Prediction of Individual Level Income: A Machine Learning Approach

BY Michael Matkowski

ADVISOR • Edinaldo Tebaldi
EDITORIAL REVIEWER • Son Nguyen

Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
APRIL 2021

Table of Contents

ABSTRACT

The use of machine learning models to improve prediction problems and handle increasingly large datasets is a rising trend in economics. Prediction plays a particularly important role in applied economics because it provides critical insights to assess market outcomes. This study builds on previous literature to showcase the relative power of these modelling methodologies in economics through the prediction of income. This research utilizes data from the Current Population Survey from 2017 – 2020, containing 467,811 observations and 264 variables. 2017-2018 data served as training data for the models and 2019-2020 served as data for the two testing sets. The results show that machine learning models performed better than traditional prediction approaches in predicting individual total income. The high performance of the machine learning models supports that these methodologies should be utilized alongside more traditional techniques to assist in economic research focusing on prediction. With further development, these models could be used with great effect to assist in both the public and private sectors.

INTRODUCTION

The use of machine learning methodologies is still in its infancy in economics. Economics, like many other disciplines, traditionally relies heavily upon traditional statistical techniques including parametric and non-parametric methods to conduct empirical research. For instance, regression analysis is an important and powerful tool, especially for determining relationships between variables and modelling trends in the data. Regression is popular and used extensively throughout a variety of fields from economics and business to medical research (Schneider, Hommel, and Blettner, 2010; Ramcharan, 2006). OLS regression does not require excessive computing and provides robust results that can be easily interpreted. More methodology has been built on regression throughout the years, including dimension reduction techniques. However, traditional regression relies on strong statistical assumptions that cannot always be met when working with real world data, which is often messy and may not satisfy the necessary assumptions.

As a result of the limitations with the traditional methods and the increase in the size of datasets, there has been an increase in recent years in the use of machine learning methodologies aimed at improving the analysis of data, especially in the area of prediction. Leveraging machine learning methodologies, economists can potentially improve the prediction ability of their models. In achieving a better prediction, new and improved applications of data-driven decisions can be made.

Machine learning models, however, are black box algorithms, where data is fed into the algorithm with the goal of predicting a *target* variable (dependent variable) based on *features* (independent variables or covariates) (Brownlee 2016). These methods were designed to deal with the large dimensionality that exists with the increasing amount of data that is collected and can work with new types of data that previous statistical methods were unable to work with in the past (e.g. text mining). The most important distinction is that machine learning methodologies strive to create models that predict the final output as best as possible. No single method will work the best for all data, but rather each model needs to be tuned and re-run with a range of values for parameters of the model. The resulting best model can be different for each set of data. There are some methods that are more consistent in high performance, but it requires testing to find the best possible model. In this research, for example, the prediction of income given the data available for each individual is the main goal of the models.

Machine learning models are built on testing data and verified with training data. Often, the testing and training data come from the same data source and are split randomly into the two categories, with the amount of observations going into each set being determined by the researcher. Common splits include 50-50, 70-30, and 80-20 for the ratio between testing and training data. The overarching goal is to have the best prediction on the testing, or out-of-sample, data based on the models that were created using the training data. This is done because the models can easily be overfit to the training data, and would be useless if applied to other observations or for future predictions – which is the goal of the models.

While machine learning methods often achieve a higher level of prediction than the traditional methods like regression, there are some critical limitations with these models. The biggest limitation is that these algorithmic models are a black box, where the inputs are entered into the model and an output is produced. One can understand the ideas and concepts of the model, but cannot see the actual process that the data goes through to get the prediction. The combination of this black box along with the ideas behind each model type, results in significant decreases in the interpretability of these models. These methods can allow for better predictions to be found, but do not allow for the same degree of interpretation and relationships between variables that can be found with the more traditional methods.

This study builds off previous literature to showcase the prediction power of these modelling methodologies in the sphere of economics through the prediction of income. Prediction of individual income has relevant applications in the public and private sectors.

In the public sector, the proper reporting of individual income is extremely important in calculating taxes. However, tax fraud in all forms has always been a massive problem but has continued to increase throughout the last two decades. The latest report from the Internal Revenue Service (IRS) found that the average gross tax gap was estimated at $441 billion for

2011-2013 and that the percent of taxes paid each year, even after enforcement, is at about 85.8% of expected revenues (IRS 2019). This percentage has remained consistent with previous years' estimates of 83.7% in 2001 and 82.3% in 2006. One of the causes of this gap is the false reporting of incomes by individuals. An income predicting model could assist in flagging potential cases of fraud to assist in government efforts. Furthermore, an income prediction model could be leveraged to help identify individuals that may need further assistance. This can allow public policy initiatives to better target individuals in need of help.

Income prediction is also integrally important for a variety of areas in the private and nonprofit sectors. One critical area this affects is marketing, where income segmentation of the population is an extremely important tool. Businesses may make different variations of their items designated for certain subgroups of the population, and these subgroups often include the income of individuals. Similar to use for the government, the models could help identify fraud in the private sector in areas like false reporting of income on credit card applications, loans, and other forms. It can further assist nonprofits in identifying potential individuals that have disposable income that may be able to donate to their cause. Inversely, an income-predicting model could identify those individuals who are of a lower income that may need the most assistance, who some nonprofits strive to identify and assist. Companies and nonprofits often have basic information about their customers through surveys, rewards programs, and other means. The ability to predict the income of individuals from this information has far-reaching impacts for every industry. This paper utilizes a combination of traditional statistical techniques paired with machine learning methodologies to create effective models to predict income to better assist both the public and private sectors.

As a result of the importance of the prediction of income for multiple uses across various sectors, this paper utilizes machine learning methodologies and traditional least squares regression to predict the factor at the individual level. The data utilized is from the Current Population Survey from 2017 – 2020, containing 467,811 observations and 264 variables. 2017-2018 data served as training data for the models and 2019-2020 served as two testing sets. This research leverages a combination of the more traditional methods along with machine learning methodologies to assist in understanding of individual income and create effective models to predict income at the individual level.

LITERATURE REVIEW

The application of machine learning methods is a currently emerging movement in economics, providing economists with new opportunities for research and analysis throughout the field and allowing for improved prediction power over traditionally used models. Hall (2018), Bajari *et al.* (2015), and Saltzman and Yung (2018) showcase the current uses of machine learning in economic research to further empirical investigations in diverse areas within the field of economics. Hall (2018) compares the machine learning model of elastic net to that of a commonly used and widely accepted model (created by the Economic Indicators and Survey of Professional Forecasters), an autoregressive model, and a random walk model (using last value as estimate for future values). The elastic net model, which is a machine learning methodology that assists in variable selection, used data from the FRED-MD database to make a prediction on unemployment. It was found that this model gave better predictions of the monthly U.S. unemployment rate as compared to the other model types. It demonstrates the ability of machine learning methods to be used in economic research to potentially improve prediction capabilities.

Bajari *et al.* (2015) and Saltzman and Yung (2018) differ from Hall (2018) in that the authors use machine learning methodology to advance research and understanding in areas where there is not already a known or widely accepted model for the data. Bajari *et al.* (2015) uses grocery store data to create a model to predict demand for salty snacks (quantity sold per week) for one chain using six years of data. The models created included linear, stepwise, forward stepwise, LASSO (variable selection), random forest, SVM, bagging, logit, and ensemble (weighted combination of the others).[1] The ensemble model gave the best out-of-sample and validation root mean squared error values, which is one way to measure the success of the predictions of the models. The random forest model lagged behind as the best single model. The ability of machine learning methodology to combine the other model types, some traditional and some machine learning

[1] Please refer to methodology for further explanation of terms and models.

based, through the ensemble model along with the use of the large and highly dimensional dataset illustrate an application of machine learning in economics.

Saltzman and Yung (2018) also use a different combination of machine learning and traditional statistical techniques to conduct their research in identifying uncertainty. The study used text from the Federal Reserve Beige Books from 1970-2018 with text mining, classification of text into usable data, to denote positive and negative connotations for uncertainty and to divide comments into different subgroups. Principal Component Analysis, which assists in reducing the large dimensionality of the output, was then performed on the results to create two clusters. One cluster was focused on politics and government and the other on business and economics. It found that, at the time, the increases in uncertainty were resulting from politics, as opposed to the economy. These two articles work in tandem to demonstrate the applicability of combining traditional techniques and machine learning in different combinations to further advance economic research in new areas and provide better predictions that can assist economists in policy recommendations and overall analysis.

