Rotman

Executive Summary

This report explores the development and evaluation of an Artificial Neural Network (ANN) model to predict whether an individual's income exceeds $50,000 per year based on demographic attributes from a US Census Bureau dataset. The preprocessing phase ensured data quality by addressing redundancy, consolidating categorical predictors, handling missing values, and scaling inputs. A single-hidden-layer ANN with optimal hyperparameters (64 neurons, ReLU activation, Adam optimizer) was tuned via grid search and cross-validation, achieving an 84.81% accuracy. Sensitivity analysis identified "capital-net," "education-num," and "hours-per-week" as the most impactful predictors, underscoring their intuitive association with income levels. While the model demonstrated strong overall performance metrics, including 84% accuracy and a 0.89 ROC-AUC score, its tendency toward false negatives suggests potential biases against high earners. Insights drawn from important categorical variables like education, marital status, and occupation align with real-world trends, providing actionable guidance for stakeholders seeking data-driven income predictions.

Rotman

Section 1: Data Preprocessing

The dataset was split into 25,000 records for training and 7,561 records for testing. Each record represents an individual's information from the US Census Bureau, with 14 demographic attributes and one target variable indicating whether income exceeds $50K per year. The goal is to identify demographic characteristics that best predict income using an Artificial Neural Network (ANN) model. Before modeling, extensive preprocessing was conducted to ensure data quality and compatibility.

Sec 1.1: Eliminating Redundant Attribute

The attributes *"education"* and *"education-num"* represent the same concept. The attribute *"education-num"* encodes the education level using ordinal numbers, as shown in **Figure 1**. To avoid redundancy and overemphasizing the effect of an individual's education level, we removed the *"education"* column and retained *"education-num"*, treating it as an ordinal categorical variable.

Sec 1.2: Consolidating Levels of Categorical Predictors

To prevent high dimensionality from excessive dummy variables, we merged minority levels (fewer than 100 records out of 25,000 in the training set) into the mode class. For instance, in *"workclass"*, levels like *"Without-pay"* and *"Never-worked"* were merged into the mode class *"Private"*. Similarly, levels like *"Married-AF-spouse"* in *"marital-status"* and *"Armed-Forces"* in *"occupation"* were consolidated into their respective mode classes.

For *"native-country"*, most records (22,421 out of 25,000) were from the United States, 488 were from Mexico, while other countries had fewer than 100 records each. To address this, we grouped countries by continent, except for the United States and Mexico, to create a new variable, *"native-origin"*. This transformation reduced the levels of this predictor from 42 to 5, with each level containing over 400 records in the training dataset (as shown in **Figure 2**).

Sec 1.3: Handling Missing Values

The dataset contains missing values in both categorical and numeric predictors, indicated by "?" and "99999," respectively. For categorical variables such as *"workclass"* and *"occupation"*, which have over 1,000 missing records in the training set, mode imputation was avoided to prevent overrepresentation of the mode class. Instead, a proportional distribution approach was used, imputing missing values based on the observed distribution of other levels in the dataset. This ensured a balanced representation across categories. For numeric predictors *"capital-gain"* and *"capital-loss"*, which represent an individual's return on capital, a distinct pattern was observed: when *"capital-gain"* was missing (126 cases), *"capital-loss"* consistently had a value of zero. Leveraging this relationship, we combined these two variables into a single variable, *"capital-net"*, representing net capital return. In cases where *"capital-gain"* was missing, *"capital-net"* was set to zero since *"capital-loss"* was also zero.

This transformation simplified the data and retained its interpretive value. The distribution of the new variable *"capital-net"* in the training set is illustrated in **Figure 3**.

Sec 1.4: Input and Output Encoding

To prepare for the ANN model, all inputs and outputs were scaled between 0 and 1. Dummy variables were created for all nominal predictors including *"workclass"*, *"marital-status"*, *"occupation", "relationship", "race", "sex"* and *"native-origin"*, increasing the number of input variables to 41. The binary target variable *"income"* was encoded as 1 for incomes over $50K and 0 otherwise.

Numeric predictors (*"age"*, *"demogweight"*, *"capital-net"*, *"hours-per-week"*) and the ordinal variable *"education-num"* were normalized using max-min scaling to ensure their values fell within the range of 0 to 1.

Section 2: ANN Model Fitting

After completing data preparation, all 41 input variables and the binary target variable were successfully normalized to values between 0 and 1. The goal was to predict whether an individual's income exceeds $50K per year using the 41 input variables with an Artificial Neural Network (ANN) model. To estimate the model's predictive performance, the original training dataset with 25,000 records was further split into a training set and a validation set in an 8:2 ratio. This allocation reserved 20% of the training data for assessing model performance, while the remaining 80% was used for model fitting.

