

**Northeastern University**

**Department of Electrical and Computer Engineering**

08

**Fall**

April 22th , 2019

**Predicting Income From Socioeconomic Data**

Roberto Rojas – Thomas Benda – Ryan Beiter

Professor: Joseph Robinson

EECE 2300: Computational Methods for Data Analytics

Final Project

**Introduction**

Income disparity in the United States continues to be a pressing issue. There are various socio-economic factors contributing to the breach between classes. Recognizing these and using them to predict annual income enables analysis on which elements cause the wealth discrepancy. The 1994 Census Bureau Database provides an extensive look into various descriptors of the population, with Boolean logic for each sample describing whether or not each person makes over or under $50,000 per year. In order to predict in which of these two categories a person lies, both K-Nearest-Neighbors (KNN) and Naïve Bayes Classifier (NBC) will be trained and used for testing. Each model will be deemed successful depending on whether it meets at least an 80% accuracy metric.

**Data Collection and Treatment**

The data was collected from the Census Bureau’s 1994 and 1995 *Current Population Surveys* by Ronny Kohavi and Barry Becker. Some controls were pre-determined to qualify as a valid sample for the data set. These include a minimum age of 16 years old and under 100 years old. There are around 30,000 samples within the data set.

The data had to be cleaned up given blank entries for some cells. Primarily, the Occupation and Work Class features had multiple gaps. This was handled by removing the entire sample. This brought the data from around 32,000 records to roughly 30,000, which should not impact the final accuracy of the model. The Final Weight feature, which describes how representative a sample was of the population, was removed as it was found to greatly hurt the performance of the model

The final treatment the data received was regarding the Capital Gain and Capital Loss features. If a person has a value for Capital Gain, then the Capital Loss value would be 0, and vice versa. This entails that both columns can be combined by subtracting the Capital Loss from the Capital Gain and thus obtaining positive and negative values to represent whether the sample gained or lost.

**Data Features**

The logic used to describe whether or not a sample earns over or under $50,000 per year is expressed through <=50K for those who earn less than or exactly $50,000, and the others are expressed through >50K. This feature, called Income, will be most useful when attempting to predict. The set contains 11 additional attributes (after the combination of Capital Gain and Capital Loss), either continuous and categorical. Each feature is described as follows:

* Age
  + Continuous metric, ages between 16 and 100 years old (exclusive)
* Working Class
  + Nominal, describes which sector of the working class each sample lies
* Final Weight
  + Continuous metric, value assigned to each sample depending on representativeness.
* Education Number
  + Continuous metric, assigns a value to education level from 1 to 16 (16 is the highest level)
* Marital Status
  + Nominal, describes the current marital status of each sample
* Occupation
  + Nominal, groups samples based on their current occupation
* Race
  + Nominal, describes the race of each sample
* Sex
  + Nominal, categorizes each sample in male or female
* Capital Gain or Loss
  + Continuous metric, profit or loss form sale of property
* Hours Worked per Week
  + Continuous metric, values range from 1 hour to 99 hours worked per week
* Country of Origin
  + Nominal, describes the country of origin for each sample

**Descriptive Statistics for Continuous Variables**

To gain further insight into the data set at hand, descriptive statistics are useful for measures of central tendency and spread. These statistics transform the raw data into valuable insight to determine where each feature or attribute tends towards, and whether there are outliers that must be treated. Table 1 depicts these measures in a presentable manner, and thus enables further analysis of the data prior to modeling KNN and NBC.

