Muzahidul Islam & Mildred Nwachukwu-Innocent
DAT300 Fall 2024

# Predicting Total Family Income Based on PSID Data

## Abstract

Predicting family income is a critical task with wide-ranging applications in public policy, business, and nonprofit sectors. Accurate income predictions can aid in detecting tax fraud, optimizing marketing strategies, targeting social assistance programs, and supporting informed decision-making. This study employs SAS Enterprise Miner to compare traditional statistical methods, such as linear and polynomial regression, with advanced machine learning models, including decision trees, gradient boosting, neural networks, and ensemble methods, to predict family income.

The research highlights the strengths and limitations of these approaches, particularly the trade-off between the interpretability of traditional methods and the predictive accuracy of machine learning techniques. By leveraging diverse models and evaluating their performance, the study identifies strategies for effective income prediction across different use cases. The findings demonstrate the potential of combining traditional and machine learning methodologies to address economic challenges, providing a robust foundation for practical applications and future research in predictive analytics.

## Introduction

The use of machine learning in economics is rapidly expanding, offering new opportunities to enhance traditional statistical methods like regression analysis. These traditional methods rely on strong statistical assumptions, such as linearity and homoscedasticity, which are often difficult to satisfy when working with real-world data. As data grows in complexity and size, machine learning methodologies are becoming increasingly relevant in predictive analytics because of their ability to handle high-dimensional datasets and non-linear relationships.

This study utilizes data from the Panel Study of Income Dynamics (PSID), a comprehensive longitudinal survey that tracks economic, social, and demographic changes across households. Specifically, data from the 2020 wave of the survey, comprising 9,207 observations. 17 different variables were selected from this dataset to aid in predicting family income.

The research explores predictive modeling for family income, a critical economic variable with significant implications across multiple sectors. Public agencies can utilize income prediction to improve tax compliance and target social assistance programs, while private companies and nonprofits rely on such insights for marketing, fraud detection, and donor outreach. Accurate income predictions can also aid in enabling better-informed policy decisions and resource allocation.

Using SAS Enterprise Miner, this study evaluates several predictive models, including linear and polynomial regression, decision trees, gradient boosting, neural networks, and ensemble methods. By comparing the performance of these methodologies, the research aims to uncover the most effective approaches for income prediction. A focus on both traditional and modern techniques underscores the importance of balancing interpretability with predictive power, ensuring the models can meet diverse needs in practical applications.

This work contributes to the growing intersection of economics and machine learning by showcasing how these advanced methods, applied to PSID data, can improve the accuracy and utility of income prediction. The findings not only highlight the advantages of machine learning for handling complex datasets but also offer a foundation for future research into combining traditional and machine learning methods to solve economic challenges.

## Literature Review

To explore the factors influencing income, we reviewed several sources analyzing variables like education, geographic location, race, gender, and social characteristics. Education consistently emerged as a primary predictor of income, with Gasperi’s (2019) logistic regression analysis on income levels demonstrating a positive correlation between years of schooling and earning potential. The National Bureau of Economic Research’s report further highlights a strong relationship between years of schooling and annual income, underscoring education's impact on earning potential (Mincer, 1974). Additionally, a working paper by the IZA Institute of Labor Economics examined parental education, revealing that family background significantly affects children’s income, reinforcing education as a core predictor for our model (Black et al., 2005).

Geographic location is another significant factor. A study in *The Quarterly Journal of Economics* illustrates regional income inconsistencies, showing that some U.S. regions afford greater income mobility than others (Chetty et al., 2014). Research from PubMed Central highlights the persistent income differences across races and gender (Leopold et al., 2015). The Brookings Institution report further associates marital status with higher average income levels, suggesting that social factors like marriage can positively influence economic stability (Reeves & Sawhill, 2016). Together, these findings informed the inclusion of geographic and social variables to capture broader perspectives on income inequality.

Finally, research from IZA also noted the influence of religious affiliation on economic behaviors, indicating that personal values might impact labor supply and earnings (Guiso et al., 2004). Collectively, these studies informed the selection of key variables such as education, geographic location, race, marital status, gender, and religion, enabling a comprehensive examination of the diverse determinants of income levels.

## Data

This analysis utilizes data from the Panel Study of Income Dynamics (PSID) to model family income. The PSID is a longitudinal survey of U.S. individuals and families, capturing rich socioeconomic and demographic information. For this study, data was downloaded from the PSID Data Center for 2020 at https://simba.isr.umich.edu/default.aspx

### *Target Variable*

- **Total Family Income:** The response variable, representing the total pre-tax income of families, was log-transformed to normalize its distribution and improve model performance.

