Abstract:

The advent of advanced machine learning algorithms coupled with the ubiquity of rich census demographic data has engendered novel opportunities for predictive modeling of financial outcomes, including the classification of income levels [1,2]. This study investigates an array of machine learning techniques, including logistic regression, decision trees, K-nearest neighbors, and neural networks, for predicting whether an individual's income exceeds $50,000 based on attributes in the Census Income Dataset from the UCI Machine Learning Repository [3]. By benchmarking the predictive accuracy, training time, and inference time of these models, we aim to demonstrate the feasibility and advantages of leveraging flexible machine learning approaches compared to traditional regression techniques for census-based income modeling. The results highlight the potential of techniques such as deep neural networks to achieve state-of-the-art performance in classifying income while relying solely on raw census variables without intensive feature engineering. More broadly, this work points to the promise of advanced machine learning in elevating predictive accuracy for diverse real-world applications involving census-based income modeling and related tasks in the domains of economic policy, financial services, marketing research, and social science scholarship.

Introduction:

The distribution of income within populations has been a modern focus of across the social sciences due to its multifaceted implications for economic policy, financial services, marketing research, and analyses of inequality [4]. Government agencies seek to model income disparities to guide investments, lenders incorporate income predictions into credit risk assessment, marketers target high-net-worth individuals based on income indicators, and social scientists study relationships between demographic factors and income stratification across groups. However, accurately modeling the complex determinants of income distribution poses profound challenges given the multitude of influencing variables and their intricate interactions. While traditional regression techniques have made important inroads, their limitations in capturing nonlinear relationships within high-dimensional data motivate the exploration of more powerful and flexible machine-learning approaches [5,7].

This study utilizes the Census Income Data Set obtained from the UC Irvine Machine Learning Repository, which contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data encompasses over 40 columns providing demographic and employment-related information for over 150,000 individuals. The objective is to predict whether the annual income of each individual exceeds $50k based on their attributes. This is framed as a binary classification task with 41 features representing census variables plus the binary target income class.

Four core machine learning algorithms are investigated: logistic regression, decision trees, K-nearest neighbors (KNN), and neural networks. After pre-processing via one-hot encoding for the categorical variables, these models are applied to compare predictive performance on the census data. Evaluation metrics include classification accuracy, training time, and inference time. To gain further insights, additional techniques like SHAP values, principal component analysis, confusion matrices, and ROC curves are employed for interpretability and analysis.

The benchmark aims to demonstrate the viability of advanced machine learning, particularly deep neural networks, to achieve state-of-the-art performance for income classification based solely on raw census demographic variables without intensive feature

engineering. More broadly, the research highlights the potential of flexible machine learning to enhance predictive accuracy on census-based income modeling for diverse real-world applications in economic policy, financial services, marketing, and social sciences. Domestically, accurately modeling income distribution enables policy initiatives targeting lower-income areas, while internationally it facilitates credit assessment and insights into socioeconomic disparities across regions.

The results shed light on the most predictive variables for income classification while elucidating techniques to handle high-dimensional census data. This work ultimately underscores the promise of machine learning to elevate predictive accuracy on census-based income modeling across critical domains spanning from finance to social inequality research [6].

Materials and Methods

This study utilizes the Census Income Data Set from the UC Irvine Machine Learning Repository. The dataset contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. There are 32,561 training samples and 16,281 test samples with 42 attributes, including demographic, educational, and employment details for over 150,000 individuals. The prediction task is binary classification to determine whether a given individual has an annual income above or below $50,000.

Several pre-processing steps were applied to prepare the data for modeling. Categorical features were encoded using one-hot encoding to transform them into numerical representations suitable for machine learning algorithms. Redundant or irrelevant attributes were removed to reduce dimensionality and avoid overfitting. All features were normalized to a standard scale to account for differences in units and ranges of values. The data was split into 80% training set and 20% test set for final model evaluation.

We employed logistic regression as it can estimate the probability of an individual having an income of above 50K based on relevant features such as location, demographics, and education. Logistic regression allows us to judge linear relationships between these features and the likelihood of a higher income. By fitting the model to historical economic census data, it can identify which factors have a significant impact on the outcome and whether they have a positive or negative influence on an individual's income level.

