Rotman Commerce
UNIVERSITY OF TORONTO

Course Number: RSM 8413

Lecture Section #: 1

Course Name: Machine Learning Analytics

Professor Name: Prof. Gerhard

Assignment Title: Neural Netwotk Group Assignment

Team Name/Number**: 8

Team Members – FIRST NAMES ONLY:

Example:
1. Jennifer Brad

1. Tracy
2. Hans
3. Anna
4. Eric
5.
6.

In submitting this **team project or assignment** 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 the team is collectively responsible for assignments found to be plagiarized in any way, and every member will be subject to sanctions under the University's Code of Behaviour on Academic Matters.

Each group member must check the box below and record her/his student number to indicate that they have read and abide by the statements above. *Example:*

[✓] 999999900

[ ] 1005202157
[ ] 1005828197
[ ] 1011337938
[ ] 1006787746
[ ] 1004356779
[ ]

**** Team Name/Number**: If your instructor has assigned a unique name/number to your team, please indicate. If you do not have an assigned name/number, please identify one to facilitate the collection and return of your assignment.

RSM8413 Group Assignment 3

Hans Bisoo, Ruiyun Chao, Anna Mao, Eric Ni, Brad Zhang

Nov 22 2024

1 Introduction and Motivation

Census data aggregated by governments have long been used to uncover the changes in demographics and provide information about a population. In this paper, the team will review the anonymized individual-level census data to predict income classifications using a neural network. Information such as age, education, work and investing habits are made available for this classification task and can identify the drivers of higher income classification. Drivers such as education, work and investing habits are influenced by government policy, and can provide a data-backed framework for economic mobility.

2 Executive Summary

The construction and subsequent fine tuning of the neural network model was able to elevate the model accuracy from 84.9% to 87.5%. This presents a fairly performant model for income classification for future purposes. The key predictors from the anonymized census data are intuitive and correspond to investment and workplace metrics on the numerical side, and education, occupation and country of origin for the categorical side. According to comparison, neural network's findings are consistent with the actual data patterns. However, tuning the threshold to limit the false positives remains a domain-dependent endeavor and is further discussed within.

3 Data Preparation

Two datasets that contain the income classifications and 13 predictors from the US Census are provided for the endeavour. It has already been split into a testing and training set, containing 7500+ and 25000 records. The following will deal with the training set, while the testing dataset will be revisited in Part 2. A preliminary descriptive scan of the data is performed and it is observed that 6 predictors are numerical while the remainder are categorical. In sampling the data, it is observed that several columns contain a default value of "?" for missing data. The feature description is shown in Appendix Table 2.

Once turned into null values, the Missing Rate identifies that workclass, occupation, and native_country field have between 1.7% to 5.6% of its data missing. Given that all the affected columns are categorical, as seen from the number of unique values and the sample - categorical field imputation can be considered. In this event, maintaining the statistical distribution could be useful, and thus, the mode of all three affected fields are used to impute the missing values.

Upon examining education, and education_num, from the data dictionary provided that similar information

about a person's education background is provided across both fields. It is noted that both have 16 values, and we find that each education value corresponds exclusively to a given education_num - which provides the discrete years in education. As each education value is associated strongly to the education_num, we drop the education_num column and retain the categorical education roles in order to facilitate feature importance later in the process.

Given the model choice described below, the categorical predictors need to be transformed into numerical values using dummy variables, causing a sharp rise from 13 predictors to 97 predictors on account of each categorical value being exploded onto the dataframe.

The target variable income in the USTraining dataset is distributed in a 76% - 24% split across less than $50K and more than $50K respectively. While the dataset is not balanced, a useful distribution still exists, and can be used. This income column is transferred into 0 and 1 for ease of modelling where a binary value of 1 is assigned to classifications of where income is more than $50K, and 0 otherwise.

4 Part 1

4.1 Model Training

A Neural Net Model is considered to train and predict income classifications. As the backpropagation process within these models are sensitive to magnitude of values, the whole dataset is standardized using a min-max scaler ensuring that all numerical values are recast between 0 and 1, complementing the binary values of the categorical data.

