Exploring U.S. Labor and Housing Trends: ACS 2018 Insights

By: Twinkle Patel
Instructor: Dr. Lizhong Peng

Objective

The goal of this analysis is to generate some insights into the U.S. labor and housing markets by focusing on labor force participation among individuals aged 18-64. Using the American Community Survey (ACS) dataset, we will identify significant demographic, economic, and geographic factors that influence labor force participation. Additionally, the analysis will evaluate the accuracy of predictive models in forecasting labor force trends. The insights generated will help to guide policy discussions and enhance understanding of the relationship between labor force participation and housing market dynamics.

- Outcome : Labor force participation (labforce)

Data preparation

Subsetting the Data: The original dataset was filtered to include only individuals aged in between 18-64.

Sampling: A 10% random sample was taken from the filtered data to perform analysis.

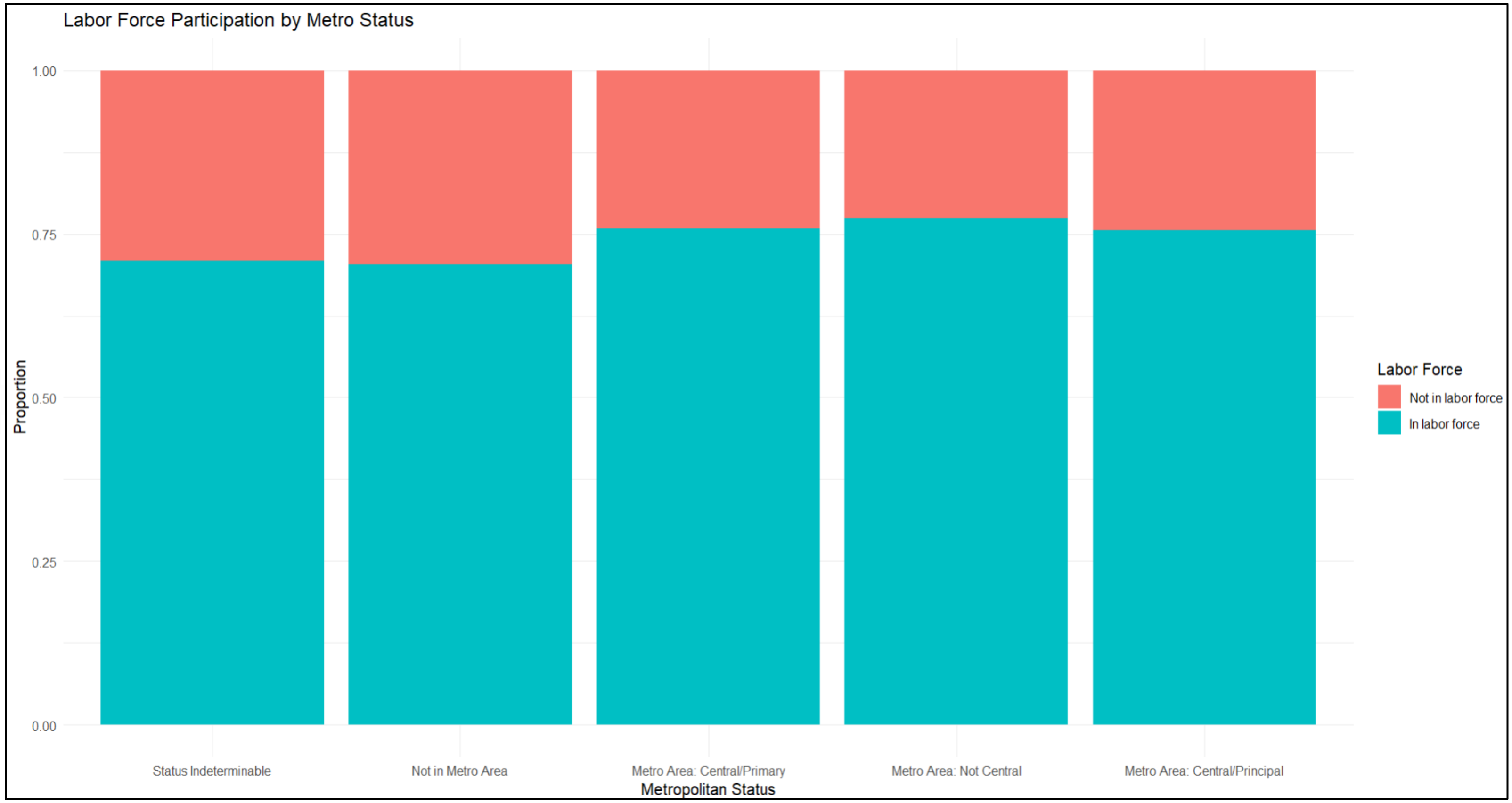
Feature Engineering: Created a binary variable, labforce_dummy, to represent labor force participation (1 = no, 2 = yes).

