A Case Study on The Efficacy of Machine Learning Models in Predicting Income

Channing Schwaebe, Dhaval Sharma, Shruti Sarle, Dhruvil Shah

***Abstract*—This project evaluates the effectiveness of multiple machine learning models in predicting income based on data extracted from the 1994 US Census database by Ronny Kohavi and Barry Becker and hosted by UCI Machine Learning.[1] The models tested include Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forests, Naïve Bayes Classifiers, and Artificial Neural Networks. Through our analysis of these models on this problem instance, we found that a Random Forest generated the most accurate predictions. Additionally, the data was separated by category to determine the relative strength of each category in predicting income. In doing so, we found that investment performance alone (capital gains and losses) was the most accurate predictor of income.**

I. INTRODUCTION

THIS Machine Learning project uses data from the 1994 US Census to predict income based on a set of 14 parameters. In this dataset, income was binned into two classes, less than $50,000 and greater than or equal to $50,000. This binary classification problem was evaluated using six popular Machine Learning algorithms: Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forests, Naïve Bayes Classifiers, and Artificial Neural Networks. Of these models, a Random Forest was found to be the most accurate. This is largely due to its ability to accurately predict the minority class (income greater than or equal to $50,000).

II. TASK DESCRIPTION

Our task was to build and train several different models on the dataset. The models were then given a test set and tasked with predicting the income for individuals in the test set. The results were analyzed to determine the best model for this binary classification problem.

Once we were able to determine the most accurate model, we set out to determine the most accurate subset of data. To do this, we split the parameters into four bins and ran the model on each subset.

III. MAJOR CHALLENGES AND SOLUTIONS

Our first challenge was to remove incomplete entries from the dataset in order negate the effects of possible regularities in the incomplete entries. The solution to this was to simply remove all incomplete entries from the dataset, reducing the total number of entries from 48,842 to 45,222.

The 14 parameters in the dataset consisted of 8 categorical variables (workclass, education, marital status, occupation, relationship, race, sex and native country). In the preprocessing of the data, these parameters were expanded into multiple binary columns. For example, race contained five possible values (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other) so it was expanded into five columns, one for each categorical value. After this expansion, the dataset contained 104 parameters.

Unfortunately, the dataset is not balanced, with 75.22% of samples earning less than $50,000 and 24.78% of samples earning greater than or equal to $50,000. This invariably leads to the model being inadequately trained on the minority class, and accounts for the substantial differences in recall between the two classes. In order to fix this issue

more data would need to be gathered, which is beyond the scope of this project.

IV. Experiments

A. Dataset Description

In this dataset, income is defined as "income received on a regular basis (exclusive of certain money receipts such as capital gains) before payments for personal income taxes, social security, union dues, Medicare deductions, etc."[2] Income was separated into two classes:

1.) Less than $50,000
2.) Greater than or equal to $50,000

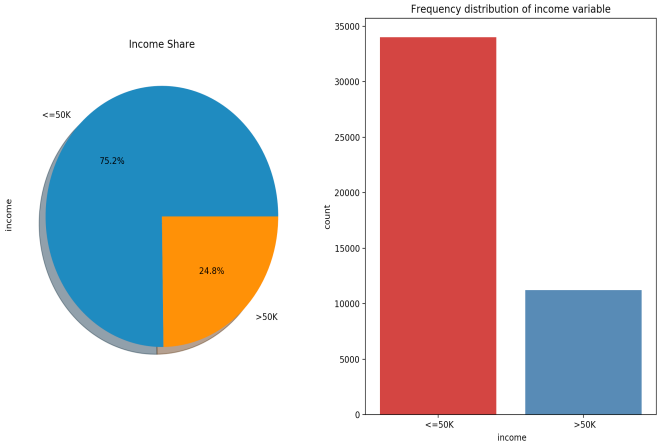
The original dataset contained 14 features, which were used to predict the aforementioned label, income. The features are as follows:

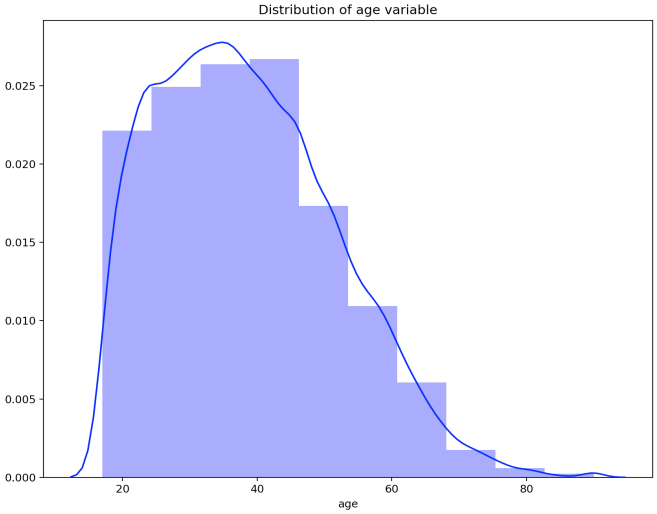
1.) Age - Continuous
2.) Workclass – Categorical
3.) Fnlwgt - Continuous
4.) Education - Categorical
5.) Education-num - Continuous
6.) Marital-Status - Categorical
7.) Occupation - Categorical
8.) Relationship - Categorical
9.) Race - Categorical
10.) Sex - Categorical
11.) Capital-Gain - Continuous
12.) Capital-Loss - Continuous
13.) Hours-per-Week - Continuous
14.) Native-Country – Categorical

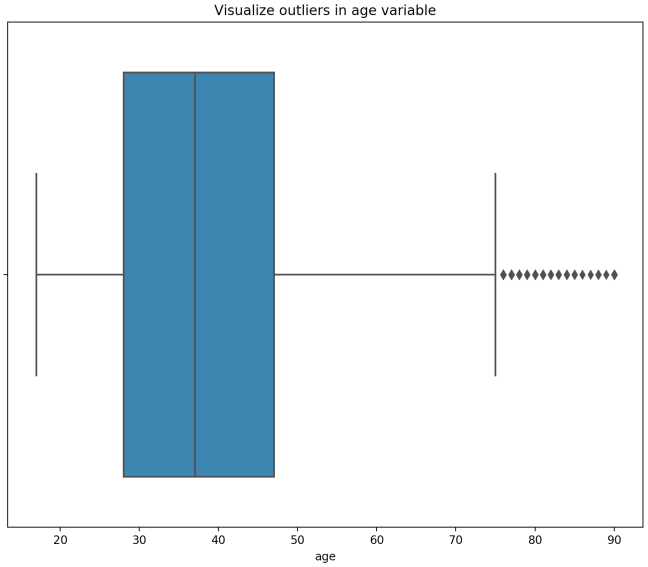
The secondary evaluations of subsets of the original data in predicting income consisted of four categories:

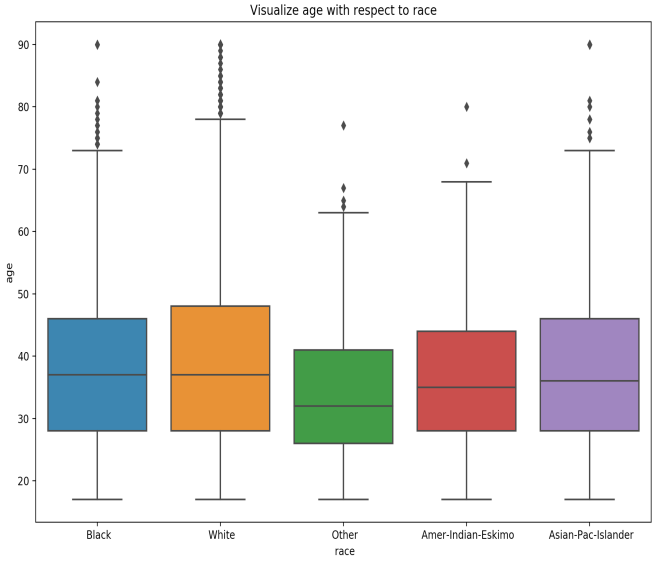
1.) Education – (Education-num)
2.) Occupation – (Workclass, Occupation, Hours-per-Week)
3.) Demographic – (Age, Marital-Status, Relationship, Race, Sex, Native-Country)
4.) Investments – (Capital-Gain, Capital-Loss)

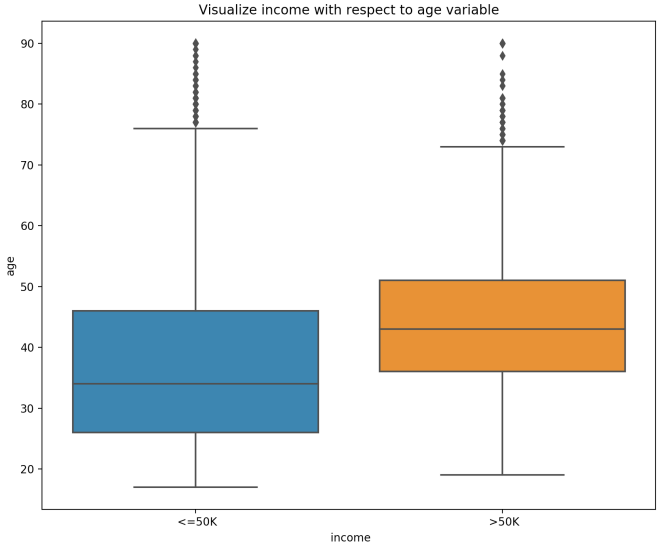
The following visualizations of the dataset were constructed as well.

