Will You Make More Than 50k?

Filip Jevic
Computational Data Science
College of Engineering
jevticfi@msu.edu

Johnny Danstrom
Data Science
College of Natural Science
danstro2@msu.edu

Brady Berg
Computational Data Science
College of Engineering
bergbrae@msu.edu

Abstract

This paper describes a project to predict the salary range of an individual using existing demographic data and two different machine learning methods. The datasets we use for this project are the Adult dataset (adu, 1996) and ACSIncome dataset (Ding et al., 2021). These datasets are collections of demographic information on over 48,000 and 1.6 million individuals, respectively. They include data such as age, education level, work class, martial status, race and sex. The goal of this project is to use this data to predict whether or not an individual makes over $50k per year using logistic regression and a multi-layer perceptron. We evaluate our models by calculating accuracy, precision, recall, F1 and AUC, among other metrics. Parameter searches are run over datasets of varying numbers of features.

1 Introduction

Demographic data consists of an array of socioeconomic information to help describe a population. This information contains a wide range of factors, including gender, age, marital status, race, and income. Demonstrated by previous work, these factors can be used to create predictive machine learning models that aid in identifying patterns and new information from this data. Using these patterns, we predict the salary range of an individual using existing demographic data and two different machine learning methods.

It is important for college students to know about the industry that follows their education process. The ACSIncome (Ding et al., 2021) dataset is appealing for many reasons and to all different types of people. Important knowledge about the job market is important for certain people like students, the unemployed, or employers and companies themselves. What does a "typical" worker from a certain occupation look like? Which types of jobs tend to earn the highest salary? What level of education is typically required for a certain job? These are some of the many questions that are associated with finding a job. We hope to use this dataset to help answer these questions.

Previously, academic papers focusing on algorithmic fairness tended to involve the UCI Adult Dataset (adu, 1996). Created from 1994 census data, this dataset has been used in three hundred research papers related to fairness interventions (Ding et al., 2021). There are several issues with the Adult dataset which should prevent its further use in research. First, the dataset is now 28 years old. As such, significant data drift can be expected. The average salary for a woman in 1994 was 72% that of the average male. In 2019, the gap was recorded at 82% (on Pay Equality,). Today, the median income is $70,784 whereas it was $35,492 in 1994 after adjusting for inflation (Statista, 2022). Just considering inflation, the value of a USD has decreased by 50% in that time (inf, 2022). Evidently, there are compelling reasons to source a more recent dataset.

In 2018, (Ding et al., 2021) derived new datasets from the US Census data. These

datasets were directly made to replace the use cases of the Adult dataset, and provided a python package to enable easier querying of Census data sources. There were five datasets created, but here we only use ACSIncome. Their motivation for creating a new dataset like the Adult dataset in founded in beliefs that disparities in error rates between demographics can be remedied by collecting larger datasets and more data reflecting social progress. The Adult dataset and it's predecessor were created mainly to benchmark bias mitigation methods. In these methods, it is customary to analyze models fit to this data to understand the underlying bias. In support of these efforts, we provide two models which could be used to understand the nature of this socioeconomic disparity.

2 Literature Review

There are two main areas of similarity between published literature and this project. The first selection of papers we select to review utilize the ACIncome dataset in the course of their study. The second group considers the problem of training a binary predictor on dataset of socioeconomic data.

(Le Quy et al., 2022) provide a survey of datasets used in fairness-aware machine learning. As one of these datasets, ACSIncome is used to benchmark fairness-aware approaches. To evaluate bias in the data, relationships are identified between different attributes using a Bayesian network. (Gultchin et al., 2022) cite the ACSIncome article only to emphasize that the Adult dataset should not be used anymore. (Hort et al., 2022) provide another survey of bias mitigation in machine learning. They state that 77% of reviewed publications used the Adult dataset, whereas only 1% used the ACSIncome dataset.

(Yamnampet, 2022) uses a neural network, support vector machine, ordinal regression, Poisson regression, linear regression, Bayesian linear regression, decision forest regression, and boosted decision tree regression to predict either a binary target or the true income of a person, depending on the model used. They found that a neural network was the best at predicting income in the regression case. (Chakrabarty and Biswas, 2018) train a classifier to predict income greater than or less than $50,000 using the Adult dataset. They achieve 88.16% accuracy using a gradient boosting classifier. (Tare et al., 2019) use the Adult dataset to predict if an individual earns more than $50,00 using a decision tree and random forest. (Chen, 2021) also uses the Adult dataset to predict across a $50,000 threshold using quadratic discriminant analysis, support vector machine, and a random forest.