Handling Missing Data: The dataset was cleaned by removing missing values.

Descriptive analysis

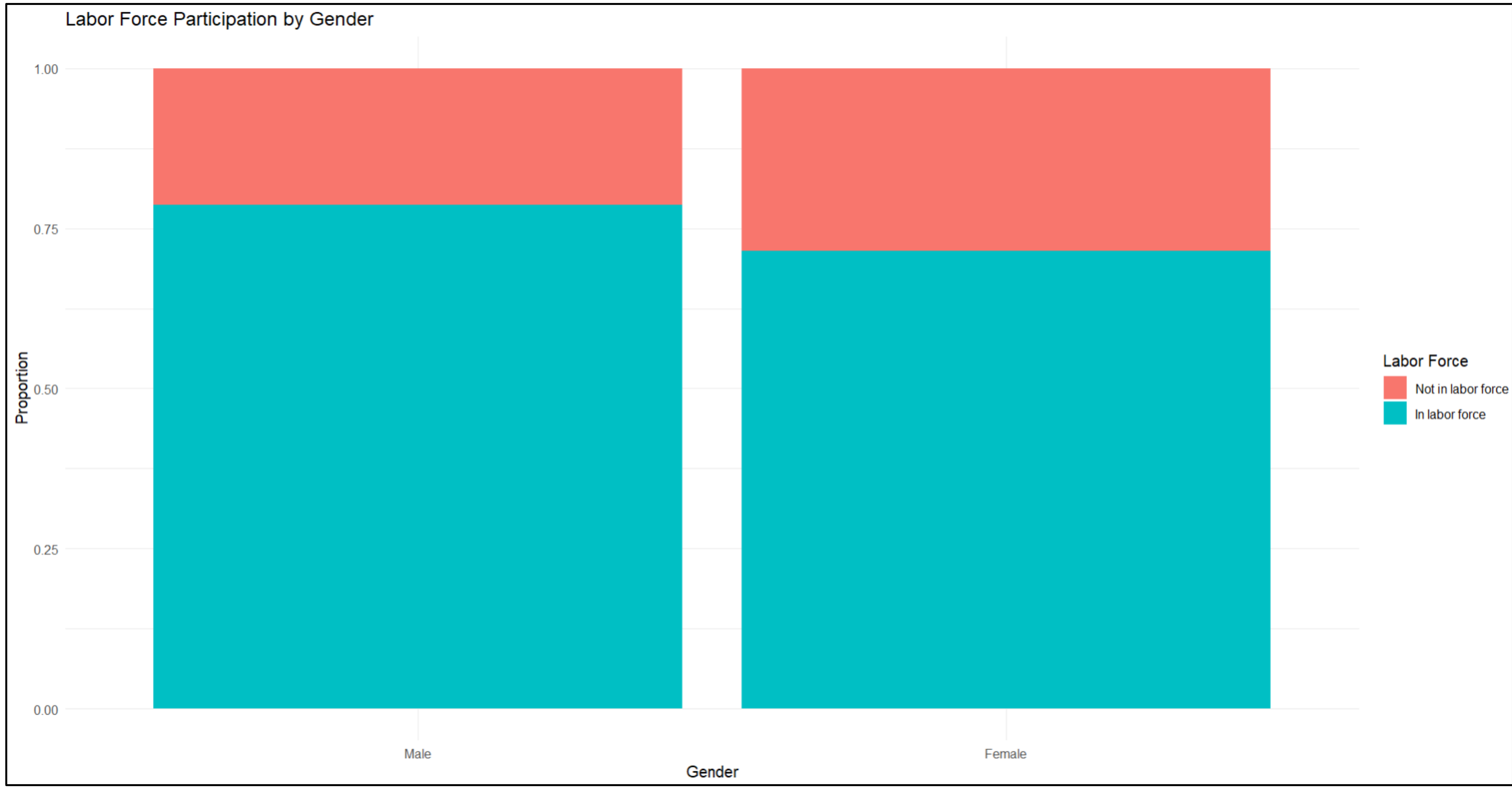
Labor Force Participation by Metropolitan Status

Understanding these patterns can help for employment strategies and economic policy. This graph will help us to understand how our outcome variable labor force related to the different location and as you can see location plays a significant role in employment trend .



