Contents

ABSTRACT SUMMARY

The inequality of wealth and income has always been a huge concern factor for the United States Government. The likelihood of diminishing poverty is a valid reason to reduce the world's surging economic inequality. Governments of different countries across the worlds are searching for an optimal solution. This project helps to study the details about the Incomes of the US citizens and the factors that influence's them along with the predictive algorithm to determine the factors that might have influence on the annual incomes of US citizens and gives the expected values of the annual income of them. The Kaggle US Adult income data set is been used for this purpose. Classification has been done to predict the induvial person's yearly income falls between less than 50K dollars or greater 50K dollars. The Probability Neural Network is been applied to predict the classification for the income values. The project aims to determine the optimal solution by predicting the income category values (less/greater 50K) with less incorrect percentage and more good prediction using the neural network tools.

DATASET

US Adult income dataset has the anonymous information of people staying in US. The dataset contains the data of people who are earning less than 50K dollars and more than 50K dollars. The dataset consists of collective variable mentioned below in the Table1. The dataset is been distributed with 35% of data with greater 50K and 65% of data with less than 50K the Here the salary depends on various factors such as age, occupation, education, gender, race, marital-status, work class and final weight.

Column Name	Variable Type
Age	Numerical
Work class	Categorical
Forward weight	Numerical
Education	Categorical
Education-num	Numerical
Marital status	Categorical
Occupation	Categorical
Relationship	Categorical
Race	Categorical
Sex	Categorical
Capital-gain	Numerical
Capital-loss	Numerical
Hours-per week	Numerical
Native country	Categorical
Income	Categorical

GOAL

The primary goal of the project is to predict the annual income salary of the individua using the factor variables in the dataset also providing the visual details of the income of the US citizens along with the categories like age, occupation, education dataset. The model will help us to predict the categorical value like US citizen's income either more than 50K or less than 50K

TOOLS

Palisade Neural Tool: The palisade Neural tool use the collective predictive algorithms like linear regression, logistic regression, Naïve basis to analyze and predict the optimal value for the problems. We will be utilizing PNN to predict the annual income salary of the US citizen using depended and independent variable in the dataset

Tableau: Tableau helps to visualize and understand the data in efficient way. We use the tool to visualize the dataset and understand about the factors that have the influence over the citizens annual income.

SCOPE

The Project aims to determine the following task mentioned bellow

1. We will see which occupations are having highest income and which are having the lowest.
2. Education qualification of people who are earning more than 50K dollars a year.
3. Check the qualification of highest earning occupation.
4. We will analyze which age group has the highest and the lowest income.
5. At the end we will predict the income of new individuals using Neural tool.

ASSUMPTIONS

- The data has missing values in the income and will be used to predict the expected values.
- The data set had some rows containing null values in independent variable columns, after clearing out data it is reduced to 30,162 rows.
- Assuming the demographics in the data set would not deviate over a period of 10 years to the current scenario.
- Assuming the data does not contain sparse input values (0) for the entire rows which can create problems with prediction

LIMITATIONS

Neural tools/Neural networks:

- Neural networks can over-generalize the embeddings and suggest the less relevant factors when interacting with sparse values.
- Neural networks are very slow and require a very high processing power to give an output.

Tableau:

- Tableau does not provide us the options to create a custom visual chart for the data.
- They have limited data preprocessing and it is a high pricing tool.

Business Logical:

- Choosing variables which has least impact on the dependent variable won't be that effective in order to find a conclusion.

DATA PROCESSING

1. Download the .csv file for the dataset required for this project from https://www.kaggle.com/uciml/adult-census-income
2. Open Tableau application. Under connection create a connection for the dataset
3. Select other data source and select the csv data set which will load the data in the data sources where the tables and columns are displayed.

4. Drag the tables from sheet to the right-hand side of the console to create joins (inner, outer, left and right) if we have more tables linked to each other.
5. In the sheets create the visuals by dragging the values from the table and drop it in the rows and columns in the sheet.

MODEL BUILDING

Cleaning:

1. The actual dataset consists of 15 variables where some column values has invalid values like '?' and null values have been removed in the csv file before loading in the tableau.
2. Once the dataset is cleaned the csv file can be loaded in the tableau and neural tool where in the neural tool, we can remove the categorical/numerical value which does not have influence over the income values.

Visualization:

To gain the more useful insights about the data we will look at the features and distribution of the income over the variables with the income factors less than 50K or greater than 50K.

1. After cleaning the dataset, csv file can be loaded into the tableau by creating a connection in the tableau and load the income dataset as data source.
2. Once the data is loaded in the tableau, we can use the variable values to create the charts and visuals that represent the useful insights from the dataset
3. Figure 4 represent the distribution of the income range of less/greater than 50k among the different age groups of US citizens.
4. Figure 5 gives us the visual comparison of distribution of incomes of the US people with the independent variable factors like age and Education of the citizens.
5. Figure 6 gives the comparison between the annual income over the variables like occupation and relationships of the US citizens.

OBSERVATION

Form the below visuals we were able to determine the insights of the data and able to understand the factors that have a potential influence over the income factors

1. From Figure 3 we able to analyze that the dataset consists of 30,162 population our of which 20,380 are male and 9,782 are female from 41 different countries living in US.
2. Figure 4 represent the distribution of income over the different range of people from 10 – 90 where range of people from 20 – 40 contributes about 60% of the annual income of less than 50K dollars. Where people age ranges from 35 – 50 contributes about 50% of income more than 50K dollars
3. Figure 5 compares between the different categories of income over the factors of age and education. Where we can observe the drop in the age factor from 50 in both categories of income. People with HS grade has more annual income but people with bachelors has 13% more greater than 50K income than HS grad peoples.
4. Figure 6 compares between the different categories of income over the variables like occupation and relationship where occupations like Prof-specialty, craft repair and exec-managerial contributes more on the annual income of 45% on the total income also these three occupations contribute more for the income more than 50K. Where in relationship husbands contribute more on both category of incomes.

