Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Group Number: 7

Assignment Title: Group Assignment 3

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:

☒	1002183031	☒	1006604701
☒	1002897378	☒	1007045118
☒	1007554745	☒	1005627403

Executive Summary

This report discusses the use of artificial neural network algorithms and Keras Classifier in order to accurately predict whether an individual has an income over $50,000 per year in the US. The first step involved understanding the given data by generating a series of plots. Followed by, exploratory data analysis which was conducted to examine the quantitative variables and assess their significance. The categorical data were transformed into sets of dummy variables which created a binary variable representing each of the possible categories. After the data was explored, pre-processed, and split between training and validation sets using a 60%-40% proportion, it was used to fit the neural network. The initial ANN was trained on 28 predictors and it was found that the network could accurately predict the target variable around 85% of the time. Further, this model was tweaked based on its performance on the validation set; the best model included only the significant predictors which led to an increase in model accuracy to around 86%, while the false positive rate was minimized (0.0893) which is deemed important for the process of granting personal loans to individuals by assessing their income levels.

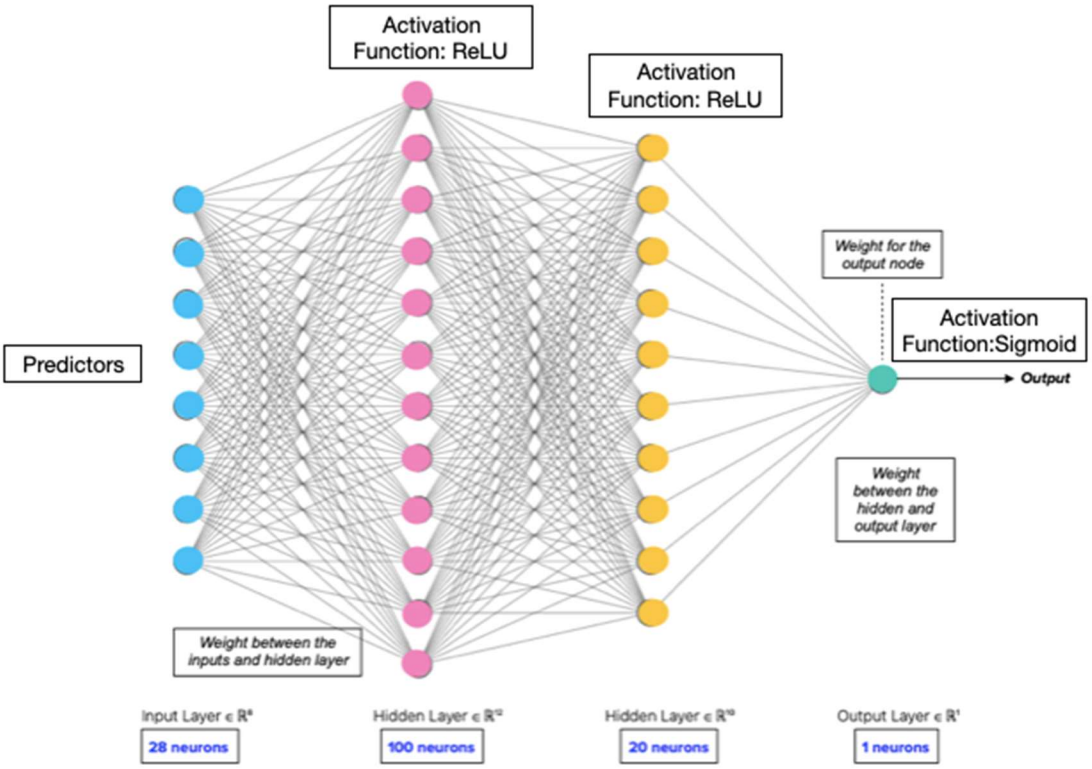
Part One: Training Data Set

Data Preprocessing

Due to the model definition, all input and output variables should be numeric values from 0 to 1. Therefore, we converted all numeric variables using a min-max scaler, and created dummy variables for all categorical variables.

