Income Patterns & Prediction Using 1990 Census Data

Kartikeya Puri
Aditi Mehta

Abstract: **In this report, we used a sample of the 1990 US Census Data, which includes anonymous data from around the country. We try and find patterns and connections between a person's income and other characteristics, and try and identify if incomes could be predicted using said features. To predict incomes, we use logistic regression and classifiers, and to identify patterns, we use *k*-means clustering. At the end of the report, we look at existing solutions to similar problems, and discuss what we learned and how our model could be improved.**

Exploratory Analysis

Dataset:

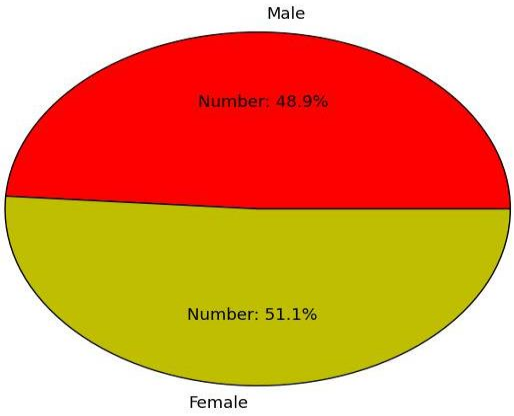
This data set was derived from the USCensus1990raw data set available at the UCI Machine Learning repository[1]. This is a multivariate dataset with categorical attributes. The original dataset has 2,458,285 instances and 125 attributes. There are no missing values in the dataset. For faster computation purposes, we created a new dataset with only the first 200,000 complete instances from the original dataset. This also served as our training set. We considered the next 50,000 complete instances for our test set. We considered the following 16 attributes for predictive tasks: age, citizenship, class of worker, English proficiency, number of children, immigration year, multilingual, mode of transport, marital status, looking for work, birthplace, workplace, race, income, sex and education. We only considered these attributes since we wanted to study their relation to the income of a person and create a model based on this. Also, most of the other attributes were just flags that indicated whether this data was available in each record. We would then use this model to predict the incomes of persons from our test set.

Dataset	Instances (number of people)	Attributes (characteristics)
Original	2,458,285	125
New (custom)	200,000	16

Patterns and data distribution:

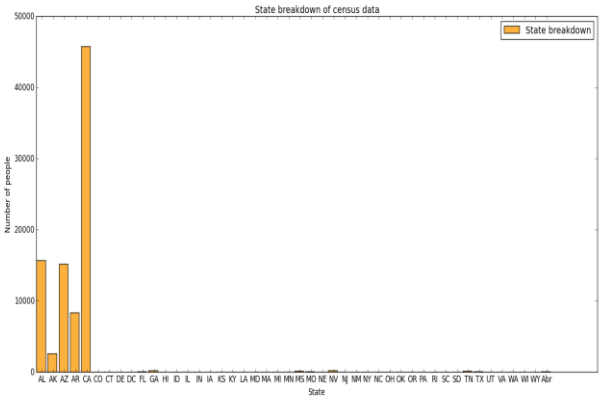
Initially, we looked at how the data was distributed. We noticed a few things. To create these graphs, we used the Python library matplotlib. Please zoom in for graph clarity:

a. The data was almost equally divided between the sexes. There is a slightly higher number of females.

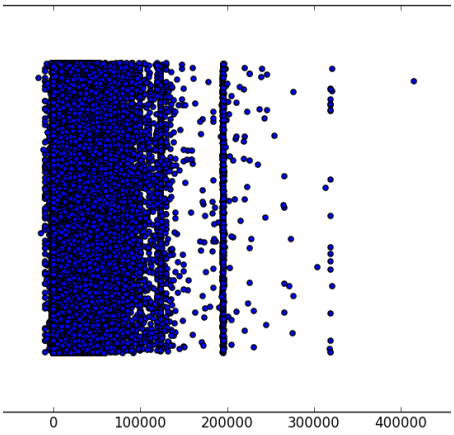


b. The data is not divided well among states people work in. There are some states, such as California, with a lot of representation (~46000), and a lot of states with less than 10 samples. Around half the population had no stable place of work, hence this graph only depicts the roughly 90,000 people who had a job.

