MGT 6203 Project Progress Report

Ngoc Le, Carlos Moncada, Brandon Ritchie, Usha Sharma, Hal Zhang

GitHub Repository: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-23

Introduction

In an era where data-driven decisions are paramount, understanding the nuances behind employment and income levels has never been more critical. This project delves into the intricate dynamics of labor statistics to identify factors influencing income levels, particularly focusing on the dichotomy between those earning above and below $50,000 annually. The importance of this study is underscored by the analysis provided by Autor, Katz, & Kearney [1], which highlights the growing wage inequality in the United States and its implications on socioeconomic mobility. Our approach brings a fresh perspective by employing advanced data cleaning techniques, feature selection, and a comparative analysis of various predictive models, addressing the urgent need to understand the forces behind this growing disparity.

Through meticulous data cleaning and preparation, we address challenges such as skewed distributions and inconsistent missing data encoding, ensuring our analysis rests on a solid foundation. By evaluating a suite of models including Logistic Regression, Decision Tree, Naive Bayes, and Random Forest, we aim to uncover the most significant predictors of income levels. Our project stands out by not only identifying these predictors but also by interpreting their implications in the context of current labor market trends. This endeavor is not just an academic exercise; it provides actionable insights for policymakers, economists, and social scientists seeking to understand and mitigate income disparities. Actions into improving income disparity can lead to lasting impacts in people’s lives especially children [2].

Data Cleaning

Before applying any analytical models, we needed to do preliminary analysis and cleaning of the data. We initially analyzed the response variables that we wanted to use, “wage_per_hour”, a quantitative variable expressing hourly salary and “income_above_limit”, a categorical variable expressing whether a person makes a salary greater or less than 50,000 dollars a year. The values of wage per hour were right skewed ranging from 0 to 9999 with a mean of 1398.5. Due to the dataset not providing the units for wage per hour and the abnormal spread, we decided to not use wage per hour for our response variable; however, we are able to use income_above_limit and converted that column to binary values with 1 being people who made above 50,000 dollars. As a result of using income_above_limit as our only response variable, we must discard the test dataset that was provided and split up our training dataset to do training, validation, and testing.

We then cleaned the predictor variables. Our dataset has 43 columns. We cleaned the columns by first looking at which columns have missing values. Unfortunately, this dataset had different values for missing values in different columns, so we changed all the missing values to “NA” to keep everything consistent. Then we looked up all the columns that had more than 20 percent of NA values and eliminated them as we believed the data would be too sparse to interpolate. We were left with 27 predictor variables after taking out those columns.

For the 27 predictor variables left many of them were categorical columns. We went through each and calculated the percent above income threshold to determine which factors were most similar. Factors with similar classifications and like percentages were combined to avoid multi-collinearity in the modelling step. For example, we combined all the categories for education that were under high school (1st grade, 2nd grade, etc.) into one unifying category of "under high school". To make these categorical variables easy for the models to use, we "one hot encoded" them. Finally, we filtered the cleaned data into two separate data frames of employed and unemployed. We will analyze the employed data frame to see which variables are significant predictors of income. We will analyze the unemployed data frame to see what differences these people might have compared to the people who are employed.

Data Analysis

For the data analysis, we evaluated the data on logistic regression, decision tree, random forest, and naive bayes model to see which ones would produce the best accuracy and precision. In addition, these models can be used for variable selection as well. The choice of employing a Random Forest model, in particular, was informed by Breiman's foundational work on this methodology [3], which has been noted for its robustness in handling complex datasets like ours. Random Forest's ability to deal with unbalanced data and its importance in variable selection made it an invaluable tool in our analysis, aiming to uncover the most significant predictors of income levels.

Logistic Regression

From the cleaned dataset, the employed dataset was used to perform a logistic regression model with response variable as "income_above_limit" and rest of the variables as predictors.

Conclusion

The outcome is mostly as we hypothesized. Parameters which were found to be most significant are higher education, and it has a positive coefficient indicating that it raises income. Race and gender are also significant. Females have lower income as compared to males. Non-white races also have lower income as compared to white. Some outcomes are interesting, for example being born in the United States is not significant, even though being a U.S. citizen does increase income.

