PHASE 2 MODELLING & ANALYTICS (Hima Varsha)

1. **Overview**

   The objective of Phase 2 was to build predictive models to understand which economic factors influence **AI Exposure (AIGE)** across U.S. counties.

   Our goal was not only to fit models but also to evaluate how much variance in AIGE can actually be explained by available data, identify key patterns, and provide realistic insights based on statistical evidence.

   The models included:
   - Multiple Linear Regression
   - Decision Tree Regression
   - Neural Networks (1- Layer & 2- Layer)
   - Feature engineering of economic indicators
   - Comparison of model performance

---

2. **Tools & Methods**

Tools:
- R (Rstudio)
- tidyverse
- caret
- rpart
- Neuralnet
- Ggplot2
- ggcorrplot

Methods Used:
- Data preprocessing
- Feature engineering
- Correlation analysis
- Regression diagnostics
- Tree-based modeling
- Neural Network modeling
- Model comparison using RMSE,MAE,R², ME
- Exporting results for Phase 3

---

3. Data Preparation

The dataset provided initially contained AIGE, county identifiers, labor force variables, employment counts, and median household income.

Key steps performed

- Checked structure, missing values, and numeric fields
- Removed Identifiers unnecessary for modeling
- Normalized variables specifically for neural networks
- Split data into training and testing sets (80/20)

---

4. Feature Engineering (New Variables Added)

To strengthen the model and generate deeper insights, not originally present in the raw file, we engineered two additional variables using existing columns:

Employment Rate(%):

Measures the share of employed individuals within each county's labor force.

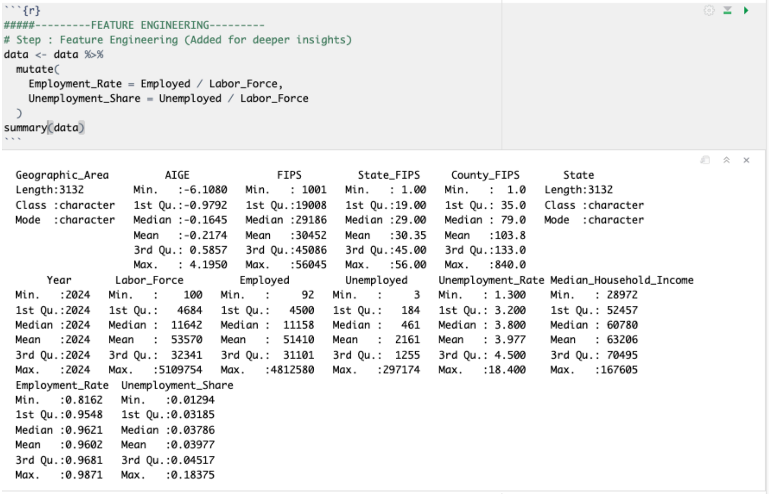
Employment Rate = Employed / Labor Force

Unemployment Share (%):

Measures the share of unemployed individuals within the labor force.

Unemployment Share = Unemployed / Labor Force

These variables help quantify workforce stability in a county and allow us to interpret labor-driven patterns more clearly.



