Mini Project: Socio-Economic Factors that Impact Potential Earning Income

Vathana Him

November 1, 2021

1 Abstract

The purpose of this mini research project is to examine future income based socio-economic factors based on the census data set gathered from the UCI Machine Learning Repository. This dataset contains many features that are intuitively considered to be predictors of income earning potential. These features are age, education, workclass, marital-status, occupation, relationship, race, capital-gains, capital-loss, hours-per-week worked, and native country. This data set will be used to train two-class classification models to predict which features will likely lead to an income of greater than fifty thousand dollars in USD and less than fifty thousand dollars in USD. Two machine learning models that will be trained and used to predict this scenario. The two machine learning models that will be used are Support Vector Machine (SVM) and Random Forest Classifiers. The result of this mini-project will be determined in two-scope: the determination of factors that influence potential income earning and the anaylsis of the result between SVM and Random Forest Models. The analysis of the results between the two models will be analyzed in order to understand the accuracy of the two models for this dataset, and thus, will indicate which machine learning model will be best for this dataset.

2 Introduction

Income inequality has been one of the most prominent social problems in America during modern times. It is often one of the main focuses of debates during midterm and presidential as well as through lenses of both the Republican and Demoncratic party. The increase in wealth generated for the upper-class of American household had come at the expense of the middle class as the wealth of the middle class of American families had experienced a decline since the late 1970s [1] Income has been one of the main contributors to wealth and financial security as it serves to protect American families from the cycle of economics ups and downs that is the characteristics of a capitalist economy.

The more income the average American families had, the more they're able to put into their savings and investments, and thus provided a a financial security blanket for times of economic recessions. It can be seen time and time again that throughout the history of recessions in America, the middle class had been hit the hardest. [2] Examples of this can be seen through the lenses of the most recent recessions during the 2008 Financial Crisis and the 2020 Corona Virus pandemic. These recent recessions slashed into the already thining of the American middle class as it caused many to lose their homes and jobs, and at worst their businesses and livelihood.

As cycles of elections continue to happen, many presidential candidates and lawmakers have debated their ideas and proposed solutions to solve this growing gap of the American dying middle class. The recent polarization of American politics had left no room for this problem to be solved in a timely manner as many conservative and liberal politicians continued to enforce their own point view and thus, hindering the process of providing a solution to fix this issue for the actual victims; the American middle class families.

Therefore, the purpose of this project was to look through the main contributing factors of income through a purely logical and statistical perspective based on the machine learning methods learned in class. The methods used were the training of machine learning models to find the variables that can be seen as correlating to income with the use of support vector machine learning and random forest machine learning. This

[1] Horowitz, J. M., Igielnik, R., Kochhar, R. (2020, August 17). Trends in U.S. income and wealth inequality. Pew Research Center's Social Demographic Trends Project.

[2] Weller, C. (2021, January 6). The recession hits an already hollowed-out middle class. Forbes.

was aimed to understand important factors that can contribute to income and if these factors can be predicted to determine a persons potential income based on the targed variables. Finally, the comparisons of the results from these two models were also analyzed.

3 Data Exploration

The targeted dataset was retried from the UCI ML repository known as the "Adult Data Set" or "Census Income Data Set". [3] This dataset contains multivariant relationship to predict whether an individual income exceeds or below 50 thousand dollars per year. The features of this dataset includes Age, Workclass, Income-bracket, Education, Marital Status, Occupation, Relationship, Race, Sex, Captial-gain, Hours-per-week, and Native-Country.

Before the machine learning model was built, the analysis of the overall dataset occured. By examining an overview of the characteristics of our dataset, it can be seen that the age of our population was mostly distributed between the ages of 25 to 40 years old based on 3.1 Figure 1. Additionally, the majority of the population had 9 and 10 years of total education. Capital gains, which is one of the indicators of wealth, had a heavily skewed distribution to the right as only a few percentage of the population of our dataset had a capital gains of 25000 or greater shown in 3.2 Figure 4. Finally, it can be seen in 3.1 Figure 3 that the majority of the American population had a 40 hours work week. Because of the skewness presented in the numerical data of this dataset, the min-max scaler was used on the numerical datatypes before it was fed into the machine learning model.