As evident, the Adult dataset and its predecessor, the ACSIncome dataset, have been used extensively to benchmark for bias mitigation approaches. In each of these studies, bias is studied through a model which captures the bias patterns. As such, a trained classifier is valuable as a first step towards bias evaluation. This is the focus of the second group of papers. As the ACSIncome dataset is much newer than the Adult dataset, there are fewer publications which seek to model it's income patterns. Studies that sought to predict the income of the Adult dataset used various methods including random forests and neural networks.

3 Data

The preliminary dataset used for this project, the Adult dataset, contains census data from 1994-1995. The data consists of 48,842 instances of 14 attributes, which include demographic information such as age, education, occupation, race, gender, and marital status, as well as information about an individual's income. The income attribute is a binary attribute that indicates whether or not an individual makes over $50,000 a year. The Adult dataset was originally obtained from the UCI Machine Learning Repository. It was created by Ronny Kohavi and Barry Becker (adu, 1996) for a paper on scaling up machine learning (Kohavi and others, 1996). The numerical features in the dataset are age, sample weight, an ordinal value representing education, cap-

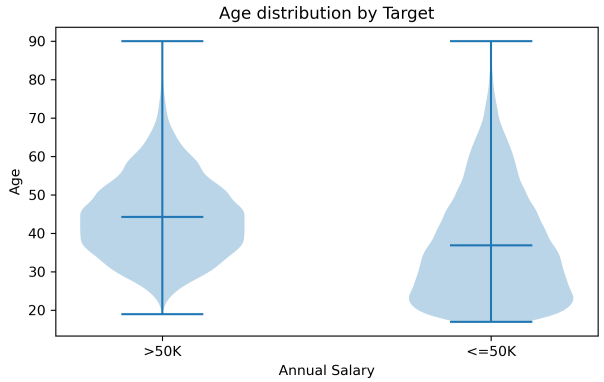

Figure 1: Distribution of ages by target value for the Adult dataset.

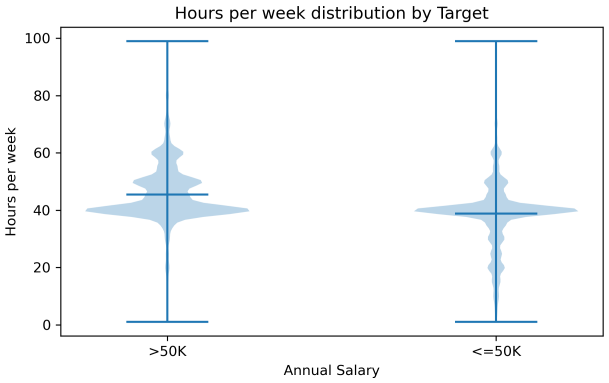

Figure 2: Distribution of number of hours worked per week for the Adult dataset.

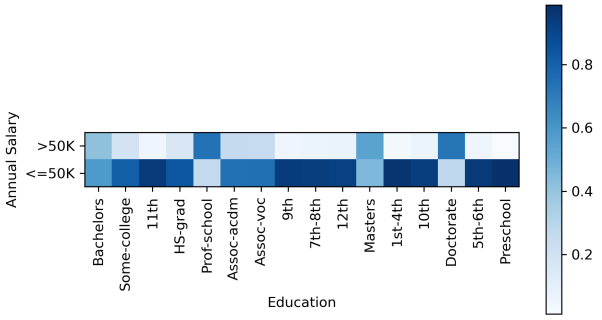

Figure 3: Distribution of number of hours worked per week for the Adult dataset.

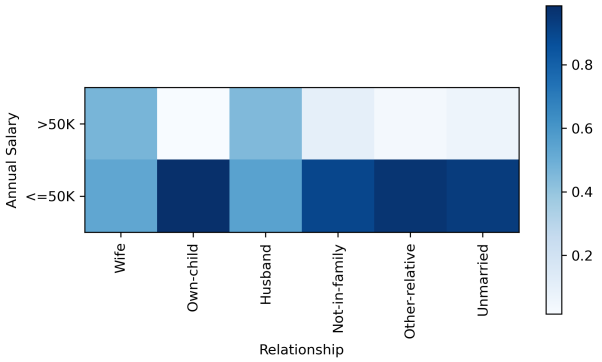

Figure 4: Distribution of observations by salary and relationship to head of household for the Adult dataset. Values are normalized along salary.