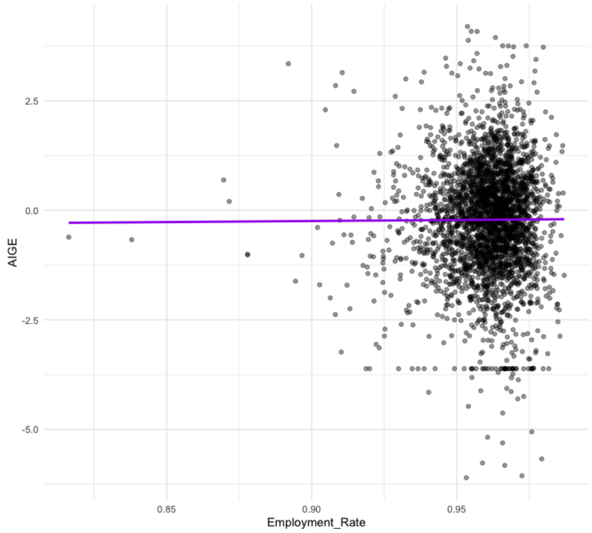
~~~
```{r}
#####---------FEATURE ENGINEERING---------
# Step : Feature Engineering (Added for deeper insights)
data <- data %>%
  mutate(
    Employment_Rate = Employed / Labor_Force,
    Unemployment_Share = Unemployed / Labor_Force
  )
summary(data)
```
~~~

```
 Geographic_Area         AIGE              FIPS          State_FIPS      County_FIPS       State          
 Length:3132        Min.   :-6.1080   Min.   : 1001   Min.   : 1.00   Min.   :  1.0   Length:3132       
 Class :character   1st Qu.:-0.9792   1st Qu.:19008   1st Qu.:19.00   1st Qu.: 35.0   Class :character  
 Mode  :character   Median :-0.1645   Median :29186   Median :29.00   Median : 79.0   Mode  :character  
                    Mean   :-0.2174   Mean   :30452   Mean   :30.35   Mean   :103.8                     
                    3rd Qu.: 0.5857   3rd Qu.:45086   3rd Qu.:45.00   3rd Qu.:133.0                     
                    Max.   : 4.1950   Max.   :56045   Max.   :56.00   Max.   :840.0                     
      Year       Labor_Force          Employed          Unemployed       Unemployment_Rate Median_Household_Income
 Min.   :2024   Min.   :    100   Min.   :     92   Min.   :     3   Min.   : 1.300    Min.   : 28972         
 1st Qu.:2024   1st Qu.:   4684   1st Qu.:   4500   1st Qu.:   184   1st Qu.: 3.200    1st Qu.: 52457         
 Median :2024   Median :  11642   Median :  11158   Median :   461   Median : 3.800    Median : 60780         
 Mean   :2024   Mean   :  53570   Mean   :  51410   Mean   :  2161   Mean   : 3.977    Mean   : 63206         
 3rd Qu.:2024   3rd Qu.:  32341   3rd Qu.:  31101   3rd Qu.:  1255   3rd Qu.: 4.500    3rd Qu.: 70495         
 Max.   :2024   Max.   :5109754   Max.   :4812580   Max.   :297174   Max.   :18.400    Max.   :167605         
 Employment_Rate  Unemployment_Share
 Min.   :0.8162   Min.   :0.01294   
 1st Qu.:0.9548   1st Qu.:0.03185   
 Median :0.9621   Median :0.03786   
 Mean   :0.9602   Mean   :0.03977   
 3rd Qu.:0.9681   3rd Qu.:0.04517   
 Max.   :0.9871   Max.   :0.18375   
```

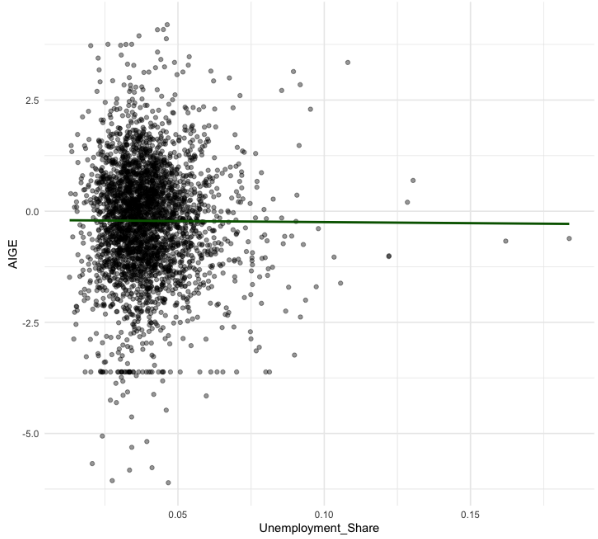
Use in Analysis:

These engineered variables were used during exploratory analysis (correlation matrix and scatterplots) to test whether employment dynamics improved interpretability of AIGE variations. However, due to extremely **high collinearity with the original unemployment rate, they were not included in the final regression or ML models.**