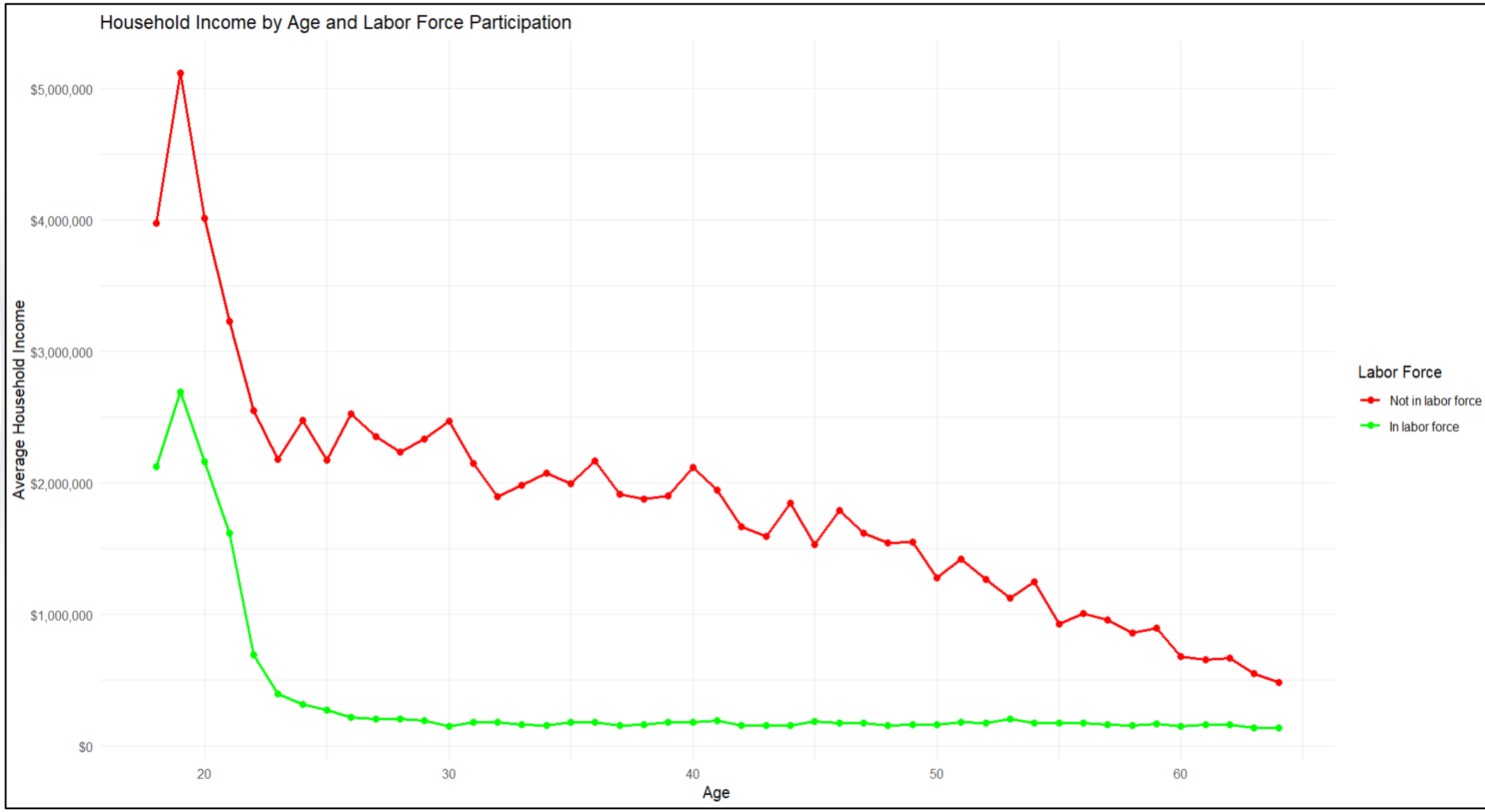
Labor Force Participation by Gender

A bar chart showing labor force participation by gender visually represents the count of individuals in the workforce.