ital gain, capital loss, and hours worked per week. First, in figure 1, we notice that the age distribution for those that make less than $50,000 is lower. This may be explained by the fact that people that make less than $50,000 have less experience in the workforce and thus are younger. We also note that in the later years of individuals' lives, they tend to more commonly make less than $50k. An evaluation of hours worked per week in 2 shows in shows that while the majority of people work around 40 hours per week, working overtime is more common with those making more than the salary threshold. Of the categorical features in this dataset, we present the three most significant. From distributions of target values across education levels in figure 3, it appears that those who attended a professional school or with a Doctorate most frequently earned over $50k. Those with a master's or bachelor followed closely behind with a large proportion exceeding the threshold. Interestingly, those who attended some college but did not graduate earned more than $50k at about the same rate as those whose highest education is high school. Considering occupation, executive and managerial positions, professional roles such as a lawyer, doctor, or engineer, as well as tech support were most likely to earn over $50k. Private household service and other work such as craft repair and farming were among the lowest-earning occupations. Lastly, figure 4 shows salary by relationship to the head of household revealing that those who are married or in a relationship are likely to earn more than $50k than those who are not married.

Our second dataset is the ACSIncome dataset. While is was made to replace the Adult dataset, is contains much more data

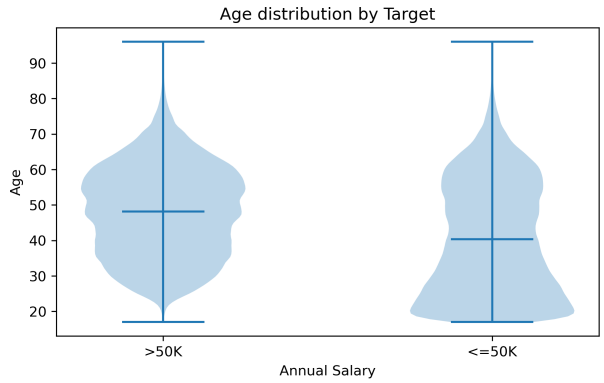

Figure 5: Distribution of ages by target within the ACSIncome dataset

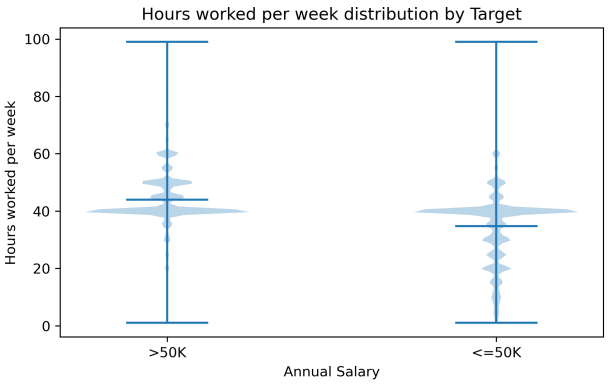

Figure 6: Distribution of hours worked per week by target within the ACSIncome dataset

with more granular categorical features. It contains two numerical features for age and the number of hours worked per week. The categorical columns are working class, education level, marital status, occupation, place of birth, relationship to head of household, sex, race, and state. The income feature was originally a continuous measure, but was converted to a binary indicator at $50,000 to remain consistent between datasets. Figure 5 shows the distribution of ages by target for this dataset. It can be seen that there is a similar pattern as the Adult dataset, we can also note that the general frequency of people aged 50+ is more common in the ACSIncome dataset. Shown in figure 6, the number of hours worked per week follows a similar pattern. One noticeable difference exists where those who earned less than 40 hours per week earned less than $50k. Compared

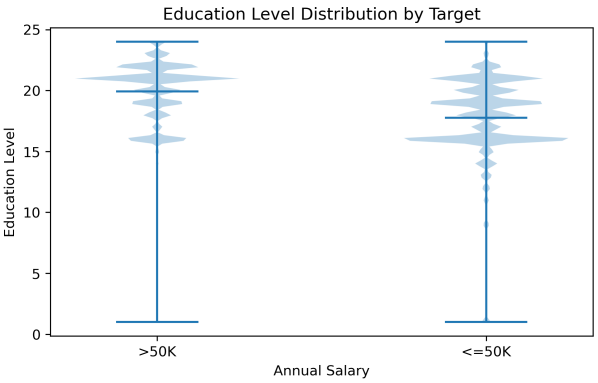

Figure 7: Distribution of education levels by target within the ACSIncome dataset