SELECTING VARIABLES

After analyzing the dataset, we can identify the variable for the prediction values that can be used in the neural tools. Considering the income as our dependent variable with 2 categories like <=50k and >50k. where other variables are independent variables like age, education, occupation, relationship, gender, hour worked and native country.

PREDECTION

1. From the dataset we have identified some invalid data like null values and "?". Where the data and removed.
2. We are initializing the whole cleaned dataset in the dataset manager in neural tools.
3. Once the dataset is initialized, we can train/test the dataset in the tool by selecting the percentage of data for training and testing
4. Our predictive model which we have created will predict those incomes for individuals. In this model dependent variable is the income because we are predicting that variable, rest are independent variables.
5. We are determining two predictive models with the different percentage of training and test data.
6. The first prediction will have 20% of dataset in training t and 80% data in test set where second prediction will contain 70% data in the training set and 30% in the test-set.

Predicting new income

Once the model is trained it displays the details about the values in the training and auto test predicated values along with incorrect percentage and either good/bad. When the incorrect percentage is less that determines the prediction is good else bad. From both prediction models we can figure out the predicated income category. Figure 7 represent the predicted values with less incorrect percentage gives good prediction which is marked with green box. Where high incorrect percent gives bad prediction marked with red box.

From the first model which consists of 80% training and 20% test data gives 89.58% good prediction in the training and 84.19% good prediction in testing in the summery below table. We can see it contains 10.41% bad prediction in the training model and 15.80% bad prediction for the testing model. The way in prediction 2 using another model consist of 70% data in training and 30% in the training, it gives 9.92% bad prediction for the training model and 16.58% for the testing model which implies 90.08% and 83.42% good prediction for the training and the testing model.

Observing the details from both the analysis of training and testing models the chance of reducing the bad prediction by increasing the percentage of testing data.

Information's	Prediction 1	Prediction 2
% Bad Predictions for training	10.4187%	9.9023%
% Bad Predictions for testing	15.008%	16.5874%
Number of training cases	24101	21088
Number of testing cases	6025	9037
% Good prediction in training	89.58%	90.09%
% Good prediction testing	84.99%	83.42%

CONCLUSION

From the analysis we have observed education does not matter to have an annual income more than 50k a year because people who have income more than 50k most of them are bachelor degree holders and are from some college instead of masters or Phds. Young crowd from age 23-24 are more likely to have income less than 50k annually however people who are senior and who fall under the age 37 -47 and working as an executive manager or a professor are more likely to have an income more than 50k annually.