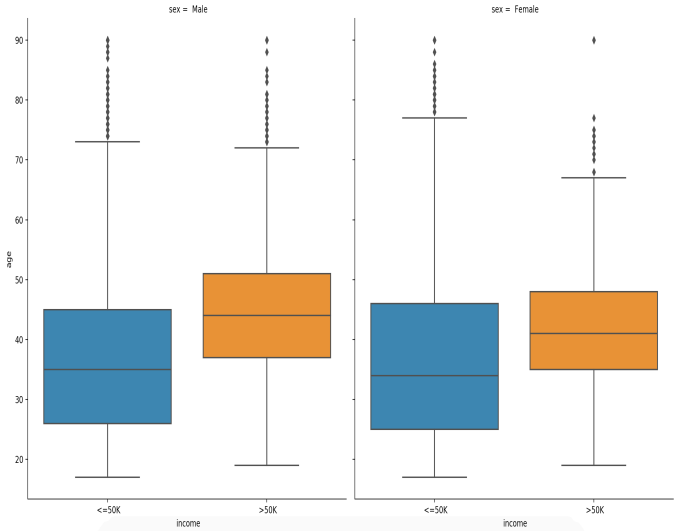


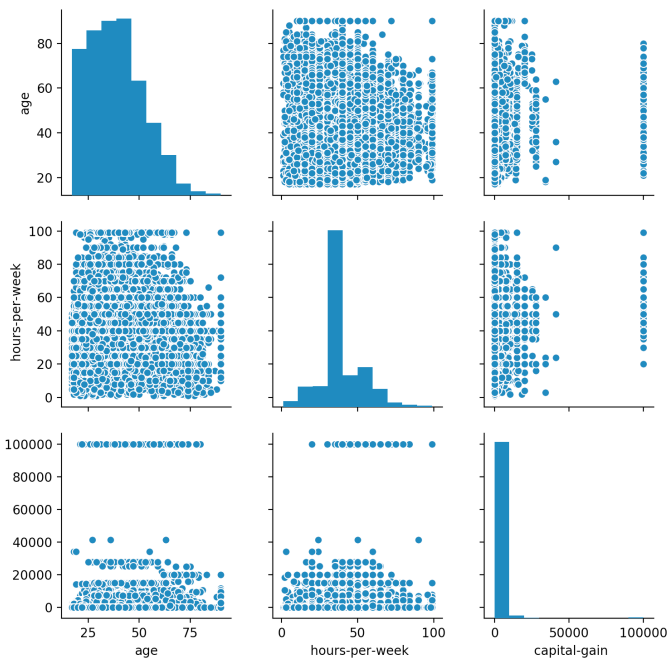


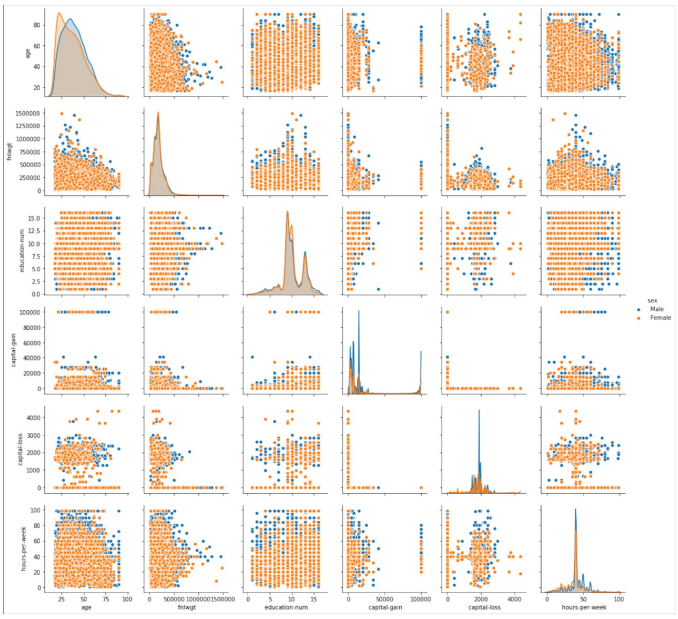












B. *Evaluation Metrics*

After the removal of incomplete entries, the 45,222 complete entries were split 80/20 into a training set and testing set. Each model was given the same training set and test set. The training set consisted of 36,177 samples and the testing set contained 9,045 samples.

The models were trained on the training set and their performance on the test set was evaluated using the following metrics:

1.) Accuracy

The sum of the correct predictions for each class divided by the total number of samples in the test set.

