University of Arizona

MIS 545 – DATA MINING

PROJECT REPORT

Team – NEGATIVELY SKEWED

Dalvi, Vedashree
Kotiyal, Iti Shri
Sabharwal, Anmol
Sehgal, Madhvika
Singh, Viraj
Sinha, Abhilasha

Contents

INTRODUCTION:

In this project we will analyze US census data to perform a predictive classification of data.
The project has the following stages:

- Data cleaning and preprocessing
- Predictive Modelling

The final goal of the project is to build a model that can predict whether the salary of an individual in the United states is greater or less than $50000 a year. The prediction will be based on several factors like their age, occupation, gender, education, race etc.

PROBLEM STATEMENT:

There are different opinions about what are the major factor that contribute towards high earnings of people in the US, but there is no definite answer to it till now. High percentage of people feel that education has a very significant impact on the income level whereas there are others who feel capital gains can lead to higher income generation. Finding a definite answer for this will help solve the problem of problem of income inequality which has been of great concern in the recent years.

The internal factors like Marital Status of the citizen can also affect individual's salary, same way his relationship in the family and the gender. We all know about the gender bias that exists in the industry and a person who is a husband might earn more money than a person who is a wife. The race of a person might also affect the money he earns however there is a chance that it might not affect his salary at all. Since race is not something that is considered as a parameter for a job, we will still take it into account to rule out any possibility of any kind of bias that might exist for a race.

This project aims to conduct a thorough analysis to find out the factors that contribute majorly in improving an individual's income. This information can be used by the government to focus on specific areas that need improvement which will significantly improve the income levels of the people.

PRACTICAL USE OF PREDICTION ANALYSIS:

Let us suppose that the data mining results indicate that people with higher the number of years of education have income higher than $50000 a year, it means that the government needs to focus on making higher education more accessible and affordable to people so that they can earn higher wages per year.
Similarly, there might be certain public or private sectors where people are underpaid, government can focus on increasing the pay scale of those sectors.

DATA DESCRIPTION:

U.S. census data is taken from the UCI (University of California at Irvine) Machine Learning Repository. The dataset includes 32560 records and 15 attributes. Attributes are divided into 14 independent variables and 1 dependent variable.

The 14 independent variables comprise of 8 categorical and 6 continuous variables that are explained below.

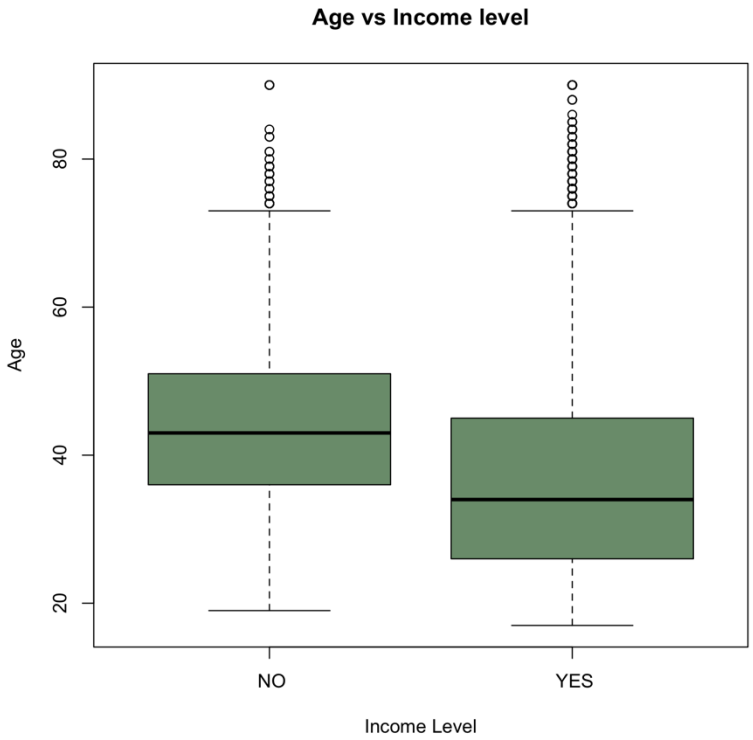
ATTRIBUTE	TYPE	DESCRIPTION	LEVELS
AGE	Continuous	This attribute specifies the age of each adult.	NA
WORKCLASS	Categorical	It specifies the employment status of the individual.	9
FNLWGT	Continuous	This is the number of adults that represent that row.	NA
EDUCATION	Categorical	It specifies the education level last reached by the adult.	16
MARITAL_STATUS	Categorical	Specifies the marital status of adult.	7
EDUCATION_NUM	Continuous	Specifies the number of years of education completed.	NA
OCCUPATION	Categorical	Identifies the occupation type that the adult works in.	15
RELATIONSHIP_FAMILY	Categorical	This explains the individual's primary relationship in his family.	6
RACE	Categorical	Defines the race of the individual.	5
SEX	Categorical	Specifies the gender of the adult.	2
CAPITAL_GAIN	Continuous	Capital gain species the profit gained from the sale of a property or investment	NA
CAPITAL_LOSS	Continuous	Capital loss specifies the loss obtained from the sale of a property or investment	NA
HOURS_PER_WEEK	Continuous	Mentions the number of hours the individual works per week	NA
NATIVE_COUNTRY	Categorical	This attribute names the country of origin of the individual.	41
INCOME	Categorical	This attribute is for income prediction	2