Neural Network

Using the quantitative and qualitative predictors, a sequential neural network was developed to predict the income of individuals based upon their attributes. Shown in figure below, is the network topology. The quantitative predictors (blue nodes) and qualitative predictors (orange nodes) represent the input layer. Next, two hidden layers were constructed with nodes, and finally an output layer contained a single node. The connections between the various layers represented the weights, and as the model reiterated through back-propagation, the weights were adjusted until they eventually reached a global minimum; resulting in the lowest squared error between the predicted output and actual (test/validation) output.



Initially, the training dataset was split into a training set and test set (60% - 40%, respectively), in order to obtain a test accuracy for the model before introducing the previously unseen data (US Census Test Data) and making predictions. Furthermore, during the training of the model, k-fold cross-validation was performed in order to obtain an estimate of the test error rate of the model. Here we used early stopping of training of the model via a callback function called *EarlyStopping.* This callback function allows us to specify the performance measure to monitor the trigger, and once triggered, it will stop the training process when generalization error begins to degrade. This allows us to avoid the problem of overfitting.

Important Predicting Variables

EDA

For a better understanding of the given dataset, exploratory data analysis was conducted. Firstly, it was found that the variables education and education-num are perfectly correlated. Below table, which shows a cross-tabulation between both variables; as seen for each string representation of education (variable: education) there is one corresponding numerical class that represents it (variable: education-num). Hence, we decided to drop the variable education from the model to ensure that all predictors

are providing independent information in the model, which helps to reduce the overall variance for the model. Furthermore, it was found that 1404 observations had "?" values for Occupation & WorkClass, shown in the cross-tabulation of the variables in below tables. We decided to create dummy variables for those observations and not neglect their patterns in the data. This approach was chosen since those values represented around 5.6% of the overall data, which was significant enough to not delete and would have affected the test accuracy of the model.

education-num / education	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
10th	0	0	0	0	0	721	0	0	0	0	0	0	0	0	0	0
11th	0	0	0	0	0	0	909	0	0	0	0	0	0	0	0	0
12th	0	0	0	0	0	0	0	323	0	0	0	0	0	0	0	0
1st-4th	0	120	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5th-6th	0	0	244	0	0	0	0	0	0	0	0	0	0	0	0	0
7th-8th	0	0	0	491	0	0	0	0	0	0	0	0	0	0	0	0
9th	0	0	0	0	394	0	0	0	0	0	0	0	0	0	0	0
Assoc-acdm	0	0	0	0	0	0	0	0	0	0	0	801	0	0	0	0
Assoc-voc	0	0	0	0	0	0	0	0	0	0	1059	0	0	0	0	0
Bachelors	0	0	0	0	0	0	0	0	0	0	0	0	4140	0	0	0
Doctorate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	315
HS-grad	0	0	0	0	0	0	0	0	8120	0	0	0	0	0	0	0
Masters	0	0	0	0	0	0	0	0	0	0	0	0	0	1300	0	0
Preschool	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Prof-school	0	0	0	0	0	0	0	0	0	0	0	0	0	0	430	0
Some-college	0	0	0	0	0	0	0	0	0	5597	0	0	0	0	0	0

occupation / workclass	?	Adm-clerical	Armed-Forces	Craft-repair	Exec-managerial	Farming-fishing	Handlers-cleaners	Machine-op-inspct	Other-service	Priv-house-serv	Prof-specialty	Protective-serv	Sales	Tech-support	Transport-moving
?	1399	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Federal-gov	0	246	7	49	132	6	19	10	30	0	139	25	11	56	20
Local-gov	0	220	0	121	172	23	32	8	158	0	529	230	6	33	92
Never-worked	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Private	0	2246	0	2420	2036	355	931	1465	2116	121	1771	148	2260	544	973
Self-emp-inc	0	26	0	79	312	37	1	11	19	0	122	3	225	2	20
Self-emp-not-inc	0	40	0	412	300	331	10	30	136	0	296	6	306	19	92
State-gov	0	196	0	40	133	11	6	11	96	0	323	91	7	49	30
Without-pay	0	1	0	1	0	4	1	1	0	0	0	0	0	0	1

