MAS 4106 – Matrix Methods and Data Science Project Report

Florida Gulf Coast University

Fort Myers, Florida

**Predicting and Modernizing Income Status Using Qualitative and Quantitative Features from U.S. Census Data**

Submitted by researchers

*Paul Hill* reachable at [pchill9748@eagle.fgcu.edu](mailto:pchill9748@eagle.fgcu.edu)

*Andrew Krupp* reachable at [aakrupp1968@eagle.fgcu.edu](mailto:aakrupp1968@eagle.fgcu.edu)

Under instruction of *Dr. Albero Condori, Ph.D.*

*April, 2023*

**Abstract**

Using 1994 census data, we will create and assess the accuracy of a created predictive model. We will also modernize the model, by inflation-adjusting the estimated income level being predicted and using a modern sample. The purpose of this research is to assess descriptive features’ ability to indicate income levels of a country’s population sample.The census data requires feature engineering to create a mathematically valuable form, and this was done using Boolean variables for each qualitative string found in the sample. The model will be created using a regression equation and QR decomposition of the data matrix. 80% of the data will be used to create the model, and the remaining 20% will be used to test the model. The original model will be predicting whether an individual exceeds an annual income of $50,000. Modern census data will be collected by the researchers through anonymous sampling, and the inflation adjusted income will be found using simple CPI-Inflation adjustment. Accuracy is found by comparing the amount of correctly predicted income levels to the total individuals in the treated data matrix. The model was significantly accurate on seen data, correctly assessing 84% of individuals. The model was much less accurate on unseen data, predicting only 57% of individuals correctly. The $50,000 in 1994 inflated to approximately $100,000 in 2023, and a sample of 9 individuals was used to create the modern data matrix. The model predicted every single individual to make more than $100,000 a year, grossly overestimating the sample and succeeding 22% of the time. The results are likely negative due to inevitable overfitting, as the model must be created using over 100 features. In preserving the data, no descriptive features should be omitted, forcing the model to face said overfitting. The overestimation of modern data reflects issues not in over fitting, but in the changing frequency of wealthy individuals and the features that would previously indicate one.

**Introduction**

The United States Department of Commerce, under the United States Census Bureau, administers a comprehensive Decennial Census on its population every 10 years. The Decennial Census collects granular, individual information on many economic, geographical, sociological, and anthropological descriptive features in attempt to profile the population. As expected, and despite extensive efforts, obtaining a response and representing every member of the population has proven impossible *(Bureau, 2021)*. To use the collected census data, mathematical and statistical analysis are required to create meaningful models that can broaden the reach of the information.

To create and test a few of these models, we used a relatively clean and clearly defined sample of 48,000 individuals provided by the University of California, Irvine Machine Learning Repository (referred to here as just the *UCI*). After engineering features into a mathematically relevant form, methods of data analysis and interpretation will be employed to create a model that can be used to predict the income level of individuals of the sample. Because the data provided by the UCI is aging, we will also employ simple financial mathematics to modernize the predictive model and render it relevant for modern use.

**Background**

Using census data collected by the United States Census Bureau, we will be attempting to predict the income status (that is, above or below $100,000/year) with reasonable accuracy. Using relatively clean census data provided by the UCI *(Becker, 1996)* we will create a model to predict the income level of U.S. citizens based on several unique descriptive features.

The data contains approximately 48,000 instances, each defined by 15 attributes. Many of the attributes are quantitative, including years of education, hours worked per week, and capital gains/losses. The more descriptive attributes are qualitative and will require feature engineering to properly fit a prediction engine.

This will be done through both binary Boolean variables and the creation of more precise categories derivative of the originals *(Pennsylvania State University)*. Some features included in the data can be excluded because they have no effect on the predictive outcome of one individual and are used to generalize the data set (i.e., final weight).

As the data set is dated to 1994, we will also attempt to “modernize” the prediction by using current currency and inflation values. When converted, the data set should be used to predict not only current income levels, but future ones. A more recent sample will be created and used to assess the accuracy of the prediction model, while modernizing the results. The data will be collected independently, by sampling individuals and applying their features to the created predictive model.