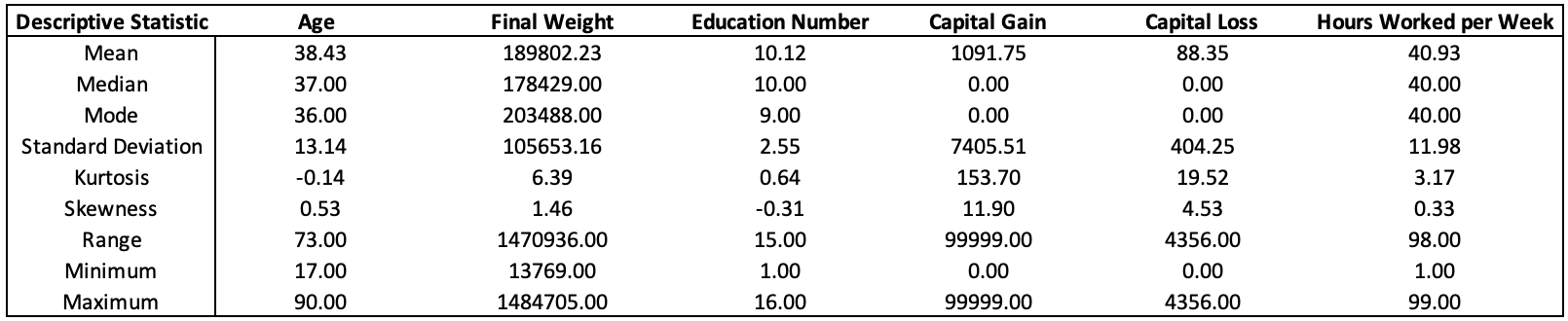
The skewness of the data set tends towards 0 for most attributes except Capital Gain and Capital Loss. A value of 0 for skewness would represent a perfectly symmetrical data set, following a normal distribution.

Table 1 - Descriptive Statistics for Continuous Variables

Other measures provide even more insight as to whether there must be further treatment to the data set. For example, for the hours worked per week the minimum and maximum are 1 hour and 99 hours, respectively. These are most likely not representative and could be considered outliers. After the use of KNN and NBC, it’s crucial to bear in mind that the data set may contain some outliers. If the accuracy of the model does not meet expectations, these outliers will be treated and the model will be tested again.

**Model Options**

1. **Naïve Bayes Classifier (NBC)**

A commonly used classification algorithm, NBC works off of theorems of probability to predict which class each sample of the data set pertains to. The only consideration which might affect its accuracy is the fact that NBC assumes conditional independence, which is not the case with the Income Census Dataset. Some of the predictors or attributes from the dataset may contain correlation between them, such as Age and Education or Hours Worked and Working Class, thus hindering the ability for NBC to perform appropriately, as it assumes they are unrelated.

1. **Kth-Nearest Neighbor (KNN)**

KNN is most commonly used for classification and regression systems. The model must determine which is the best value for K in order to obtain the Kth closest samples of data for training and testing. The output of the model should determine classes depending on the nearest neighbor for the data under consideration. The K value provides the boundaries each class will have and groups based on distance to the neighbors. KNN is the model most likely to present an accuracy score that meats the pre-determined objective of an 80% result for accuracy, and ideally a value for the F1 score higher than 60%.

**Results and Visualizations**

Data analysis is enabled by visual depictions of large data sets. Processing results in a visual manner will enable in depth examination of the data at hand, and how various factors attribute to the results obtained.

**A picture containing screenshot

Description automatically generated**

Figure - Determining the Value of K

As previously mentioned, KNN is the preferred method to use for the Income Census Data Set, as there is no assumption of independent variables. Figure 1 demonstrates the ideal value for K based on the optimal value for the performance metrics, which are Accuracy, Precision, Recall, F1 Score and NMI (normalized mutual information). The y axis indicates simply the percentage of these scores. Figure 1 demonstrates how the optimal value for K is equal to 23, as shown by the vertical yellow line. The corresponding values for the performance measures are demonstrated in Table 2.

|  |  |
| --- | --- |
| **Metric** | **Score** |
| Accuracy | 85.90% |
| Precision | 75.59% |
| Recall | 63.63% |
| F1 Score | 69.09% |
| NMI | 30.57% |

Table - Performance Measures

All of the measures depicted in Table 2 meet the requirements for a successful model, therefore meeting the null hypothesis that KNN would be a good fit for the data set.