Select Variables

In order to determine the important variables, as the first step we conducted sensitivity analysis to measure the relative influence of each feature on the output result. We created a new observation as X_mean to calculate the mean of various features for all the records in the testset. We obtained the model output for Xmean and named it as Y_mean. Finally, for each feature we changed them from their minimum to maximum amount and compared the output result to Y_mean. Based on this analysis we found the following features that have higher influence on the output (income):

- Education-num

- Capital-gain
- Capital-loss
- Hours-per-week
- Demographic
- Sex
- Marital-status_Divorced
- Marital-status_Never-married
- Marital-status_Married-civ-spouse
- Relationship_Husband
- Relationship_Not in family
- Relationship_Own Child
- Relationship_Unmarried
- Occupation_other services
- Occupation_Prof-specialty
- Occupation_Handlers-cleaners
- Occupation_Craftg-repairing
- Occupation_Adm-clerical
- Occupation_Farming-fishing
- Work class_Private
- C_Germany
- C_United-States
- Race_black
- Race_white

In order to identify more features that might have impact on the output (income), we used "Permutation Importance Analysis" which is randomly shuffle a single feature in the data, leaving the target and all others in place, to see how would that affect the final prediction performances. We have used the permutation importance package from ELI5. Based on this analysis we found the following features (weight >= 0.005) that have higher influence on the output (income):

- Relationship_Never-married
- Education _num
- Capital_gain
- Relationship_Husband
- Age
- Marital-status_Divorced
- Occupation_Craftg-repairing
- Work class_Private

- Occupation_other services
- Work class_Self-emp-not-inc
- Hours-per-week
- Capital_loss
- Occupation_Machine-op-inspect
- C_United-States
- Occupation_Handlers-cleaners
- Race_white
- Occupation_Adm-clerical

Parameter Tuning

To create the optimal model, we used the GridSearchCV function to determine what optimal parameters should be in terms of learning rate, neurons, batch size, and epochs. We additionally used 10-folds cross validation to shuffle the training set. Additionally, due to the computing power required to run this, we implemented early stopping based on the validation set and implemented model checkpoints to have our model saved after the best one is found. This allowed for the model to stop running at the epoch it finds the optimal model based on a patience of 5 additional epochs.

After running the grid search, we found that the optimal model would have 100 neurons, batch_size of 10, epochs of 100, and learning rate of 0.001. For future use, we replicated these parameters in a single model so that the parameter grid would not have to be run each time. Due to the limitations of our computing power, we could not explore as many parameters as we would like, but the results showed that overfitting had not occurred.

Accuracy Analysis

With best parameters and using all the features with cross validation, our model gives a 0.852 accuracy for the train set, it shows that 85.2 percent of the time, the model correctly predicts whether a person's income is over $50,000. The test set has an accuracy of 0.851 showing that our model is not overfitting.

After the feature selection, the new model has an accuracy of 0.8596 for the train set and an accuracy of 0.8528 for the test set. Both classification summaries are shown below. A value of 1 indicated the income is predicted as over $50K, while "0" means below $50K.

