Consulting Project

Title: The Warren Group Consulting (Foreclosure Property)

**Summary:** Used this project for mainly one reason that is to know the foreclosure rate and the types of property get foreclosed most so banks would be more cautious giving them a loan.

**Business understanding:**

Foreclosure is a critical legal action wherein a lender seeks to recoup funds from a borrower who has defaulted on their mortgage.

In Massachusetts, banks don't need a court order to accomplish this. After taking the property, banks frequently sell it for less than it's worth. It is critical for homeowners to understand what may be done to prevent or limit this process while protecting their interests.

**Potential Clients:**

Banks and Credit Unions: market risk and patterns assessment.

Real Estate Agencies: insights into market dynamics.

Real Estate Investors: strategic investment opportunities at a lower price.

Property Management Comp.: understanding market trends.

**Research Question**

Short Term: Which county exhibits the highest rate of foreclosures?

Mid Term: Which classification of property and price range is most frequently subject to foreclosure?

Long Term: What factors contribute to the escalation of foreclosure rates in 2023 compared to the baseline of 2021?

Data Dictionary:

1. County
2. PreForclosureType
3. Town
4. UnitType
5. Buyer1
6. City
7. Lender
8. Price (Target)
9. Style
10. Proptype (Target)
11. Neighborhood
12. SignDate
13. UnitNumber
14. Address
15. Buyer2
16. CountyCode
17. MapID
18. Street
19. Towncode

**Data Understanding:**

New merged was created by combining MAPC24 & Preforeclosure to get a new data set on which models can be created. (Merged on Mapid)

The number of observations derived after combining & cleaning the files was 30K.

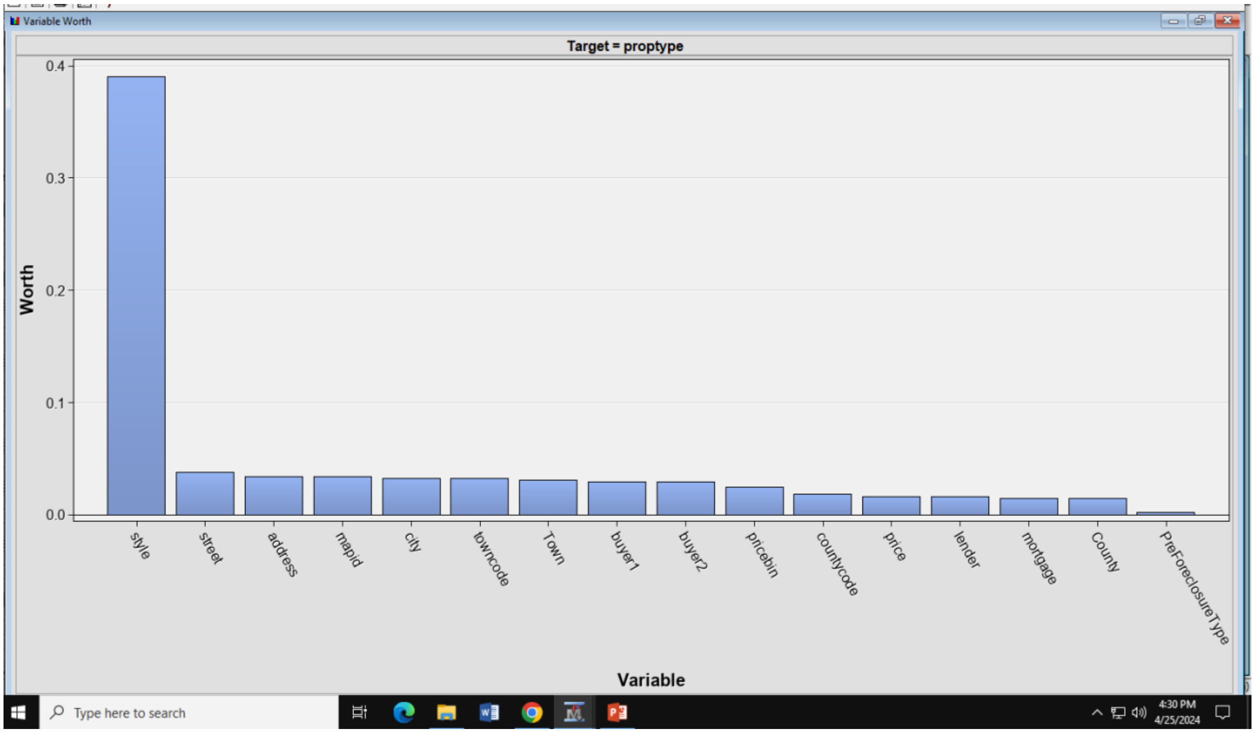
20 variables

Duplicate values were present which were corrected by keeping the original one & removing the others, Missing values were significantly present, but they were handled by drop node as values that were missing more than 50% of the entire data.

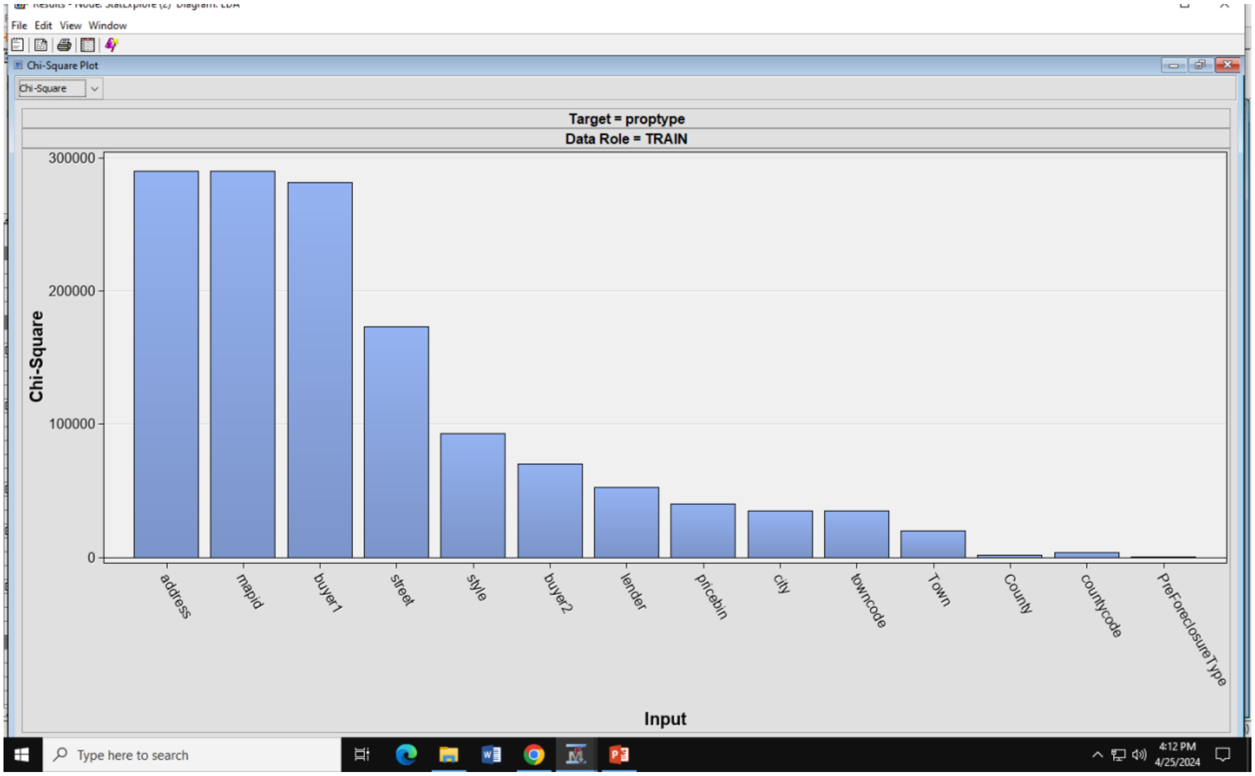
Data Partition was used in 70:30 ratio, Where 70% was for training and 30% for validation and 0% for testing.