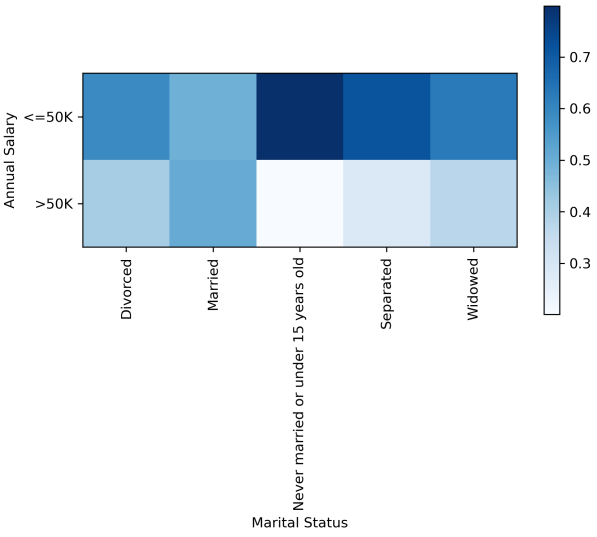

Figure 8: Distribution of marital status by target within the ACSIncome dataset. Values are normalized along the target

to the Adult dataset, more frequently people are working 20, 25, or 30 hours per week. An additional similarity shows between the income differences between levels of education. Lastly, a mentionable difference between the two dataset's marital status is observed in figure 8. In the Adult dataset, those who were divorced rarely made more than $50k. In the ACSIncome dataset it is more common to make more than $50k after being divorced. We argue that this is significant, even given the rise in inflation, due the other marital statuses maintaining the same general patterns between datasets.

4 Methods

The first, simpler, model we implemented was logistic regression. We designed a wide hyperparameter search to leverage the computationally inexpensive model, but we used a random search over the domain as it has shown to achieve results just as good as grid search, but in a fraction of the time. We also sought to reduce computational load because this search would be run five times, one for each l. For each l, the data was preprocessed to enforce the new limit, and the random search was run for 30 iterations. For each sample, 5 fold cross-validation was performed and evaluated on AUC. Each limit's best parameters were chosen as those which produced the highest mean AUC on the cross-validation hold-out sets. The following were the hyperparameter distributions from which the searches were sampled:

- C = 10e-4 - 10e4
- Penalty = elasticnet, l1, l2, none
- When the penalty was elasticnet, L1 Ratio was sampled from uniform 0-1

A model was trained for both the Adult and ACSIncome datasets. The Adult dataset trained in a reasonable amount of time, however the ACS Income dataset, due to the increased number of features and sheer number of samples, took over 10 hours to train and complete parameter search. Increasing the feature dimension size search posed massive runtime issues as well. The results for both datasets are shown here:

Adult:

- Accuracy: 80.5%
- F1_weighted: 67.1%
- Recall: 84.3%
- ROC_AUC Score: 90.4%

With parameters:

- Best Penalty: No Penalty

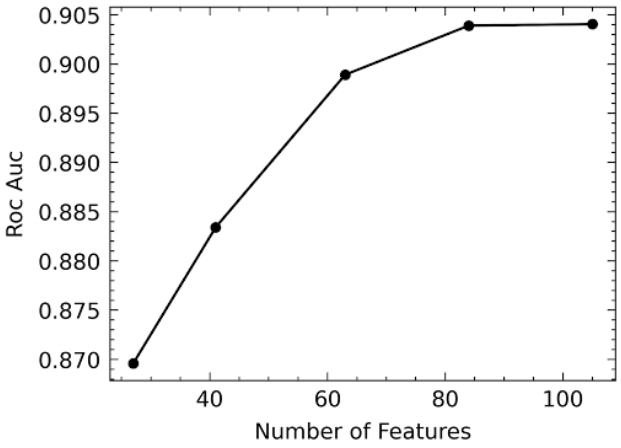

Figure 9: Roc Auc Score for Logistic Regression (Adult Dataset)

- Best C Value: 10
- Solver: 'saga'
- Max_iter: 300
- Best Dimensions: max

ACS Income:

- Accuracy: 79.7%
- F1_weighted: 71.3%
- Recall: 68.4%
- ROC_AUC Score: 87.3%

With parameters:

- Best Penalty: L2
- Best C Value: 0.1
- Solver: 'saga'
- Max_iter: 300
- Best Dimensions: max