Training

Confusion Matrix (Accuracy 0.8596)

Actual	Prediction 0	1
0	10396	1019
1	1087	2498

Test

Confusion Matrix (Accuracy 0.8528)

Actual	Prediction 0	1
0	6914	687
1	785	1614

According to the classification summary table, the model is prone to make Type 2 error (false negative), so the proportion of false negative is more than that of false positive. The best model we reached will generate a false positive rate of 0.0893 while a false negative rate of 0.3032 for the train set; it will generate a false positive rate of 0.1020 while a false negative rate of 0.2986 for the test set. It is definitely more protective of banks, who would rather misclassify someone with >$50k as having < $50K (false negatives) than to misclassify someone with <50$K as having >$50K (false positives). This is worse for loan applicants, as the banks try to be more cautious than necessary when predicting someone's income.

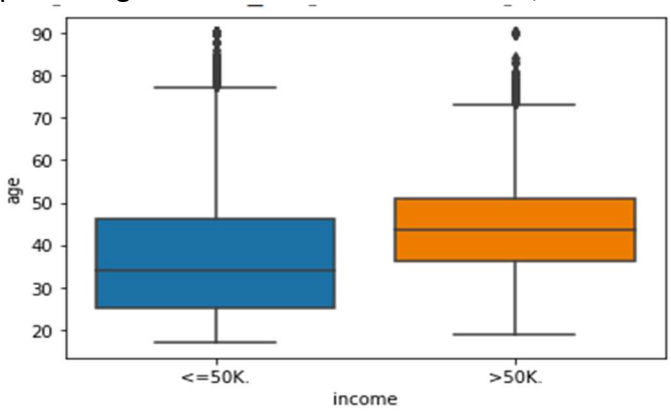
Top Three Categorical Predictors

Exec-Managerial, Prof-Specialty, followed by sales and Craft-repair are the occupations most associated with a predicted income over $50,000. This is also shown in the countplot graph in the Appendix section from the EDA earlier on. This is intuitive as managerial and executive positions tend to have more responsibilities and so they are compensated more. Professors that work in specialty areas also tend to earn more if there is a high demand for them in their field. However, after permutation importance analysis in our model, we find that Craft-repairing, other services, Machine-op-inspect, Handlers-cleaners, Adm-clerical were the occupations that had a larger effect on the predictions.

By our plots during EDA, the education level most associated with an income over $50,000 is bachelor followed by some college, high school and then masters. This plot is in Appendix. This is in line with the view that more education can lead to higher paying jobs and at some point, further return to higher education starts to diminish. NOTE: Later on, we removed education and used only 'education-num' instead for our model as the two variables are correlated. After sensitivity analysis and permutation importance analysis in our model we found that education-num, which shows the number of years of education, was important for predicting who made incomes above $50,000.

The age range that is most associated with above $50000 income is the range of ages 36 to 47. From the histogram in the Appendix, we can see the distribution of above 50k

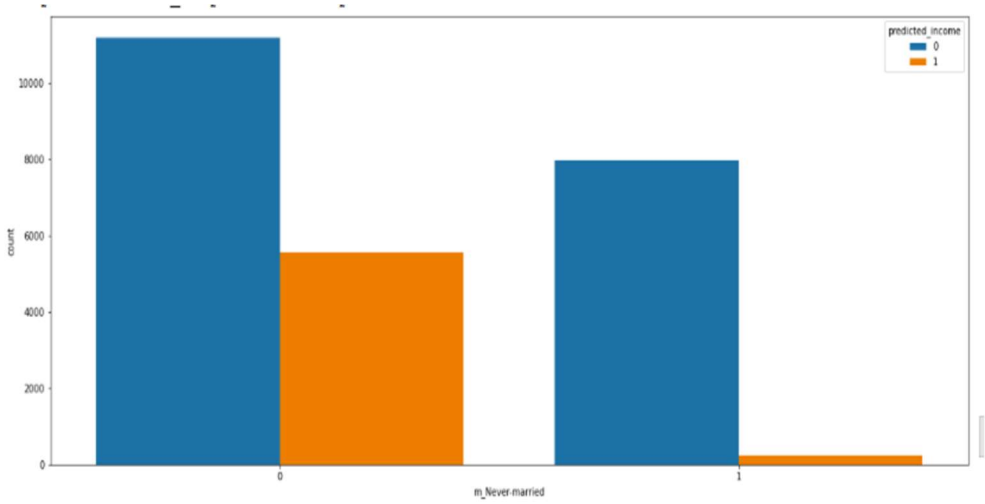
income earners vs that of below income earners ranging from different ages. This is also evident in the box plot below. The below graph shows that there is a higher concentration of higher income earners around older people. This implies that older ages are associated with higher incomes. This makes sense as we would expect people to advance in their careers with time and as such, achieve higher levels of pay. Also, later on in life, people retire so we would expect their income to reduce again. After sensitivity analysis and permutation importance analysis in our model, we found that the variable age was important for predicting who made incomes above $50,000.