Later many of variables were dropped as they were not as useful for our models. i.e: Buyer1, Lender, Neighborhood, Sign date, Unit Number, Address, Buyer2, County code, Mapid, and Street.

## Modeling and model evaluation:

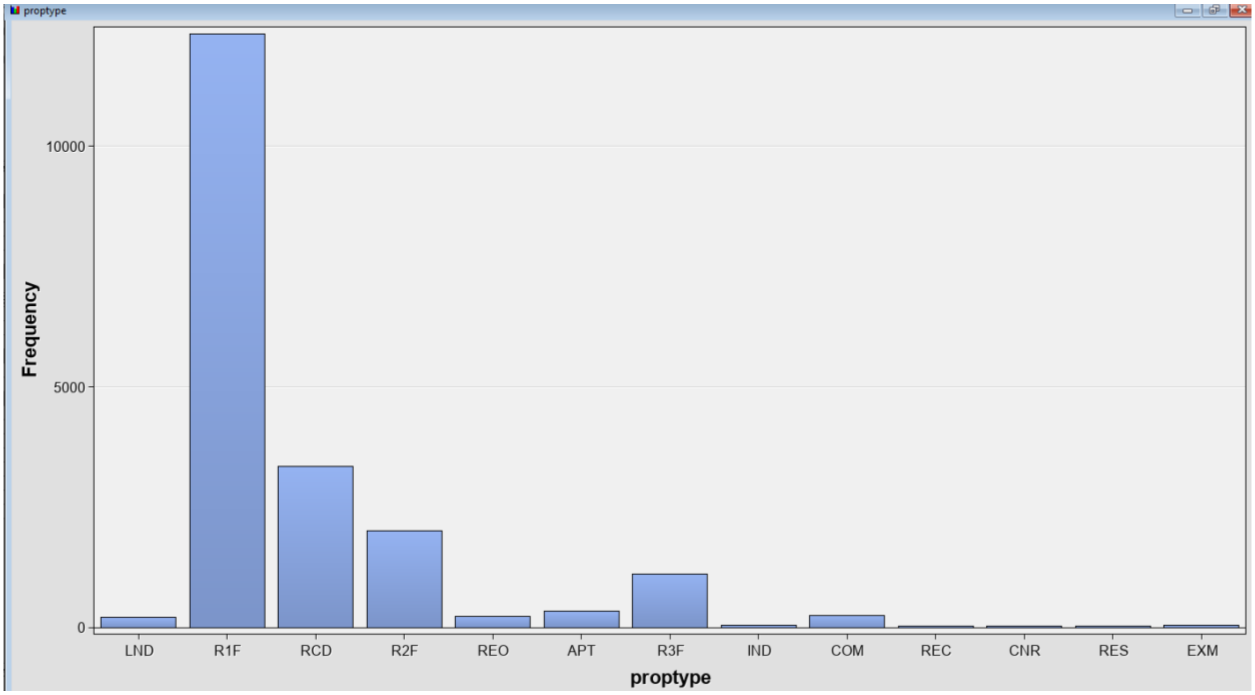


The chart displays the importance of various variables for predicting the property type for foreclosure. Style stands out as the most important variable, while others are less influential.