AIGE vs Employment Rate Scatterplot



AIGE vs Unemployment Share Scatterplot



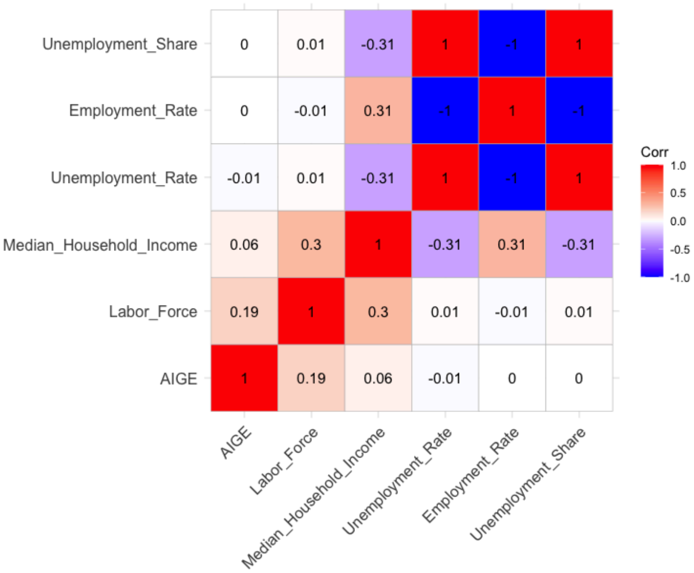
5. Exploratory Data Analysis (EDA)

Correlation Analysis

We computed correlations among all numeric variables including the new engineered ones.

Key Observations:

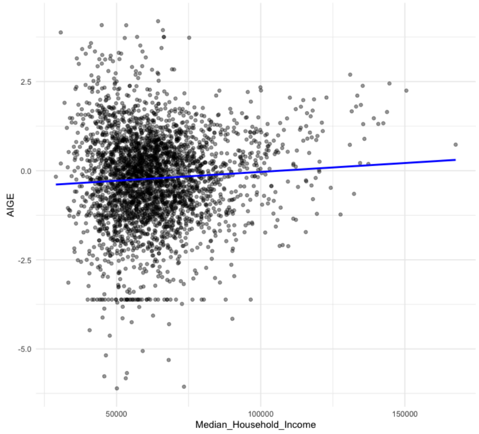
- AIGE has weak linear correlations with income, unemployment rate and employment metrics.
- Labor force shows a small but stronger correlation compared to others.
- The engineered variables also show weak correlations.



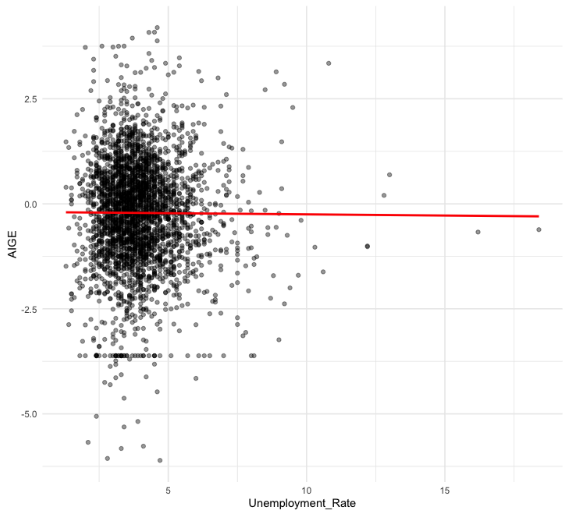
Scatter Plots

Scatter plots were generated to visually inspect relationships

- AIGE vs Median Household Income



- AIGE vs Unemployment Rate



---

6. Modeling Approach

REGRESSION MODELING

Objective

To assess whether county-level economic indicators help explain variation in **AI Exposure (AIGE)**, and to identify the strongest predictors for subsequent modeling.

Predictors evaluated:

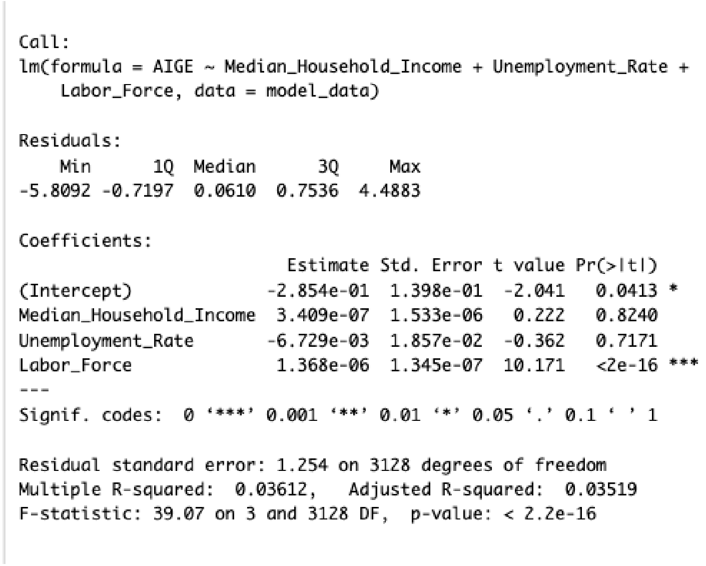
- Median_Household_Income
- Unemployment_Rate
- Labor_Force

Initial Regression Model

Model:

$$AIGE=\beta0+\beta1(Income)+\beta2(Unemployment)+\beta3(Labor\ Force)$$

Summary:



```
Call:
lm(formula = AIGE ~ Median_Household_Income + Unemployment_Rate +
    Labor_Force, data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8092 -0.7197  0.0610  0.7536  4.4883

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.854e-01  1.398e-01  -2.041   0.0413 *
Median_Household_Income  3.409e-07  1.533e-06   0.222   0.8240
Unemployment_Rate       -6.729e-03  1.857e-02  -0.362   0.7171
Labor_Force              1.368e-06  1.345e-07  10.171   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.254 on 3128 degrees of freedom
Multiple R-squared:  0.03612,	Adjusted R-squared:  0.03519
F-statistic: 39.07 on 3 and 3128 DF,  p-value: < 2.2e-16
```

Key Points:

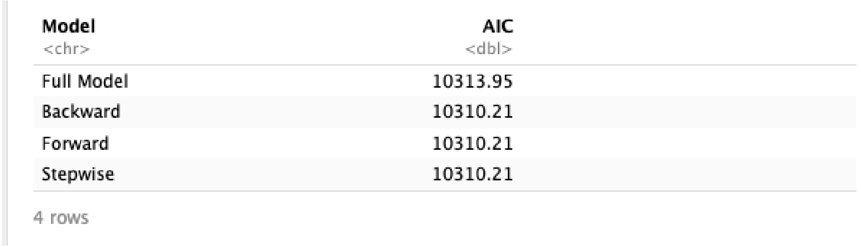
- **Labor_Force is the only statistically significant variable (p < 0.001)**
- Income and Unemployment Rate are not significant
- Adjusted **R² ≈ 0.035**, meaning the predictors explain very little in AIGE

Stepwise Model Selection (Backward,Forward,Both)

Three model- selection procedures were applied to confirm predictor relevance.

Results (all three methods):

- The stepwise consistently retained Labor_Force
- Adding or removing the other variables did not improve model quality (AIC based)



Model <chr>	AIC <dbl>
Full Model	10313.95
Backward	10310.21
Forward	10310.21
Stepwise	10310.21

4 rows

Conclusion:

Labor_Force is the only stable predictor across all selection techniques.

Final Regression Model

The final chosen model remains the 3 variable model, with Labor_force as the only meaningful driver.

