DSC 630 Final Paper

Income As It Relates To Individual Characteristics

Lara Clasen

**[A close up of a logo

Description automatically generated](https://drive.google.com/file/d/1quT3zTgYvpNEiePPgwoAaQpwJaJUVWnG/view?usp=sharing)**

Abstract

This project reviews United States Censes income data as it applies to related characteristics. The utilized data set contains variables describing individuals as well as their income level of either higher than or lower than fifty-thousand dollars annually. Variables are assessed as to their abilities to predict income level. Analysis is run on the variables and two of these is highlighted as being highly able to predict income level based on their own levels. Further research is required to determine the ability of less accurate variables to predict the same measurement.

Intro/Background

For my project I chose to look at United States Adult Census income data as it relates to social factors such as age, race, education, et cetera. The problem I will address is whether any one of these characteristics has a significant effect on income level in the United States. I look at each characteristic individually as it applies to income, with efforts to isolate them from other characteristics. With this I aim to conclude basic assumptions of income level based on other aspects of a person’s being. I selected this data set because I found the U.S. Census to be a reliable source of data, but continued to look to outside sources for comparison with what I found in my analysis. The analysis of this data also presents a unique opportunity to answer many socioeconomic questions about income disparity in our country. We generally know that factors such as race can correlate with income level, but the goal is to see how that effect changes when effects from other characteristics are being controlled.

Methods

My project plan of action was to identify the most efficient methods of analysis and apply a trial and error process as I worked through my data. I attempted to utilize methods we learned about during the course. This was an entirely new data set for me as I have not previously pulled data from the US Census.

The U.S. Census data set that I utilized came with pre-designated test and train sets. I began by viewing each of my included data sets using the head() function. By doing this we can see consistency between columns to realize that we are working with closely related data “chunks”. Following this step, I ran summary statistics on each data set using the describe() function. This function provides various metrics for our set including mean, standard deviation, and min/max values (see Figure 1 below for example of this function as applied to train set).

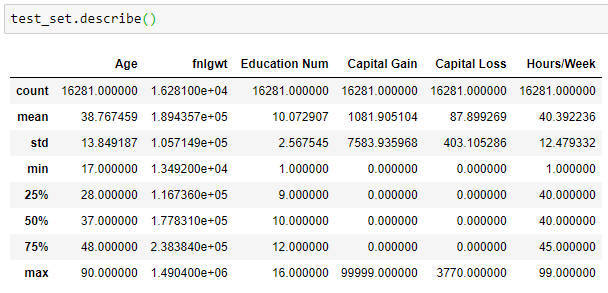


Figure 1

In the following step I obtained several visualizations of my data. This allows me to better grasp potential relationships or lack thereof, as well as to identify variables that might not be meaningful for this analysis. I was able to create histograms for appropriate variables, including the below Figures 2 & 3: Hours Worked Per Week, and (our target variable) Income Level. From these visualizations we can see that 40 Hours Worked Per Week is by and far the most common number of hours worked, and that the majority of our data includes an Income Level of less than $50,000 per year.

A screenshot of a cell phone

Description automatically generated

Once I have reviewed the data content and looked at initial visualizations, it is time to begin to clean the data. This will produce data that is easier to analyze and fits within the questions we want to answer. The first step that I took to clean the data was to delete the first rows since they are erroneous to our analysis. This was done using the drop function and an indexing value of [0]. We can then proceed with our analysis without worrying that the first, invalid, row will conflict our results.

An important step when cleaning a new data set is to search for and resolve any missing values. During our search for missing values, it reported that none were found in our data. This means that we do not have to account for any of those values as we move forward. We created a function to search for these missing values, and none were returned based on those outlined criteria (see Figure 4 for the resulting counts). I also write a function to separate numerical values from string values in the data. This will help to differentiate the two going forward.

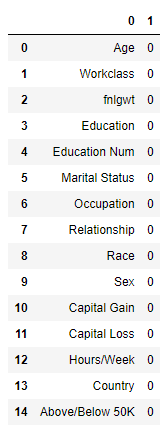


Figure 4

Another important step, related to finding missing values, is to find and resolve all other *null* values in the data. Null value can be invalid characters or values that do not match the rest of the data set as they are expected to, or even missing values. In our case, we find that we have some null values in the form of an invalid character, or a question mark (“?”). We must resolve these in order to process with an accurate analysis of the data. To do so, we write a function to locate any of these particular characters and essentially drop them. See Figure 5 for the function used and the resulting counts.

