DISSERTATION

PREDICTING THE PAYCHECK: USING MACHINE LEARNING TO UNDERSTAND DETERMINANTS OF INCOME

Submitted by

Annika Benson

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Josh Prasad

Danielle Gardner
Mark Prince
Samantha Conroy

ABSTRACT

PREDICTING THE PAYCHECK: USING MACHINE LEARNING TO UNDERSTAND DETERMINANTS OF INCOME

Income is a variable of interest in industrial/organizational psychology due to its relationship with outcomes like turnover, motivation, and psychological well-being. However, current research on income has generally assumed a linear relationship between predictors and income, not accounting for potential curvilinear effects or variable interactions. Further, studies on income indicate that large amounts of variance are unaccounted for, suggesting there are predictors yet to be identified. This study addresses those gaps in the research by using machine learning techniques and a large archival data set to investigate the strength and nature of how variables contribute to predicting income. Results demonstrate the effectiveness of machine learning techniques over traditional OLS regression and identifies variables not found currently in the literature. Findings from this research can be used both to create more effective organizational compensation systems as well as indicate targets for interventions to address income inequality.

DEDICATION

This work is dedicated to those who have shaped me into the person I am today.

It is for my mom, dad, sister, and brother, who gave me all the love and support I needed to believe I could succeed in graduate school.

For my cohort, for countless hours of studying, camaraderie, and teaching me there were different ways to be right.

For my friend and mentor, Kelsie, who taught me how to find value when it seemed there was none and showed me compassion I did not know I needed.

And for my partner, Nathaniel, who has been by my side every step of the way. He always believed I could do it, even if I did not.

I hope I have made them proud

TABLE OF CONTENTS

LITERATURE REVIEW

Primarily inductive in nature, exploratory research may be the best methodological approach when the process has been generally expanded using only predictive methods or has not received much empirical scrutiny (Stebbins, 2001). The goal of this method of research is to discover generalizations that can inform future theories and hypotheses (Stebbins, 2001). Researchers have started to incorporate more exploratory and predictive statistical methods in recent years, partially driven by the replication crisis of studies largely using statistical inference methods (Orrù et al., 2020) and the plethora of data now easily accessible (Grimmer et al., 2021). In the field of industrial/organizational psychology, researchers are playing catch-up as organizations deploy machine learning tools to remove bias from job descriptions, screen resumes faster, test applicant cognitive abilities, and evaluate person-job fit often without a full understanding of a tool's reliability, validity, or fairness (Gonzales et al., 2019).

Income, a key component of the employment relationship between employee and organization, has received scant consideration within industrial/organizational psychology despite its impact on workers, businesses, and society. An individual's income (wages or earnings) is a means to pursue a satisfactory quality of life (Leana & Meuris, 2015), as well as impacts one's ability to achieve security, pursue goals, maintain well-being, and attain status (Lawler, 1971). Within industrial/organizational psychology specifically, income is both an important predictor and outcome due to its relationship with workplace performance (Bryson et al., 2010), turnover (Conroy et al., 2022; Trevor et al., 1997), motivation (Conroy & Gupta, 2018) and worker health and well-being (Sayre & Conroy, 2023; Whelan, 1992). Further, worker compensation is one of the largest expenses for organizations (Conroy, 2019). As such,

organizations strive to create effective compensation structures that motivate and retain workers without compromising the financial health of the business.

At a societal level, wage inequality in the United States has been persistent and continues to increase with a greater proportion of income accrued in upper-income households compared to middle and lower-income households, with the most economic growth since the 1980s going to the top 5% of families (Horowitz et al., 2020). Income inequality is associated with lower mobility for the economically disadvantaged, poorer health and emotional well-being, higher mortality, and occupational outcomes like absenteeism (Leana & Meuris, 2015). This relationship is explained by a framework proposed by Sayre and Conroy (2023) where factors such as pay level and performance pay may be insufficient to meet basic needs or misaligned with our perceived efforts, increasing allostatic load (i.e. the cumulative effects of body system dysregulation (Stephan et al., 2016)) and thus potential adverse health outcomes. As such, individuals and organizations have a vested interest in understanding the factors which influence income.

Understanding income has proven to be a complicated endeavor as there are numerous variables, both individual and contextual, that are related to wages such as industry (Mokre & Rehm, 2020), education (Mishel, 2012), skills (Yerger, 2017), personality (Ng et al., 2005), race (Castilla, 2008), and gender (Säve-Söderbergh, 2019). Even with the breadth of predictors already identified in the literature, studies report substantial variance that is unexplained by included variables, suggesting there are predictors yet to be identified. In a 2005 meta-analysis by Ng and colleagues, thousands of studies on compensation were analyzed and variables considered to be important predictors of salary like race, gender, and education, had corrected correlation coefficients of .11, .18, and .29, respectively. Similarly, a study by Spurk and Aberle

(2011) demonstrated only 12% of the variance in annual salary was explained by all demographic and personality variables included in a multiple mediation model. This problem is compounded by the difficulty of including all relevant predictors in a model in a field that highly values both explainability and parsimony.

The current literature on wage prediction has also failed to consider numerous important factors. First, most studies fail to account for how relevant predictors may interact with each other. For example, Castilla (2008) reports that non-U.S.-born employees make 4.5% less than U.S.-born employees and that salary growth is less for women than men, but does not address how the country of origin and gender might influence each other. This problem is partially a result of traditional methodologies being unable to incorporate known antecedents of income concurrently as well as a lack of literature investigating variable interactions within this domain. Second, explanatory methodologies are often less generalizable than predictive methods. Generally, the quality of a proposed model is evaluated on a "goodness of fit" statistic between the proposed model and the data in hand. This methodology does not consider a model's ability to generalize to out-of-sample data and can reduce the likelihood of generating a model with good predictions (Yarkoni & Westfall, 2017). This means that while a proposed model might work well within a specific study, it likely cannot effectively predict wages outside of the specified study condition. Even if there is good model fit, without a model that incorporates the numerous known predictors of income, there is a lack of evidence to indicate which factors matter most. As evidenced by the content discussed here, current methods for predicting wages can be improved.

Research neglects to consider the broad scope of predictors related to wages, and as such, this study heeds the call to identify a larger set of heterogeneous predictors (Ng et al., 2005),

explore new and untested variables (Nyhus & Pons, 2005) use large sample sizes (DeHaro et al., 2020), and examine how variables interact with each other and the strength of their contributions (Nyhus & Pons, 2005) in predicting wages. Results from this study can assist in identifying variables that significantly contribute to wage predictions such that individuals better understand what influences their compensation, organizations can construct more effective compensation systems, and interventions can be created at a societal level to address wage inequality.

The current study aims to use machine learning techniques to uncover predictors of wages not already found in the literature. By focusing on prediction rather than explanation, predictors that may have traditionally been overlooked or seemed unrelated can be investigated. Further, by incorporating models that can account for the nonlinearity and interaction of predictor variables, the nature of how predictors impact income can be better understood as previously identified relationships have generally been assumed to have a linear relationship. In this study, I use archival data from the American Community Survey from 2018 and 2019 to explore 1) whether machine learning techniques result in improved prediction over ordinary least squares (OLS) regression, 2) if there are differences in algorithm utility, and 3) which variables are the most relevant predictors of wages and whether there are variables in that set of predictors not already represented in the literature.

This research will be valuable to a variety of academic researchers, even those who are not interested in income. In this context, wage prediction serves as a larger example of how machine learning techniques can be applied to an area in which there is already a strong research program to glean new insights. Prediction is underused in social sciences and can help address the criticism that studies lack real world relevance and are overly simplistic (Verhagen, 2022). This study can clarify for social scientists how to approach a methodology that may be

unfamiliar and apply it to their own work. Further, the outcomes of this study are likely to be valuable to wage researchers and organizations. Results may provide new evidence for theory building and emphasize variables that may need further research or were previously not considered to be important. Organizations can also use these outcomes to evaluate their compensation outcomes to see if organizational predictors of wages match larger national trends.

Figure 1 below provides an advanced organizer of the methodological steps used in this paper such that readers can better understand the introduction and easily replicate these steps themselves. The letter "v" denotes the number of variables included in each step.

Figure 1. The letter "v" denotes the number of variables included in each step.

New Era of Prediction

The field of psychology has aimed to understand human behavior and historically this has meant trying to both explain the causal underpinnings of behavior and predict behaviors that have not yet been observed. There are three basic types of research in the social sciences: explanation, description, and exploration (Strydom, 2013). Explanatory modeling is the process of testing causal hypotheses about theoretical constructs generally using observed data and causal inference methods, while predictive modeling applies a model to current data to predict future observations (Shmueli, 2010). Common methods used in explanatory modeling are t-tests, ANOVA, and structural equation models, while predictive modeling falls within the machine learning domain with methods like random forest and elastic net regression. Theoretically, explanation should facilitate prediction, such that a proposed model that best fits observed behavior should in turn also be the best model to predict future behavior. In practice, however, explanation and prediction often are in statistical tension with one another (Yarkoni & Westfall, 2017) as measured data are not always accurate representations of the constructs they are supposed to represent, making it unlikely that an explanatory model will be accurate in its predictions (Shmueli, 2010). Explanatory models also must have a limited number of predictors included such that the model can be both estimated and clearly explained. This means that not every variable that may impact an outcome can be included, further limiting an explanatory model's ability to predict (Putka et al., 2018). Even so, explanation has received the majority of scientific focus over the past 40 years (Douglas, 2009).

What is machine learning?

Models with a predictive focus fall under the umbrella of machine learning (ML). ML is a term for a broad set of strategies used to make predictions or find patterns in new data after

building their “knowledge” on a training set of data (Goretsko & Israel, 2022). The types of models can be broken down into two categories: supervised and unsupervised models. In supervised models, the training data can be meaningfully divided into predictors and labeled outcome(s). The goal then is for the algorithm to “learn” how predictors relate to the outcome(s). Unsupervised models, on the other hand, focus on finding structure (e.g., groups, dimensions), in unstructured data (Orrù et al., 2020). The goal of the present work is to adopt a supervised machine learning approach whereby several predictors are used to estimate income. Supervised machine learning includes conventional approaches like multiple regression, which are used in the present work. However, additional modern prediction methods are applied (Putka et al. 2018).

One issue all models must contend with is the bias-variance tradeoff, which is named for the two ways that a model’s predictions can vary from their “true” scores. A highly biased model underutilizes the available predictor data and the model is underfit. In other words, an unnecessarily simplistic set of relationships is estimated between predictors and the outcome, resulting in poor predictions. A model with high variance does the opposite in which the predictor data is overfit, and the model is too highly sensitive to the dataset from which it learned. This produces overly complex models and erroneously attempts to model error variance. The outcome from both of these scenarios is the same – poor prediction of the outcome of interest in future samples (Putka et al., 2018; Yarkoni & Westfall, 2017). This issue creates a key difference from traditional explanatory statistical models, such that a model must be “tuned” to work towards the optimal balance of bias and variance.

Model Types

One of the most commonly used methods of ML is ordinary least squares (OLS) regression, where the goal of the algorithm is to minimize the differences (measured by the sum of squared deviations) between the observed scores and the predicted scores (Yarkoni & Westfall, 2017). Many other ML algorithms build off and add to this basic OLS function as OLS on its own is poor in predicting outcomes and difficult to interpret (Zou & Hastie, 2005). For example, Zou and Hastie (2005) proposed a method known as "elastic net" regression, in which previous components of other regression iterations, ridge and lasso, are combined to handle the multicollinearity (i.e. high correlations between multiple variables in a regression model; Kuhn & Johnson) in predictors that may exist as well as incorporating variable selection features to create a parsimonious model. Ridge models keep all predictors in the model and penalize (reduce the weight attributed to a predictor) regression coefficients based on their multicollinearity. When a variable is comparatively less important in predicting the outcome and/or high multicollinearity, the regression weight is pushed toward zero. Models that are overfit, meaning the predictors are too overweighted to the dataset they are trained on, are less likely to generalize well to a new dataset. Thus, penalization is always a reduction in regression weight. A final ridge model is likely to have clusters of related items with the regression coefficients more evenly distributed across them rather than one item chosen arbitrarily with an inflated regression weight (Putka et al., 2018, Zou & Hastie, 2005). Lasso models are largely similar but automatically select predictors for inclusion in a final model by allowing the regression weights to become zero, rather than just near zero. However, lasso models are not as successful at dealing with sets of multicollinear predictors (Putka et al., 2018).

Elastic net is a combination of both ridge and lasso regression (Zou & Hastie, 2005). Elastic net is like OLS in that the model provides regression coefficients where the predictors are linearly related to the outcome but adds the ability to tune parameters (add penalties to regression weights) which increases model generalizability. Tuning parameters are empirically informed model features that can scale the model's complexity and/or put constraints on model estimates, making it less sensitive to a particular dataset. Elastic net also incorporates the variable selection feature from lasso regression. The addition of tuning parameters is a critical difference between OLS and elastic net as fitting a model becomes a two-step process, first finding optimal values for the tuning parameters to maximize generalizability and then model estimation (Putka et al., 2018).

One limit of OLS regression and elastic net is capturing non-linear relationships and interactions, which are often not captured in organizational sciences (Putka et al., 2018). Tree-based methods, often called classification and regression trees (CART), are more equipped to model these complex relationships. CART models are created by identifying split points at specific levels of various predictors to create subgroups of the sample based on those split points. Each split becomes a node, at which a new split point may be evaluated in either the same predictor or a new predictor such that the final model resembles a tree. Ultimately, this process creates groups that are defined by a specific set of split points of identified predictor(s) that seek to meaningfully distinguish groups based on the outcome.

Tree-based algorithms must also manage the bias-variance trade-off and do so via tuning parameters (Putka et al., 2018). If a tree becomes fully grown, in that predictors are split as many times as possible, it can become overfit such that it is difficult to generalize beyond the training sample. Trees can be "pruned" where decision points that do not meaningfully contribute to

outcome prediction can be removed (Putka et al., 2018). However, basic CART models are not very predictive, and small changes in data can change the final tree substantially, impacting generalizability. To increase generalizability, ensemble methods can be used which combine multiple models to create a more powerful model (James et al., 2023). Random forests are an example of a more powerful ensemble model based on the weaker CART model. In a random forest, several iterations of the tree are created and then aggregated. Importantly, randomization impacts these tree iterations in two ways. First, each tree is estimated with a random subsample of the data. Second, only a random sample of the training predictors are used for each tree, which reduces the computational burden, redundancy of predictions, and addresses the complexity-parsimony tradeoff. Commonly, the predictions from a random forest model are the aggregation of predictions from 500 to 1,000 trees (Breiman 1996; Putka et al., 2018). Random forests can be especially helpful when dealing with correlated predictors as is common in psychology (James et al., 2023), but choosing a model should be driven by the features of the data. If the relationship between the predictors and the outcome is linear, then OLS or elastic net will likely outperform a tree-based method like random forest that does not rely on a linear structure (James et al., 2023).

Interpreting output

One of the cautions of ML use is that the results of some models may be difficult to explain or interpret. The issue of interpretability is referred to as the "black box" problem, where a model may be a strong predictor of a relevant outcome, but how the outcome is produced is unknown to the person who built the model (Yarkoni & Westfall, 2017). The "black box" problem led to calls for more explainable ML (Goretzko & Israel, 2022) with the underlying goal of transparency in how the model is making decisions. Difficulty with interpretation is not necessarily a property of predictive models as a whole, but rather a feature of some specific

approaches (even regression coefficients cannot be straightforwardly interpreted; Yarkoni & Westfall, 2017). However, with tools that help with interpretation, there is less of a divide between prediction and explanation.

For example, in a dataset with many predictor variables, partial dependence plots are a way to visualize the influence of individual predictors holding all other predictors constant as well as determine if there are deviations from linearity (Friedman, 2001). Similarly, the relative importance of a variable can be calculated, where a zero to 100 score represents the magnitude of a predictor's impact on the model and can be interpreted similarly to how one might interpret a standardized regression weight (Johnson & LeBrenton, 2004). In many cases, however, one variable is likely to impact another, creating an interaction effect. In predictive models, an interaction between predictors cannot be expressed as the sum of the individual effects as the effect of one predictor may depend on others. Friedman's H statistic can be calculated to measure the interaction between two predictors or between a predictor and all other predictors. The H statistic indicates the amount of variance explained by the interaction(s) with zero being no interaction, a one meaning each function is constant and the only impact on predictions happens when they interact, and a statistic larger than one indicating a high amount of variance within the interaction (Molnar, 2022). Even models that can account for interactive effects do not explain the nature or extent of an interaction, making the H statistic relevant to understanding the underlying relationship of predictor variables. Collectively, these techniques can be applied to capitalize on the strengths of predictive models while also producing results that facilitate explanation.

