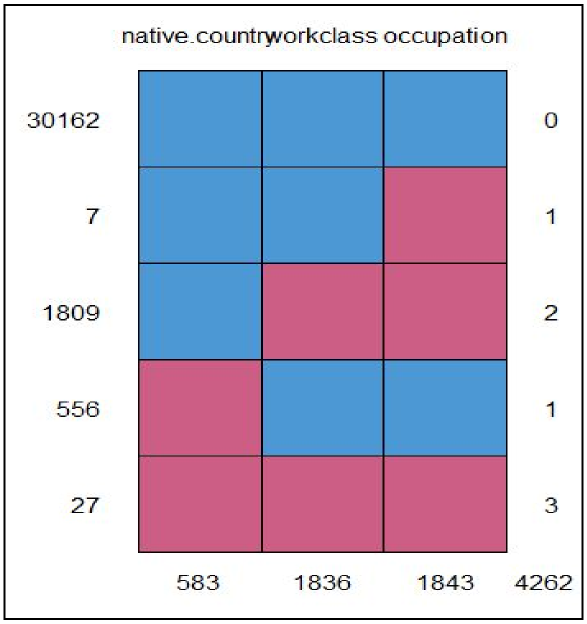


The first chart shows that the majority of this dataset's observations native location is North America. The second chart shows that each location has a different proportion of high class earners with North America, Europe, and Asia. However, based on the how imbalanced this factor is, the overall predictive power of the variable should be low.

Pre-Processing

Once the initial EDA was completed the variables ‘fnlwgt’ and education num’ were removed due to reasons listed prior. Then the following integer variables were transformed into categorical variables ‘capital gains’ ‘capital losses’, and ‘hours per week’. Specifically, capital gains/ losses were split into 0s and 1s with 0 representing no capital gain/loss and one representing a capital gain/loss of any amount. Hours per week was also split into 0 and 1 with 0 representing a person that works less than 40 hours and 1 representing a person who works equal to or greater than 40 hours. Categorical/ factor variables were then transformed into dummy variables to the (k-1). These variables included ‘workclass’, ‘education’, ‘marital. status’, ‘occupation’, ‘relationship’, ‘race’, ‘sex’, ‘native.country’, ‘income.class’, ‘capital.gain.yes’, ‘capital.loss.yes’, and ‘hours.per.week.over40’.These dummy variables versions were assigned as their own dataset. Creating two versions of the dataset (one categorical/numeric and one dummy/numeric) allows for the use of models based on distance such as KNN and models that work with classes/factors such as the C5.0 and CART decision trees. To create the dummy variables the model.matrix function was used.

Before the continuing to the modeling section, the missing data in workclass,occupation, and native country was looked at a bit closer. The following graphic on the next page displays the dispersion of missing data nicely. Notice how workclass and occupation are almost always missing together (1809). This means that when attempting to impute the data it would be frivolous to include occupation to predict workclass and vice versa. The distribution of the other variables were examined as well using the table function to examine if their seemed to be an underlying pattern in the missing data. Nothing seemed too out of the ordinary but there was substantial difference with a the missing data having a lower percentage of high income class workers. A reason for this being the case might be it was harder to obtain information from people who earned a lower income. Imputation was attempted using C5.0 and CART decision trees using all complete observations.



The performance of the models were poor when compared to the base proportions of largest category in workclass and native country. This means that the same error rate could be achieved or improved by selecting the level with the most observations which would be private class and North America respectively. In fact the CART decision tree did not even split on workclass as it found it was best to just classify all missing data as private. Based on these models it seems that there is no strong relationship with these variables and the rest of the variables used as predictors. The occupation trees had an error rate of 51% on the training set making testing the test set not worth while. Based on these results the missing data was excluded from our model. Even though native country and workclass could be imputed at a acceptable accuracy it will be later shown through the modeling section that these two variables will have little effect on prediction and it would be unwise to use occupation imputations as it would increase garbage in due to the high error rate.