Newly emerging research conducted in the subfield of financial economics has adopted these machine learning methodologies to assist with research. Research conducted by Hanh and Viviani (2017) and Barzarbash (2019) illustrate the applicability of this methodology in the subfield, which was applied to predict outcomes at the firm level for financial institutions. Hanh and Viviani (2017) applied machine learning methods to the banking industry specifically, using methods such as neural networks and support vector machines to predict bank failure. Meanwhile, the newer study of Barzarbash (2019) applied similar machine learning methods such as support vector machines, neural networks, random forests, and gradient boosting to effectively predict credit ratings of firms. In both of these cases, machine learning methodologies were effective in predicting credit ratings or bank failure. As such, these papers further advance the applicability of machine learning models to predict risk and outcomes at the firm level, which can be used to assist in policy decisions for economists.

In the same way that financial economics has applied machine learning concepts to predicting risk and outcomes at the firm level, other economists have applied machine learning methods to help with the prediction of income and labor market outcomes at the individual level. Research conducted by Mareckova and Pohlmeir (2017), Matz, *et al.* (2019), and Lazar (2004) all demonstrate the applications of machine learning in economics in the area of income prediction and analysis through various methods. Mareckova and Pohlmeir (2017) utilize noncognitive skills data to assist in addition to basic demographic data to analyze the impact that these skills, built in childhood, have on long term labor outcomes. Specifically, the study focuses on the outcomes regarding employment status and labor force participation. In order to bring noncognitive skills into the model to test the impact, three main variable selection types were used – human selection, PCA selection (traditional statistical selection), and LASSO (machine learning selection). It was found that the R-squared values for each age group performed better with LASSO selection as opposed to the other methods. In this case it provides evidence that machine learning applications, especially in variable selection, can be extremely useful in research.

Matz, *et al.* (2019) also used machine learning methods with new data, but focused on social media data as opposed to noncognitive skills data to predict income. The study used Facebook data in tandem with basic demographic information to assist in predicting income of individuals. Facebook data was extracted using machine learning methods like text mining on status updates, while demographic data contained traditional variables measuring aspects including zip code, age, gender, ethnicity, educational attainment, and more. It was found that the best model used the likes, status updates, and socio-demographic data to get a correlation between the prediction and actual income values of r = .47, while only socio-economic data gave a correlation of r=.42. In both of these cases, the correlation between the predicted and actual values remains weak, with only a limited increase considering the large amount of data from Facebook. In both of these cases, machine learning methods were used to manipulate data to supplement basic demographic data in order to attempt to reach an improved prediction. The results found that the noncognitive skills had a very significant impact on prediction when using LASSO selection, yet the Facebook data, while assisting in prediction, only had a modest impact on the results.

An earlier paper by Lazar (2004) uses only demographic data from the Current Population Survey (CPS) in conjunction with machine learning methodology to demonstrate the potential ability of these variables to effectively predict income. The study combines principal component analysis (PCA) with support vector machines (SVM) to demonstrate the use of statistical methods in tandem with machine learning methods to assist in predictions. The dataset contained 48,842 observations from the 1994 CPS Dataset that contains variables on a plethora of different demographic data. The results found that the use of PCA with modelling yielded accuracy values as high as 84%, and reduced computational time by 60%. Through the combination of the traditional statistical and machine learning models, computational time on the model was shorter and resulted in high accuracy. In addition to these important ideas in methodology, the paper provides several other key findings. When comparing Lazar (2004) to its later counterparts such as Mareckova and Pohlmeir (2017) and Matz, *et al.* (2019), it demonstrates that high prediction output can be found even without the use of outside data, as in those two studies. Outside data beyond basic demographic data may still be useful to consider, but the paper offers evidence that methodology is a key factor in prediction ability. Selection of methodology will be extremely important in prediction and working with limited computational power.

Lazar (2004), in tandem with Chase, Kozma, and Matkowski (2019), demonstrate the effectiveness of CPS data with machine learning methodology to achieve significant results in prediction. Lazar (2004) demonstrated the potential ability to predict income on old 1994 data, while Chase, Kozma, and Matkowski (2019) demonstrate a more recent study to predict labor force participation using recent 2018-2019 data. In both cases, results demonstrated the ability of machine learning methodology to be effective in assisting prediction in key labor market outcomes. Conducting research on a newer release of the CPS data or a similar set such as the American Community Survey data, should allow for the application of new methodology and models to predict individual level income.

Another commonality between much of the recent literature is the utilization of new types of data to assist in analysis, which are only easily accessible using newer data techniques, such as machine learning methodology. Studies such as Bajari *et al.* (2015), Saltzman and Yung (2018), and Matz, *et al.* (2019) all utilize data that could not have been effectively analyzed without using these newer methodologies. In these cases, the data was simply too multidimensional or in a form like text that could not be analyzed with traditional methods. Einav and Levin (2014) highlighted the potential applicability of these new datasets before these later studies were completed. It highlights that newer data sources are accessible, especially public administrative data and private sector data. These new types of data often include increased dimensionality and larger datasets to utilize, along with not suffering from issues with missing values. An example that Einav and Levin (2014) uses to demonstrate the applicability of these new datasets is the Billion Prices Project (BPP). This project uses the prices and product attributes from online retailers to create a daily price index, which matches fairly well with the Consumer Price Index (CPI) provided by the BLS. While the CPI is calculated monthly and has a lag of several weeks, the BPP is calculated daily with a lag of several hours. The use of new data sources to conduct both new research and to improve on existing models is only possible with the new machine learning and other methodologies to work with new forms and larger datasets.

In addition to the utilization of new data sources, another major theme in methodology is the careful selection of models chosen to test. This careful selection of methodology was one of the strongest themes permeating the literature. Several papers, (Varian, 2014; Mullainathan and Speiss, 2017; Athey and Imbens, 2019), focused only on the technical description of the methodology, which strived to clarify methodology for the reader and suggest potential applications of the concepts for research, as opposed to provide a case study or unique research using the methods. Apparent in all of these articles, to varying extents, was the careful selection of models. All the other papers used a blend of traditional and newer machine learning methodologies in their research, either comparing them or blending them together in analysis.

Brieman (2001) and Cornell-Farrow and Garrard (2019) both highlight the importance of picking the most appropriate methodology for the research question and dataset being used. Breiman, having pioneered many of the early machine learning techniques, did not advocate purely for its use over traditionally proven methods. Rather, it served to complement existing methods to assist with analysis. After demonstrating the applicability of all different types of models in past research, Brieman makes the important conclusion that the key focus of a researcher is to find a good solution to the research problem and to use any model that gives a good solution, either algorithmic (machine learning) or data (traditional). This piece of literature, written before all other literature referenced, serves an integral purpose to reinforce the idea that methodology selection should be based on achieving the best solution, instead of the blind application of new methodology. The vast majority of the papers do indeed use traditional statistical models in addition to the newer algorithmic-based (machine learning) models, but almost universally found that machine learning techniques provided improved prediction power.

While this was largely the case, Cornell-Farrow and Garrard (2019) serves as an example of traditional methodologies providing better results than newer machine learning models. In this case, a traditional logistic model gave the best prediction for student success on the Australian National Assessment Program, which is a standardized test. Compared to machine learning methods such as elastic net, decision tree, random forest, and neural network models, logistic regression provided better prediction accuracy across all age groups that took the test. The importance of this paper is that it foils much of the other existing literature in demonstrating that traditional models may be superior in prediction power to machine learning methods depending on the data. The combination of Brieman (2001) and Cornell-Farrow and Garrard (2019) demonstrate the theoretical and actual application of selecting models that provide the best final solution and how this has served as a standard for analysis for almost two decades.