The top 3 categorical variables according to our permutation importance analysis are: relationship_Never-married, relationship_husband, and marital-status_divorced.

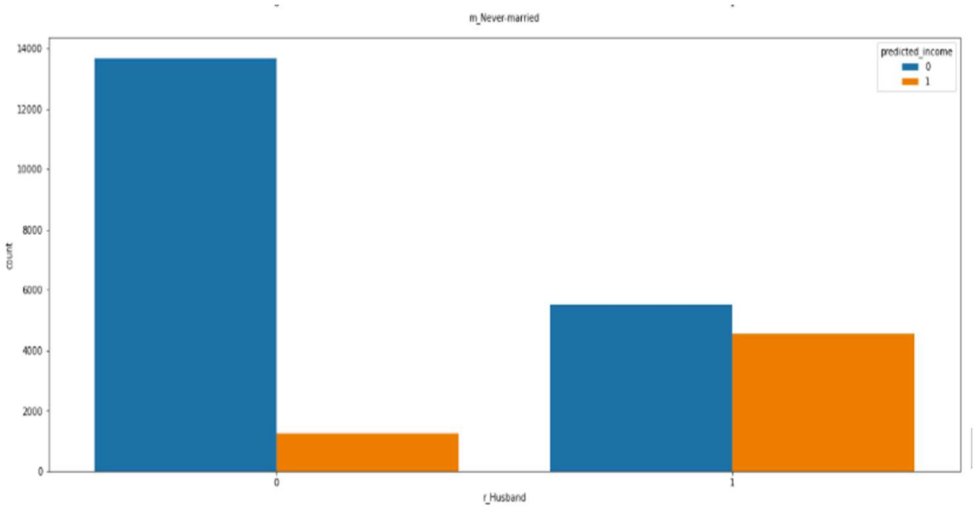
Relationship_Never-married

Below graph implies that never married people are less likely to have incomes above $50000



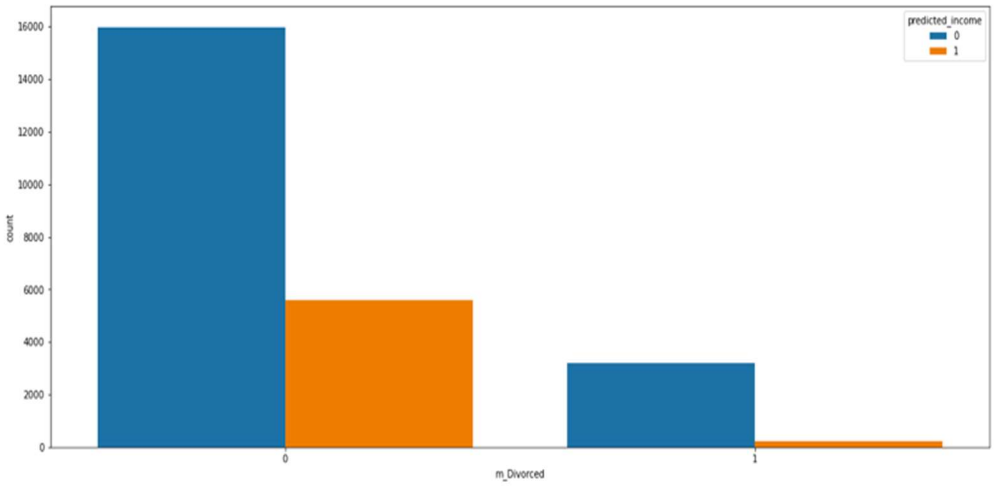
Relationship_husband

Below graph implies that those who fall under the husband category are more likely to have incomes above $50000.



