Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Group Number: 3

Assignment Title: Group Assignment 3: ANN

Course Code: RSM8413

Instructor Name: Gerhard Trippen

In submitting this **group** work for grading, we confirm:

- That the work is original, and due credit is given to others where appropriate.
- That all members have contributed substantially and proportionally to each group assignment.
- That all members have sufficient familiarity with the entire contents of the group assignment so as to be able to sign off on them as original work.
- Acceptance and acknowledgement that assignments found to be plagiarized in any way will be subject to sanctions under the University's Code of Behaviour on Academic Matters.

Please **check the box and record your student number** below to indicate that you have read and abide by the statements above:

☐ 1006527741
☐ 1002140541
☐ 1006507512
☐ 1004654527
☐ 1006604934
☐ 1005605374

Assignments are to be submitted using Student ID Numbers *only*; do not include your name. Assignments that include names or that do not have the box above checked **will not be graded.**

Please pay attention to Course Outline for specific formatting requirements set by instructors.

If submitting this assignment online please use the following "Standard File Naming Convention":

Full Course Code *(including Section)*–Group Name or Number–Assignment Title

Example: RSM1234HS.2016-0101-Group1-Homework1

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The overarching goal of the analysis is to generate an Artificial Neural Network (ANN) that can be used to predict whether or not an individual has an income of over $50000 per year. The analysis will make use of US Census Bureau data. As analysts we want to identify the demographic characteristics that are best suited for this prediction. For each observation in the US Census Bureau training data, there are fourteen given predictors such as age and work-class, and the target variable - income.

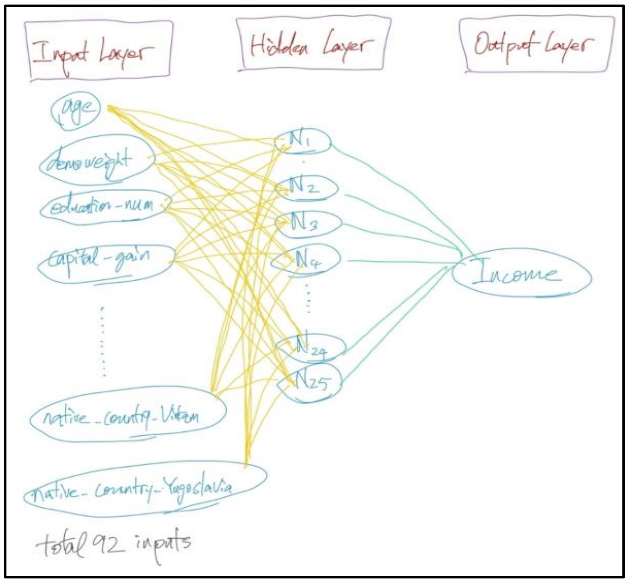
On the training set, we achieved a maximum prediction accuracy of 84.848% with 25 neurons which was selected by conducting cross validation on the data set. The most important predictors in the ANN were *"education-num,"* which is the number of years of education, and *"capital-gain."* In part two, when we graduated to the big leagues of data mining, we split the training data into training and validation datasets. We achieved an accuracy rate of 85.06% with 29 neurons on validation set.

2. NEURAL NETWORK

i. Description of Topology

The topology of the resulting neural network is described below as follows:
One input layer containing 92 predictors, one hidden layer having 25 neurons (*note: we initially began with 12 neurons, however the results of our Grid Search CV, which are discussed in a later section, indicated that 25 is the optimal number of neurons)* and last but not the least one output later.

ii. Illustration



3. DATA PREPROCESSING

The first phase of our analysis was data preprocessing which included two crucial steps. Firstly, we explored the data to determine variables that were best suited for our analysis. Secondly, we performed variable transformation. The details of each step are highlighted below.

i. Variable Selection

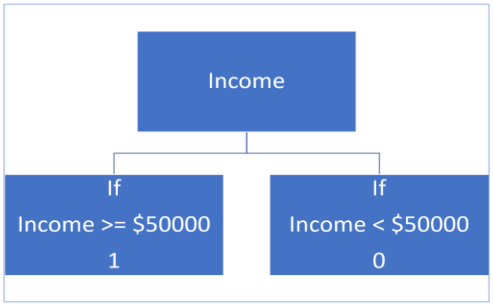
After thorough exploratory data analysis we decided to exclude the variable 'education' from our model. This is due to the fact that 'education' seems to be highly correlated with another variable 'education-num'. 'Education' is a categorical variable that described the level of education attained (such as high school or college), and 'education-num' is continuous showcasing the number of years of education. Correlated variables should generally not be included together in ANN models, as they introduce redundancies and unnecessary complexities into the model. Thus, for this reason, we decided to eliminate one of them that is 'education' variate from our model. The rest of the variables were included in the ANN.