Overall, the literature demonstrates the increasing applications of machine learning methodologies in economics, allowing for new research across the field. It illustrates applications from demand estimation and predicting unemployment rates to financial applications like bank failure and firm credit worthiness. The literature also shows the potential to apply these methods to labor force outcomes like labor force participation and income prediction, which is the main focus of this paper. The literature also covers standards regarding the use of machine learning methodologies in the field, especially focusing on the idea that the goal of methodology selection is to provide the best solution – rather than simply applying new methodology for no apparent purpose. In some cases, traditional methods may provide better results, and in others machine learning or the combination of the methodologies may provide the best output. This research utilizes a wide range of methodologies to find the best possible prediction. This project serves as another application of machine learning in the field of economics and strives to improve the prediction of individual income. A gap in the literature exists for the application of these methodologies to the area of income using highly dimensional demographic data, such as the Current Population Survey.

DATA & METHODOLOGY

This analysis leverages data from the Current Population Survey from 2017-2020 to predict income using a variety of features that capture personal characteristics. It considers traditional linear regression and a selection of machine learning models. This allows for comparison between the various model types to investigate what model type can provide the best income prediction.

Data

The data used in this paper is from the Current Population Survey, which is conducted and sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (U.S. Census Bureau). The data was extracted from IPUMS and includes the data from 2017-2020. Data from 2017 and 2018 is used to train the models, while 2019 and 2020 data is used to as the validation set to test the models. This uses typical methodology from machine learning, in which the models are trained on one subset of data and then evaluated on the predictions produced for a different subset called the validation data. The actual and predicted values for the validation data are compared, allowing for reasonable comparison of the models. The CPS is a voluntary survey conducted each month for approximately 60,000 households. The survey was chosen for its high dimensionality, large number of observations, completeness of data, and its reputation and usage in the field.

The Current Population Survey contains a plethora of information from each respondent and their household. The approximately 150 variables extracted capture a variety of characteristics for individuals from categories such as work, income, education, ethnicity, tax status, poverty status, disability status, migration status, family interrelationships, welfare benefits, and veteran status. These original variables were recoded to create 264 unique variables that were used in the analysis. These new variables created include a large number of dummy variables created for categories such as state of residence, and occupation of the individual. A full listing of variables used is available in Appendix A and further details can be provided at the reader's request. The data used in the analysis contained 467,811 observations and 264 recoded variables after cleaning. The large dimensionality is appropriate for the machine learning techniques, as many of the models can effectively deal with the high dimensionality through variable selection and other techniques.

The response, or target, variable for the analysis is the natural log of real income for individuals. The information from the variable originally came from INCTOT, which is the self-reported total personal income for an individual. According to the documentation, "INCTOT indicates each respondent's total pre-tax personal income or losses from all sources for the previous calendar year" (CPS IPUMS, 2020). While inflation may not have a large impact on the analysis since the data is over a period of four consecutive years, it is important to recognize that the incomes reported are in nominal terms and not real terms. Therefore, a deflator is used to account for inflation so that the incomes can be effectively compared. The deflator used was the Consumer Price Index, as provided by the Bureau of Labor Statistics. Using this adjustment allows the nominal values to be compared in real terms, which in this case has the baseline of 1999 dollars. Only individuals with positive real incomes were then used in the analysis, such that the natural log of the variable could be taken. This assisted with helping to normalize the data as well as increase interpretability of values for the metrics of evaluation. The predictions made on the variable can still be transformed back into real income, which still allows for the same overall interpretability for the final results of the models.

Summary Statistics

It is important to make sure the populations in the testing and training data sets are similar in composition for important variables. This is indeed the case, as demonstrated below. The target of the log of real income had a similar mean and distribution for the three groups.

Table 1: Summary Statistics of Income Variable (Target)

Real Income	Mean (Untransformed)	Mean (Transformed)	Standard Deviation	Minimum	Maximum
2017 & 2018	$ 30,082.46	9.751	1.561	.307	13.979
2019	$ 33,919.63	9.811	1.475	.282	14.125
2020	$ 36,051.57	9.859	1.488	.265	14.123

The testing and training data sets were also similar for other important variables that were identified in previous literature. Average education differed by less than half a year, and gender, disability status, and race were all within 1% of each other for each of the data subsets.

Table 2:Summary Statistics of Key Features

Variable	2017-2018	2019	2020
Observations	241,489	119,493	106,829
Average Education (Years)	13.5	13.6	13.7
% Female	52.61%	52.44%	52.28%
Disability (Any)	11.86%	11.76%	11.98%
Race			
White/Caucasian	76.93%	77.29%	77.39%
African American	12.44%	12.03%	11.84%
Asian	6.62%	6.62%	6.91%

Methodology

This paper utilizes a variety of machine learning and traditional methods to find the best possible prediction for individual income and compare results from the different methods. Methodologies selected are similar to those applied in previous literature. The general framework of analysis assumes that:

$$(1)\ Y_i = G(x_1, x_2, x_3, \ldots, x_k),$$

Where Y is real income, x_i denotes features, and k indexes features (variables) in the dataset. Variable selection and relationships are dictated by the machine learning algorithms or literature for the linear model.

Linear Model: Ordinary Least Squares

A traditional ordinary least squares linear regression serves as the baseline model for analysis, as it represents the most widely used traditional methodology in economics. The linear model finds the linear relationship between independent variables (x_i), and the target or dependent variable (y). Each independent variable (or feature) has a coefficient (β) that measures the relationship between the dependent and independent variables. The OLS method minimizes the residual sum of squares through the selection of coefficients (βs). The residual sum of squares is equal to the squared sum of the actual values for the dependent variable, yi, subtracted by the predicted values of y from the model (or $\beta^T x_i$). This allows for a measure of how well the model created fits the data, and the model aims to fit the data as best as possible and minimize this measure. Mathematically, the OLS estimator is obtained by solving:

$$(2)\ y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon \qquad \text{Goal: } min \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

In this analysis, the linear model is run without using a variable selection method, thus relying on human selection of variables. Based on theory and previous literature, initial independent variables selected for the model include education, age, race, geographic location, and disability among others. Several linear model specifications were attempted, which selected variables based on previous literature. The final model specification used the following independent variables: age, education, state, race, occupation, industry, sex, employment status, difficulty (any), food stamp (dummy), and metropolitan status.

Machine Learning Models

Seven machine learning methods are considered in this research, and were run using sci-kit learn library in Python. LASSO, ridge, and the elastic net model assist in dealing with large dimensionality and help prevent overfitting of the data.

LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a shrinkage method to minimize the penalized residual sum of squares. The addition of a penalty to the traditional model assists in preventing overfitting to the training data. LASSO estimates solving the following algorithm:

(3) $\min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \lambda \sum_{j}^{k} |\beta_j|$, subject to $\sum_{j}^{p} \beta_j^2 \leq s$.

The penalty is denoted by the second part of the equation. LASSO's penalty is the sum of the absolute value of magnitudes of the coefficient multiplied by $\boldsymbol{\lambda}$**,** a tuning parameter that determines the extent of the penalty. The penalty is placed on the square of the coefficients, which puts further restraints on the parameters in the model, as increases in the coefficients lead to corresponding increase in the penalization. With the penalty on of the square of the coefficients, smaller coefficients, especially removing a coefficient, will assist in making the overall penalty term smaller. The actual value for $\boldsymbol{\lambda}$ is determined through cross validation and other information criteria. This model results in dropping regressors as the penalization pushes the coefficients to zero, thus creating a sparser final model.

Ridge

The Ridge regression is very similar to LASSO, but has a key difference in the penalty term. Ridge uses the squared magnitudes of the coefficient in the penalty, while LASSO uses the absolute value of the magnitudes. While LASSO penalization pushes the coefficients to zero, thus dropping variables out of the model, Ridge minimizes the coefficients of the independent variables, leading them to remain in the model, but with reduced coefficients.

$$(4)\ \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \lambda \sum_{j}^{k} \beta_j^2, \text{ subject to } \sum_{j}^{p} \beta_j^2 \leq s$$

Elastic Net

Elastic net is a compromise between LASSO and Ridge methods.

$$(5)\ \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} \left((1-\alpha)|\beta_j|^2 + \alpha|\beta_j| \right)$$

It allows for the balancing of the two methods through the parameter α. α can range from zero to one, with the extremes of one resulting in LASSO and zero resulting in Ridge. This is represented by the parameter L1, which will be seen in later charts. This model combines the prior two methods and is more flexible that the other two models. The combination of the prior methods leads to the creation of a new hyper parameter to potentially tune, while still allowing for the previous models.