Next the dummy variables dataset and the regular data set was broken down into training and test sets at a split of 80% of the data used in the training set and the remaning 20% of the data used as a validation or test set. With the size of the complete dataset being 30,000 this proportion seemed reasonable. A seed was used so that the results could be replicated. Next, the training and test data set was tested to make sure that the splits are representative of each other. This was accomplished by running the t-test on numeric variables and Chi-square test on categorical variables. Once a good split was confirmed the one remaining integer variable age was normalized with min/max. The remaining integer variables of capital gains, capital losses, and hours per week were transformed into categorical variables.

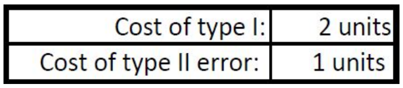
Modeling Section and Variables Used

To evaluate the performance of all models, their accuracy in predicting the respective test data set will be used. To resolve the concerns of overfitting, the models will be judge using their accuracy on the validation dataset which does not include the target variable income class. The models that were used were logistic regression, C50 and CART decision trees, KNN, and neural networks. Theses models used the following variables age, education, occupation, relationship, sex, capital gain, capital loss, and hours per week. This combination of predictors achieved similar and in some cases better accuracy than including race, workclass, and native country. The overall accuracy of the models including these variables can be seen in the appendix.

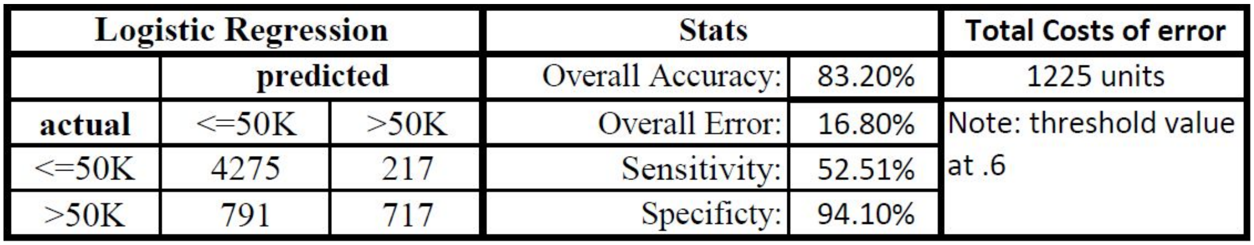
Interpretation of the Models

To evaluate the models it is important to look back on our business objective to consider the consequences of a false positive or a false negative. These models will be used for making a decision on whether or not to increase a person's credit. The consequences of a false positive can be described as selecting a person who may abuse the credit limit and significantly hurt the company. The consequences of a false negative is denying a potential customer who has the resources to spend and will bring in extra revenue to the company. Thus, even though a false negative does not actually cost the firm any money the incorrect classification can be looked at as a loss of potential revenue. The company currently does not have any information about actual costs of these errors so out team decided to make false positives be twice as costly as false negatives.

The first model that is used is a type of generalized linear model called logistic regression. Since logistic regression was not covered in the course an brief explanation of the model can be found in the appendix. The function glm() was run on the the complete training set of dummy variables. The summary output is useful as each variable’s significance as a predictor can be estimated by the p value. High p values indicate that the predictor variable does not have a significant impact on being able to predict income class. This is a useful feature not found in the other models implemented. Another useful feature of logistic regression is the ability to adjust the classification level. For example, all inputs with a value greater than .5 are classified as a 1 or income class of >50K. This .5 value is referred to as the threshold value. A higher threshold value will cause the sensitivity of the logistic regression to decrease while increasing the specificity of the regression. One way to look at sensitivity is how accurate the model is in identifying individuals who made over 50K while specificity shows the accuracy of selecting individuals who made less than 50K. It was established earlier that false positives are more costly than false negatives which means that a higher specificity will be more important to the firm to minimize cost. After experimenting with various threshold values .6 was chosen as it minimized the total cost to the firm based on each errors assigned costs. Finally, the impact that each variable had in classifying income can be estimated by looking at the absolute value of the coefficient, with a higher coefficient value indicating the weight or importance that each variable had in the classification process. The top three variables based on the coefficients are education, relationship, and age closely followed by capital gains.