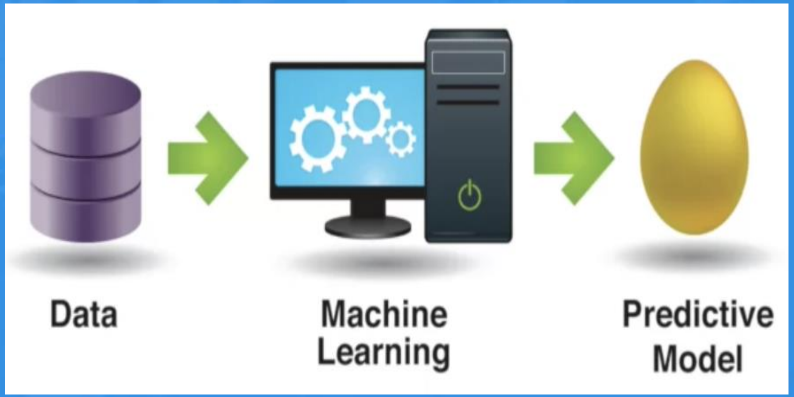


Labor Force Participation by Age and Average Household Income

This line distribution will help to understand the how age and Average household income relate to outcome variable.

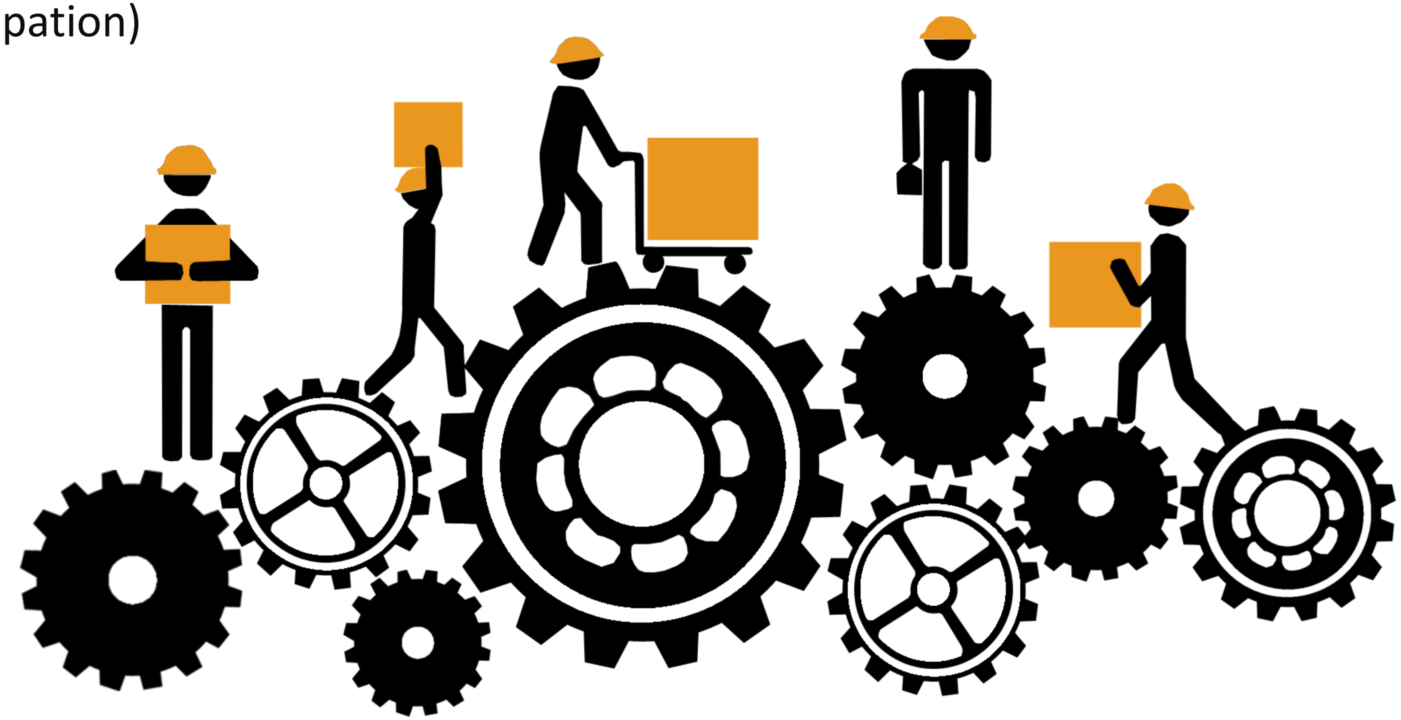


Predictive Modelling



Logistic Regression

Purpose: To determine the relationship between a binary outcome (in this case, labor force participation) and multiple predictors.



- age
- sex
- marst (marital status)
- educ (education)
- metro (metropolitan status)
- Race
- hispan (Hispanic ethnicity)
- uhrswork (usual hours worked)
- poverty
- classwkr (class of worker)

Result

summary(model)

Result

```
Coefficients:
              Estimate Std. Error  z value Pr(>|z|)    
(Intercept)  3.572e+00  6.927e-02   51.566  < 2e-16 ***
age          9.325e-03  7.134e-04   13.070  < 2e-16 ***
sex         -1.514e-01  1.877e-02   -8.068 7.16e-16 ***
marst       -5.618e-03  4.877e-03   -1.152  0.24933    
educ        -3.843e-02  4.325e-03   -8.886  < 2e-16 ***
metro1      -6.616e-02  3.672e-02   -1.802  0.07162 .  
metro2      -3.010e-01  3.553e-02   -8.470  < 2e-16 ***
metro3      -2.047e-01  2.959e-02   -6.917 4.62e-12 ***
metro4      -1.947e-01  2.853e-02   -6.825 8.79e-12 ***
race        -1.474e-02  4.925e-03   -2.993  0.00276 ** 
hispan      -8.576e-02  1.117e-02   -7.679 1.60e-14 ***
uhrswork    -1.046e-01  5.921e-04 -176.607  < 2e-16 ***
poverty     -2.116e-03  5.693e-05  -37.162  < 2e-16 ***
classwkr    -8.729e-01  1.383e-02  -63.101  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 216351  on 192399  degrees of freedom
Residual deviance:  87028  on 192386  degrees of freedom
AIC: 87056

Number of Fisher Scoring iterations: 6
```

Our logistic regression analysis shows that most variables in our model significantly affect labor force participation, with a p-value threshold of 0.05. However, marital status (**marst**) and the first category of metropolitan status (**metro1**) are not significant, suggesting that these factors may not play a substantial role in our outcome