DESCRIPTIVE STATISTICS FOR CONTINUOUS VARIABLES:

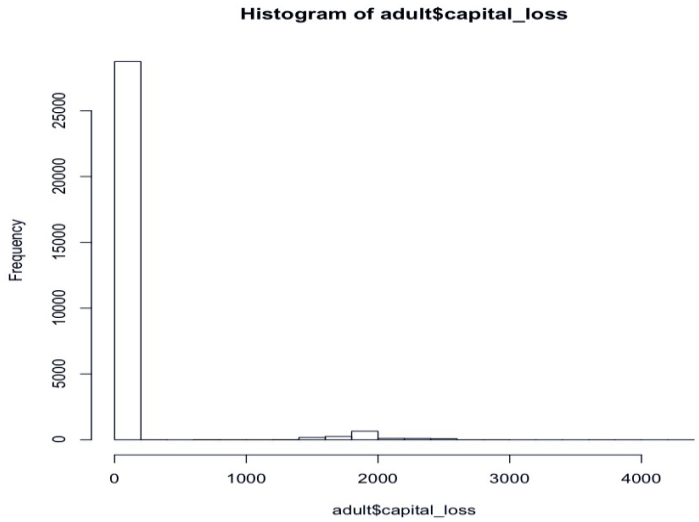
ATTRIBUTE	MEAN	MEDIAN	STANDARD DEVIATION	SKEWNESS	MIN	MAX
AGE	38.581	37	13.64	0.55	17	90
HOURS_PER_WEEK	40.43	401	12.34	0.22	1	99
EDUCATION_NUM	10.08	10	2.58	-0.31	1	16
CAPITAL_GAIN	1077	0	7385.29	11.95	0	99999
CAPITAL_LOSS	87.30	0	402.96	4.59	0	4356

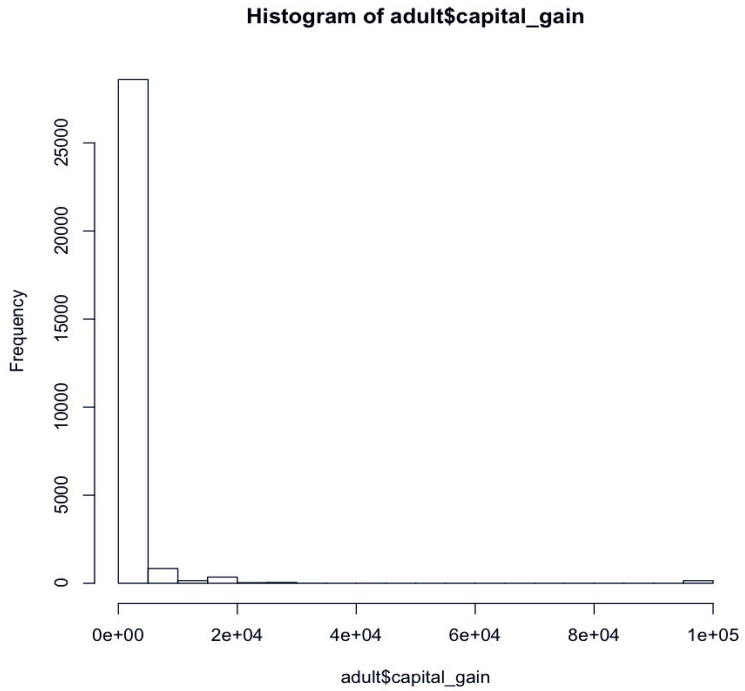
PLOTS FOR DATA BEFORE PREPROCESSING:

- When we plotted Age vs Income level, we found that there were many outliers and it indicated that we should perform statistical data binning which means converting this continuous variable to smaller number of categories.