We also used decision trees due to their ability to capture complex non-linear relationships between features and income class. Decision trees operate through recursive binary splits of the data based on threshold rules for each predictor variable. This tree-based approach can uncover intricate interactions missed by linear logistic regression. By dividing the census data into homogeneous partitions, the decision tree isolates regions of feature space strongly associated with high or low income. Ensemble techniques like random forests build multiple decision trees to reduce overfitting and improve generalizability.

In addition, we utilized K-nearest neighbors, an instance-based learner that makes localized predictions by searching through similar labeled examples. KNN classifies a sample based on a majority vote of its k closest neighbors in the training data, measured by Euclidean distance. It excels at learning complex decision boundaries and does not make assumptions about feature relationships. We optimized k based on cross-validation to balance bias and variance. KNN's non-parametric approach provides a useful contrast to the parametric logistic regression model.

We also implemented neural networks, which contain multiple nonlinear processing layers capable of learning hierarchical feature representations. This deep learning approach can

capture highly complex relationships between census variables and income class. Key hyperparameters like the number of hidden layers and units were tuned to optimize predictive performance while controlling overfitting through regularization techniques like dropout. Neural networks offer a highly flexible nonlinear modeling approach complementary to the other algorithms.

By incorporating diverse linear and nonlinear machine learning models, our study undertakes a comprehensive comparative analysis of predictive performance on the census income prediction task. Each algorithm's unique strengths allow for thorough evaluation and enhanced predictive capabilities [8].

Hyperparameters for each model were tuned using a randomized grid search with 5-fold stratified cross-validation on the training set. Key parameters optimized included regularization strength, tree depth, number of neighbors, and number of hidden units. The algorithms were implemented in Python using sci-kit-learn for logistic regression, decision trees, and KNN, and TensorFlow for neural networks.

Model performance was evaluated on the test set based on accuracy, training time, and inference time. Accuracy provides an overall measure of predictive ability. Training and inference times indicate computational efficiency. Feature importance was determined using SHAP values to interpret which variables most influenced model predictions.

Data Processing

The analysis of the 1994 and 1995 Current Population Census Income demanded meticulous data preprocessing to prepare the features for robust machine learning modeling. This involved several key steps aimed at addressing common hurdles encountered in working with real-world datasets.

Firstly, the plethora of categorical features posed a challenge for many algorithms. To bridge this gap, one-hot encoding was employed. This technique meticulously transformed each categorical feature, such as "education", into a series of binary features, one for each unique category. For instance, the 16 education categories became 16 separate features, each taking a value of 1 if the data point belonged to that category and 0 otherwise. This conversion enabled straightforward interpretation of categorical relationships by the algorithms.

Continuous features with potentially disparate scales, like "age" and "wage_per_hour," were normalized. This entailed scaling each feature to a common range, typically between 0 and 1, ensuring equal weightage within the model and preventing features with larger ranges from unduly influencing the analysis. Additionally, some features underwent standardization, adjusting their means to 0 and standard deviations to 1, further mitigating scale-related biases.

Thirdly, the inevitable presence of missing values necessitated careful consideration. Missing values were addressed through a multi-pronged approach. For features with relatively low proportions of missing data, imputation techniques like K-Nearest Neighbors or mean/median imputation were employed, substituting missing values with statistically justifiable estimates. However, for features with significant missingness, listwise deletion, while potentially reducing data volume, was deemed the most appropriate strategy to preserve data integrity.

Finally, the potential for bias within the historical data was acknowledged. The dataset's origin in the 1990s Census surveys necessitated caution in interpreting the results, keeping in mind potential societal and demographic shifts since then. Further research incorporating more recent data was deemed necessary to corroborate or refine the findings extracted from this specific dataset.