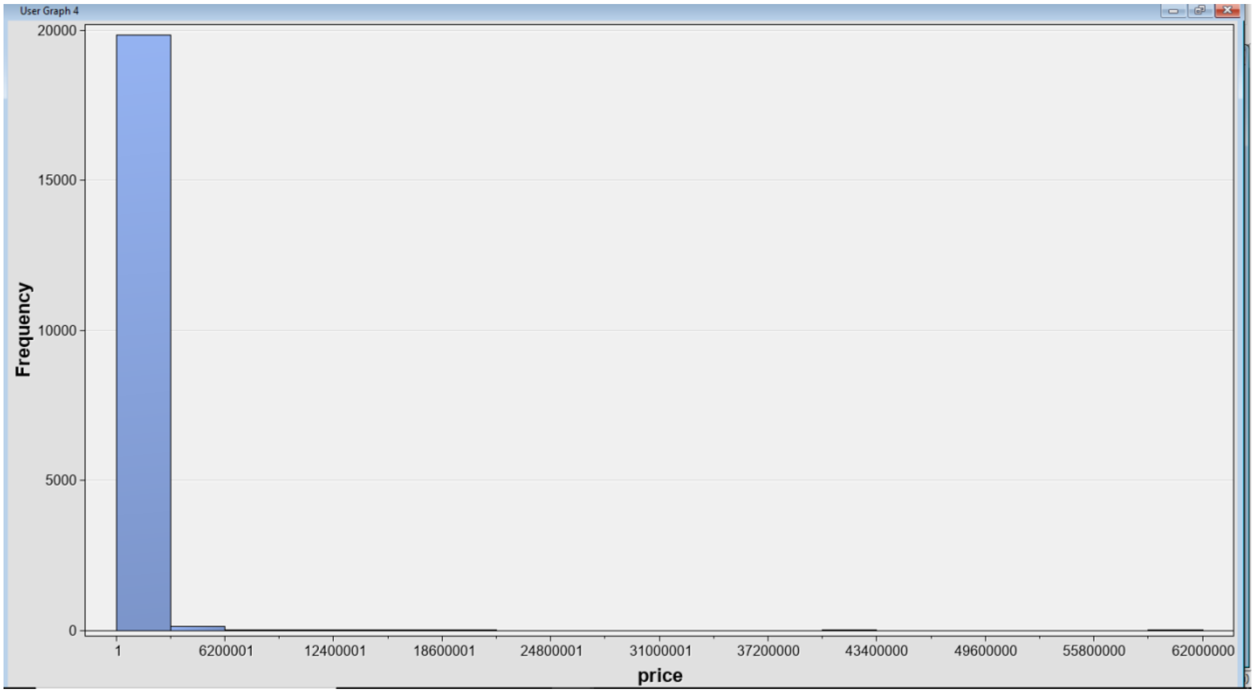


This Chi-square plot ranks variables by their statistical significance in predicting a target variable called **proptype**. The variables **address**, **Mapid**, and **Buyer1** show high Chi-square values, suggesting they are strong predictors.

(Skewness)

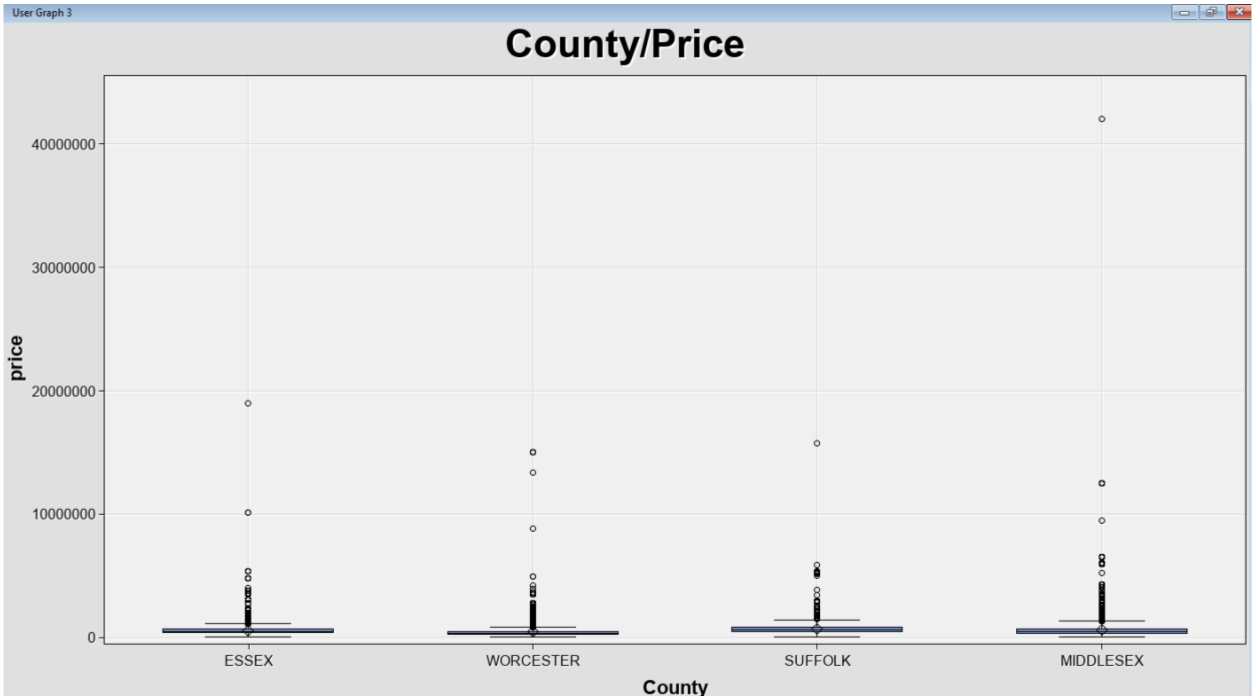


Skewness for target variable proptype is on the right side which shows that there are only some type of particular properties which are having forclosure more often.

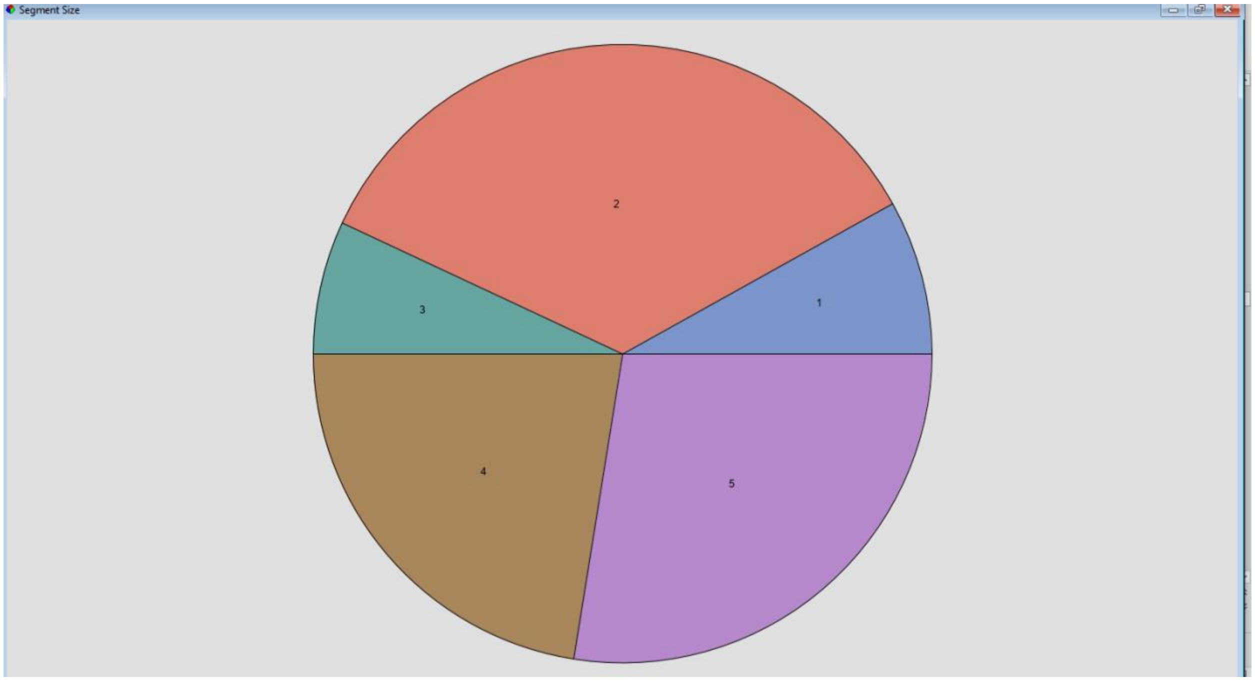


Skweness for target variale price is on the right side of the which shows some perticualr price of properties are having most of the forclosures.