Decision Tree Model/Regression Trees

Decision trees split the observations using different cutoffs of the variables in the dataset. The cutoff is often calculated using impurity gain, for which there are different indexes, or through human selection to fit theory. Using impurity gain, the goal is to use the cutoffs/splits to filter the observations into distinct, separate groups.

If left in the current continuous form, regression trees, which are a special type of decision tree that works with continuous variables, can be utilized. In a regression tree, the cutoff value for a split is selected by calculating the sum of squared residuals at each possible cutoff for the variable. The cutoff with the smallest sum of squared residuals is the best split for that variable. For each level of the tree, each variable is evaluated to see which cutoff provides the most impurity gain and is selected as the variable to split the observations. Human selection based on theory is also potentially applicable, in order to give a higher degree of interpretability and understanding. Impurity gain would still be calculated for each potential split, but only splits approved by the researcher would be allowed as possible splits. The cutoff for splits for decision tree are set for the model by the researcher. These stopping rules can include the maximum number of branches, maximum number of levels in the tree, minimum leaf size and more. After a complete tree is grown (stopping at the rules set), it then needs to be pruned. Pruning the tree means that all potential subtrees of the complete tree are considered and tested against validation data. The subtree with the lowest error on the validation data is the optimal tree. Please note that this model is not used as a prediction method. Decision trees can be unstable and cannot provide sufficient predictions for this type of prediction. However, this is included to give the reader a baseline understanding of decision trees, which is necessary to understand some models below.

Random Forest Model

Random Forest models are the collection of a large number of decision trees combined together to form one final prediction. Each tree produces its own prediction, which are then all taken into account with equal weight to produce a final prediction. Each tree produces its own prediction, because of variability within the different trees. The key to random forests is bootstrap aggregation, or bagging, which causes this variety. Each tree is grown using a user-specified number of observations, which is a subset of the training data. The samples for each tree are taken at random from the training data and are done with replacement (same observation can be in multiple trees and/or multiple times in the same tree). The number of variables each tree considers at every split is also user-specified. Typically, the number of variables to consider is less than the total number of variables, so each tree considers different variables at the splits – furthering differences between the trees. In random forests, no trees are pruned; the full grown version of each tree is used. The lack of pruning results because even if one tree overfits one subsection of data, the combination of all the trees under the one random forest will prevent aggregate overfitting. In general, based on results from the literature and past research, random forests consistently provide good predictions (Brieman 2001; Bajari *et al.* 2015; Mullainathan and Speiss, 2017).

Gradient Boosting Model

The Gradient Boosting model utilizes a large number of weak predictors, in this case decision trees, that are built off of each other to create the final prediction. After the first tree is created, all other trees are then fit onto a modified version of the dataset that places a greater weight on predicting observations that were not effectively predicted during the first version. The loss function, which is the function that is being optimized using gradient descent, utilized in the version utilized was least squares regression. The idea is the combination of many of these weak trees that build off each other to fix errors from previous trees will lead to an improved prediction. The number of trees created (boosting stages performed) is the number of estimators.

Ensemble Model

Similar to how random forests combine the predictions of decision trees, the ensemble model combines the predictions of other models to create one final prediction. The random forest model is a special ensemble model of just decision trees. The ensemble model can be a simple or weighted average of the predictions from the input models. In this case, the ensemble model created used a basic average of the other machine learning methods to create the ensemble model. Previous literature has found that ensemble models have produced final predictions that outperform those produced by any single, individual model (Varian, 2014; Mullainathan and Speiss, 2017; Bajari *et al.* 2015; Athey and Imbens, 2019).

K-Nearest Neighbors

The K-Nearest Neighbors, or KNN, model predicts the target variable by identifying similar observations using the features (commonly known as independent variables outside of the field) of the observation. Euclidean distance or Manhattan distance can be used to determine the closeness of observations to each other, but Euclidean distance was used in this analysis. The number of neighbors, or observations that are used to predict the target for the observation, is determined through tuning. A uniform weight is given to all neighbors in the prediction in the model used, although other variations including weighting the points inverse of their distance can be used.

Model Choice & Evaluation

These models were selected to represent key standards of both traditional statistical and machine learning methods in the field of economics. Each model outlined has been effective in conducting analysis in both previous research and throughout the aforementioned literature.

There are several different key metrics that are used to compare the models. The standard for machine learning techniques is to train the models on one subset of data (training data) and test the models on a different subset of data (testing/validation data). The evaluation of the models is based on how well each model can predict the target variable for the testing/validation data. Therefore, each metric is calculated for the application of the trained model on this out-of-sample testing data. One measure is using the traditional R-squared value, measuring how much variation of the target variable can be explained by the model. This is the most common measure of the performance of a continuous target variable. Another key metric is the mean squared error (MSE), which is the sum of the squared difference between the actual value and predicted value of the target for each observation in the testing data. The smaller the MSE, the better the model fits the testing data. Another metric that is common is the mean absolute error (MAE), which is very similar to MSE. Instead of the squared sum, MAE is the sum of the absolute value difference between the actual and predicted values for the target for each observation in the testing data. These three measures are the standard for evaluating a continuous target variable that is used in the analysis.

RESULTS

Model Results

The seven machine learning models performed better than the traditional OLS regression that was built using selection based on previous literature. The traditional OLS model with variable selection from literature resulted in the lowest R^2 values and highest MAE and MSE, and thus was the worst performing model. Based on all three chosen methods of evaluation, the tuned gradient boosting model performed the best. The tuned gradient boosting model used 125 estimators.

The R^2 value for the gradient boosting model was 0.70 for 2019 and 0.69 for 2020, meaning almost 70% of the variation in incomes could be explained by thwe model using the characteristic data provided in the survey. The random forest model performed second best with R^2 values of 0.68 and 0.67, which still provides a decent prediction. All other models performed worse based on this metric. The worst model was the linear regression with variable selection from literature, which provided an R^2 of 0.29 and 0.28 for 2019 and 2020 data respectively. For a full list of the performance of all models based on the R^2 metric, please see Table 3.

Table 3: Results of Models – R^2 Values

Model	Specification	R^2: Training	R^2: 2019	R^2: 2020
Linear Regression	(Selection from Literature)	0.295	0.289	0.280
LASSO	l1_ratio = 1	0.383	0.374	0.366
Ridge	l1_ratio = 0	0.573	0.565	0.552
Elastic Net	alpha: .1, l1_ratio: 1	0.497	0.484	0.474
Gradient Boosting	n_estimators: 125	0.702	**0.696**	**0.688**
Random Forest	max_features: 15, n_estimators: 100	0.957	0.680	0.666
K-Nearest Neighbors	n_neighbors: 10	0.560	0.400	0.388
Ensemble	Equal Weight – All ML Methods	0.704	0.610	0.600

The mean squared error (MSE) metric demonstrated similar results, with gradient boosting performing best and random forest performing second best. Gradient boosting had a MSE of 0.83 for 2019 and 0.87 for 2020, compared to the MSE values of 0.88 (2019) and 0.94 (2020) for the random forest model. All other models performed worse according to this metric, with linear regression performing the worst in this metric as well. The MSE values of the linear model were 1.95 for 2019 and 2.02 for 2020. For a full list of the performance of all models based on the MSE metric, please see Table 4.

Table 4: Results of Models – Mean Squared Error

Model	Specification	MSE Training	MSE 2019	MSE 2020
Linear Regression	(Selection from Literature)	1.892	1.951	2.015
LASSO	l1_ratio = 1	1.654	1.719	1.774
Ridge	l1_ratio = 0	1.146	1.194	1.253
Elastic Net	alpha: .1, l1_ratio: 1	1.349	1.417	1.470
Gradient Boosting	n_estimators: 125	0.800	**0.834**	**0.872**
Random Forest	max_features: 15, n_estimators: 100	0.116	0.880	0.936
K-Nearest Neighbors	n_neighbors: 10	1.180	1.647	1.713
Ensemble	Equal Weight – All ML Methods	0.795	1.071	1.119

The mean absolute error (MAE) metric also provided similar results, with the gradient boosting model and random forest model performing the best in this measurement. The gradient boosting model had a MAE of 0.46 for 2019 and 0.48 for 2020, which was smaller than the random forest model's values of 0.48 and 0.50 respectively. Linear regression still performed the worst, with values of 0.83 and 0.85 for 2019 and 2020. For a full list of the performance of all models based on the MAE metric, please see Table 5.