In order to address the model's performance on unseen data, a train-validation data split is used where 70% of the USCensus_Train dataset is exposed to the model for training. A target variable distribution of both groups indicate that a representative sample is drawn from the entire training dataset. The train and valid(ate) portions show a 75% distribution. The initial neural network model is fit on the train portion.

The first neural network with the following simplified architecture is constructed:

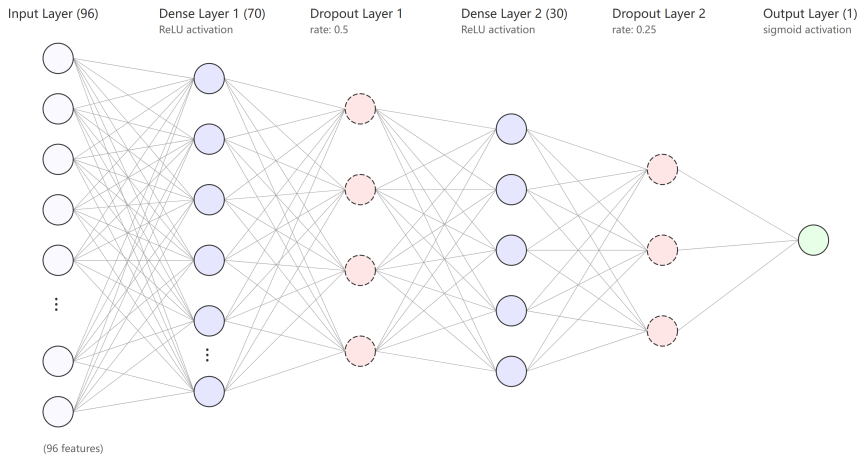

Figure 1: Simplified Neural Network Architecture for predicting income classification

The input layer consists of 96 nodes - one for each standardized predictor value. There are 2 sequences,

each consisting of a hidden layer and model dropout. The first hidden layer is set at 70 nodes, being fed information from the input node. The model dropout with a value of 0.5 drops half of the output from the hidden layer 1 before forwarding it to the second hidden layer. Hidden layer 2 is 30 nodes long and outputs to a second model dropout where a quarter of the inputs are forwarded to the output layer. The Rectified Linear Unit (ReLU) is used for both hidden layers on account of its performance in capturing complex interactions. The final output node however uses a sigmoid activation function to output the probability of being classified as 1 (indicating income over $50K). The benefit of a sigmoid function lies in its outputs being contained between 0 and 1 which can be categorized as 1 if it exceeds a threshold (0.5 in this case). This is most suitable for binary classification problems. Additional arguments such as a reduced learning rate during a plateau helps to find the optimal set of weights during training while the early stopping callback will exit the model training if no progress is made for the past 8 epochs. The model is run for 75 epochs with a batch size of 128. These arbitrary values are fine tuned in Part 2 below.

4.2 Feature Importance

For this section, we conducted a sensitivity analysis to assess the relative importance of each feature in the model. This method involves varying each feature within its observed range, from its minimum to maximum value, while keeping all other features constant. The sensitivity score for each feature is determined by the extent to which changes in the feature influence the model's predicted probability. Higher sensitivity scores indicate a greater impact on the model's predictions, providing a clear quantification of each feature's contribution. The 20 most influential features, as determined by this analysis, are presented in the accompanying bar chart.

The analysis revealed that capital_gain is the most influential variable for predicting whether an individual's income exceeds $50,000 per year, with the highest sensitivity score of approximately 0.68. This indicates that changes in capital gains have the most substantial impact on the model's predictions. The second most influential feature, capital_loss, showed a sensitivity score around 0.5, further emphasizing the importance of financial transactions in determining income level.

In addition to financial factors, hours_per_week emerged as a key predictor, which aligns with the expectation that individuals who work longer hours are more likely to earn higher incomes. Moreover, education, certain occupations and native countries were found to play a significant role, highlighting the correlation between demographics and greater earning potential.

Overall, these findings support socioeconomic theories, illustrating how education, occupation, and financial indicators influence income outcomes.

4.3 Model Accuracy

Our model achieved a training accuracy of 88.18% and a test accuracy of 84.96%. The slight difference of 3.22% between the two suggests that the model generalizes well to unseen data, with minimal signs of overfitting. The final test accuracy is determined by the total number of correct predictions, including both true positives and true negatives, divided by the total number of samples in the validation set. This means that in real-world scenarios, our model would have strong performance in accurately classifying samples.

