Income Level Predictions based on National Longitudinal Survey of Youth

ORIE5741

Cornell University

Operations Research and Information Engineering

Authors:

Abigail Ries

Gretchen Siewert

Bohan Zhang

December 5, 2021

Introduction

In the United States, a great emphasis is placed on how much income a person makes. It can define their status, their quality of life, and the opportunities available to them. Most people would like to know how much they will end up making in the future, and what steps they can take to increase their earning potential. For our project, we took a look at predicting what people's future income would be based on myriad factors from throughout their life. We took a look at aspects of their childhood, education, family, and daily life to see if we could determine how much money they would make when they were middle aged.

The ability to correctly predict future income can have a wide range of impacts. Educators could use it to analyze which students might need extra attention in order to be successful later in life. Parents could see how the choices they make for their children might affect their earning potential down the road. Even young adults could use it to orient their habits in potentially beneficial ways. This information can also be used to determine whether protected attributes such as race and sex have an impact on income, which could lead to future analysis about what is causing that in our society.

Data

The data we used for our model comes from the National Longitudinal Survey of Youth (NLSY79). This dataset was collected from over 12,000 participants in the United States, all of whom were between the ages of 14 and 22 when the study started back in 1979. While many participants have dropped from the study in the past 40 years, there are still over 10,000 participants who answer questions every year about topics such as their education, family, career, and daily activities. There are over 70,000 different features in the data set, each one representing one question that was asked to participants at one point in time. Some of these questions (such as participant race and sex) are answered for every participant. Other, more specific questions, are missing the majority of the answers, which were represented in the dataset with various negative numbers as their responses, depending on the reason for not answering.

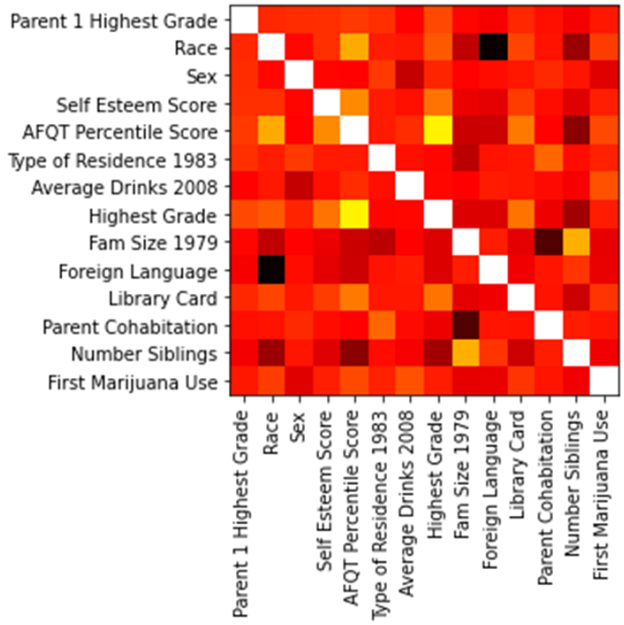
With over 4,000 different features about income and assets alone, we needed to find the one that would be the best indication of future income. We chose a question from 2018 which asked participants "(Not counting any money you received from your military service...)

In the past year, how much did you receive from wages, salary, commissions, or tips from all (other) jobs, before deductions for taxes or anything else?" This data, which was the most recent year a comprehensive income question was asked, was what we used as our *y* output space.

We decided to narrow down the thousands of questions asked to a handful of features that we would pursue for our model. We specifically looked for questions that the majority of participants had answered, which also seemed like they could have an interesting correlation with income and spanned a wide range of topics. Below are descriptions of the 14 features we ultimately pulled from the NLSY79 data:

Categorical	**Quantitative**
Race	Highest Level of Education of Participant
Sex	Highest Level of Education of Participant's Parent
Type of Residence in 1983	Self Esteem Test Score
Do Parents Live Together	AFQT (achievement test) Score
Is a Foreign Language Spoken at Home	Family Size in 1979
Does the Participant Have a Library Card	Number of Siblings
	Average Drinks per Day in 2008
	Age of First Marijuana Use

After selecting our features, we did a correlation analysis to see whether or not some of the features were very strongly correlated with each other. We found that both Foreign Language vs Race and Family Size vs Parent Cohabitation were strongly correlated. We used this information when creating our final models.