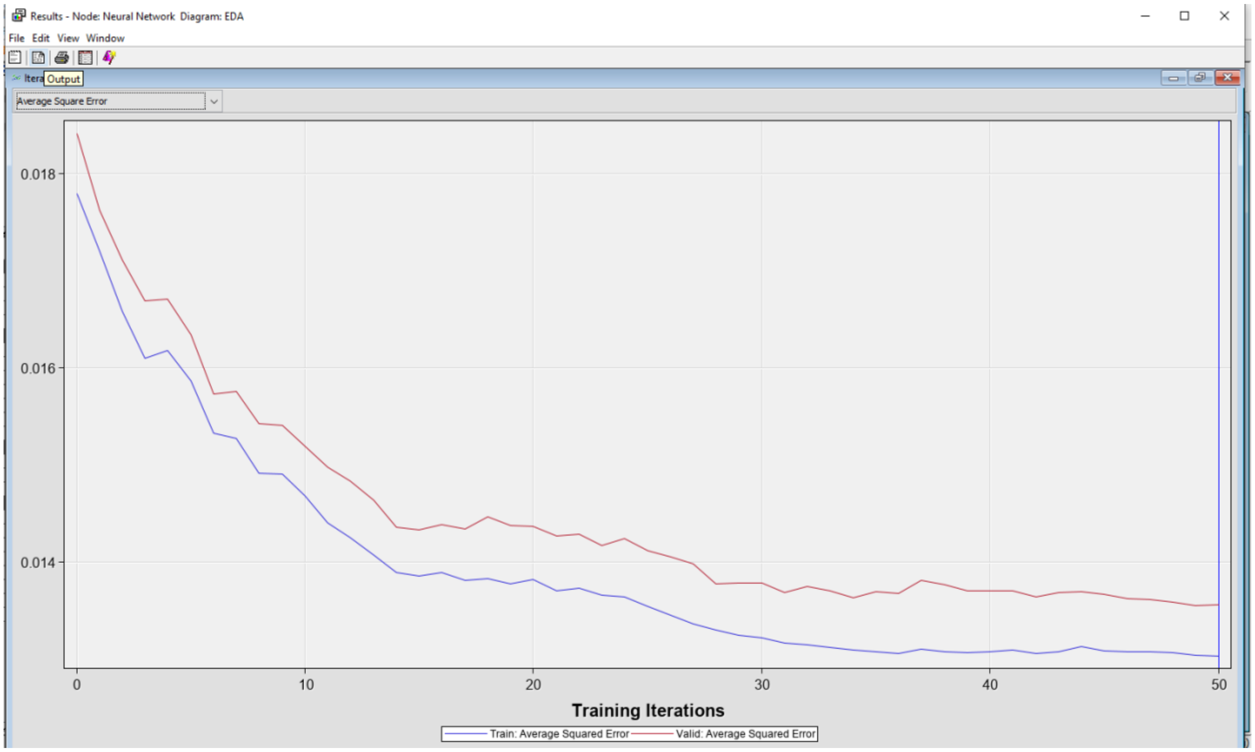
Outliers



With this boxplot we can see that outliers are present only after 1 million in prices in all the counties, which means that only very few houses in that price range are having foreclosure and below that there are most of the foreclosures.

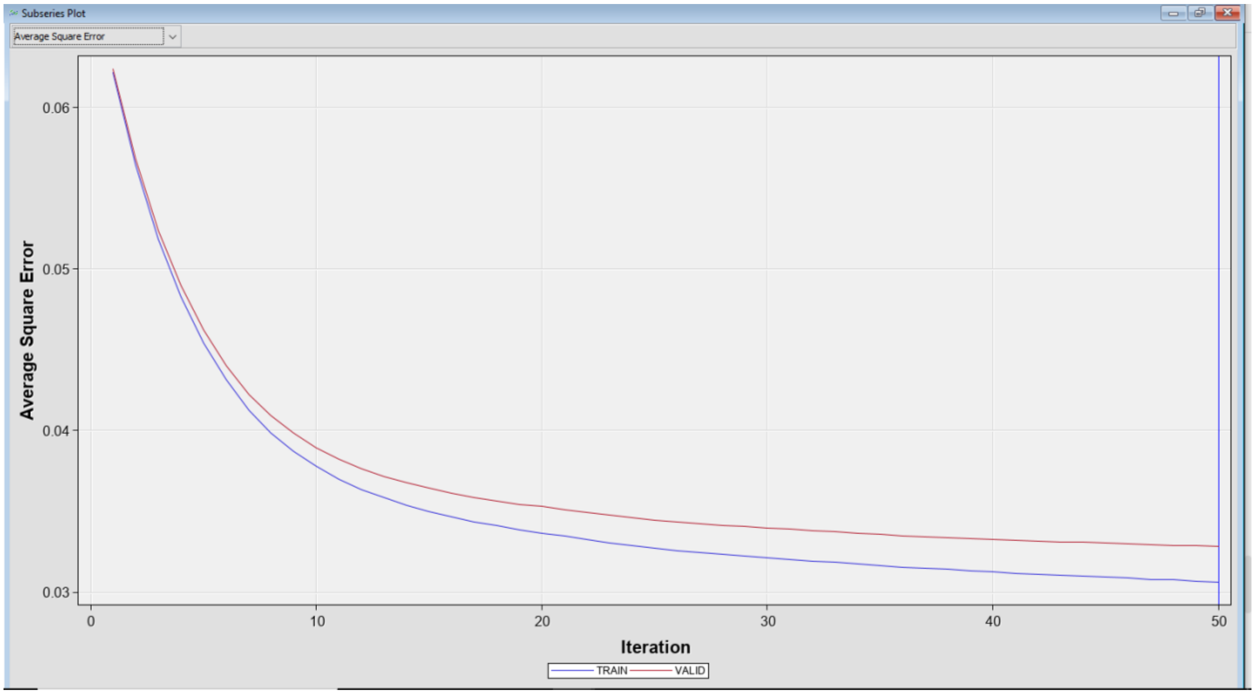


Cluster 2 is the largest group in the data, followed by Clusters 1, 3, 4, and 5. It provides a quick view of how the data is grouped into clusters, with the size of each cluster indicating its relative prevalence in the dataset.