Some types of marital status are not significant, but tax filing jointly is significant.

Model details

Coefficients: (5 not defined because of singularities)

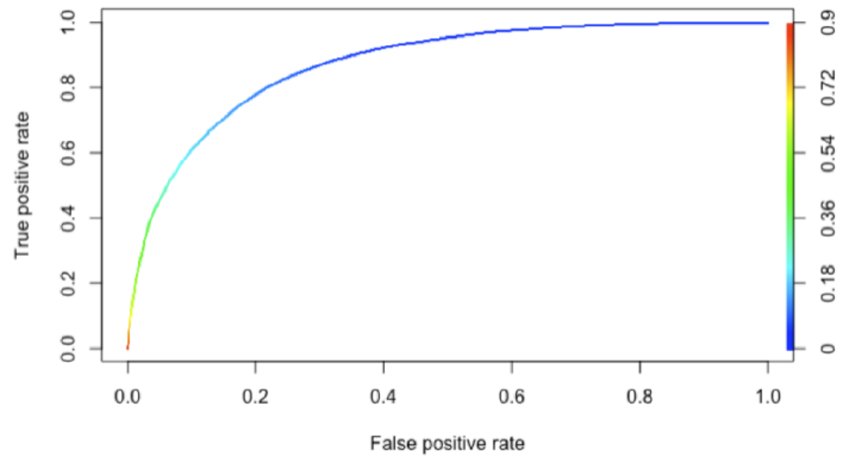
	Estimate	Std.	Error	Z value	Pr(>|z|)
(Intercept)	-5.493731	0.127153	-43.206	< 2e-16	***
age	0.034404	0.001111	30.966	< 2e-16	***
week_working_hours_above_50	1.435690	0.035538	40.399	< 2e-16	***
head_of_house	0.430331	0.028780	14.953	< 2e-16	***

us_citizen	0.377284	0.071271	5.294	1.20E-07	***
parents_from_us	-0.081289	0.023090	-3.52	0.000431	***
born_in_us	0.065078	0.061666	1.055	0.291273	
educationAssociates	0.173590	0.046735	3.714	0.000204	***
educationBachelors	1.167344	0.032797	35.593	< 2e-16	***
educationHS_Grad	-0.559574	0.035897	-15.588	< 2e-16	***
educationLess_than_HS	-1.382252	0.064753	-21.346	< 2e-16	***
educationMasters	1.731340	0.040132	43.141	< 2e-16	***
educationPhD_or_greater	2.465772	0.050236	49.084	< 2e-16	***
educationSome_College	NA	NA	NA	NA	
genderFemale	-1.167784	0.029122	-40.099	< 2e-16	***
genderMale	NA	NA	NA	NA	
raceAsian	-0.162227	0.072305	-2.244	0.024855	*
raceBlack	-0.382709	0.050301	-7.608	2.77E-14	***
raceHispanic	-0.442455	0.054037	-8.188	2.66E-16	***
raceNative_American	-0.484521	0.156076	-3.104	0.001907	**
raceOther	-0.108805	0.223920	-0.486	0.627033	
raceWhite	NA	NA	NA	NA	
marital_statusDivorced	0.228817	0.082496	2.774	0.005543	**
marital_statusMarried_Spouse_Absent	0.263548	0.142172	1.854	0.063778	.
marital_statusMarried_Spouse_Present	-1.746102	0.750443	-2.327	0.019978	*
marital_statusSeparated	0.127083	0.118935	1.069	0.285295	
marital_statusUnmarried	-0.185071	0.086223	-2.146	0.031839	*
marital_statusWidowed	NA	NA	NA	NA	
tax_statusHoH	0.214888	0.062207	3.454	0.000551	***
`tax_statusJoint_<65`	2.307298	0.749539	3.078	0.002082	**
`tax_statusJoint_>=65`	1.401283	0.751000	1.866	0.062057	.
tax_statusNonfiler	-2.317682	0.261855	-8.851	< 2e-16	***
tax_statusSingle	NA	NA	NA	NA	

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80345  on 122779  degrees of freedom
Residual deviance: 56684  on 122752  degrees of freedom
AIC: 56740