Models

The first step taken into engineering the models is to further clean up features. The raw data comes with 14 features with different strengths of correlation with income 2018. As indicated by the heatmap in the previous section, we further narrowed down our feature space to using only 7 features that have the stronger correlations to income 2018 than other 7 features. Those selected features include categorical ones such as race and sex, and quantitative ones such as AFQT scores and self-esteem scores. By reducing the number of features to be used, our goal was to reduce the chance of overfitting; this is achieved by reducing the potential complexity of the models.

Feature transformation is the intuitive step following feature selection. In our study, the features fall into two categories: quantitative ones that are already in integers which require no action of feature transformation, and categorical features that are already encoded into integers (e.g. races are encoded as integers -5 to 5 in the raw dataset). Some transformations and imputing are required for hard-encoded negative data, because negative numbers such as -5 (meaning refused to be answered) and -1 (meaning not asked) are not expected to affect the result in different ways; it would be more intuitive to group all negative numbers into one typical value and call it "not applicable." The next paragraph will address this transformation and imputing.

The next step taken was to impute missing data. The dataset is messy and composed of various missing or N/A cells; this is partly because some of the features are harder to collect (e.g. marijuana use - which might be refused to be answered, or simply not applicable) than others (e.g. race - which is relatively easier to collect from the participants). During our experiment, if we drop all rows which contain at least one missing or N/A cell (these fields are indicated by negative encodings, e.g. -1 means that the query was not asked), eventually we will only have around 60 rows. Given this observation, we took two approaches: first we abandoned all rows where income 2018 is missing; since income 2018 is our y variable, imputing this will undermine the accuracy of our prediction. We lost a small fraction of data points after this operation but the remaining data points are abundant enough to train a functioning model. On the other hand, for missing cells in features, we impute them with the average of existing values from the column; for example, if participant John Doe is missing an AFQT score, we impute his score with the average score of all participants who have provided a valid AFQT score. The post-cleaning dataset has 6571 rows.

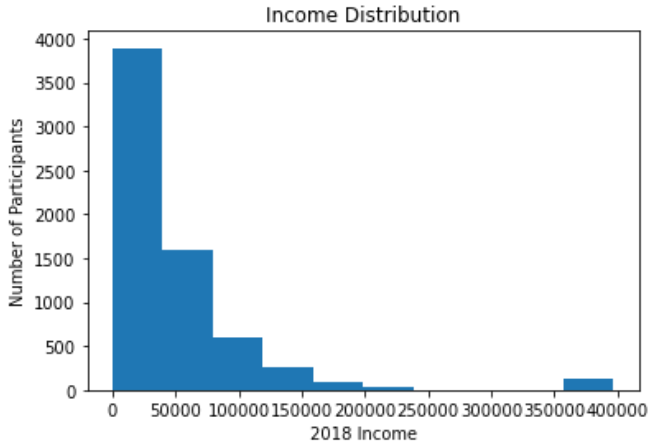
We split the dataset into training and testing subsets according to the 80% - 20% rule from class. Validation sets are not required because our project was aiming to measure the accuracy of income prediction given the study, as opposed to producing actual predictions on unlabeled data. Our model is trained on the training dataset, and predicts using the testing dataset, while being scored by comparing its predictions against the actual incomes. The training dataset has 5256 rows, whereas the testing dataset has 1315 rows.

To achieve more precise results, we experimented with different models including both regression models and classification models, with the first attempting to predict actual numbers as income, and the latter attempting to classify participants into different income groups. Prior to performing the classifications, we made decisions on dividing participants into various income groups. The first and the most intuitive approach is to divide participants into low and high income groups; participants in the low income group have less-than-average income, whereas the high income group have higher-than-average income. The average income is around $44,000, and this division successfully divides participants almost equally (see next section for the distribution histogram). The second approach is to use 5 groups, which adds to model complexity and we expected to use different classifiers on this split and compare the results. Eventually we collected various results on 6 different models: a linear regression model (MSE L2 norm), a linearSVM classifier on two categories, a logisticRegression classifier on two categories, a randomForest classifier on five categories, and an autoML regression model. The linear regression model is a continuation of our preliminary assessment from midterm, but did not achieve an ideal result; as a follow up, we removed significant outliers to assess its performance again (addressed in next section), and introduced above categorical models, and autoML package, to compare performances in terms of accuracy.