Defining Income

As noted in Sayre and Conroy (2023), there are various ways to describe the money an individual earns, and scholars may use different definitions. For this paper, I use wages, earnings, and income (but not compensation; Lazear, 1986) interchangeably to represent the money received through employment, not including other forms of what could be considered income, such as subsidies. I have made this choice for several reasons. First, general nomenclature interchanges these terms; for example, "wage inequality" and "income inequality" are generally understood to be the same. Second, many income researchers interchange the terms or use them as comparable (see Fulmer et al., 2023). Third, this definition is closely aligned with how the American Community Survey (ACS), a portion of the U.S. Census, defines income. The ACS defines "wage and salary income" as including " total money earnings received for work performed as an employee during the past 12 months. It includes wages, salary, Armed Forces pay, commissions, tips, piece-rate payments, and cash bonuses earned before deductions were made for taxes, bonds, pensions, union dues…" (U.S. Census Bureau, 2022, p. 89). When citing a source that differentiates from this definition, a specific indicator (such as salary), will be used.

Industrial, Organizational, and Occupational Predictors

Global or national events, like a recession or the COVID-19 pandemic, can also affect wages at a market level. Wages generally increase year over year to keep up with inflation, but with a volatile economic market, this wage growth can stagnate or drop. At the peak before the Great Recession, wage growth was at 4.3% in August 2007 and plummeted to 1.6% in January 2010. A drop in wage growth was also observed from 2019 to 2021, attributed to the COVID-19 pandemic (Federal Reserve Bank of Atlanta, 2023). If the wage growth rate changes, it does not mean that wages decline year over year but rather that individual wages will not rise at the same

rate as inflation, impacting their overall purchasing power. With wage growth not changing in step with inflation, it requires an individual to spend a larger proportion of their wages on things like groceries and household items that have risen in price.

Another factor affecting wages is the individual's industry of employment. Economic theorists suggest that inter-industry wage differentials are driven in part by occupational structure, the collective bargaining power of the workers in the industry, and the profitability of the industry (Mokre & Rehm, 2020). 2023 data from the Bureau of Labor Statistics (BLS) of average hourly earnings for all employees within an industry demonstrate inter-industry gaps, even among industries that employ similar numbers of workers. For example, the "Informational" industry has average hourly earnings at $48.21, while "Transportation and Warehousing" is $29.37, and "Leisure and Hospitality" is $21.30 (BLS, 2023). Some of the variance can be attributed to job characteristics, as a data scientist working in the "Information" industry likely makes more on average than a housekeeper in "Leisure and Hospitality". However, inter-industry wage differentials have not been fully explained by skill differences, labor productivity, or other individual factors (Mokre & Rehm, 2020) and thus remain relevant predictors of wages even when accounting for the aforementioned individual factors.

Wages also tend to increase with the progression in occupation level. While levels can be reported differently, generally they can be broken down into entry-level, intermediate, and experienced. Job levels are relevant to wage differences due to four job factors: necessary knowledge to perform the job, differences in controls and complexity, nature and purpose of contracts, and the environment in which the job is performed. For example, the average hourly wage for an entry-level loan offer is $25.01, intermediate is $36.82, and experienced is $52.72 (Allamani et al., 2022). In principle, the progression between levels exists to compensate

employees per their experience, credentials, or complexity of tasks. A study by Mainert and colleagues (2019) demonstrated a .46 correlation between salary and job level, indicating that job level is positively related to wages.

Characteristics of a job or occupation itself, like whether it is full-time, part-time, or unionized, impact wages as well. Differences in wages between part- and full-time workers extend beyond just annual earnings. Part-time workers with similar demographics and educational levels receive 29.3% less hourly than those who work full-time. Even when controlling for industry, occupation, education, and other demographic characteristics, the difference is still 19.8% (Golden, 2020). Similarly, workers who are covered by a collective bargaining contract or a union earn 13.6% higher wages than nonunion workers who are comparable in industry, occupation, marital status, education, and experience (Mishel, 2012). Accordingly, job characteristics account for some of the variance in wage differentials not explained by other individual or market characteristics.

Individual Predictors

While labor and organizational-level predictors impact wages, individual-level predictors likely account for the largest source of variance in wage differentials. One of the most common ways individuals can influence their wages is through education. Broadly, those with more years of education and advanced degrees earn more and have higher job levels than those with less education and fewer degrees (Mainert et al., 2019). The median weekly earnings stands at $520 for individuals without a high school diploma, rises to $1,173 for those with a Bachelor's degree, and further increases to $1,743 for individuals with a Doctoral degree. Those with the highest level of educational attainment (Doctoral and professional degrees) earn more than three times those without a high school diploma (Torpey, 2018). Field of study and degree type also impact

wages. Men who received a Bachelor of Arts in Business had gross lifetime earnings of $2,308,989, while an Education degree yielded $1,566,094, and a degree in a STEM field averaged $2,797,436. This difference in wages between areas of study grows even further when comparing graduate degrees, as men with graduate degrees in Business earned on average $3,027,151 over their lifetimes while those with Education graduate degrees earned $1,955,046 (Kim et al., 2015). Thus, both an individual's level of education and field of study contribute to predicting wages.

An individual's skills and abilities also contribute to differences in wages. Cognitive ability or general mental ability (GMA) has been investigated in relation to numerous organizational outcomes. GMA captures ability across a variety of cognitive tasks and several studies have found GMA, or related abilities like complex problem-solving, to be correlated with pay (Lang & Kell, 2020; Mainert et al., 2019). Even in occupations where bachelor's degrees are not required, cognitive ability factors like complex problem-solving, critical thinking, judgment, and decision-making are related to earnings (Yerger, 2017). Soft skills, or skills related to how you interact and work with others, are also related to wages. Ferris and Witt (2001) demonstrate that social skills, which were not correlated with GMA, were related to salary. Similarly, a study of leadership skills and wages found that individuals who exhibited a propensity for leadership in high school earned significantly more 10 years later, controlling for family background, cognitive ability, and school characteristics (Kuhn & Weinberger, 2005). Consequently, skills and abilities likely greatly contribute to wage differences across jobs, as individuals with higher skill and ability levels can attain higher skilled and thus higher paying jobs.

Workplace performance is also related to wages. Many organizations have a pay-for-performance model such that employees receive wage or salary increases for performing well or

having a high-performance rating. This is also referred to as merit-based pay. In a study of one large organization over many years, Castilla (2008) reported statistically significant salary differences between those who received different performance evaluation ratings. Those who received a rating of "requiring improvement" received salary increases of 1.7% less than average performing employees, while those judged as good received an increase of 1.4% higher than average performers. As expected, outstanding performers received increases of 2.4% higher than average performers. Thus, individuals who are higher than average performers generally earn more than lower performers.

Several personality traits, which are innate patterns of thinking and behavior, have also been investigated in relation to wages. The most commonly researched personality variables are the Big 5: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Previous research has demonstrated that employees with a specific configuration of these traits earn more money. This correlation is known as "the big-five salary link" (De Haro et al., 2020; Nyhus & Pons, 2005). A meta-analysis by Ng and colleagues (2005) reported that Neuroticism and Agreeableness were negatively correlated with salary while Conscientiousness, Openness to Experience, and Extraversion were positively correlated. It's theorized that these personality traits may lead an employee to work harder, have higher self-efficacy, and set career-advancing goals leading to them earning more money (De Haro et al., 2020; Spurk & Abele, 2011). This theory is bolstered by the finding that individuals higher in "grit" (who are persevering at work) and self-esteem, are more likely to receive a higher income (five percent and eight percent respectively), than those who are lower in those traits (Albandea & Giret, 2018). In the same vein, higher core self-evaluations (CSE) which include self-esteem, generalized self-efficacy, locus of control, and neuroticism, are positively related to higher

salaries (Judge et al., 1997; Cheung et al., 2016). Finally, risk-taking has emerged as a personality variable of interest as risk-taking is a determinant of wages across income distributions, with risk-takers earning more money on average (Albandea & Giret, 2018). With the multitude of personality traits found to be related to wages, it is clear that an individual's traits are a pronounced component of wage differences.

An individual's country of origin is also related to wages. In an investigation of one large North American organization, non-U.S.-born employees made 4.5% less money than U.S.-born employees and received smaller pay increases (Castilla, 2008). This pattern is reflected on a national level as well. In 2021, the median hourly wage for U.S.-born workers was $21.50 compared to $20.00 for foreign-born workers (Kochar & Bennett, 2021). The specific country of origin is an important component of differentiating how much wages vary for immigrants, with immigrants from Europe making almost double what immigrants from Central/South America and Asia make (Duleep & Regets, 1998). Levels of education, industry of work, and skill transferability also add to the variability, but even holding the variables commonly used in projections of individuals' wages constant, immigrants and U.S. natives had different earning profiles (Duleep & Regets, 1998). Differences between specific countries of origin remain to be fully parsed out, but overall, individuals who are born in the United States are likely to earn more than immigrants.

Disability is a class legally protected by the Americans with Disabilities Act but still appears to impact wages. In 2019, the U.S. Census Bureau reported that full-time workers with disabilities earn 87 cents on the dollar compared to those without disabilities. Some sources report that this disparity is greatly reduced when accounting for industry or job type (Day & Taylor, 2019), however, this likely does not account for the nuances of regulations surrounding

the employment of disabled people and that disabled people are more likely to work in low-wage industries in general (Longhi, 2021). For example, employers can apply for a government certificate waiving the federal minimum wage so that they can employ disabled people for subminimum wages (Friedman & Rizzolo, 2020). The type of disability an individual has is also impactful in determining wages, with smaller gaps for those diagnosed with conditions like depression in comparison to those with learning disabilities or neurological disorders (Longhi, 2021). Similarly, the wage differential is larger for those with mental disabilities rather than physical disabilities (Metcalf, 2009). The grouping of all these conditions under the category of "disability" likely obscures drastic differences in wages for many disabled workers, but there is strong evidence that at a minimum, some types of disabilities are significantly related to earnings.

Many innate characteristics of people are related to wages. Race and gender, specifically, have been the topic of research for many years. Differences in wages between individuals of different races appear across many conditions. While wage gaps have narrowed over the years, Pew Research Center (Patten, 2016) reported that college-educated Black and Hispanic men as well as Asian women, only earn about 80% of the hourly wages of college-educated white men, while college-educated Black and Hispanic women only earned about 70%. These gaps are in part due to compounding differences in salary increases. Even among equally performing employees, in one study, Black and Hispanic employees were found to have received salary increases 0.5% lower than white employees (Castilla, 2008). Ultimately, while discrimination based on race is illegal, an individual's race does appear to be related to the wages they receive.

In a similar vein to race and disability, the relationship between gender and wages has also been investigated. Overall, those who identify as women tend to earn less than men, and this

pattern is seen across levels of education and industry. In 2022, American women made 82 cents for every dollar earned by American men, on average. This pay gap has persisted over time, even as women are more likely than men to graduate from college (Kochhar, 2023). Not only are women offered lower starting salaries with the same qualifications (Säve-Söderbergh, 2019; Schuster et al., 2022), but the rate at which their wages grow is less than men's, even when controlling for workplace performance (Castilla, 2008). Some may attribute this wage difference to more women working part-time; however, wage gaps between men and women also exist in part-time positions (Houtrow et al., 2020). This relationship is not necessarily attributable to differences in promotion either, as Ng and colleagues (2003) reported in their meta-analysis on predictors of career success that men and women are promoted at almost the same rates but women are still paid less. Like race, discrimination based on gender is illegal but the link between gender and wage differences is evident.

Many of the aforementioned predictors of wages do not operate in isolation and may interact to create compounding effects. Previous research has also shown that gender impacts whether or not marital status is a predictor of wages. In the case of men, a meta-analysis demonstrated there is a marriage premium of about 9-12% in the United States when accounting for other confounding variables such as years of experience, age, parental status, and education (Leonard & Stanley, 2015). However, the same pattern is not seen for women and seems to depend more on other variables, such as race, age, and education. Married white women earn more until about their 50s when single white women begin to accrue larger annual earnings. In comparison, the annual earnings of married Black women are consistently higher than single Black women across age groups (Federal Reserve Bank of St. Louis, 2020). In the case of education, in general, married women who hold doctoral degrees appear to receive a salary

penalty. In contrast, married men receive a salary premium, with single women having higher salaries than married women (Weber & Canche, 2015). It's clear that the marriage wage premium/penalty does not impact everyone the same but impacts intersecting identities differently.

As previously discussed, there are differences in wages between part-time and full-time workers, but other characteristics either widen or narrow this gap. When the gap is adjusted for gender, men face a 25.8% penalty while women have a 15.9% penalty, meaning men who work part-time are penalized more than women. When considering race, this gap widens further with white men facing a 28.1% penalty, while the penalty for Black men is 24.6%, and 16.9% for Hispanic men (Golden, 2020). This runs counter to the other findings reported above where white men generally outearn women across conditions, highlighting the importance of considering the impact of multiple intersecting identities when investigating wage differentials. While the impact of every identity on wages cannot be covered in this literature review, it is important to acknowledge that many of the variables that influence wages do not always act independently.

Current Study

The current study aims to use a variety of techniques (e.g., elastic net, random forest) to determine which, if any, predictors of wages present in a large archival dataset (U.S. Census Bureau, 2017) have not been captured in the literature. The current literature on wages generally relies on explanatory modeling techniques like ANOVAs (Fossum & Fitch, 1985), structural equation modeling (Spurk & Abere, 2011), and general estimating equations (Castilla, 2008). Due to this focus on explanatory modeling, a limited number of predictors have been identified as they are drawn from theory (see Table 1), and it is unknown whether a larger and more

encompassing set of predictors evaluated via predictive modeling may result in a new understanding. Thus, the research questions are:

1. Do machine learning techniques result in improved prediction over ordinary least squares (OLS) regression?
2. Are there differences in algorithm utility?
3. Which variables are the most relevant predictors of income and are there variables in that set of predictors not already represented in the literature?

METHOD

Participants

Archival data from the American Community Survey (ACS) from years 2018 and 2019 were used as the training and test data. The ACS is a nationwide survey from the U.S. Census Bureau, which was formerly the "long" form data collected during each census. Since 2000, the long form of the Census became the ACS, and data is collected yearly. The Census Bureau contacts more than 3.5 million households yearly to participate, gathering "timely social, economic, housing, and demographic data" (U.S. Census Bureau, 2017). The criterion variable of interest is "Wage and salary income" in which the respondents report their total pre-tax wage and salary income over the past 12 months on a continuous scale. ACS data can be extracted from the Integrated Public Use Microdata Series (IPUMS) USA which provides Census and/or ACS data to the public in microdata form (Ruggles et al., 2023).

Measures

Variables created by the ACS for weighting, respondent linking, redundant variables, and questions pertaining to older versions of the survey were removed, yielding a total of 108 variables for inclusion. Both the 2018 (training) and 2019 (test) datasets include over three million participants. Categories of data collected can be split into eight subcategories: demographics, location, education, finances, resources, household information, workforce participation, and health.

Demographics

Demographic variables include variables such as age, sex, race, birth year, ancestry, veteran status, language spoken, if the participant has been divorced, whether or not they are part of a same-sex couple, year immigrated and naturalized, and identified tribe.

Location

Data gathered on the respondent's location include variables such as the state and county in which they reside, the location's metropolitan status, the city population, and where the individual lived a year ago.

Education

Variables encompassing educational information include whether or not the participant is currently in school, educational attainment, grade level attending, public or private school, and two items for the field of degree.

Finances

The ACS includes substantial information about the participant's financial status including their mortgage payments, property taxes and insurance, rental costs, utility costs, poverty status, and several measures of types of income (e.g. social security, welfare, interest, wages). Total income, the dependent variable, was an open-ended question where respondents answered "What was this person's total income during the PAST 12 MONTHS?". They could also report "None" or "Loss".

Resources

Participants reported the resources they had access to within their homes such as a telephone, type of internet access, smartphones, computers, cooking facilities, hot and cold water, and available vehicles.

Household Information

Respondents provided information about the household in general and those who live there. Variables include whether it is a family or non-family household, whether the inhabitants are married, cohabitating, or other living arrangements, house acreage, age of the structure, number of bedrooms, number of subfamilies or couples in the household, whether or not it is multigenerational, and number and age of children in the household.

Workforce Participation

Participants indicate workforce participation related to time to travel to work, time of arrival at work, place of work (city/state/country), if they carpool, occupational prestige score, whether they are available or looking for work, usual hours worked last week, weeks worked per year, the industry of occupation, class of worker, and labor force status.

Health

Several variables describing the participant's health include edicare status, type of health insurance (e.g. public, private, VA), whether they have difficulty living independently, and reported disabilities (e.g. hearing, cognitive, vision).

RESULTS

Data Processing

Before analysis, variables with near-zero variance were dropped from the dataset which helps improve model stability and performance (Kuhn & Johnson, 2013). The cutoff ratio for near zero variance of the most common value of the variable compared to the second most common value was set at 30. Variables with redundant information were excluded. Other variables such as household weight and strata, are also included in downloads from IPUMS, so they were removed from the dataset before analysis. Finally, cases where the criterion was unavailable or was equal to or less than zero were removed.