$$AIGE=\beta 0+\beta 1(Labor\ Force)+\beta 2(Income)+\beta 3(Unemployment)$$

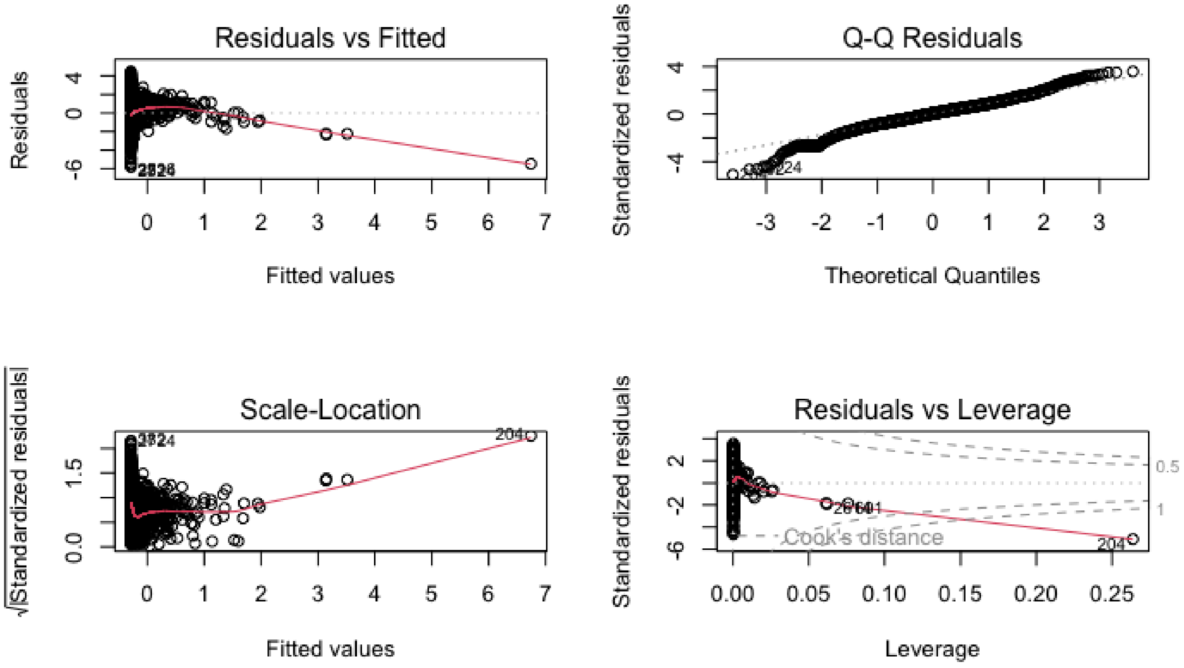
However, due to the low Rsquared, the model has limited predictive power.

Diagnostic Checks

Standard regression diagnostic plots were generated:

- Residual vs Fitted
- Q-Q plot
- Scale-Location
- Residuals vs Leverage

Residual Diagnostic Plots



Findings:

- Residuals are widely spread, indicating weak predictive fit
- Some heteroscedasticity is present
- Several high-residual counties suggest missing structures variables

Interpretation:
Counties with larger Labor Forces tend to have higher AI exposure, but economic indicators alone do not explain AIGE well.

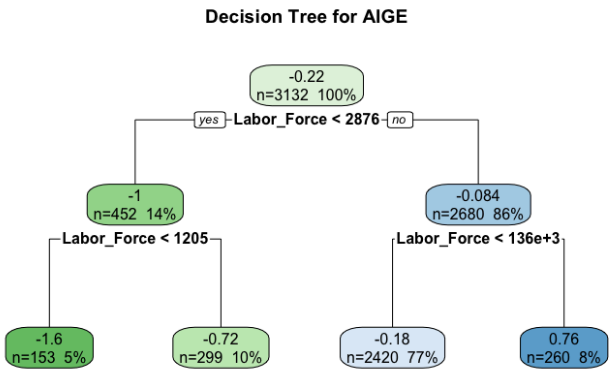
DECISION TREE REGRESSION

A small decision tree model was built to capture non-linear patterns.

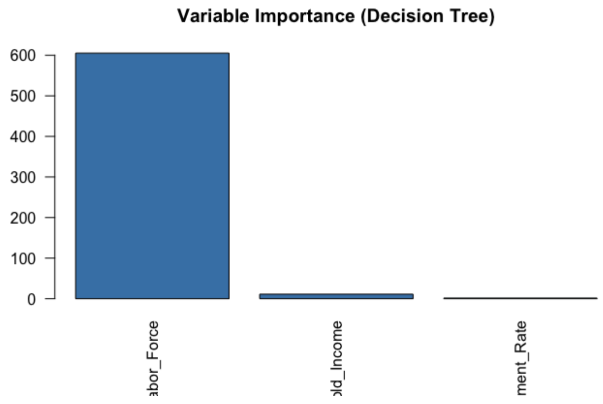
Key Actions

- Fit a regression tree using
  AIGE ~ Median_Household_Income + Unemployment_Rate + Labor_Force
- Visualized the tree and extracted variable importance
- Predicted on test data and computed performance metrics