The next section will compare results collected from these models and draw conclusions. To measure the performance of regression models, we used the average absolute errors from true incomes. To measure the performance of classifiers, we used the accuracy score (defined as the number of true predictions over the number of all predictions).

Modeling Results

One of the issues we ran into due to the nature of income distribution in the U.S. was outliers. The extent to which this distribution is right-skewed can be seen below.



When we initially ran our models including the entire distribution of income (which ranged from $0 to ~$400,000, with a mean of ~$44,000 and median of $30,000), we had a very large absolute error in our regression predictions, around $30,000. While this error is small when looking at predicting very large incomes in the upper range of the data, it is too large of error to be able to confidently predict incomes in the lower range, as small changes in income make a noticeable difference to someone in this range.

In order to create a model that would have an acceptable error for lower incomes, we ran our regression models on the dataset excluding outliers, which we considered any incomes over $150,000 (262 participants). The error after removing all such outliers becomes $24867.53. This reduced our average error by roughly 20%, and we have included both with and without outliers in our model results summary below.

For regression models, we tried a Linear regression model, as well as AutoML. The AutoML gave us the result of light gradient boosting, and did not include much of the feature engineering described above - we fed it the 6571 rows with income data, split into the 80/20 train/test sets to see how it would do. As shown in the table below, linear regression without outliers had the lowest average absolute error of the regression models we used.

Model	Error
Linear Regression	$30,000
Linear Regression (without outliers)	$25,000
AutoML (light gradient boosting)	$42,000
AutoML (light gradient boosting, without outliers)	$29,000

In an attempt to create a more accurate model, we also tried classification. With the mean of Income being $45,000, we split the data into two classification groups (above and below average income). Using these two categories we attempted a LinearSVM model, Logistic Regression model, and a Random Forest model. We also tried a Random Forest model with five different Income categories, however this did significantly decrease the accuracy, so we feel more confident with our two category results.

Model	Accuracy Score
LinearSVM (2 categories)	0.653
Logistic Regression (2 categories)	0.722
Random Forest (2 categories)	0.694
Random Forest (5 categories)	0.418

Due to the high errors found for the regression models, we have chosen Logistic Regression with two categories as the model we have the most confidence in.

Fairness/Weapons of Math Destruction

While our model does use data that may be deemed unfair such as sex and race, the goal of the model was not a concrete yes or no question. This model was not making a credit decision, or a loan decision. It is merely an attempt to predict what a child's future might look like. This model determines whether individuals might need extra assistance or resources. The use of this model should only help participants, not hurt them (such as a no on a loan decision). This is why we have allowed our model to use features such as sex and race.

We do not believe that our model will be a weapon of math destruction. Income is an easily measurable outcome. Our intention is that our model is not used in decisions that might have negative consequences. The model should hopefully not create a feedback loop because the information gained should be able to help give everyone equal opportunity.

Conclusion

We are fairly confident in our results using Logistic Regression with a 72% accuracy. While we do not recommend making life changing decisions for individuals based on these results, we believe that this model can identify groups of individuals that are more likely to be above or below average income.

Future work for this project includes expanding upon the work we did with the five categories instead of two. While we did not have much success with the five categories, if we could improve upon models using this, the results may be more useful. We could also explore using the features we ultimately ended up ignoring such as family size and type of residence. In addition we could also explore even more features from the survey that we didn't start with.

When working with missing data, we used strategies such as imputing averages. Future work could include using unsupervised learning instead to deal with missing data. In our project we identified light gradient boosting as a good option through AutoML. We could improve on this by exploring light gradient boosting itself more in depth.