[1] https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29

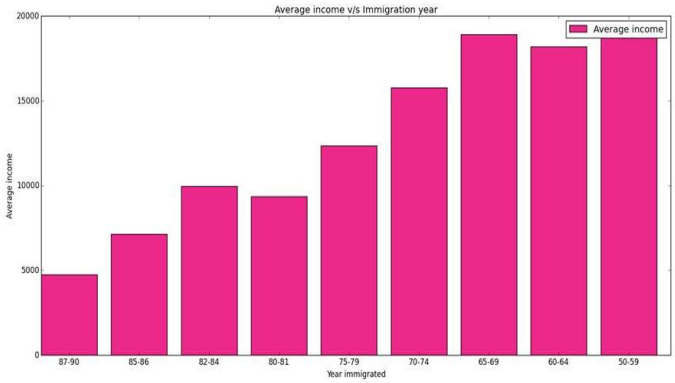


c. The income is quite clustered. The majority of the income lies between $0 and $120,000. There are no visible clusters except for one near $195,000. Data for incomes greater than $250,000 is sparse.



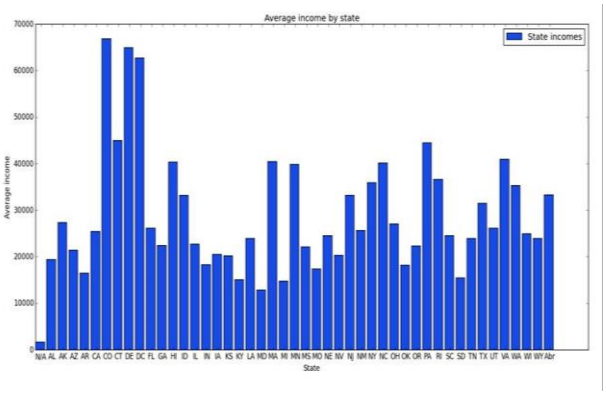
As part of our exploratory analysis, we studied how the income attribute changes against the other attributes and got some interesting results, which are explained in detail with the help of graphs and short explanations.

Average income vs Immigration year



Here we noticed that, as an immigrant, the longer one has been in the country, the higher one's income is.

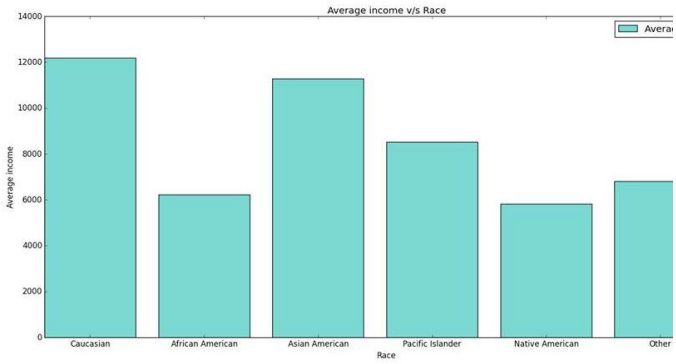
Average income by state



Notice here that Colorado has the highest average income (~ $70,000). This is not reflective of the true average income of Colorado in 1990 (which is ~ $20,000)[2]. This discrepancy occurs due to the fact that the original dataset that we obtained, while random, is only representative of 1% of the Public Use Microdata Samples (PUMS) person records drawn from the full 1990 census sample, and we are further using only 10% of that. Also, as mentioned previously, there are not a great amount of samples from every state. Hence, we do not consider state as a factor in income.