ii. Variable Transformation

The US Census Bureau dataset contained seven categorical variables namely work-class, marital status, relationship, sex, race, native-country, and occupation. In order to make these predictors useful in our model, we transformed each categorical variables into dummy variables. Once the dummy variables were created for each response within the aforementioned categorical variables, we dropped the original predictors. Further, we also created a binary variable for our outcome variable, income.

The following diagram shows the variable transformation of income.

Exhibit 1: Variable Transformation



For continuous variables, we used a min-max scaler to transform the variables. This is required for ANN, as all the inputs must be encoded in a standardized manner.

iii. Handling Missing Values

As data from the US Census Bureau was in raw format, it contained numerous missing values. The missing values were recorded as ***'?'*** in the data set. After extensive research and discussion, we decided to keep the missing variables in our model, by creating a ***'?'*** dummy variable for each predictor that contained this missing value. Our basis for selecting this approach was due to the following reasons. Firstly, the missing values accounted for approximately 6% of the total dataset. As more data is almost always better, there were too many missing observations to be simply deleted from the dataset. Secondly, an alternative method considered was to use a random forest to predict the missing values. However, an immediate cause of concern was that there were a substantial amount of missing values for numerous predictors, such as occupation and work class. This is concerning because we would be predicting values, based on predicted values. As a result, we were doubtful that this method would produce accurate predictions. It is also highly probable that this methodology would introduce inherent bias in the data set. Therefore we decided to abandon this solution. Further, the test data set also includes ***'?'*** missing values, so if we imputed these values in the training set we would have to be consistent and do the same on the test set. This would end up adding additional complexity and is impractical for classifying new observations. Thus, for these reasons we decided that the best approach is to keep the missing values in the data set by creating ***'?'*** dummy variables.

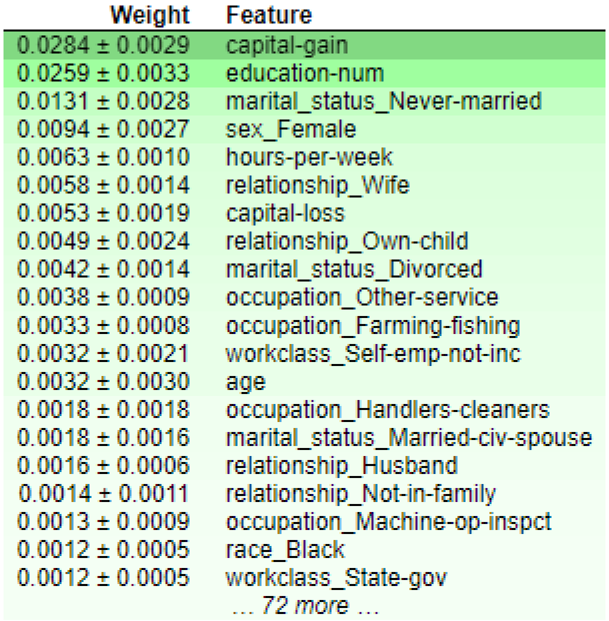
iv. Handling Outliers

When performing exploratory data analysis, we came across numerous outliers in the 'capital-gain' predictor. These outliers were recorded as '99999,' and there were 126 of these in total within the training data set. We decided to remove the 126 observations that included these outliers as they significantly affect the model's accuracy. After removing the outliers in both the training and test datasets, we were left with 24874 observations in training dataset.

4. MODEL DETAILS

i. Important Predictors

We used *eli5 package* to calculate the weights of each input variable. Higher weight means the variable is more important. As shown in the following table, the top 2 most important predictions are capital-gain and education-num with weights 0.0284 and 0.0259 and standard deviation of 0.0029 and 0.0033, respectively.