Accuracy: Our model achieved an accuracy of approximately 92.9%. This high accuracy indicates that the model is generally reliable in predicting labor force participation.

print(accuracy)

Result

```
> print(accuracy)
[1] 0.9290644
```

Confusion Matrix

This matrix compares the model's predictions to the actual outcomes.

- True Positives (Predicted to be in the labor force and actually are): 39,895
- True Negatives (Predicted not to be in the labor force and actually aren't): 138,857
- False Positives (Predicted to be in the labor force but aren't): 7,806
- False Negatives (Predicted not to be in the labor force but are): 5,842

```
Confusion Matrix and Statistics

          Reference
Prediction      1      2
         1  39895   7806
         2   5842 138857

               Accuracy : 0.9291
                 95% CI : (0.9279, 0.9302)
    No Information Rate : 0.7623
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8071

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8723
            Specificity : 0.9468
         Pos Pred Value : 0.8364
         Neg Pred Value : 0.9596
             Prevalence : 0.2377
         Detection Rate : 0.2074
   Detection Prevalence : 0.2479
      Balanced Accuracy : 0.9095

       'Positive' Class : 1
```

Cross-Validation with train()

Purpose: To assess model stability and prevent overfitting by training the model on different subsets of the data.

Accuracy: The average accuracy from the 10-fold cross-validation is 92.90749%, which is consistent with my earlier result, indicating that the model is accurate.

```
Generalized Linear Model

192400 samples
    10 predictor
     2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 173159, 173161, 173160, 173160, 173160, 173160, ...
Resampling results:

  Accuracy   Kappa
  0.9290749  0.8071348
```

Splitting Data

80% training set

20% testing set

Recursive Feature Elimination (RFE)

Purpose: To select the most important features for the model by recursively removing the least important ones.