**Methodology**

After the initial creation of our data matrix, we will begin feature engineering our data. Of the 15 features, 5 are quantitative and in integer form. The remaining 9 will need to be converted to a series of Boolean variables, made up of 0 for false and 1 for true, and an additional column added for each qualitative option the subjects have per feature. One feature, “Final Weight”, will be omitted as it does not affect individuals in the sample. The new composition of features is shown below:

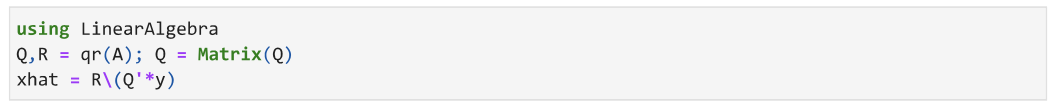
|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Columns Created (if applicable)** |
| Age | Integer | - |
| Work-class | Boolean | 8 |
| Education | Boolean | 16 |
| Education number | Integer | - |
| Marital Status | Boolean | 7 |
| Occupation | Boolean | 14 |
| Relationship | Boolean | 6 |
| Race | Boolean | 5 |
| Sex | Boolean | 2 |
| Capital Gains | Integer | - |
| Capital Losses | Integer | - |
| Hours Worked | Integer | - |
| Country of Origin | Boolean | 41 |
| Income (Above/Below $50,000) -1994 | Boolean | 1 |

After the feature engineering is complete, the resulting matrix is a 15 x 105 matrix. Missing entries will be replaced by “?” by the researchers.

To create both “seen” and “unseen” portions of our data, 80% of collected data will be used for the creation of our predictive model. The other 20% will be our treatment sample, used to assess the initial accuracy of our model. 80% of our matrix can be represented by 26,049 rows and will be used to construct our engine. 20% is represented by the remaining 6,512 rows and will be utilized after the model is created. After creating a working matrix out of the first 80%, we can begin the construction of our model.

Our model can be represented by

The regression coefficients, , can found using QR decomposition. The QR decomposition converts our matrix, A, into a product of an orthogonal matrix, Q, and an upper triangular matrix, R *(Yanovsky)*. In Julia, the QR decomposition can be found per the commands show below:



is found in line 3, or as a quotient of our created matrix R by the product of our transpose of our created matrix Q and the Boolean column of the predicted variable, “Income (above or below $50,000).”

Once the regression coefficients have been calculated, we will run a test on the created matrix using the 80% “seen” data. By comparing the actual income data collected by the census and the income data predicted by our model, we can assess the accuracy of our created model.

Next, we will test our model on the created matrix using the 20% “unseen” data. The methodology will be repeated, but we will predict information that was not used in the creation of our model. These predicted results will once again be compared to the data collected by the census to assess the accuracy of our model.

Once the initial accuracies of our model are calculated, we will need to alter the model to reflect modern dollar values of income. Using financial mathematics, we can find the modern (2023) value of our predicted income (1994). According to the Federal Reserve Bank of Minneapolis, the Consumer Price Index of All Consumers has increased from 148.2 in 1994 to 294.4 in 2022 (Federal Reserve Bank). The current value of $50,000 can be found using the equation below:

2023 Price = 1994 Price x (2023 CPI/1994 CPI)

2023 Price = $50,000 x (294.4/148.2)

2023 Price = **$99,325.25**

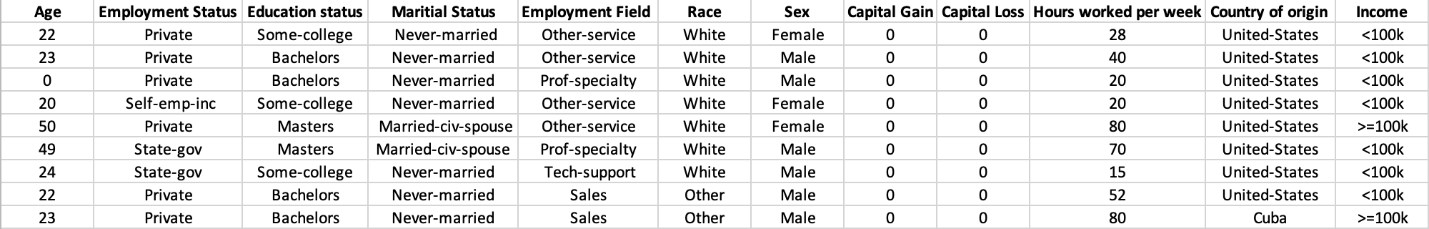
For simplicity in data collection, we will round this value to $100,000. Our updated model will therefore be attempting to predict whether a subject makes more than $100,000 today, using the original 15 collected features.