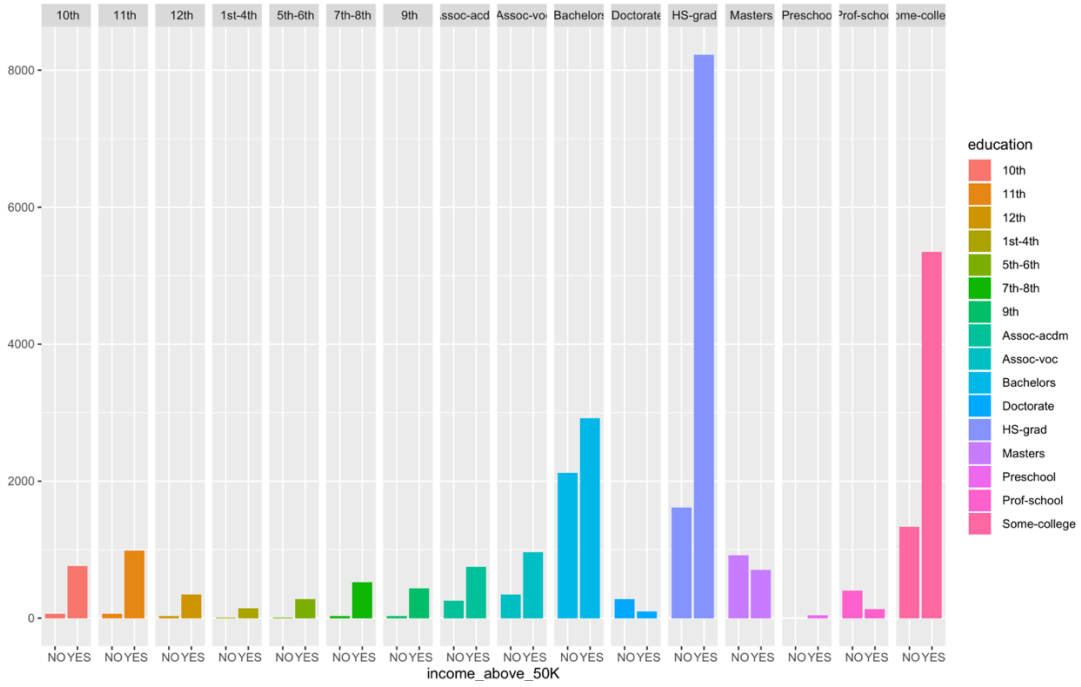


- We plotted histograms for capital gain and capital loss vs Income level and discovered that the data was right skewed, so we performed mathematical operations to reduce the skewness.





- We plotted Bar graphs of education level vs Income and we saw that there were too many levels for the education level which might impact our accuracy, so we reduced the number of levels in the preprocessing stage.



DATA PREPROCESSING:

As we discussed in the previous section there were some issues with the dataset which we fixed by preprocessing activities that are explained below.

NEED FOR PREPROCESSING:

1. The dataset contains some null values that we should remove.
2. The levels of categorical variables are high for most variables and this can lead to unnecessary complication in data analysis and might cause overfitting which means we need to create lesser levels of these variables for easy interpretation and better results.
3. The continuous variables have a lot of outliers which is why we decided to group these values into categories and reduce the number of outliers.
4. Attributes like capital gain and capital loss had anomalies like large number of zero values which led to high skewness.
5. The data set was not balanced, the number of rows with income less than $50K was 24.08% whereas the number of rows with income greater than $50K was 75.91%

PREPROCESSING ACTIVITIES AND RESULTS:

- The dataset contains some null values which were removed when the dataset was imported, the number of records reduced from 32,560 to 30,161 Column names were assigned to each column as the predefined column names were junk values.