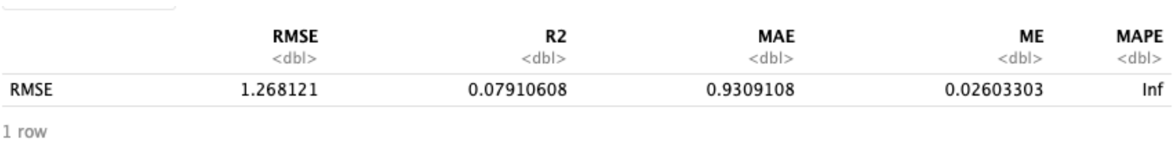
Tree Plot



Variable Importance Plot



Results:



	RMSE <dbl>	R2 <dbl>	MAE <dbl>	ME <dbl>	MAPE <dbl>
RMSE	1.268121	0.07910608	0.9309108	0.02603303	Inf

1 row

Model Behavior:

- The tree exclusively split on Labor_Force, meaning it is the strongest predictor among the available variables
- Performance is slightly better than regression but still low explanatory power Rsquared = 8%

Interpretation:

- **Labor_Force** remains the only variable with significant predictive value
- Simple socioeconomic metrics do no meaningfully explain AIGE variation at the county level
- This supports the insight that AIGE depends on deeper structural factors not present in the dataset (industry composition, digital economy, education)

Rationale:

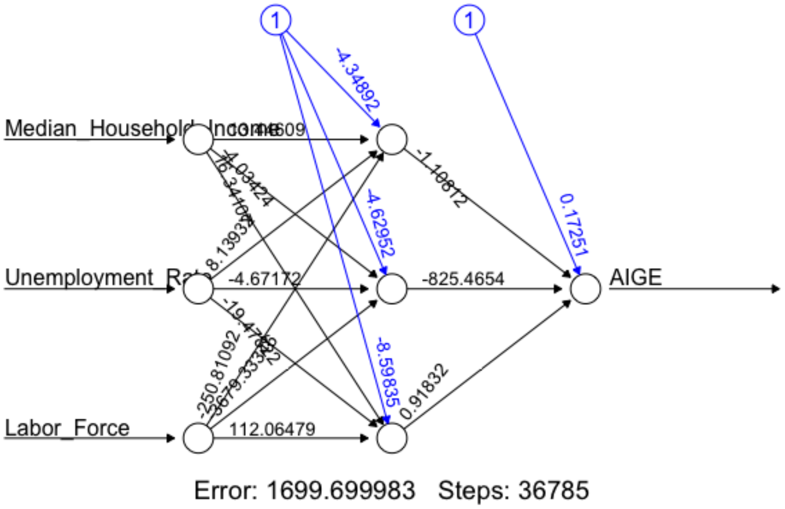
Decision trees test for non-linear and threshold effects.
Even with this flexibility, the model confirms the same conclusion as regression:
AIGE is weakly predicted by county level labor and income variables.

NEURAL NETWORK MODELS

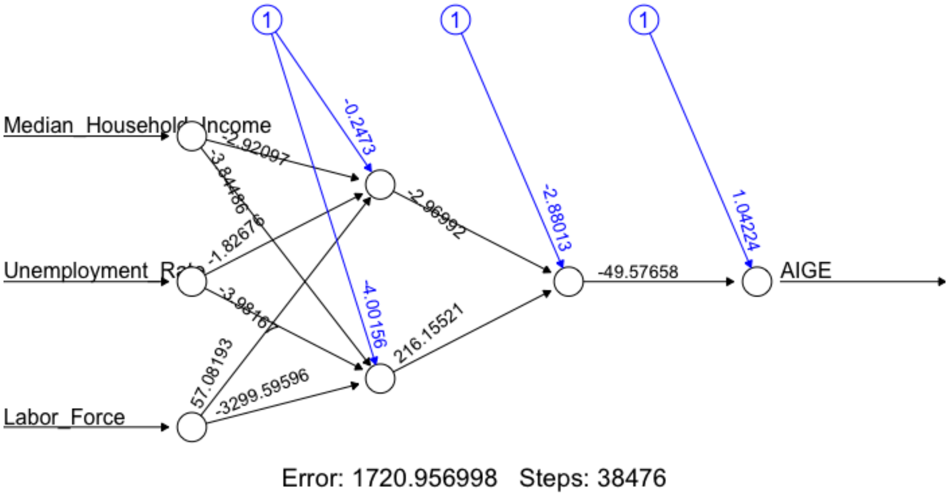
Key Actions:

- Scaled numeric predators using min-max normalization
- Trained two neural networks to test whether non-linear patterns improve prediction
- Generated predictions on test data
- Computed standard performance metrics

NN Architecture Plot (1 Layer)

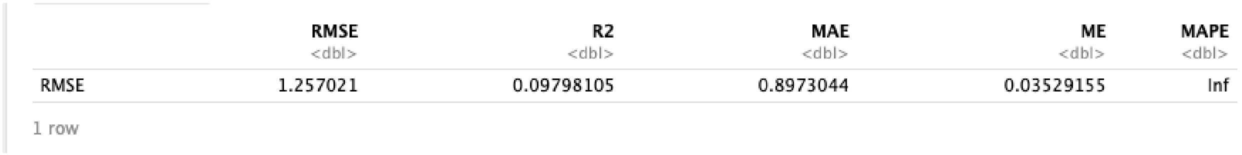


NN Architecture Plot (2 Layers)



Results:

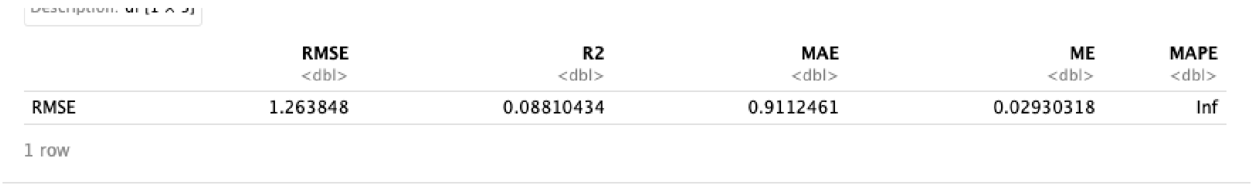
NN 1 Layer



	RMSE <dbl>	R2 <dbl>	MAE <dbl>	ME <dbl>	MAPE <dbl>
RMSE	1.257021	0.09798105	0.8973044	0.03529155	Inf

1 row

NN 2 Layers



Description: df [1 x 5]

	RMSE	R2	MAE	ME	MAPE
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
RMSE	1.263848	0.08810434	0.9112461	0.02930318	Inf

1 row

Model Behavior:

- Both neural networks perform very similarity to the regression and tree models.
- Slight improvement in Rsquared = 0.10, but still too small to indicate meaningful predictive power.

Interpretation

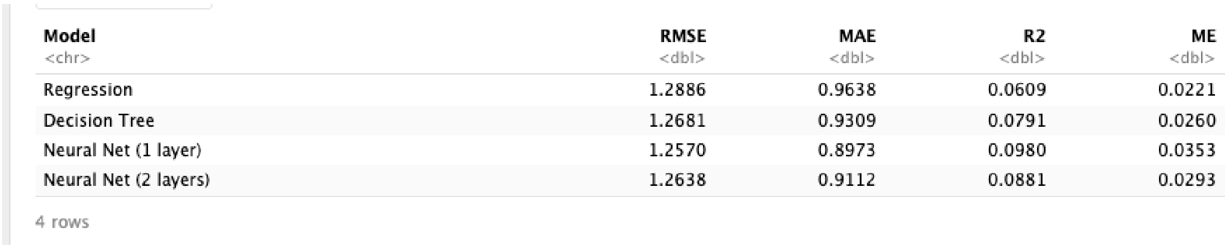
- Even with flexible non-linear modeling, the neural networks do not uncover strong relationships between AIGE and socioeconomic variables
- Neural models confirm the same insight: AIGE is not strongly explained by basic county-level economic indicators.
- Labor_Force remains the only stable signal, but still with very weak predictive strength.