Results - Node: Neural Network Diagram: EDA
File Edit View Window
Itera Output
Average Square Error
0.018
0.016
0.014
0
10
20
30
40
50
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error

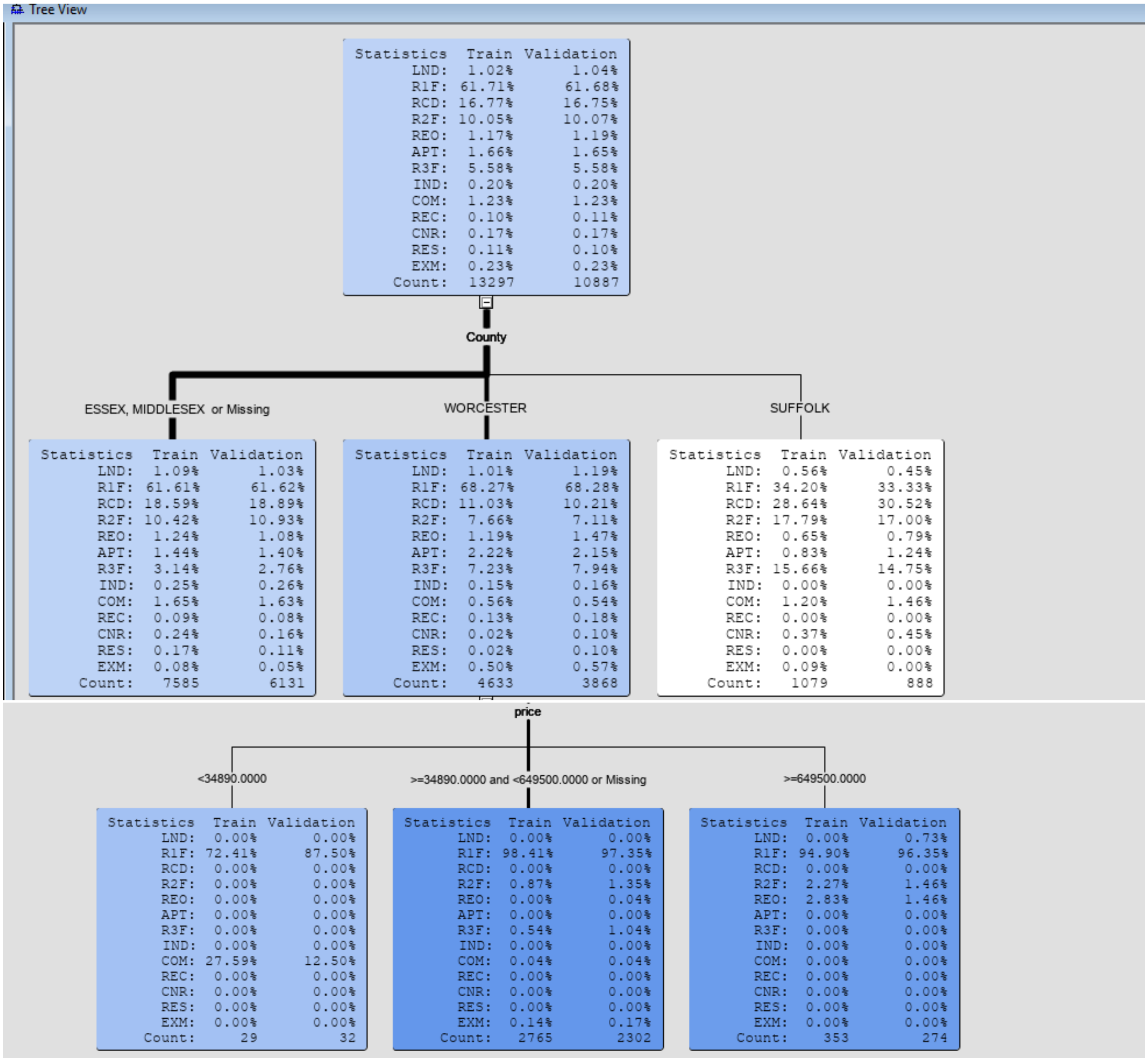
Neural Network

Iteration plot of training error decreasing over time, stabilizing after about 30 iterations, showing model learning and convergence.



Gradient Boosting

Gradient boosting model's average squared error declines sharply initially, then plateaus, indicating early learning with diminishing improvements, suggesting good model fit without overfitting, as validation error remains close to training error throughout iterations.



```
Statistics  Train  Validation
      LND:  1.02%       1.04%
      R1F: 61.71%      61.68%
      RCD: 16.77%      16.75%
      R2F: 10.05%      10.07%
      REO:  1.17%       1.19%
      APT:  1.66%       1.65%
      R3F:  5.58%       5.58%
      IND:  0.20%       0.20%
      COM:  1.23%       1.23%
      REC:  0.10%       0.11%
      CNR:  0.17%       0.17%
      RES:  0.11%       0.10%
      EXM:  0.23%       0.23%
    Count:  13297       10887
```

ESSEX, MIDDLESEX or Missing

```
Statistics  Train  Validation
      LND:  1.09%       1.03%
      R1F: 61.61%      61.62%
      RCD: 18.59%      18.89%
      R2F: 10.42%      10.93%
      REO:  1.24%       1.08%
      APT:  1.44%       1.40%
      R3F:  3.14%       2.76%
      IND:  0.25%       0.26%
      COM:  1.65%       1.63%
      REC:  0.09%       0.08%
      CNR:  0.24%       0.16%
      RES:  0.17%       0.11%
      EXM:  0.08%       0.05%
    Count:   7585        6131
```

WORCESTER

```
Statistics  Train  Validation
      LND:  1.01%       1.19%
      R1F: 68.27%      68.28%
      RCD: 11.03%      10.21%
      R2F:  7.66%       7.11%
      REO:  1.19%       1.47%
      APT:  2.22%       2.15%
      R3F:  7.23%       7.94%
      IND:  0.15%       0.16%
      COM:  0.56%       0.54%
      REC:  0.13%       0.18%
      CNR:  0.02%       0.10%
      RES:  0.02%       0.10%
      EXM:  0.50%       0.57%
    Count:   4633        3868
```

SUFFOLK

```
Statistics  Train  Validation
      LND:  0.56%       0.45%
      R1F: 34.20%      33.33%
      RCD: 28.64%      30.52%
      R2F: 17.79%      17.00%
      REO:  0.65%       0.79%
      APT:  0.83%       1.24%
      R3F: 15.66%      14.75%
      IND:  0.00%       0.00%
      COM:  1.20%       1.46%
      REC:  0.00%       0.00%
      CNR:  0.37%       0.45%
      RES:  0.00%       0.00%
      EXM:  0.09%       0.00%
    Count:   1079         888
```