4.4 Confusion Matrix

The model demonstrates a higher tendency to produce false negatives than false positives, with 746 false negatives compared to 382 false positives. Its precision rate is approximately 73.8%, meaning that when the

model predicts an income over $50K, 73.8% of these predictions are correct. However, the recall rate is only 59.1%, indicating that 59.1% of actual high-income cases (over $50K) are correctly identified. As a result, the model fails to capture a significant portion of true positives and exhibits a bias toward underestimating income, frequently predicting incomes below $50K.

When banks provide loan services to applicants, a model with high false negatives but low false positives can mitigate default risks by being more conservative in predicting high incomes. Additionally, a high precision rate allows banks to make accurate predictions for applicants with incomes above $50K, reducing the risk of approving ineligible applicants. However, the low recall rate implies risks of failing to recognize eligible high-income borrowers who could be profitable, potentially causing banks to miss out on valuable high-income loan applicants.

		Prediction	
		0	1
Actual	0	5296	382
	1	746	1076

Table 1: Confusion Matrix

4.5 Categorical Predictors of Income

To further investigate how the neural network determines the likelihood of individuals earning over $50,000, we will initially examine several pertinent variables. Based on Appendix Figure 9, the top three categorical predictors are occupation, education and native country. Therefore, we will focus on the three significant categorical variables, as well as age, the common geographical characteristics.

4.5.1 Occupation

According to Figure 2, the occupations of professor and executive manager are markedly associated with a predicted income exceeding $50,000, with each occupation accounting for over 31.6% of the total proportion. These are followed by sales, craft repair jobs, and administrative clerk, which represent 15%, 6%, and 5% respectively. This outcome aligns with conventional expectations, as professors, blue-collar workers, and white-collar workers typically command higher salaries.

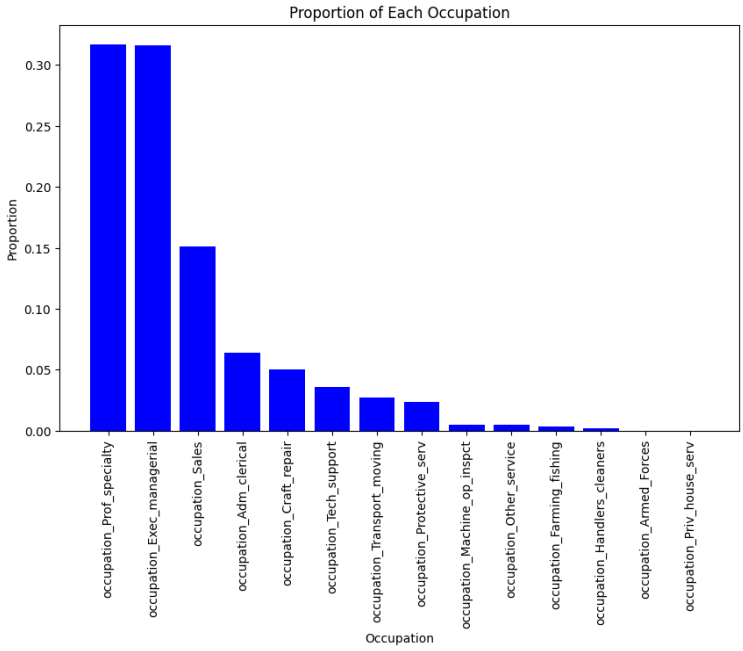

Figure 2: Proportion of Each Occupation

4.5.2 Education

Figure 3 demonstrates that a Bachelor's degree is most strongly correlated with a predicted income exceeding $50,000, accounting for approximately 35% of earners within this income bracket. Individuals with Master's degrees and those with college education are also significantly associated with higher income levels, each representing approximately 15.57% of earners within this income bracket. This is followed by high school graduates and professional school graduates, who constitute 12.7% and 8% of this group respectively.