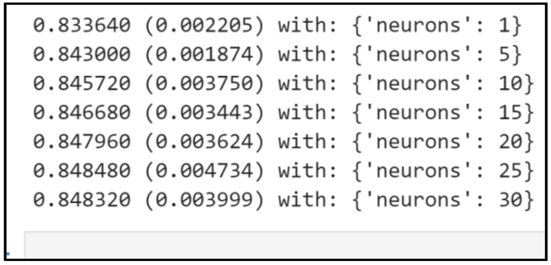
Weight	Feature
0.0284 ± 0.0029	capital-gain
0.0259 ± 0.0033	education-num
0.0131 ± 0.0028	marital_status_Never-married
0.0094 ± 0.0027	sex_Female
0.0063 ± 0.0010	hours-per-week
0.0058 ± 0.0014	relationship_Wife
0.0053 ± 0.0019	capital-loss
0.0049 ± 0.0024	relationship_Own-child
0.0042 ± 0.0014	marital_status_Divorced
0.0038 ± 0.0009	occupation_Other-service
0.0033 ± 0.0008	occupation_Farming-fishing
0.0032 ± 0.0021	workclass_Self-emp-not-inc
0.0032 ± 0.0030	age
0.0018 ± 0.0018	occupation_Handlers-cleaners
0.0018 ± 0.0016	marital_status_Married-civ-spouse
0.0016 ± 0.0006	relationship_Husband
0.0014 ± 0.0011	relationship_Not-in-family
0.0013 ± 0.0009	occupation_Machine-op-inspct
0.0012 ± 0.0005	race_Black
0.0012 ± 0.0005	workclass_State-gov
	… 72 more …

ii. Prediction Accuracy

The prediction accuracy on the full training data set is 84.848% with 25 neurons. Below we describe the process of determining the prediction accuracy. In the case of ANN, prediction accuracy represents the ability of the model to correctly classify the outcome of an observation. In this scenario, prediction accuracy refers to model correctly classifying an observation as 0 (individual makes <$50000) or 1 (individual makes >=$50000).

1. Fit the model on the training data set
2. Set the number of neurons (grid search parameters) to 1,5,10,15,20,25,30
3. Used Grid Search CV to determine the highest prediction accuracy for each of the specified values of neurons
4. The highest prediction accuracy, 84.848%, is when neurons=25

Exhibit: Prediction Accuracy using Grid Search



```
0.833640 (0.002205) with: {'neurons': 1}
0.843000 (0.001874) with: {'neurons': 5}
0.845720 (0.003750) with: {'neurons': 10}
0.846680 (0.003443) with: {'neurons': 15}
0.847960 (0.003624) with: {'neurons': 20}
0.848480 (0.004734) with: {'neurons': 25}
0.848320 (0.003999) with: {'neurons': 30}
```

iii. Discussion of Model Errors

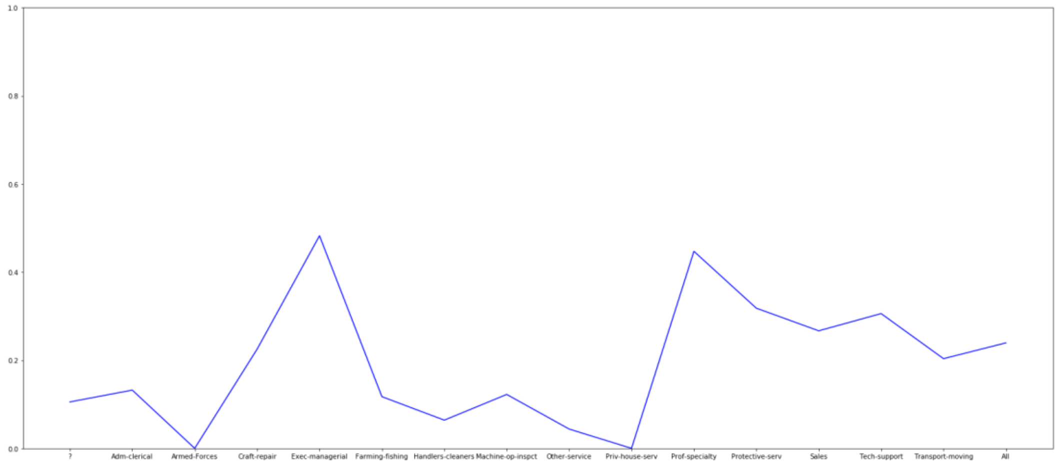
The following table shows the comparison between the predicted output and actual output. The output value is compared to the actual value of the target variable for this training set observation, and the resulting error (actual – output) is 14.86% (Based on (FN+FP)/Total). Furthermore, there are two types of error the model makes. The errors are: Type I error (something doesn't exist, but we predicted that something does) and Type II (something exists, but we didn't predict it). The Type I error we get from this model is 929 (FP)/19016 = 4.89%, and the Type II error is 2768 (FN)/ 5858 = 47.25%. Since the Type II error is way higher than the Type I error, this model tends to make Type II error more frequently.