Wages were log-transformed as in prior research (Kim et al., 2015, Nyhus & Pons, 2005) as the transformation reduces the skewness of a variable (West, 2022). This transformation was chosen over other methods, like winsorizing, which changes or removes outliers assuming that they are contaminants to the data (Pek et al., 2018). In the case of this dataset, outliers are not contaminants but representative of a wide wealth distribution. Other model distributions would have been possible given the data (e.x. GLM). The following analytic strategy was chosen because the chosen models were available in the caret package and log transformation allowed the data to meet the assumptions needed for those models.

Within the data, there were several high-cardinality variables or categorical variables with many levels but without clear ordering. Variables of this nature can be difficult for ML algorithms to understand as they are meant to interpret numerical inputs (Pargent et al., 2022). Target encoding reduces the levels within a variable by making a prediction of the outcome for each level of the variable (Pargent et al., 2022) without increasing data dimensionality (Nazyrova et al., 2022). In this case, the mean income was calculated for each level within a variable and

replaced the categorical variable. For example, within the Occupation variable, the mean income was calculated for each level (ex. Natural science managers, management analyst) and replaced every instance of the categorical level. Target encoding was used to modify the variables: Place of work (state), Birthplace, City, County, Degree field, Industry, Language, Occupation, Place of work (county), Place of work (metropolitan status), State, Tribe, and Ancestry (first response). Variables that would be considered high cardinality but were dropped due to zero or near zero variance were not recoded.

Research has shown that target encoding works best for variables with more levels (Pargent et al., 2022) so other categorical variables that had more than five levels but fewer than ten were dummy coded. These variables were whether there were grandchildren in the house, Hispanic origin, marital status, and citizenship. The variable year married was also modified as it included those who had never been married as zero, yielding a value of zero that was not comparable to calendar years. A new variable, Years Married, was created by subtracting the year married from the year the data was collected. This better represents the length of time someone had been married and yields a value of zero (to reflect never having been married) that is comparable to the other variable values. After variable exclusions and recoding, the total number of predictors was 109. See Table 1 for all variable information.

Data from 2018 served as the training set and the 2019 data as the test set. In the training data, a 10-fold K-fold cross-validated repeated three times was used, which is outlined in Putka et al. (2018). This technique is used to both evaluate model performance and select hyperparameters in the ML algorithms to estimate models with the 2018 data. Models trained on the 2018 data were then used to predict salary using the 2019 data to evaluate the generalizability of those models.

Analysis

Prediction

To determine whether machine learning techniques result in improved prediction over OLS regression, OLS regression, elastic net regression, and random forest regression were estimated using the 2018 data and all predictors. For all algorithms, R^2 can be used to quantify the proportion of variance explained of income. The entirety of the dataset (n = 2,348,223) was used to estimate the OLS and elastic net regression, however, due to the computing power needed, a random subsample of 20,000 was used to estimate the random forest. The OLS regression returned an R^2 of .58 ($p < .001$), the elastic net an R^2 of .58, and the random forest with an R^2 of .60.

The same three types of models were then estimated only using variables that had previously been found to be related to income in the literature. This smaller dataset included 42 predictors (see Table 1 for all variables). The same subsampling method was used for the random forest as described above. R^2 was smaller across all models (OLS = .50, elastic net = .50, random forest = .55) indicating that the additional 54 predictors included in the larger dataset accounted for a change in R^2 of .06 to .08 depending on the model used. As such, answering research question 1 random forest does provide increased prediction over OLS regression with random forests providing the best prediction regardless of dataset used.

Determining whether there are differences in algorithm utility can be evaluated by how well the models trained on the 2018 data apply to a new dataset. In the test data, the observed R^2 can be used to assess how well models formed on the training dataset apply to the test dataset. In alignment with the a priori decision rule to proceed with the reduced dataset (i.e., variables previously found to predict income) if those models yielded R^2 values within .01 of models using

all predictors, analysis of generalizability proceeded with the models trained on all predictors. This .01 threshold is based on a minimally acceptable difference as shown in prior work (Schmidt & Hunter, 1998). When using the model built on 2018 data to generalize to 2019 data, the OLS model returned an R^2 of .45, the elastic net.47, and the random forest .49 (Table 2). As for research question two, an initial indicator of utility is the quality of predictions made with new data. Given these results, the predictions made from the random forest model are more useful as they explain more variance than those from the elastic net and OLS models.

Model Interpretation

The following steps were then used to evaluate which predictors are the most important predictors of income. Using the full dataset, the relative importance function in the R package *caret* was used to calculate relative importance (i.e., the predictor's influence on the resulting prediction) of each predictor within an algorithm (Kuhn, 2008). The top ten predictors in the OLS model were 1) weeks worked last year (intervalled), 2) whether the individual worked last year, 3) usual hours worked per week, 4) occupation, 5) age, 6) sex, 7) education, 8) class of worker, 9) industry, and 10) school type. In the weeks worked variable, the measurement is given in intervals (1-13 weeks, 14-26 weeks), instead of the precise number of weeks. For the elastic net model, the top ten were 1) whether the individual worked last year, 2) the state they live in, 3) occupation, 4) sex, 5) weeks worked last year (intervalled), 6) county they live in, 7) class of worker, 8) school type, 9) industry, and 10) marital status (if someone was widowed). Finally, the top ten most important predictors of income in the random forest model were 1) weeks worked last year (intervalled), 2) usual hours worked per week, 3) occupation, 4) age, 5) industry, 6) education, 7) degree field, 8) years married, 9) place of work (state), 10) ancestry. Figures 2-4 show the variable importance plots across models for comparison.

Partial dependence plots (PDPs) were then created using the *pdp* package in R (Greenwell, 2017) of each of the top ten most important predictors from the random forest model, which visualize the linearity (or lack thereof) of the predictor and income relationship, as the random forest model is the only model to permit nonlinear relationships of the algorithms evaluated here (Putka et al., 2018). All plots are included in Figures 5-14. Based on the PDPs, eight variables exhibited nonlinear effects suggesting that subsequent OLS and elastic net models could benefit from additional terms added into those models. A quadratic term was deemed appropriate for weeks worked last year (intervalled), usual hours worked per week, age education, degree field, place of work (state), years married, and ancestry. Additionally, cubic terms were deemed appropriate for place of work (state) and ancestry.

The top ten predictors from the random forest model were also evaluated using the H statistic to better understand interaction effects. Interactions with an H statistic of .01 or greater (indicating the share of variance attributable to a given predictor that is shared with another predictor) are reported in Figure 15. The *hstats* package (John & Mills, 2023) produces the strongest overall interactions with an individual variable, the joint effect variability of the strongest pairwise interactions, and the variance unexplained by the sum of all main effects. The variables with the highest prediction variability due to interactions with other variables were weeks worked last year (intervalled), usual hours worked per week, age, and each accounted for between three and six percent of the model's overall variance explained. Variable interactions that exceeded the .01 threshold were between age and usual hours worked per week (.02), age and weeks worked last year (intervalled) (.02), and weeks worked last year (intervalled) and usual hours worked per week (.01). The total interaction strength (H^2) was .32, meaning 32% of the prediction variance in the model is unexplained by the main effects.

The three interactions that exceeded the .01 threshold were then added into the data along with the quadratic and cubic terms mentioned above, and OLS and elastic net models were re-run with the goal of retaining the fewest number of predictors within an R^2 of .01 of the original models. With the top 20 most important variables retained, the R^2 of the OLS model was .57 and the elastic net was .51. When again generalized to the 2019 dataset, the R^2 for the OLS model was .45 (i.e., the same level of predictive performance as the original model with over 90 predictors). When this strategy was applied to the elastic net model, the R^2 was .25, which was a sharp decline in performance for the elastic net model with fewer predictors.

Variable importance was then recalculated for the final OLS model with the top ten most important predictors being 1) weeks worked last year (intervalled), 2) whether someone worked last year, 3) age, 4) usual hours worked, 5) occupation, 6) education, 7) sex, 8) when the occupant moved into their residence, 9)if they are current married with a present spouse, 10) and class of worker.

In summary, the machine learning methods (elastic net and random forest) outperformed the OLS regression model when generalized to the test data set. The edge of elastic net over OLS was not seen when considering just the training set (RQ1). In all cases, random forest outperformed both elastic net and OLS, indicating differences in algorithm utility (RQ2). Finally, while many of the important predictors have been identified previously in the literature, novel predictors of income and non-linear variable relationships were identified (RQ3). These results will be discussed in detail below.

DISCUSSION

The purpose of this study was to investigate the effectiveness of a novel prediction methodology using data from the American Community Survey. Study results suggest that generally, machine learning models do outperform OLS in terms of prediction with these data. Further, this greatly improves upon other studies predicting income with an R^2 of .49 in a holdout set of data compared to .12 using explanatory methods (Spurk and Aberle, 2011). The data used in the present work also lacks the personality predictors incorporated in Spurk and Aberle (2011), yet still provides greater predictive utility. It is likely that given a data set that includes a greater number of predictors from the literature that model performance would perform even better. The methodology presented here can overcome some of the downfalls of more traditional methods by incorporating many more variables, a large sample size, and using models able to account for the non-linearity of predictors.

In both the full test and reduced training set of data, OLS and elastic net performed similarly. However, when the full model was generalized to a new dataset, the elastic net model outperformed the OLS model by an R^2 difference of .04. The random forest model outperformed OLS and elastic net in both the training and the test sets. This indicates there are differences in algorithm utility as the random forest was able to account for more variance even when using fewer cases (i.e., the random subsample of $N = 20{,}000$).

Surprisingly, the elastic net model did not outperform the OLS model in all cases. When the dataset was reduced to just the top 20 most important predictors, the OLS model outperformed the elastic net model in the 2018 training dataset, explaining 6% more variance. This difference was greatly exacerbated when the reduced models were generalized to the 2019

data, as the OLS model then explained 20% more variance, indicating the elastic net model was overfit to the 2018 data. Comparing these results to the initial full model results suggests that the elastic net performs better with a greater number of predictors. This may be because of elastic net's inherent variable selection features (Putka et al., 2018) which may produce unnecessary modeling complexity given a reduced set of the top twenty performing predictors and a large enough sample that many of the drawbacks of OLS regression are avoided (e.g., overfitting, multicollinearity; Kuhn, 2008).

While the newer machine learning methods did generally outperform OLS regression, it is important to contextualize what a difference of .04 means practically (OLS compared to random forest in the 2019 data). This change may seem small, but it does not suggest that the same results would be found if you were to use OLS and random forest on any other dataset, rather than in a data set with more than 2,000,000 more cases, the OLS model came close to the performance of the elastic net. The strong performance of OLS is due to the large sample size the model was trained on and it likely would not have the same level of success in a more common sample size of a few thousand individuals or less. These results should not just be considered in the context of which algorithm was used but the size of the sample as well, as more data almost always leads to more accurate predictions.

Considering the nature of the predictor variables themselves also provided new insight. Using the *hstats* package also allowed for the identification of important variable interactions that were previously not mentioned in the literature. Both age and variables related to the amount of time spent working appeared in two of the three interactions accounting for over 1% of the model's variance explained. To better understand the interactions, stratified partial dependence plots were created to visualize the nature of the interactions. For age and weeks worked last year

(intervalled), older workers make more than younger workers, but the relationship rises steadily at first and then levels off when individuals are working more weeks per year (Figure 16). Similarly, for age and usual hours worked, older individuals are paid more with a mostly smooth relationship until a jump where hours transition from part time into full time (Figure 17). Finally, in the interaction between usual hours worked per week and weeks worked last year (intervalled), there is a weaker relationship between weeks worked and pay, while those who work more hours per week see a greater benefit to working more weeks per year (Figure 18).

Another nuance of the predictor variables was that the partial dependence plots indicated that eight of the top ten variables identified in the random forest model were non-linear in nature. Thus, using these two investigation techniques provided new information about how income is predicted that would have been difficult to come by using more traditional methods.

Variable Importance

In terms of which predictors were most important in predicting income, many of the results of this study replicate previous findings from studies of wage prediction. Considering the variable importance indices, weeks worked last year (Golden, 2020), occupation (BLS, 2023), and industry (Mokre & Rehm, 2020) were present in the indices of all three models which is consistent with the literature. Several variables were also present in two model importance indices. Age and usual hours worked per week appeared in the OLS and random forest models, and location-based variables appeared in both elastic net and random forest. Elastic net models can handle multicollinearity well which may be why two location predictors (state and county) were both important predictors. However, it is likely the best location-based predictor of income is curvilinear (place of work (state)) and that may be why the random forest model was able to identify it within the model. The ability to detect non-linear predictors is also likely why years

married, ancestry, and degree field were only identified within the random forest relative importance indices as they all were identified as non-linear through the partial dependence plots. Elastic net models may also reduce the regression weights of highly correlated or redundant variables, which may be why age and location were not in the model's top ten most important predictors but were included in both the OLS and random forest models (Putka et al., 2018). When considering the strengths of each model type and the information gathered from the PDP's, it is clearer why there are differences in variable importances across models.

When considering which variables are the most important predictors of income, the random forest model gives the most information as it allows for non-linear predictors. The relative importance indices from the full random forest model showed that usual hours worked and weeks worked last year (indicator of part-time workforce participation; Golden, 2020), occupation (BLS, 2023), education (Mishel, 2012), industry (Mokre & Rehm, 2020), age (Cardoso et al., 2011), degree field (Kim et al., 2015), and state (Frank, 2007) comprised eight of the top ten predictors of income, all of which have been reported in the literature as important predictors of income. Income rises steadily until people are in their 50s and then starts to decline. For those who had some schooling but did not finish high school, the average income was about the same, but with a high school degree, income steadily increased with additional years of education. It is also logical that someone working as an engineer probably makes more than a clergyperson given their occupations within their industries and degree fields. Weeks worked last year, and usual hours worked per week followed a consistent predictable pattern as well, with the more weeks or hours worked, the higher the average income. Once usual hours per week reached about 50, the increase leveled off, likely due to the individuals who work over 40 hours a week

being on a salaried schedule where more work does not equate to more pay. As such, many of the findings of this study are in alignment with previous literature.

One predictor that has not received much previous attention in the income literature but was one of the top ten predictors in the random forest model, is years married. Previous studies have investigated the marriage premium (Leonard & Stanley, 2015), but years married have not been previously considered to be an important predictor of income. Age is likely related to how long someone has been married, as generally people get married around the age of 28 (U.S. Census Bureau, 2021). However, if years of marriage were just a proxy for age, the random forest model would have been able to avoid the use of redundant predictors, suggesting there is something within the identity of being married and the length of marriage that is related to income. Also, marital status related variables appeared in the elastic net relative importance indices as well as the indices from the final parsimonious model, suggesting that there is some relationship between being married and income. It may be that those who are married versus unmarried have different personality traits that are not captured in the larger ACS data, making marriage status and its duration a valuable predictor in the model (Hoan & MacDonald, 2024).

Previous research has demonstrated a relationship between marital status and SES. Couples who are at higher SES levels often get married later as they pursue education and have lower divorce rates, while those at lower SES levels are less likely to marry at all (Karney, 2021). Marriage rates also differ greatly by race and ethnicity, with 54% of white individual over 18 married compared to 61% of Asians, 46% for Hispanics, and 30% for Black individuals. Given the numerous variables that are relevant to marital status and how long they are married, more investigation into years married as a predictor of income is needed to understand why it is such a strong predictor.

The other variable not well captured in the literature but in the top predictors of random forest is ancestry. In the same way that years married overlaps with other variables in the dataset, ancestry likely shares variance with data set variables such as race and citizenship, yet ancestry is the variable that is highlighted as important. The ancestry variable is more granular than the race variable, with over 100 categories compared to less than ten, such that it is likely more informative to the model as there is more information available. Again, this indicates that there is some underlying relationship between ancestry and income that is not captured in the relationship between race and income.

However, there are many variables, such as race and sex, that have been touted as important predictors of income differentials that were not present in the top ten predictors in the random forest model. While the ACS dataset was not all-encompassing, many predictors that were included and have been researched extensively did not appear to be important predictors of income. Other than race and sex, examples include veteran status, disability, whether an individual spoke English, and immigration status. Study results should not be interpreted as suggesting that these variables are not related to income. Rather, it is likely that the predictors found to be the most important are more proximal to the outcome. Occupational segregation by gender is related to education, suggesting there are differences in the industries and occupations cisgender men and women choose to pursue (Blau et al., 2013). Similarly, someone's disability, whether they spoke English, or other identity-based characteristics may impact access to education or someone's ability to work full-time.