### *Predictor Variables*

Based on the literature review and data availability, the following variables were selected as potential predictors of family income:

- **Demographic Variables:**
  - Age of Head/Spouse
  - Race of Head/Spouse
  - Sex of Head/Spouse
- **Educational Variables:**
  - College degree attainment (Head and Spouse)
  - Parent education levels (Head's mother, Head's father, Spouse's mother, Spouse's father)
- **Geographic Variables:**

  - State
- **Social Variables:**
  - Marital Status
  - Religion of Head/Spouse

Various literature was read online and repeated that important variables in determining income would be Age, Education Level, Race, Marital Status, Gender, Occupation, and others. For some of these variables, when browsing the dataset, not enough data existed for them, or they were not included in the data. After browsing the data and finding the relevant variables with enough data that could have missing values be imputed or adjusted, we came to this conclusion of variables. When building our models SAS might determine some of our variables aren't statistically significant and not include them, which is fine. We would like to give SAS a wide range of potentially relevant variables so that it can go through the variable selection process.

## *Data Preparation*

1. **Handling Missing Values:**
   a. The codebook accompanying the data download allowed us to indicate which values should be treated as missing in our dataset. Codes such as 99, 9, or 0 usually showed up as N/A in the dataset. Specifically, for our data Head/Spouse degree 9 was coded as missing, Head/Spouse race 9 was coded as missing, Head/Spouse Father/Mother Education 99 were coded as missing, Head/Spouse religion 99 were coded as missing.
   b. Geographic region codes of 0 were considered missing.
   c. Missing values for numerical variables were imputed with the mean, while mode imputation was used for categorical variables.
2. **Variable Transformation:**
   a. **Log of Income:** To normalize the income distribution and see if it allows the software to create more accurate models
   b. **Age Squared:** Captures non-linear effects of age on income.
   c. **Dummy Variables:** Created for categorical variables (e.g., race, religion, region). SAS Enterprise Miner automatically handles this when variables are labeled as categorical.
   d. **Interaction Terms:** Explored combined effects of education and region on income. These were explored but no other interaction terms were included in

the final models because they did not show up as statistically significant during the variable selection process.

3. **Outlier Handling:**
    a. Outlier thresholds for income were calculated using the IQR method.
    b. Outliers beyond calculated upper and lower bounds were excluded.
    c. For normal total family income target variable, we calculated lower range as –$101483 and upper range as $268174.
    d. For log total family income target variable, we calculated lower range as 12.7447 and upper range as 13.3992.

### *Validation Plan*

- **Data Splitting:** The dataset was divided into training and testing sets to build and evaluate the models. 70% of the data was used for training, and 30% of the data was used for validation.
- **Diagnostics:**
    - Check for multicollinearity among variables. SAS automatically does this to help ensure that model assumptions (e.g., homoscedasticity, normality of residuals) are satisfied.
- **Performance Metrics:**
    - Adjusted R-squared
    - Mean Squared Error (MSE)
    - Schwarz's Bayesian Criterion (SBC)

# Methodology

This study applied both traditional regression techniques and advanced machine learning models to predict family income. All modeling was conducted using SAS Enterprise Miner, which provides a comprehensive suite of tools for data mining and predictive modeling.

## Traditional Models

### *Linear Regression*

- **Ordinary Least Squares (OLS) Regression**: Served as a baseline model to establish a reference for predictive performance.

- **Variable Selection Methods**: We employed stepwise, forward, and backward selection methods to identify significant predictors:
  - **Stepwise Selection**: Variables were added or removed based on statistical significance, balancing model complexity and performance.
  - **Forward Selection**: Variables were added sequentially based on their statistical significance until no additional variables met the entry criteria.
  - **Backward Elimination**: Starting with all candidate variables, non-significant variables were removed sequentially.
- **Polynomial Regression**: Polynomial terms were added to capture non-linear relationships where appropriate. SAS Enterprise Miner facilitated the inclusion of polynomial terms and evaluated their significance during the variable selection process.

## Machine Learning Models

### *Decision Tree*

- **Algorithm Used**: Classification and Regression Trees (CART) algorithm within SAS Enterprise Miner.
- **Pruning**: Trees were pruned based on validation errors to prevent overfitting. The optimal tree size was determined by evaluating the model's performance on the validation set.
- **Splitting Criteria**: The algorithm used variance reduction as the splitting criterion to select the best splits at each node.