Table 5: Results of Models – Mean Absolute Error

Model	Specification	MAE Training	MAE 2019	MAE 2020
Linear Regression	(Selection from Literature)	0.827	0.825	0.854
LASSO	l1_ratio = 1	0.757	0.758	0.777
Ridge	l1_ratio = 0	0.604	0.604	0.622
Elastic Net	alpha: .1, l1_ratio: 1	0.674	0.676	0.694
Gradient Boosting	n_estimators: 125	0.461	**0.459**	**0.475**
Random Forest	max_features: 15, n_estimators: 100	0.179	0.479	0.501
K-Nearest Neighbors	n_neighbors: 10	0.585	0.669	0.691
Ensemble	Equal Weight – All ML Methods	0.491	0.547	0.567

These results demonstrate that the machine learning models outperformed the traditional baseline OLS model in predicting individual income. The tuned gradient boosting performed the best across all three metrics, with the random forest model performing second best. The OLS regression with variable selection from literature performed the worst across the three metrics out of all the models performed. This demonstrates that machine learning models could be effective to assist more traditional models in the area of regression. However, the machine learning models are "black box" algorithms that do not allow the same level of interpretability as other methods that are used in economics. Therefore, these models are complementary with existing methods in the field as opposed to substitutes, as each have their own unique strengths.

The ability of the best models to provide an R-squared value of around 70% also illustrates that models can be created that can predict individual income with modest accuracy. These models performed well using only demographic data captured by the Current Population Survey. Private firms and the government should be able to utilize their own datasets to predict income. While each firm or government agency does have access to different amounts of information on individuals, this research demonstrates that high performing models could be created using similar data. In some cases, proprietary data on individuals collected by companies or agencies could be used to produce improved estimates of income. With further investment, firms could develop proprietary models based on their own data that can effectively predict income.

PROPER USE OF MODELS

It is often enticing to leverage models like these beyond their intended purpose. While these models successfully predict an estimated income for an individual based on the characteristics captured in the questions posed by the Current Population Survey, they should not be used to set salaries or be used for any means that discriminate against individuals. Using similar models to assign salaries would be gross misuse of the models and can lead to discriminatory practices. The goal was to academically demonstrate that a survey such as the CPS can effectively allow businesses or the government to predict an individual's income without them actually disclosing the value. It can help companies segment the market to allow for more efficient marketing of products to individuals with certain incomes. Similarly, this could be used by the government as one metric to help identify and investigate potential fraud on income tax returns. However, this should not be used as the only metric and is not perfect. While these predictions are fairly reasonable for most observations, there is still significant error present in the predictions. Therefore, it should only be used as an indicator, alongside other metrics and methods, as an indicator to help with market segmentation and fraud detection. Abuse of these models to set salaries or conduct other practices is abuse of these models and can lead to systematic discrimination and unfair practices. Furthermore, these models perform well for most recent years, but new trends in society, both positive and negative, can greatly impact the effectiveness of prediction.

LIMITATIONS & EXTENSIONS

One key limitation in the research was the technology available to complete the analysis. The data analysis and models were run on a student laptop, which has a lower speed and overall capability compared to higher powered computers available at other institutions. The machine did not have the capability to run the models with as high of a speed, which led to a reduction in the tuning range for the parameters that could be conducted. Models were run overnight and sometimes ran for several days to mitigate this limitation, but it is worth noting that limitation.

Another limitation is the limited focus that this research investigated. More research could be conducted to assist in creating further improved models that could be beneficial to government or private firms. Other model types or methodologies could assist in model creation, which could possibly improve upon this research. There are many other modelling methodologies that could be used for both machine learning models as well as more traditional models in economics. Furthermore, these models relied solely on the Current Population Survey data, which was chosen for its size and quality. Other survey data, such as the American Community Survey, could also be utilized to conduct this analysis and could contain different variables that may lead to significantly different results. Furthermore, macroeconomic data could be added to the analysis, which could assist with the longevity of the models. The focus of the research was to show that successful models could be created within the confines of data collected in a survey, but this could be a useful extension that could assist in improving the ability of the models to continue to predict effectively in the future.

This analysis also only focuses on the sole target of individual income, which is only one metric among of other variables related to earnings for individuals. Similar research could be conducted with wages or family income, which could have similar positive uses by both the public and private sectors.

CONCLUSION

This research finds that machine learning methodologies outperformed the traditional OLS model with variable selection from literature in the area of prediction. The OLS regression performed worse across all three metrics relative to the machine learning models for prediction. This demonstrates that machine learning methodologies could be effective to supplement other models to assist with research focusing on prediction. However, it is important to note that the lower performance of the OLS regression also comes as a result of key differences between OLS and machine learning methods. The machine learning models were able to consider and utilize more features (or independent variables) relative to the OLS regression. The mere fact that more variables were potentially included in the machine learning models is one reason why they outperformed the OLS regression. However, the OLS regression was chosen as the baseline comparison with this in mind. The traditional OLS regression still has many benefits, including its interpretability and lower computing needs, but often requires individuals to create several regressions and self-select variables to be included based on previous literature or other factors. There are methods like stepwise regression that can help with the variable selection, but oftentimes variables are selected by the researcher and fewer variables are considered. The more traditional techniques in economics and machine learning actually complement each other rather than serve as substitutes for one another. The improved prediction power of machine learning methods can be used with tested techniques to further advance research and lead to new outcomes.

Another key finding is that income could be predicted with modest accuracy by the models. The best models, gradient boosting and random forest, were able to account for almost 70% of the variation in income with the personal characteristic features available. This demonstrates the feasibility of firms and individuals to create models that could predict income. While the Current Population Survey does contain necessarily contain the same information that firms may have, companies and others may have collected different information on individuals or could create a form that asks individuals to provide information similar to that of the survey. This 70% threshold was reached using modest computing, publicly available software, and publicly available data. With further investments in time and resources, companies could develop extremely effective proprietary models that can predict income.

The ramifications of the development of such models are extensive. These models could further improve fraud detection, evolve marketing strategies, and even change the purchasing experience of individuals for expensive items like cars. Similar models are already in use by many companies in other areas like credit card use fraud or to market to customers using knowledge of previous purchases. However, this prediction of income could further impact the actions of firms in the future, thus changing the experience of consumers in the future. Government and nonprofit use of similar models could also impact policy decisions and distribution of funds. These models could help verify individuals that request benefits from nonprofits and the government. This could help mitigate many types of fraud including benefits and tax fraud. Reducing fraud could allow for the better allocation of resources and money by both the government and nonprofits, which can be used in a more positive way to help society.

This research does not provide a perfect model that can predict income from personal characteristics. Rather, it demonstrates the applicability of these methods in economic research and the feasibility for such models to be developed by those willing to put large amounts of labor and capital into its development. These implications have far-reaching effects that, if applied ethically, could have positive benefits for society.