3.1 Data Exploration

Figure 1: Age Distribution

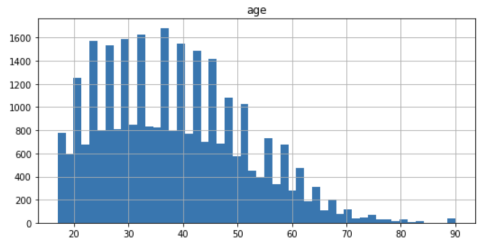
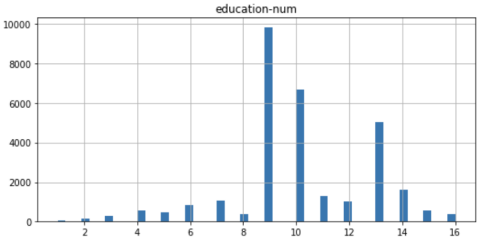


Figure 2: Education Distribution



[3] UCI Machine Learning Repository: US Census data (1990) Data Set. (n.d.). Retrieved November 21, 2021

Figure 3: Hours Work Distribution

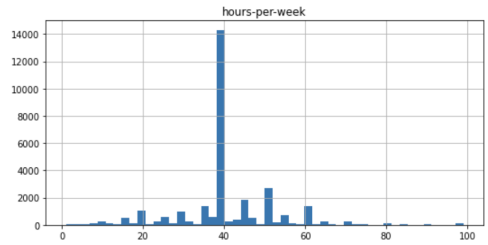
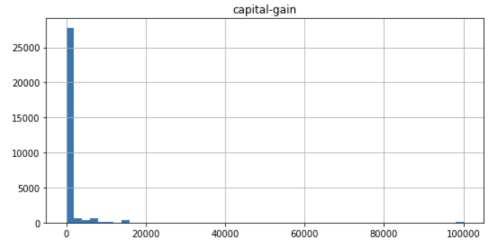


Figure 4: Capital Gains Distribution



4 Data Cleaning and Preparation

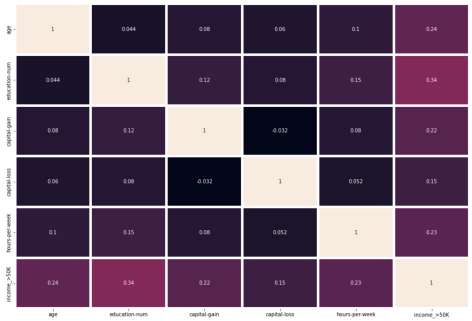
Aside from the numerical datatypes of this dataset, it also contained categorical datatypes. These categorical datatypes included occupation types, and marital/relationship statuses. Thus, in order to transform these categorical datatypes, the dummy encoding method was used. This method transformed categorical data by encoding them and transforming them into a set of binary variables. This transformed the dataset from a long format into a wide format. Subsequently, once the transformation was done, this dataset became a high dimensional dataset in which there were more features than observations. After this transformation was done, a heatmap was constructed to analyze if there were possible co-variances and co-correlated features that need to be eliminated in order to maintain the integrity of the machine learning model, removed potential biases within the dataset, and removed the high dimensional aspect of the transformed dataset.

Based on the heatmap of the numerical features in 4.1 Figure 5, if can be seen that there were some co-correlated and co-variance features to the number of income. Capital-gains and capital-loss for example had a co-correlating features to income. Additionally, for the categorical features of occupation and marital status, the heatmap was produced to determine if there were any co-correlating features. In Figure 4.2 Figure 6 that many marital and relationship statuses had co-correlating features towards income and thus, some of these features would need to be eliminated to prevent a bias in the model that could cause it to factor in relationship and marital status as the heaviest weight. Finally, based on the the occupation heatmap of 4.2 Figure 7, it can also be seen that the majority of occupation had co-correlating and co-variance relationships to income made based on the similarity in colors of the heatmap. There-

fore, in order to eliminate these co-correlating features, feature elimination would be used in order to remove potential biases of one feature over the other when training the machine learning models.

4.1 Data Pre-Processing

Figure 5: Numerical Features HeatMap



4.2 Data Pre-Processing

Figure 6: Categorical Features HeatMap

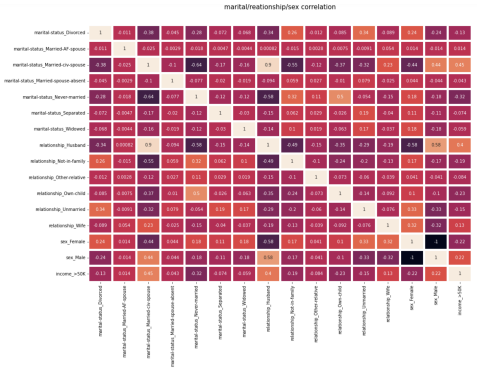
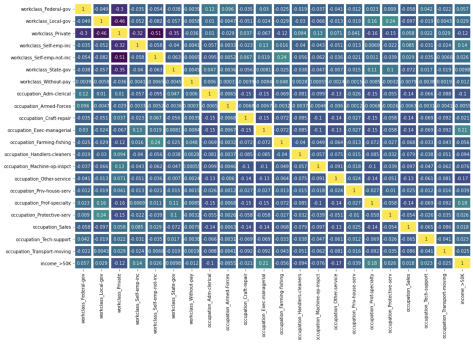