Many previous studies have attempted to parse out the nature and proximal and distal income predictors. For instance, Casilla (2008) includes both race and gender variables as well as a more proximal predictor of income to examine if demographic differences still exist within the

model. Similarly, Kochlar (2023) examined if income differences persisted between men and women (more distal predictor) while accounting for differences in education (more proximal predictor). Other variables that may be more proximal predictors of income than were not included in the ACS data include cognitive ability (Yerger, 2017) and personality traits (De Haro et al., 2020). This study provides disparity researchers with a longer list of proximal variables to consider in future work in addition to those previously identified in the literature. For example, the findings of this study have not clarified why an income gap between men and women exists, but in addition to education, performance, part-time status, and promotions (which have been identified as contributing to the gap) this study has yielded additional information that would be relevant to the future study of gender disparities in income, such as years married.

Several predictors appear in the final parsimonious model's relative importance indices that were not in any of the prior models' top ten. Upon further investigation, these new variables (when the occupant moved into their residence and if they are currently married with a present spouse) are present in the variable importance indices, just outside the top ten. Changes to the variable importance indices in the OLS model come after non-linear variable terms were added into the model. This indicates that the added terms model accounted for variance that was overlapping with variables that were in the top ten in the variable importance indices. When the non-linear variable terms could account for more of the variance, some of the variables in the original top ten for OLS then dropped out. For example, a non-linear term added to the age variable may have overlapped with the variance explained by the school type variable, as that variable is indicative of whether someone is still in school, which often is related to age. With the new term explaining more of the shared variance, school type then dropped out of the top ten list, allowing a unique predictor to take its place. Considering this, changes in the variable

importance after variable modification suggests that these new predictors are accounting for something better than those who were bumped down lower on the variable importance list.

The variable importance findings highlighted here should inform both the substantive and control variables used in future studies. In addition to the variables already identified as predictors of income in the literature, years married should be considered in future models. While the underlying nature of the relationship is not understood, it still emerged as an important predictor over many other predictors, like gender. Given the variable importance indices and the interaction effects, weeks worked in the past year, hours worked per week, and age, should be included as control variables. The inclusion of these variables in models is likely to aid in prediction and help researchers better understand how the variance is distributed across included predictors.

Implications for Research

While this study has focused on prediction rather than theory building, theory is a necessary component of I/O. With the most important predictors of income identified, academics can incorporate findings into existing theories or develop new ideas on why these predictors are related to income. The novel predictors found in this study, such as years married and when an occupant moved into their house, are unlikely to be considered in studies not focused exclusively on prediction as their relationship to income does not seem to align with prior theory nor follow a logical basis (like hours worked per week). Developing theory surrounding these predictors is likely to encourage more investigation of the relationships found here, leading to a clearer overall picture of income prediction.

These results should also be considered when designing future income studies for control purposes. Spurke and Aberle (2011) investigated the effects of personality, motivation, and hours

of work per week on income, controlling for prior salary, occupational field, GPA, organizational tenure, gender, and training and skill development. Current findings suggest that variables like age, industry, education, degree-field, and state of the workplace should be added as controls as they greatly impact income. Controlling for variables known to be related to income in future studies will more clearly differentiate what variance can be attributed to the predictors of interest.

Implications for Practice

Unfortunately, the results of this study are not all directly applicable, as what is driving model performance is not usable in operational settings (ex. ancestry). However, people analytics practitioners often have access to large datasets, not unlike this study, with data at the individual-, team-, and organizational-level. Using this methodology (see also Putka et al., 2018), organizations can build more accurate and generalizable models to predict outcomes like turnover, job satisfaction, and performance. This type of modeling could also be used to assess how pay structures differ across organizational units, like teams or departments, to evaluate organizational fairness and alignment with compensation policies.

Organizations can also use this information to audit their own compensation systems to see how well their own predictors of income correspond with those reported here. This is not to say that organizations should strive to match the most important predictors described here, especially as several of the predictors identified are indicators of a protected class. However, it may be a useful exercise to compare how many protected classes appear in their variable importance indices so that steps can be taken to reduce the reliance on this type of information. Similarly, if variables like education or degree field are not strong predictors of performance, that might suggest an under reliance on this information given the findings of this study and the

literature on income. In an ideal world, income should be related to job-relevant factors. Organizations should not interpret these findings as suggesting that income should be related to any protected classes.

If organizations want to use this methodology to determine if there is equity across groups, careful consideration must be taken to ensure that all relevant predictors that should be related to income are included. If they are not, the true relationships may be obscured by more distal variables that are included (e.g. race and gender) rather than more proximal predictors that should be related to income (e.g. education, tenure). Considering the pay equity study literature would be valuable to organizations in determining the types of variables they need to capture (see Taylor et al., 2020). If all relevant variables have been included and protected class variables are still some of the most important predictors, then there is likely a pay equity issue.

Limitations and Future Directions

This study, while leveraging a large data set and a robust methodological approach, is not without limitations. First, the entirety of the ACS data relies on an individual to accurately report for themselves and others in their household. Self-report data can often be skewed by social desirability (Donaldson and Grant-Vallone, 2002), so participants may have overinflated variables like income and education, while underreporting variables like their inability to care for themselves. Self-reporting data on race and ethnicity can also be difficult as many individuals do not feel that the traditional “check a box” approach represents them or feel that they are a combination of different races and ethnicities. The ACS has updated the way it collected this data twice since the years the study data was collected (2020, 2024), suggesting the standard for measurement is still a moving target (Marks et al., 2024). Studies that use non-self-report data,

like internal company information, may find stronger relationships between the predictors and income.

The ACS dataset was chosen for this study due to its inclusion of strictly biographical self-report data, a thorough item development process, and detailed data quality checks (U.S. Census Bureau, 2022). This type of self-report data does not rely on the foundation of construct validity and thus removes some of the issues that arise when dealing with validity evidence. However, that means that constructs that are known to be important in predicting income like personality (De Haro et al., 2020), workplace performance (Castilla, 2008), and skills (Yerger, 2017) are not considered in this model. Therefore, the current model is incomplete as it lacks the inclusion of predictors that have been demonstrated to be related to income in previous research.

The levels at which variables are measured may also be obscuring the underlying nature of the variable and income relationship. For example, place of work (state) was one of the top ten most important predictors of income, but it is not clear how the state where someone works is related to income. It may be that large cities with high salaries and big labor markets are driving the prediction or that the tax structure across states creates incentives for different pay structures. While this is not a limitation unique to a study using machine learning, it is important to consider for future studies to be deliberate on how they measure variables of interest.

The goal of reducing the overall predictors while keeping the R^2 within 1% of the overall model was to demonstrate how the methodology employed here could maximize prediction while creating a simple model based on the most important predictors. In this study, a backward wrapper method was used to select predictions, meaning a full model was first created and then a measure of variable importance was used to rank predictors with non-important predictors removed (Kuhn & Johnson, 2013). However, in linear models, the inclusion of non-important

variables can hurt overall model performance, suggesting that methods of removing these variables may be beneficial (Kuhn & Johnson, 2013). Further, focusing strictly on parsimony though may not be beneficial. Fit propensity, a model's ability to fit a wider range of data, should also be considered when selecting a model. Models that can fit a wider range of data are less parsimonious as such there is tension between parsimony and fit propensity (Falk & Muthukrishna, 2023). With the goal of explaining the most variance, the final model reported here has heavily leaned toward parsimony and that choice likely reduces the model's fit propensity. The tension between these concepts cannot be resolved but a resolution must be chosen by the researcher depending on the goals of the study.

Previous research has shown the models chosen (elastic net, random forest) can predict and generalize better than OLS (Putka et al., 2018), but many other models are newer and more robust than the ones used here. Xgboost, for example, is a tree-based model that can combine weaker trees to create a stronger one and can handle both classification and regression problems (Sakar & Natarajan, 2019). While many other models may be equipped to answer the research questions here, a researcher in this field must make choices that consider the type of data, prior literature on the topic, and the models that apply to the research question. In the I/O literature, we have not fully applied the functionalities of the random forest model. Thus, it is unwise to go beyond that until the models that are the foundations of more complex models, like elastic net and random forest, are utilized. The benefit of the models used here is that they are all available within the caret package, so the syntax remains generally the same across models. The methods in this study were also chosen such that they could be easily replicated by individuals who may not have a background in machine learning.

Another limitation related to the selection of models is that the complexity of the random forest model required a reduced dataset to run. A sample of 20,000 was randomly selected without replacement to be large enough to be representative of the larger dataset. Not using the full dataset is a limitation, as the entirety of the sample of could be used for OLS and elastic net. However, a comparison of means and standard deviations in the full and 20,000 case dataset of the top twenty predictors from the random forest model showed that the reduced model was almost exactly equivalent (see Table 3), indicating the smaller sample set was representative.

While this study was successful in demonstrating the utility of machine learning methods in predicting income and identifying novel predictors, it has also created many avenues for future research. First, the ACS data captured many important predictors of income but hardly captures the literature fully, especially individual predictors like job performance, personality, and skills. Inclusion of more of these variables will likely increase model performance as well as help shed light on how these individual predictors compare in importance to those discussed here. Second, there are several predictors (i.e. years married) and variable interactions identified in this study that are not discussed in the current literature. Researchers should examine these findings at a more granular level to more clearly understand the underlying mechanisms driving these relationships. Finally, policymakers should consider how these results align with the current interventions for addressing income inequality and use the variable importance indices to inform where new areas for action may be.

Finally, these results should not be interpreted to suggest that the predictors described here are the best predictors of income in all scenarios, rather, that they are the best predictors of income as it is defined by the American Community Survey. The “total income” variable used as the dependent variable here was also log transformed, which resolves some of the issues with

non-linearity but reduces overall variable discrepancy such that income variables that differentiate top earners would be reduced in magnitude. For this study, it was the best approximation of income and was continuous which allowed for more variability in the outcome, but also has some downsides as it likely conceals some important aspects of income, such as whether someone was salaried or hourly which matters in determining how relevant hours worked is in predicting income.

Conclusion

An R^2 value is considered indicative of how much is understood about a phenomenon. With less variance explained (a low R^2), there is still much to be understood about the relationship between predictors and outcomes. Using elastic net and random forest models, the findings of this study greatly improve upon previous income prediction studies (Spurk & Abele, 2011), with R^2 above .40, even when generalized to a new dataset. Focusing on prediction versus explanation has allowed for a greater understanding of what variables are most important in predicting income. These increases in variance explained over other studies suggest we now have a clearer picture of how to predict income. While important personality and job-related variables known to be related to income were not included in this dataset, future studies can incorporate these predictors using the machine learning methods outlined in the paper to build upon the foundations laid here and continue to improve upon our understanding of predicting income.

REFERENCES

Allamani, J., Hudak, M., & Issan, A. (2022). *Introducing Modeled Wage Estimates by grouped work levels*. Monthly Labor Review, U.S. Bureau of Labor Statistics. https://doi.org/10.21916/mlr.2022.23

Baker, J. L., Rotimi, C. N., & Shriner, D. (2017). Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, 7(1), 1572–10. https://doi.org/10.1038/s41598-017-01837-7

Blau, F. D., Brummund, P., & Liu, A. Y.-H. (2013). Trends in Occupational Segregation by Gender 1970-2009: Adjusting for the Impact of Changes in the Occupational Coding System. *Demography*, *50*(2), 471–494. http://www.jstor.org/stable/42920534

Bryson, A., Buraimo, B., & Simmons, R. (2011). Do salaries improve worker performance? *Labour Economics*, 18(4), 424–433. https://doi.org/10.1016/j.labeco.2010.12.005

Bureau of Labor Statistics (2023*). Employment and average hourly earnings by industry*. https://www.bls.gov/charts/employment-situation/employment-and-average-hourly-earniNgs-by-industry-bubble.htm

Castilla, E. (2008). Gender, Race, and Meritocracy in Organizational Careers. *The American Journal of Sociology*, 113(6), 1479–1526. https://doi.org/10.1086/588738

Cheung, Y., Herndon, N. C., & Dougherty, T. W. (2016). Core self-evaluations and salary attainment: the moderating role of the developmental network. *International Journal of Human Resource Management*, 27(1), 67–87. https://doi.org/10.1080/09585192.2015.1042897

Conroy, S.A. (2019). Setting base pay rates: integrating compensation practice with human

capital value creation and value capture. Handbook of Research on Strategic Human Capital Resources.

Conroy, S. A., & Gupta, N. (2018). Disentangling horizontal pay dispersion: Experimental evidence. *Journal of Organizational Behavior*, 40(3), 248–263. https://doi.org/10.1002/job.2323

Conroy, S. A., Roumpi, D., Delery, J. E., & Gupta, N. (2022). Pay Volatility and Employee Turnover in the Trucking Industry. *Journal of Management*, 48(3), 605-629. https://doi-org.ezproxy2.library.colostate.edu/10.1177/01492063211019651

Day, J. C., & Shin, H. B. (2005). *How does ability to speak English affect earnings?* U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2005/demo/2005-Day-Shin.pdf

Day, J. & Taylor, D. (2019*) Do People with Disabilities Earn Equal Pay?* United States Census Bureau. https://www.census.gov/library/stories/2019/03/do-people-with-disabilities-earn-equal-pay.html

De Haro, J. M., Castejon, J. L., & Gilar, R. (2020). Personality and salary at early career: the mediating effect of emotional intelligence. *International Journal of Human Resource Management,* 31(14), 1844–1862. https://doi.org/10.1080/09585192.2017.1423365

Douglas, H.E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, 76(4), 444–463. https://doi.org/10.1086/648111

Duleep, H. & Regets, M. (1998). Projecting Immigrant Earnings: The Significance of Country of Origin. *Social Security Bulletin*, Vol., 61, No. 4. https://www.ssa.gov/policy/docs/ssb/v61n4/v61n4p32.pdf

Falk, C. F., & Muthukrishna, M. (2023). Parsimony in Model Selection: Tools for Assessing Fit

Propensity. *Psychological Methods*, *28*(1), 123–136. https://doi.org/10.1037/met0000422

Federal Reserve Bank of Atlanta. (2023). *Wage Growth Tracker*. https://www.atlantafed.org/chcs/wage-growthtracker#:~:text=The%20Atlanta%20Fed%27s%20Wage%20Growth%20Tracker%20is%20a,developed%20by%20colleagues%20at%20the%20San%20Francisco%20Fed.

Federal Reserve Bank of St. Louis.(2020). *Taking a Closer Look at Marital Status and the Earnings Gap*.https://www.stlouisfed.org/on-the-economy/2020/september/taking-closer-look-marital-status-earnings-gap

Ferris, G.R., Witt, L. A., & Hochwarter, W. A. (2001). Interaction of Social Skill and General Mental Ability on Job Performance and Salary. *Journal of Applied Psychology*, 86(6), 1075–1082. https://doi.org/10.1037/0021-9010.86.6.1075

Frank, M.W. (2009). Inequality and Growth in the United States: Evidence from a New State-level Panel of Income Inequality Measures. *Economic Inquiry,* 47: 55-68. https://doi-org.ezproxy2.library.colostate.edu/10.1111/j.1465-7295.2008.00122.x

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. T*he Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, C. & Rizzolo, M. C. (2020). Fair-Wages for People With Disabilities: Barriers and Facilitators. *Journal of Disability Policy Studies*, 31(3), 152–163. https://doi.org/10.1177/1044207320919492

Fulmer, I. S., Gerhart, B., & Kim, J. H. (2023). Compensation and performance: A review and recommendations for the future. *Personnel Psychology*, *76*(2), 687–718. https://doi.org/10.1111/peps.12583

Golden, L. (2020). *Part-time workers pay a big-time penalty*. Economic Policy Institute.

https://www.epi.org/publication/part-time-pay-penalty/#:~:text=There%20is%20a%20penalty%20for%20working%20part%20time,characteristics%20and%20education%20levels%20who%20work%20full%20time.

Gonzalez, M., Capman, J., Oswald, F., Theys, E., & Tomczak, D. (2019). "Where's the I-O?" Artificial Intelligence and Machine Learning in Talent Management Systems. *Personnel Assessment and Decisions*, 5(3). https://doi.org/10.25035/pad.2019.03.005

Goretzko, D. & Israel, L. S. F. (2022). Pitfalls of Machine Learning-Based Personnel Selection: Fairness, Transparency, and Data Quality. *Journal of Personnel Psychology*, 21(1), 37–47. https://doi.org/10.1027/1866-5888/a000287

Greenwell, B. (2017). pdp: an R package for Constructing Partial Dependence Plots. *The R Journal,* 9:1, 421-436.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Guzman, G. (2020). *Household Income by Race and Hispanic Origin: 2005-2009 and 2015-2019.* American Community Survey Briefs. https://www.census.gov/content/dam/Census/library/publications/2020/acs/acsbr19-07.pdf

Hernandez-Murillo, R., & Owyang, M. T. (2017). *Comparing income, education, and job data for immigrants vs. those born in U.S*. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/publications/regional-economist/second-quarter-2017/comparing-income-education-and-job-data-for-immigrants-vs-those-born-in-us

Hoan, E., & MacDonald, G. (2024). Personality and Well-Being Across and Within Relationship

Status. *Personality & Social Psychology Bulletin*, 1–16. https://doi.org/10.1177/01461672231225571

Houtrow, A. J., Pruitt, D. W., & Zigler, C. K. (2020). Gender-Based Salary Inequities Among Pediatric Rehabilitation Medicine Physicians in the United States. *Archives of Physical Medicine and Rehabilitation*, 101(5), 741–749. https://doi.org/10.1016/j.apmr.2019.11.007

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J. (2023). *Tree-Based Methods. In: An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, Cham. https://doi-org.ezproxy2.library.colostate.edu/10.1007/978-3-031-38747-0_8

John, D., & Mills, J. (2023). hstats: Univariate and multivariate conditional independence testing. Comprehensive R Archive Network (CRAN). Retrieved from https://cran.csail.mit.edu/web/packages/hstats/hstats.pdf

Johnson, J.W. & Lebreton, J. M. (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods*, 7(3), 238–257. https://doi.org/10.1177/1094428104266510

Judge, T., Locke, E., & Durham, C. (1997). The dispositional causes of job satisfaction: A core evaluations approach. *Research in Organizational Behavior*, 19, 151– 188.