APPENDICES

Appendix A: Variable List & Summary Statistics

Variables Included in Analysis						
Variable	Variable Description	Obs	Mean	Std. Dev.	Min	Max
ln_real_inctot	Natural Log of Real Individual Total Income (Target)	467,811	9.791	1.523	0.265	14.125
heatval	Value of Energy Assistance Received	467,811	12.420	109.267	0	8000
stampno	Number of People Covered by Food Stamps in HH	467,811	0.194	0.790	0	9
stampmo	Number of Months HH Received Food Stamps in Previous Year	467,811	0.898	3.064	0	12
stampval	Value of Food Stamps Received by HH in Previous Year	467,811	225.450	1045.416	0	30000
nfams	Number of families in HH	467,811	1.090	0.423	1	16
ncouples	Number of Married Couples in HH	467,811	0.718	0.492	0	5
nmothers	Number of Mothers in HH	467,811	0.573	0.592	0	4
nfathers	Number of Fathers in Household	467,811	0.467	0.538	0	5
age	Age (Years)	467,811	48.235	17.546	18	85
famsize	Number of Other Family Members in HH	467,811	2.915	1.571	1	16
nchild	Number of Own Children in HH	467,811	0.809	1.133	0	9
nchlt5	Number of Own Children Under 5 Years Old in HH	467,811	0.148	0.444	0	5
famunit	Number of Family in HH	467,811	1.035	0.247	1	16
nsibs	Number of Siblings Residing in HH	467,811	0.100	0.463	0	9
ptweeks	Number of Weeks Working Part Time Last Year	467,811	5.477	14.320	0	52
durunem2	Duration of Unemployment Spell	467,811	0.174	1.282	0	16
wkswork1	Weeks Worked Last Year	467,811	35.027	22.958	0	52
numemps	Number of Employers Last Year	467,811	0.832	0.603	0	3
strechlk	Number of Stretches Looking for Employment Last Year	467,811	0.076	0.411	0	4
mthwelfr	Number of Months Received Welfare Income Last Year	467,811	0.071	0.885	0	12
health	Health Status (Rating 1-5)	467,811	2.322	1.071	1	5
atelunch_r	Number of Children in HH Who Ate School Lunch	467,811	0.385	0.825	0	9
freelunch_r	Number of Children in HH Who Ate Free or Reduced Lunches	467,811	0.189	0.628	0	9
real_ftotval	Real Total Family Income	467,811	65206.710	73001.780	-6396	1585789
educ_r	Years of Education	467,811	13.742	2.811	1	21
ahrsworkt_r	Number of Hours Worked in Previous Week	467,811	25.405	21.171	0	99
uhrsworkly_r	Usual Hours Worked Per Week Last Year	467,811	28.741	19.812	0	99

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
sdummy1	State of Residence Dummy - AL	467,811	0.020
sdummy2	State of Residence Dummy - AK	467,811	0.012
sdummy3	State of Residence Dummy - AZ	467,811	0.017
sdummy4	State of Residence Dummy - AR	467,811	0.018
sdummy5	State of Residence Dummy - CA	467,811	0.098
sdummy6	State of Residence Dummy - CO	467,811	0.013
sdummy7	State of Residence Dummy - CT	467,811	0.010
sdummy8	State of Residence Dummy - DE	467,811	0.011
sdummy9	State of Residence Dummy - DC	467,811	0.018
sdummy10	State of Residence Dummy - FL	467,811	0.047
sdummy11	State of Residence Dummy - GA	467,811	0.022
sdummy12	State of Residence Dummy - HI	467,811	0.017
sdummy13	State of Residence Dummy - ID	467,811	0.016
sdummy14	State of Residence Dummy - IL	467,811	0.028
sdummy15	State of Residence Dummy - IN	467,811	0.016
sdummy16	State of Residence Dummy - IA	467,811	0.011
sdummy17	State of Residence Dummy - KS	467,811	0.012
sdummy18	State of Residence Dummy - KY	467,811	0.011
sdummy19	State of Residence Dummy - LA	467,811	0.022
sdummy20	State of Residence Dummy - ME	467,811	0.008
sdummy21	State of Residence Dummy - MD	467,811	0.013
sdummy22	State of Residence Dummy - MA	467,811	0.021
sdummy23	State of Residence Dummy - MI	467,811	0.022
sdummy24	State of Residence Dummy - MN	467,811	0.013
sdummy25	State of Residence Dummy - MS	467,811	0.018
sdummy26	State of Residence Dummy - MO	467,811	0.013
sdummy27	State of Residence Dummy - MT	467,811	0.018
sdummy28	State of Residence Dummy - NE	467,811	0.012
sdummy29	State of Residence Dummy - NV	467,811	0.014
sdummy30	State of Residence Dummy - NH	467,811	0.013
sdummy31	State of Residence Dummy - NJ	467,811	0.020
sdummy32	State of Residence Dummy - NM	467,811	0.019
sdummy33	State of Residence Dummy - NY	467,811	0.041
sdummy34	State of Residence Dummy - NC	467,811	0.023
sdummy35	State of Residence Dummy - ND	467,811	0.013
sdummy36	State of Residence Dummy - OH	467,811	0.024
sdummy37	State of Residence Dummy - OK	467,811	0.015
sdummy38	State of Residence Dummy - OR	467,811	0.015
sdummy39	State of Residence Dummy - PA	467,811	0.026
sdummy40	State of Residence Dummy - RI	467,811	0.009
sdummy41	State of Residence Dummy - SC	467,811	0.016
sdummy42	State of Residence Dummy - SD	467,811	0.011
sdummy43	State of Residence Dummy - TN	467,811	0.018
sdummy44	State of Residence Dummy - TX	467,811	0.059
sdummy45	State of Residence Dummy - UT	467,811	0.015
sdummy46	State of Residence Dummy - VT	467,811	0.012

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
sdummy47	State of Residence Dummy - VA	467,811	0.019
sdummy48	State of Residence Dummy - WA	467,811	0.019
sdummy49	State of Residence Dummy - WV	467,811	0.019
sdummy50	State of Residence Dummy - WI	467,811	0.013
metro2	Metropolitan Status Dummy - In Metro, Central City	467,811	0.259
metro3	Metropolitan Status Dummy - In Metro, Outside Central City	467,811	0.396
cbsasz2	Metropolitan Size Dummy - 100,000 - 249,999	467,811	0.075
cbsasz3	Metropolitan Size Dummy - 250,000 - 499,999	467,811	0.077
cbsasz4	Metropolitan Size Dummy - 500,000 - 999,999	467,811	0.136
cbsasz5	Metropolitan Size Dummy - 1,000,000 - 2,499,999	467,811	0.149
cbsasz6	Metropolitan Size Dummy - 2,500,000 - 4,999,999	467,811	0.115
cbsasz7	Metropolitan Size Dummy - 5,000,000+	467,811	0.205
race1	Race Dummy - Caucasian	467,811	0.778
race2	Race Dummy - African American	467,811	0.119
race3	Race Dummy - Native American	467,811	0.014
race4	Race Dummy - Asian	467,811	0.064
gq_d	Group Quarter Status (Dummy)	467,811	0.001
relate2	Relation to HH Head - Spouse (Dummy)	467,811	0.262
relate3	Relation to HH Head - Child (Dummy)	467,811	0.088
relate4	Relation to HH Head - Parent (Dummy)	467,811	0.022
relate5	Relation to HH Head - Sibling (Dummy)	467,811	0.012
relate6	Relation to HH Head - Grandchild (Dummy)	467,811	0.004
relate7	Relation to HH Head - Other Relative (Dummy)	467,811	0.016
relate8	Relation to HH Head - Unmarried partner (Dummy)	467,811	0.033
relate9	Relation to HH Head - Roommate (Dummy)	467,811	0.017
relate10	Relation to HH Head - Lodger (Dummy)	467,811	0.003
relate11	Relation to HH Head - Foster Child (Dummy)	467,811	0.000
relate12	Relation to HH Head - Other nonrelative (Dummy)	467,811	0.006
ftype2	Family Type Dummy - Nonfamily Householder	467,811	0.171
ftype3	Family Type Dummy - Related Subfamily	467,811	0.028
ftype4	Family Type Dummy - Unrelated Subfamily	467,811	0.002
ftype5	Family Type Dummy - Secondary Individual	467,811	0.056
famkind2	Family Kind Dummy - Male Reference	467,811	0.158
famkind3	Family Kind Dummy - Female Reference	467,811	0.230
famrel2	Relationship to Family Dummy - Reference Person	467,811	0.387
famrel3	Relationship to Family Dummy - Spouse	467,811	0.272
famrel4	Relationship to Family Dummy - Child	467,811	0.079
famrel5	Relationship to Family Dummy - Other Relative	467,811	0.036
citizen2	Citizenship Status Dummy - Naturalized Citizen	467,811	0.085
citizen3	Citizen Status Dummy - Not a Citizen	467,811	0.078
hispan_d	Hispanic Status (Dummy)	467,811	0.171
marst2	Marital Status - Married (Dummy)	467,811	0.565
marst3	Marital Status - Formerly Married (Dummy)	467,811	0.186
empstat2	Employment Status Dummy - Unemployed	467,811	0.025
empstat3	Employment Status Dummy - Not in Labor Force	467,811	0.298