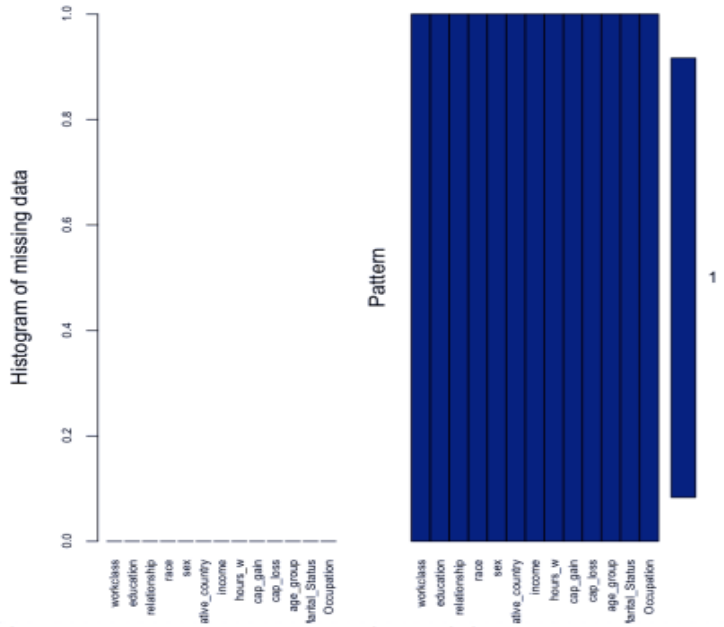
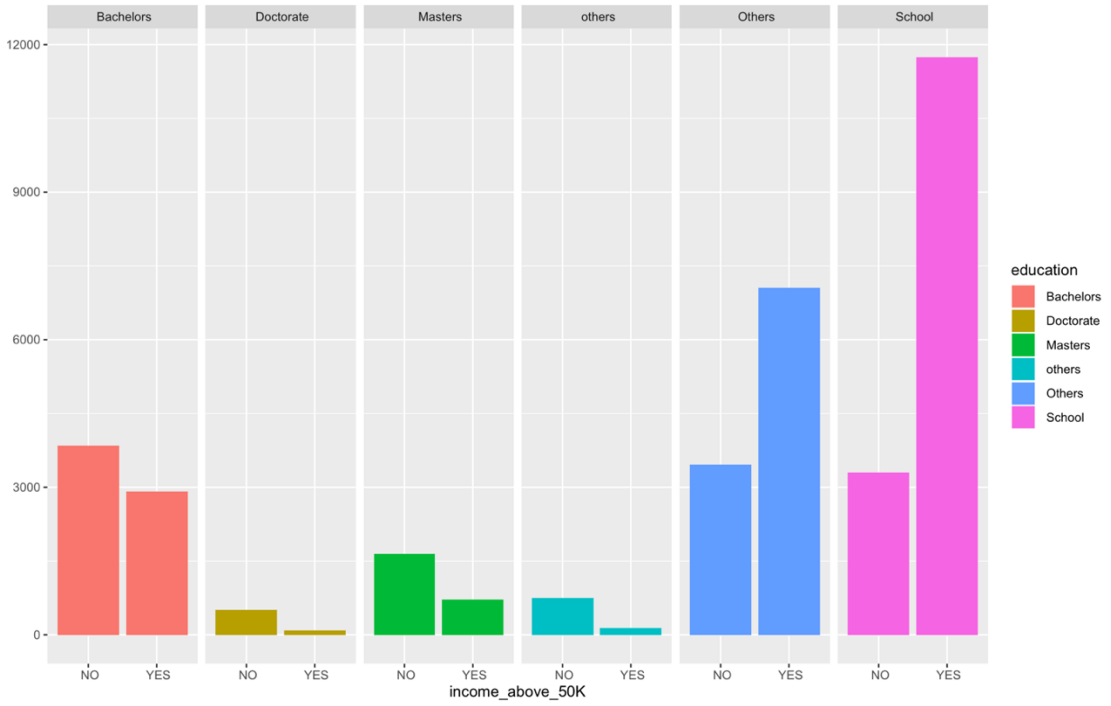
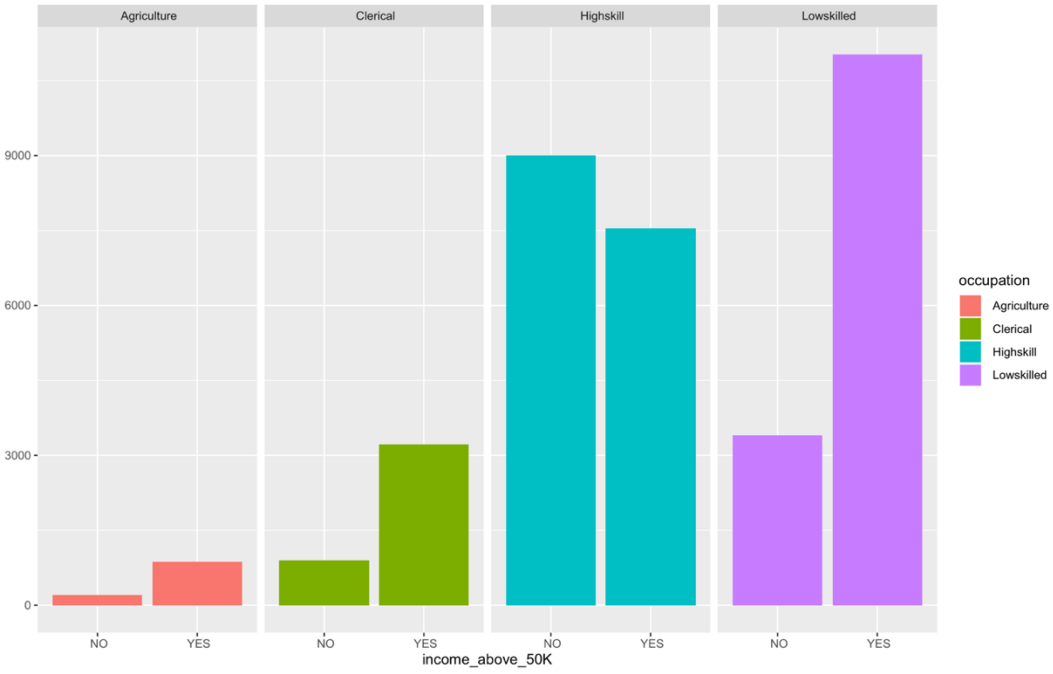


Fig1-Summary of dataset and graph showing no missing values.

- We reduced levels of categorical variable like Education, marital status and Occupation.