A multilayer perceptron model (MLP) was one of the models chosen for classification. The MLPClassifier from Scikit-learn machine learning library was the base model used. For this model, we considered its computational expense and designed a narrower, dense

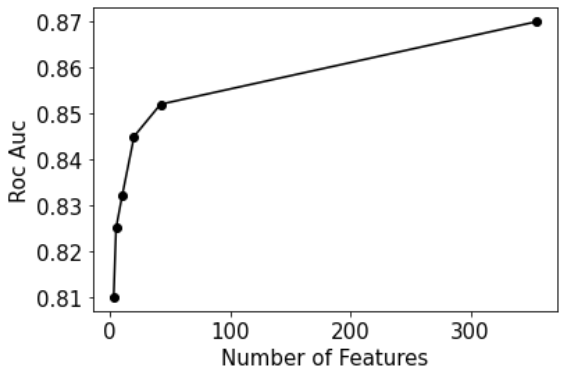

Figure 10: Roc Auc Score for Logistic Regression (ACS Income Dataset)

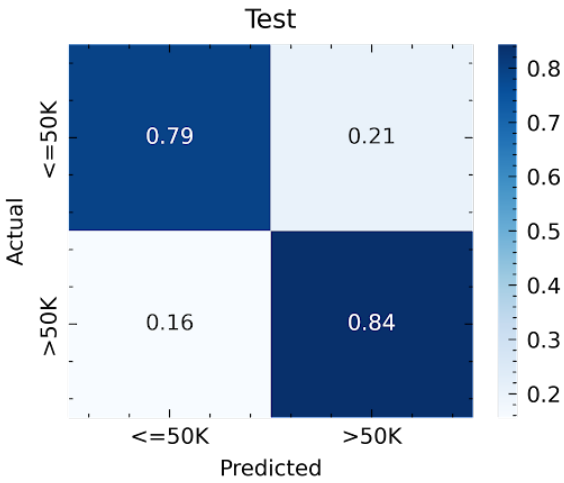

Figure 11: Confusion Matrix of the Test Split Predictions for Logistic Regression (Adult Dataset)

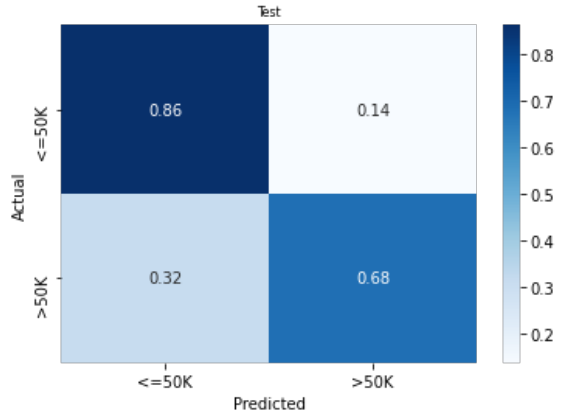

Figure 12: Confusion Matrix of the Test Split Predictions for Logistic Regression (ACS Income Dataset)

search. Throughout this small and exhaustive search, the models maintained relu activation functions and an adaptive learning rate, as well as an Adam solver. After the data was cleaned and mapped to the categorical features, the categorical features needed to be encoded. After the encoding, the data was split into a testing dataset and a training dataset. Parameter optimization was used to find the best set of parameters. Similar to the logistic regression model, this search defined the best estimator as the model with the highest mean AUC across a 5-fold cross-validation. The following hyperparameters were tested: hidden layer size and number of them, as well as the activation function. Using a list of different values/selections for each hyperparameter it matched every combination and also used cross validation with 5 folds. The Adult dataset trained relatively fast, with the best model generating a test accuracy of 80.6%, as well as yielding the following results:

- Accuracy: 80.6%
- F1_weighted: 67.7%
- Recall: 85.3%
- ROC_AUC Score: 90.4%

With parameters:

- Hidden_layer_sizes: (100,)
- Activation: 'relu'
- Solver: 'adam'
- Max_iter: 300

For the ACSIncome dataset, the first MLP classification model was the base model with standard inputs. It took a while to run due to the large dataset and output an accuracy of 79%, not too bad for a first run. After this, the activation function was changed and all tested with different layers of the network. Most of the random testing yielded an accuracy of 79%, but one model (named Model A) outputted 80.11% accuracy. This was the

highest accuracy model achieved. This model (Model B) ran overnight and output a score of 79%. This was unexpected because the features that made Model A were included in the parameters searched through in the optimization algorithm run. Looking deeper through different error tests the following error measurements were found.