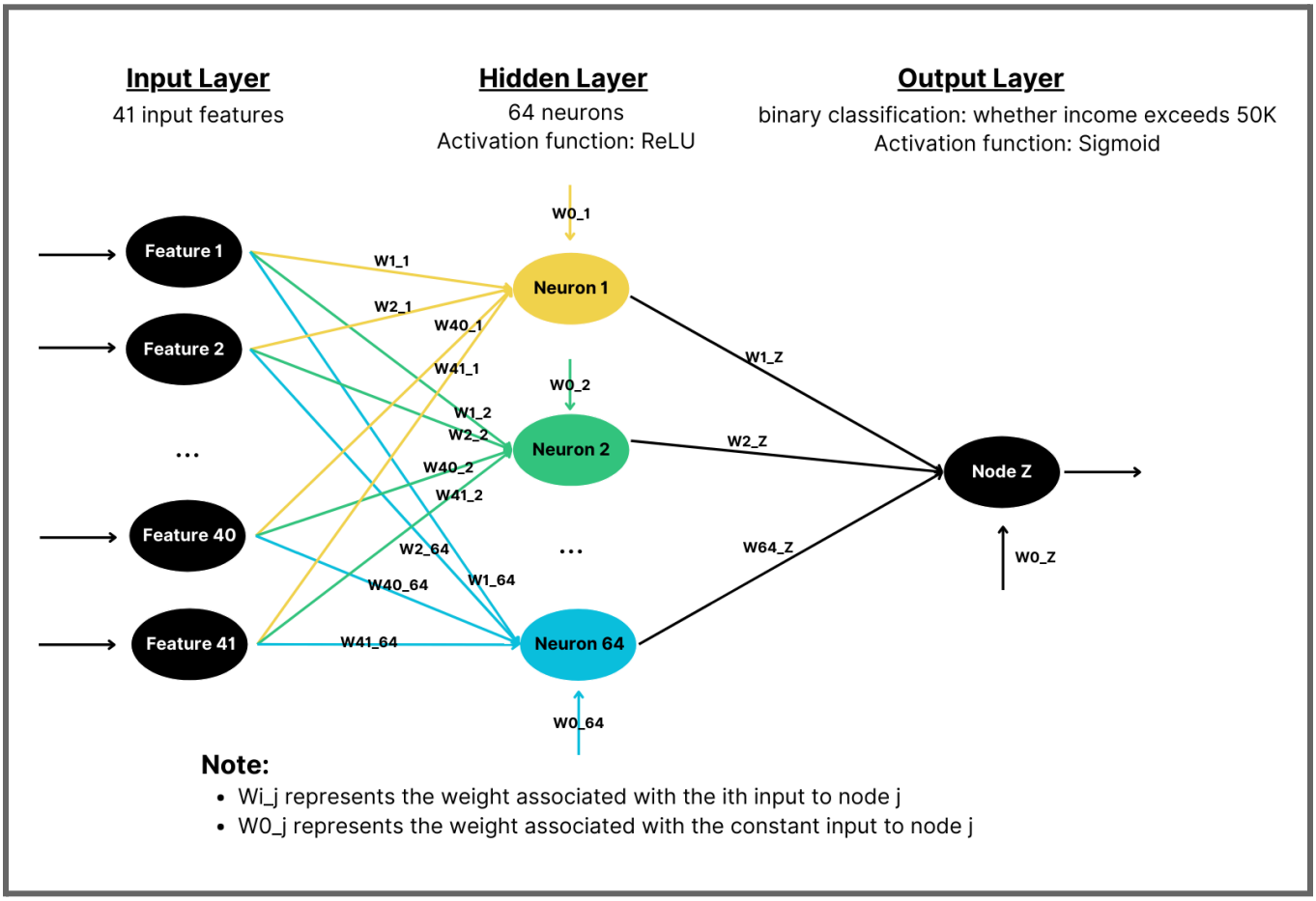
Our objective is to identify the key demographic characteristics for predicting income. To maintain model interpretability, the ANN model was designed with a simple architecture: one input layer with 41 features, a single hidden layer, and an output layer containing a single node for binary classification using the Sigmoid activation function. To optimize the model's performance, grid search with 5-fold cross-validation was employed to tune the following parameters:

- Number of neurons in the hidden layer: 32, 64
- Activation function in the hidden layer: ReLU, Sigmoid
- Optimizer: RMSprop, Adam (used to adjust weights and biases during training to minimize the loss function)
- Epochs: 10, 20 (number of complete passes through the entire training dataset)
- Batch size: 64, 128 (number of training examples processed together in a single forward and backward pass). Larger batch sizes were chosen to accommodate the relatively large dataset.

The best estimated test accuracy from 5-fold cross-validation is 84.8%, achieved with the following optimal parameters:

- Number of neurons in the hidden layer: 64
- Activation function in the hidden layer: ReLU
- Batch size: 64
- Epochs: 20
- Optimizer: Adam

The topology of the resulting network, based on the optimal parameters, is as follows:



- Input Layer: Consists of 41 input features (*Feature1, Feature2, ..., Feature41*).
- Hidden Layer: Comprises 64 neurons, fully connected to the input layer and utilizing the ReLU activation function (*Neuron1, Neuron2, ..., Neuron64*).
- Output Layer: Contains a single node for predicting the binary target variable, using the Sigmoid activation function (*Node Z*).

After selecting the optimal parameters, the model was fitted using the training set, and evaluated the model's performance using the testing set in the next step.

Section 3: Predictive Performance

Sec 3.1: Variable Importance

To identify the most important variables in the model for predicting income, we analyzed which features drive the predictions most strongly. First, we calculated a baseline prediction by setting all input features to their mean values and obtaining the model's output. Then, by varying each feature in the test dataset from its mean to its maximum and minimum values while keeping all other features constant, we performed sensitivity analysis to determine the most influential predictors in the neural network model. The impact of these changes on the model's predictions was measured and recorded in terms of 'Max vs. Mean,' 'Min vs. Mean,' and 'Max vs. Min,' where it is notable that 'Min vs. Mean' + 'Max vs. Mean' equals 'Max vs. Min.'