APPENDIX

Appendix – A List of Figures

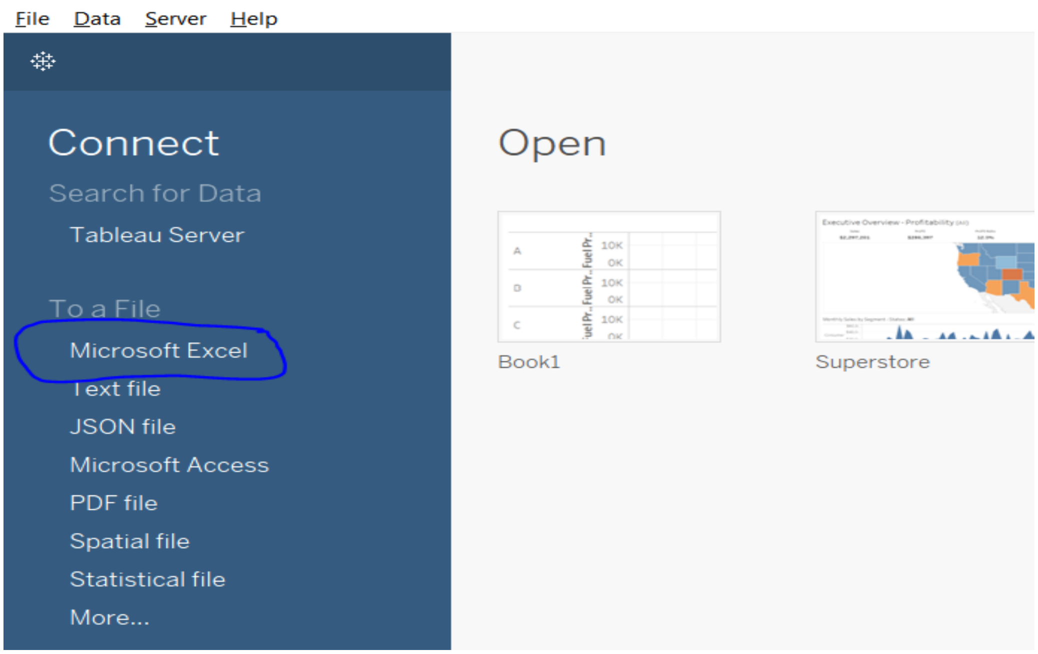

Figure 1: Connecting to tableau

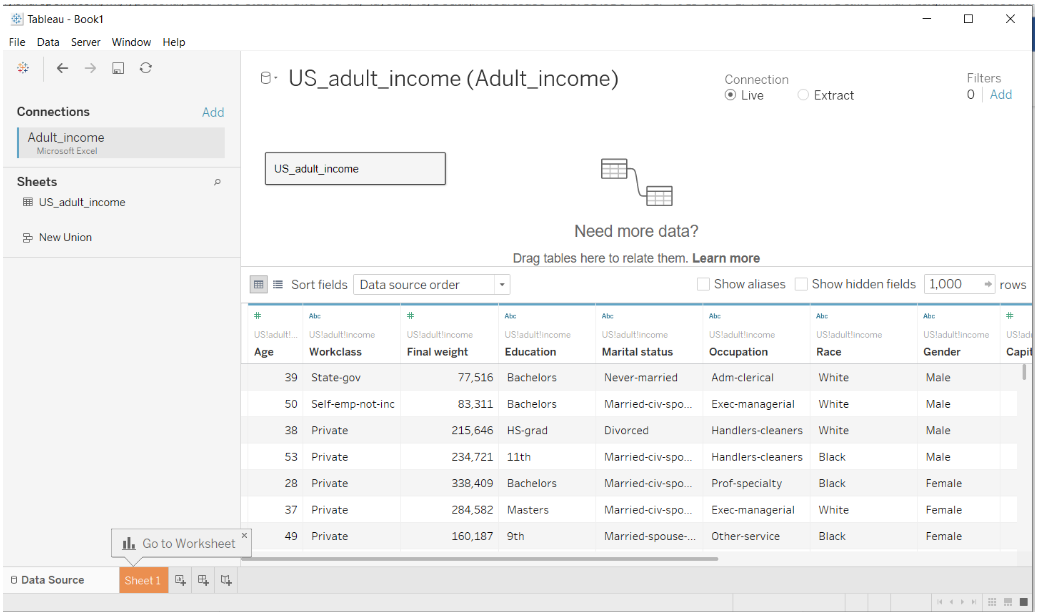

Figure 2: Setting up connections

Gender	
Female	9,782
Male	20,380

Figure 3: Total population

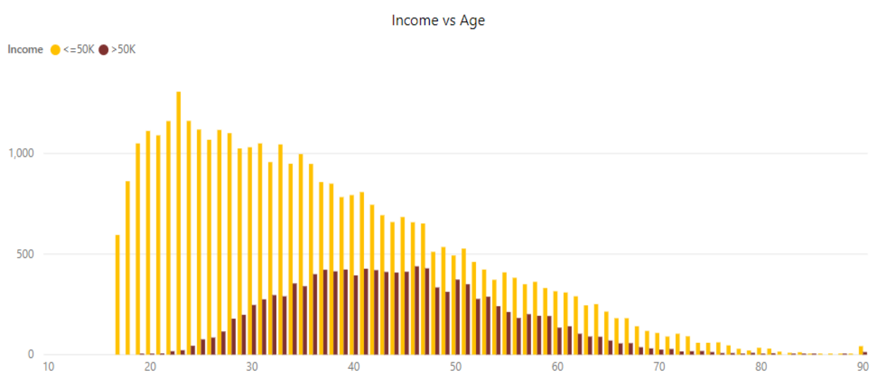

Figure 4: Distribution of income with ages

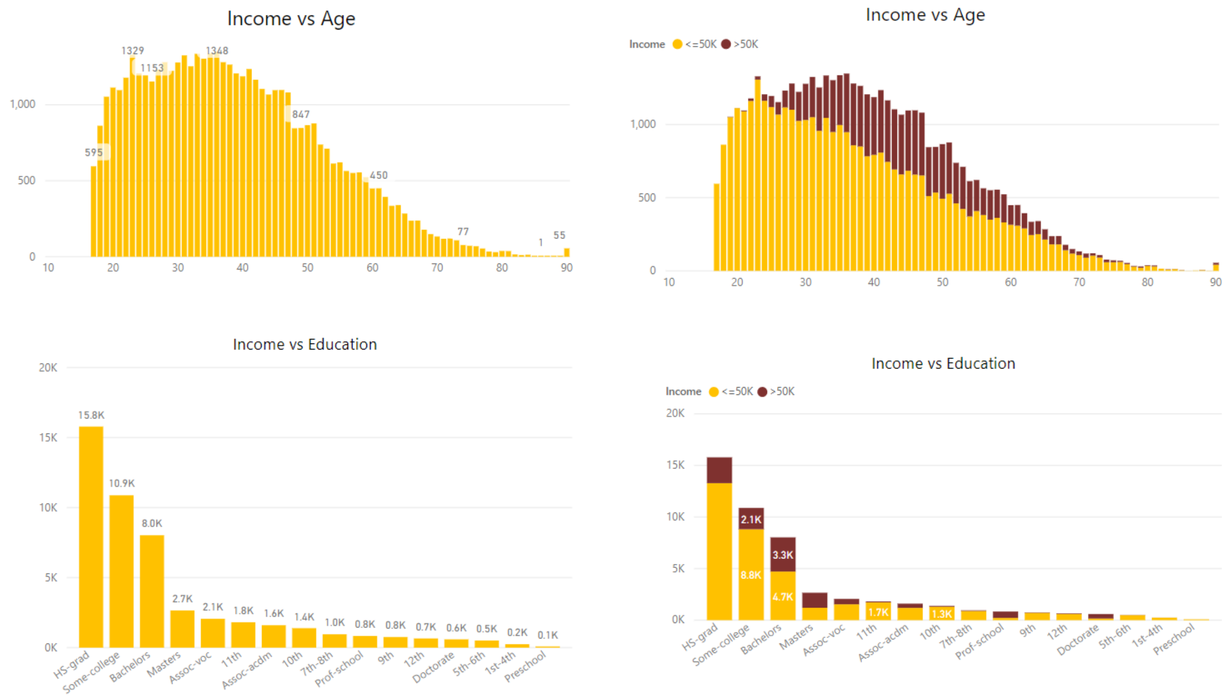

Figure 5: Comparison of incomes

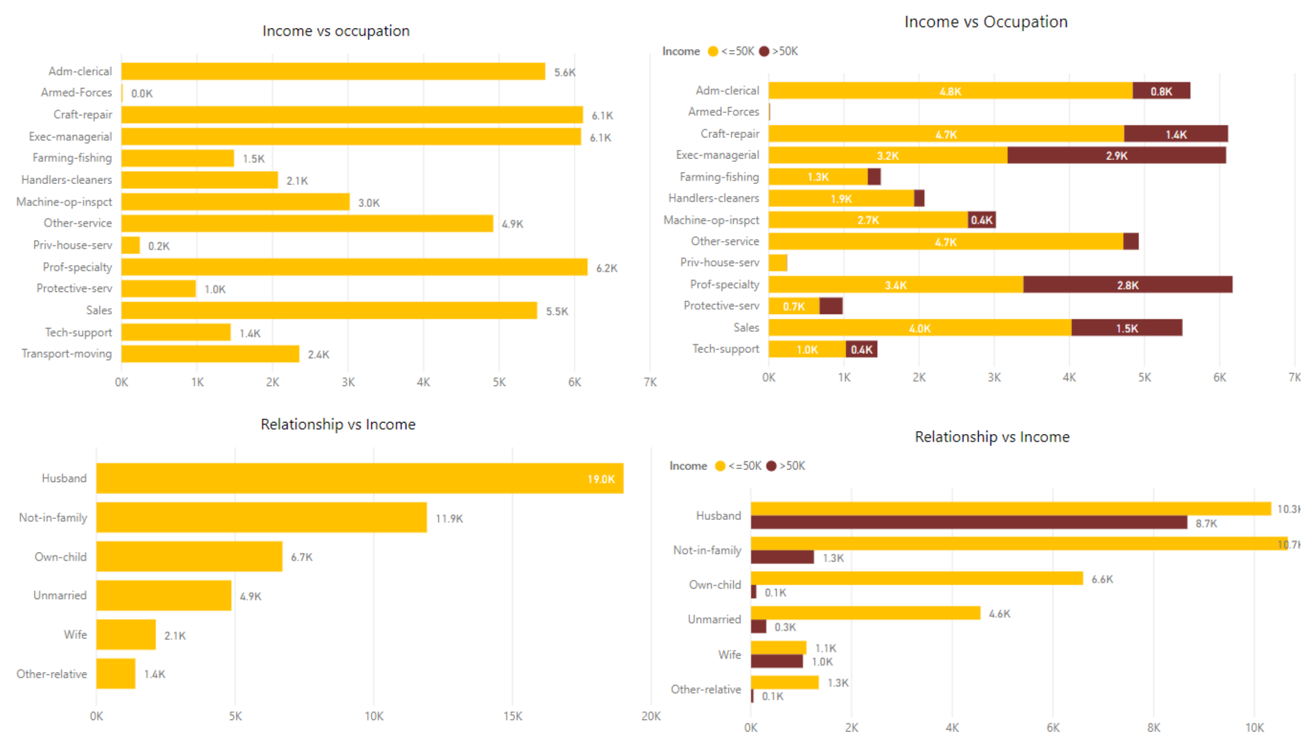

Figure 6: Comparison of incomes

Appendix – B Prediction Models

Prediction

In order to do the prediction we will have 2 models one would be using 80% using training set and 20% in the test set and another would be using 70% in the test set and 30% will be in the test set.

gender	hours-per	income
Male	40	<=50K
Male	50	<=50K
Male	40	>50K
Male	40	>50K
Female	30	<=50K
Male	30	<=50K
Male	40	<=50K
Male	32	>50K
Female	40	<=50K
Male	10	<=50K
Male	40	>50K