- Key Findings of performing model is the best subset size is eight variables, which has the highest accuracy of 0.9455 and a kappa of 0.8498.
- The top five variables (out of ten) selected by the RFE process are, Uhrswork, poverty, educ, classwkr, age

<u>Result using Train dataset</u>

```
Recursive feature selection

Outer resampling method: Cross-Validated (5 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD  KappaSD Selected
         1   0.9394 0.8300   0.002282 0.006810
         2   0.9437 0.8456   0.001994 0.005805
         3   0.9441 0.8467   0.002328 0.006903
         4   0.9435 0.8451   0.002129 0.006078
         5   0.9446 0.8481   0.002354 0.006762
         6   0.9453 0.8495   0.002400 0.006983
         7   0.9453 0.8495   0.002307 0.006726
         8   0.9455 0.8498   0.002243 0.006542        *
         9   0.9448 0.8484   0.002093 0.005966
        10   0.9451 0.8489   0.002119 0.006090

The top 5 variables (out of 8):
   uhrswork, poverty, educ, classwkr, age
```

RFE Model Evaluation: Confusion Matrix

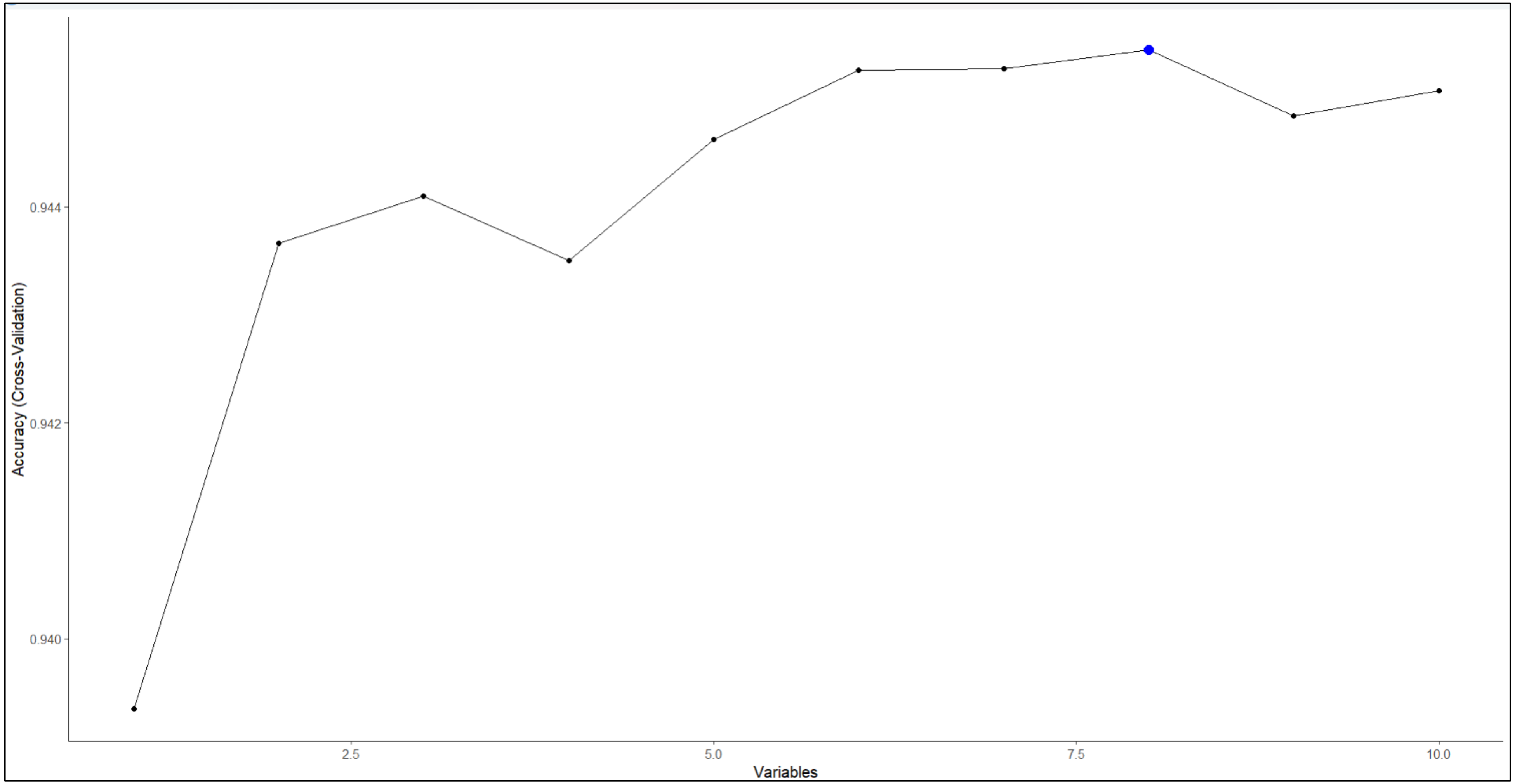
- **Accuracy**: The overall rate of correct predictions. In this case, the accuracy is 94.57%, indicating a high level of correct predictions.