The results from our sensitivity analysis are presented in the appendix, **figure 4**. We observe that the variable '**capital-net'** is the most sensitive, with a 'Max vs. Min' score of 0.95, demonstrating extreme sensitivity even when the range is normalized to 1. This variable is the combination of Capital Loss and Capital Gain, representing an individual's return on capital. Its high sensitivity is logical, as before normalization, the capital-net variable was set to zero if Capital Gain was missing. If Capital Loss had a value, capital-net was adjusted by subtracting Capital Loss, and if Capital Gain had a value, capital-net was equal to Capital Gain. These operations resulted in a wide range for this variable, explaining why it is the most sensitive in the analysis.

Following 'capital-net,' other variables identified as sensitive and important include 'education-num,' 'hours-per-week,' and 'relationship-own-child.' Intuitively, it makes sense for these variables to play a significant role in predicting an individual's income category. People with higher education levels, more working hours per week, and those with children are likely to be in later stages of their lives, where earning a higher income is typically expected. These factors reflect their increased responsibilities and opportunities for higher-paying roles.

Sec 3.2: Model Results

Sec 3.2.1 : Classification Matrix

To assess the model's performance, we ran a confusion matrix on the 20% test data from our initial split before training the model. Referring to **Figure 5**, we observed a sensitivity of 0.54, a specificity of 0.94, a false positive rate of 0.05, and a false negative rate of 0.46.

The model's higher rate of **False Negatives (FN)** indicates that individuals earning >50K are often mistakenly predicted as earning ≤50K. This could result in banks denying credit or loans to eligible individuals. While this is a **conservative error** that helps protect banks from lending to risky applicants, it may negatively impact customer acquisition. Conversely, the model's low **False Positive (FP)** rate means fewer individuals earning ≤50K are incorrectly predicted as earning >50K, further protecting the bank from financial risks.

Overall, the model is more prone to False Negatives (FN), which benefits banks by reducing financial risks but can disadvantage qualified loan applicants. In terms of classification, the model correctly identified 54% of individuals earning >50K and 94% of individuals earning ≤50K.

Sec 3.2.2 : Other Metrics

We also evaluated the model's performance using additional metrics (refer to **Figure 6,7**). The model achieved **84% accuracy**, **76% precision**, **63% F1 Score**, and a **0.89 ROC-AUC score**.

- 84% of all predictions made by the model (both for ≤50K and >50K) were correct.
- 76% of individuals predicted to earn >50K actually earn >50K.
- 63% represents the balance between how well the model identifies actual >50K earners (recall) and how often its >50K predictions are correct (precision).
- 89% probability of correctly ranking a randomly chosen >50K earner higher than a randomly chosen ≤50K earner.

Sec 3.3: Detailed Variable Analysis against Predicted Income

The model built in the previous step with one hidden layer is used for analysis in this step. To find out the association between certain variables and predicted income over $50,000, the max and min value for each column is identified. Percentages of target value = 1 (i.e. income > $50,000) for max and min are calculated respectively. The larger the percentage change, the more sensitive the variable is, thus more association between the variable and the income.

From **Figure 8**, it shows the percentages for income greater than $50,000 for max and min values with each categorical variable ordered from highest to lowest. Note that for categorical variables, max value would always be 1 and min values would always be 0. That is, whether the person is under this specific category or not.

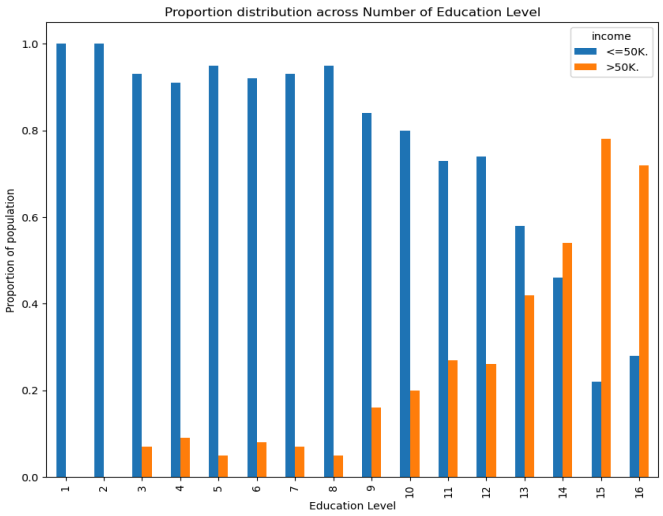
- **Occupations:** By looking only at occupations from **Figure 8**, occupations with high percentage changes are Exec-managerial, and Prof-specialty, with a percentage change of income > $50k of 34.02%, and 22.261% respectively. The result is intuitive as people who are executives or at manager level make more money in a company than others in the real world. People who are in prof-specialty (i.e. professors in colleges and universities) also have a high salary.