Karney B. R. (2021). Socioeconomic Status and Intimate Relationships. *Annual Review of Psychology,* 72, 391–414. https://doi.org/10.1146/annurev-psych-051920-013658

Kim, C., Tamborini, C. & Sakamoto., A. (2015). Field of Study in College and Lifetime Earnings in the United States. *Sociology of Education*, 88:4, 320-339. https://doi-org.ezproxy2.library.colostate.edu/10.1177/0038040715602132

Kochhar, R. (2023). *The Enduring Grip of the Gender Pay Gap*. Pew Research Center.

https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/

Kochar, R. & Bennett, J. (2021). *Immigrants in U.S. experienced higher unemployment in the pandemic but have closed the gap*. Pew Research Center. https://www.pewresearch.org/short-reads/2021/07/26/immigrants-in-u-s-experienced-higher-unemployment-in-the-pandemic-but-have-closed-the-gap/

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1-26.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kuhn, P & Weinberger, C. (2005). Leadership skills and wages. *Journal of Labor Economics*, 23(3), 395–436. https://doi.org/10.1086/430282

Lang, J. & Kell, H. (2020). General Mental Ability and Specific Abilities: Their Relative Importance for Extrinsic Career Success. *Journal of Applied Psychology*, Vol 105, No. 9 1047-1061.

Lawler, E. L. (1971). *Pay and organizational effectiveness: A psychological view*. McGraw-Hill.

Lazear, E. P. (1986). Salaries and Piece Rates. *The Journal of Business* (Chicago, Ill.), 59(3), 405–431. https://doi.org/10.1086/296345

Leana, C. R., & Meuris, J. (2015). Living to work and working to live: Income as a driver of organizational behavior. *The Academy of Management Annals, 9*(1), 55–95. https://doi-org.ezproxy2.library.colostate.edu/10.1080/19416520.2015.1007654

Longhi, S. (2017). *The disability pay gap*. Equality and Human Rights Commission. https://www.equalityhumanrights.com/sites/default/files/research-report-107-the-disability-pay-gap.pdf

Mainert, J., Niepel, C., Murphy, K. R., & Greiff, S. (2019). The Incremental Contribution of Complex Problem-Solving Skills to the Prediction of Job Level, Job Complexity, and Salary. *Journal of Business and Psychology*, 34(6), 825–845. https://doi.org/10.1007/s10869-018-9561-xhttps://doi-org.ezproxy2.library.colostate.edu/10.1007/s10869-018-9561-x

Mattock, M. G., Hosek, J., Trott, D. M., Miller, L. L., & Asch, B. J. (2022). *How veterans fare in the civilian labor market*. RAND Corporation. https://www.rand.org/pubs/articles/2022/how-veterans-fare-in-the-civilian-labor-market.html

Metcalf, H. (2009). *Pay Gaps Across the Equality Strands: A Review*. Equality and Human Rights Commission.https://www.equalityhumanrights.com/en/publication-download/researchreport-14-pay-gaps-across-equality-strands-review

Mishel, L. (2012). *Unions, inequality, and faltering middle-class wages*. Economic Policy Institute. https://www.epi.org/publication/ib342-unions-inequality-faltering-middle-class/

Mokre, P. & Rehm, M. (2020). Inter-industry wage inequality: persistent differences and turbulent equalization. *Cambridge Journal of Economics*, Volume 44, Issue 4, 919-942.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.).christophm.github.io/interpretable-ml-book/

Ng, T.W. H., Ebay, L. T., Sorensen, K.L., & Feldman, D.C. (2005) Predictors of Objective and Subjective Career Success: A Meta-Analysis. *Personnel Psychology*, 58(2), 367–408. https://doi.org/10.1111/j.1744-6570.2005.00515.x

Nyhus, E., & Pons, E. (2005). The effects of personality on earnings. *Journal of Economic Psychology*, 26(3), 363–384. https://doi.org/10.1016/j.joep.2004.07.001

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. *Frontiers in Psychology*, 10, 2970–2970. https://doi.org/10.3389/fpsyg.2019.02970

Ozhamaratli, F., Kitov, O. & Barucca, P. (2022) A generative model for age and income distribution. *EPJ Data Sci.* 11, 4. https://doi.org/10.1140/epjds/s13688-022-00317-x

Patten, E. (2016). *Racial, gender wage gaps persist in U.S. despite some progress*. Pew Research Center. https://www.pewresearch.org/short-reads/2016/07/01/racial-gender-wage-gaps-persist-in-u-s-despite-some-progress/

Putka, D., Beatty, A. S., & Reeder, M. C. (2018). Modern Prediction Methods: New Perspectives on a Common Problem. *Organizational Research Methods*, 21(3), 689–732. https://doi.org/10.1177/1094428117697041

Ruggles, S., Flood, S., Sobek, M., Brockman, D., Cooper, G., Richards, S., & Schouweiler, M. (2023) IPUMS USA: Version 13.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V13.0

Sarkar, D., & Natarajan, V. (2019). *Ensemble Machine Learning Cookbook* (1st edition). Packt Publishing.

Säve-Söderbergh. J.(2019). Gender gaps in salary negotiations: Salary requests and starting salaries in the field. *Journal of Economic Behavior & Organization*, 161, 35–51. https://doi.org/10.1016/j.jebo.2019.01.019

Sayre, G. M., & Conroy, S. A. (2023). The Other Side of the Coin: An Integrative Review Connecting Pay and Health. *Journal of Applied Psychology*. Advance online publication. https://dx.doi.org/10.1037/apl0001151

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel

psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi-org.ezproxy2.library.colostate.edu/10.1037/0033-2909.124.2.262

Schuster, C., Sparkman, G., Walton, G. M., Alles, A., & Loschelder, D. D. (2023). Egalitarian norm messaging increases human resources professionals' salary offers to women. *Journal of Applied Psychology*, 108(4), 541–552. https://doi.org/10.1037/apl0001033

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Speach, M. E. P., Badura, K. L., & Blum, T. C. (2023). Everything is negotiable, but not for everyone: The role of disability in compensation. *Journal of Applied Psychology*, 108(4), 571–594. https://doi.org/10.1037/apl0001039

Spurk D., & Abele A. E. (2011). Who earns more and why? A multiple mediation model from personality to salary. *Journal of Business and Psychology*, 26: 87-103.

Stephan, Y., Sutin, A. R., Luchetti, M., & Terracciano, A. (2016). Allostatic Load and Personality: A 4-Year Longitudinal Study. *Psychosomatic Medicine*, *78*(3), 302–310. https://doi.org/10.1097/PSY.0000000000000281

Tax Policy Center (2023). *Historical Higher Marginal Income Tax.* https://www.taxpolicycenter.org/statistics/historical-highest-marginal-income-tax-rates

Taylor, L. L., Lahey, J. N., Beck, M. I., & Froyd, J. E. (2020). How to Do a Salary Equity Study: With an Illustrative Example From Higher Education. *Public Personnel Management*, 49(1), 57-82. https://doi-org.ezproxy2.library.colostate.edu/10.1177/0091026019845119

Torpey, E. (2018). *Measuring the value of Education.* U.S. Bureau of Labor Statistics.

https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm

Trevor, C. O., Gerhart, B., & Boudreau, J. W. (1997). Voluntary Turnover and Job Performance: Curvilinearity and the Moderating Influences of Salary Growth and Promotions. *Journal of Applied Psychology*, 82(1), 44–61. https://doi.org/10.1037/0021-9010.82.1.44

U.S. Bureau of Economic Analysis. (2023*). Personal income by county, metro, and other areas*. https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas

U.S. Census Bureau. (2021). *Marriages and divorces*. U.S. Department of Commerce. https://www.census.gov/newsroom/press-releases/2021/marriages-and-divorces.html

U.S. Census Bureau. (2022). *American Community Survey: Design and methodology report* (2022). U.S. Department of Commerce. https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2022/acs_design_methodology_report_2022.pdf

U.S. Census Bureau. (2022). *American Community Survey: Subject definitions*. Retrieved from _https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2022_ACSSubjectDefinitions.pdf

U.S. Department of Commerce. (2023). *Personal Income by County, Metro, and Other Areas,* https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas

Verhagen, M. D. (2022). A pragmatist's guide to using prediction in the social sciences. *Socius: Sociological Research for a Dynamic World, 8*, 1-17. https://doi.org/10.1177/23780231221081702

West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry* ;59(3):162-165.

Whelan, C.T. (1992), The role of income, life-style deprivation and financial strain in mediating

the impact of unemployment on psychological distress: Evidence from the Republic of Ireland. *Journal of Occupational and Organizational Psychology*, 65: 331-344. https://doi-org.ezproxy2.library.colostate.edu/10.1111/j.2044-8325.1992.tb00509.x

Yarkoni, T. & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100 - 1122. https://doi.org/10.1177/1745691617693393

Yerger, D. (2017). Skills and earnings in less than bachelor's occupations. *International Journal of Social Economics*, 44(1), 60–74. https://doi.org/10.1108/IJSE-03-2015-0048

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*. Series B, Statistical Methodology, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

APPENDIX

Table 1

IPUMS Predictor Use

Variable Title	Description	Variable Inclusion	Reason for Exclusion	Inclusion in Theoretical Dataset	Evidence for Inclusion
ABSENT	Absent from work last week	Y			
ACREHOUS	House acreage	Y			
AGE	Age	Y		Yes	Ozhamaratli et al, 2022
ANCESTR1 (general)	Ancestry, first response [general version]	Y		Yes	Construct overlap with race (Baker et al., 2017)
ARRIVES	Time of arrival at work	Y			
AVAILBLE	Available for work	Y			
BEDROOMS	Number of bedrooms	Y			
BIRTHQTR	Quarter of birth	Y			
BPL (general)	Birthplace [general version]	Y			
BUILTYR2	Age of structure, decade	Y			
CARPOOL	Carpooling	Y			

CIDATAPLN	Cellular data plan for a smartphone or other mobile device	Y			
CIDIAL	Dial-up service	Y			
CIHISPEED	Broadband (high speed) Internet service such as cable, fiber optic, or DSL service	Y			
CILAPTOP	Laptop, desktop, or notebook computer	Y			
CINETHH	Access to internet	Y			
CIOTHCOMP	Other computer equipment	Y			
CISAT	Satellite internet service	Y			
CISMRTPHN	Smartphone	Y			
CITABLET	Tablet or other portable wireless computer	Y			
CITIZEN	Citizenship status	Y		Yes	Hernandez-Murillo & Owyang (2017)
CITY	City	Y			
CITYPOP	City population	Y			
CLASSWKR (general)	Class of worker [general version]	Y			

CNTRY	Country	Y			
COSTELEC	Annual electricity cost	Y			
COSTFUEL	Annual home heating fuel cost	Y			
COSTGAS	Annual gas cost	Y			
COSTWATR	Annual water cost	Y			
COUNTYFIP	County (FIPS code, identifiable counties only)	Y		Yes	U.S. Bureau of Economic Analysis (2023)
DEGFIELD (general)	Field of degree [general version]	Y		Yes	Kim et al., 2015
DEPARTS	Time of departure for work	Y			
DIFFCARE	Self-care difficulty	Y			
DIFFEYE	Vision difficulty	Y			
DIFFHEAR	Hearing difficulty	Y		Yes	Longhi, 2021
DIFFMOB	Independent living difficulty	Y		Yes	Longhi, 2021
DIFFPHYS	Ambulatory difficulty	Y			
DIFFREM	Cognitive difficulty	Y		Yes	Longhi, 2021
DIFFSENS	Vision or hearing difficulty	Y		Yes	Longhi, 2021

DIVINYR	Divorced in the past year	Y			
EDUC (general)	Educational attainment [general version]	Y		Yes	Mishel, 2012
ELDCH	Age of eldest own child in household	Y			
FAMSIZE	Number of own family members in household	Y			
FARM	Farm status	Y			
FARMPROD	Sales of farm products	Y			
FERTYR	Children born within the last year	Y			
FRIDGE	Refrigerator	Y			
FUELHEAT	Home heating fuel	Y			
GCHOUSE	Own grandchildren living in household	Y			
GCMONTHS	Months responsible for grandchildren	Y			
GCRESPON	Responsible for grandchildren	Y			
GQ	Group quarters status	Y			
HHTYPE	Household Type	Y			
HISPAN (general)	Hispanic origin [general version]	Y		Yes	Guzman, 2020

HOMELAND	American Indian, Alaska Native, or Native Hawaiian homeland area	Y		Yes	
HOTWATER	Hot and cold piped water	Y			
INCTOT	Total personal income	Y			
IND	Industry	Y		Yes	Mokre & Rehm, 2020
KITCHEN	Kitchen or cooking facilities	Y			
LABFORCE	Labor force status	N	Criterion Related		
LANGUAGE (general)	Language spoken [general version]	Y		Yes	Day & Shin, 2005
LINGISOL	Linguistic isolation	Y			
LOOKING	Looking for work	N	Criterion Related		
MARRINYR	Married within the past year	Y			
MARRNO	Times married	Y			
MARST	Marital status	Y		Yes	Leonard & Stanley, 2015
METRO	Metropolitan status	Y			
MIGCOUNTY1	County of residence 1 year ago	Y			
MIGPLAC1	State or country of residence 1 year ago	Y			

MIGRATE1 (general)	Migration status, 1 year [general version]	Y		Yes	Duleep & Regets, 1998
MOVEDIN	When occupant moved into residence	Y			
MULTGEN (general)	Multigenerational household [general version]	Y			
NCHILD	Number of own children in the household	Y			
NCHLT5	Number of own children under age 5 in household	Y			
NCOUPLES	Number of couples in household	Y			
NFAMS	Number of families in household	Y			
NSIBS	Number of own siblings in household	Y			
OCC	Occupation	Y		Yes	BLS, 2023
OWNERSHP (general)	Ownership of dwelling (tenure) [general version]	Y			
PHONE	Telephone availability	Y			
PLUMBING	Plumbing facilities	Y			
PWCOUNTY	Place of work: county	Y			
PWSTATE2	Place of work: state	Y		Yes	Frank, 2007

PWTYPE	Place of work: metropolitan status	Y			
RACAMIND	Race: American Indian or Alaska Native	Y			
RACASIAN	Race: Asian	Y		Yes	Castilla, 2008
RACBLK	Race: black or African American	Y		Yes	Castilla, 2008
RACE (general)	Race [general version]	Y		Yes	Castilla, 2008
RACHSING	Race: Simplified race/ethnicity identification	N			
RACOTHER	Race: some other race	Y			
RACPACIS	Race: Pacific Islander	Y			
RACWHT	Race: white	Y		Yes	Castilla, 2008
REGION	Census region and division	N	Repeated measure		
RESPMODE	Response mode	Y			
RIDERS	Vehicle occupancy	Y			
ROOMS	Number of rooms	N	Criterion related		
SCHLTYPE	Public or private school	Y			
SEX	Sex	Y			
SHOWER	Bathtub or shower	Y			
SINK	Sink with faucet	Y			

SPEAKENG	Speaks English	Y		Yes	Day & Shin, 2005
SSMC	Same-sex married couple	Y			
STATEFIP	State (FIPS code)	Y		Yes	Frank, 2007
STOVE	Stove or range	Y			
TRANTIME	Travel time to work	Y			
TRANWORK	Means of transportation to work	Y			
TRIBE (general)	Tribe [general version]	Y			
UHRSWORK	Usual hours worked per week	Y		Yes	Golden, 2020
UNITSSTR	Units in structure	Y			
VACANCY	Vacancy status	Y			
VEHICLES	Vehicles available	Y			
VETDISAB	VA service-connected disability rating	Y			
VETSTAT (general)	Veteran status [general version]	Y		Yes	Mattock et al., 2022
WIDINYR	Widowed in the past year	Y			
WKSWORK2	Weeks worked last year, intervalled	Y			
WORKEDYR	Worked last year	Y			