- The results suggest that the Random Forest model is highly accurate in predicting the labor force participation, with strong performance in terms of sensitivity, specificity, and other relevant metrics.

<u>Result using Test dataset</u>

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 28233   678
         1  1413  8156

               Accuracy : 0.9457
                 95% CI : (0.9433, 0.9479)
    No Information Rate : 0.7704
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8507

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9523
            Specificity : 0.9233
         Pos Pred Value : 0.9765
         Neg Pred Value : 0.8523
             Prevalence : 0.7704
         Detection Rate : 0.7337
   Detection Prevalence : 0.7513
      Balanced Accuracy : 0.9378

       'Positive' Class : 0
```

RFE: Plotting Accuracy



Decision Tree

Purpose: To create a simple model that divides the data into subsets based on the most informative way.

The first model is selected based on lowest cp value (0.00015)with Hight accuracy of 94.47%.

Result using Train dataset

```
CART

153920 samples
    10 predictor
     2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 138528, 138528, 138528, 138527, 138527, 138529, ...
Resampling results across tuning parameters:

  cp            Accuracy   Kappa
  0.0001573482  0.9447765  0.8480494
  0.0001835729  0.9447570  0.8481027
  0.0002097975  0.9446791  0.8478592
  0.0002884716  0.9445231  0.8473452
  0.0003409210  0.9444971  0.8475762
  0.0003802581  0.9445296  0.8476450
  0.0004370782  0.9444582  0.8473624
  0.0011801112  0.9440034  0.8459887
  0.0057956572  0.9421062  0.8400594
  0.7551924892  0.8453157  0.4136379

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.0001573482.
```

Decision Tree: Confusion Matrix

- **Accuracy**: The overall accuracy of the Decision Tree model is 94.48%, indicating that it correctly predicted the outcome for nearly 95% of the test data.

<u>Result using Test dataset</u>

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 28205  1419
         1   706  8150

               Accuracy : 0.9448
                 95% CI : (0.9424, 0.947)
    No Information Rate : 0.7513
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8484

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9756
            Specificity : 0.8517
         Pos Pred Value : 0.9521
         Neg Pred Value : 0.9203
             Prevalence : 0.7513
         Detection Rate : 0.7330
   Detection Prevalence : 0.7699
      Balanced Accuracy : 0.9136

       'Positive' Class : 0
```

Best model and conclusion

<u>Comparing Models based on Accuracy</u>

Logistic Regression Model : 92.91%
Recursive Feature Selection (RFE) : 94.57%
Decision Tree Model : 94.48%

- Based on these comparisons, **the Recursive Feature Selection (RFE)** seems to perform the best overall with 94.57% accuracy

sales pitch

High Predictive Accuracy: Recursive Feature Selection (RFE) model achieves an impressive 94.57% accuracy in predicting labor force participation. This high accuracy means that we can consistently forecast whether individuals are likely to be part of the labor force or not.

Valuable Variables Identified: This analysis identified the top predictors of labor force participation, including working hours, poverty level, education, and class of worker. Understanding these variables allows you to focus on key areas that influence workforce trends, providing actionable insights for your business strategies.

My analysis isn't just about numbers—it's about giving you a strategic edge. By understanding the factors that influence labor force participation, you can make smarter decisions that drive business success and create a positive impact in your community.

Thank you!