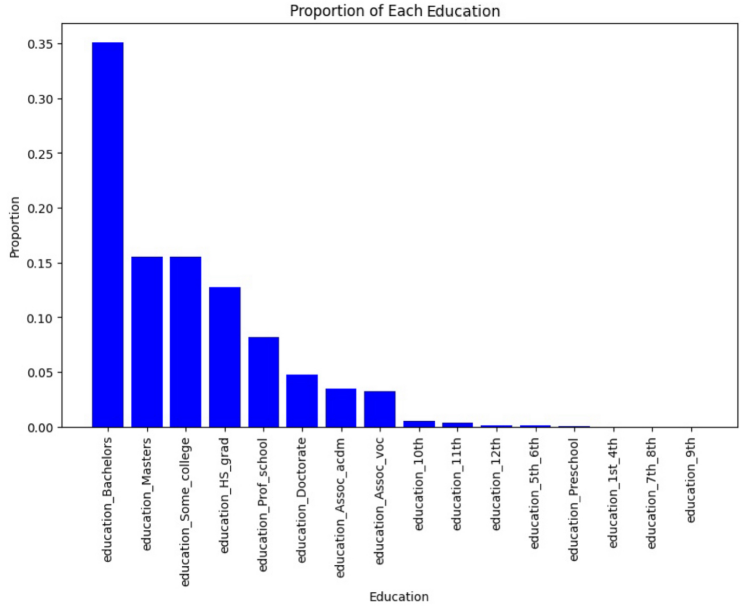

Figure 3: Proportion of Each Education

4.5.3 Native Country

Figure 4 demonstrates that the United States exhibits the strongest association with high-income earners compared to other native countries, accounting for over 90%. Other leading nations predominantly include developed countries such as Germany and Canada, as well as major developing countries like China and India.

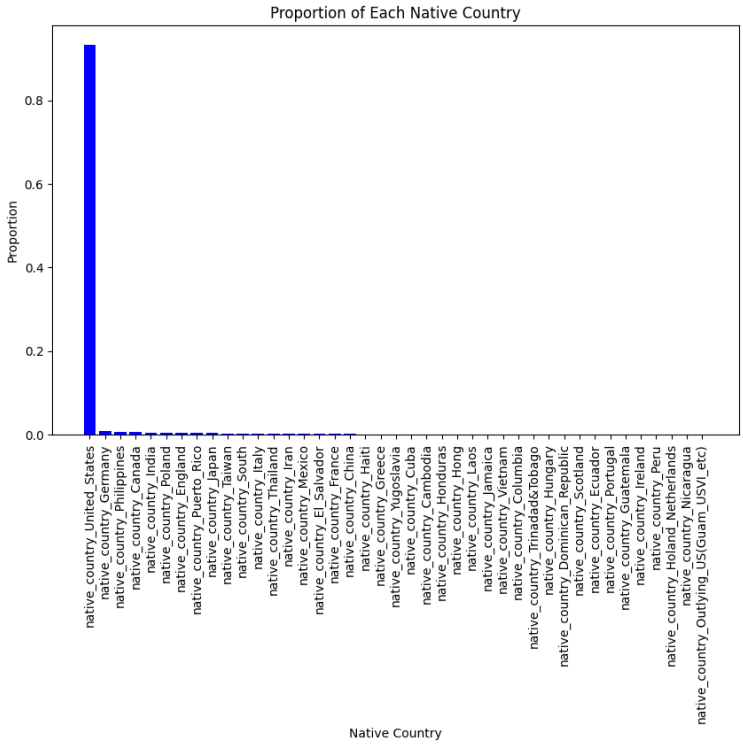

Figure 4: Proportion of Each Native Country

4.5.4 Age

Figure 5 indicates the relationship between various age groups and a predicted income exceeding $50,000, showing a bell-shaped curve. This curve indicates that individuals over the age of 60 captured in the green shared area constitute a smaller segment of the predicted high-income earners, reflecting a trend of diminishing income as age advances. Similarly, younger individuals under 30 in the orange shaded area represent a modest proportion of the high-income demographic, typically due to their engagement in entry-level positions that require further experience and time for career advancement. Conversely, the data reveals that the largest proportion of high-income earners is among individuals aged 39 to 50. Statistical summaries corroborate these observations, indicating that individuals aged 39 to 50 each year represent slightly over 3% of high-income earners. Specifically, individuals aged 41, 39, 50, and 43 each constitute 4% of this income bracket. These findings are consistent with expectations: middle-aged individuals, possessing peak productivity and greater corporate responsibilities, command the highest income proportions, whereas older individuals, potentially retired or exiting high-intensity roles, earn less.