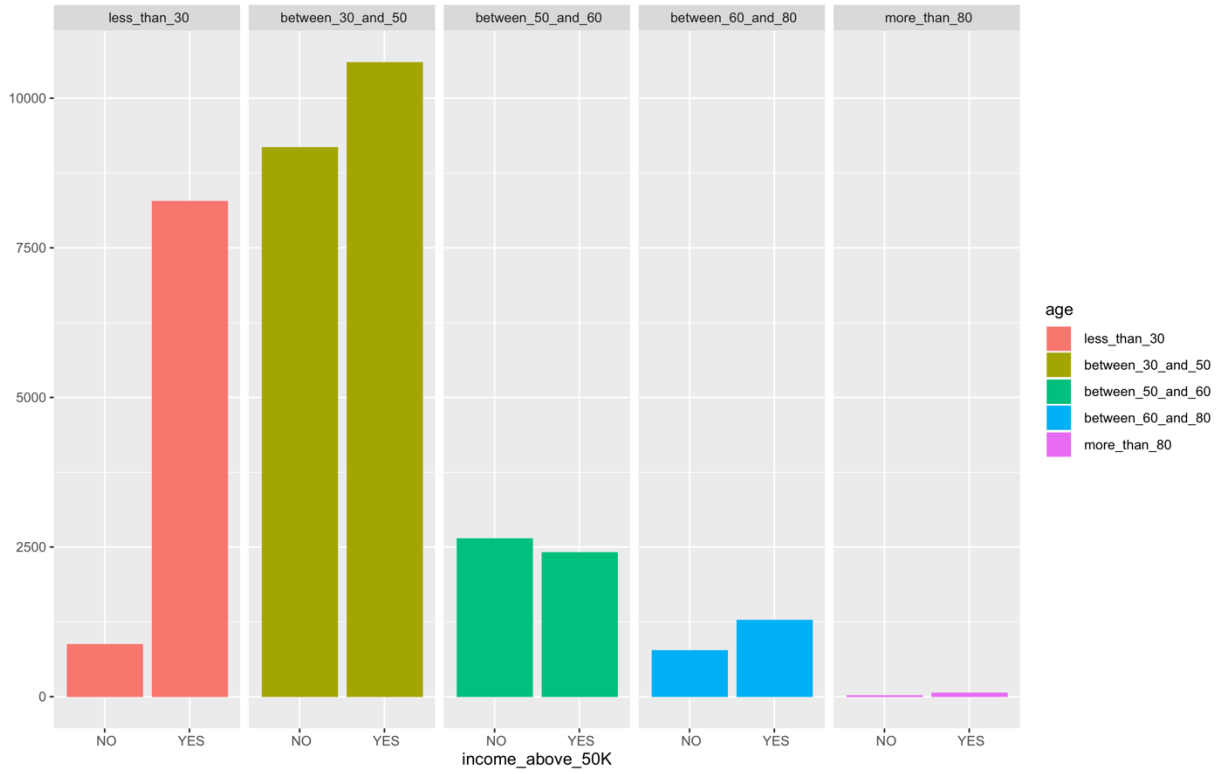
The Bar graphs of few variables after preprocessing are shown below:



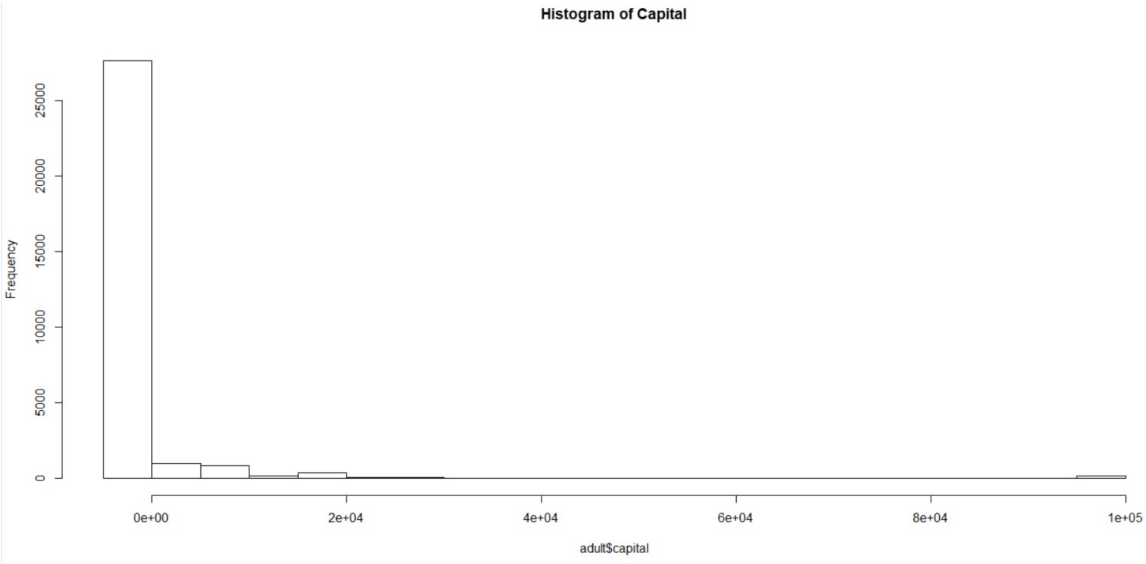


- The continuous variable age was binned to smaller number of categories as shown below, we divided the value of age in to categories like :
  - Less than 30