Male	40	<=50K
Female	39	<=50K
Male	35	<=50K
Male	48	>50K
Male	50	>50K
Male	25	<=50K
Female	30	<=50K

Testing Report: "Net Trained on Data Set #1"

Tag Used	Prediction	Prediction%	Incorrect%	Good/Bad
test	<=50K	99.19%	0.81%	Good
test	<=50K	86.62%	13.38%	Good
test	<=50K	58.57%	58.57%	Bad
test	<=50K	62.91%	62.91%	Bad
test	<=50K	99.42%	0.58%	Good
test	<=50K	99.84%	0.16%	Good
test	<=50K	97.07%	2.93%	Good
test	>50K	98.38%	1.62%	Good
test	<=50K	97.81%	2.19%	Good
test	<=50K	99.43%	0.57%	Good
test	<=50K	62.92%	62.92%	Bad
test	>50K	50.27%	50.27%	Bad
test	<=50K	98.02%	1.98%	Good
test	<=50K	64.68%	35.32%	Good
test	<=50K	68.43%	68.43%	Bad
test	>50K	87.56%	12.44%	Good
test	<=50K	99.78%	0.22%	Good
test	<=50K	61.60%	38.40%	Good

Figure 7: Predicting Good/ Bad values

Prediction 1

This prediction shows in figure 8 has been done using 20% of data in the test set and the training set comprises 80% of the data.