### *Gradient Boosting*

- **Methodology**: Gradient boosting models were built by iteratively adding decision trees to correct the errors of previous trees.
- **Parameters**:
  - **Number of Trees**: Determined through experimentation to balance bias and variance.
  - **Learning Rate**: Set to control the contribution of each tree to the final model.
  - **Tree Depth**: Limited to prevent overfitting and improve generalization.

- **Implementation**: SAS Enterprise Miner provided tools to adjust these parameters, such as the Gradient Boosting node, allowing for fine-tuning and evaluation of their impact on model performance.

*Neural Networks*

- **Architectures Tested**: Multiple neural network architectures with different numbers of hidden units (neurons) were tested:
  - **3 Hidden Units**: Tested with and without variable selection, with up to 500 training iterations (the chosen model converged at 287 iterations).
  - **4 Hidden Units**: Forward regression variable selection was used, with up to 500 training iterations (the chosen model converged at 193 iterations).
  - **5 Hidden Units**: Tested with forward regression variable selection, with up to 500 training iterations (the chosen model converged at 25 iterations).
  - **6 Hidden Units**: Also utilized forward regression variable selection, with up to 500 training iterations (the chosen model converged at 34 iterations).
- **Reason for Multiple Architectures**:
  - The performance of neural networks can be sensitive to the architecture, including the number of hidden units. Since determining the optimal architecture is often an experimental process, we tested multiple configurations to identify the model that performs best on our data.
- **Variable Selection**:
  - Models were optimized using forward regression for variable selection within SAS Enterprise Miner. This process helps in identifying the most significant input variables for the neural network.
- **AutoNeural Consideration**:
  - Although SAS Enterprise Miner offers an AutoNeural node that automates the selection of network architecture and training parameters, we found that manually configuring the neural networks yielded better performance. This is because neural network design is often considered an art, requiring experimentation and domain knowledge to fine-tune.
- **Training Details**:
  - **Activation Functions**: Logistic (sigmoid) function was used in the hidden layers, and a linear activation function was used in the output layer.
  - **Training Algorithm**: Models were trained using backpropagation with gradient descent optimization.

  - **Convergence Criteria**: The number of training iterations varied depending on convergence criteria and model complexity.

***Ensemble Models***

- **Purpose**: To enhance predictive performance by combining predictions from multiple models.
- **Models Included in the Ensemble**:
  - Decision Tree
  - Gradient Boosting
  - Forward Selection Regression
  - Polynomial Regression (Forward Selection)
  - Neural Network (without variable selection)
- **Ensemble Method**:
  - Predictions from the individual models were averaged to produce the final prediction.
  - The ensemble approach leverages the strengths of different models and mitigates their individual weaknesses, often resulting in improved accuracy.