To get 2023 data, we will have to create our own sample. The researchers will create an anonymous survey that collects the information of 9 current subjects, described by the same 15 features as the original sample. A new matrix is created of the collected 2023 data, and the predictive model is tested a final time, this time on its current, inflation adjusted accuracy.

**Results**

On seen data, the predictive model was able to predict over 84% of the data correctly, accurately assessing the income level of 21,883 subjects correctly. On unseen data however, the model was only approximately 57% accurate, only able to accurately assess 3,709 subjects correctly in the unseen (20%) matrix.

The sample the researchers collected can be seen below:



In collected modern data, the model did not work. The engine predicted that every single subject makes more than $100,000, the equivalent of $50,000 in 1994. Only two of our collected nine subjects reach said minimum – a success rate of 22.2% and grossly overestimating the sample.

**Conclusions**

Recall the purpose of this research, to both create a predictive engine, and modernize said engine, of an individual’s annual income (above or below $50,000) based on descriptive features collected in the 1994 census. We created the engine using Boolean feature engineering, a regression equation, and QR decomposition of approximately 80% of the data. The model’s accuracy was tested on both the seen data matrix, and an unseen matrix populated with remaining 20% of the data. This process showed that the model was significantly accurate on seen data, 84%, and only marginally better than random chance, 57%, on unseen data. Based on these results, we conclude our model was harmed by over-fitting, and was created using too many independent variables. Due to the nature of the economic data, we cannot confidently omit any individual data or features of the sample. This is a major limitation of the findings, as the model has little choice but to face this over fitting.

We then collected a sample of anonymous volunteer information and tested the model’s accuracy on the inflation-adjusted income, $100,000/year. Of the 9 collected subjects, only 2 claimed to exceed the income minimum. Despite this, our adjusted model indicated that all 9 subjects reflect the features to make greater than $100,000, overestimating the sample. Though the overfitting of the original model influenced these results, we also face additional implications.

The first is that young professionals, those with a college education under 25 (as reflected in the collected sample), were much wealthier on average in 1994. This might reflect the depreciating value of a bachelor’s degree, as the model predicts those in possession of such to be much wealthier. We can also conclude it is increasingly difficult to establish wealth, as the gross over estimation reflects the model’s leniency to assume a higher income. In other words, reasonably wealthy individuals were much more common 30 years ago.

Future research should address over fitting, and work to create a more succinct model. Work should also be done to establish more specific dependent variables, predicting income to smaller intervals and not upon one minimum value. This can be done first through access to more highly descriptive census data, as recent data is increasingly difficult to find.

**Researcher Background and Acknowledgments**

*Researcher Paul Hill is an undergraduate Finance student with a graduate focus on statistics, offering insight into the context of the economic nature of the data and assumptions being made.*

*Researcher Andrew Krupp is an undergraduate Software Engineer, with research experience in the field of data science and predictive modeling and will provide these learned skills to the project.*

*Special thanks to Dr. Alberto Condori, Ph.D. for the oversight, advisement, and support given throughout the project.*

**References**

Becker, Barry. “Census Income Data Set.” Edited by Ronny Kohavi, *UCI Machine Learning Repository: Census Income Data Set*, 1 May 1996, <https://archive.ics.uci.edu/ml/datasets/Census+Income>

Bureau, US Census. “U.S. Census Bureau at a Glance.” *Census.gov*, 16 Dec. 2021, <https://www.census.gov/about/what/census-at-a-glance.html>

Federal Reserve Bank, Minneapolis. “Consumer Price Index, 1913.” *Federal Reserve Bank of Minneapolis*, <https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913->

Pennsylvania State University. “3.2.2 - Binomial Random Variables: Stat 500.” *PennState: Statistics Online Courses*, Pennsylvania State University, <https://online.stat.psu.edu/stat500/lesson/3/3.2/3.2.2>

Yanovsky, Igor. “QR Decomposition with Gram-Schmidt - UCLA Mathematics.” *UCLA Mathematics*, <https://www.math.ucla.edu/~yanovsky/Teaching/Math151B/handouts/GramSchmidt.pdf>