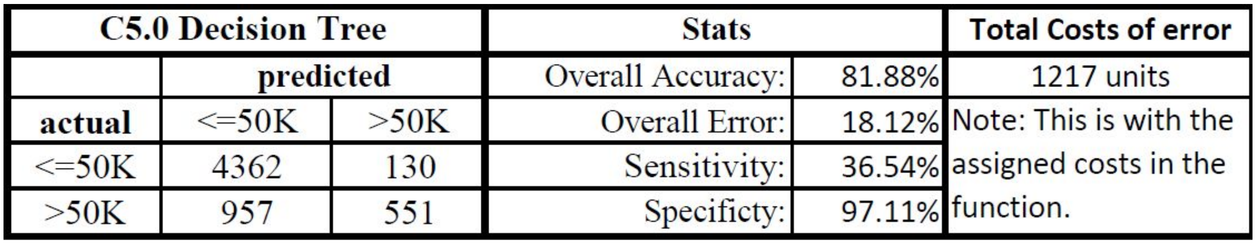


Cost of type I:	2 units
Cost of type II error:	1 units



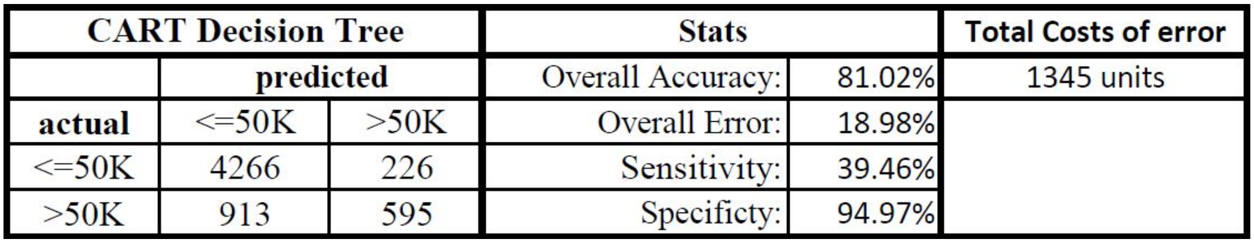
Logistic Regression			Stats		Total Costs of error
	predicted		Overall Accuracy:	83.20%	1225 units
actual	<=50K	>50K	Overall Error:	16.80%	Note: threshold value at .6
<=50K	4275	217	Sensitivity:	52.51%	
>50K	791	717	Specificty:	94.10%	

The second models that were used were decisions trees. Our team used two types of decision trees which were the C5.0 and CART decision trees. The main difference between these trees is that the CART only does binomial splits while the C5.0 can create leaves or nodes with more than two. One benefit that the C5.0 decision tree has over the CART tree is its ability to assign weights to error so the model can develop splits that minimize the cost of those errors. The variables that the C5.0 decision tree found most useful were relationship, education, and capital gains. Using the weights of 2 for a false positive and 1 for a false positive the following results were achieved. (The results of the non-weighted C5.0 are in the appendix.).



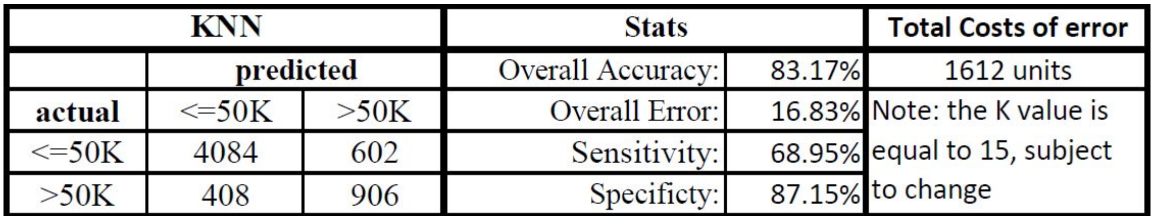
C5.0 Decision Tree			Stats		Total Costs of error
	predicted		Overall Accuracy:	81.88%	1217 units
actual	<=50K	>50K	Overall Error:	18.12%	Note: This is with the assigned costs in the function.
<=50K	4362	130	Sensitivity:	36.54%	
>50K	957	551	Specificty:	97.11%	