<34890.0000

```
Statistics  Train  Validation
      LND:  0.00%       0.00%
      R1F: 72.41%      87.50%
      RCD:  0.00%       0.00%
      R2F:  0.00%       0.00%
      REO:  0.00%       0.00%
      APT:  0.00%       0.00%
      R3F:  0.00%       0.00%
      IND:  0.00%       0.00%
      COM: 27.59%      12.50%
      REC:  0.00%       0.00%
      CNR:  0.00%       0.00%
      RES:  0.00%       0.00%
      EXM:  0.00%       0.00%
    Count:     29          32
```

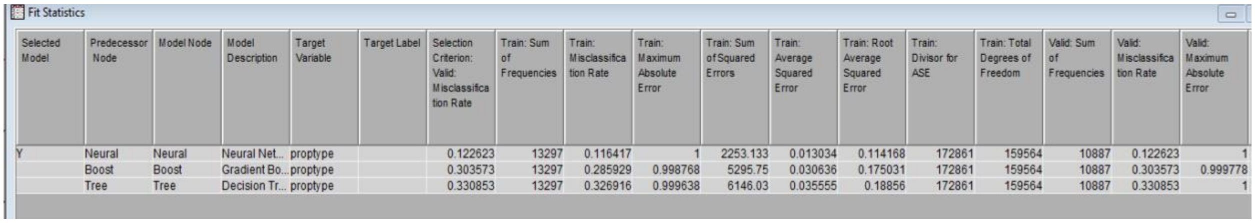
>=34890.0000 and <649500.0000 or Missing

```
Statistics  Train  Validation
      LND:  0.00%       0.00%
      R1F: 98.41%      97.35%
      RCD:  0.00%       0.00%
      R2F:  0.87%       1.35%
      REO:  0.00%       0.04%
      APT:  0.00%       0.00%
      R3F:  0.54%       1.04%
      IND:  0.00%       0.00%
      COM:  0.04%       0.04%
      REC:  0.00%       0.00%
      CNR:  0.00%       0.00%
      RES:  0.00%       0.00%
      EXM:  0.14%       0.17%
    Count:   2765        2302
```

>=649500.0000

```
Statistics  Train  Validation
      LND:  0.00%       0.73%
      R1F: 94.90%      96.35%
      RCD:  0.00%       0.00%
      R2F:  2.27%       1.46%
      REO:  2.83%       1.46%
      APT:  0.00%       0.00%
      R3F:  0.00%       0.00%
      IND:  0.00%       0.00%
      COM:  0.00%       0.00%
      REC:  0.00%       0.00%
      CNR:  0.00%       0.00%
      RES:  0.00%       0.00%
      EXM:  0.00%       0.00%
    Count:    353         274
```

The decision tree splits data by counties and property types, assessing model predictions with statistics like lift and response rates. Performance is consistent across training and validation sets, indicating a well-fitting model.



| Selected Model | Predecessor Node | Model Node | Model Description | Target Variable | Target Label | Selection Criterion: Valid: Misclassification Rate | Train: Sum of Frequencies | Train: Misclassification Rate | Train: Maximum Absolute Error | Train: Sum of Squared Errors | Train: Average Squared Error | Train: Root Average Squared Error | Train: Divisor for ASE | Train: Total Degrees of Freedom | Valid: Sum of Frequencies | Valid: Misclassification Rate | Valid: Maximum Absolute Error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | Neural | Neural | Neural Net... | proptype | | 0.122623 | 13297 | 0.116417 | 1 | 2253.133 | 0.013034 | 0.114168 | 172861 | 159564 | 10887 | 0.122623 | 1 |
| | Boost | Boost | Gradient Bo... | proptype | | 0.303573 | 13297 | 0.285929 | 0.998768 | 5295.75 | 0.030636 | 0.175031 | 172861 | 159564 | 10887 | 0.303573 | 0.999778 |
| | Tree | Tree | Decision Tr... | proptype | | 0.330853 | 13297 | 0.326916 | 0.999638 | 6146.03 | 0.035555 | 0.18856 | 172861 | 159564 | 10887 | 0.330853 | 1 |

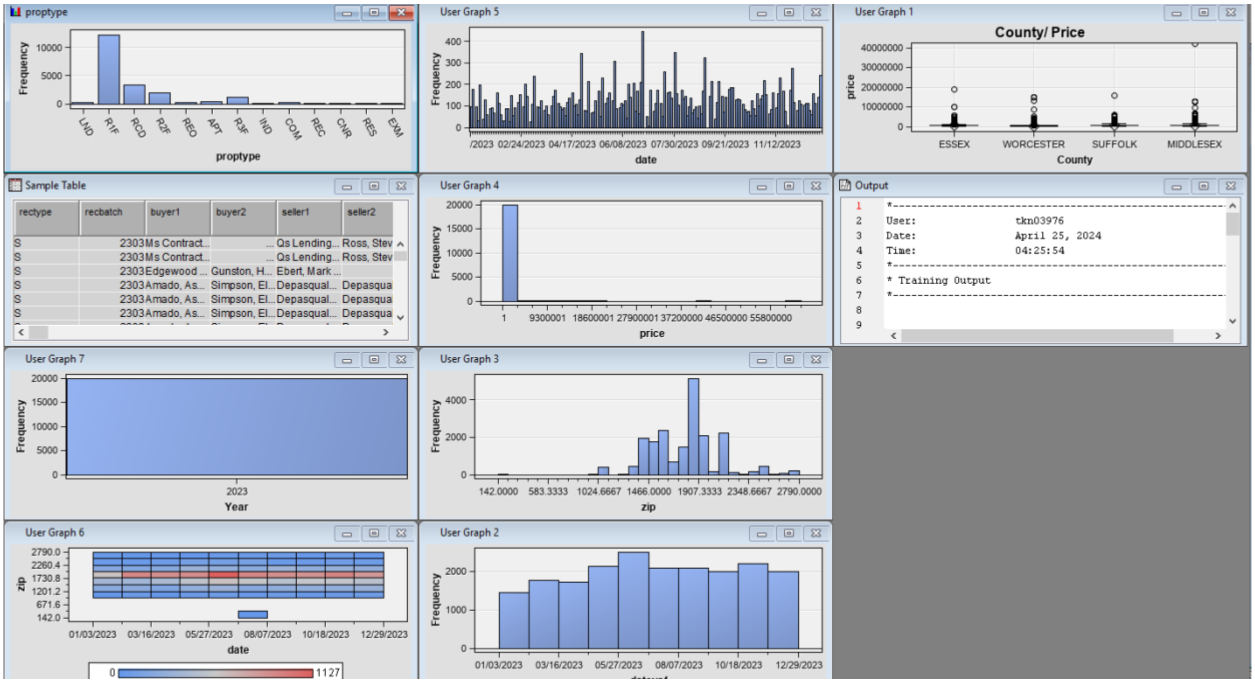
The model with the lowest validation misclassification rate, sum of squared errors, and maximum absolute error would be considered the best in terms of predictive performance. It appears that the

Neural Network model has the lowest validation misclassification rate, suggesting it might be the best performing model among the three.