This does not imply that the model is poor, rather the threshold for Type I and Type II errors depends on the business context. To contextualize Type I and Type II errors, let's apply the model errors to a bank loan application. In this scenario, an analyst is trying to predict whether or not an individual will default on their loan. Here, 4.89% ratio (Type I error) represents the error corresponding to wrong prediction that an individual will default on their loan, when they actually don't. Contrary, the 47.25% error corresponds to wrongly prediction that an individual doesn't default on their loan, when actually they do. In this application, the errors don't align with the business context. As a loan officer or bank, you want to mitigate the risk of Type II errors. This is because the bank doesn't want to be accountable for individuals defaulting on their loan. Thus in this context, we would want to adapt the model to minimize the risk of Type II errors. However, here, the errors that model are more prone to are justifiable.

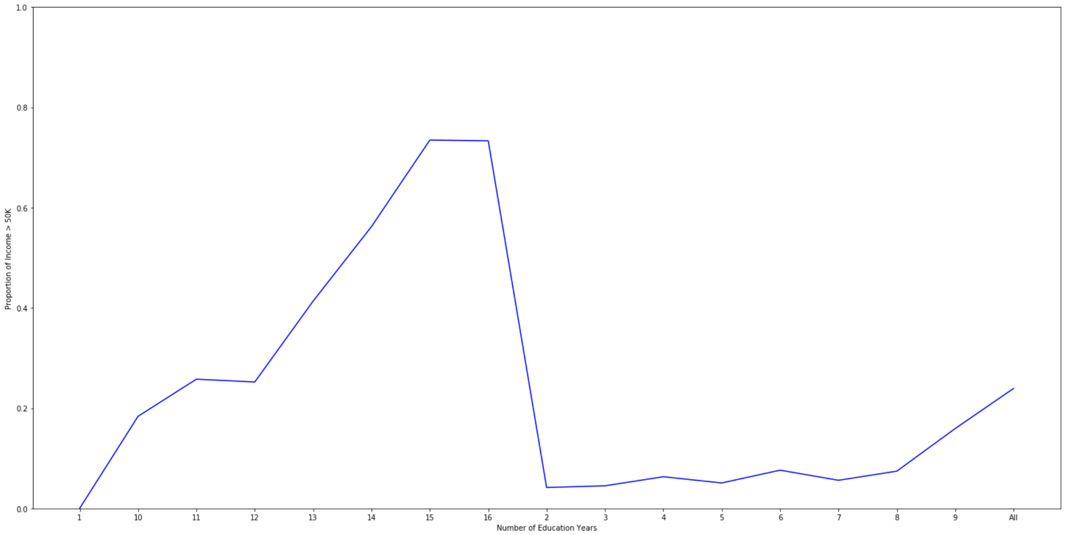
		Predicted Y		
Actual Y		**0**	**1**	**Total**
	0	18087 (TN)	929 (FP)	19016
	1	2768 (FN)	3090 (TP)	5858
	Total	20855	4019	24874