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
usftptlw_d	Usually work Full Time if Worked Part-Time Last Week (Dummy)	467,811	0.046
whynwly2	Reason for Not Working Last Year - Could Not Find Work (Dummy)	467,811	0.002
whynwly3	Reason for Not Working Last Year - Ill/Disabled (Dummy)	467,811	0.056
whynwly4	Reason for Not Working Last Year - Taking Care of Family (Dummy)	467,811	0.025
whynwly5	Reason for Not Working Last Year - Education (Dummy)	467,811	0.014
whynwly6	Reason for Not Working Last Year - Retired (Dummy)	467,811	0.164
whynwly7	Reason for Not Working Last Year - Other (Dummy)	467,811	0.002
actnlfly2	Activity When Not in Labor Force - Ill/Disabled (Dummy)	467,811	0.012
actnlfly3	Activity When Not in Labor Force - Taking Care of Family (Dummy)	467,811	0.021
actnlfly4	Activity When Not in Labor Force - Education (Dummy)	467,811	0.025
actnlfly5	Activity When Not in Labor Force - Retired (Dummy)	467,811	0.012
actnlfly6	Activity When Not in Labor Force - Other (Dummy)	467,811	0.025
actnlfly7	Activity When Not in Labor Force - No Work Available (Dummy)	467,811	0.011
ownershp_d	HH Owned (Dummy)	467,811	0.701
pubhouse_d	Public Housing Status (Dummy)	467,811	0.023
rentsub_d	Subsidized Housing (Dummy)	467,811	0.009
heatsub_d	Heat Subsidy Received (Dummy)	467,811	0.025
foodstmp_d	Food Stamps Received (Dummy)	467,811	0.087
lunchsub_d	Children Received Free or Reduced Lunches (Dummy)	467,811	0.108
unitsstr2	Housing Structure - 2 Units (Dummy)	467,811	0.041
unitsstr3	Housing Structure - 3-4 Units (Dummy)	467,811	0.038
unitsstr4	Housing Structure - 5-9 Units (Dummy)	467,811	0.041
unitsstr5	Housing Structure - 10+ Units (Dummy)	467,811	0.108
phone_d	Telephone Availability in HH (Dummy)	467,811	0.975
sex_d	Sex, (Dummy; Female = 1)	467,811	0.512
nativity2	Nativity - Born in U.S., 1 Parent Native to U.S. (Dummy)	467,811	0.037
nativity3	Nativity - Born in U.S., Both Parents Foreign (Dummy)	467,811	0.048
nativity4	Nativity - Foreign Born (Dummy)	467,811	0.179
unitsstr2	Housing Structure - 2 Units (Dummy)	467,811	0.041
unitsstr3	Housing Structure - 3-4 Units (Dummy)	467,811	0.038
unitsstr4	Housing Structure - 5-9 Units (Dummy)	467,811	0.041
unitsstr5	Housing Structure - 10+ Units (Dummy)	467,811	0.108
phone_d	Telephone Availability in HH (Dummy)	467,811	0.975
sex_d	Sex, (Dummy; Female = 1)	467,811	0.512
vetstat_d	Veteran Status (Dummy)	467,811	0.079
wkstat_current_2	Current Work Status - Part-Time (Dummy)	467,811	0.148
wkstat_current_3	Current Work Status - Full Time (Dummy)	467,811	0.505
wkstat_typical_2	Typical Work Status - Part-Time (Dummy)	467,811	0.118
wkstat_typical_3	Typical Work Status - Full Time (Dummy)	467,811	0.584
schlcoll_d	Currently Attending School (Dummy)	467,811	0.065
dependent_d	Dependent Status (Dummy)	467,811	0.038

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
diffhear_d	Hearing Difficulty (Dummy)	467,811	0.036
diffeye_d	Vision Difficulty (Dummy)	467,811	0.018
diffrem_d	Memory Difficulty (Dummy)	467,811	0.036
diffphys_d	Physical Difficulty (Dummy)	467,811	0.070
diffmob_d	Mobility Difficulty (Dummy)	467,811	0.040
diffcare_d	Personal Care Limitation Difficulty (Dummy)	467,811	0.020
diffany_d	Any Difficulty (Dummy)	467,811	0.119
gov_worker	Government Worker Last Year (Dummy)	467,811	0.109
selfemployed	Self Employed Last Year (Dummy)	467,811	0.071
privateworker	Private Worker Last Year (Dummy)	467,811	0.556
workly_d	Worked Last Year (Dummy)	467,811	0.736
fullpart_d	Part Time Worker (Dummy; 1 = Part Time)	467,811	0.136
pension_d	Receive Pension Plan at Work (Dummy)	467,811	0.242
wanttowork_d	Want A Job, Not in LF (Dummy)	467,811	0.717
disabwrk_d	Work Disability (Dummy)	467,811	0.093
quitsick_d	Quit Job or Retired for Health Reasons (Dummy)	467,811	0.041
srcearn2	Source of Earnings from Longest Job - Wage/Salary (Dummy)	467,811	0.693
srcearn3	Source of Earnings from Longest Job - Self Employment (Dummy)	467,811	0.040
srcearn4	Source of Earnings from Longest Job - Farm Self Employment (Dummy)	467,811	0.003
gotvdisa_d	Received Veterans' Disability Compensation (Dummy)	467,811	0.013
gotveduc_d	Received Veterans' Education Assistance (Dummy)	467,811	0.001
gotvothe_d	Received Other Veterans' Payments (Dummy)	467,811	0.001
gotvpens_d	Received Veterans' Pension (Dummy)	467,811	0.004
gotvsurv_d	Received Veterans' Survivor Benefits (Dummy)	467,811	0.001
paidgh2	Employer Paid for Part of Health Plan (Dummy)	467,811	0.296
paidgh3	Employer Paid for All of Health Plan (Dummy)	467,811	0.078
himcaidly_d	Covered by Medicaid Last Year (Dummy)	467,811	0.124
himcarely_d	Covered by Medicare Last Year (Dummy)	467,811	0.226
hichamp_d	Covered by Military Insurance Last Year (Dummy)	467,811	0.043
phinsur_d	Covered by Private Health Insurance Last Year (Dummy)	467,811	0.699
phiown_d	Covered by Private Health Insurance in Own Name Last Year (Dummy)	467,811	0.502
caidly_d	Covered by Medicaid Last Year Based on Qualifications (Dummy)	467,811	0.121
anycovnw_d	Covered by Health Insurance at Time of Interview (Dummy)	467,811	0.910
gotwic_d	Received WIC Benefits in Previous Year (Dummy)	467,811	0.012
union_d	Union Membership (Dummy	467,811	0.012
occ1	Occupation Dummy - Management	467,811	0.084
occ2	Occupation Dummy - Business and Financial Operations	467,811	0.036
occ3	Occupation Dummy - Computer and Mathematical Science	467,811	0.022
occ4	Occupation Dummy - Architecture and Engineering	467,811	0.014
occ5	Occupation Dummy - Life, Physical, and Social Science	467,811	0.007
occ6	Occupation Dummy - Community and Social Service	467,811	0.009
occ7	Occupation Dummy - Legal	467,811	0.009