Model A yielded the following:

- Accuracy: 80.14%
- F1_weighted: 80.03%
- Recall: 70.53%
- ROC_AUC Score: 78.14%

With parameters:

- Hidden_layer_sizes: (50,50,)
- Activation: 'relu'
- Solver: 'adam'
- Max_iter: 300

For the 'optimized' model B:

- Accuracy: 79.92%
- F1_weighted: 79.94%
- Recall: 73.34%
- ROC_AUC Score: 78.55%

With parameters:

- Hidden_layer_sizes: (100,50)
- Activation: 'logistic'
- Solver: 'adam'
- Max_iter: 300

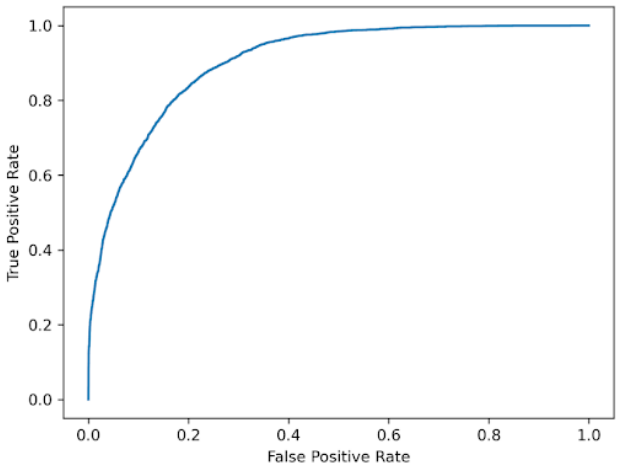

Figure 13: ROC AUC curve for Test Split of MLP Model (Adult Dataset))

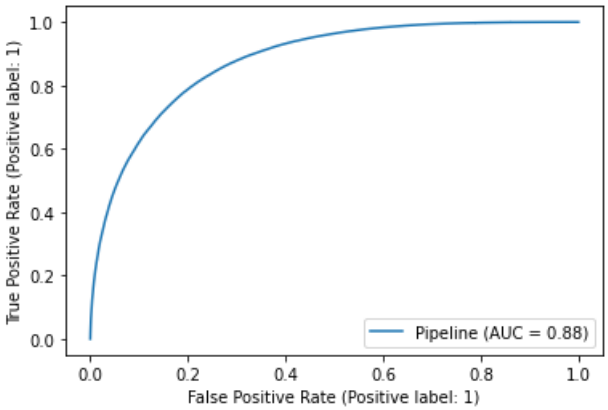

Figure 14: ROC AUC curve for Test Split of MLP Model A (ACS Income Dataset)

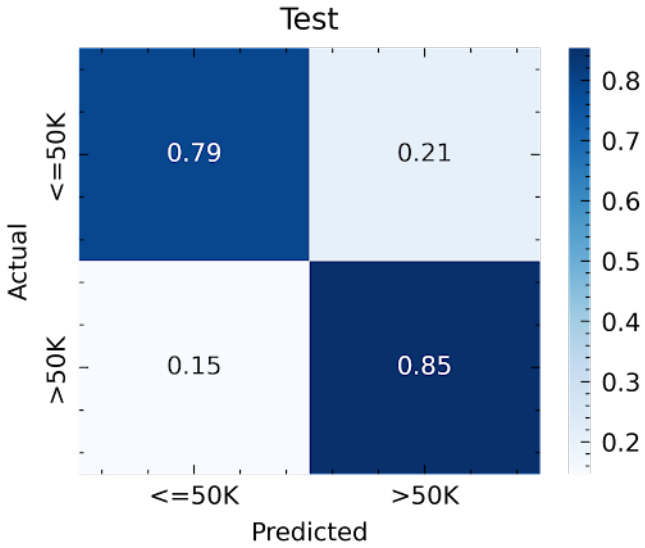

Figure 15: Confusion Matrix of the Test Split Predictions for the MLP Model (Adult Dataset)