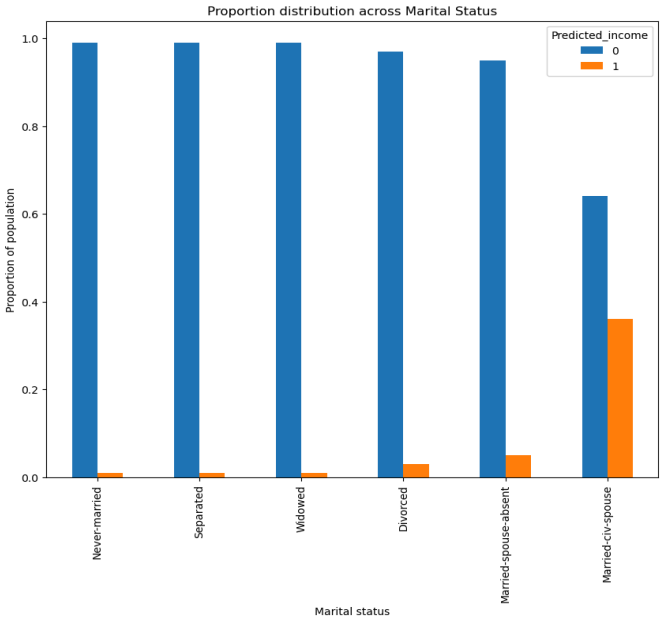
- **Education levels:** From **Figure 9**, the table shows the percentages for income greater than $50,000 for each education level ordered from highest to lowest. Since education level and education_num have a one-to-one relationship, the table is built off education_num. Based on the table, it is obvious that the longer the years of education (i.e. the higher the education level), the higher the percentage of predicted income is greater than $50,000. There is a big gap starting in year 13 which is Bachelor's degree, followed by Masters (14 years), Doctorate (16 years), and Prof-school (15 years). This is intuitive as people with a higher degree tend to have a higher chance to make more money in the real world.

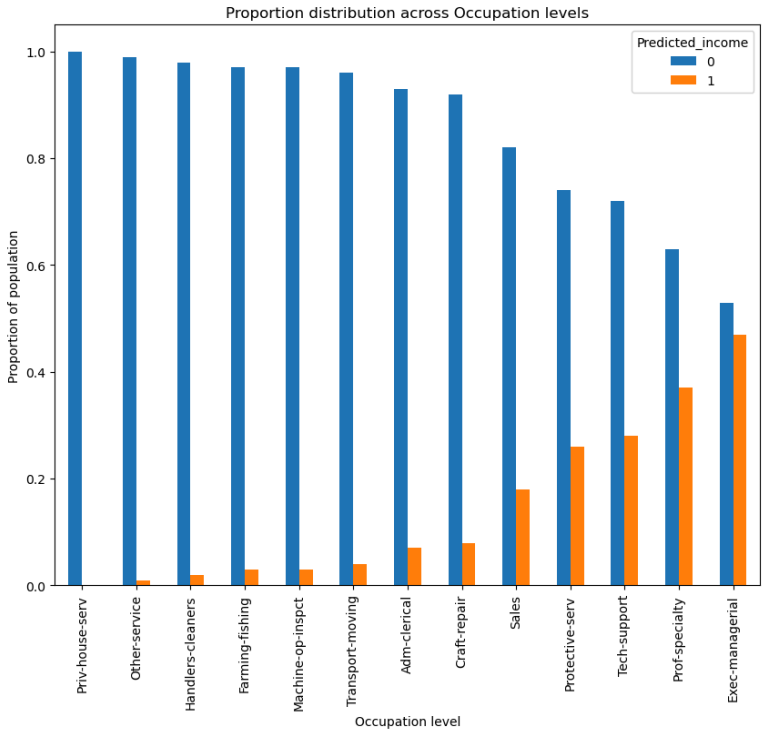
- **Age:** From **Figure 10**, the table shows the percentages for income greater than $50,000 for each age ordered from highest to lowest. Even though the range for age varies, people who are older than 40 years old generally make more money, especially for people that are in their 40s and 50s, which is an intuitive result because many people who are at a higher level in a company or own a company are around this age.

Sec 3.3.1 : Top Three Categorical Variables

Figure 11 contains proportion distribution plots for each categorical variable with predicted income as hue. Among all the categorical variables, the top three predictors are as follows:





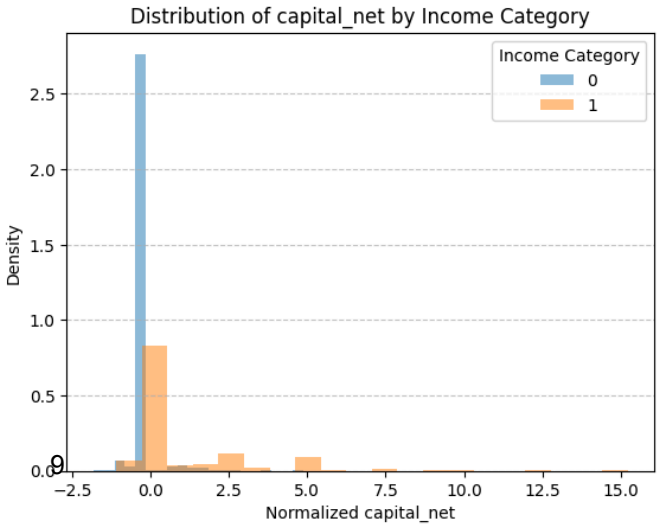


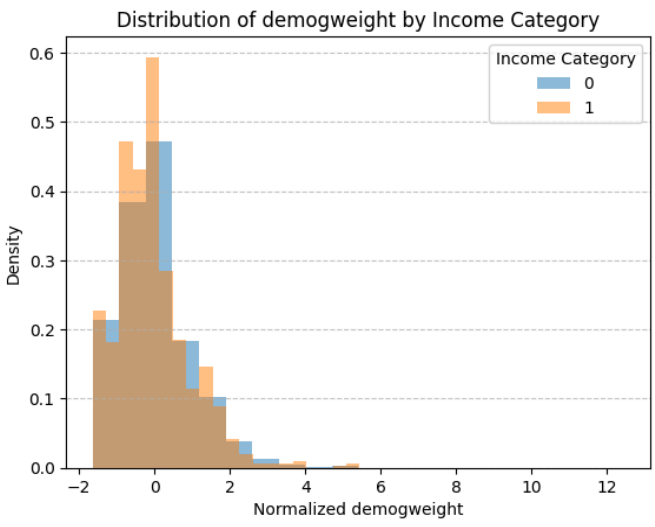
- **Education Level:** From the plot, it is obvious that the higher the education level is, the more likely the person's income is higher than $50k. There is a significant increase in the proportion of people with income greater than $50k when their degrees are Bachelor and above. Most of the people with a low education level have an income less than or equal to $50k.
- **Marital status:** From both marital status and relationship plots against predicted income, it is clear that almost all people who are not married earn less than or equal to $50k. Being married with spouse present significantly increases the proportion of income greater than $50k.
- **Occupation:** From the plot, people who are in Other services and Private house services make less than or equal to $50k. For people who are in the category of manual labor such as Handlers-cleaners, Machine-op-inspect, etc., most of them make less than or equal to $50k. There is a higher proportion of people with income higher than $50k when they are executives, at manager level or are Prof-specialty.

Sec 3.3.2: The Most and Least Important Numeric Variables

To get the importance of each numeric variable in the model, we first extract the importance calculated by weights of each variable (figure 12) from the model. Then, a SHAP summary(figure 13) diagram has been constructed. After validating the min-max analysis in the previous section, the importance of each variable and the SHAP summary, the most important variable is capital-net, and the least important variable is demogweight.

After normalizing the data, two histograms were generated. The first histogram, depicting the distribution of capital-net, reveals a distinct separation across income categories. This highlights the significant influence of capital-net on income distribution. The relatively clear separation of distributions between higher and lower income groups suggests that capital-net is a valuable feature for classifying individuals by income level. In contrast, the second histogram, illustrating demogweight, shows substantial overlap between the income categories. This aligns with the findings from the artificial neural network (ANN), indicating that demogweight has limited predictive value for income classification.





Conclusion

This study successfully implemented an Artificial Neural Network (ANN) to classify individuals' income levels based on key demographic variables, addressing significant challenges in data preprocessing, model optimization, and interpretability.

The preprocessing steps ensured a robust dataset by reducing redundancy, managing missing values through distributional imputation, and consolidating categorical predictors to minimize dimensionality. Normalization and encoding ensured compatibility with the ANN, resulting in an efficient input structure with 41 features.

The ANN model, featuring a single hidden layer with 64 neurons and optimized parameters, achieved a test accuracy of 84.81% through rigorous grid search and cross-validation. Performance evaluation highlighted the model's strong classification capabilities, with high specificity (94%) and precision (76%), but moderate sensitivity (54%). This reflects its conservative bias, favoring false negatives over false positives—a scenario advantageous for risk-averse applications like credit assessment but potentially restrictive in customer acquisition.

Key insights from sensitivity analysis revealed that "capital-net," an aggregated measure of capital gains and losses, was the most influential predictor. Other significant variables included education level, weekly working hours, and age, all of which align intuitively with income prediction. Categorical analysis further highlighted education, marital status, and occupation as critical indicators of income disparities, with advanced degrees and executive roles correlating strongly with higher income probabilities. Notably, individuals aged over 40 showed a marked tendency toward higher earnings, reflecting the accumulation of experience and seniority in professional roles. Numerical comparisons emphasized the predictive strength of "capital-net" in contrast to the low relevance of "demogweight", supporting its prioritization in model design.

Despite its strong performance, the model has limitations. Its high false-negative rate underscores a potential bias against identifying high-income earners, which may necessitate calibration for applications prioritizing equitable predictions. Additionally, while the ANN offers predictive power, its relatively simple architecture was chosen for interpretability, possibly at the expense of capturing complex interactions.