The CART decision tree on the other hand, does not have the cost function built into it. This naturally lead to results that were more undesirable than the C5.0 decision tree as the results show below. The two variables that the CART tree used were relationship and education number



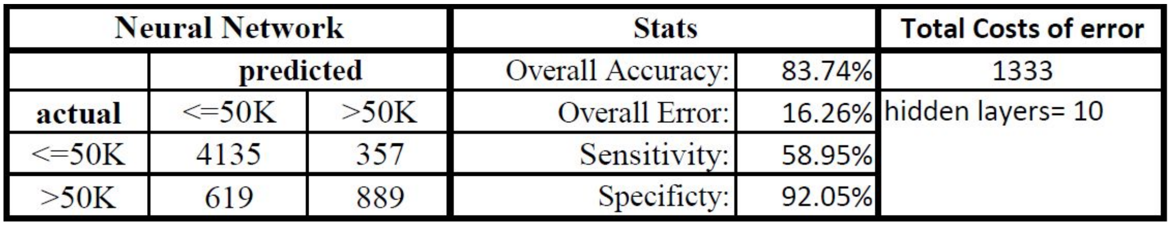
CART Decision Tree			Stats		Total Costs of error
	predicted		Overall Accuracy:	81.02%	1345 units
actual	<=50K	>50K	Overall Error:	18.98%	
<=50K	4266	226	Sensitivity:	39.46%	
>50K	913	595	Specificty:	94.97%	

The third model or algorithm was the KNN model. There are a couple of issues with the results generated by the model. One is that everytime it is run different accuracy measurements are generated due to ties being broken at random. A potential solution to this problem and one that might increase the accuracy of the KNN algorithm is to stretch the axises of the important predictors like occupation, capital gains/losses, and occupation. However, this was not done leaving ties to be broken at random which is not advised. Next came the issue on determining which K would result in the highest accuracy on the validation set. To address this a for loop was created and run to see the accuracy of k values 1-30. The accuracy value around 83% percent across all k values or an error rate of about 17% which is similar performance seen in the other models. Due to the nature of how the KNN algorithm was set up, all variables carried the same weight of importance due to the min/max normalization. To get a look at the performance of this model we selected a 'k' value of which was the highest accuracy percentage for the current instance and re-ran the knn algorithm to determine the error costs and its specifictivity which is shown below.



KNN			Stats		Total Costs of error
	predicted		Overall Accuracy:	83.17%	1612 units
actual	<=50K	>50K	Overall Error:	16.83%	Note: the K value is equal to 15, subject to change
<=50K	4084	602	Sensitivity:	68.95%	
>50K	408	906	Specificty:	87.15%	

Finally, the last model that was implemented was the neural network. There are few points to mention to before getting to the final results. Due to the randomized weights it is not guaranteed that the the global minimum will be found. Then an optimal amount of hidden layers had to be established. After experimenting with a varying number of hidden layers 10 was selected as it provided a lower error rate on the validation data set. A drawback of using neural networks is that they are a black box algorithm meaning its results are difficult to interpret which leads to not knowing the most important variables. The error costs are shown below.