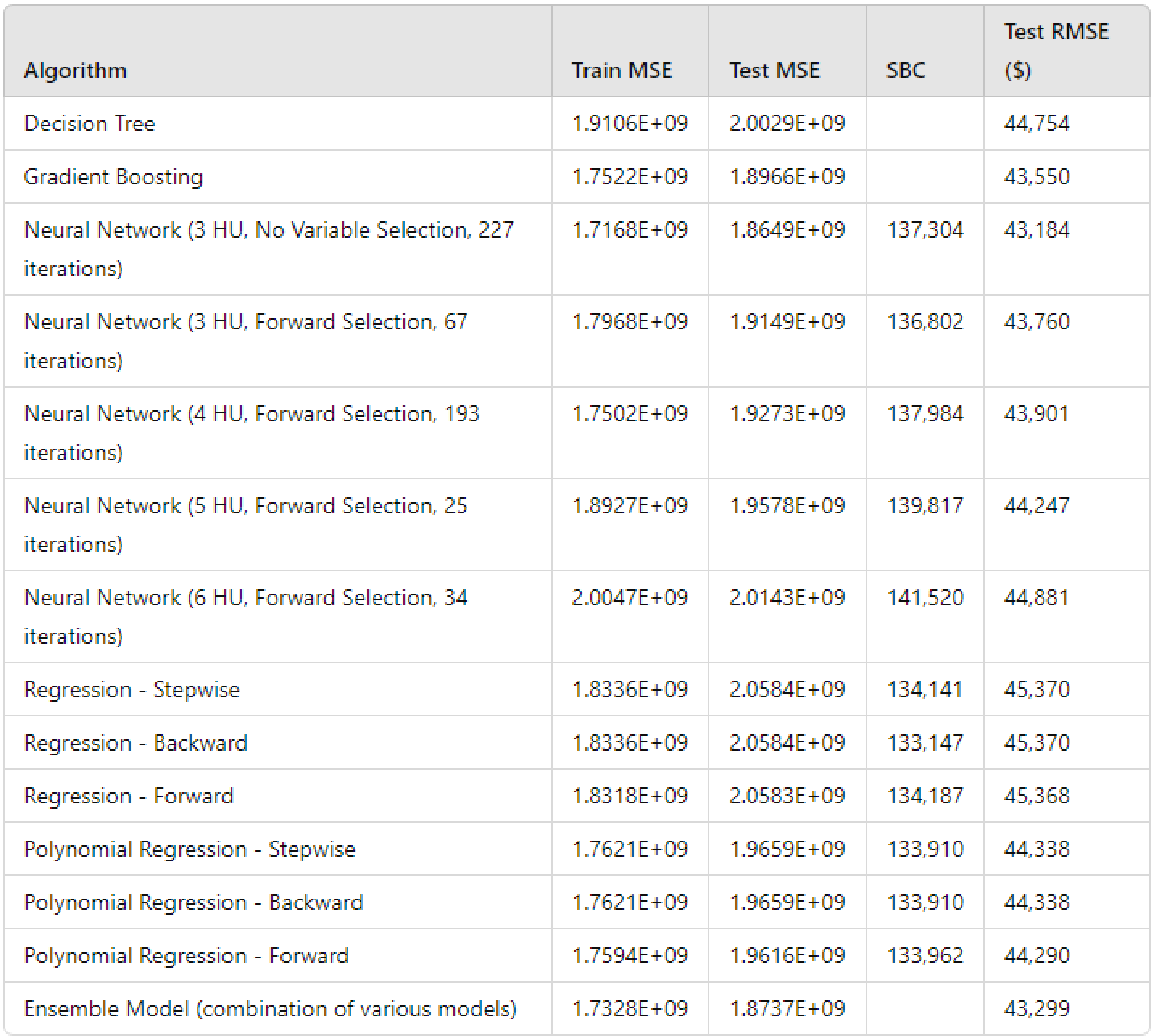
# Results

## Comparison of Models Predicting Total Family Income vs. Log Total Family Income

We developed multiple models to predict family income using both the actual income values and their logarithms. The models were evaluated based on several performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Schwarz's Bayesian Criterion (SBC). The $R^2$ metric is not included for certain models, such as neural networks and tree-based models, because it is less interpretable for non-linear and complex models. Instead, we focus on MSE, RMSE, and SBC to compare model performance.

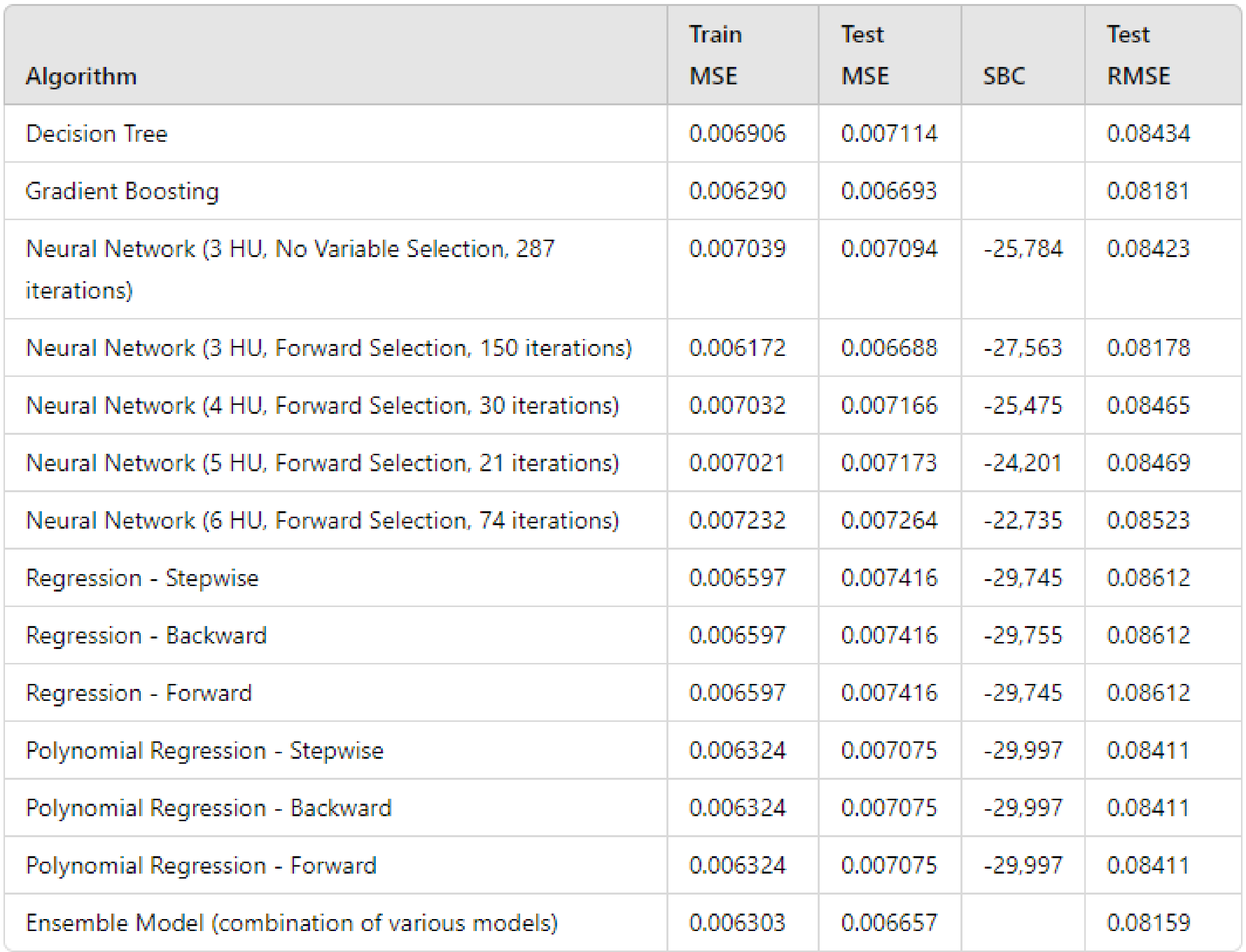
Muzahidul Islam & Mildred Nwachukwu-Innocent
DAT300 Fall 2024