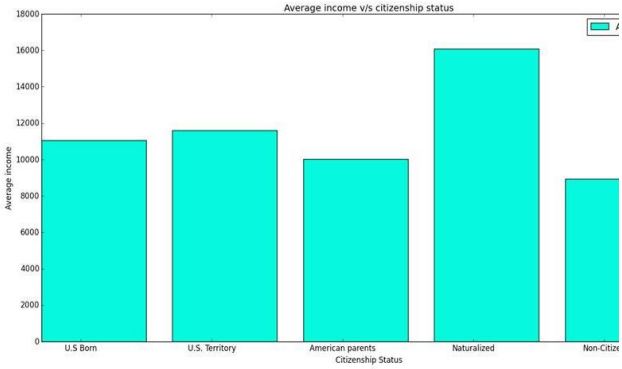
[2] http://www.statista.com/statistics/205218/per-capita-personal-income-in-colorado/

Average income vs race



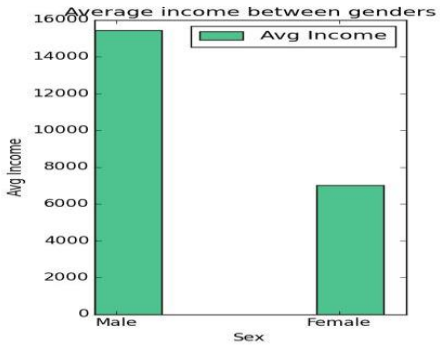
Here, we notice that the average income for Caucasian and Asian Americans is nearly twice that of African Americans and Native Americans. Hence, we do consider race as a characteristic while predicting incomes.

Average income vs citizenship status



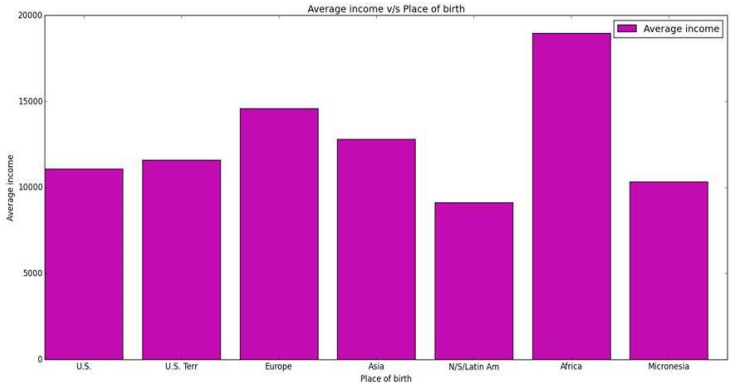
Surprisingly, naturalized citizens earned the most in our sample. This data was generally inconclusive, though, hence we ignore it while predicting incomes.

Average income based on gender



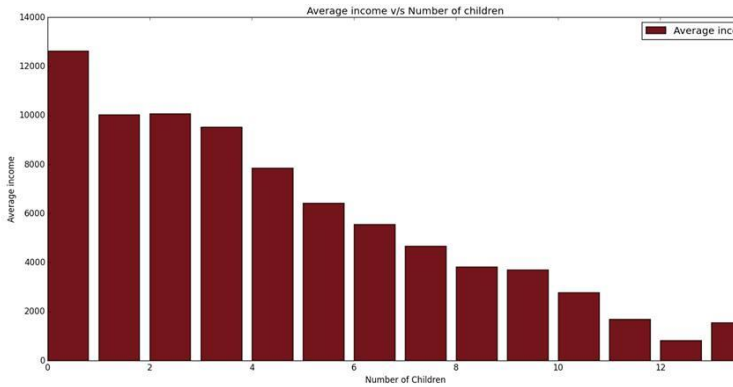
As expected, males earn nearly twice as much as females. Pattern noted for prediction.