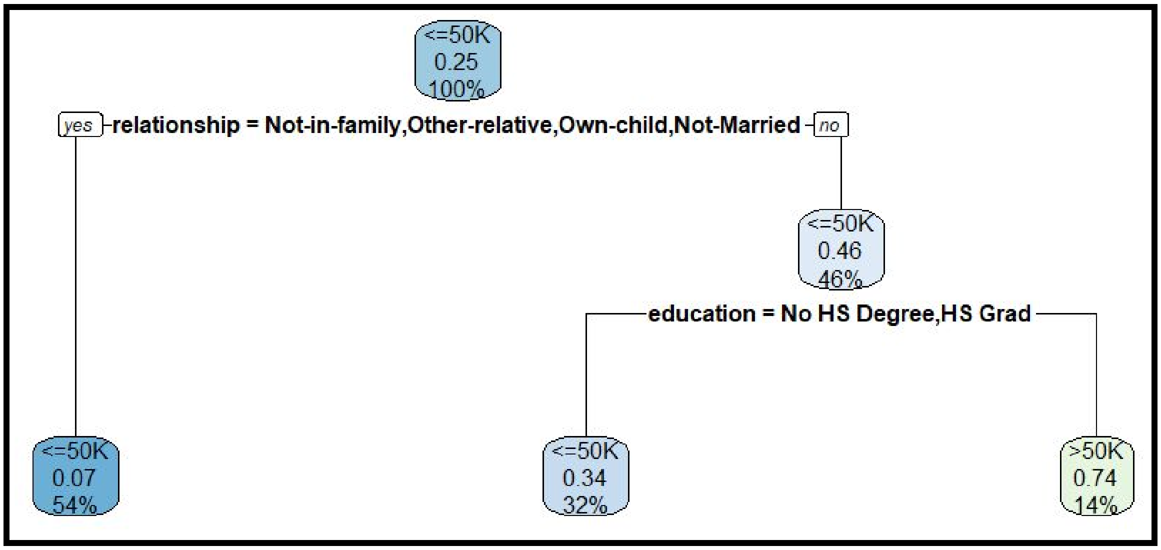
Neural Network			Stats		Total Costs of error
	predicted		Overall Accuracy:	83.74%	1333
actual	<=50K	>50K	Overall Error:	16.26%	hidden layers= 10
<=50K	4135	357	Sensitivity:	58.95%	
>50K	619	889	Specificty:	92.05%	

Conclusion

Looking at the variable importance across models (logistic regression and decision trees) there seemed to be agreement that ‘relationship’, ‘education’ and ‘capital.gains.yes’ were the three most important variables for predicting the target variable. When evaluating performance there are a few factors to consider. One is how much worse is a type I error compared to a type II error. The cost of type one error has a significant effect on the performance of our models and will need to be adjusted accordingly before implementation. Also note that based on the fact that a type I error is more costly it will be more beneficial to the firm to make sure the models used have a high specificity which will minimize the type I error costs at the expense of sensitivity to an extent.

The Neural Network and KNN models performed the worse in terms of total unit cost. This was due to the inability or lack of adjusting the penalty of a type one error. This lead them to having a higher sensitivity at the expense of a lower specificity which which lead to them having the most expensive total error costs. These models also do not provide any insight as to which variables are the most important. It is because of these reasons that our team does not recommend implementing them to aide in the final decision making process.

The CART decision tree also lacked the ability to adjust to the specificity of the tree which contributed to the high error cost as stated earlier. The Cart Tree also only has two splits which are at the variable ‘relationship’ and ‘education’ which can be plotted by this tree.


<=50K
0.25
100%
yes
relationship = Not-in-family,Other-relative,Own-child,Not-Married
no
<=50K
0.46
46%
education = No HS Degree,HS Grad
<=50K
0.07
54%
<=50K
0.34
32%
>50K
0.74
14%

A quick interpretation of this tree is that the decimals represent the ratio of earners above <=50K in that decision node or leaf. The percentage represents percentage of the dataset that redside at that split or node.The CART decision tree provides some insight into confirming the importance of 'relationship' and 'education' but lacks in its specifictivity and thus is not recommended to be used in the implementation stage.