### *Models Predicting Total Family Income*



| Algorithm | Train MSE | Test MSE | SBC | Test RMSE ($) |
|---|---|---|---|---|
| Decision Tree | 1.9106E+09 | 2.0029E+09 | | 44,754 |
| Gradient Boosting | 1.7522E+09 | 1.8966E+09 | | 43,550 |
| Neural Network (3 HU, No Variable Selection, 227 iterations) | 1.7168E+09 | 1.8649E+09 | 137,304 | 43,184 |
| Neural Network (3 HU, Forward Selection, 67 iterations) | 1.7968E+09 | 1.9149E+09 | 136,802 | 43,760 |
| Neural Network (4 HU, Forward Selection, 193 iterations) | 1.7502E+09 | 1.9273E+09 | 137,984 | 43,901 |
| Neural Network (5 HU, Forward Selection, 25 iterations) | 1.8927E+09 | 1.9578E+09 | 139,817 | 44,247 |
| Neural Network (6 HU, Forward Selection, 34 iterations) | 2.0047E+09 | 2.0143E+09 | 141,520 | 44,881 |
| Regression - Stepwise | 1.8336E+09 | 2.0584E+09 | 134,141 | 45,370 |
| Regression - Backward | 1.8336E+09 | 2.0584E+09 | 133,147 | 45,370 |
| Regression - Forward | 1.8318E+09 | 2.0583E+09 | 134,187 | 45,368 |
| Polynomial Regression - Stepwise | 1.7621E+09 | 1.9659E+09 | 133,910 | 44,338 |
| Polynomial Regression - Backward | 1.7621E+09 | 1.9659E+09 | 133,910 | 44,338 |
| Polynomial Regression - Forward | 1.7594E+09 | 1.9616E+09 | 133,962 | 44,290 |
| Ensemble Model (combination of various models) | 1.7328E+09 | 1.8737E+09 | | 43,299 |