Conclusion: The conclusion of our project is if someone is applying for a loan from any bank or financial institutions specially for Residential 1 family house and below 1 million there should be a throw background check as these kinds of properties are having most of the foreclosures.

Limitations: The limitations we faced were that only data of 2023 was given to us so for comparing our findings we had to use some data from outside sources.

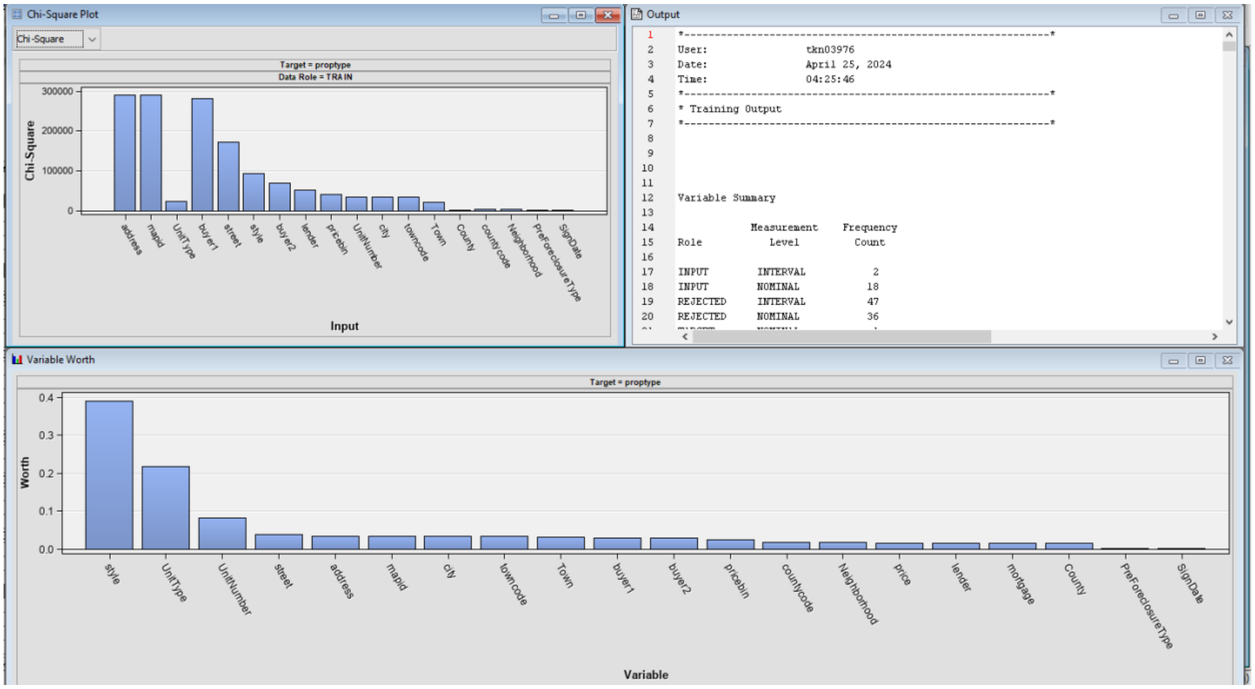
EDA and Appendix:



| rectype | recbatch | buyer1 | buyer2 | seller1 | seller2 |
|---|---|---|---|---|---|
| S | 2303 | Ms Contract... | ... | Qs Lending... | Ross, Stev |
| S | 2303 | Ms Contract... | ... | Qs Lending... | Ross, Stev |
| S | 2303 | Edgewood ... | Gunston, H... | Ebert, Mark ... | |
| S | 2303 | Amado, As... | Simpson, El... | Depasqual... | Depasqua |
| S | 2303 | Amado, As... | Simpson, El... | Depasqual... | Depasqua |
| S | 2303 | Amado, As... | Simpson, El... | Depasqual... | Depasqua |

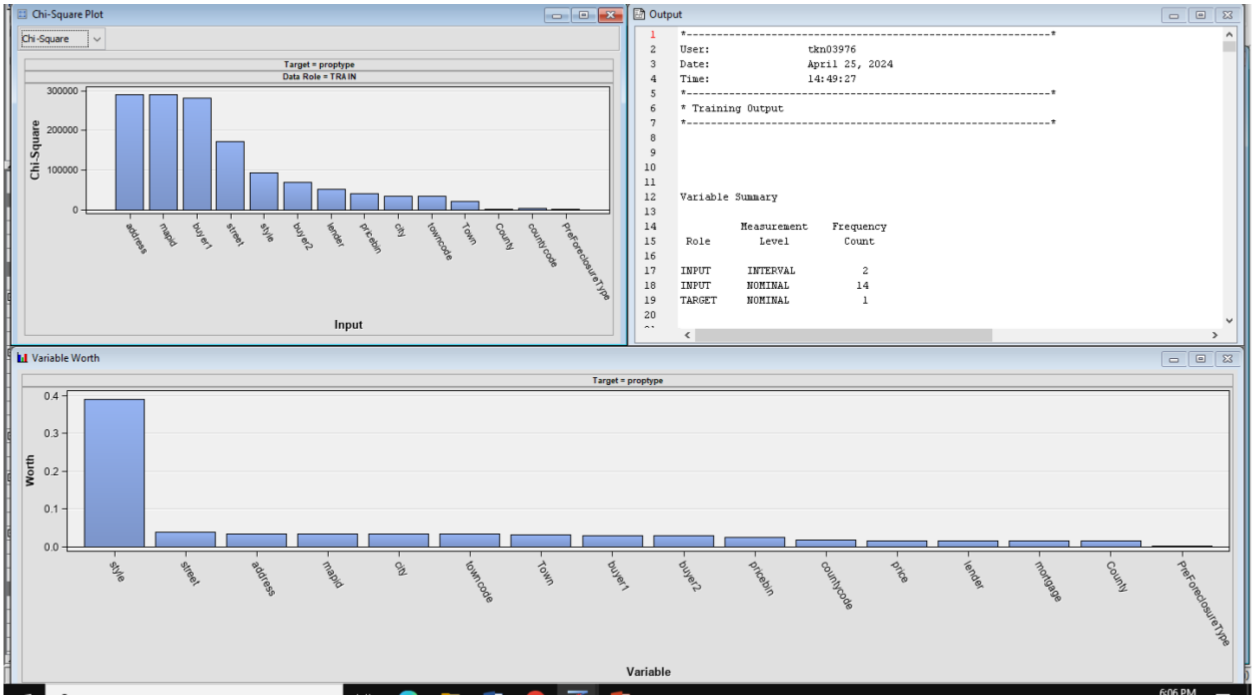
```
*------------------------------------------------------------
User:                    tkn03976
Date:                    April 25, 2024
Time:                    04:25:54
*------------------------------------------------------------
* Training Output
*------------------------------------------------------------
```