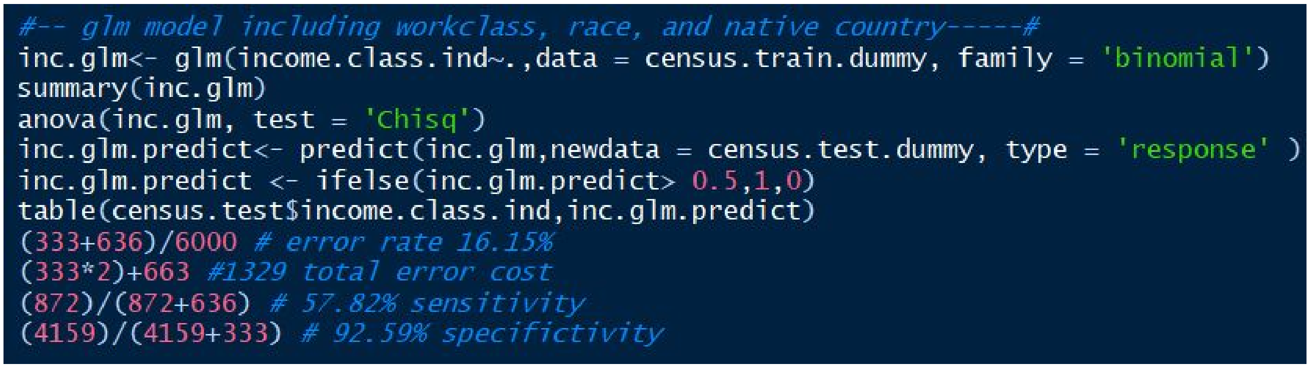
The models that our team found preformed the best were logistic regression and the C5.0 decision tree. The major feature that these models had over the other models was the ability to adjust the results to increase specificity. This allowed these models to reduce their type 1 error rate which our team deemed costly than type 2. Of course, this came at a cost of reducing their sensitivity which means their effectiveness in identify true positives is reduced. This allowed for them to achieve the total lowest costs with the C5.0 decision tree having the lowest total cost of error with 1217 just barely beating logistic regression which had a total cost of error of 1225.We believe that both of the models are would be a solid choice to implement but would recommend using the logistic regression. Even though logistic regression had a higher total cost then the C5.0 decision tree, logistic regression had a sensitivity of 52.21% compared the C5.0 36.54%. This resulted in logistic regression predicting 166 more true positives or credit-extension worthy customers at the cost of 97 more false positives. Also note that positives of a successfully identified true positive should be considered as well. It will be up to management to decide the cost of each error officially and whether or not those potential extra good customers are worth the risk of potentially extending credit to more risky customers.