Number of Fisher Scoring iterations: 9
```

Accuracy and AUC

The accuracy of the model with prediction threshold as 0.5 is pretty good with AUC as 0.8716539.



Random Forest

We then created a random forest model by splitting the employed dataset into training and testing and trained the random forest model on the training dataset with “income_above_limit” as the response and the rest of the variables as predictors. We predicted the test data set and the confusion matrix and accuracy statistics are below. We also looked at the importance of variables of the random forest model.

Conclusion

It is seen that age, being female and working more than 50 hours a week are the most important factors.

Model details

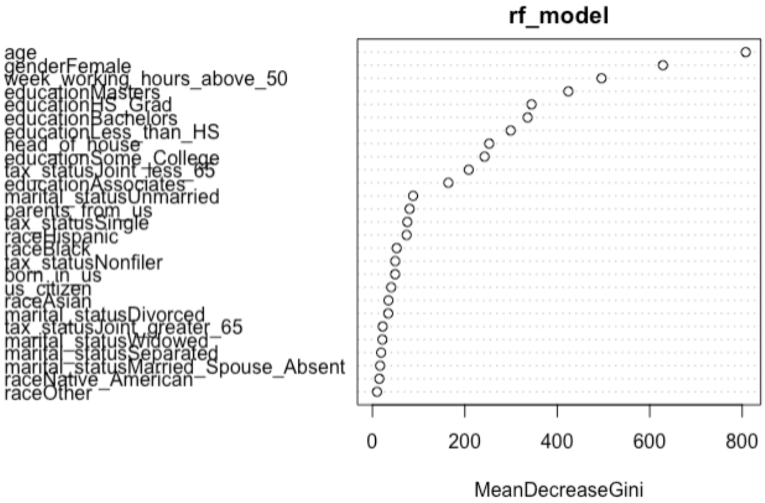
```
Number of trees: 500
No. of variables tried at each split: 5

 OOB estimate of  error rate: 9.24%
Confusion matrix:
     0    1 class.error
0 85414 1323  0.01525301
1  7591 2142  0.77992397
```

Confusion Matrix and Statistics

```
          Reference
Prediction     0     1
         0 23303  2049
         1   344   614

               Accuracy : 0.909
                 95% CI : (0.9055, 0.9125)
    No Information Rate : 0.8988
    P-Value [Acc > NIR] : 1.121e-08
```

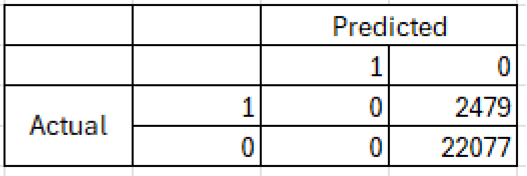
```
               Kappa : 0.3017

 Mcnemar's Test P-Value : < 2.2e-16

          Sensitivity : 0.9855
          Specificity : 0.2306
       Pos Pred Value : 0.9192
       Neg Pred Value : 0.6409
           Prevalence : 0.8988
       Detection Rate : 0.8857
 Detection Prevalence : 0.9636
    Balanced Accuracy : 0.6080