WRKLSTWK	Worked last week	Y			
WRKRECAL	Informed of work recall	N	Criterion related		
YEAR	Census year	Y			
YNGCH	Age of youngest own child in household	Y			
YRIMMIG	Year of immigration	Y			
YRMARR	Year married	Y			
YRNATUR	Year naturalized	Y			
YRSUSA1	Years in the United States	Y			
BIRTHYR	Year of birth	N	Repeated measure		
FOODSTMP	Food stamp recipiency	N	Criterion Related		
GRADEATT (general)	Grade level attending [general version]	N	Repeated measure		
SCHOOL	School attendance	N	Repeated measure		
ADJGINC	Tax unit's adjusted gross income	N	Criterion Related		
ANCESTR1D (detailed)	Ancestry, first response [detailed version]	N	Repeated measure		
ANCESTR2 (general)	Ancestry, second response [general version]	N	Repeated measure		
ANCESTR2D (detailed)	Ancestry, second response [detailed version]	N	Repeated measure		

BPLD (detailed)	Birthplace [detailed version]	N	Repeated measure		
CBHHTYPE	Census bureau household type (with cohabiting)	N	Repeated measure		
CBNSUBFAM	Number of subfamilies in household (original Census Bureau classification)	N	Repeated measure		
CBSERIAL	Original Census Bureau household serial number	N	Participant linking		
CBSUBFAM	Subfamily number (original Census Bureau classification)	N	Repeated measure		
CIOTHSVC	Other internet service	N	Repeated measure		
CLASSWKRD (detailed)	Class of worker [detailed version]	N	Repeated measure		
CLUSTER	Household cluster for variance estimation	N	Weighting		
CONDOFEE	Monthly condominium fee	N	Temporal precedence		
COUNTYICP	County (ICPSR code, identifiable counties only)	N	Repeated measure		
DEGFIELD2 (general)	Field of degree (2) [general version]	N	Repeated measure		
DEGFIELD2D (detailed)	Field of degree (2) [detailed version]	N	Repeated measure		

DEGFIELDD (detailed)	Field of degree [detailed version]	N	Repeated measure		
DENSITY	Population-weighted density of PUMA	N	Weighting		
EDUCD (detailed)	Educational attainment [detailed version]	N	Repeated measure		
EMPSTAT (general)	Employment status [general version]	N	Repeated measure		
EMPSTATD (detailed)	Employment status [detailed version]	N	Repeated measure		
FAMUNIT	Family unit membership	N	Repeated measure		
FTOTINC	Total family income	N	Criterion Related		
GQTYPE (general)	Group quarters type [general version]	N	Repeated measure		
GQTYPED (detailed)	Group quarters type [detailed version]	N	Repeated measure		
GRADEATTD (detailed)	Grade level attending [detailed version]	N	Repeated measure		
HCOVANY	Any health insurance coverage	N	Temporal precedence		
HCOVPRIV	Private health insurance coverage	N	Temporal precedence		
HCOVPUB	Public health insurance coverage	N	Temporal precedence		

HCOVSUB2	Subsidized marketplace insurance coverage (original)	N	Temporal precedence		
HHINCOME	Total household income	N	Criterion Related		
HHWT	Household weight	N	Weighting		
HINSCAID	Health insurance through Medicaid	N	Temporal precedence		
HINSCARE	Health insurance through Medicare	N	Temporal precedence		
HINSEMP	Health insurance through employer/union	N	Temporal precedence		
HINSIHS	Health insurance through Indian Health Services	N	Temporal precedence		
HINSPUR	Health insurance purchased directly	N	Temporal precedence		
HINSTRI	Health insurance through TRICARE	N	Temporal precedence		
HINSVA	Health insurance through VA	N	Temporal precedence		
HISPAND (detailed)	Hispanic origin [detailed version]	N	Repeated measure		
HWSEI	Socioeconomic Index, Hauser and Warren	N	Temporal precedence		
INCEARN	Total personal earned income	N	Criterion Related		

INCINVST	Interest, dividend, and rental income	N	Criterion Related		
INCOTHER	Other income	N	Criterion Related		
INCRETIR	Retirement income	N	Criterion Related		
INCSS	Social Security income	N	Criterion Related		
INCSUPP	Supplementary Security Income	N	Criterion Related		
INCWAGE	Wage and salary income	N	Criterion Related		
INCWELFR	Welfare (public assistance) income	N	Criterion Related		
INSINCL	Mortgage payment includes property insurance	N	Temporal precedence		
LANGUAGED (detailed)	Language spoken [detailed version]	N	Repeated measure		
MEDICAREB	Person's Medicare Part B premium	N	Temporal precedence		
MET2013	Metropolitan area (2013 OMB delineations)	N	Previous survey		
MIGMET131	Metropolitan area of residence 1 year ago (2013 delineations)	N	Repeated measure		
MIGRATE1D (detailed)	Migration status, 1 year [detailed version]	N	Repeated measure		

MIGTYPE1	Metropolitan status 1 year ago	N	Repeated measure		
MOBLHOME	Annual mobile home costs	N	Temporal precedence		
MOMLOC	Mother's location in the household	N	Participant linking		
MOMLOC2	Second mother's location in the household	N	Participant linking		
MOOP	Person's medical out of pocket expenses, other than premiums	N	Temporal precedence		
MORTAMT1	First mortgage monthly payment	N	Temporal precedence		
MORTAMT2	Second mortgage monthly payment	N	Temporal precedence		
MORTGAG2	Second mortgage status	N	Temporal precedence		
MORTGAGE	Mortgage status	N	Temporal precedence		
MULTGEND (detailed)	Multigenerational household [detailed version]	N	Repeated measure		
NFATHERS	Number of fathers in household	N	Repeated measure		
NMOTHERS	Number of mothers in household	N	Repeated measure		
NSUBFAM	Number of subfamilies in household	N	Repeated measure		

NUMPREC	Number of person records following	N	Participant linking		
OCCSCORE	Occupational income score	N	Criterion Related		
OCCSOC	Occupation, SOC classification	N	Repeated measure		
OFFPOV	Official poverty status	N	Temporal precedence		
OWNCOST	Selected monthly owner costs	N	Repeated measure		
OWNERSHPD (detailed)	Ownership of dwelling (tenure) [detailed version]	N	Repeated measure		
PERNUM	Person number in sample unit	N	Participant linking		
PERWT	Person weight	N	Weighting		
POPLOC	Father's location in the household	N	Participant linking		
POPLOC2	Second father's location in the household	N	Participant linking		
POVERTY	Poverty status	N	Temporal precedence		
PRENT	Occupational prestige score, Nakao and Treas	N	Repeated measure		
PRESGL	Occupational prestige score, Siegel	N	Repeated measure		
PROPINSR	Annual property insurance cost	N	Temporal precedence		

PROPTX99	Annual property taxes, 1990	N	Previous survey		
RACED (detailed)	Race [detailed version]	N	Repeated measure		
RACNUM	Number of major race groups	N	Repeated measure		
RENT	Monthly contract rent	N	Temporal precedence		
RENTGRS	Monthly gross rent	N	Temporal precedence		
RENTMEAL	Meals included in rent	N	Temporal precedence		
SAMPLE	IPUMS sample identifier	N	Participant linking		
SEI	Duncan Socioeconomic Index	N	Temporal precedence		
SERIAL	Household serial number	N	Participant linking		
SFRELATE	Relationship within subfamily	N	Repeated measure		
SFTYPE	Subfamily type	N	Repeated measure		
SPLOC	Spouse's location in household	N	Participant linking		
SPMPOV	SPM poverty status	N	Temporal precedence		
STATEICP	State (ICPSR code)	N	Repeated measure		
STRATA	Household strata for variance estimation	N	Weighting		
SUBFAM	Subfamily membership	N	Repeated measure		

TAXINCL	Mortgage payment includes property taxes	N	Temporal precedence		
TRIBED (detailed)	Tribe [detailed version]	N	Repeated measure		
VALUEH	House value	N	Temporal precedence		
VET01LTR	Veteran, served 2001 or later	N	Repeated measure		
VET47X50	Veteran, served 1947-1950	N	Repeated measure		
VET55X64	Veteran, served 1955 to 1964	N	Repeated measure		
VET75X90	Veteran, served May 1975 to July 1990	N	Repeated measure		
VET90X01	Veteran, served 1990-2001	N	Repeated measure		
VETKOREA	Veteran, served during Korean conflict era	N	Repeated measure		
VETOTHER (general)	Veteran of other period [general version]	N	Repeated measure		
VETOTHERD (detailed)	Veteran of other period [detailed version]	N	Repeated measure		
VETSTATD (detailed)	Veteran status [detailed version]	N	Repeated measure		
VETVIETN	Veteran, served during Vietnam era	N	Repeated measure		
VETWWII	Veteran, served during WWII era	N	Repeated measure		

YRSUSA2	Years in the United States, intervalled	N	Repeated measure		

Note. There were variables that would have met the criteria for inclusion in the theoretical dataset that were not included due to having zero or near zero variance. “Repeated measure” refers to a variable already represented in the dataset and thus it’s inclusion would be redundant. A label of “Temporal Precedence” was given to variables that would be preceded by income, rather than vice-versa and thus should not be included as a predictor. Variables with the label “Weighting” were removed as the were not actually variables but just created to weight variables. Similarly, those with the “Participant Linking” variable not variables but a way to track answers across ACS participants and thus needed to be removed. Previous survey questions marked with “Previous survey” were also removed as there was no data for the chosen years. “Criterion related” variables were also removed as they were too similar or another representation of the outcome of interest.

Table 2

Training and Test Results

Algorithm	Predictors- 2018 Data		Predictors- 2019 Data
	Theory	All	All
OLS	.50	.58	.45
ENET	.50	.58	.47
RF	.55	.60	.49

Note. Table presents R^2 estimates. R^2 estimates presented with 2018 data are K-fold cross-validated estimates using the training data. R^2 estimates presented under 2019 data represent variance explained in the outcome when using predictions based on 2019 predictor data, using models estimated from 2018 data.

Table 3

Variable Comparisons- Representativeness of Reduced Dataset

Variable	Reduced Sample		Full Sample	
	Mean	SD	Mean	SD
Weeks Worked Last Year	3.68	2.69	3.67	2.70
Usual Hrs Worked Per Week	27.37	20.58	27.19	20.58
Occupation (mean)	10.17	.66	10.17	.66
Age	50.26	19.08	50.34	19.12
Industry (mean)	10.17	.54	10.17	.54
Education	7.43	2.39	7.43	2.41
Degree Field (mean)	10.17	.45	10.17	.45
Years Married	19.77	19.22	19.82	19.27
Place of Work State (mean)	10.17	.46	10.17	.46
Ancestry (mean)	10.17	.20	10.17	.20
School Type	1.11	.38	1.12	.39
State of Residence (mean)	10.17	.10	10.17	.10
Birthplace (mean)	10.17	.13	10.17	.13
Travel Time to Work	10.92	17.40	10.84	17.23
No Grandchildren in House	.80	.40	.80	.40
Time of Departure for Work	486.76	483.92	488.61	487.42
County of Residence (mean)	10.17	.11	10.17	.11
Place of Work: Metropolitan State (mean)	10.17	.46	10.17	.46
Age of Structure, Decade	5.56	4.21	5.58	4.22
When Occupant Moved into Residence	2.18	2.50	2.16	2.50

Note. Table compares means and standard deviations for the reduced (n=20,000) and full sample for the top 20 most important predictors from the random forest model.

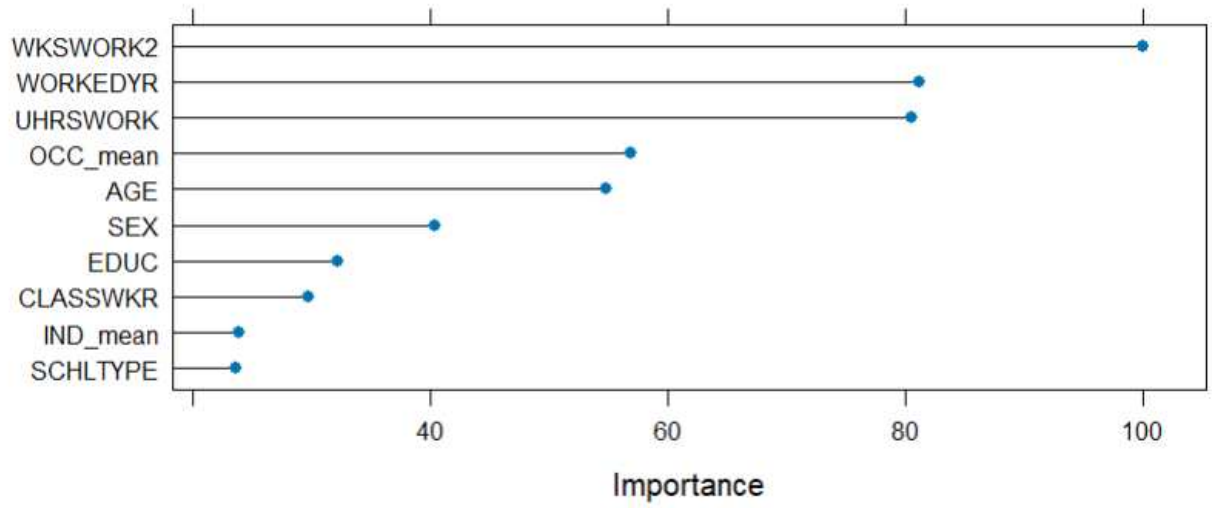

Figure 2

Variable Importance for OLS Model

Note. This variance importance plot estimates the contribution of each variable to the model with a maximum of 100 using the absolute value of the t-statistic for each model parameter, as the model is linear.

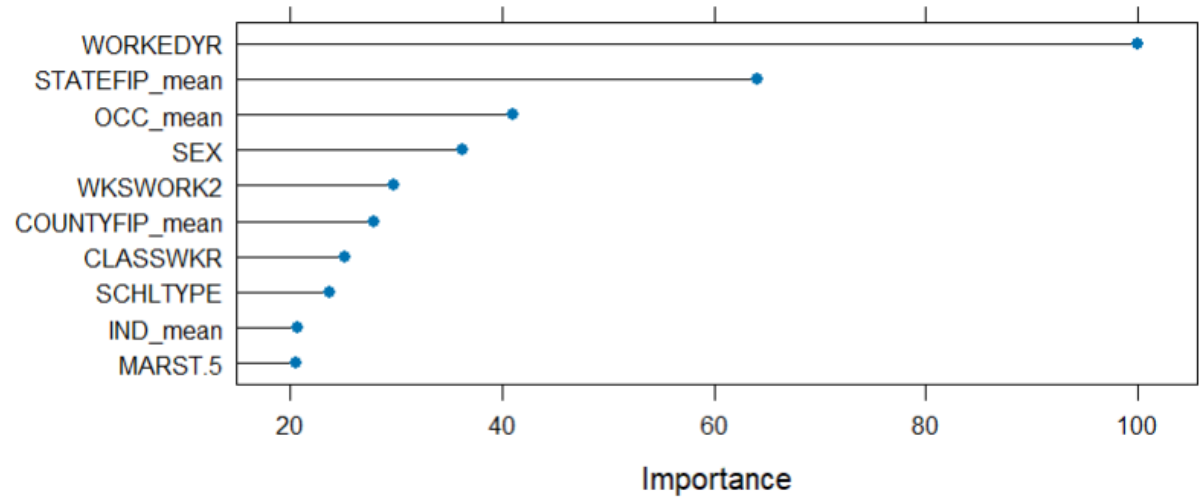

Figure 3

Variable Importance for Elastic Net Model

Note. This variance importance plot estimates the contribution of each variable to the model with a maximum of 100 using the absolute value of the t-statistic for each model parameter, as the model is linear.

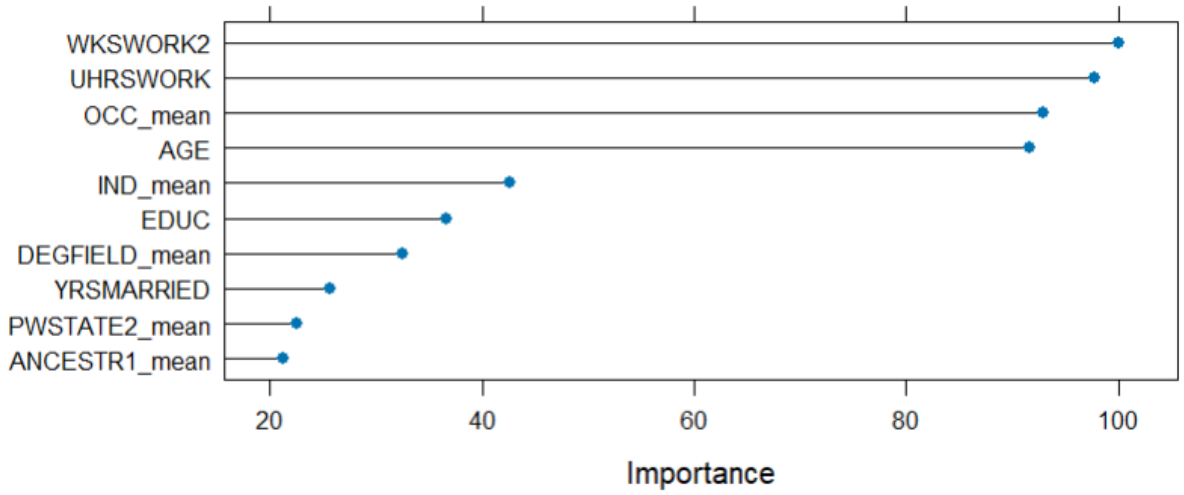

Figure 4

Variable Importance for Random Forest

Note. This variance importance plot estimates the contribution of each variable to the model with a maximum of 100. First, the MSE is computed for each tree on the out of the bag data and then again after variables permutation. Differences are averaged and normalized using the standard error.