Appendix

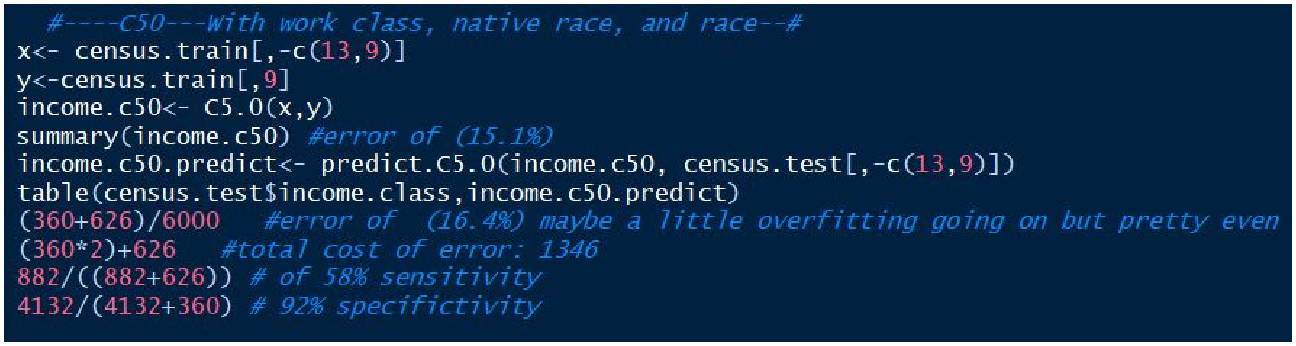
1. Logistic Regression explanation

Logistic regression outputs the probability that an input belongs to a specific class with the output being within 0 and 1. The probability between income class being greater than 50K can be written as p(X) = Pr(income.class=1|X) with X being a predictor variable. Then this probability is rewritten in linear format put into a sigmoid activation function so the output is always between zero and one p(X)= [e^(intercept coefficient(X))]/ [1+e^(intercept coefficient(X))]. Finally, p(X) is then take and used in the following log function written as log([p(X)]/(1-p(X)). There are also some assumptions that the model is based off of. The main ones are there is no multicollinearity and the predictor variable is separated into 0 and 1 (Le).

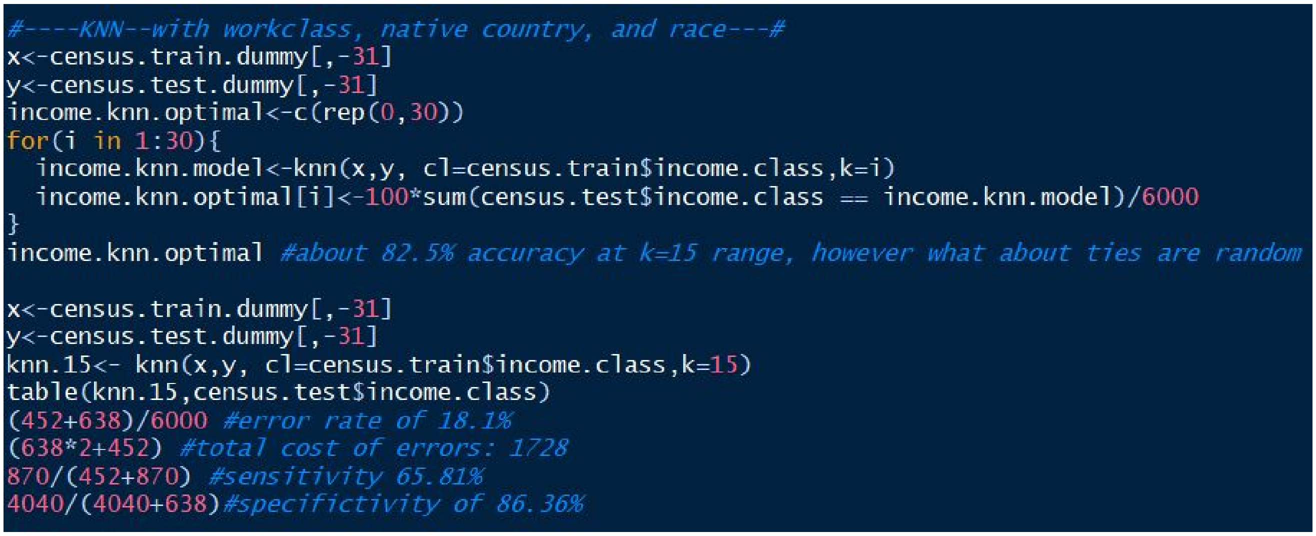
2. Model results including Workclass, Race, and Native Country



```
#-- glm model including workclass, race, and native country-----#
inc.glm<- glm(income.class.ind~.,data = census.train.dummy, family = 'binomial')
summary(inc.glm)
anova(inc.glm, test = 'Chisq')
inc.glm.predict<- predict(inc.glm,newdata = census.test.dummy, type = 'response' )
inc.glm.predict <- ifelse(inc.glm.predict> 0.5,1,0)
table(census.test$income.class.ind,inc.glm.predict)
(333+636)/6000 # error rate 16.15%
(333*2)+663 #1329 total error cost
(872)/(872+636) # 57.82% sensitivity
(4159)/(4159+333) # 92.59% specifictivity
```