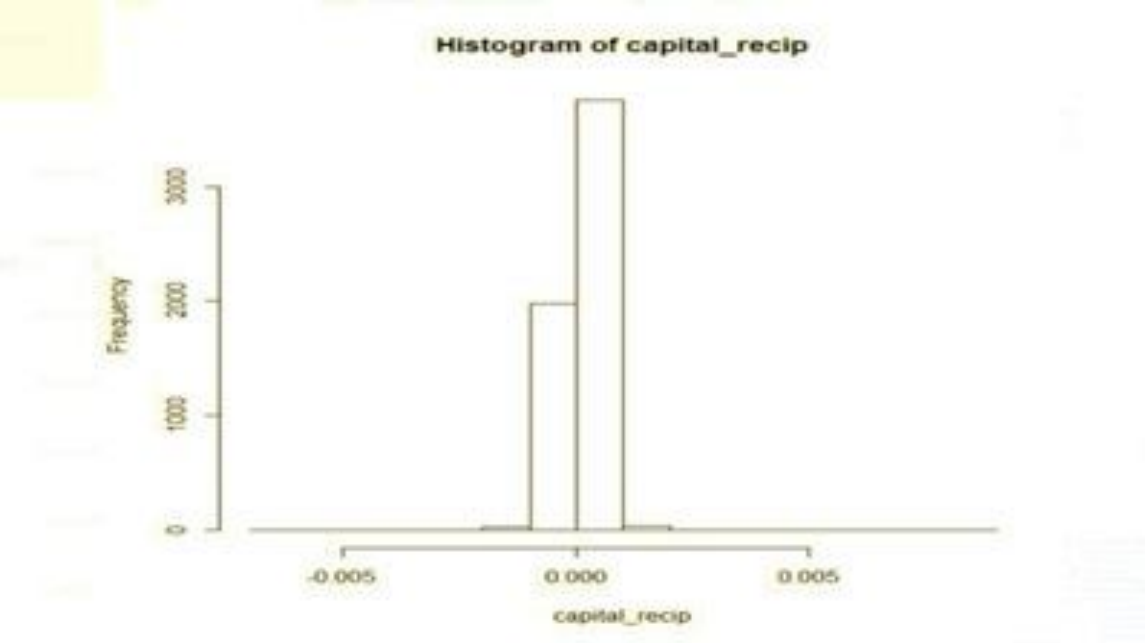
- Between 30 and 50
- Between 50 and 60
- Between 60 and 80
- More than 80



- To deal with skewness in variables capital gain and capital loss we first combined them into variable 'capital' which is the difference of Capital gain and capital loss.



- The result we got after combining both variables to one was also right skewed so then we took reciprocal of this variable to remove the skewness of the variable.



- The number of Yes and No in the dataset were not balanced so we performed oversampling with replacement to create balance in the number of Yes and No. Oversampling increased the number of rows in the dataset from 30161 to 36168.

VARIABLE LEVELS AFTER PREPROCESSING:

Variable name	Variable type	Number of levels in the original dataset	Number of levels after level reduction
age	continuous	NA	5
education	categorical	16	6
marital_status	categorical	7	5
occupation	categorical	14	4
hours_per_week	continuous	NA	5
native_country	categorical	41	8

INCORPORATION OF PREVIOUS ROUNDS OF FEEDBACK:

As mentioned above, we used predictive algorithms, such as Support Vector Machine, Neural Networks, Naïve Bayes and C5.0 Decision Tree in the previous stages predicting whether the annual income of a person would be less or more than $50,000 per year using the features like age, education level, number of years of education, marital status etc. We obtained variant accuracies for the above algorithms for a random set of features.

To improvise on our results, we were given the below suggestions to get a more accurate result from using the algorithms:

1) **Perform Feature Selection**: To select those features which contribute most to our prediction variable or output instead of using random characteristics from the data set.

For feature selection, we used the random forest ensemble algorithm. This is typically used for running prediction over predictive algorithms and favoring the features that contribute most to the data set. This algorithm is run over a random set of training data which then aggregates the votes from different decision trees to decide the final class of the test object.

After performing feature selection on a training data set out of our original data, we found the attributes that had the most weight and their importance:

Attribute	Importance (in %)
relationship	5.12
marital_status	3.35
capital_gain	2.79
education_num	2.31
occupation	1.92
education	1.87
age	1.36
sex	0.72
capital_loss	0.64
hours_per_week	0.63
workclass	0.61
race	0.09
native_country	0.09

2) **Perform Oversampling**: To tackle the imbalanced data with a larger number ‘Yes’ with an aim to increase the number of instances from the underrepresented class in the data set which is ‘No’.

The accuracy was almost the same after using this oversampled data set. The problem here is that accuracy is not a good measure of performance on unbalanced classes. It may be that our data is too difficult, or the capacity of our classifier is not strong enough. Since we did not get any significant increase in our accuracy, we discarded the over-sampled data set.
We could have also tried under-sampling, but we did not have the luxury to work on a data set with lesser number of rows.

3) **Perform Correlation Tests**: Correlation coefficients are used to measure the strength of the relationship between two variables. Positive correlation is a relationship between two variables in which both variables move in tandem—that is, in the same direction.
The correlation tests are performed on the columns, such as Age, Number of Education Years etc. We were suggested to perform correlation tests between the variables: "Education Class" for example, School, Bachelors, Masters and Doctorate and "Number of Education Years". This was suggested to check if both gave the same results, then either of them could be redundant.
Since the "Education Level" is a categorical variable, we could not perform the correlation tests.