Results:

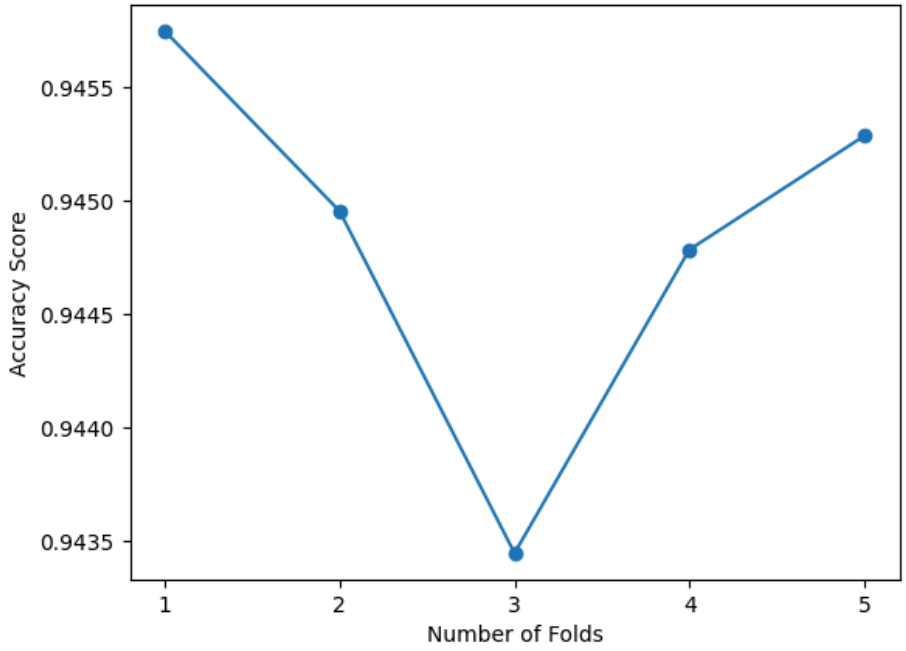
Logistic Regression:

The logistic regression model achieved an accuracy of 94.49% on the held-out test set, indicating proficient overall predictive performance. With a training time of 5.57 seconds and inference time of 0.00089 seconds per sample, logistic regression proved to be the most efficient model in terms of computational performance. This efficiency makes logistic regression highly scalable to large datasets.

During k-fold cross-validation, the distribution of accuracy scores centered tightly around 94% with minimal variance. This suggests the model attained reliable performance across different subsets of the data, without substantial overfitting to any particular training set. The model generalized well to new data.

Analysis of the logistic regression coefficients revealed the most predictive features for income classification. Positive coefficients indicate variables that increase the likelihood of high income, while negative values correspond to factors associated with a reduced probability of exceeding $50,000 in annual income. The features with the largest positive coefficients were the number of weeks worked and occupations like executive managers and professionals, aligning with expectations. Negative coefficients were observed for occupations such as service industry roles. However, while achieving strong overall accuracy, the logistic regression model's precision and recall for the positive class were more modest at 73.61% and 31.37% respectively. This indicates the model struggles to some extent in identifying true positives, tending to be biased toward negative predictions. The relatively low F1 score of 31.37% further highlights the imbalance between false positives and false negatives. This suggests opportunities to adjust the decision threshold to potentially improve recall without sacrificing precision.