2.) Per Class Precision

The number of correct predictions of a class divided the total number of times the model predicted that class.

3.) Per Class Recall

The number of correct predictions of a class divided by the total number of times that class appeared in the test set.

Following the evaluation of the models, the most accurate model (a Random Forest) was used to measure the accuracy, precision, and recall of the four subsets of parameters in predicting income.

C. Major Results

In the model evaluation phase, we found that all models struggled to accurately predict the minority class. This is evidenced by the difference in the precision and, most notably, recall rates between the classes. The Random Forest was the most accurate model, and although it did not have the highest scores in all measured metrics, it did exhibit a drastic improvement in its ability to predict the minority class.

In the subset evaluation phase, we found that investments were the best predictor of income, even though capital gains or losses do not directly contribute to the income measure used.[2] Education, Occupation, and Demographics all performed only slightly better than chance, with Demographics being the worst predictor of income. All subsets also exhibited the same struggle to correctly predict the minority class. However, Investments showed a marked increase in precision compared to the other subsets.

The following tables display the results in full.

MODEL EVALUATION RESULTS

Model	Accuracy	Precision (<50K)	Recall (<50K)	Precision (>= 50K)	Recall (>= 50K)
Logistic Regression	79.64%	80.54%	96.48%	70.56%	26.58%
K Nearest Neighbors	77.27%	80.84%	91.82%	54.96%	31.42%
Decision Tree	81.62%	87.64 %	88.22%	62.10%	60.80%
Random Forest	**84.88%**	**87.70%**	**93.14%**	**73.15%**	**58.85%**
Naïve Bayes	78.90%	80.96%	94.41%	63.08%	30.09%
Artificial Neural Network	80.41%	81.01%	96.91%	74.46%	28.43%

SUBSET EVALUATION RESULTS

Category	Accuracy	Precision < 50K	Recall < 50K	Precision ≥ 50K	Recall ≥ 50K
Education	77.10%	78.44%	95.83%	62.31%	20.74%
Occupation	77.65%	80.34%	92.96%	59.83%	31.56%
Demographics	76.72%	84.61 %	84.19%	53.74%	54.54%
Investments	82.51%	81.31%	99.76%	97.57%	29.30%

D. Analysis

As mentioned in the results section, all models struggled to predict the minority class. The uneven distribution of income in the dataset caused the models to exhibit a bias toward predicting income to be less than $50,000. This tendency is revealed by the difference in precision and recall rates of the majority class. Recall rates were extremely high, meaning that the overwhelming majority of samples in the majority class were correctly predicted. However, the drop-off in precision shows that the model predicted the majority class too often, which led to incorrect predictions. Logistic Regression and Artificial Neural Networks exhibited this tendency to overpredict the majority class the most, whereas the Random Forest and Decision Tree traded slightly lower recall rates for the majority class in exchange for drastic improvements in recall and precision of the minority class. The Decision Tree and Random Forest predicted the minority class substantially more often than all other models, and although this led to a lower recall rate for the majority class, it yielded a higher overall accuracy. It would be worth investigating the performance of these algorithms on a more balanced version of this dataset.

The subset evaluation yielded the most unexpected results. One would naively assume that Occupation or Education would be the best predictors of income, but neither of these categories were as good of indicators as Investments. We cannot make a claim as to whether individuals invest because they make more money and have more disposable income, or if there is some characteristic that investors share that also leads them to find higher paying jobs. Further research would need to be done to see if this correlation holds true for other income datasets. Our data is too limited to make any substantial claims on the relevance or importance of this finding but it is an interesting finding nonetheless.

V. CONCLUSION AND FUTURE WORKS

Through our analysis of the income classification dataset, we found that a Random Forest was the most accurate model for predicting income. However, the dataset was highly unbalanced, which caused every model to be undertrained for the minority class. Due to this fact, we cannot conclusively state that a Random Forest is the optimal model for this type of classification problem. It is possible that with a larger, more balanced dataset, another model would have been more accurate. But for this dataset, the optimal model was a Random Forest.

The findings in our subset evaluations provide an intriguing hypothesis for future research. Much more data would need to be gathered and a more even distribution of samples would be beneficial for training the models. A dataset with more granularity in income would also be an improvement as the current classification groups a wide array of socio-economic classes into just two groups.

References

[1] Kohavi, Ronny, and Barry Brecker. "Adult Data Set." *UCI Machine Learning Repository: Adult Data Set*, 1996, archive.ics.uci.edu/ml/datasets/adult.

[2] US Census Bureau. "About." *The United States Census Bureau*, 29 Feb. 2016, www.census.gov/topics/income-poverty/income/about.html.