Average income vs place of birth



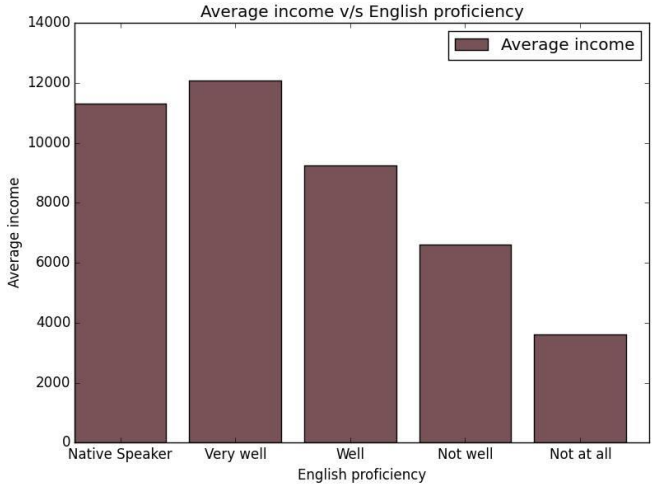
Data here was quite inconclusive, hence ignored.

Average income vs number of children



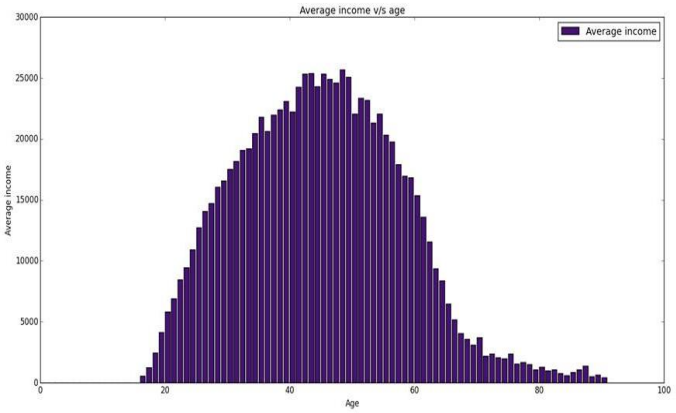
The result here is that as the number of children increases, the average income goes down. It seems like this relates to the fact that educated couples are more likely to have fewer children. Noted for prediction.

Average income vs English proficiency



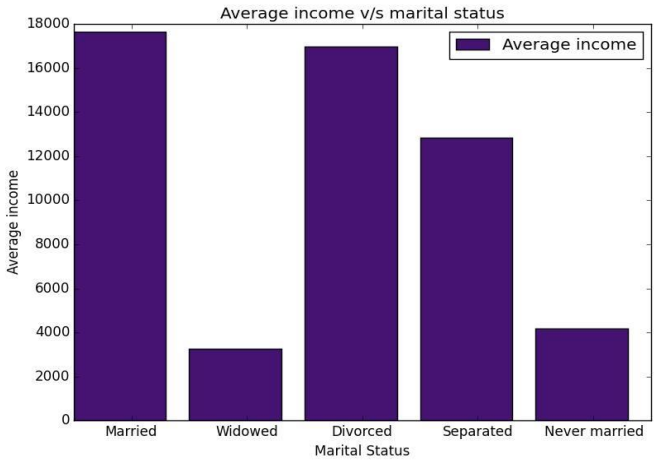
As expected, native/proficient speakers generally have higher incomes than those who aren't as proficient in the language. Noted for prediction.

Average income vs age



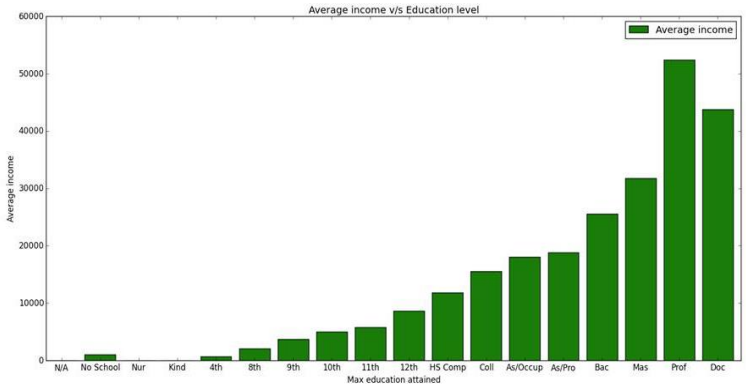
The peak income age is ~50 which seems reasonable since not many people retire before the age of 50. For the regressor that we train, we divide the age feature in the feature vector as ages 10-20, 20-30, 30-40, 40-50, 50-60, etc.