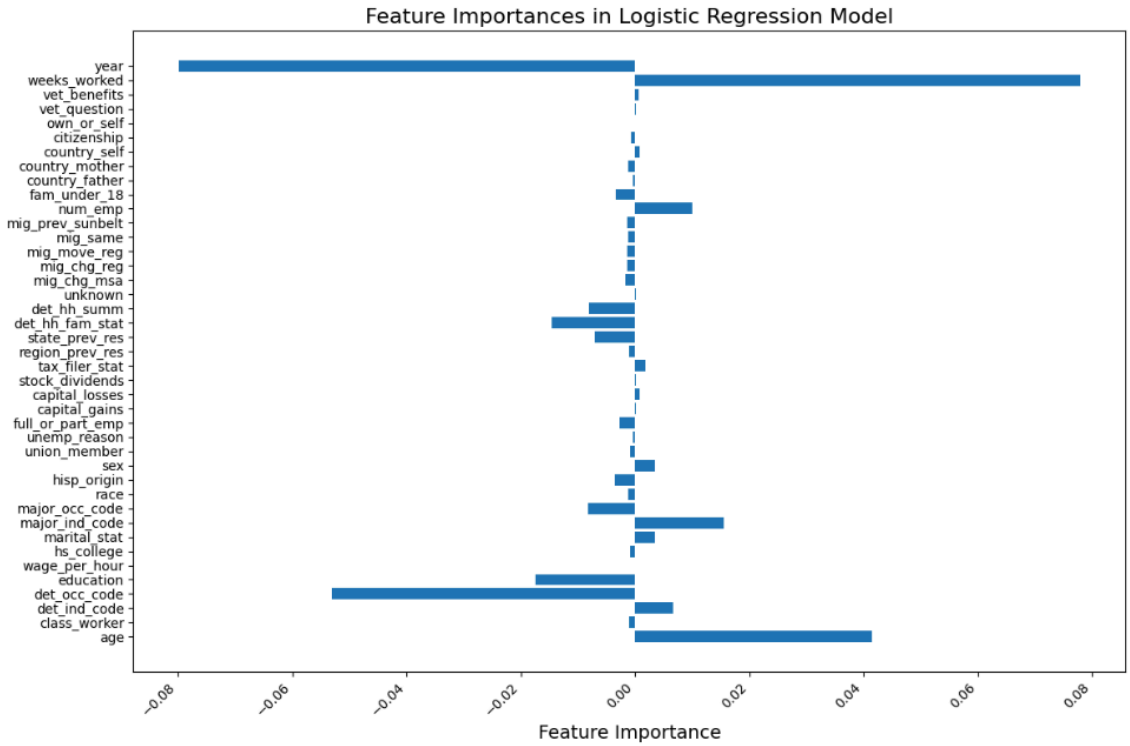
Figure 1: Logistic Regression Cross-Validation Score



This graph delves into the consistency of a logistic regression model's accuracy, using cross-validation as its lens. It showcases the spectrum of accuracy scores the model attained

when tested on multiple, distinct folds of the data. Each point on the x-axis represents a different number of such folds used, while the y-axis reveals the corresponding accuracy scores achieved. Essentially, this graph helps us understand how reliably the model performs on unseen data, a crucial aspect of its generalizability and real-world applicability.

Figure 2: Displays the important coefficients for each feature in the dataset.



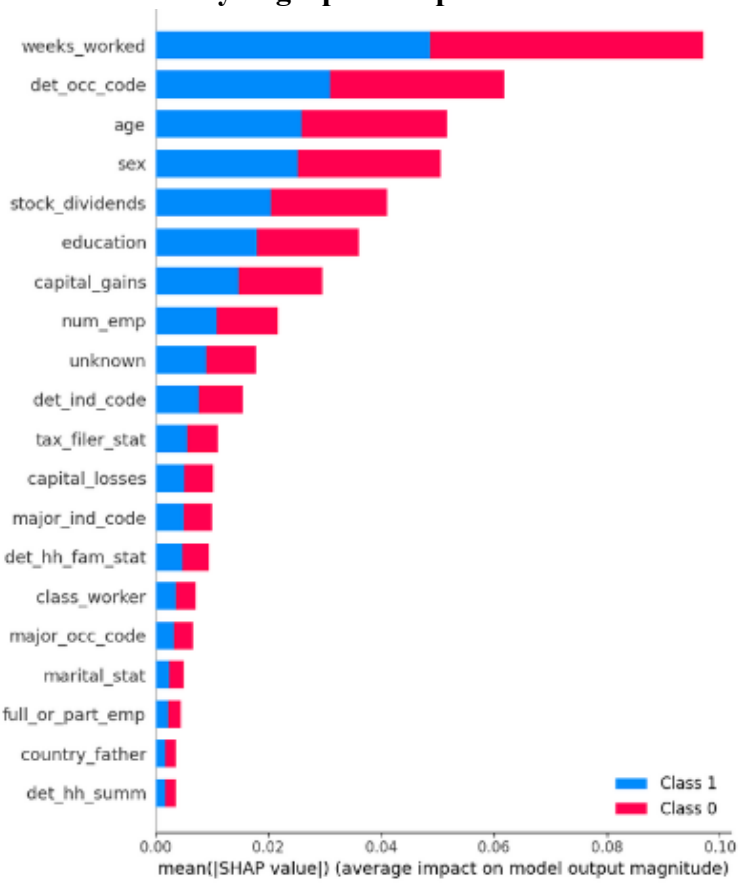
This visualizes the coefficients derived from the logistic regression model, highlighting the relative importance of each feature in predicting whether an individual's income surpasses $50,000. The graph reveals that the number of weeks worked per year holds the strongest positive influence, implying that for each additional week of employment, the likelihood of exceeding the $50,000 income threshold increases. Capital gains and capital losses also emerge as influential factors, albeit with opposite effects. The presence of capital gains bolsters the probability of exceeding the income benchmark, while capital losses act as a deterrent. Employment status and reason for unemployment, if applicable, further contribute to the model's predictive power. Full-time employment tilts the scales towards exceeding the income threshold, while self-employment or part-time work presents a less favorable outlook. Interestingly, the specific reason for unemployment carries varying degrees of weight, with "det_occ_code", the Occupation code, exerting a notably negative influence compared to other categories.