RESULTS OF INDIVIDUAL PROJECTS OF TEAM MEMBERS(STAGE 2):

Algorithm	**Preprocessing Done Y/N?**	**Feature Selection Done Y/N?**	**Number of Attributes used after feature selection**	**Accuracy**	**Error Rate**
Decision Tree (C50)	Yes	Yes	10	84.5%	15.5%
SVM	Yes	Yes	10	83.9%	16.1%
Neural Network	Yes	Yes	10	83.5%	16.5%
Naïve Bayes	Yes	Yes	10	78.61%	21.39%

The results from stage two of the project helped us understand the following :

- SVM gives a good accuracy for our dataset but it cannot be used because the variables are dependent on each other. For example, years of education and education level are attributes that have strong correlation, so we decided to not consider Naïve Bayes for the third stage of the project.

- We also noticed that even though we increased the hidden layers of Neural network still the accuracy could not be as good as the accuracy achieved using Decision Tree C50 algorithm. So, we decided to move ahead with C50 algorithm for stage 3.
- We tried to increase the accuracy of Neural network after oversampling, but it did not increase the accuracy of the algorithm.
- We ran neural networks with and without feature selection, meaning we ran the algorithm with few selected variables and ran it using all the variables. We discovered that accuracy was better when we used fewer variables, so we understood we need to run feature selection to find the most important attributes.

ALGORITHMS USED AND RATIONALE:

Based on the results of stage 2, we will be using decision tree C50 for the following reasons.

1)Best Accuracy

2)Least computational time

3)Maximum Sensitivity

We choose 80% of the data randomly and build a training set and the rest 20% is the testing set. We select the attributes capital, relationship, education, occupation, age, workclass, sex, marital_status based on the result we got from random forest.

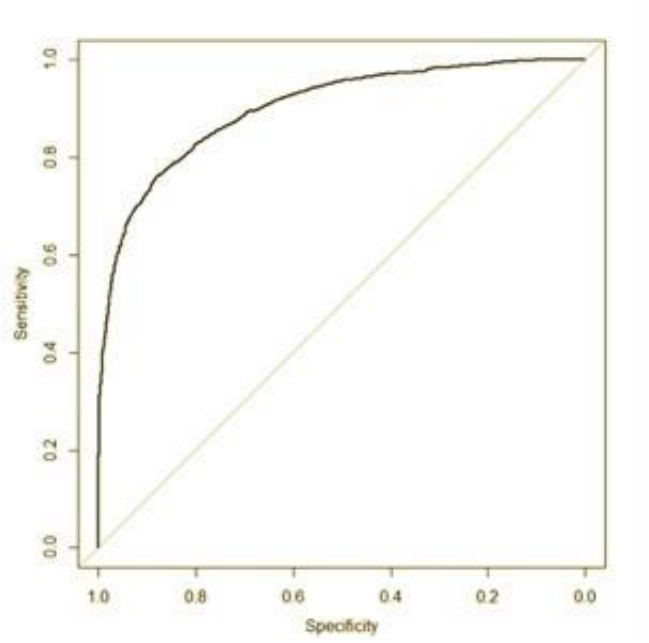
We exclude the dependent variable 'income_above_50K'

```
Evaluation on training data (24128 cases):

            Decision Tree
          ----------------
          Size      Errors

            43 3263(13.5%)   <<


           (a)   (b)    <-classified as
          ----  ----
          3766  2272    (a): class NO
           991 17099    (b): class YES


        Attribute usage:

        100.00% capital
         95.61% relationship
         40.68% education
         37.87% occupation
         18.47% age
         11.77% workclass


Time: 0.0 secs
```

We observe that after running the algorithm, the accuracy increased because of the extra preprocessing that was done on the dataset in this stage. Following are the other performance parameters.

Measure	Value	Derivations
Sensitivity	0.9395	TPR = TP / (TP + FN)
Specificity	0.6231	SPC = TN / (FP + TN)
Precision	0.8856	PPV = TP / (TP + FP)
Negative Predictive Value	0.7685	NPV = TN / (TN + FN)
False Positive Rate	0.3769	FPR = FP / (FP + TN)
False Discovery Rate	0.1144	FDR = FP / (FP + TP)
False Negative Rate	0.0605	FNR = FN / (FN + TP)
Accuracy	0.8624	ACC = (TP + TN) / (P + N)
F1 Score	0.9117	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.6066	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

We plot the ROC curve and the area under the curve is 90.11. A strong ROC curve tells us that the model can differentiate between the people who have annual income above 50K and people who don’t have income above 50k.



RESULTS AND INTERPRETATION:

Real world data tend to be incomplete and inconsistent. In order to remove the missing values and smooth the noise while removing the outliers, data preprocessing was performed.

- It was observed that the data set contained some missing (“ NA”) values, this was done using ‘na.strings’ function.
- Also, it was also observed that the large number of values in the attributes reduces the accuracy of the algorithm. Hence, data preprocessing was performed to reduce the number of levels in both categorical and continuous variables. This also reduced the number of outliers in our dataset.