Average income based on marital status



For our regressor, we choose 2 categories here: Married/divorced/separated, and widowed/never married. This gives us better results.

Average income based on Education level



(Nur – nursery, Kind – Kindergarten, HS Comp – High school complete, Coll – College, As/Occup – Associate/occupational, As/Pro – Associate/Professional, Bac – Bachelor's degree, Mas – Master's degree, Prof – Professional Degree, Doc – Doctorate)

As expected, a rise in education levels sees a rise in incomes as well. We use this information to train our predictor and classifier.

Clustering

We tried to identify clusters in income data, however, the data was quite contiguous. We tried *k*-means clustering with the greedy algorithm and $k = 4$. Here are the clusters we get:

Centroid	Members in Cluster
1500	47146
21495	135924
50502	1543
149469	15388

Since these clusters are of varied sizes and are not quite distinct, we cannot really use these findings in our predictive analysis.

Predictive Tasks

Task 1

For the first predictive task, we trained a linear regressor to predict the income of persons from the test set and calculated the MSE to get an estimate of the accuracy of our regressor. In the feature vector, the features used were chosen from the attributes described above. The reason for choosing this model over other models was that it suits our

task better than clustering based models since those try and identify patterns rather than actually making predictions. The predictor tries to predict the income up to the closest 1000, 5000 and 10,000 mark. The baseline we tried to beat was simply predicting the average income from the training data. To calculate the MSE, we can use the test set incomes since those are actually provided.

Our model performs well on the training set but wasn't initially performing well on the test since we were overfitting. We used Occam's razor to battle overfitting and got far better results by ignoring a few hypotheses that contained a lot of assumptions. Finally, here are the features we use:

a. Sex
b. Number of children
c. Education attained
d. Marital Status
e. English proficiency
f. Age
g. Race
h. Immigration year

Initially, to use the attributes described above as features, we first had to create a new dataset from the given dataset to cater to our needs. This is because the original dataset consisted of 125 attributes, but we were only interested in a few of those. So we created a new list of dictionaries as our dataset.

For some features, we could not use direct numerical values, hence we used flags. For age, we created 9 flags, one for each 10 year increment in age. For race, the census data was quite detailed, hence we created groups (Asian/Caucasian/Pacific Islander, etc.) and used those as flags. For marital status, we used 2 flags to indicate whether a person was single or married.

Hence our final predictor was of the form:

$$X\theta = y$$

Here, θ is a 24 dimensional matrix.

Our final MSEs for this task using our predictor were:

Predicting to closest	MSE
1000	169.278
5000	23.5239
10000	1.06818

As evident, we did not get very low MSEs, since the data we got was very random and hence it was difficult to find patterns.

Task 2

For the second task, we trained a classifier using logistic regression to predict whether a particular person earned more than a certain amount. We used classification accuracy to determine how good our classifier was, using actual incomes from the test set. Our classifier predicts whether a person made more than $50,000, and we also used it to predict whether a person made more than the average income.

