**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 1

Understanding Income Inequality

Jeff Hackmeister

5/25/2025

**Introduction**

For this week’s project, we are working with demographic data from the US census. The data contains 15 different attributes for 48,841 individuals. Our goal is to utilize 14 of the variables to construct a prediction model for income to better understand the attributes that lead to prosperity and affluence in the United States in hopes of provide guidance to organizations working to shrink the wealth gap and confront income inequality. This particular data set does not contain raw values for income, rather it is demonstrated in a categorical variable that indicates an annual income of up to $50,000 a year or great than $50,000. The explanatory variables include age (in numeric years), work class, education, marital status, occupation, relationship, race, sex, capital gain and loss values and country of origin. With this mix of numeric and categorical variables, a nearest neighbors model has been deemed the most appropriate.

**Methodology**

Before a model can be fit to the data, we must first take several data cleaning steps to prepare the dataset. First, we drop redundant or unnecessary variables from the data. In this is instance fnlwgt (a estimation of each respondent’s representation in data) and education (which is redundant with the inclusion of education-num, a numeric representation of year of education) were removed. Next, we looked for missing or null values. With a dataset of over 48,000 observations, missing values below 6% were deemed insignificant to the analysis and were removed, this resulted in a reduction of 3,620 rows in our dataframe. We also ensured the sting variable data was consistent but performing a text cleaning process.

To better understand the dataset, we performed an exploratory data analysis (EDA). We looked at the distribution and correlation of numeric values; visualizations of this analysis can be found in the appendix.

To further prepare for to fit our data to a nearest neighbor model, we created categorical representations of the numeric values by assigning them to groups. For example, ages were binned in Young, Young Adult, Middle Age, Senior and Elderly. Then we separated our data into test and training subsets. The training dataset contained 36,176 observations, which left 9,045 for testing. Checking the distribution of our target variable, income, we see that over 75% of our training data observations belong to the under $50,000 per year group. This closely aligns with the distribution from the entire dataset, as seen in the EDA.

**K Nearest Neighbors**

The final step in preparing to fit our model is to identify the optimal K value. The nearest neighbors technique relies on calculating the distance between every observation in the data (a multidimensional distance, based on each of the explanatory variables, usually an E*uclidean* distance) and then deciding on a neighborhood value, which is the number of nearest values, the K value, to the observation to be predicted. For example, if we choose a K values of 5 we would find the closest 5 observations to our test and then whichever income category was most prevalent among those 5, that category would be predicted. Since the K values plays such a critical role in the performance of the model, we will use a loop function to actually train our model multiple times, each with a different K value, to find our optimal K value for accuracy. This process, a 5-fold stratified cross validation, returned an optimal K values of 29. In the visual found in the appendix, you can see that the model accuracy exceeds 0.84 at K = 6 and hits 0.85 at K = 29. With this in mind, we fit the KNN model with K = 29 to our training dataset and then use the model to make predictions on our testing data. The resulting confusion matrix, available in the appendix, shows that we have a strong model with a test accuracy of 84.2%. When compared to the training accuracy of 85.6%, this leaves a 1.4% overfitting gap, which is well withing the acceptable range.

**Model Evaluation**

Calculating the performance of the model, we find the model demonstrates excellent performance in identifying low-income individuals (≤$50K) with 91.7% specificity, correctly classifying over 9 out of 10 low earners. For high-income detection, the model achieves 62.9% sensitivity, successfully identifying approximately 2 out of 3 high earners. When the model predicts someone as a high earner, this prediction is accurate 71.4% of the time, indicating good precision despite the class imbalance challenge. The model rarely misclassifies a low earner as a high but can misclassify a high earner as a low earner. When the model predicts a high earner, it is very reliable. Given that nearly ¾ of the dataset contained information on low earners, these are not unexpected results.

These results can provide meaningful insights for groups working to lower income inequality. By creating programs and outreach for people that match the characteristics the model predicts to be low earner, advocates can be assured that most of the people matching those characteristics will be low earners. Conversely, individuals matching the profile predicted to be high earners and very unlikely to be low earners who could, potentially, benefit from these programs if they were not misclassified.

To improve the performance of the model, additional observations could be included to the dataset - specifically to add more high-income individuals to better identify the separation between the categories.

**Next Steps**

With this model established and limitations noted, it is important to understand which variables have the biggest impact on the earnings of the individual. We utilized feature importance analysis to determine which variables have the biggest impact on income. Surprisingly, education ranked much lower (10th) than conventional wisdom would suggest. The biggest determinants were access to capital, as seen in captain gain and capital loss, followed by self-employment. While education is still an important factor, this analysis indicates that in the dynamic US economy, access to capital and entrepreneurship opportunities have the potential for larger impact on affluence. Given the high rates of student debt and proliferation of college graduates, it is understandable to see the data prove out that higher education is no longer the guarantor of financial success that is had been for previous generations.

**References**

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[2] Luke, T. (2022, April 9). *Create Your Own k-Nearest Neighbors Algorithm in Python | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/create-your-own-k-nearest-neighbors-algorithm-in-python-eb7093fc6339/

[3] Rukshan Pramoditha. (2024, February 27). *Choosing the Right Number of Neighbors (k) for the K-Nearest Neighbors (KNN) Algorithm | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/choosing-the-right-number-of-neighbors-k-for-the-k-nearest-neighbors-knn-algorithm-fbc635279ec7/

‌[4] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

**Appendix A - Images**