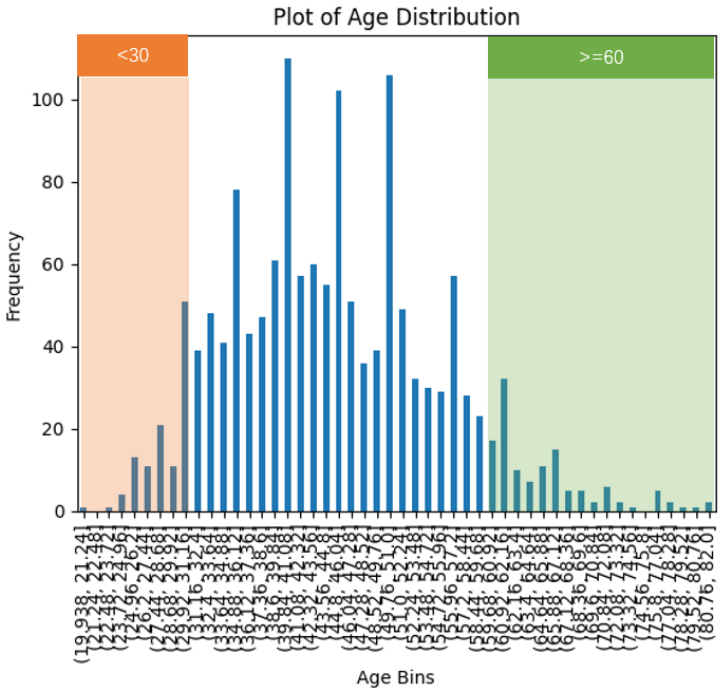

Figure 5: Plot of Age Distribution

4.5.5 Comments

However, the results are not immediately intuitive due to the imbalance in the database. Within each category, the distribution of various types varies significantly. The large proportion of predicted high-income earners in certain categories may be attributed to their prevalent representation in the validation set.

4.6 Numerical Predictors of Income

Based on Appendix Figure 9, capital gain is identified as the most significant feature. Conversely, demographic weight is considered a less important feature, as it does not rank prominently in the feature sensitivity analysis. We apply the original dataset to explore the relationship between income and the two numerical variables.

4.6.1 Capital Gain

Figure 6 displays the distribution of income across various capital gains bins, showing a clear and intuitive relationship with income. The distribution is notably imbalanced, underscoring its significance in predicting income levels. The plot reveals that individuals with capital gains below 5013 are predominantly low-income earners. Beyond the threshold of 5013, there is a significantly higher likelihood of being classified as high-income earners. The small proportion of the low-income population appearing in the final few bars may be attributed to their engagement in other business activities, such as investments, where income is not the primary source of remuneration. Thus despite registering high capital gains, they are classified as "low-income".

4.6.2 Demographic Weight

Figure 7 illustrates that the proportions of individuals earning above and below $50,000 remain relatively stable across demographic weights. Since the histogram indicates demographic weight has no strong pattern with income, this indicates that demographic weights do not significantly influence income predictions, as various income levels consistently exhibit similar proportions throughout most demographic categories, thereby offering limited utility in forecasting an individual's income level.

Therefore, we find that the comparison above demonstrates that the neural network's findings are consistent with the actual data patterns.