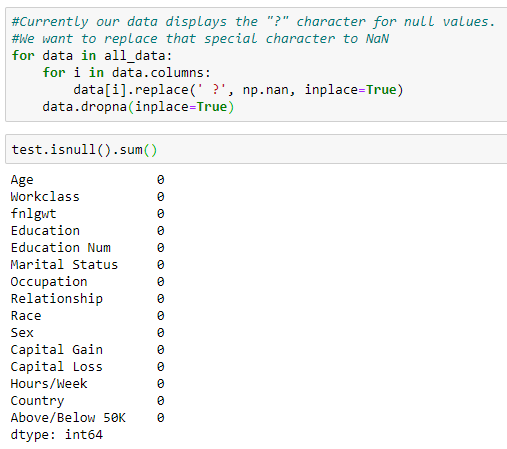


Figure 5

In the next step we define our target variable as Income Level. We have now also been able to see that the Education and Hours per Week variables appear to provide a good look into where the Income Level variable will fall. These will be our expected predictors of the target variable.

We can continue with some important cleaning steps that involve the concept of binning. By binning our variables, we can “group” them into like classes based on averages that will allow us to get a more clean-cut picture of the relationships that do or do not exist. A couple examples of this is with our two important variables: Education and Hours per Week. The Education variable can be binned into level of education within three groups: Low, Medium, and High. Similarly, we can bin the Hours per Week variable into levels of number of hours per week worked. Initial exploration suggests that 40 hours is the most frequent value, which is equivalent to a 40-hour work week. We will bin this variable around our most frequent value and from there we will bin into Low, Medium, High, and VeryHigh. We continue by binning our Occupation, Age, Marital Status, Race, and Work Class variables into relevant groupings. See Figure 6 for one example of this process.

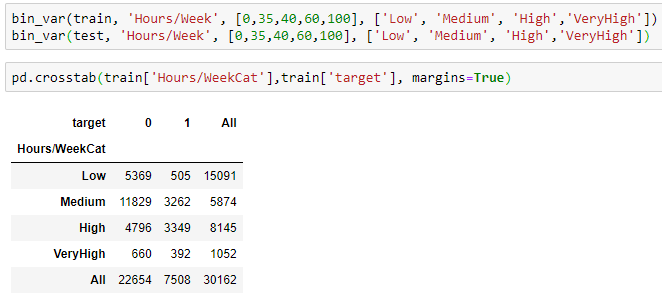


Figure 6: Binning Hours per Week variable

Finally, we begin with feature selection and prediction algorithms. Variance Threshold allows us to remove features where the variance doesn’t meet a certain threshold. It removes all zero-variance features, so those that have the same value in all samples. For our data we are aiming to remove variables which have more than 80% values that are either 0 or 1. After this removal, the number of total columns has been reduced to 15 because of the specified variance threshold. The removed columns had the same value in 80% of the observations.

We then work to create a confusion matrix for our data. A confusion matrix is a “summary of prediction results on a classification problem” (Sharma, 2018). It essentially shows ways in which your classification model is “confused” when it makes predictions. See Results for our obtained confusion matrix along with other prediction results.

Results

As mentioned in our Methods section, we chose to create a confusion matrix for our data. See Figure 7 for the resulting confusion matrix.

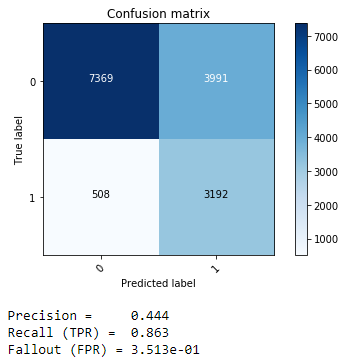


Figure 7: Confusion matrix and metrics

Precision of the confusion matrix is the ability of a classification model to identify only the relevant data points. Precision is defined as the number of true positives divided by the number of true positives, plus the number of false positives. False positives are cases which the model incorrectly labels as positive that are actually negative. Recall is the ability of a model to find all of the relevant cases within a dataset. While Recall expresses the ability to find all relevant instances in a dataset, Precision expresses the proportion of the data points our model says was relevant actually were relevant. Our Recall here is .863 which is fairly good. However, our Precision is just .444 which shows room for improvement.

I will begin by applying several cleaning techniques on my data set, such as accounting for any missing or null valued. I will select certain variables, such as work week length or skill level, as indicator variables to compare to other features in the set. From here I can run linear regressions to find any potential effects between variables.