Age	Workclass	Final_wei	Education	Marital_st	Occupatio	Race	Gender	Capital_ga	Capital_lo	Hour_wor	Country	Income	Tag Used	Prediction	Prediction%	Incorrect%	Good/Bad
39	State-gov	77516	Bachelors	Never-ma	Adm-cler	White	Male	2174	0	40	United-St	<=50K	train				
50	Self-emp	83311	Bachelors	Married-c	Exec-man	White	Male	0	0	13	United-St	<=50K	train				
38	Private	215646	HS-grad	Divorced	Handlers-	White	Male	0	0	40	United-St	<=50K	train				
53	Private	234721	11th	Married-c	Handlers-	Black	Male	0	0	40	United-St	<=50K	train				
28	Private	338409	Bachelors	Married-c	Prof-spec	Black	Female	0	0	40	Cuba	<=50K	test	<=50K	60.55%	39.45%	Good
37	Private	284582	Masters	Married-c	Exec-man	White	Female	0	0	40	United-St						
49	Private	160187	9th	Married-s	Other-ser	Black	Female	0	0	16	Jamaica	<=50K	train				
52	Self-emp	209642	HS-grad	Married-c	Exec-man	White	Male	0	0	45	United-St	>50K	train				
31	Private	45781	Masters	Never-ma	Prof-spec	White	Female	14084	0	50	United-St	>50K	train				
42	Private	159449	Bachelors	Married-c	Exec-man	White	Male	5178	0	40	United-States		predict	>50K	72.14%		
37	Private	280464	Some-col	Married-c	Exec-man	Black	Male	0	0	80	United-States		predict	<=50K	55.72%		
30	State-gov	141297	Bachelors	Married-c	Prof-spec	Asian-Pac	Male	0	0	40	India		predict	<=50K	53.70%		
23	Private	122272	Bachelors	Never-ma	Adm-cler	White	Female	0	0	30	United-States		predict	<=50K	98.13%		
32	Private	205019	Assoc-acc	Never-ma	Sales	Black	Male	0	0	50	United-St	<=50K	test	<=50K	90.27%	9.73%	Good
34	Private	245487	7th-8th	Married-c	Transport	Amer-Ind	Male	0	0	45	Mexico	<=50K	train				
25	Self-emp	176756	HS-grad	Never-ma	Farming-f	White	Male	0	0	35	United-St	<=50K	train				
32	Private	186824	HS-grad	Never-ma	Machine-	White	Male	0	0	40	United-States		predict	<=50K	95.51%		
38	Private	28887	11th	Married-c	Sales	White	Male	0	0	50	United-St	<=50K	train				
43	Self-emp	292175	Masters	Divorced	Exec-man	White	Female	0	0	45	United-St	>50K	train				
40	Private	193524	Doctorate	Married-c	Prof-spec	White	Male	0	0	60	United-St	>50K	train				
54	Private	302146	HS-grad	Separated	Other-ser	Black	Female	0	0	20	United-St	<=50K	train				
35	Federal-g	76845	9th	Married-c	Farming-f	Black	Male	0	0	40	United-St	<=50K	train				
43	Private	117037	11th	Married-c	Transport	White	Male	0	2042	40	United-States		predict	>50K	69.87%		
59	Private	109015	HS-grad	Divorced	Tech-sup	White	Female	0	0	40	United-St	<=50K	train				
56	Local-gov	216851	Bachelors	Married-c	Tech-sup	White	Male	0	0	40	United-St	>50K	train				
19	Private	168294	HS-grad	Never-ma	Craft-rep	White	Male	0	0	40	United-St	<=50K	train				
39	Private	367260	HS-grad	Divorced	Exec-man	White	Male	0	0	80	United-St	<=50K	train				
49	Private	193366	HS-grad	Married-c	Craft-rep	White	Male	0	0	40	United-St	<=50K	train				
23	Local-gov	190709	Assoc-acc	Never-ma	Protective	White	Male	0	0	52	United-St	<=50K	train				
20	Private	266015	Some-col	Never-ma	Sales	Black	Male	0	0	44	United-St	<=50K	train				
45	Private	386940	Bachelors	Divorced	Exec-man	White	Male	0	1408	40	United-St	<=50K	train				
30	Federal-g	59951	Some-col	Married-c	Adm-cler	White	Male	0	0	40	United-States		predict	<=50K	60.09%		
22	State-gov	311512	Some-col	Married-c	Other-ser	Black	Male	0	0	15	United-St	<=50K	train				
48	Private	242406	11th	Never-ma	Machine-	White	Male	0	0	40	Puerto-Ri	<=50K	train				
21	Private	197200	Some-col	Never-ma	Machine-	White	Male	0	0	40	United-St	<=50K	train				

Figure 8: Prediction 1- Training dataset

Summary	
Net Information	
Name	Net Trained on Data Set #1
Configuration	PNN Category Predictor
Location	This Workbook
Independent Category Variables	7 (Workclass, Education, Marital_status, Occupation, Race, Gender, Country)
Independent Numeric Variables	5 (Age, Final_weight, Capital_gain, Capital_loss, Hour_worked)
Dependent Variable	Category Var. (Income)
Training	
Number of Cases	24101
Training Time	0:38:11
Number of Trials	0
Reason Stopped	Auto-Stopped
% Bad Predictions	10.4187%
Mean Incorrect Probability	19.3376%
Std. Deviation of Incorrect Prob.	20.0947%
Testing	
Number of Cases	6025

% Bad Predictions	15.8008%
Mean Incorrect Probability	24.2264%
Std. Deviation of Incorrect Prob.	22.7893%
Prediction	
Number of Cases	35
Live Prediction Enabled	YES
Data Set	
Name	Data Set #1
Number of Rows	30162
Manual Case Tags	NO

Table 1: Summary of prediction 1 training/testing

Classification Matrix			
(for training cases)			
	<=50K	**>50K**	**Bad (%)**
<=50K	17503	603	3.3304%
>50K	1908	4087	31.8265%

Classification Matrix			
(for testing cases)			
	<=50K	**>50K**	**Bad (%)**
<=50K	4241	282	6.2348%
>50K	670	832	44.6072%