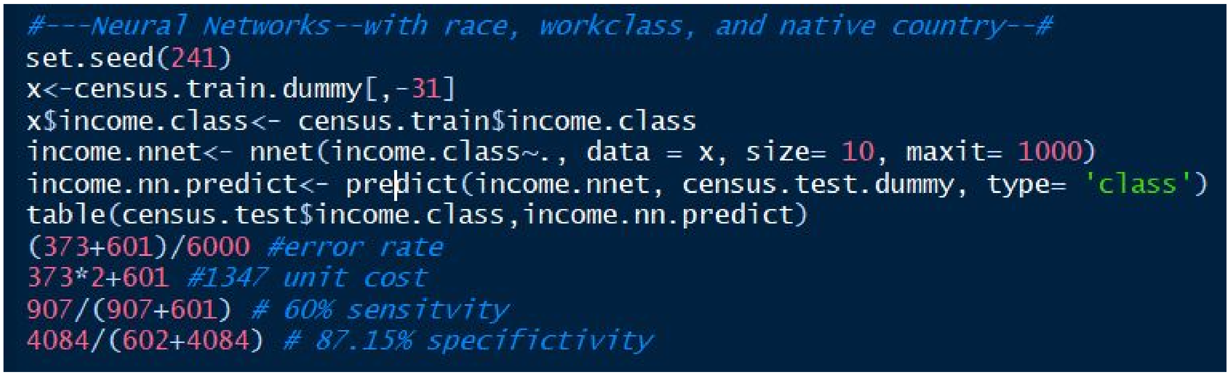
```
  #----C50---With work class, native race, and race--#
x<- census.train[,-c(13,9)]
y<-census.train[,9]
income.c50<- C5.0(x,y)
summary(income.c50) #error of (15.1%)
income.c50.predict<- predict.C5.0(income.c50, census.test[,-c(13,9)])
table(census.test$income.class,income.c50.predict)
(360+626)/6000   #error of  (16.4%) maybe a little overfitting going on but pretty even
(360*2)+626   #total cost of error: 1346
882/((882+626)) # of 58% sensitivity
4132/(4132+360) # 92% specifictivity
```



```
#----KNN--with workclass, native country, and race---#
x<-census.train.dummy[,-31]
y<-census.test.dummy[,-31]
income.knn.optimal<-c(rep(0,30))
for(i in 1:30){
  income.knn.model<-knn(x,y, cl=census.train$income.class,k=i)
  income.knn.optimal[i]<-100*sum(census.test$income.class == income.knn.model)/6000
}
income.knn.optimal #about 82.5% accuracy at k=15 range, however what about ties are random

x<-census.train.dummy[,-31]
y<-census.test.dummy[,-31]
knn.15<- knn(x,y, cl=census.train$income.class,k=15)
table(knn.15,census.test$income.class)
(452+638)/6000 #error rate of 18.1%
(638*2+452) #total cost of errors: 1728
870/(452+870) #sensitivity 65.81%
4040/(4040+638)#specifictivity of 86.36%
```

*note: this was one occurence run and if replicated would be different. Also note how including race, workclass, and native country reduce the accuracy compared to without them



```
#---Neural Networks--with race, workclass, and native country--#
set.seed(241)
x<-census.train.dummy[,-31]
x$income.class<- census.train$income.class
income.nnet<- nnet(income.class~., data = x, size= 10, maxit= 1000)
income.nn.predict<- predict(income.nnet, census.test.dummy, type= 'class')
table(census.test$income.class,income.nn.predict)
(373+601)/6000 #error rate
373*2+601 #1347 unit cost
907/(907+601) # 60% sensitvity
4084/(602+4084) # 87.15% specifictivity
```

References

CIA. (2019, April 23). The World Factbook: United States. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/us.html

Le, J. (2018, April 10). Logistic Regression in R Tutorial. Retrieved from https://www.datacamp.com/community/tutorials/logistic-regression-R