First, we created a new dataset that had the attributes we were considering, as a list of dictionaries. Then we found the average income, which came to $11,127. Hence we decided to partition the data around that figure (and $50,000) to find patterns among the data. Here are the attributes we discovered that made a difference:

a. Sex (more females below the average)
b. Education attained (associate degrees and above tend to earn more)
c. Marital Status (more married/divorced/separated people make above average)
d. Age (Our sample contained data from a lot of children, hence people below ~30 and greater than 80 made below average)
e. English proficiency (English proficiency of 0 or 1 – native or very well speaking tended to make more than average)

For the figure of $50,000, sex and age did not give us great results, but since we want to make a general model, we considered those factors. Using these attributes, we ranked each sample in the test data and classified it using the following equation:

$$y_i = \begin{cases} 1 \; if \; X_i.\theta > q \\ 0 \; if \; X_i.\theta \leq q \end{cases}$$

Where q is our partitioning figure

For our classifier, we used a logistic regression model instead of a Naïve Bayes or Support Vector Machine classifier because:

a. Naïve Bayes assumes features are independent, and we have discovered while exploring the data that there are relationships between a lot of the features
b. Support Vector Machines attempt to find a plane separating 2 sets of data, but a lot of our data is clustered, especially around the average income, and finding such a plane was proving to be nearly impossible

Here is the classification accuracy for our model:

Partitioning income	Classification accuracy
$11,127	0.8351
$50,000	0.6024

As evident, the model works rather well for $11,127 considering the data is scattered. For the $50,000 mark, it was harder to find patterns for samples labeled negatively. Our model was predicting too many positives.

Related Literature

We obtained our dataset from the UCI machine learning repository, who in turn listed their source as follows: The USCensus1990raw data set was obtained from the (U.S. Department of Commerce) Census Bureau website using the Data Extraction System. This system can be found at http://dataferrett.census.gov/. Donors include: Chris Meek, Bo Thiesson and David Heckerman from Microsoft.[3]

In trying to identify similar datasets that have been studied in the past, we found an individual short study in an online blog conducted on the 1994 US census data (also available from the UCI machine learning repository). Ilan Man, a member of the Data Science team at Squarespace, studied this data, which is similar to ours. He tried to predict whether an adult in 1994 will have an income greater or less than $50,000. For this purpose, he used logistic regression. Ilan didn't use all of the variables (attributes) available in the dataset as some of them weren't relevant to determining income level and also because using them all might lead to overfitting (which is also why we didn't consider all the attributes in our dataset). He only chose the following 4 variables: age, education level, number of hours worked per week and gender. He concluded (as part of one his findings) that married individuals tend to make more than their single counterparts (which is also true of our study).

He also concluded that men make more than women (both having similar qualifications) on average (which is true of our data, and unfortunately, even true today (in 2015) to an extent).

The conclusions from Ilan's work are similar to ours. This is probably an indicator that the factors affecting income between in the United States 1991 and 1994 were roughly the same.[4]

(Note: We have been using the terms variable and attribute interchangeably here).

The current methods used to study this type of data usually include using graphs and other visual aids to determine how certain factors affect income.

Results and Conclusions

In this project, we studied the variation of income of people in the United States in 1990 with factors such as their age, sex, place of birth, work place, education level, race, etc. and got some interesting results.

We did a predictive analysis of income by training a regressor that used features such as sex, age, and marital status. This tells us that these 8 are the most important features that affected a person's income back then. We tried other features, but they did not give us very conclusive results. For this task, we chose a linear regression based model because that was the best choice for the kind of data that we had. We tried to consider alternative methods such a mixture of a classifier and *k*-means, but due to a lack of clustering in the data, it did not give us good results. There were some feature representations that did not work well, such as those that contained attributes like place of work, place of birth, etc. This proves that given the sample of data, there are trends in income earning ability amongst different groups in society.

Even for our classifier, we observed these trends and used those to predict what groups of people tend to make more than the average income ($11,127). Hence we used similar features to train our classifier as well.

---

[3] Relevant paper is: Meek, Thiesson, and Heckerman (2001), "The Learning Curve Method Applied to Clustering", to appear in The Journal of Machine Learning Research.

Link: http://rexa.info/paper/0071aac48d5e8154c0f7197433ed828aae00d4b2

[4] Source: http://ilanthedataman.com/understanding-the-data-game/2013/05/02/us-census-analyzed