Rotman

Appendix

Figure 1 Mapping of *"education"* to *"education-Num"* Values

education	education-num
Preschool	1
1st-4th	2
5th-6th	3
7th-8th	4
9th	5
10th	6
11th	7
12th	8
HS-grad	9
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

Figure 2 New Attribute *“native-origin”* with 5 Levels

native-origin	count
United-States	22866
America	743
Asia	503
Mexico	488
Europe	400

Figure 3 Distribution of the New Attribute *"capital-net"*

	capital-net
count	25000.000000
mean	498.084160
std	2598.832402
min	-4356.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	41310.000000

Figure 4 Top 10 Variable Importance Plot

Ranking	Attribute	Max vs Min	Min vs Mean	Max vs Min
1	Capital-Net	0.825	0.125	0.95
2	Education-Num	0.436	0.163	0.599
3	Hours-Per-Week	0.363	0.129	0.492
4	Relationship-Own-Child	0.154	0.072	0.226
5	Relationship-Not-In-Family	0.132	0.087	0.219
6	Age	0.082	0.124	0.206
7	Race-White	0.035	0.166	0.201
8	Relationship-Unmarried	0.153	0.031	0.184
9	Occupation-Priv-House-Serv	0.172	0.005	0.176
10	Race-Black	0.146	0.025	0.171

Figure 5 Confusion Matrix

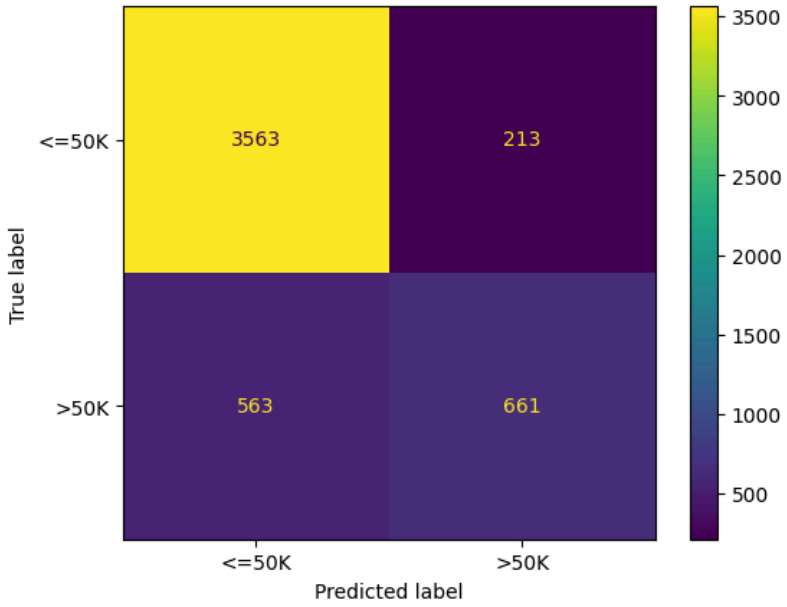


Figure 6 Model's Performance Metrics

Metric	Score
Accuracy	0.8448
Precision	0.7563
F1 Score	0.6301

Figure 7 ROC Curve

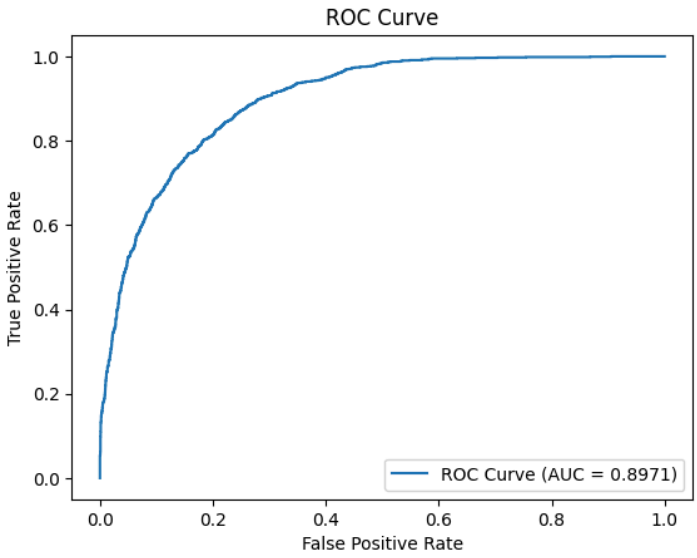


Figure 8 Percentage of Target Variable change for Max and Min values for Categorical Variables

(Note: all dummy variables range between a minimum value of 0 and a maximum value of 1.)

Col_names	Percentage of Income > 50k for maximum	Percentage of Income > 50k for minimum	Change in target variables when encoding changes
workclass_Self-emp-inc	56.25	16.161	40.089
marital-status_Married-civ-spouse	36.555	1.946	34.609
occupation_Exec-managerial	47.304	13.284	34.02
marital-status_Never-married	1.274	25.776	24.502
occupation_Prof-specialty	36.671	14.41	22.261
relationship_Own-child	0	20.937	20.937
relationship_Not-in-family	3.146	23.001	19.855
occupation_Other-service	0.353	19.914	19.561
occupation_Priv-house-serv	0	17.796	17.796
relationship_Unmarried	1.754	19.523	17.769
relationship_Wife	34.231	16.793	17.438
sex_Male	23.499	6.291	17.208
marital-status_Widowed	1.205	18.266	17.061
occupation_Handlers-cleaners	1.942	18.377	16.435
occupation_Machine-op-inspct	2.624	18.81	16.186
marital-status_Separated	2.069	18.167	16.098
relationship_Other-relative	2.128	18.152	16.024
native-origin_Mexico	2.222	17.984	15.762
occupation_Farming-fishing	2.649	18.169	15.52

race_Other	2.564	17.819	15.255
marital-status_Married-spouse-absent	4.054	17.905	13.851
occupation_Transport-moving	5.19	18.467	13.277
workclass_Private	14.421	26.926	12.505
native-origin_Asia	28.736	17.505	11.231
occupation_Protective-serv	28.736	17.505	11.231
occupation_Tech-support	28.276	17.384	10.892
race_Black	8.447	18.588	10.141
occupation_Craft-repair	8.861	18.979	10.118
workclass_State-gov	24.378	17.42	6.958
race_Asian-Pac-Islander	24.375	17.479	6.896
race_White	18.602	12.104	6.498
native-origin_Europe	22.581	17.607	4.974
workclass_Local-gov	20.649	17.486	3.163
native-origin_United-States	17.866	15.865	2.001
workclass_Self-emp-not-inc	16.169	17.834	1.665
occupation_Sales	19.031	17.533	1.498