***Models Predicting Log Total Family Income***



| Algorithm | Train MSE | Test MSE | SBC | Test RMSE |
|---|---|---|---|---|
| Decision Tree | 0.006906 | 0.007114 | | 0.08434 |
| Gradient Boosting | 0.006290 | 0.006693 | | 0.08181 |
| Neural Network (3 HU, No Variable Selection, 287 iterations) | 0.007039 | 0.007094 | -25,784 | 0.08423 |
| Neural Network (3 HU, Forward Selection, 150 iterations) | 0.006172 | 0.006688 | -27,563 | 0.08178 |
| Neural Network (4 HU, Forward Selection, 30 iterations) | 0.007032 | 0.007166 | -25,475 | 0.08465 |
| Neural Network (5 HU, Forward Selection, 21 iterations) | 0.007021 | 0.007173 | -24,201 | 0.08469 |
| Neural Network (6 HU, Forward Selection, 74 iterations) | 0.007232 | 0.007264 | -22,735 | 0.08523 |
| Regression - Stepwise | 0.006597 | 0.007416 | -29,745 | 0.08612 |
| Regression - Backward | 0.006597 | 0.007416 | -29,755 | 0.08612 |
| Regression - Forward | 0.006597 | 0.007416 | -29,745 | 0.08612 |
| Polynomial Regression - Stepwise | 0.006324 | 0.007075 | -29,997 | 0.08411 |
| Polynomial Regression - Backward | 0.006324 | 0.007075 | -29,997 | 0.08411 |
| Polynomial Regression - Forward | 0.006324 | 0.007075 | -29,997 | 0.08411 |
| Ensemble Model (combination of various models) | 0.006303 | 0.006657 | | 0.08159 |

## Interpretation of Results

***Comparing Models Within Each Target Variable***

### **Total Family Income Models**:

- **Neural Network with 3 Hidden Units and No Variable Selection** achieved the lowest Test MSE (1.8649E+09) and the lowest Test RMSE ($43,184) among all models predicting total family income directly.
- **Polynomial Regression (Forward Selection)** showed competitive performance, with a Test MSE of 1.9616E+09 and a Test RMSE of $44,290.

- The error difference between the neural network and the polynomial regression model is approximately $1,106, which is relatively small considering the scale of income values.
- **Standard Regression Models** had higher Test MSE values around 2.0584E+09 and higher Test RMSE values around $45,370.

## **Log Total Family Income Models**:

- The **Neural Network with 3 Hidden Units and Forward Variable Selection** achieved the lowest Test MSE (0.006688) and Test RMSE (0.08178).
- **Polynomial Regression (Forward Selection)** had a Test MSE of 0.007075 and a Test RMSE of 0.08411, performing comparably to the neural network model.
- The difference in RMSE between the neural network and the polynomial regression model is only 0.00233, which translates to a minimal difference in percentage error.

### ***Comparing Total Family Income vs. Log Total Family Income Models***

- Models predicting **Log Total Family Income** generally achieved lower MSE and RMSE values due to the logarithmic scale, which reduces the impact of extreme values and stabilizes variance.
- The **Neural Network with 3 Hidden Units and Forward Variable Selection** performed best among the log income models, while the **Neural Network with 3 Hidden Units and No Variable Selection** performed best among the total income models.
- The **Ensemble Models** for both target variables showed improved performance by combining the strengths of individual models.

## **Interpretation of Test RMSE for Log Income**

The Test RMSE for the log-transformed income models represents the average deviation of the predicted log income from the actual log income. To interpret this value in terms of percentage error, we can use the following approximation:

$$\text{Percentage Error} \approx \left(e^{\text{RMSE}} - 1\right) \times 100\%$$

**Example Calculation for the Best Log Income Model - Polynomial Regression Model (Forward Selection):**

- **Test RMSE**: 0.08411
- **Percentage Error**:

$$(e^{0.08411} - 1) \times 100\% = (1.0877 - 1) \times 100\% = 8.77\%$$

This indicates that on average, the polynomial regression model's predictions are within approximately **8.77%** of the actual income values, which is only slightly higher than the neural network's 8.52%.

## Interpreting the Regression Model

### *Polynomial Regression Model for Log Total Family Income*

The polynomial regression model predicting **Log Total Family Income** includes several significant variables. Below is the model equation with the most significant variables, and a note that other variables are included but not shown for brevity.

#### Model Equation

$$\begin{aligned}\hat{Y} = 12.9623 &+ 0.00647 \times \text{Head Age} - 0.00006 \times (\text{Head Age})^2 \\ &- 0.0243 \times D_{\text{Head Degree}=0} + 0.0303 \times D_{\text{Head Degree}=1} \\ &- 0.0207 \times D_{\text{Head Race}=2} + 0.0364 \times D_{\text{Head Race}=4} \\ &+ 0.00731 \times \text{Head Sex} - 0.0612 \times D_{\text{Marital Status}=2} \\ &+ [\text{Other Terms}]\end{aligned}$$

Note: $D_{\text{Variable}=k}$ is a dummy variable equal to 1 if the variable equals category $k$, and 0 otherwise. $[\text{Other Terms}]$ include additional variables and coefficients not listed here, such as other education levels, spouse's characteristics, religions, and state indicators.