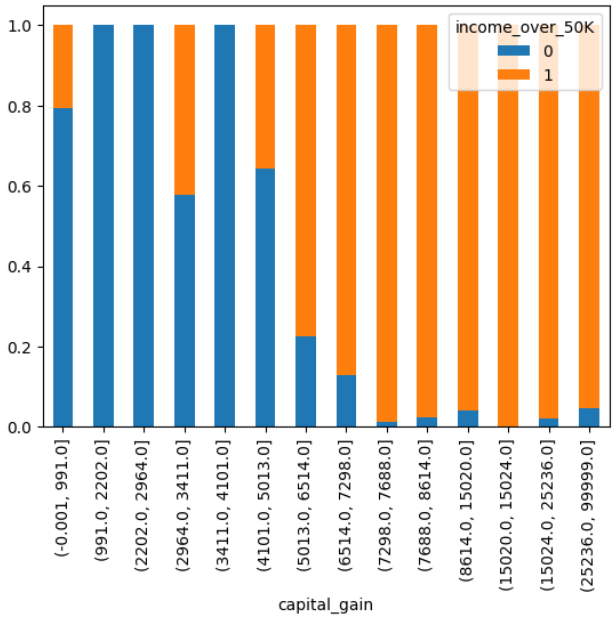

Figure 6: Capital Gain

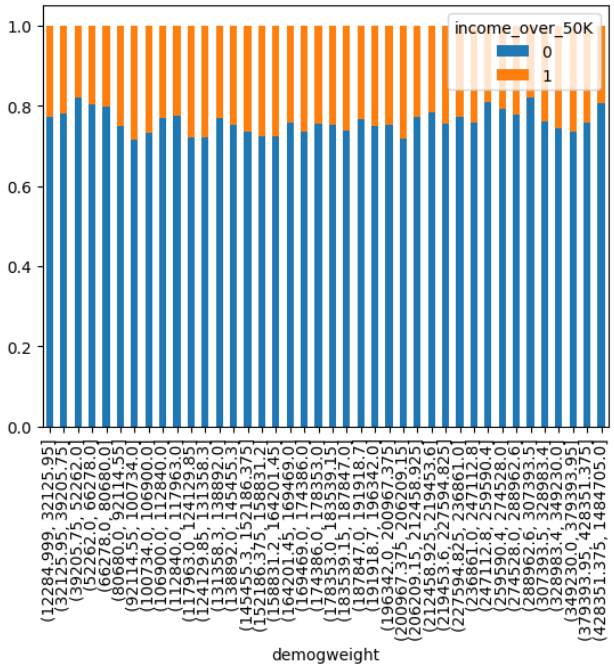

Figure 7: Demographic Weight

5 Part 2

5.1 Model Fine Tuning

To determine the optimal neural network architecture, we conducted an exhaustive hyperparameter optimization using k-fold cross-validation across the complete parameter space. The optimization encompassed architectural parameters (number of layers, neurons per layer), training parameters (batch size, epochs, learning rate), and regularization parameters (dropout rates). The hyperparameter search space was systematically explored through grid search with stratified 10-fold cross-validation to ensure robust model selection.

The optimal architecture converged to a two-layer neural network with 31 and 32 neurons in the first and second layers respectively, both employing ReLU activation functions. Regularization was implemented through two dropout layers with rates of 0.2 and 0.1, which proved effective in mitigating overfitting. The network was trained using the Adam optimizer with a learning rate of 0.001, batch size of 88, and maximum

epochs set to 54. To further prevent overfitting, we implemented early stopping with model checkpointing, which monitored validation loss and restored the model weights to the epoch with minimal validation error.

The optimized architecture demonstrated robust performance, achieving a mean classification accuracy of 85.4% (±0.5%) across validation folds. This represents a statistically significant improvement over our baseline models. Figure 1 presents a schematic visualization of the optimal neural network architecture.