Figure 9 Percentage of Target Variable change for Each Education Level

Number of education years	Percentage of Income > 50k at this age
15	79.167
16	70
14	56.809
13	40.732
12	26.543
11	17.157
10	10.607
1	10
9	5.127
8	3.279
3	2.273
7	1.613
6	0
4	0
2	0
5	0

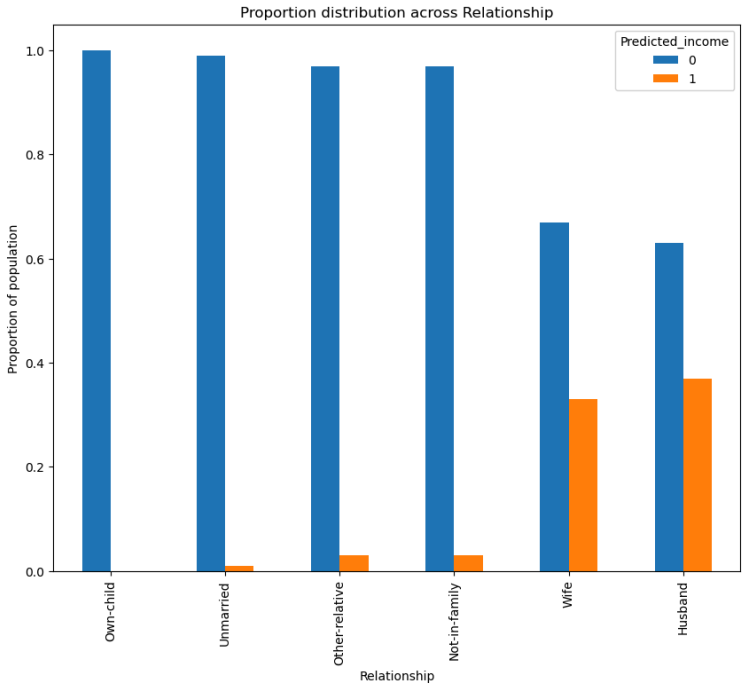
Figure 10 Percentage of Target Variable change for Each Age

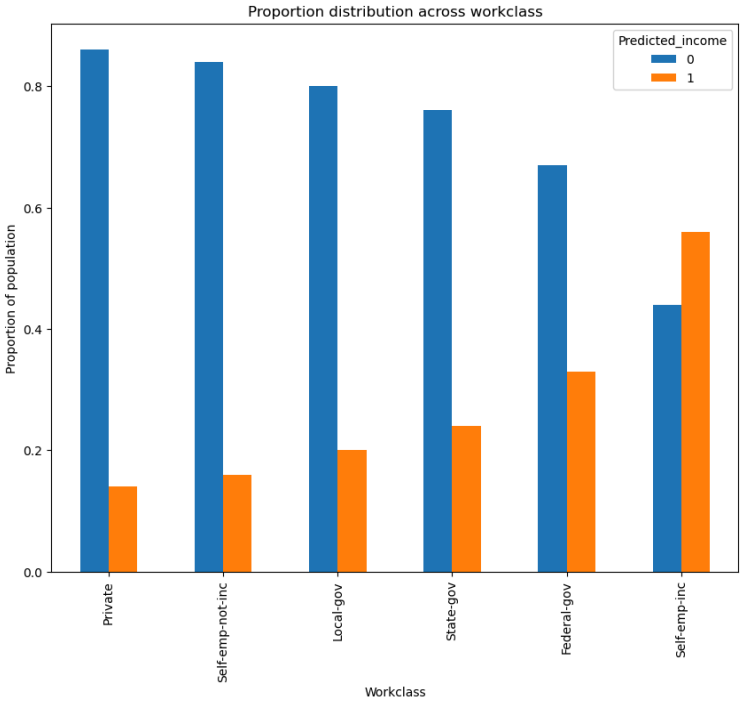
Ages	Percentage of Income > 50k at this age
79	40
57	38.333
45	35.514
42	34.483
65	34.375
44	33.636
68	33.333
50	32.258
53	32.143
43	31.897
59	31.111
38	30.973
61	29.412
52	28.916
71	28.571
49	27.381
41	26.891
51	26.829
37	26.357
39	25.581
54	25.424
48	25