Figure 7: Categorical Features HeatMap



The feature selection method used to further pre-process the data was feature importance elimination. This method refers to the techniques that assign a score to the input features based on how useful they are in predicting the target variable through statistical correlation scores and coefficients and then eliminate any features that is below the average threshold value.[4] The feature importance elimination method resulted in a large reduction in number of co-correlated and co-variance features and eliminated the high number of features of the dataset after its categorical dummy enconding. The total number of features choosen was 11 features out of 87 features. Many of the features included variables such as age, hours work per week, education, and capital gains as these features can intuitively be determined to have a large impact on income as seen

[4]Sklearn-feature-selection-Selectfrommodel. scikit. (n.d.). Retrieved November 22, 2021, from https://scikit-learn.org

in 4.3 Figure 8.

4.3 Data Pre-Processing

Figure 8: Features Choosen

```
Index(['age', 'education-num', 'capital-gain', 'capital-loss',
       'hours-per-week', 'marital-status_Married-civ-spouse',
       'marital-status_Never-married', 'occupation_Exec-managerial',
       'occupation_Prof-specialty', 'relationship_Husband',
       'relationship_Not-in-family'],
      dtype='object')
```

5 Machine Learning Training/Predicting

In order to process our data for machine learning, the dataset was split into a train and test set of 80:20 ratio. Susequently, for the parameters of the random forest machine learning model, parameter tuning was done in order to find the optimal parameter for the random forest model for the given dataset. GridSearchCV was used to find the optimal parameters for the random forest model. With the given input of a list of maxdepth, boostrap, maxfeatures, minsamplesleaf, minsamplessplit, and nestimator, the GridSearchCV had found the optimal parameter for the random forest model with input paremeters seen in 5.1 Figure 9. For the SVM model, the linear kernel classifier would be used to train on the same dataset in order to compare the results.

5.1 Random Forest Figures

Figure 9: Parameters Tuning

```
param_grid={'bootstrap': [True], 'max_depth': [80, 90, 100, 110],
            'max_features': [2, 3], 'min_samples_leaf': [3, 4, 5],
            'min_samples_split': [8, 10, 12],
            'n_estimators': [100, 125]},
verbose=2)
```

```
grid.best_params_
```

```
{'bootstrap': True,
 'max_depth': 110,
 'max_features': 3,
 'min_samples_leaf': 4,
 'min_samples_split': 10,
 'n_estimators': 100}
```

6 Model Results

The random forest model resulted in a training and testing accuracy of 0.93 and 0.84 as shown in 6.1 Figure 10. The marginal differences in between the training and testing can be attributed to the models slight overfitting of the training dataset. Nevertheless, this marginal differences between the training and testing accuracy was not considered to be signficant enough to reject the results of the model. Class0 and class1 respresents those that made under 50K and at/above 50k per year. Because this dataset was taken from the census bareau, it was noted that there were more class0 than that of class1. This means that there were more observations who's income was less than 50k than that of at/above 50k as the dataset represents the overall socio-economic status of the United States Population. Therefore,

in order to alleviate the differences in classes, class1 was assigned the same weight as class0. However, the adjustments of class weights still affected the prediction of class1. This can be seen in in 6.1 Figure 11 as the true positive score for class0 is signficantly higher at a recall rate of 0.93 compared to class1 with a recall rate of 0.55. The adjustments of weights between class0 and class1 help alleviate the inbalanced dataset for training the random forest model based on the the precision value of 0.86 and 0.73 for class0 and class1, however, it still created a bias towards class0 based on the recall rate.

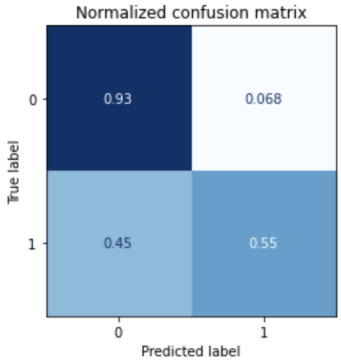
6.1 Random Forest Figures

Figure 10: Random Forest Score

	precision	recall	f1-score	support
0	0.86	0.93	0.90	4532
1	0.73	0.55	0.63	1501
accuracy			0.84	6033
macro avg	0.80	0.74	0.76	6033
weighted avg	0.83	0.84	0.83	6033

Training Accuracy: 0.9327779849973061
Test Accuracy : 0.8378915962207857

Figure 11: Random Forest Confusion Matrix



The SVM model resulted in a training and testing accuracy of 0.82 and 0.82 respectively based on 6.2 Figure 12. There was no difference between the accuracy of the train set and the testing set, which indacated that the SVM model had no biases towards the train set or testing set. However, the recall rate was similar to that of the random forest model as it had a high true positive rate for class0 and low true positive rate for class1 based on 6.2 Figure 13. This exposed the imbalance nature of the datase. Subsequently, it also showed that the adjustment of class weights to alleviated the dataset's imbalanced characteristics did not imporve the model to a large extent. Additionally, the precision rate for class0 and class1 was 0.86 and 0.66. There was a larger degree of difference in precision rate of class0 and class1 in comparison to the precision rate of the random forest model. Thus, indicating that the SVM model was more biased towards class0 to a larger extent than the ran-

dom forest model.