After the preprocessing stage, the feature selection step was performed. The purpose of feature selection is to choose the variables that are useful in predicting the response. The main objective of this step is to find the most important predictor variables (or independent variables) and explain the major variance on the dependent variables.

For the purpose of performing feature selection on the selected dataset, Random forest Method is used. This is a method for classification and regression. It works by constructing numerous decision trees (in this case 500) at training time and hence, resulting into those classes or mean prediction of individual trees.

The package used for Random forest algorithm is “party”. cforest() refers to a function for this algorithm which is a default type of forest. This gives us those attributes

After performing feature selection through this method, we choose those attributes that might significantly affect our prediction. This included relationship, marital_status, capital_gain, education_num, etc.

Now we subset our data into 80% data for training and 20% data for testing our trained model. The final step of this project was to run this data on various algorithms. These algorithms include, Decision Tree, Naives Bayes, Neural Network and SVM. As a result of this step, it was observed that we get the best accuracy from the Decision tree algorithm.

Oversampling of the dataset was also performed. But it was observed that the accuracy of the dataset did not improve instead it got reduced from 84.5% to 84.3%. Hence, we decided to do the prediction without oversampling.

Inferring the results, we see that most of the variables have a good correlation with the dependent variable and have significant impact on the results of the accuracy of the model when removed from the predictors.

To conclude – We would like to mention that the data pre-processing by reducing the levels of categorical variables and feature selection showed us the impact, independent variables have on the dependent variable and how the model is able to predict an accurate result. The most accurate results are generated when we train our sampled data with Decision tree with all predictors and hence, decided to show our final prediction with the Decision Tree algorithm.

RECOMMENDATIONS:

1. During the process of feature selection, it was observed that 'education_num' is one of the top 5 factors that determine an individual's salary. Hence, we recommend the Government to focus more on improving the education system within the country.
   To increase the earning capacity of individuals government should make education more affordable and accessible to everyone.

2. Also, occupation was another major deciding factor that affects an individual's salary. So, we recommend Governments should work towards increasing the number of public and private sector industries. In order to maintain a fair and just society, Governments should work with public and private-sector businesses to accomplish a fair balance in terms of wages paid to the employees.

3. Relationship was a major factor that affected our prediction. This signifies that an individual's role in the family has a significant impact their income. There are chances that in a family husband has a higher income than his wife who has the same educational qualification. These findings can be justified by the gender bias that exists in the society and we recommend that governments should take actions to ensure that wages offered to any individual should be free from any gender bias and be based purely on talent.

EVALUATION:

We started with the raw dataset where we found a lot of missing values and the names of the columns were also not correct. The first step that we did was to get rid of the null values and name the columns appropriately. We moved ahead, with limited knowledge of the concepts we used all the attributes and C50 algorithm in the first stage. After the first presentation and the feedback that we got, we tried to implement most of them in the second stage. While working separately in the second stage, we noticed a lot of problems with our database. One of the mistakes that we did in the first stage was to choose all variables instead of running a correlation test or feature selection algorithm. All of us used one or the other feature selection algorithm and results improved. We also noticed that our variables had too many levels with the highest levels being 41. We performed pre-processing activities and reduced the number of levels for most of the variables. We had two important variables capital_gain and capital_loss that were affecting the result by a lot but they were both right-skewed so we calculated the net gain or net loss and took the reciprocal of the final value which was also right-

skewed, to remove skewness. 'Age' being a continuous variable had to be binned into groups. Now that most of the attributes were categorical, we were not able to run correlation and hence some variables like education and education_num which by definition were very similar, we concluded that they are correlated. The same correlation can be observed between marital_status and relationship. These correlations between variables forced us not to use Naive Bayes for which the assumption is that we need the variables to be independent. The next problem we thought could affect our model performance was the skewness of the dependent variable. We had 76% of people who were earning more than 50K and only 24% of people earning less than 50K, we corrected this imbalance by performing oversampling with replacement. We again ran the algorithm, since the number of records increased the computation time increased by a bit and the overall accuracy went down. We figured that the time spent to perform oversampling was producing results that were not worth the time spent and hence we decided to get rid of oversampling as we wanted our model to be effective in terms of both accuracy and computational time. In the final stage, after performing all these steps the accuracy, sensitivity and computational time improved and the model became more effective.

Finally, our recommendation is that in order to bridge the income gap in the country governments should focus on improving the education system, improving the pay scales or public and private sector and lastly ensure that there is no gender wage gap existing in the country. Governments should implement minimum income policies to ensure a decent standard of living to every individual.

REFERENCES:

http://rstudio-pubs-static.s3.amazonaws.com/265200_a8d21a65d3d34b979c5aafb0de10c221.html#1_introduction

https://arxiv.org/pdf/1810.10076.pdf

https://www.investopedia.com/terms/e/economic-forecasting.asp

http://www.dataminingmasters.com/uploads/studentProjects/Earning_potential_report.pdf

https://cloudxlab.com/blog/predicting-income-level-case-study-r/

https://d2l.arizona.edu/d2l/le/content/820111/viewContent/7685221/View

https://d2l.arizona.edu/d2l/le/content/820111/viewContent/7685246/View