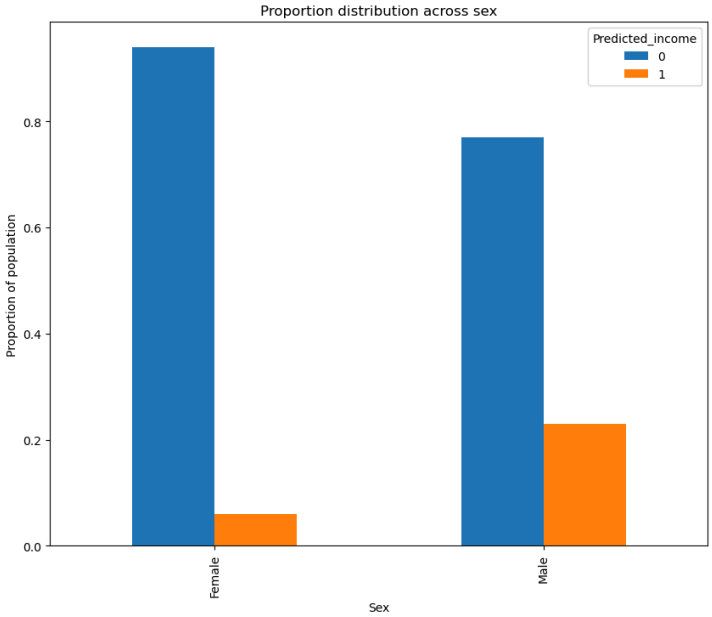
33	25
66	25
70	25
40	24.59
72	23.077
46	23.077
60	22.857
36	21.854
55	21.667
63	21.429
34	21.233
78	20
75	20
77	20
47	19.672
58	17.308
62	16.667
69	16.667
64	15.625
76	14.286
32	14.179
56	13.208
67	12.5

31	12.245
35	11.94
90	11.111
30	10.938
29	10
27	8.036
28	7.752
73	6.667
26	3.125
25	2.778
24	2.419
23	1.418
20	0.813
19	0
17	0
21	0
18	0
22	0
74	0
80 ~ 84	0

Figure 11 Proportion distribution for All Other Categorical Variables







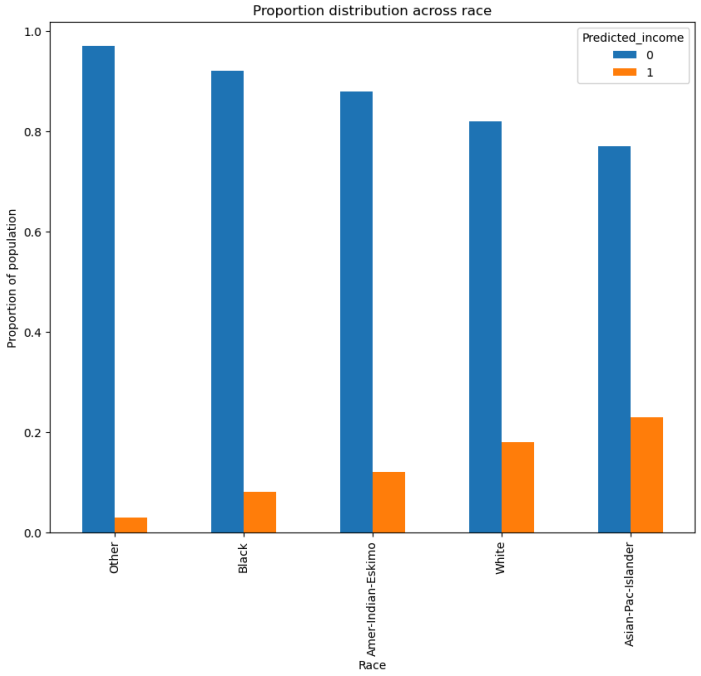


Figure 11 Continued Proportion distribution for All Other Categorical Variables

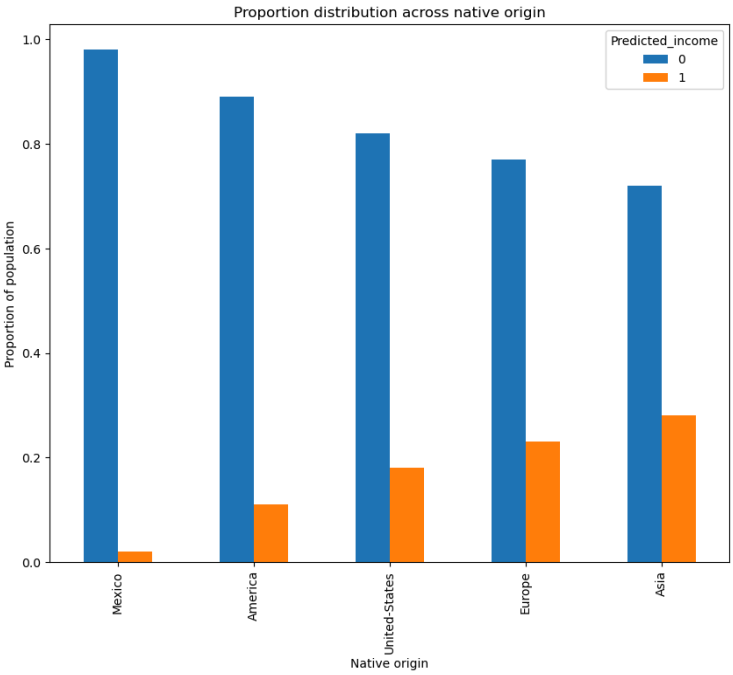


Figure 12 Numeric Variables Importance Calculated by Weights

Numeric Variable	Importance
capital-net	36.58
age	18.55
hours-per-week	12.01
demogweight	10.409

Figure 13 SHAP Summary