iv. An Analysis of Demographic Predictors

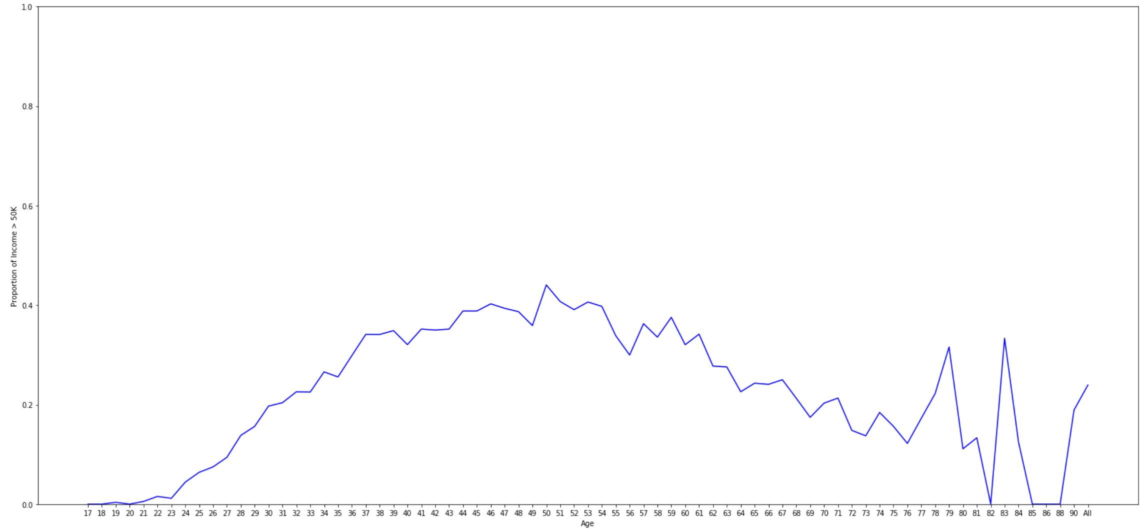
We see that almost all professions have cases of income >= $50,000 except for Armed forces and House servents. Exec-managerial and Prof-speciality have the highest frequency of occupations having salaries greater than $50,000. On plotting the frequency chart for age vs income >= 50,000, we see people in the age range of 23 - 82 have incomes greater than 50k. Frequency of these people monotonically increase from 23 years and then gradually decreases after around 50 years. Further, the number of people with income greater than 50k and education of 15 years or more is the highest. All the graphs are intuitive and relevant to our expectations.



Occupation vs Income >= $50,000



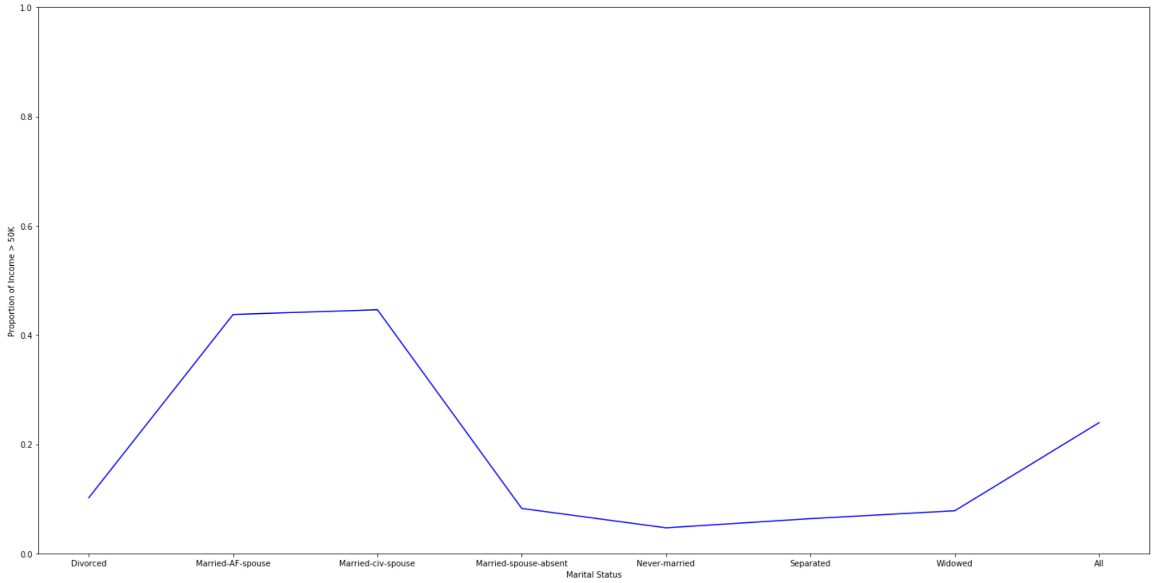
Number of Education Years vs Income >= $50,000



Age vs Income >= $50,000

The top 3 categorical predictor graphs are:

1. Education-num (graph above)
2. Occupation (graph above)
3. Marital status: Marital statuses with highest number of people having incomes greater than 50k are Married-civ-spouse and then Married-AF-spouse.