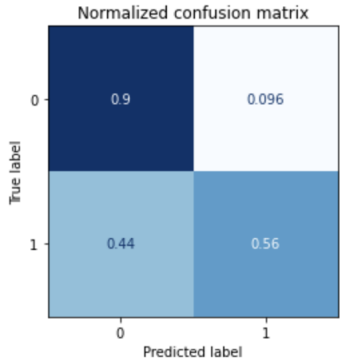
6.2 SVM Figures

Figure 12: SVM Score

	precision	recall	f1-score	support
0	0.86	0.90	0.88	4532
1	0.66	0.56	0.61	1501
accuracy			0.82	6033
macro avg	0.76	0.73	0.74	6033
weighted avg	0.81	0.82	0.81	6033

Linear SVM Test Score Scale : 0.8186640145864412
Linear SVM Train Score Scale : 0.8239877326039206

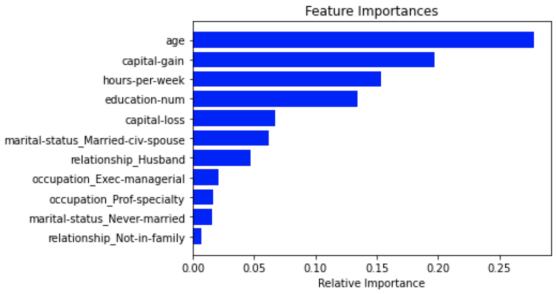
Figure 13: SVM Confusion Matrix



Finally, feature importance ranking was done to see which features were indicative in determining potential income earning. Based on the 6.3 Figure 14, it was found that age, capital-gain, hours-per-week, and education-num were, to a large extent, strong indicators of potential earning incoming as each of the four features scored highest in terms of its relative importance to the predicted varaible. This would give lawmakers a quantitative idea of what variables can help alleviate the shifting wealth gap in the United States as they could design their social and economic policies around those variables.

6.3 Features

Figure 14: Feature Score



7 Conclusion

In contrast to the random forest multiple, the support vector machine model took longer to train in terms of its time complexity. The SVM model took aproximately 6 hours to train in comparison to 1.2 seconds of time that it took to train the random forest model. Thus, for speed and efficiency, the random forest model performed better than that of the SVM model. In terms of its accuracy and precision, the SVM model tend to show more biased towards class0 than that of the random forest model. Additionally, the random forest model also scored a higher accuracy in both its training and testing dataset. The random forest model re-

call rate for both class0 and class1 was also significantly higher than that of the SVm model, and therefore, the random forest model performed better in its accuracy and precision and should be used for predicting potential income in order to shed an insight into the current social-economic problem.

Within the scope of socio-economic variables that impacted potential earning income in this dataset; age, capitalgain, hourswork, and number of years in education were variables that were indicative of determing potential income earning. Although age can be an ambiguous variable in determining income, capital gain, hours work and number of years in education were not. Lawmakers can design their policy to adjusting capital gains, and wages in order to lessen the divide in wealth inequality between socio-economic classes in America. Education was also seen as an important indicator to income, thus, policies that are designed to reduce education expenses and provide equal opportunity to access higher education would also play a pivotal role in bridging the gap of socio-economic classes in the United States and strenghtening the middle class.

References

[1] Horowitz, J. M., Igielnik, R., Kochhar, R. (2020, August 17). Trends in U.S. income and wealth inequality. Pew Research Center's Social Demographic Trends Project.

[2] Weller, C. (2021, January 6). The recession hits an already hollowed-out middle class. Forbes.

[3] Income inequality. Inequality.org. (2021, August 30). Retrieved November 20, 2021, from https://inequality.org/facts/income-inequality/.

[4] Schaeffer, K. (2021, May 27). 6 facts about economic inequality in the U.S. Pew Research Center. Retrieved November 20, 2021, from https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/.

[5] UCI Machine Learning Repository: US Census data (1990) Data Set. (n.d.). Retrieved November 21, 2021

[6] Sklearn-feature-selection-Selectfrommodel. scikit. (n.d.). Retrieved November 22, 2021, from https://scikit-learn.org