Here we can see that many of the variables were not working for us that's why we had to drop them.



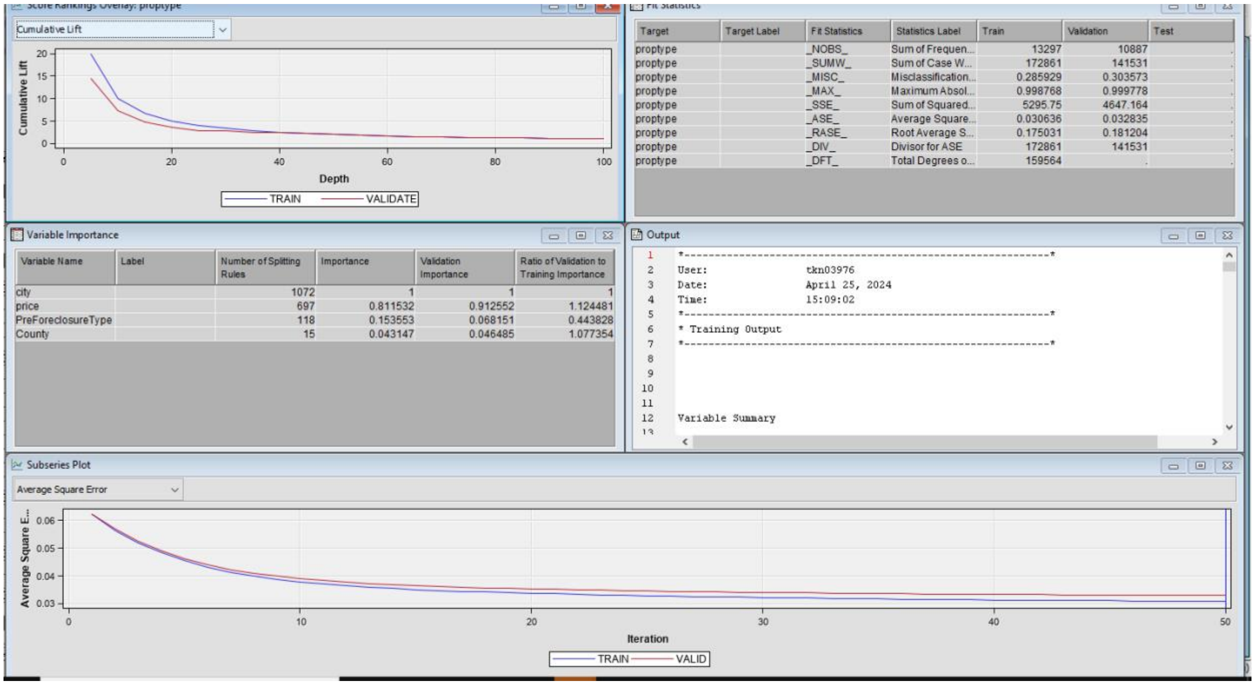
```
*------------------------------------------------------------*
User:                  tkn03976
Date:                  April 25, 2024
Time:                  04:25:46
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL          2
INPUT        NOMINAL          18
REJECTED     INTERVAL         47
REJECTED     NOMINAL          36
```

This is the chi-square and variable worth of target 1 (prop type) from this we can see which variables are worth more than others and which could give us an optimal results for different models.



```
*------------------------------------------------------------*
User:                  tkn03976
Date:                  April 25, 2024
Time:                  14:49:27
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          2
INPUT      NOMINAL          14
TARGET     NOMINAL           1
```

This is the chi-square and variable worth of target 2 (price) here we  can se that the variable Style is most prominent and others are somewhat equally worth. In comparison of both targets style of houses was most worthy variable for both of the targets which we set.



| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| proptype | | _NOBS_ | Sum of Frequen... | 13297 | 10887 | . |
| proptype | | _SUMW_ | Sum of Case W... | 172861 | 141531 | . |
| proptype | | _MISC_ | Misclassification... | 0.285929 | 0.303573 | . |
| proptype | | _MAX_ | Maximum Absol... | 0.998768 | 0.999778 | . |
| proptype | | _SSE_ | Sum of Squared... | 5295.75 | 4647.164 | . |
| proptype | | _ASE_ | Average Square... | 0.030636 | 0.032835 | . |
| proptype | | _RASE_ | Root Average S... | 0.175031 | 0.181204 | . |
| proptype | | _DIV_ | Divisor for ASE | 172861 | 141531 | . |
| proptype | | _DFT_ | Total Degrees o... | 159564 | . | . |

| Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance |
|---|---|---|---|---|---|
| city | | 1072 | 1 | 1 | 1 |
| price | | 697 | 0.811532 | 0.912552 | 1.124481 |
| PreForeclosureType | | 118 | 0.153553 | 0.068151 | 0.443828 |
| County | | 15 | 0.043147 | 0.046485 | 1.077354 |

```
*--------------------------------------------------------------*
User:                 tkn03976
Date:                 April 25, 2024
Time:                 15:09:02
*--------------------------------------------------------------*
* Training Output
*--------------------------------------------------------------*




Variable Summary
```

Here we can see the **Cumulative Lift Chart**:This chart compares cumulative lift across data depths for training and validation important for evaluating model effectiveness over various sample sizes.) datasets

**Variable Importance Table**:Displays variables sorted by importance with metrics like number of splitting rules, validation importance, and their ratio to training importance. Key variables include **city**, **price**, **PreForeclosureType**, and **County**.

**Subseries Plot**: Represents average square error across iterations for both training and validation, showing model performance and convergence over iterations.