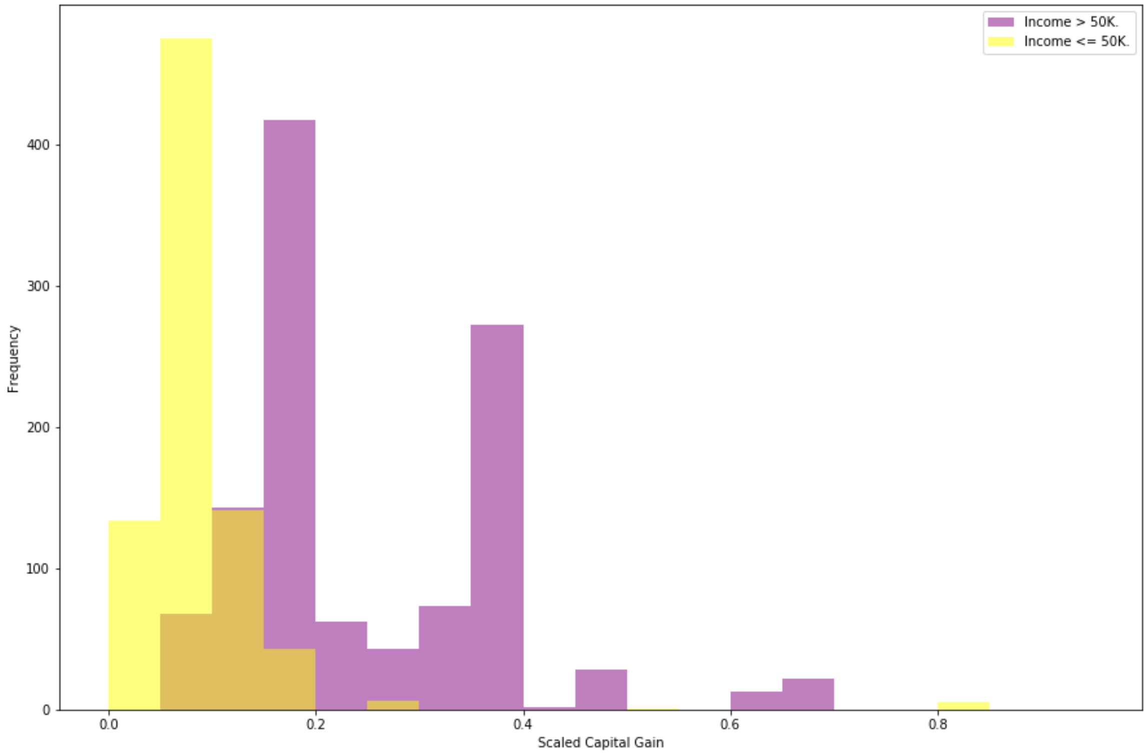


Marital Status vs Income >= $50,000

v. Histograms of Numeric Variables

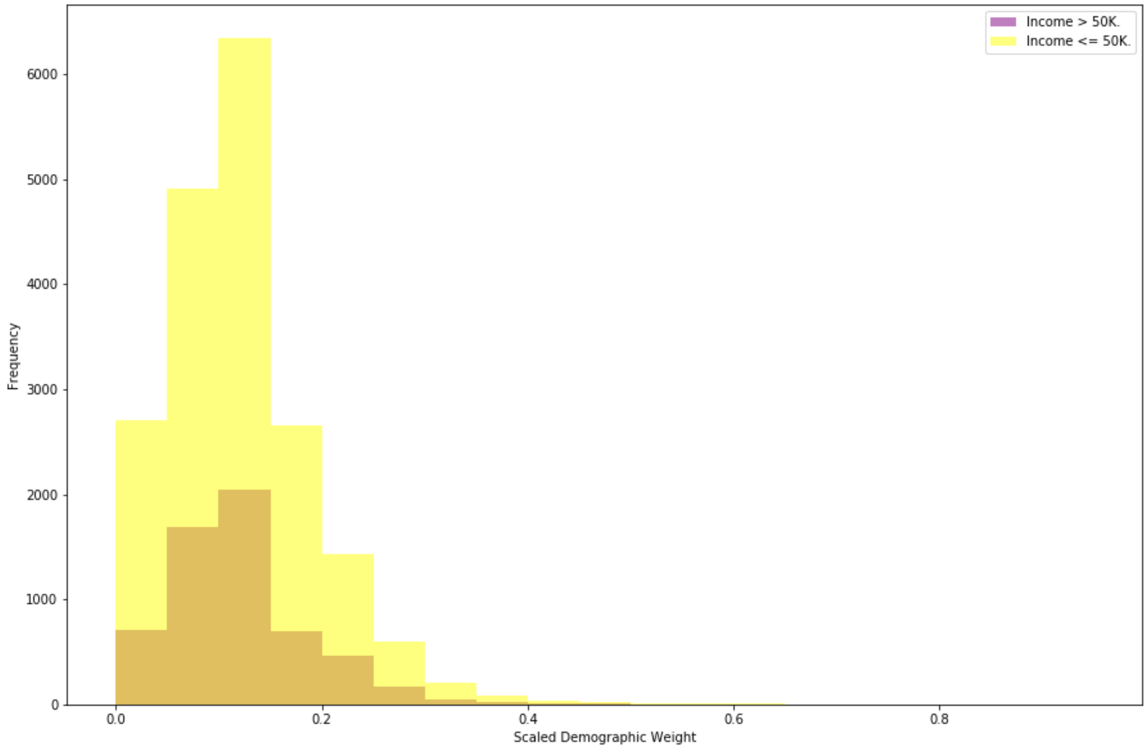
We construct two histograms, one of one numeric variable which is important in the model, and one of one numeric variable which is not important in the model. Based on the analysis in section 4.i. above, we choose the important numeric variable to be “capital-gain”, while we choose the not important numeric variable to be “demogweight”. We scaled the 2 numeric variables using the min-max normalization to increase the contrast between income over $50K and income below $50K. In the histograms below, both individuals below $50K and above $50K are normalized using the same scale for each of the numeric variable (capital gain and demographic weight).

a. Histogram of Capital Gain (Scaled):



In the histogram above, we delete the records with capital gain equal to zero to better illustrate the contrast between capital gains among individuals with income higher than $50K (purple color) versus individuals with income lower than $50K (yellow color). From the histogram, we can observe that the purple boxes are skewed more to the right than the yellow boxes, indicating that individuals with income over $50K tends to have higher capital gains than individuals with income below $50K. This is consistent with the findings of the neural network, which states that “capital-gain” variable plays the most important part in determining the income of the individuals.

b. Histogram of Demographic Weight (Scaled):



In the histogram above, we can observe that the purple boxes and yellow boxes have the same shapes and areas of dominance, except for the bigger scale/frequency of yellow boxes (as number of individuals with income below $50K is higher than income above $50K). This indicates that individuals with income over $50K and below $50K tend to have similar demographic weights. This is consistent with the findings of the neural network, which states that "demogweight" variable does not play an important part in determining the income of the individuals.