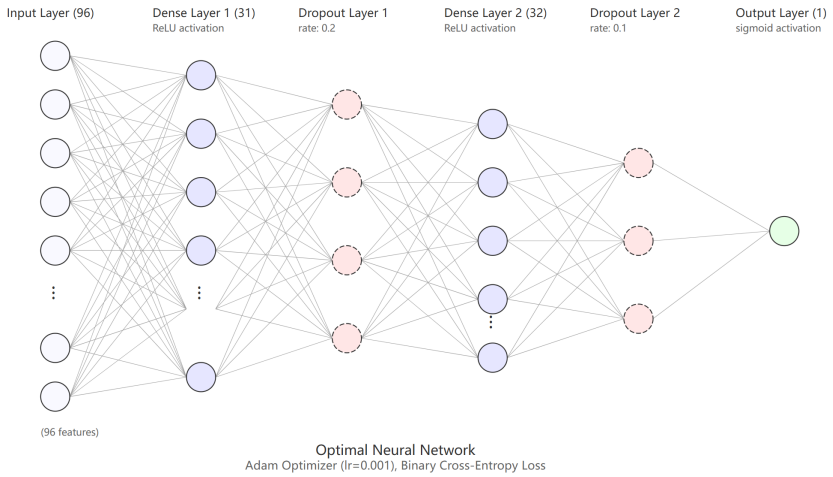

Figure 8: Fine Tuned Model for Income Classification

The model selection process was primarily constrained by computational resources and temporal limitations. Cross-validation for neural networks is computationally intensive, our available computational infrastructure limited concurrent model fitting to approximately 120 parallel processes per cross-validation iteration, significantly constraining our hyperparameter search space. Given access to high-performance computing infrastructure and extended computation time, a more comprehensive exploration of the parameter space would be feasible. This would allow finer discretization of parameter ranges, investigation of more complex architectures, and potentially yield marginally superior model performance.

5.2 Feature Importance for Fine Tuned Model

To evaluate feature importance and enhance model interpretability, we implemented a Classification and Regression Tree (CART) analysis. Feature importance scores were computed based on the Gini importance metric, which quantifies each feature's cumulative contribution to node impurity reduction across all tree splits. This approach was selected for its robustness and interpretability in handling both categorical and continuous variables, while providing insights into the model's decision-making process. The resulting feature importance rankings are presented in Appendix Table 3, ordered by their importance scores.

The feature importance rankings derived from the CART analysis provide empirical validation for the neural network architecture's design choices. The observed distribution of feature importance, where marital_status_Married_civ_spouse accounts for 39.94% of predictive power and only 11 features exceed 1% importance, suggests an effective dimensionality significantly lower than the raw feature count of 96. This empirical finding aligns with the architectural decision to implement 31 nodes in the initial hidden layer, providing sufficient capacity to capture the relevant feature interactions while maintaining computational efficiency. The implementation of dropout layers serves as an effective regularization mechanism, particularly given the highly skewed feature importance distribution, mitigating potential overreliance on dominant pre-

dictors. The dual hidden layer architecture, coupled with ReLU activation functions, facilitates the learning of complex non-linear relationships between the empirically significant features identified through CART analysis, such as the interactions between marital status, capital gains/losses, and educational attainment. This complementarity between tree-based feature importance and neural architecture provides methodological validation for the network's design, suggesting appropriate capacity and regularization for the underlying data structure's effective dimensionality. The alignment between CART-derived feature importance and neural network architecture offers robust evidence for the model's capacity to efficiently capture the inherent patterns within the dataset while maintaining resistance to overfitting through targeted regularization strategies.

6 Appendix

Feature	Nnull	Nunique	Total Rows	MissingRate	Dtype	Sample
age	0	72	25000	0.00000	int64	27
workclass	1399	8	25000	0.05596	object	Private
demoweight	0	17824	25000	0.00000	int64	187981
education	0	16	25000	0.00000	object	HS-grad
education_num	0	16	25000	0.00000	int64	9
marital_status	0	7	25000	0.00000	object	Never-married
occupation	1404	14	25000	0.05616	object	Handlers-cleaners
relationship	0	6	25000	0.00000	object	Own-child
race	0	5	25000	0.00000	object	White
sex	0	2	25000	0.00000	object	Male
capital_gain	0	117	25000	0.00000	int64	0
capital_loss	0	89	25000	0.00000	int64	0
hours_per_week	0	94	25000	0.00000	int64	40
native_country	445	41	25000	0.01780	object	United-States
income	0	2	25000	0.00000	object	<=50K