#### Interpretation of Significant Variables

- **Intercept (12.9623)**: Represents the baseline log income when all predictors are at their reference levels.
- **Head Age**:
    - **Coefficient**: 0.00647
    - **Interpretation**: Holding all else constant, each additional year of the head's age is associated with a **0.647%** increase in total family income $(e^{0.00647} - 1 \approx 0.0065)$
- **Head Age Squared**:
    - **Coefficient**: −0.00006
    - **Interpretation**: The negative coefficient indicates diminishing returns to income with increasing age.
- **Head Degree**:
    - **Degree Level 0 (No Degree)**:
        - **Coefficient**: -0.0243
        - **Interpretation**: Heads without a degree have a **2.4%** lower income compared to those with a degree, holding all else constant.
- **Head Race**:
    - **Race Category 2(Black)**:
        - **Coefficient**: −0.0207
        - **Interpretation**: Heads identified as Black earn **2.0%** less than the reference race category (White), holding all else constant.
    - **Race Category 4(Asian)**:
        - **Coefficient**: 0.0364
        - **Interpretation**: Heads identified as Asian earn **3.7%** more than the reference race category (White), holding all else constant.
- **Head Sex**:
    - **Coefficient**: 0.00731
    - **Interpretation**: Families with a male head earn **0.73%** more than those with a female head, holding all else constant.
- **Marital Status**:
    - **Status 2(Single)**:
        - **Coefficient**: 0.0612
        - **Interpretation**: This status is associated with a **6.0%** decrease in income compared to the reference marital status(married), holding all else constant.

***Polynomial Regression Model for Total Family Income***

**Model Equation**

$$\begin{aligned}\hat{Y} = \$14,515.50 + \$3,330.30 \times \text{Head Age} - \$32.4748 \times (\text{Head Age})^2 \\ - \$12,234.50 \times D_{\text{Head Degree}=0} + \$15,590.50 \times D_{\text{Head Degree}=1} \\ - \$9,893.90 \times D_{\text{Head Race}=2} + \$18,624.40 \times D_{\text{Head Race}=4} \\ - \$17,171.20 \times D_{\text{Spouse Degree}=0} + \$16,328.80 \times D_{\text{Spouse Degree}=1} \\ + \$3,799.20 \times \text{Head Sex} - \$28,072.60 \times D_{\text{Marital Status}=2} \\ + [\text{Other Terms}]\end{aligned}$$

### *Interpretation of Significant Variables*

- **Intercept ($14,515.5)**: Represents the baseline family income when all predictors are at their reference levels.
- **Head Age**:
  - **Coefficient**: $3,330.30
  - **Interpretation**: Holding all else constant, each additional year of the head's age increases family income by **$3,330.30**.
  - **Squared Term (−32.4748)**: Indicates that the rate of income increase diminishes with age; income increases at a decreasing rate.
- **Head Degree**:
  - **Degree Level 0 (No Degree)**:
    - **Coefficient**: −$12,234.50
    - **Interpretation**: Heads without a degree earn **$12,234.50** less than those with degree, holding all else constant.
- **Head Race**:
  - **Race Category 2**:
    - **Coefficient**: −$9,893.90
    - **Interpretation**: Heads identified as race category 2(Black) earn **$9,893.90** less than the reference race category(White), holding all else constant.
  - **Race Category 4**:
    - **Coefficient**: $18,624.40

    - **Interpretation**: Heads identified as race category 4(Asian) earn **$18,624.40** more than the reference race category(White), holding all else constant.

- **Spouse Degree**:
    - **Degree Level 0 (No Degree)**:
        - **Coefficient**: −$17,171.20
        - **Interpretation**: Spouses without a degree contribute to a **$17,171.20** lower family income compared to those with degree, holding all else constant.
- **Head Sex**:
    - **Coefficient**: $3,799.20
    - **Interpretation**: Families with a male head earn **$3,799.20** more than those with a female head, holding all else constant.
- **Marital Status**:
    - **Status 2(Single)**:
        - **Coefficient**: −$28,072.60
        - **Interpretation**: This status is associated with a **$28,072.60** decrease in income compared to the reference marital status(married), holding all else constant.

## Summary of Key Findings

- **Education**: Both the head's and spouse's educational attainment are strong predictors of family income. Higher education levels are associated with significantly higher income.
- **Age**: Family income increases with the head's age but at a decreasing rate, indicating that income growth slows down as the head gets older.
- **Gender**: Male heads of household earn more than female heads, even when controlling for other factors.
- **Race**: Certain race categories are associated with significant differences in income, highlighting racial disparities in earnings.
- **Marital Status**: Marital status has a substantial impact on family income, with some statuses associated with significantly lower income.
- **Spouse's Family Background**: The spouse's degree and father's education level also play a role in family income, suggesting the influence of family background.