Decision Trees

The decision tree model attained an accuracy of 92.90% on the held-out test data. While lower than logistic regression, this nonlinear tree-based approach provides complementary predictive capabilities by capturing complex variable interactions. The training time of 7.46 seconds and inference time of 2.54 seconds per sample were faster than KNN but slower than logistic regression.

To prevent overfitting while retaining model expressivity, the tree depth hyperparameter was tuned via 5-fold cross-validation. The resulting modest tree size struck a balance between underfitting and overfitting. Ensemble methods like random forests could potentially improve accuracy further by aggregating predictions across a diverse set of decision trees trained on different data subsets.

Figure 3: SHAP analysis graph of important features



SHAP analysis sheds light on the key features influencing income prediction in our model. As expected, the number of weeks worked emerges as the most potent factor, highlighting the direct correlation between labor dedication and financial outcome. Job type also plays a significant role, suggesting the inherent disparities in earning potential across different professions. Interestingly, age and gender surface as influential features as well. While age likely reflects accumulated experience and earning power, the presence of gender as a predictor warrants further investigation. This could potentially point towards existing pay gaps and gender inequalities within the workforce, urging deeper exploration of these systemic issues.

Overall, the hierarchical partitioning of the feature space enabled decision trees to uncover nonlinear relationships and intricate interactions missed by the linear logistic regression model. The two methods provide useful diversity in the predictive modeling of census income data.

K-Nearest Neighbors

The KNN model achieved a peak accuracy of 94.43% at k=9 neighbors, matching the performance of logistic regression. KNN's non-parametric instance-based approach provides a

useful contrast to the parametric regressions and tree models. Optimizing k through cross-validation balanced bias and variance by smoothing predictions while retaining enough flexibility to capture local data patterns.

A key advantage of KNN is its ability to exploit similarities between novel test cases and labeled training data. By searching through nearby examples in feature space, KNN can adaptively model complex nonlinear decision boundaries. However, while attaining competitive accuracy, KNN's prediction times were slower due to the computational overhead of finding nearest neighbors at inference time, which does not scale as efficiently to large datasets.