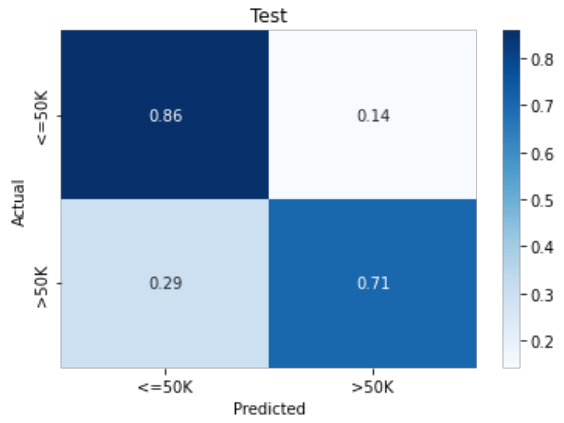

Figure 16: Confusion Matrix of the Test Split Predictions for MLP Model A (ACS Income Dataset

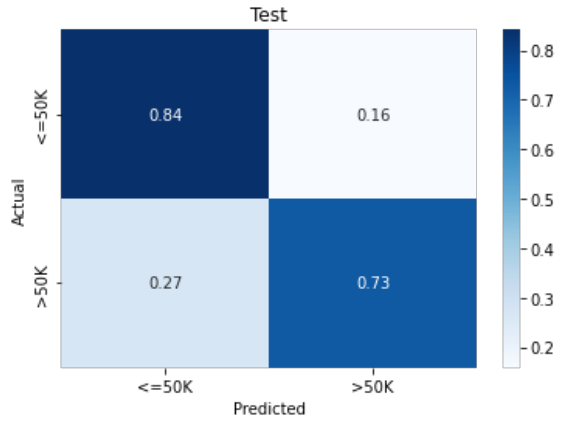

Figure 17: Confusion Matrix of the Test Split Predictions for MLP Model B (ACS Income Dataset)

5 Analysis

After collecting all of the information for the models, there was no model significantly more robust and better performing than the rest. However, each model has a uniquely good performance for different datasets/prediction tasks. The first aspect that was most noticeable is runtime, where MLP could train and optimize a model up to 10x faster. The runtime of the time to train each model was compared using the ACS Income dataset, which contained both a large number of samples and features. While Model A and B for the MLP method each took between 1 - 2 hours to train and optimize, the logistic regression model took 10 hours to train and optimize 1 model. The time complexity of training a logistic regression model is O(n*d), where n = training samples and d = feature dimensions. However, for MLP, the time complexity is entirely dependent on the number of layers in the model, which never exceeds 2. The hidden dimension sizes also never exceeded 100, which made the model much more efficient to train. While time complexity is an important factor to data scientists looking to train many models quickly for large datasets, it is not the only determining factor when choosing the best machine learning method to use for a prediction task. Test accuracy is a great metric to use when the classes have a similar number of samples, whereas Roc Auc is the ideal metric for datasets with a class imbalance. In the case of the ACS Income dataset, about 40% of both the training and testing samples are positive (>50K) while 60% of the training/testing samples are negative (<= 50K). Since the classes are not perfectly balanced through the dataset, it is justifiable to assume Roc Auc is a better measure of model performance than test accuracy.

With this in mind, the Roc Auc score for the logistic regression model is about 8.75% higher than the best MLP model. However, the MLP model has a slightly higher accuracy percentage than the logistic regression model (0.44% higher), implying that if a dataset had balanced classes then the MLP model could

possibly be the better model choice. The last aspect to consider when comparing the models is the confusion matrix for each model (Figures 11, 12, 15, 16, 17). The confusion matrix is another metric that is meant to handle class imbalance and can give a more detailed understanding of where the error in the model is coming from. In the case of this prediction task, the type 1 error, or the false positive error, represents samples labeled as '<=50K' that are predicted as '>50K'. Type 2 error, the false negative error, represents samples labeled as '>50K' but predicted as '<=50K'. The lowest type 1 error on the ACS Income dataset was the same for both MLP and logistic regression at 0.14, indicating that both models can very accurately classify negative samples. The lowest type 2 error for the logistic regression was 0.32 while Model B for the MLP model had an error of 0.27, indicating that the MLP model is better at classifying positive samples than logistic regression. This is very interesting because the Roc Auc Score for logistic regression was much higher than for the MLP model. Overall, both models showed very good performance in different ways.

When comparing this performance to that of the previous work, however, both models seem to perform worse than the original model when predicting on the Adult dataset. The test accuracies of both models were less than the 88% accuracy achieved by previous studies. One aspect that wasn't mentioned in the previous works was the Roc Auc Score, as well as the confusion matrix, so it can't be finitely determined that the methods in this project performed worse than previous ones. However, based on the information given, previous works' models seem to significantly outperform the ones proposed in this paper.