Table 2: Income classification for prediction 1 training/testing

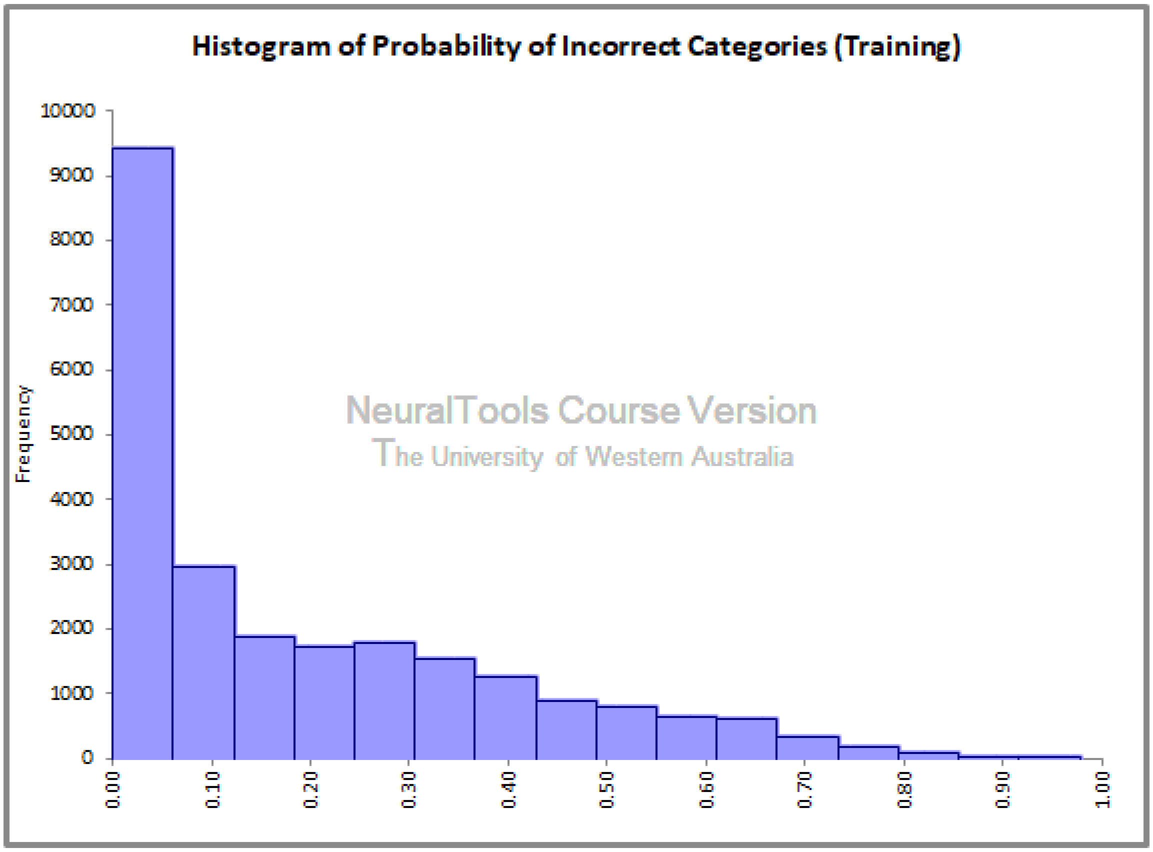

Figure 9: Prediction1 Training incorrect probability

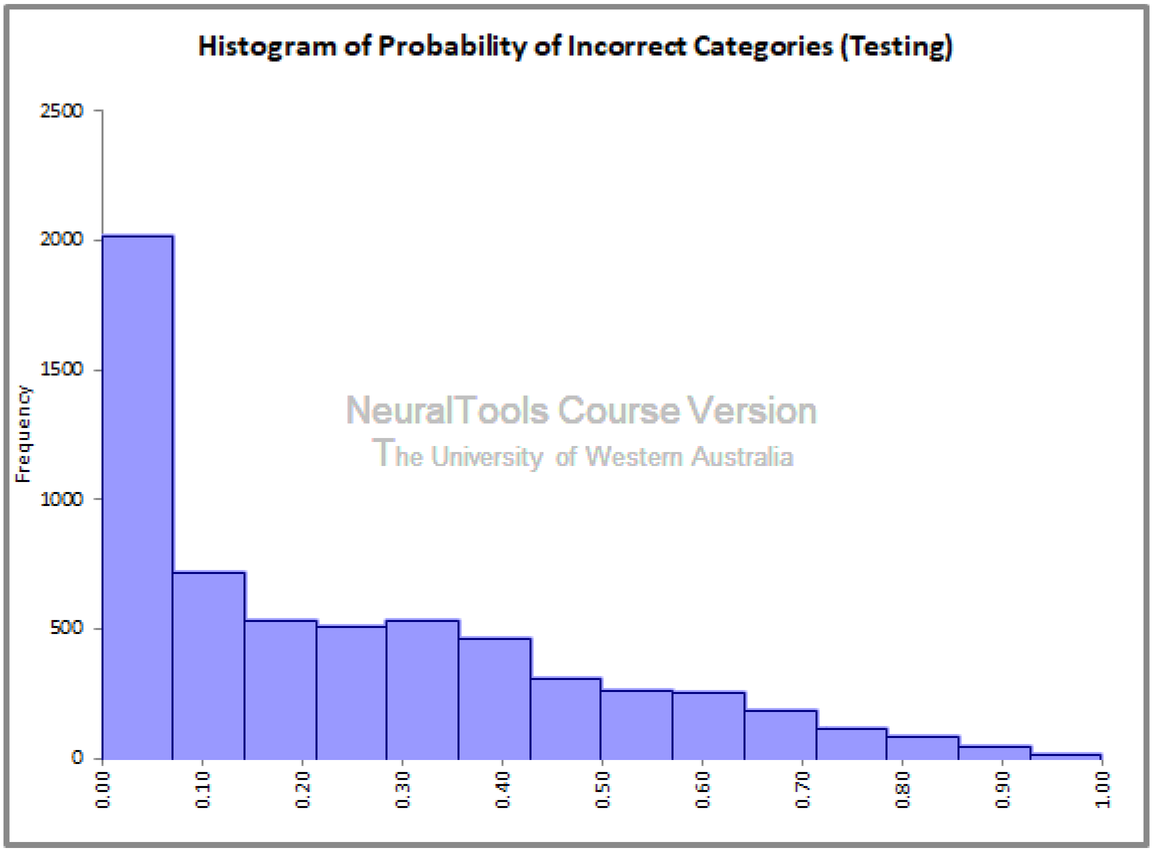

Figure 10: Prediction1 Testing incorrect probability

Prediction 2:
This prediction shows in figure 9 has been done using 30% of data in the test set and the training set comprises 70% of the data.