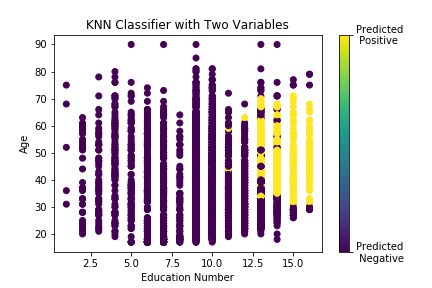


Figure - KNN with Two Variables for Positive and Negative Predictions

Education Number and Age are likely to be highly correlated with high income levels, so it was used to visualize KNN in two dimensions in Figures 2 through 4. The higher the education number, the more likely a person is to earn over $50,000 a year, same goes with age although that trend follows more of a bell curve. After the age of 70 the annual income should start to go down due to physical constraints or simply due to retirement. As can be seen in Figure 2, the yellow values represent those who earn more than $50,000 a year, with results being 85.9% accurate. This fits the theory that education number and age are related with regards to income. If that was not the case, the yellow and purple values would be more scattered all throughout the figure.

A screenshot of a cell phone

Description automatically generated

Figure - KNN with Two Variables for False Positives and False Negatives

Figure 3 depicts the same two variables, education number and age, and where the model might have missed the mark. As shown by the legend, the more light green/blue dots there are the better the model is at predicting positive and negative (whether or not a person makes over or under $50,000 a year) results. Yellow dots depict a Type II error, or a false negative, and purple dots signify a Type I error, or a false negative. A false positive refers to the model predicting that a person does earn more than $50,000 when they really don’t, and a false negative is simply the opposite. Figure 3 clearly demonstrates a large majority of light green/blue dots, which demonstrates that the model is in fact highly accurate.

A screenshot of a cell phone

Description automatically generated

Figure - Income by Age and Education

Figure 4 simply demonstrates what the actual true labels are for both education number and age. By comparison to Figure 2, Figure 4 shows a similar pattern of yellow and purple dots, which is highly encouraging about KNN. There are some yellow dots to the left side of the graph, but those are few exceptions to the higher-level hypothesis. Again, these visualizations simply further demonstrate the fact that an 85.9% accuracy rate does indeed meet the requirements of the model.

**Challenges**

The data had a truly raw format, with over 2,000 empty records which had to be completely cleaned. Some of the features present in the data were useless or repetitive for the purpose of predicting annual income, and determining how to delete these was a true obstacle. A small amount of the features present in the data had strange nomenclature, and figuring out what each feature represented was not a simple feat. For example, the variable “fnlwgt” was incomprehensible at first, and required some research to determine that it was a feature representative of each record.

A separate challenge was utilizing KNN to treat non-continuous variables. To solve it, dummy variables had to be brought into place to represent each record of ordinal or nominal variables, such as country of origin, gender, etc. The final hurdle became recognizing that Naïve Bayes Classifiers would not work particularly well due to the fact that they assume conditional independence, when some of the variables may be interlinked.