marital-status_divorced

Below graph implies that those that have been divorced less likely to have incomes above $50000

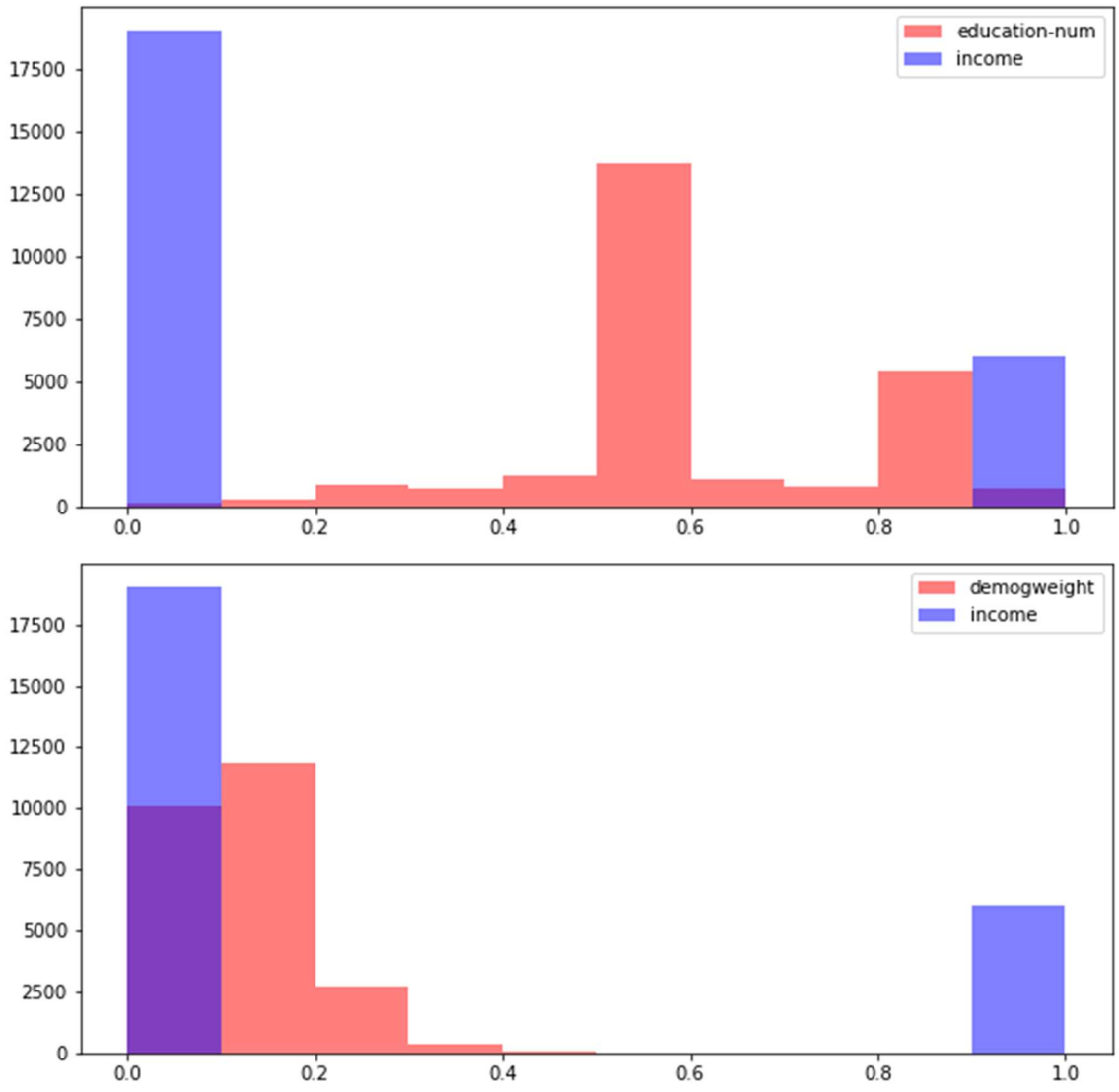


Histogram Analysis

Based on sensitivity analysis and the importance weight calculated, we select education-num as the numeric variable that is important in the model; and select demogweight as the numeric variable that is not important in the model.

The following graphs show the histogram between selected variables and income. Apparently, histograms support the findings of the neural network. When education-num is large, which means the individual is more educated, it is more likely to have an income of over $50,000 per year. For demogweight, when the individual has an income of over $50,000 per year, there are no demogweight data at all. So, this variable is not related to

income at all. Even though demogweight is relatively important based on sensitivity analysis, it is not based on important weight levels, which is the same as the graph shown below.



Conclusion