Age	Workclass	Final_wei	Education	Marital_st	Occupatio	Race	Gender	Capital_ga	Capital_lo	Hour_wor	Country	Income	Tag Used	Prediction	Prediction%	Incorrect%	Good/Bad
39	State-gov	77516	Bachelors	Never-ma	Adm-cler	White	Male	2174	0	40	United-St	<=50K	train				
50	Self-emp	83311	Bachelors	Married-c	Exec-man	White	Male	0	0	13	United-St	<=50K	test	<=50K	57.35%	42.65%	Good
38	Private	215646	HS-grad	Divorced	Handlers-	White	Male	0	0	40	United-St	<=50K	train				
53	Private	234721	11th	Married-c	Handlers-	Black	Male	0	0	40	United-St	<=50K	train				
28	Private	338409	Bachelors	Married-c	Prof-spec	Black	Female	0	0	40	Cuba	<=50K	test	<=50K	58.61%	41.39%	Good
37	Private	284582	Masters	Married-c	Exec-man	White	Female	0	0	40	United-St						
49	Private	160187	9th	Married-s	Other-ser	Black	Female	0	0	16	Jamaica	<=50K	train				
52	Self-emp	209642	HS-grad	Married-c	Exec-man	White	Male	0	0	45	United-St	>50K	train				
31	Private	45781	Masters	Never-ma	Prof-spec	White	Female	14084	0	50	United-St	>50K	train				
42	Private	159449	Bachelors	Married-c	Exec-man	White	Male	5178	0	40	United-States		predict	>50K	71.26%		
37	Private	280464	Some-col	Married-c	Exec-man	Black	Male	0	0	80	United-States		predict	<=50K	59.59%		
30	State-gov	141297	Bachelors	Married-c	Prof-spec	Asian-Pac	Male	0	0	40	India		predict	>50K	50.21%		
23	Private	122272	Bachelors	Never-ma	Adm-cler	White	Female	0	0	30	United-States		predict	<=50K	98.08%		
32	Private	205019	Assoc-acc	Never-ma	Sales	Black	Male	0	0	50	United-St	<=50K	test	<=50K	88.60%	11.40%	Good
34	Private	245487	7th-8th	Married-c	Transport	Amer-Ind	Male	0	0	45	Mexico	<=50K	train				
25	Self-emp	176756	HS-grad	Never-ma	Farming-f	White	Male	0	0	35	United-St	<=50K	train				
32	Private	186824	HS-grad	Never-ma	Machine-	White	Male	0	0	40	United-States		predict	<=50K	95.92%		
38	Private	28887	11th	Married-c	Sales	White	Male	0	0	50	United-St	<=50K	train				
43	Self-emp	292175	Masters	Divorced	Exec-man	White	Female	0	0	45	United-St	>50K	train				
40	Private	193524	Doctorate	Married-c	Prof-spec	White	Male	0	0	60	United-St	>50K	train				
54	Private	302146	HS-grad	Separated	Other-ser	Black	Female	0	0	20	United-St	<=50K	train				
35	Federal-g	76845	9th	Married-c	Farming-f	Black	Male	0	0	40	United-St	<=50K	test	<=50K	75.76%	24.24%	Good
43	Private	117037	11th	Married-c	Transport	White	Male	0	2042	40	United-States		predict	>50K	70.86%		
59	Private	109015	HS-grad	Divorced	Tech-sup	White	Female	0	0	40	United-St	<=50K	test	<=50K	88.93%	11.07%	Good
56	Local-gov	216851	Bachelors	Married-c	Tech-sup	White	Male	0	0	40	United-St	>50K	test	>50K	51.15%	48.85%	Good
19	Private	168294	HS-grad	Never-ma	Craft-rep	White	Male	0	0	40	United-St	<=50K	train				
39	Private	367260	HS-grad	Divorced	Exec-man	White	Male	0	0	80	United-St	<=50K	train				
49	Private	193366	HS-grad	Married-c	Craft-rep	White	Male	0	0	40	United-St	<=50K	train				
23	Local-gov	190709	Assoc-acc	Never-ma	Protective	White	Male	0	0	52	United-St	<=50K	train				
20	Private	266015	Some-col	Never-ma	Sales	Black	Male	0	0	44	United-St	<=50K	train				
45	Private	386940	Bachelors	Divorced	Exec-man	White	Male	0	1408	40	United-St	<=50K	train				
30	Federal-g	59951	Some-col	Married-c	Adm-cler	White	Male	0	0	40	United-States		predict	<=50K	51.19%		
22	State-gov	311512	Some-col	Married-c	Other-ser	Black	Male	0	0	15	United-St	<=50K	test	<=50K	99.28%	0.72%	Good
48	Private	242406	11th	Never-ma	Machine-	White	Male	0	0	40	Puerto-Ri	<=50K	train				
21	Private	197200	Some-col	Never-ma	Machine-	White	Male	0	0	40	United-St	<=50K	train				

Figure 11: Prediction2 Training/Testing

Summary	
Net Information	
Name	Net Trained on Data Set #1
Configuration	PNN Category Predictor
Location	This Workbook
Independent Category Variables	7 (Workclass, Education, Marital_status, Occupation, Race, Gender, Country)
Independent Numeric Variables	5 (Age, Final_weight, Capital_gain, Capital_loss, Hour_worked)
Dependent Variable	Category Var. (Income)
Training	
Number of Cases	21088
Training Time	0:27:29
Number of Trials	0
Reason Stopped	Auto-Stopped
% Bad Predictions	9.9203%
Mean Incorrect Probability	18.9184%
Std. Deviation of Incorrect Prob.	19.9077%
Testing	
Number of Cases	9037
% Bad Predictions	16.5874%
Mean Incorrect Probability	24.6794%
Std. Deviation of Incorrect Prob.	22.9416%
Prediction	
Number of Cases	35

Live Prediction Enabled	YES
Data Set	
Name	Data Set #1
Number of Rows	30162
Manual Case Tags	NO

Table 3: Summary prediction 1 training/testing

Classification Matrix			
(for training cases)			
	<=50K	**>50K**	**Bad (%)**
<=50K	15421	485	3.0492%
>50K	1607	3575	31.0112%

Classification Matrix			
(for testing cases)			
	<=50K	**>50K**	**Bad (%)**
<=50K	6303	419	6.2333%
>50K	1080	1235	46.6523%