5. PART TWO: BIG LEAGUE DATA MINING

The best model was determined by splitting training data into training and validation datasets (60-40 ratio). We then used GridSearchCV on KerasClassifier to determine the most efficient number of neurons. We initially started with a range of 1,5,10,15,20,25,30 neurons in the GridSearchCV and then narrowed down it to 28, 29 and 30. By tweaking the model with different ranges we got highest accuracy of 85.06% with 29 neurons. Once the model with 29 neurons was trained using our training dataset, we ran it on validation dataset which gave an accuracy of 84.9%. Now that we have the best model with highest accuracy, we used it to predict the income outputs for observations in the test data saved in Team3predictions.txt.

6. CONCLUSION

The goal of our analysis was to build an ANN to predict whether or not an individual would have an income over $50000. Our analysis made use of US Census Bureau data. In our first model, we were able to achieve a prediction accuracy of 84.848% with 25 neurons. In this model, the most important demographic predictors were 'educational-num' and 'capital-gain'. In the second model, we partitioned the data into training and validation sets (60:40 ratio) and achieved a prediction accuracy of 85.06%, and observed 29 neurons. Furthermore, when this model was run on validation set we achieved an accuracy of 84.9%. This same model was thereafter used to predict the outcomes of whether the income for observations in the test dataset was higher than $50K or not and results of the same are attached in Team3predictions.txt for your reference.