## Discussion

The results demonstrate that while advanced machine learning models like neural networks offer high predictive accuracy, traditional regression models—specifically the **polynomial regression with forward selection**—provide a valuable balance between accuracy and interpretability. Although the neural network models achieved slightly lower Test MSE and RMSE values, the differences were minimal.

**Importance of Interpretability:**

- The polynomial regression model allows for clear insights into how individual variables affect family income.
- This interpretability is crucial for policy analysis, socioeconomic studies, and decision-making processes where understanding the impact of specific factors is as important as the prediction itself.

**Polynomial Regression Advantages:**

- **Captures Non-Linear Relationships:** Including polynomial terms helps model the non-linear effects of variables like age on income.
- **Variable Selection:** Forward selection identifies the most significant predictors, simplifying the model without sacrificing much accuracy.

**Model Complexity vs. Performance:**

- While neural networks can model complex patterns, they act as "black boxes," making it difficult to interpret the influence of individual variables.
- The polynomial regression model, though slightly less accurate, provides transparency in how predictions are made.

### *Interpretation of Predictive Factors*

The regression models provided insights into factors influencing income:

- **Education**: Both the head's and spouse's educational attainment are strong predictors of income.
- **Age**: Income increases with age but at a decreasing rate, highlighting the importance of experience.

- **Gender**: Gender disparities exist, with males earning more on average.
- **Marital Status**: Marital status impacts income, possibly reflecting differences in household responsibilities or social factors.
- **Race and State**: Certain race categories and states are associated with significant differences in income, indicating regional and demographic disparities.

## Conclusion

The comparative analysis of various predictive models for family income indicates that the **polynomial regression model with forward selection** is the preferred choice for this study. Despite neural networks achieving marginally better predictive performance—with the Neural Network model attaining a Test RMSE of 0.08178—the polynomial regression model's Test RMSE of 0.08411 is comparably close. The minimal difference in error rates (approximately 0.25% in percentage error) does not outweigh the benefits gained from the interpretability of the regression model.

When predicting family income, especially for applications requiring policy analysis and strategic planning, understanding the impact of individual variables is essential. The polynomial regression model offers this interpretability, providing clear insights into how factors such as age, education, race, and marital status influence income.

The log-transformed models effectively reduce the impact of outliers and stabilize variance, enhancing model performance. The polynomial regression with forward selection captures non-linear relationships and focuses on significant predictors, ensuring both accuracy and simplicity.

In the initial stages of this study, variables were selected based on their presumed relevance from existing literature and domain knowledge. This approach, while focused, limited the amount of data available for model training. In future research, incorporating a broader range of variables—even those that may not seem immediately relevant—could enhance model performance. Machine learning models can uncover hidden patterns and relationships that are not apparent through traditional analysis. Including more variables could provide the models with additional information to improve predictions and potentially reveal novel insights into the determinants of family income.

Overall, the polynomial regression model with forward selection strikes an optimal balance between predictive accuracy and interpretability. It is recommended for predicting family income when both precise predictions and an understanding of variable impacts are desired. This approach ensures that the models are not only accurate but also actionable, providing valuable insights for stakeholders.

**References**

- Black, S. E., Devereux, P. J., & Salvanes, K. G. (2005). Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *IZA Discussion Paper* No. 1390. https://docs.iza.org/dp926.pdf
- Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics, 129*(4), 1553–1623. https://doi.org/10.1093/qje/qju022
- Gasperi, A. (2019). Predicting income levels: A logistic regression analysis in R. *Medium*. https://medium.com/@alessio.gasperi81_25968/predicting-income-levels-a-logistic-regression-analysis-in-r-bdacdb64735d
- Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review, 94*(3), 526–556. https://www.aeaweb.org/articles?id=10.1257/0002828041464498
- Greenman, A., Xie, Y. (2015). Double Jeopardy? The Interaction of Gender and Race on Earnings in the U.S. *Social Science Research, 53*, 363–374. https://pmc.ncbi.nlm.nih.gov/articles/PMC4631221/
- Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research. https://www.nber.org/system/files/chapters/c1765/c1765.pdf
- Reeves, R. V., & Sawhill, I. V. (2016). Men's earnings and children's poverty. *Brookings*. https://www.brookings.edu/wp-content/uploads/2016/06/disparities.pdf