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
occ8	Occupation Dummy - Education, Training, and Library	467,811	0.041
occ9	Occupation Dummy - Arts, Design, Entertainment, Sports, and Media	467,811	0.014
occ10	Occupation Dummy - Healthcare Practitioner and Technical	467,811	0.043
occ11	Occupation Dummy - Healthcare Support	467,811	0.011
occ12	Occupation Dummy - Protective Service	467,811	0.014
occ13	Occupation Dummy - Food Preparation and Serving Related	467,811	0.036
occ14	Occupation Dummy - Building and Grounds Cleaning/Maintenance	467,811	0.028
occ15	Occupation Dummy - Personal Care and Service	467,811	0.025
occ16	Occupation Dummy - Sales and Related	467,811	0.067
occ17	Occupation Dummy - Office and Administrative Support	467,811	0.079
occ18	Occupation Dummy - Farming, Fishing, and Forestry	467,811	0.006
occ19	Occupation Dummy - Construction and Extraction	467,811	0.039
occ20	Occupation Dummy - Installation, Maintenance, and Repair	467,811	0.023
occ21	Occupation Dummy - Production	467,811	0.040
occ22	Occupation Dummy - Transportation	467,811	0.026
occ23	Occupation Dummy - Material Moving	467,811	0.019
occ24	Occupation Dummy - Armed Forces	467,811	0.000
ind1	Industry Dummy - Agriculture	467,811	0.011
ind2	Industry Dummy - Forestry, Logging, Fishing, Hunting, and Trapping	467,811	0.001
ind3	Industry Dummy - Mining	467,811	0.005
ind4	Industry Dummy - Construction	467,811	0.051
ind5	Industry Dummy - Nonmetallic Mineral Products	467,811	0.002
ind6	Industry Dummy - Primary Metals and Fabricated Metal Products	467,811	0.007
ind7	Industry Dummy - Machinery Manufacturing	467,811	0.005
ind8	Industry Dummy - Computer and Electronic Products	467,811	0.005
ind9	Industry Dummy - Electrical Equipment, Appliance Manufacturing	467,811	0.002
ind10	Industry Dummy - Transportation Equipment Manufacturing	467,811	0.011
ind11	Industry Dummy - Wood Products	467,811	0.002
ind12	Industry Dummy - Furniture and Fixtures Manufacturing	467,811	0.002
ind13	Industry Dummy - Miscellaneous and Not Specified Manufacturing	467,811	0.007
ind14	Industry Dummy - Food Manufacturing	467,811	0.010
ind15	Industry Dummy - Beverage and Tobacco Products	467,811	0.001
ind16	Industry Dummy - Textile, Apparel, and Leather Manufacturing	467,811	0.003
ind17	Industry Dummy - Paper and Printing	467,811	0.004
ind18	Industry Dummy - Petroleum and Coal Products	467,811	0.001
ind19	Industry Dummy - Chemical Manufacturing	467,811	0.006
ind20	Industry Dummy - Plastics and Rubber Products	467,811	0.002
ind21	Industry Dummy - Wholesale Trade	467,811	0.016
ind22	Industry Dummy - Retail Trade	467,811	0.073
ind23	Industry Dummy - Transportation and Warehousing	467,811	0.032
ind24	Industry Dummy - Utilities	467,811	0.006

Appendix A Continued: Variable List & Summary Statistics

Dummy Variables Included in Analysis			
Variable	Variable Description	Obs	Mean
ind25	Industry Dummy - Publishing Industries	467,811	0.002
ind26	Industry Dummy - Motion Picture and Sound Recording Industries	467,811	0.002
ind27	Industry Dummy - Broadcasting	467,811	0.002
ind28	Industry Dummy - Internet Publishing and Broadcasting	467,811	0.000
ind29	Industry Dummy - Telecommunications	467,811	0.004
ind30	Industry Dummy - Internet Service Providers and Data Processing Services	467,811	0.000
ind31	Industry Dummy - Other Information Services	467,811	0.001
ind32	Industry Dummy - Finance	467,811	0.020
ind33	Industry Dummy - Insurance	467,811	0.010
ind34	Industry Dummy - Real Estate	467,811	0.009
ind35	Industry Dummy - Rental and Leasing Services	467,811	0.002
ind36	Industry Dummy - Professional and Technical Services	467,811	0.053
ind37	Industry Dummy - Management of Companies and Enterprises	467,811	0.001
ind38	Industry Dummy - Administrative and Support Services	467,811	0.027
ind39	Industry Dummy - Waste Management and Remediation Services	467,811	0.002
ind40	Industry Dummy - Educational Services	467,811	0.068
ind41	Industry Dummy - Hospitals	467,811	0.026
ind42	Industry Dummy - Health Care Services (Except Hospitals)	467,811	0.037
ind43	Industry Dummy - Social Assistance	467,811	0.016
ind44	Industry Dummy - Arts, Entertainment, and Recreation	467,811	0.014
ind45	Industry Dummy - Accommodation	467,811	0.008
ind46	Industry Dummy - Food Services and Drinking Places	467,811	0.040
ind47	Industry Dummy - Repair and Maintenance	467,811	0.009
ind48	Industry Dummy - Personal and Laundry Services	467,811	0.012
ind49	Industry Dummy - Membership Associations and Organizations	467,811	0.010
ind50	Industry Dummy - Private Households	467,811	0.003
ind51	Industry Dummy - Public Administration	467,811	0.037
ind52	Industry Dummy - Armed Forces	467,811	0.000

REFERENCES

"Current Population Survey (CPS)." The United States Census Bureau. United States, September 25, 2019. https://www.census.gov/programs-surveys/cps.html.

Athey, Susan, and Guido W. Imbens. "Machine Learning Methods That Economists Should Know About." Annual Review of Economics 11, no. 1 (March 1, 2019): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Bajari, Patrick, Denis Nekipelov, Stephen Ryan, and Miaoyu Yang. "Machine Learning Methods for Demand Estimation." American Economic Review: Papers & Proceedings 105, no. 5 (May 1, 2015): 481–85. https://doi.org/10.3386/w20955.

Barzarbash, Majid. "Fintech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk." IMF: Money and Capital Department, May 1, 2019.

Breiman, Leo. "Statistical Modeling: The Two Cultures." Statistical Science 16, no. 3 (December 24, 2001): 199–231. https://doi.org/10.1214/ss/1009213726.

Brownlee, Jason. "Supervised and Unsupervised Machine Learning Algorithms." Machine Learning Mastery, August 12, 2019. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

Cornell-Farrow, Sarah, and Robert Garrard. "Machine Learning Classifiers Do Not Improve The Prediction Of Academic Risk: Evidence From Australia." Applied Economics, April 2, 2019.

Chase, Audrey, McKenzie Kozma, Michael Matkowski. "Labor Force Participation: A Machine Learning Approach." Presentation at Eastern Economic Conference, Boston, MA, February 28-March 1.

Einav, Liran, and Jonathan Levin. "Economics in the Age of Big Data." Science 346, no. 6210 (November 7, 2014). https://doi.org/10.1126/science.1243089.

Hall, Aaron Smalter. "Machine Learning Approaches to Macroeconomic Forecasting." The Federal Reserve Bank of Kansas City Economic Review 103, no. 4 (2018). https://doi.org/10.18651/er/4q18smalterhall.

Johnson, Barry W., Peter J. Rose, Theodore Black, Andrew Johns, Patrick Langetieg, Kara Leibel, Mark Payne, Alan Plumley, Mary-Helen Risler, and Eric Spitzer. Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011–2013 , Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011–2013 § (2019).

Lazar, Alina. "Income Prediction via Support Vector Machine." 2004 International Conference on Machine Learning and Applications, December 16, 2004. https://doi.org/10.1109/icmla.2004.1383506.

Le, Hong Hanh, and Jean-Laurent Viviani. "Predicting Bank Failure: An Improvement by Implementing a Machine-Learning Approach to Classical Financial Ratios." Research in International Business and Finance 44 (February 17, 2017): 16–25. https://doi.org/10.1016/j.ribaf.2017.07.104.

Mareckova, Jana, and Windried Pohlmeier. "Noncognitive Skills and Labor Market Outcomes: A Machine Learning Approach." University of Vienna Conference on Alternate Money and Financial Structure, May 1, 2017.

Matz, Sandra C., Jochen I. Menges, David J. Stillwell, and Andrew H. Schwartz. "Predicting Individual-Level Income From Facebook Profiles." Plos One 14, no. 3 (March 28, 2019): 1–13. https://doi.org/10.1371/journal.pone.0214369.

Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives 31, no. 2 (2017): 87–106. https://doi.org/10.1257/jep.31.2.87.

Ramcharan, Rodney. "Regressions: Why Are Economists Obessessed with Them?" Finance & Development Magazine: IMF 43, no. 1 (March 2006).

Saltzman, Bennett, and Julieta Yung. "A Machine Learning Approach to Identifying Different Types of Uncertainty." Economics Letters 171 (July 11, 2018): 58–62. https://doi.org/10.1016/j.econlet.2018.07.003.

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles and J. Robert Warren. Integrated Public Use Microdata Series, Current Population Survey: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D030.V7.0

Schneider, Astrid, Gerhard Hommel, and Maria Blettner. “Linear Regression Analysis.” Series on Evaluation of Scientific Publications: NIH 14 (December 5, 2010).

Varian, Hal R. “Big Data: New Tricks for Econometrics.” Journal of Economic Perspectives 28, no. 2 (2014): 3–28. https://doi.org/10.1257/jep.28.2.3.