Table 4: Income classification for prediction 2 training/testing

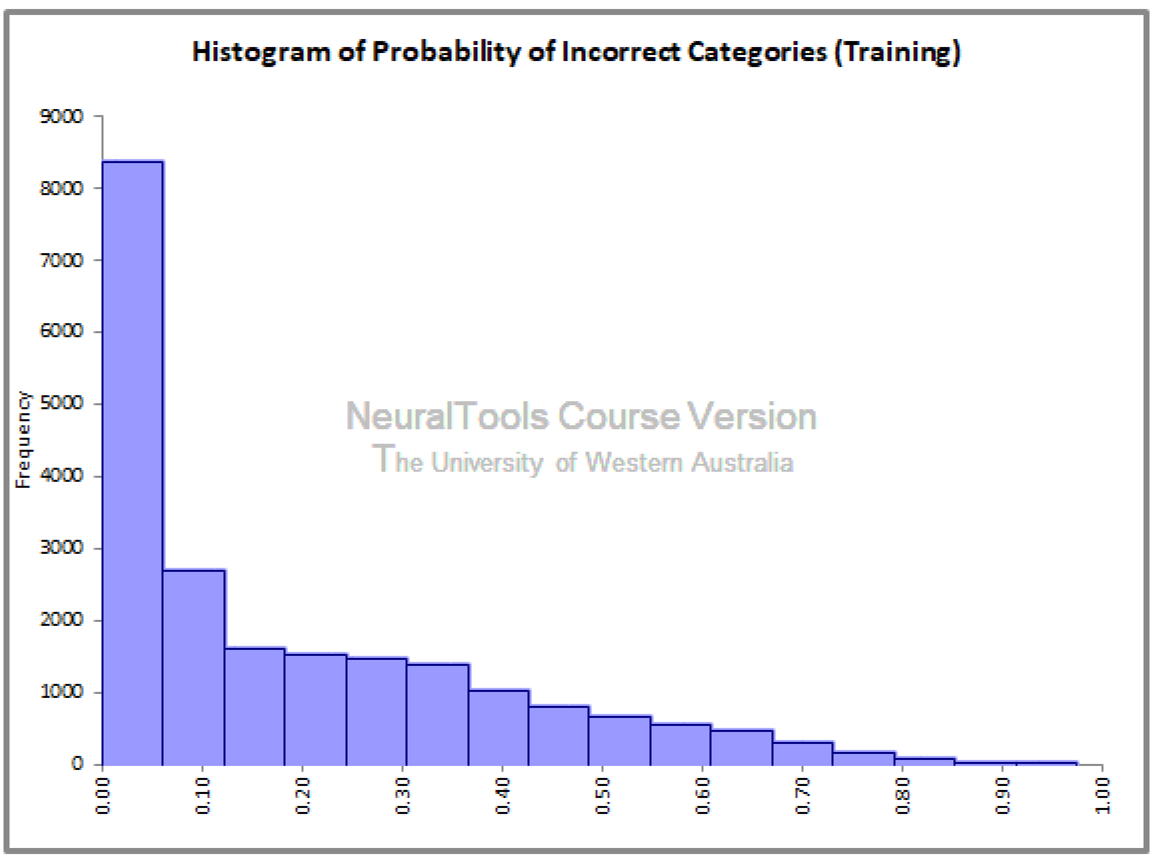

Figure 11: Prediction2 Testing incorrect probability

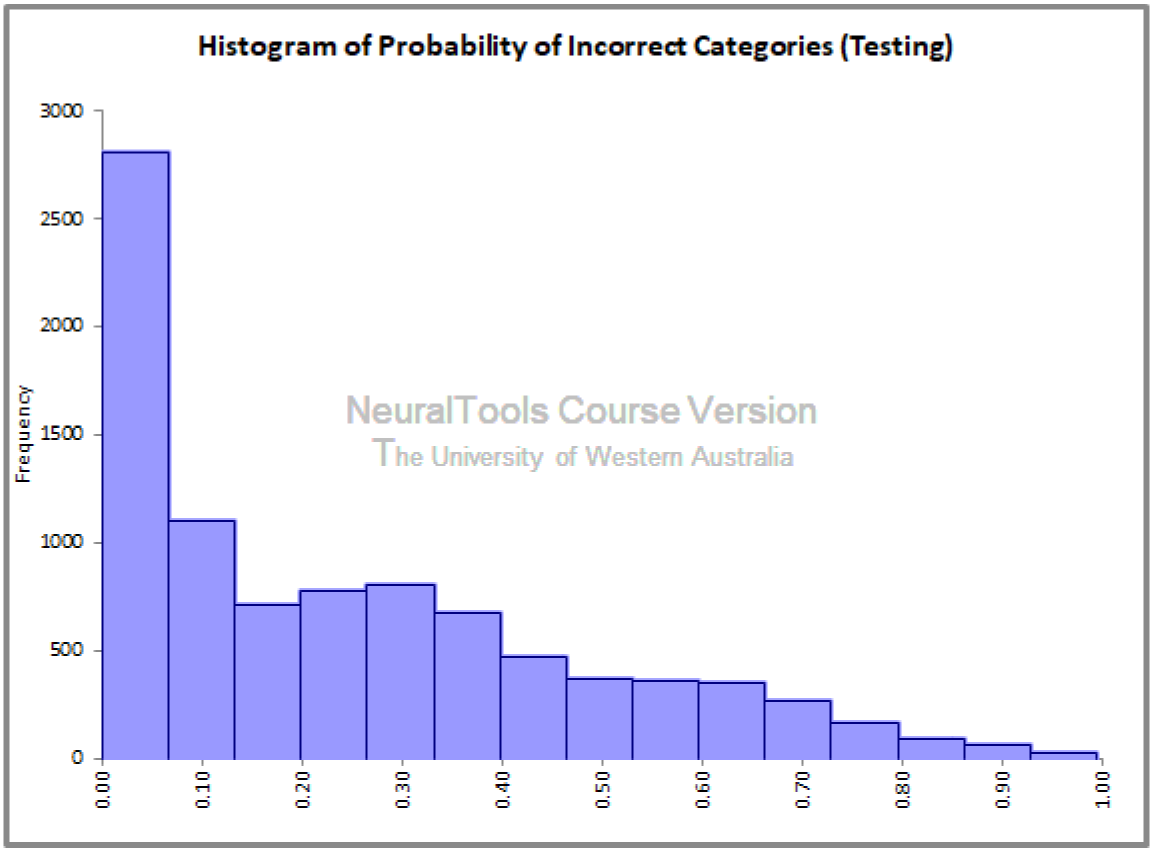

Figure12: Prediction2 Testing incorrect probability