In a final step we attempt to understand which of all of our variables are important features for High and Low Paying Jobs. See Figure 8 for the resulting numbers. We can see that positive coefficients are for high paying and negative coefficients are for low paying jobs. A High number of years of education along with high work hours per week are important for obtaining high income.

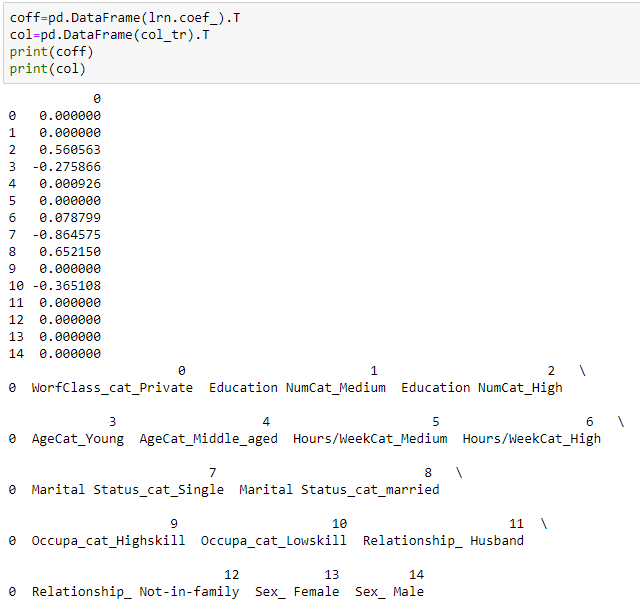


Figure 8

Discussion/Conclusion

Overall, we can confidently say that several of our included characteristics play an important role in predicting income level. Specifically, Education Level and Hours Worked Per Week can pretty accurately predict income level. This is valuable information, however, in my opinion the combination of characteristic variables used are somewhat limiting. Of course we are likely to see that Hours Worked Per Week directly contribute to higher income level, as there is a direct and mathematically understandable link between the two. However, the variables such as Race or Marital Status are more interesting variables to compare. For this reason, I would choose to remove the more obviously connected variables from this data set if I were to run my analysis again. This way we would be more likely to see lesser known links between unexpected variables and income level, which is a more valuable result for society.

Another concern of the data is the overwhelming number of data points that are categorized as White. It would be imperative to expand this categorization to include a greater number of other races in order to obtain a more relevant outcome of the analysis. I fear that with the current data used, only so many inferences can be made to the larger population as it stands in modern times. This could likely be resolved simply by using a more recent data set of the U.S. Census.

Acknowledgements

For the completion of this paper and project I would like to acknowledge a few parties and resources. First, the texts for our course have become an invaluable partner in this process. Both Applied Predictive Analytics by David Abbott and Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel have proven to be crucial to our learning. Each text has its own unique way of walking through the power of predictive analytics and they are beneficial individually and together.

Another acknowledgement must go to our educator in this course as well as my fellow students. Both parties have been central to success in this class. I have found the ability to connect with and bounce ideas off of others in the class to be incredibly helpful over the course of this term and I hope to be able to utilize that type of open communication in future courses.

References

Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Indianapolis, IN: John Wiley & Sons.

Grinstein, G., Trutschl, M., & Cvek, U. (n.d.). High-Dimensional Visualizations. Retrieved from https://pdfs.semanticscholar.org/43f7/66c06e2a7770d9f37dcd9cfff5bd5dcfc22f.pdf

Ippolito, P. P. (2019, October 11). Feature Extraction Techniques. Retrieved from https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be.

Moffitt, C. (2019, October 14). Binning Data with Pandas qcut and cut. Retrieved from https://pbpython.com/pandas-qcut-cut.html.

Reddy, M. A. (2019, September 6). A Comprehensive guide on handling Missing Values. Retrieved from https://medium.com/bycodegarage/a-comprehensive-guide-on-handling-missing-values-b1257a4866d1.

Sarkar, D. (2018, January 5). The Art of Effective Visualization of Multi-dimensional Data. Retrieved from https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

Sharma, A. (2018, February 7). Confusion Matrix in Machine Learning. Retrieved from https://www.geeksforgeeks.org/confusion-matrix-machine-learning/.

Siegel, E. (2016). Predictive analytics: the power to predict who will click, buy, lie, or die. Hoboken: Wiley.

Swaminathan, S. (2019, January 18). Logistic Regression - Detailed Overview. Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.