Table 2: Dataset Features Description

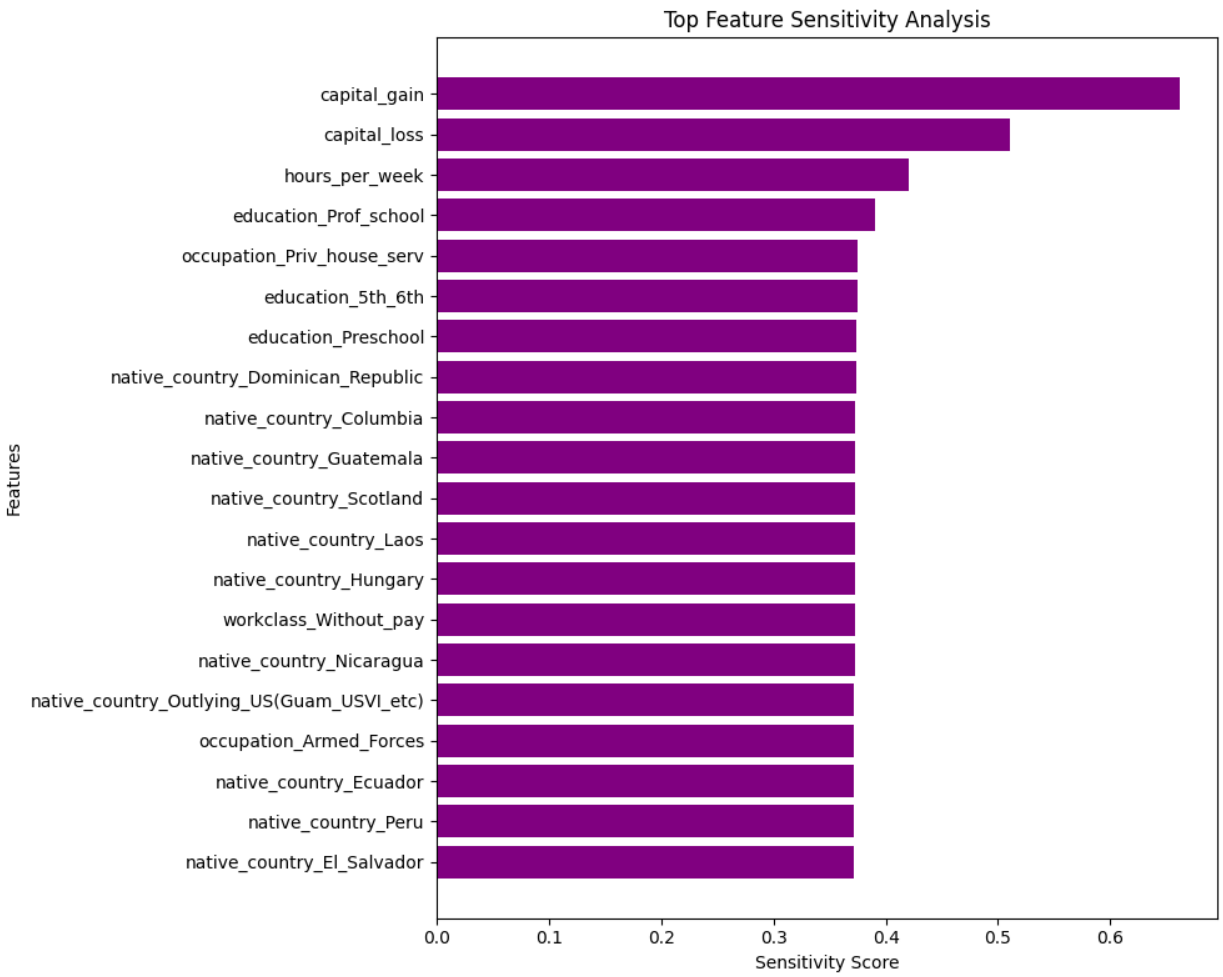

Figure 9: Sensitivity Analysis

Feature	Importance (%)
marital_status_Married_civ_spouse	39.9436
capital_gain	22.4421
capital_loss	9.1932
age	5.8332
education_Bachelors	4.9096
education_Masters	3.9824
hours_per_week	3.8197
occupation_Exec_managerial	1.8907
education_Prof_school	1.6970
education_Doctorate	1.5228
demoweight	1.3891
workclass_Self_emp_not_inc	0.6910
relationship_Wife	0.6110
education_HS_grad	0.3625
occupation_Other_service	0.2528
native_country_India	0.1785
workclass_Local_gov	0.1461
native_country_United_States	0.1335
sex_Male	0.0990
occupation_Transport_moving	0.0884
Total number of features	96
Features with Importance $> 1\%$	11
Cumulative importance of top 10 features	95.23%

Table 3: Feature Importance Rankings