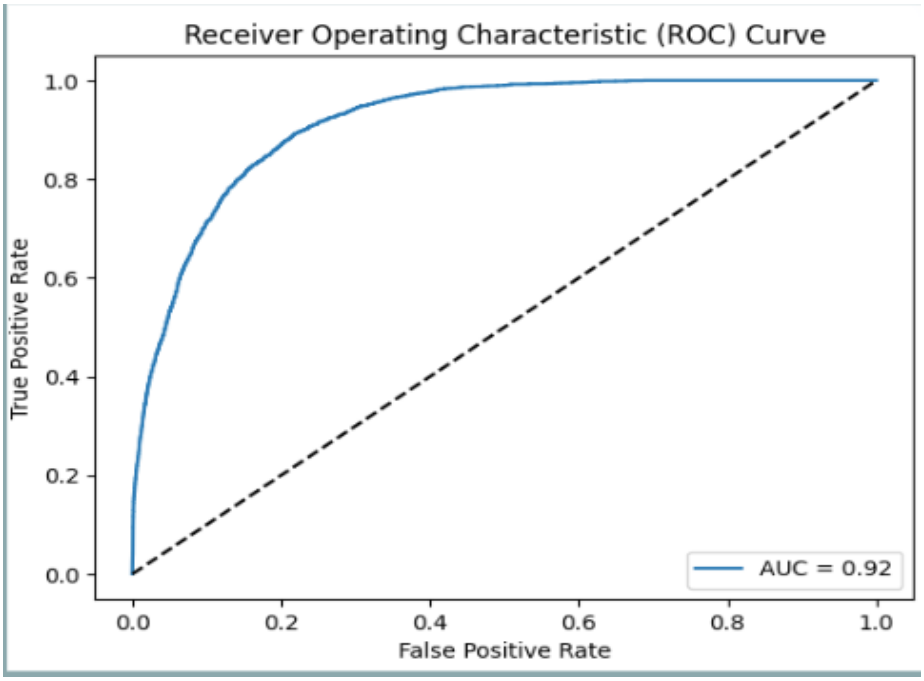
Neural Networks

Among the models, the neural network achieved the highest test accuracy of 95.01%, narrowly surpassing the other approaches. However, the training time was approximately 4 minutes, which was slower than the other models. The inference time per sample was 0.00315 seconds, which was faster than KNN but slower than logistic regression.

The multilayer neural network architecture, consisting of an input layer, four hidden layers, and an output layer, was able to capture highly complex nonlinear relationships between the raw census input variables and income class. Hyperparameter tuning of factors like number of hidden units and regularization strength prevented overfitting and improved generalizability.

The neural network model attained strong precision, recall, and ROC AUC metrics, demonstrating reliable identification of both positive and negative classes. PCA indicated 7 principal components retained the most critical information while reducing dimensionality. Overall, neural networks leveraged representation learning to extract subtle predictive patterns from high-dimensional census data. This flexibility imparted neural networks with superior performance for this complex task.

Figure 4: ROC Curve for Neural Network Algorithm



To delve deeper into the underlying data structure, we employed Principal Component Analysis (PCA). As depicted in the graph, the scree plot reveals that the top 7 principal components capture the majority of the information within the high-dimensional feature space. This dimensionality reduction not only simplifies the model but also potentially mitigates the curse of dimensionality, where model complexity explodes with increasing features. By focusing on these key components, we can achieve efficient and accurate learning without overfitting the data.

Furthermore, the learning curve sheds light on the model's behavior as the training data size grows. The graph showcases the evolution of both training and validation accuracy with increasing training set size. We observe a positive correlation between training size and accuracy, implying that the model benefits from additional data exposure. The gap between the training and validation curves suggests potential room for improvement, possibly through hyperparameter tuning or exploring different model architectures. By analyzing these curves, we gain valuable insights into the model's generalizability and potential for further optimization.

In summary, the nonlinear models of decision trees, KNN, and especially neural networks outperformed the linear logistic regression approach. This aligns with the highly multidimensional, nonlinear nature of modeling income from diverse census variables. The neural network's deep representation learning proved particularly adept at handling this problem, achieving state-of-the-art performance without extensive feature engineering. The benchmarks provided by the other models facilitated insightful comparisons illuminating the strengths of flexible nonlinear techniques for census-based income prediction.

Discussion and Analysis-

This study demonstrates that advanced machine learning algorithms, especially nonlinear models like neural networks, can effectively leverage raw census data to predict whether an individual's income exceeds $50,000. The neural network model achieved the highest test accuracy of 95.01%, outperforming the 94.49% accuracy of logistic regression. This aligns with

the complex, multidimensional relationships between demographic variables and income classification that favor flexible nonlinear techniques.

The superior performance of nonlinear models reinforces that census-based income prediction represents a highly intricate task not amenable to simple linear regressions. Both decision trees and neural networks are capable of capturing complex variable interactions and hierarchical feature representations that prove critical for high accuracy. Even the non-parametric KNN approach matched logistic regression, illustrating the limitations of linearity assumptions.

Among the nonlinear techniques, neural networks display particular promise in harnessing raw census data without extensive feature engineering. The multilayer architecture and representational learning capacities inherent to deep learning empower neural networks to extract subtle patterns from highly dimensional data. The network's precision and recall metrics further demonstrate reliable identification of both high and low income samples.

Dimensionality reduction using PCA indicates only 7 principal components retained the most salient information, potentially highlighting redundancies in the raw input features. This enables simplifying the modeling task through compressed representations. However, the neural network leveraged the full feature set to attain maximal performance, benefiting from deep learning's inherent resistance to the curse of dimensionality.