After conducting feature selection using sensitivity and ELI5, as well as parameter tuning, we arrived at a model with an accuracy of above 85% for both the training and test set. We found that the most important categorical variables were never-married, relationship-husband and divorced. The relationship between these variables and having over $50,000 are in line with intuitive thinking as we would expect married over never-married to have higher incomes as they can pool their efforts. Also, the husband category is more associated with higher incomes which is not that surprising as there are some

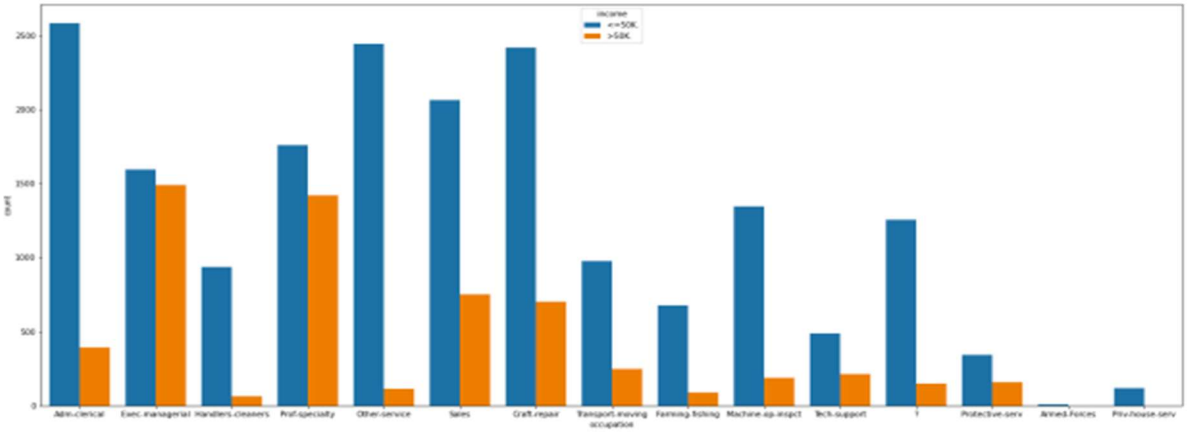
families with male breadwinners. Finally, we would expect non-divorcees to have higher incomes as divorces usually involve sharing of incomes. In terms of numerical variables, Education-num was the most important for our predictions which is in line with the idea that there are returns to education that lead to increased salaries. Reducing the number of variables dealt with issues such as multicollinearity and the curse of dimensionality. A limitation we had in this assignment was the lack of computing power making it near impossible to experiment with a large range for the parameters chosen.