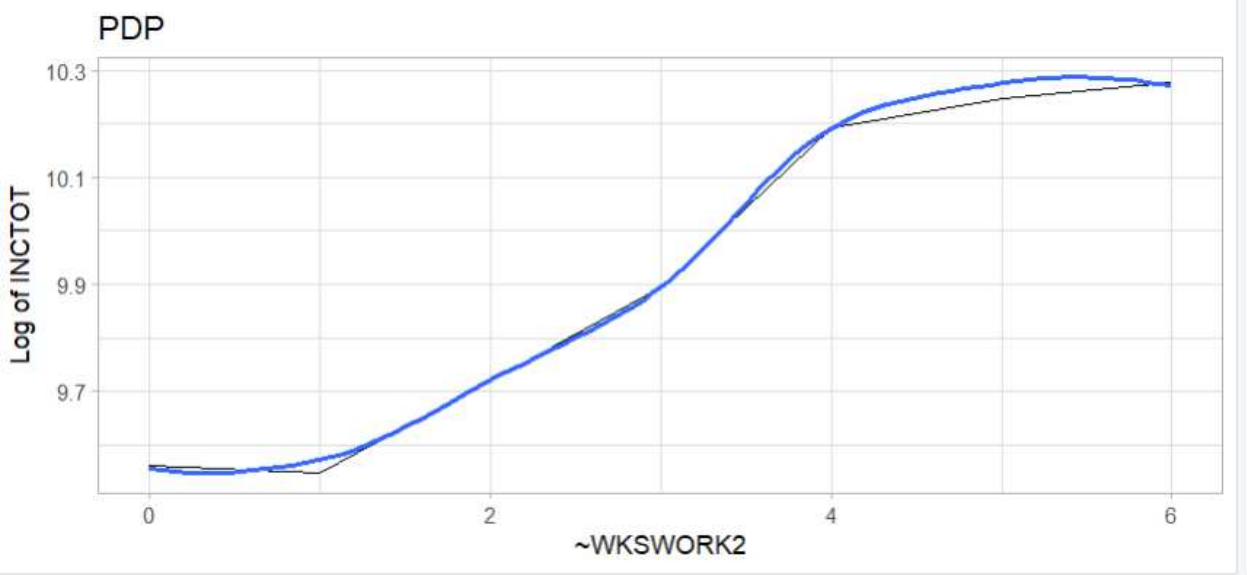

Figure 5

Partial Dependence Plot of Weeks Worked Last Year (Intervalled)

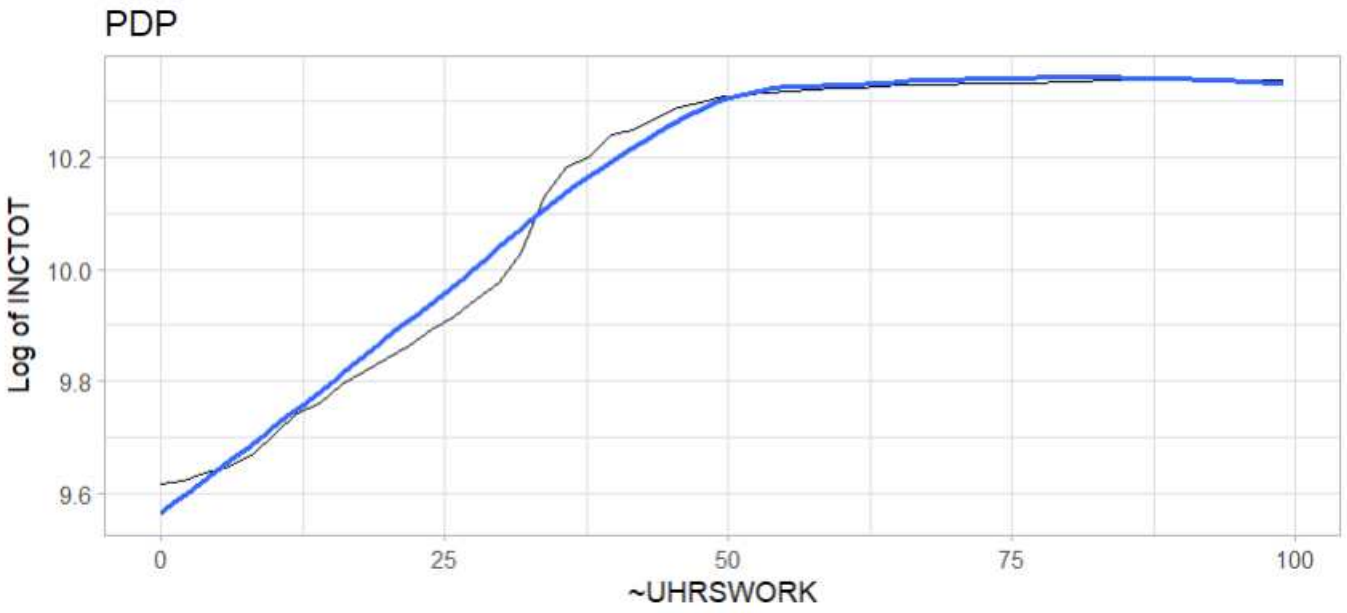

Figure 6

Partial Dependence Plot of Usual Hours Worked

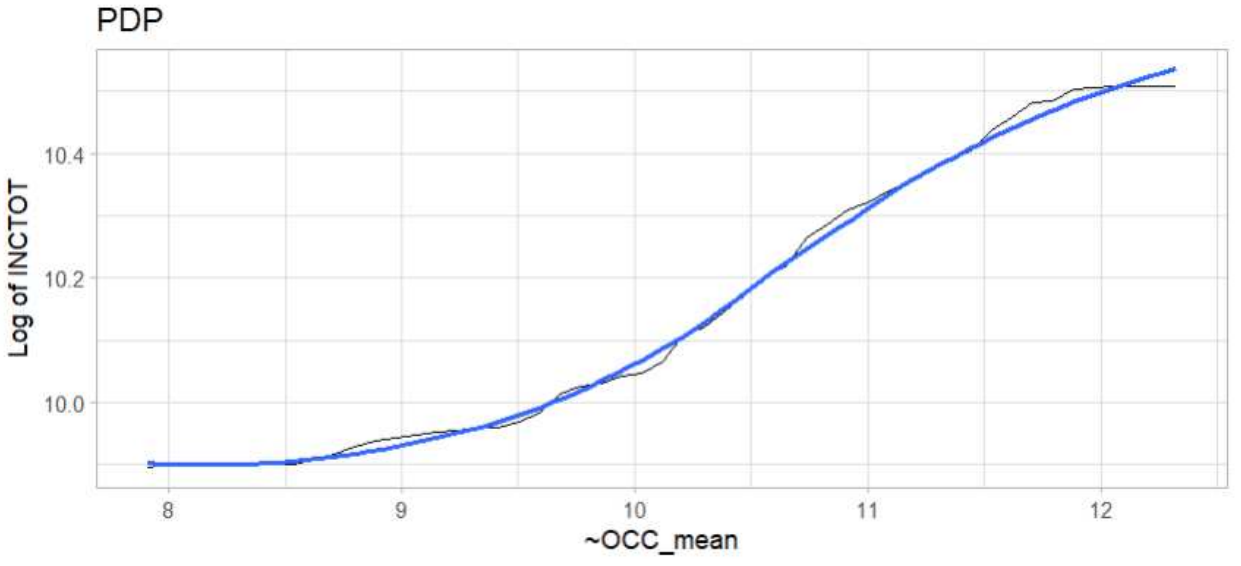

Figure 7

Partial Dependence Plot of Occupation (target encoded)

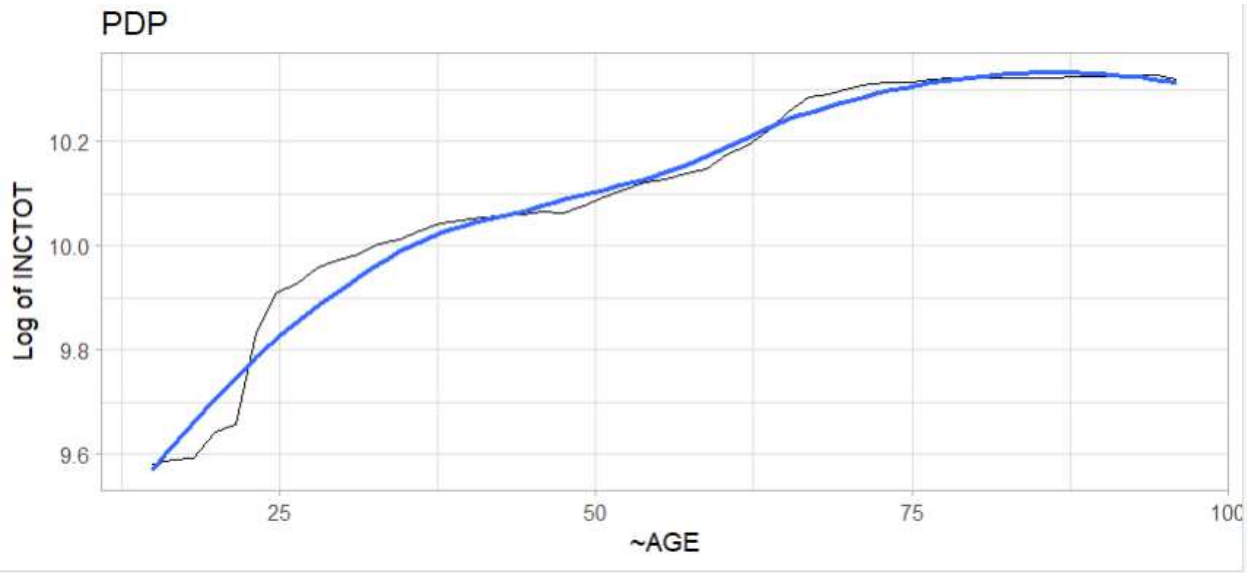

Figure 8
Partial Dependence Plot of Age

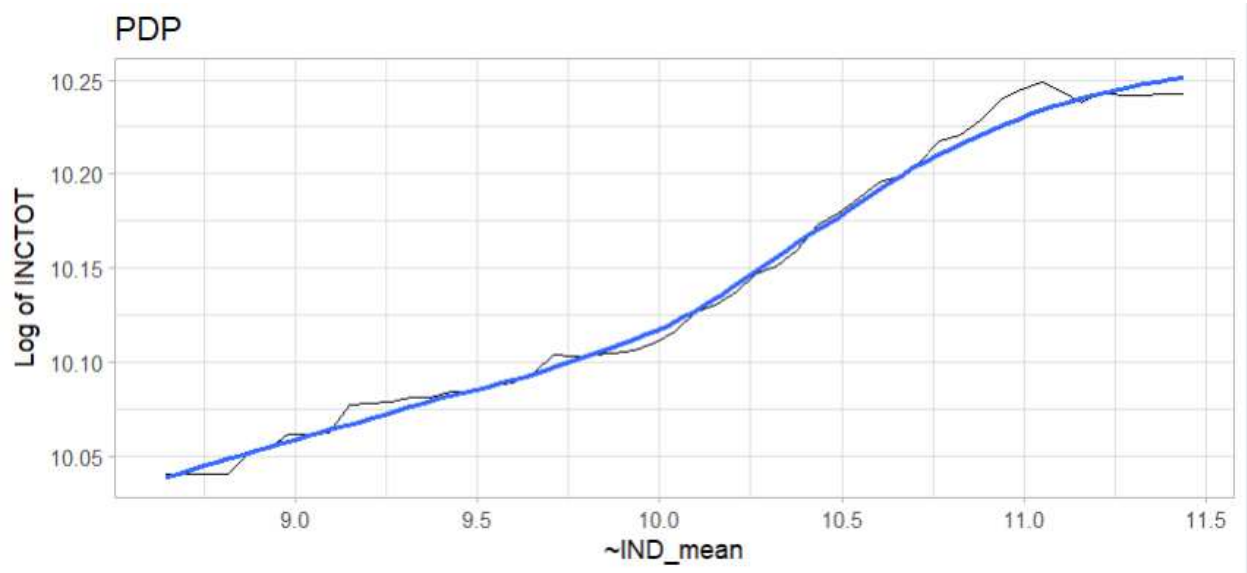

Figure 9
Partial Dependence Plot of Industry (target encoded)

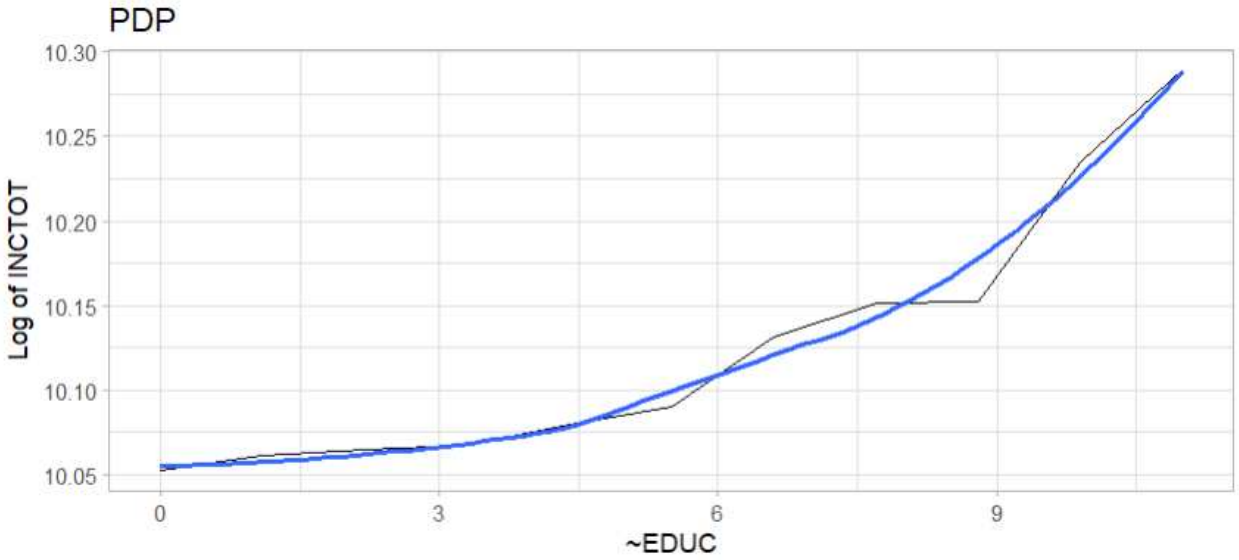

Figure 10

Partial Dependence Plot of Education

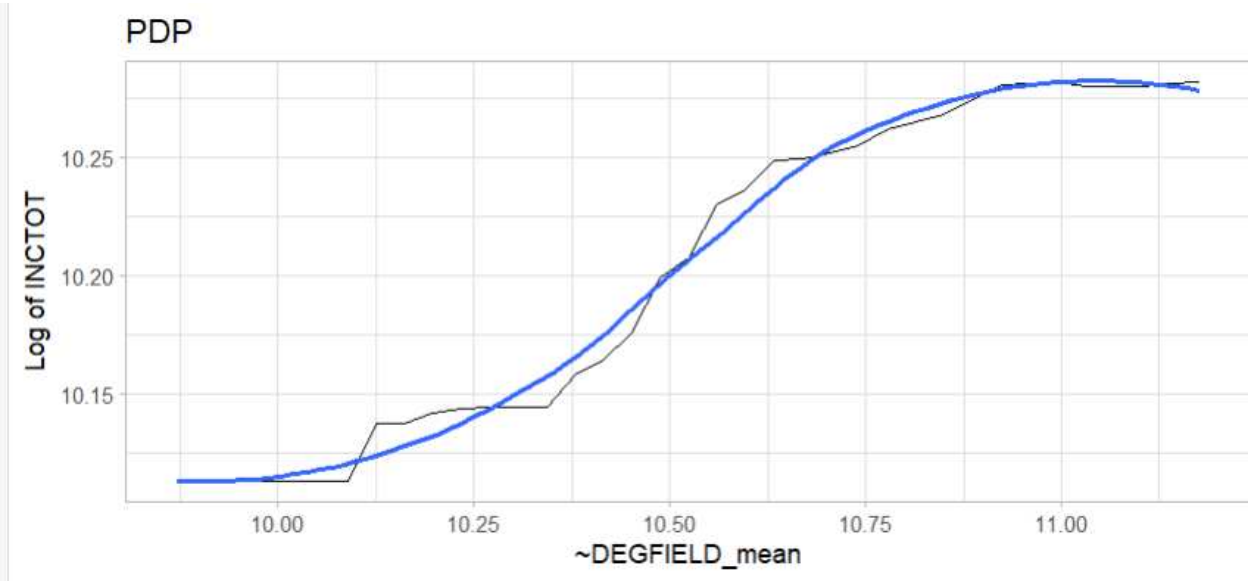

Figure 11

Partial Dependence Plot of Degree Field (target encoded)