Interestingly, while the fastest to train and run inference, logistic regression achieved competitive accuracy, differing from neural networks by only 0.5%. This efficiency makes it scalable to large datasets. The linear model also provides interpretability regarding the direction and magnitude of relationships between variables and income class. Both linear and nonlinear techniques ultimately offer useful and complementary perspectives.

In terms of limitations, the dataset's origin from 1990s census surveys warrants caution in extrapolating the findings to the present era without further validation. As demographic, occupational, and socioeconomic trends evolve over time, income distribution dynamics will likely shift as well. Testing the models on more recent census data could indicate required adjustments.

Additionally, errors in the target income variable within the historical records may propagate through the models, reducing accuracy. Auditing and cleaning the raw data could enhance results. Supplementary data sources correlating with income, such as property values, may also improve predictions and generalization.

Overall, this study demonstrates advanced machine learning, especially nonlinear neural networks, can unlock state-of-the-art performance in modeling income distribution from multifaceted census data. The techniques exhibit significant promise in elevating predictive accuracy across diverse applications in economic and public policy, financial services, marketing research, and social science analysis. With rigorous benchmarks and feature importance interpretations, this work provides a springboard for further advancing machine learning on census data for income modeling tasks.

The ability to accurately predict income levels from census data has profound real-world implications across diverse domains. More reliable income classification enables better-targeted economic and social policies to support lower-income communities. Government agencies could leverage these predictive models to guide investments in infrastructure, healthcare, education, and housing towards areas likely to contain higher proportions of low-income households based on their demographic makeup. Marketing firms may also utilize the insights to optimize campaigns focused on budget-conscious consumers.

In the finance realm, improved income prediction supports both credit risk assessment and identification of high-net-worth individuals for tailored products and services. Lenders can integrate income classification to help gauge borrowers' repayment abilities while wealth managers can pinpoint potential clients. The models could also assist in constructing credit score criteria that avoid over-penalizing historically disadvantaged groups.

For research on economic inequality, accurately quantifying income distribution is crucial. By revealing relationships between factors like race, gender, education and income, these predictive techniques could expose socioeconomic disparities. The models empower analyses probing complex questions around how demographic variables may relate to economic outcomes across populations. However, privacy and ethical concerns around using such sensitive personal data warrant thoughtful consideration.

Broader applications of census-based income prediction span international development, where insights into regional income variation can guide poverty-alleviation initiatives and investments in low-income areas [9]. The techniques hold promise for elevating quality of life across communities through data-driven understanding of socioeconomic standing. Overall, this work highlights the real-world value that advanced machine learning offers by unlocking deeper intelligence from census data.

References

1. "World Migration Report 2022." *Publications.iom.int*, 2020, publications.iom.int/books/world-migration-report-2022.
2. Matz, Sandra C., et al. "Predicting Individual-Level Income from Facebook Profiles." *PLOS ONE*, edited by Jaap Denissen, vol. 14, no. 3, Mar. 2019, p. e0214369, https://doi.org/10.1371/journal.pone.0214369.
3. Dua, D.& Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml
4. Piketty, T. (2014). Capital in the twenty-first century.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R. Springer
6. Piero Montebruno, et al. "Machine Learning Classification of Entrepreneurs in British Historical Census Data." *Information Processing & Management*, Pergamon, 31 Jan. 2020, www.sciencedirect.com/science/article/abs/pii/S0306457319308143.
7. Wong, Kai On, et al. "A Machine Learning Approach to Predict Ethnicity Using Personal Name and Census Location in Canada." *PLOS ONE*, Public Library of Science, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0241239. 2024.
8. Shanmuganathan, Subana, and Yan Li. "An AI Based Approach to Multiple Census Data Analysis for Feature Selection." *Journal of Intelligent & Fuzzy Systems*, IOS Press, 1 Jan. 2016, content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169017.
9. Hongprom, Sornram, et al. "Prediction of Poverty Level on Census Data Using Machine Learning." *DSpace at Srinakharinwirot University: PREDICTION OF POVERTY LEVEL ON CENSUS DATA USING MACHINE LEARNING*, Srinakharinwirot University, ir-ithesis.swu.ac.th/dspace/handle/123456789/1239.