Rationale

Neural networks were included to test whether AIGE exhibits hidden nonlinear structure The consistently low Rsquared across all models predict AIGE

---

7. Combined Model Performance



Model	RMSE	MAE	R2	ME
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Regression	1.2886	0.9638	0.0609	0.0221
Decision Tree	1.2681	0.9309	0.0791	0.0260
Neural Net (1 layer)	1.2570	0.8973	0.0980	0.0353
Neural Net (2 layers)	1.2638	0.9112	0.0881	0.0293

4 rows

---

8. **Final Interpretation (Phase 2 Conclusion)**
    - Across regression, decision trees and neural networks; Labor_force consistently emerged as the strongest predictor of AIGE
    - Income and unemployment rate did not meaningfully improve predictions
    - Non-linear models did not add predictive power suggesting a stable and linear relationships
    - Counties with larger labor forces tend to have higher AI exposure
    - Neural Network (1 layer) performs best statistically

---

1. How does AI exposure (AIGE) vary across U.S. counties?

Answer:
AI exposure varies widely across counties, but most counties have moderately low to average AIGE values, with a small group of counties showing substantially higher exposure. The distribution shows that only a minority of counties have very high AIGE, indicating uneven adoption across regions.

---

2. Do socioeconomic indicators such as labor force size, household income, and unemployment rate explain variation in AI exposure?

Answer:
Socioeconomic indicators explain very limited variation in AIGE. Regression R^2 values are extremely low (≈ 3–9%), meaning that these factors account for only a small fraction of AI exposure differences across counties.

---

3. Which socioeconomic variable is the strongest predictor of AI exposure at the county level?

Answer:
Across all models (regression, decision tree, neural networks), Labor_Force consistently emerges as the only stable predictor, though its predictive power is weak.
Income and unemployment rate provide minimal or no predictive value.

---

4. Can traditional statistical models (regression) reliably predict AI exposure?

Answer:
No.
Linear regression yields very low R² (~0.03) and high RMSE relative to the range of AIGE, indicating that the linear model cannot reliably explain or predict AI exposure.

---

5. Do machine learning methods (decision trees, neural networks) improve prediction accuracy?

Answer:
Only slightly, but not meaningfully.

- **Decision tree: small improvement in R² (~0.07)**
- **Neural network (1-layer): best model but R² only 0.098**
  No model provides strong predictive performance.

---

6. What does model performance indicate about the complexity of AI exposure patterns?

Answer:
Model performance suggests that AIGE is not strongly driven by basic linear or nonlinear relationships with socioeconomic indicators.

AI exposure likely depends on other factors not included in the dataset (e.g., industry composition, tech infrastructure, education levels).

---

7. Do U.S. counties naturally form distinct clusters based on AI exposure and economic characteristics?

Answer:
Yes.
Clustering reveals clear segmentation when using standardized socioeconomic features.

---

8. What socioeconomic profiles define high-exposure vs low-exposure county clusters?

Answer:
From Phase 3 results (k=3 clusters):

- **Cluster 1 (Largest group)**
  Mid-size counties with moderate labor force and average income → moderate AIGE.
- **Cluster 2 (Smaller counties)**
  Low labor force, low income → lowest AIGE.
- **Cluster 3 (Very large counties)**
  Extremely high labor force, high income → highest AIGE.

---

9. Does high AI exposure consistently correspond to larger labor markets or higher-income regions?

Answer:
Yes, partly.

Counties with larger labor forces and higher incomes tend to show higher AIGE, although income alone is not a strong predictor in modeling.
Labor force size shows the clearest association.

10. What county-level characteristics are most associated with higher AI exposure?

Answer:

- **Large labor force**
- **Higher economic activity**
- **Higher employment counts**

These counties are typically urban, economically active, and have broader job diversity, which aligns with more AI adoption.

11. Are there nonlinear or hidden patterns in the data that advanced models can uncover?

Answer:
Not significantly.
Neural networks (even 2-layer) do not uncover meaningful hidden structure.
This indicates that AIGE is not strongly structured by the available socioeconomic features.

12. What practical implications do these patterns have for workforce development and policy planning?

Answer:

- **AI exposure is not evenly distributed across the U.S.**

- **Counties with larger, dynamic labor markets may need more reskilling and AI readiness programs.**

- **Smaller, low-income counties currently face lower exposure, but may also have less capacity to adopt AI in the future.**