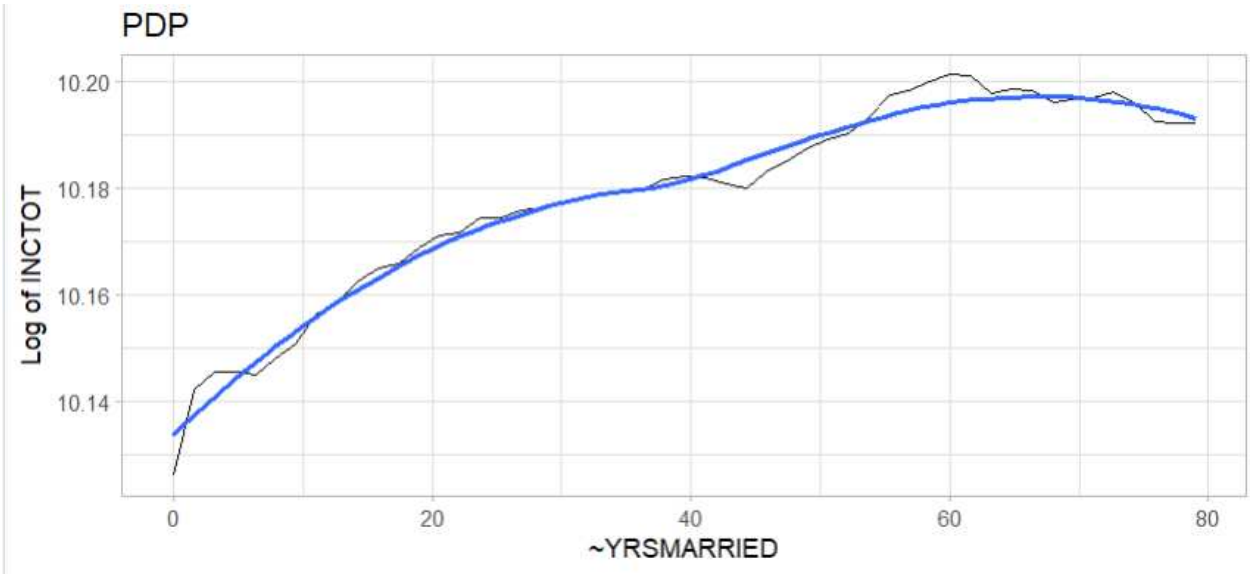

Figure 12

Partial Dependence Plot of Years Married

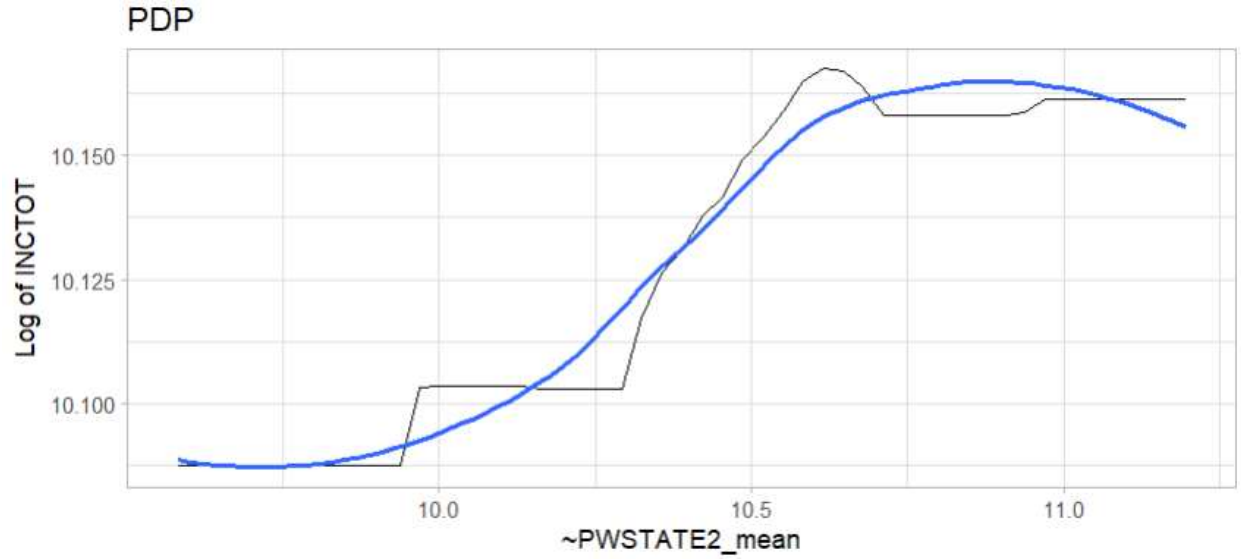

Figure 13

Partial Dependence Plot of Place of Work (State) (Target Encoded

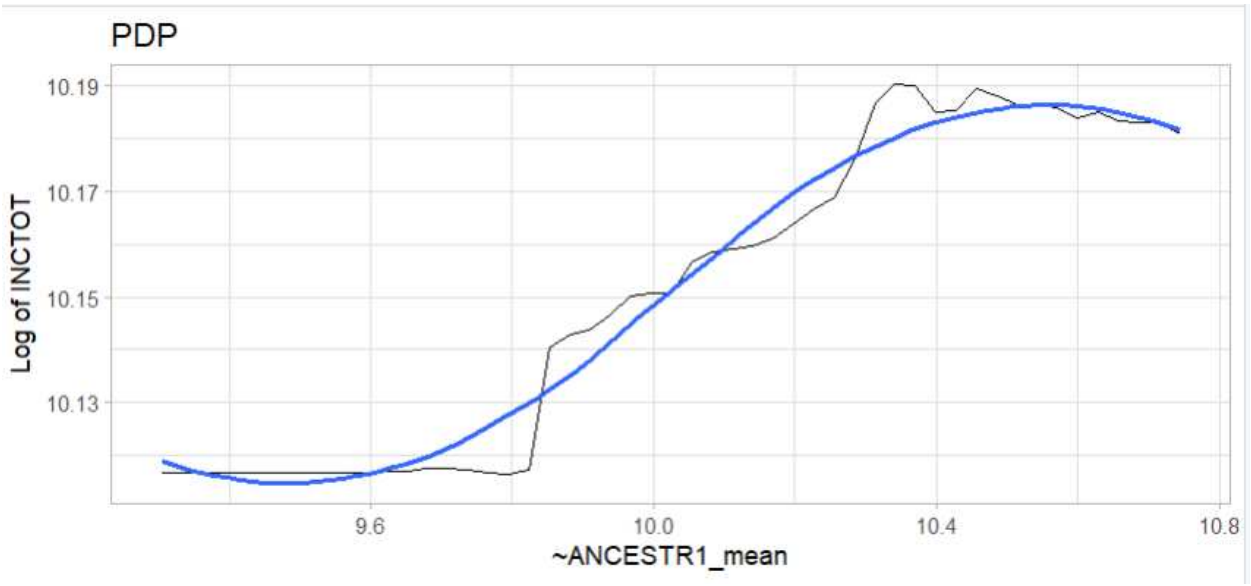

Figure 14

Partial Dependence Plot of Ancestry (Target Encoded)

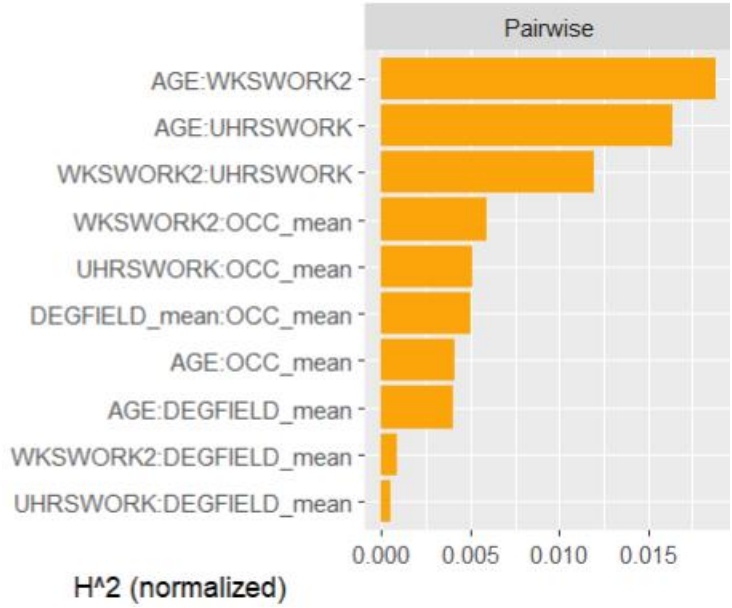

Figure 15

Hstats Interaction Plot

Note. Scale indicates the share of variance attributable to a given predictor that is shared with another predictor.

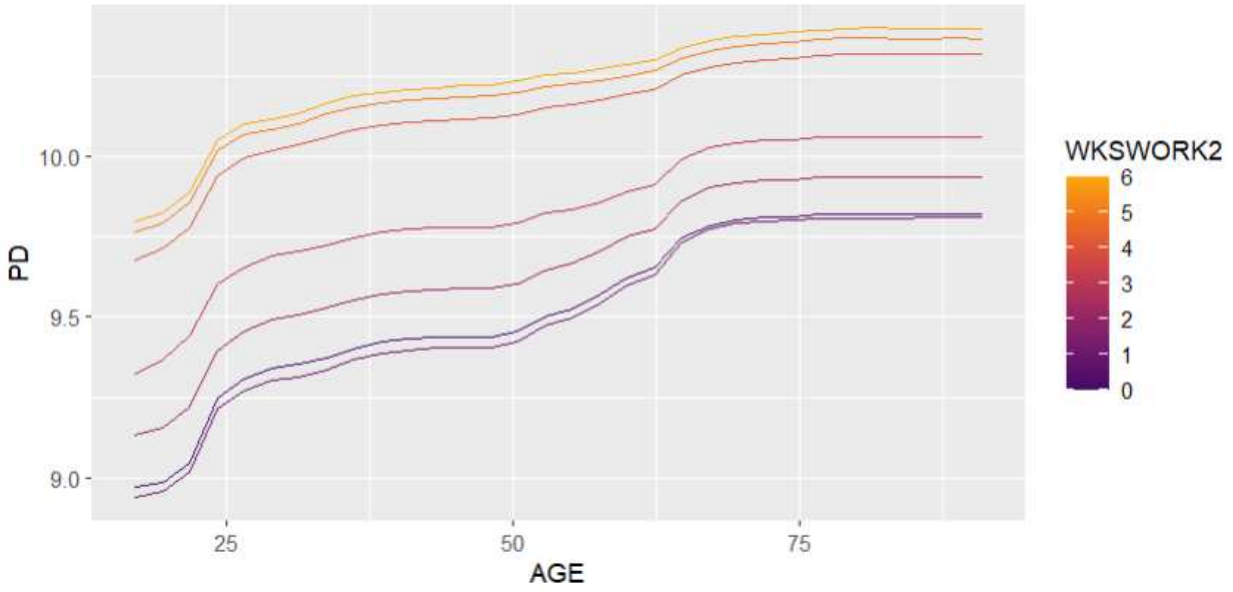

Figure 16

Interaction PDP Between Age and Weeks Worked Last Year (intervalled)

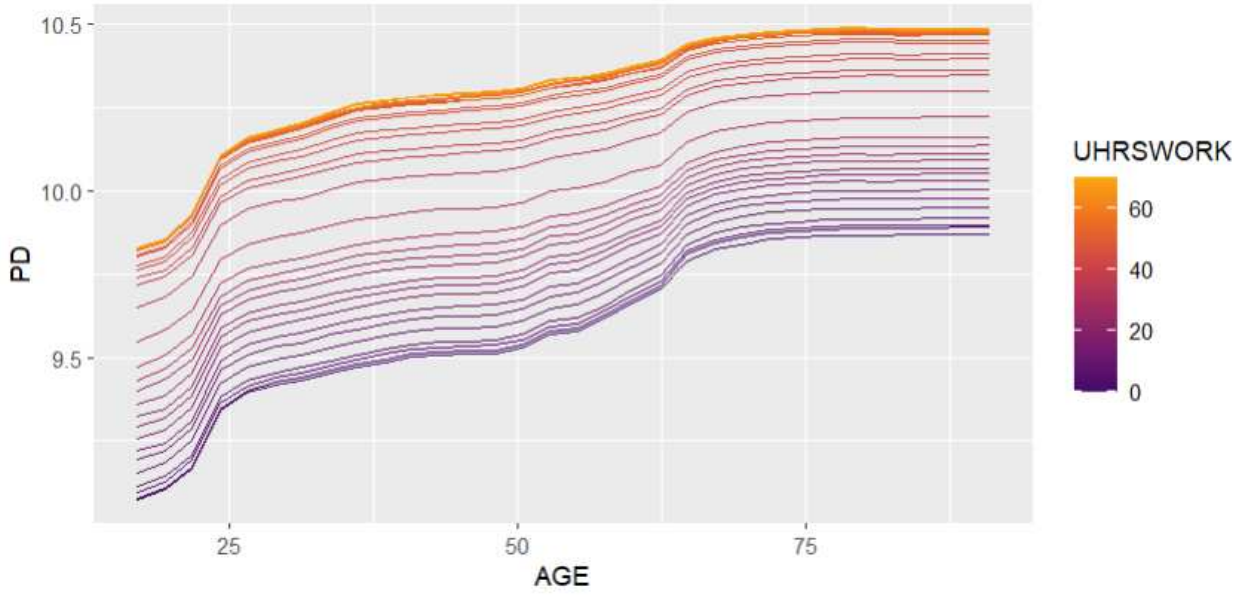

Figure 17

Interaction PDP Between Age and Usual Hours Worked

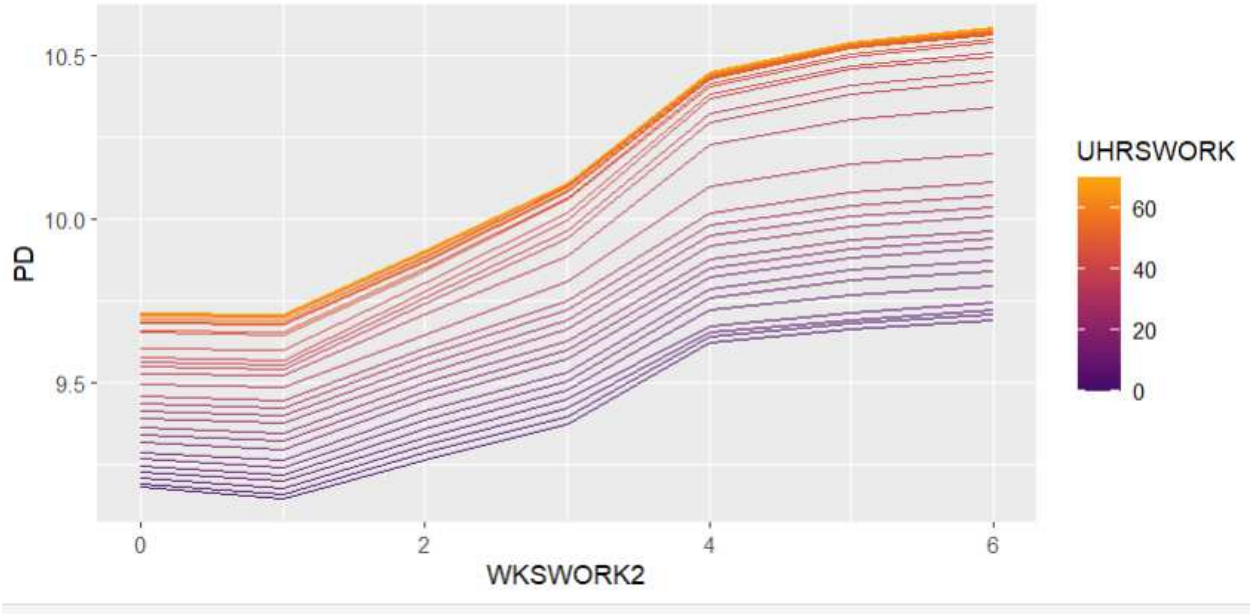

Figure 18

Interaction PDP Between Usual Hours Worked and Weeks Worked Last Year (intervalled)