      'Positive' Class : 0

Pos Pred Value
     0.919178
Sensitivity
  0.9854527
```

Variable Importance plot



Decision Tree

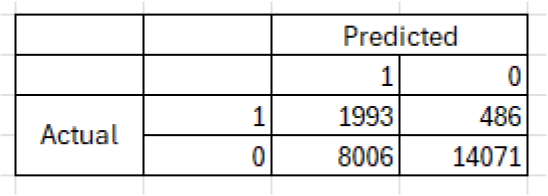
The decision tree model on the training data showed that working_hours_above_50, genderFemale, age less than 33.5, educationHS_Grad, educationLess_than_HS and education_Some_College were the significant variables used for the branching in the tree. The accuracy of the tree on the validation data was good at 0.899; however, the precision was 0 as the tree did not predict any true or false positive. The confusion matrix produced is below.



		Predicted	
		1	0
Actual	1	0	2479
	0	0	22077

Naive Bayes model

Finally, we built a Naive Bayes model with income_above_limit as the dependent variable and the rest of the features as the independent variables. The accuracy of the model in predicting the validating data set was 0.654 with a precision of 0.804. The confusion matrix produced is below.



		Predicted	
		1	0
Actual	1	1993	486
	0	8006	14071

Discussion

The results of the four models as seen above showed what we expected, being a female, having less education, being of non-white race all led to decreased incomes. This was consistent throughout our models which is a good sign that these variables are clearly significant. All our models obtained good accuracy on the test data set however, an issue we did not expect was how poorly some of our models predicted precision especially the decision tree and the random forest. This is indicative that our dataset is imbalanced with almost a 10:1 ratio of negative to positive predictor responses. Handling this problem will be something we talk about in the next steps section below.

Next Steps

A primary issue that we are experiencing with some of our initial models is a tendency to classify everyone as below the $50 K threshold. This is expressed through high accuracy and low precision scores. This is likely because our predictor column is very unbalanced in favor of people below the threshold as mentioned above.

One solution to this problem is to utilize bagging and boosting ensemble learning methods. With bagging (bootstrap aggregation) base models (most likely decision trees) are trained on samples of the data with replacement. This diversity of samples allows for a more robust aggregation than a single sample would provide.

Boosting algorithms like AdaBoost and Gradient Boosting iteratively train weak learners, focusing more on the misclassified instances in subsequent iterations. This allows the model to pay more attention to the minority class.

Along with the algorithmic techniques described above, we could also try some resampling techniques and evaluate the results on a holdout dataset. These could include:

- **Undersampling:** Randomly remove observations of the majority class.
- **Oversampling:** Randomly replicate instance of the minority class.
- **Synthetic Minority Over-sampling Technique (SMOTE):** Generate synthetic instances for the minority class to balance the dataset, preserving the underlying characteristics of the minority class (probability distributions etc.) [4,5].

After using these techniques above to balance out our dataset, we will rerun the models we have run above in the data analysis section on our dataset split into training, validation, and test datasets. We will use the model with the best accuracy and precision to determine which variables contribute the most to income disparity. We will analyze these variables to see which ones can lead to action that will improve income disparity.

Conclusion

As we reach this stage in our exploration of the factors influencing income levels, our project has made significant strides in understanding the intricate dynamics of employment and income disparities. Our progress is marked by the successful employment of advanced data cleaning techniques, feature selection, and the comparative analysis of various predictive models. This rigorous approach has enabled us to uncover preliminary insights, such as the importance of education, gender, and race in determining income levels, which align with some of our hypotheses while also presenting new avenues of inquiry.

Our endeavor has not been without challenges. One of the primary issues we encountered was the tendency of our initial models to skew towards predicting lower income levels, a reflection of the inherent imbalance in our dataset. This challenge has provided us with a valuable learning opportunity, leading us to consider the implementation of bagging and boosting ensemble methods, as well as resampling techniques to achieve a more balanced and accurate model representation.

Looking forward, our immediate next steps involve refining our predictive models through these advanced methods. We aim to not only improve the accuracy and precision of our predictions but also to deepen our understanding of the variables that contribute most significantly to income disparity. This phase of our work will be critical in identifying actionable insights that can inform policies and interventions to mitigate income inequalities.

In conclusion, this progress report reflects both the achievements and the hurdles we have encountered in our journey to dissect the factors contributing to income inequality. While we have made considerable progress, we recognize that our work is ongoing. We are committed to continuing our research to contribute meaningful insights to the discourse on economic equality and informing strategies that can lead to tangible improvements in addressing income disparities.

Works Cited

1. Autor D, Katz L, Kearney M. Trends in U.S. Wage Inequality: Revising the Revisionists. Rev Econ Stat. Available from: https://scholar.harvard.edu/lkatz/publications/trends-us-wage-inequality-revising-revisionists. Accessed March 13, 2024.
2. Ratcliffe C, Kalish E. Escaping Poverty: Predictors of Persistently Poor Children's Economic Success. Urban Inst; May 2017. Available from: https://urban.org.
3. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
4. Birla S, Kohli K, Dutta A. Machine Learning on imbalanced data in Credit Risk. In: Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); 2016 Oct; Vancouver, BC, Canada. IEEE; 2016. p. 1-6.
5. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–232. Available from: https://doi.org/10.1007/s13748-016-0094-0.