**Practical Applications of the Model**

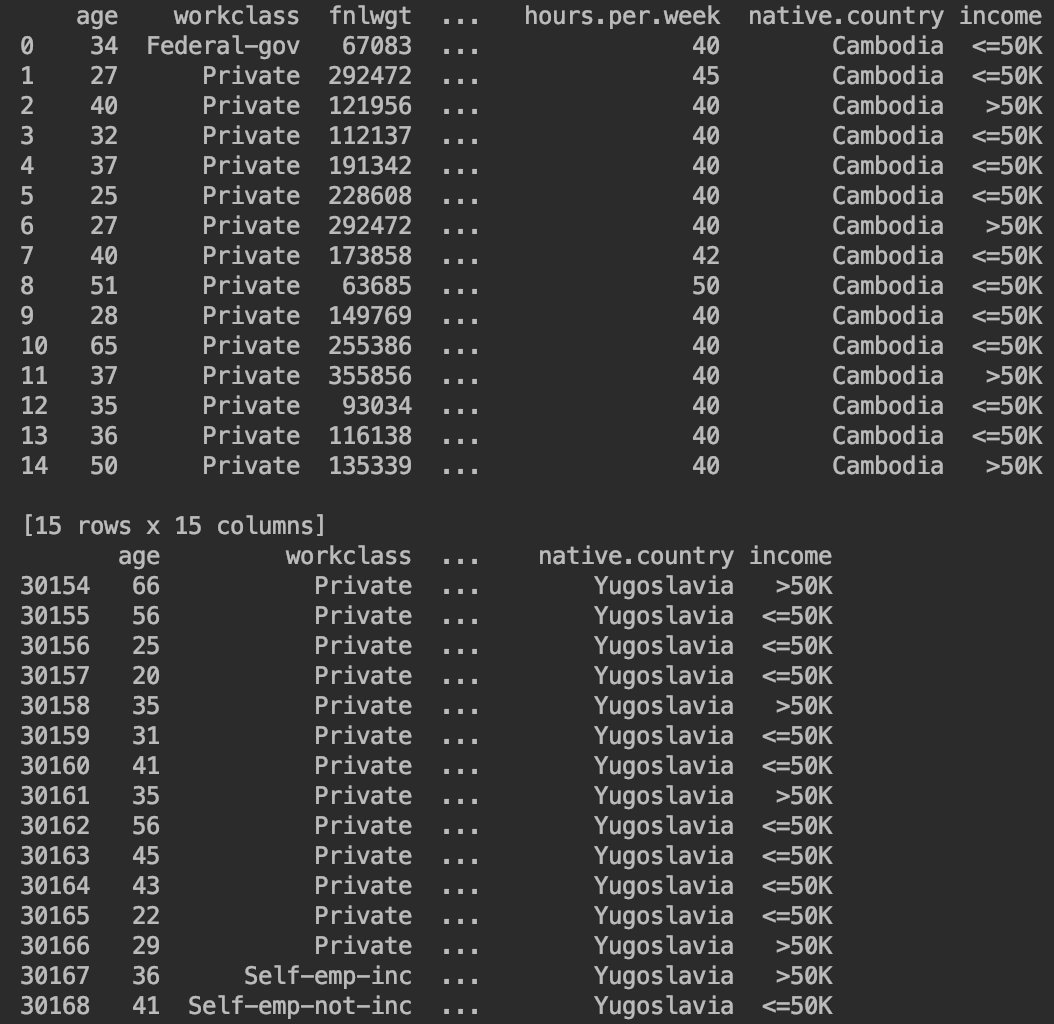
The model itself along with the logic from KNN can be applicable for any classification problems present in any industry. It does not make any assumptions regarding the raw data file, so it can be used for a wide range of topics. KNN can be used, for example, for determination of credit loans banks use based on historical data of successful returns. KNN can classify each applicant based on similarity of features given that financial details are standardized across the country.

Another application for KNN could be to determine whether or not a student will drop out of school, whether a person will vote republican or democrat, recommendation systems, amongst others. KNN can be used any time there are “similar” features between each subject under study, and thus the applications are vast.

**Conclusions**

The Income Census Data set contained various features representing the American population, as seen in the Data Dictionary present in Appendix B. The data had to be cleaned up and manipulated before application. Originally, a successful model was deemed to output a higher than 80% accuracy, and the final value obtained was 85.9%. Dummy variables were used in treating the data as nominal or ordinal values would not have been useful for KNN. A key takeaway is that KNN works best for this classification problem as it has no underlying assumptions regarding the data, and simply obtains similar features in order to classify whether or not a person will earn more or less than $50,000.

**Appendix A – Raw Data**

****

**Appendix B – Data Dictionary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature Name** | **Display Name** | **Format** | **Data Type** | **Example** | **Description** |
| Age | age | NNN | Integer | 34 | Age of the record |
| Work Class | workclass |  | String | Federal-gov | Work Class category |
| Final Weight | fnlwgt | NNNNNNN | Integer | 67083 | Representativeness of record |
| Education | education |  | String | Bachelors | Level of education |
| Education Numbered | education.num | NN | Integer | 13 | Number assigned to level of education |
| Marital Status | marital.status |  | String | Never-married | The marital status |
| Occupation | occupation |  | String | Exec-managerial | Current occupation |
| Relationship | relationship |  | String | Unmarried | Human based relationship |
| Race | race |  | String | Asian-Pac-Islander | Race of the record |
| Sex | sex |  | String | Male | Sex of the record |
| Capital Gain | capital.gain | NNNNNNN | Integer | 1471 | Assets gained |
| Capital Loss | capital.loss | NNNNNNN | Integer | 0 | Assets lost |
| Hours Worked per Week | hours.per.week | NN | Integer | 40 | How many hours the record worked per week |
| Native Country | native.country |  | String | Cambodia | Country of origin |
| Income | income |  | String | <=50K | Annual income |

**Appendix C – Python Script**

See included Python notebook