6 Conclusion

The goal of this project was to use demographic and employment data to predict whether one would make above or below $50k. The models used to test this out all had similar results. The time it took to train the models was long. This could be due to the fact that the dataset was extremely large. With an accuracy of 80%, model A was our best model. This wasn't as high as initially desired but, there are still takeaways. There are many entries into the data, this will require higher computing time and thus models that can be tested frequently can be better for fitting the model. Also, these models can further be used to evaluate bias. It was important to have many ways of evaluating the model. The Accuracy, F1 score, Recall, and Roc Auc score were important for weighing the strength and the weaknesses of the models. It was tough to define a "best model" as the models were similar. The goals were defined and Model A was selected by those constraints. Transitioning to future steps with the lower than desired accuracy in this project the next steps would look towards trying other models to evaluate this dataset.

The accuracy for Model A was the highest throughout the three models, it had a much shorter training time and was a simpler model. The comparison between MLP Model A and Model B waters down to how important each metric is. If the roc auc and recall were the desired metrics then model B would yield the best results but, at the cost of computing time. Due to these factors, Model A is the best model selected for this project for our prediction. The goal of predicting whether one will earn more than $50,000 is best reached by model A. Model A has the highest accuracy with a much smaller training time.

7 Future Work

While the results we have compiled from our current methods are substantial, there is still much that can be done to further improve this work. The first approach to possibly implement is an active learning approach. This process would involve iteratively adding labels and retraining a model. The labels to add are chosen as the most uncertain predictions out of the unlabeled data. It is possible that this approach will be able to generate similar performance to the current model while only us-

ing a small subset of the labels. This is applicable because other problems may not have as much data to work with. This would be relevant if the current dataset is replaced with a recently updated, but smaller dataset.

The second improvement approach involves reducing demographic identity bias within the dataset to ensure that this model cannot be used for unethical purposes. Since factors such as race, gender, and age can be used to discriminate against people of certain identities, including these features in a predictive model to determine salary could negatively mislead the correlations and predictions made. For example, if this prediction task was used by credit card or insurance companies, then individuals could be negatively impacted by the model due to their identity. If machine learning perpetuates human bias instead of limiting it, then it can only negatively impact peoples' lives. The goal of this prediction task is to inform college aged students of their possible salary based on their career and education, so removing discriminatory bias is absolutely necessary to make the model practical. If discriminatory bias were removed and massively reduced the performance of the model, then a complementary approach to improving the model would be to find data with non discriminatory featuresl. Overall, there are still many developments that can enhance the performance and relevance of this predictive task.

Authors' Biographies

Filip Jevtic Is a senior computational data science major in the College of Engineering. His roles in this project were to write the introduction and literature review for the first two reports, and creating models for the final report.

Johnny Danstrom Is a senior data science major in the College of Natural Science. His roles in this project were to write the introduction and literature review for the first two reports, and creating models for the final report.

Brady Berg Is a Senior computational data science major in the College of Engineering. His role in this project was to create the clean and prepare the data, create the models for the initial reports, and write the introduction and literature review for the final report.

References

1996. Adult. UCI Machine Learning Repository.

Navoneel Chakrabarty and Sanket Biswas. 2018. A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 207–212.

Li-Pang Chen. 2021. Supervised learning for binary classification on us adult income. *Journal of Modeling and Optimization*, 13(2):80–91.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *CoRR*, abs/2108.04884.

Limor Gultchin, Vincent Cohen-Addad, Sophie Giffard-Roisin, Varun Kanade, and Frederik Mallmann-Trenn. 2022. Beyond impossibility: Balancing sufficiency, separation and accuracy.

Max Hort, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2022. Bias mitigation for machine learning classifiers: A comprehensive survey.

2022. Inflation Calculator — Find US Dollar's Value from 1913-2022, 11.

Ron Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452.

National Committee on Pay Equality. The wage gap over time: In real dollars, women see a continuing gap.

Statista. 2022. U.S. median household income 1990-2021, 10.

Prachi Tare, Satyam Kumar Mishra, Mukul Lakhotia, and Kushagra Goyal. 2019. Bias variance tradeoff in classification algorithms on the census income dataset.

Ghatkesar Yamnampet. 2022. Comparative analysis of classification models on income prediction. *International Journal on Recent and Innovation Trends in Computing and Communication*, 5(4):451–455.