Part Two: Test Data Set

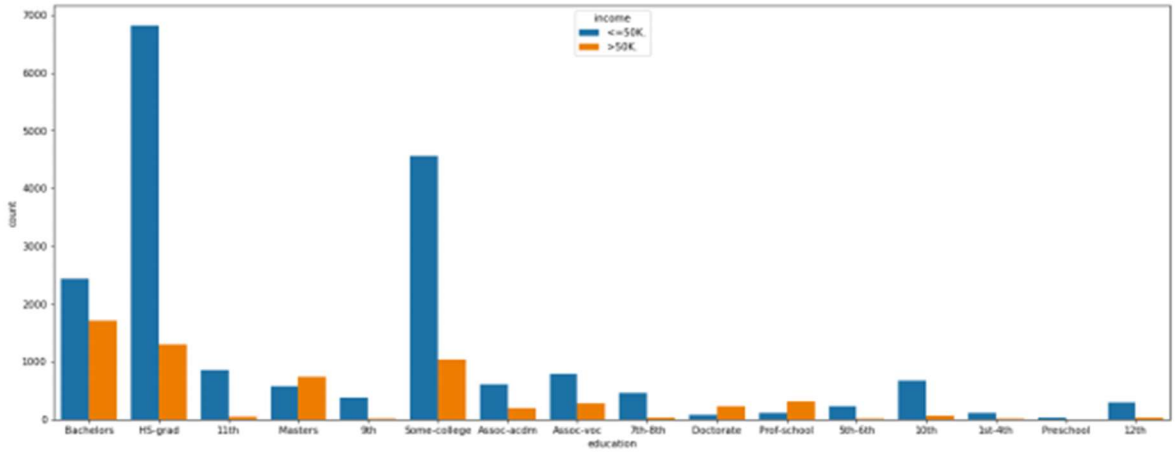
For part two, we used the final model to predict income greater than 50K and saved the numpy array into a text file. We did the same data preprocessing as before to make sure the model can be used. Additionally, we loaded the text file back into the code for future reference when testing the accuracy.

Appendix

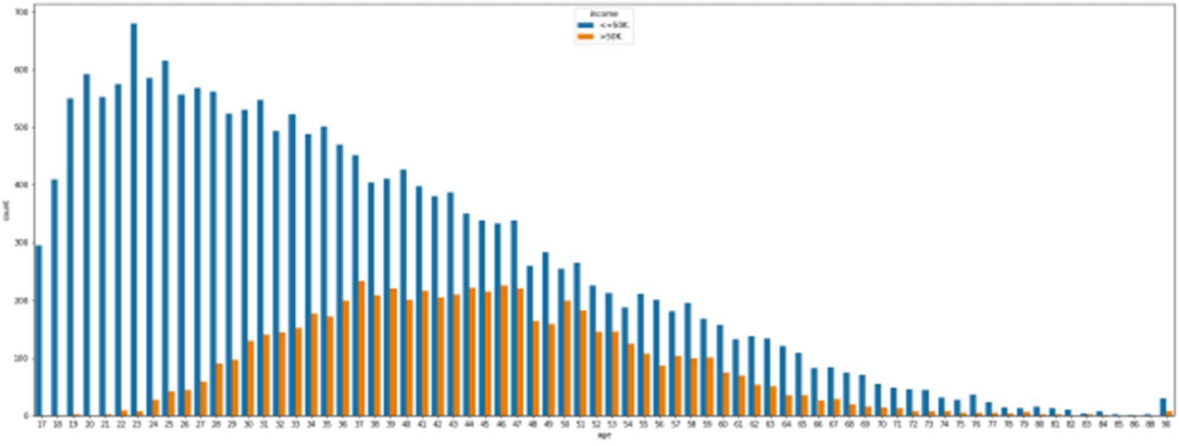
Occupation Countplot



Education Countplot



Age Countplot



Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Final Page

Grade: _____