TECHNIQUES IN HOUSE PRICE PREDICTION

Alessandro Pala - *alessandro.pala@studenti.unipd.it*

Domenico Tremaggi - *domenico.tremaggi@studenti.unipd.it*

OVERVIEW

- We tried predicting house prices with:
  - A linear regression
  - A regularized Poisson regression
  - A group lasso regression
  - A neural network

- Enforce sparsity through feature selection – variables or groups.

WHY THESE MODELS?

- **Linear Regression:** Baseline with a simple model.
- **Poisson Regression with Elastic Net:** Introduce sparsity and reduce multicollinearity.
- **Group Lasso:** Test grouped predictors (real estate features vs highly correlated variables).
- **Neural Network:** Establish an upper bound of predictive performance.

WHAT WE EXPECTED

- Feature regularization would outperform others due to the dataset's dimensionality and multicollinearity.

- Group Lasso to outperform other models given the nature of real-world price behavior.

DATASET

- Attributes include temporal, physical, scenic, and locational characteristics.
- Scale and preprocess features for machine learning compatibility.
- Removed duplicates, dropped irrelevant columns, and scaled the data.
- Created engineered variables for richer feature representation.

MULTIVARIATE LINEAR REGRESSION

- An extension of linear regression where the goal is to model the relationship between multiple independent variables and a dependent variable
- **Equation**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

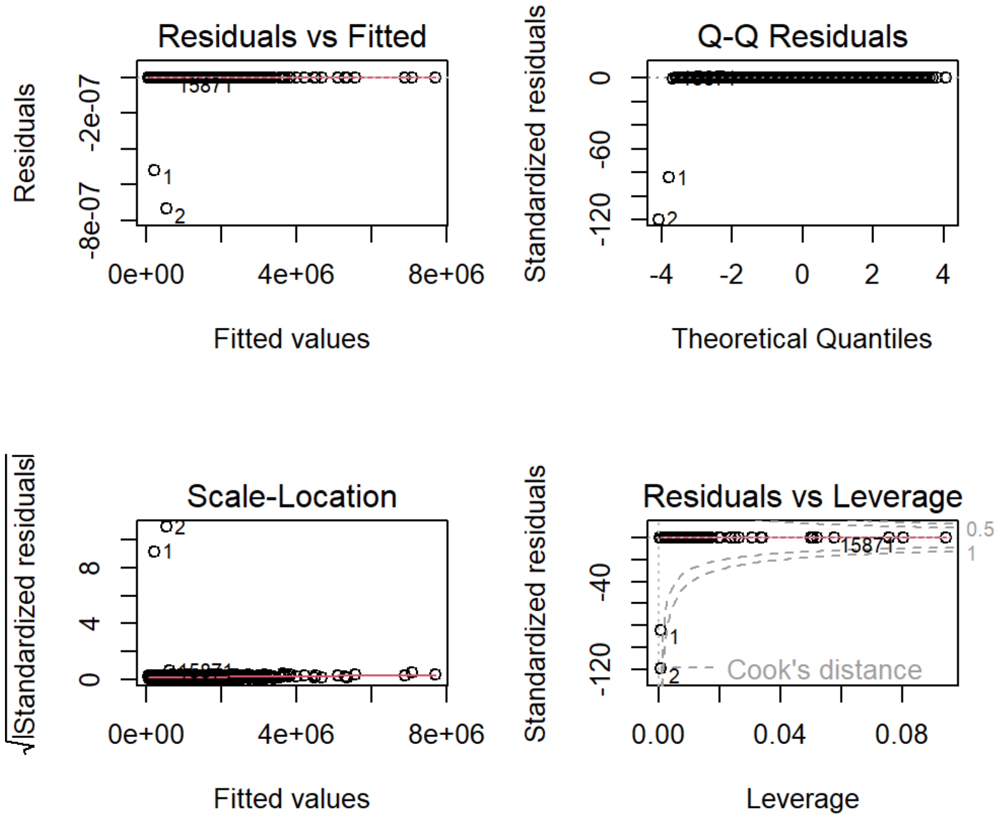
- where: $\boldsymbol{Y}$ is the dependent variable, $\boldsymbol{(X_1, X_2,\ldots, X_k)}$ are the independent variables, $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{(\beta_1,\ \beta_2,\ldots,\ \beta_k)}$ are the coefficients and ϵ is the error term.

MULTIVARIATE LINEAR REGRESSION

Results:

- Adjusted $R^2 = 1$, low residual standard error.
- Coefficients for bedrooms and intercept are statistically significant; others are not.

MULTIVARIATE LINEAR REGRESSION



Variable	Estimate	Std. Error	t value	p value
(Intercept)	5.405e+05	4.201e-11	1.287e+16	< 2e-16 ***
bedrooms	3.677e+05	7.398e-11	4.970e+15	< 2e-16 ***
bathrooms	7.485e-11	5.423e-11	1.380e+00	0.168
sqft_living	3.946e-11	7.260e-11	5.440e-01	0.587
sqft_lot	-1.557e-10	1.267e-10	-1.229e+00	0.219
floors	7.671e-12	6.084e-11	1.260e-01	0.900
waterfront	-1.241e-11	5.815e-11	-2.130e-01	0.831
view	-6.033e-12	4.719e-11	-1.280e-01	0.898
condition	1.576e-11	5.097e-11	3.090e-01	0.757
grade	3.389e-11	4.487e-11	7.550e-01	0.450
sqft_above	7.269e-11	7.782e-11	9.340e-01	0.350
sqft_basement	-4.531e-11	1.110e-10	-4.080e-01	0.683
zipcode	NA	NA	NA	NA
lat	-6.270e-11	5.391e-11	-1.163e+00	0.245
long	-2.108e-11	4.867e-11	-4.330e-01	0.665
sqft_living15	-1.568e-13	5.590e-11	-3.000e-03	0.998
sqft_lot15	5.429e-11	7.241e-11	7.500e-01	0.453
age	5.543e-12	6.138e-11	9.000e-02	0.928

MULTIVARIATE LINEAR REGRESSION

Interpretation

- Overfitting evident in "too perfect" results.
- Unsuitable.
- But 'suggests' to take interest in the *bedrooms* variable.

POISSON REGRESSION WITH ELASTIC NET

- **Poisson distribution:**

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- **Poisson regression:**

$$\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- where: $\boldsymbol{Y}$ is the dependent variable, $\boldsymbol{(X_1, X_2, \ldots, X_k)}$ are the independent variables, $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{(\beta_1, \beta_2, \ldots, \beta_k)}$ are the coefficient.

POISSON REGRESSION WITH ELASTIC NET

- **Elastic Net:** This kind of regularization combines L_1 and L_2 regularizations and applies penalty to coefficients

$$\text{Penalty} = \gamma \left[\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{k} \beta_j^2 \right]$$

- Where: $\boldsymbol{\gamma}$ controls the strength of the regularization, $\boldsymbol{\alpha}$ controls the balance between Lasso and Ridge
- **Poisson regression with Elastic Net:**

$$\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \\ + \gamma \left[\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{k} \beta_j^2 \right]$$

POISSON REGRESSION WITH ELASTIC NET

Results:

- Significant bedroom coefficient, effective sparsity enforcement.
- Regularization mitigated overfitting observed in MLR.

POISSON REGRESSION WITH ELASTIC NET

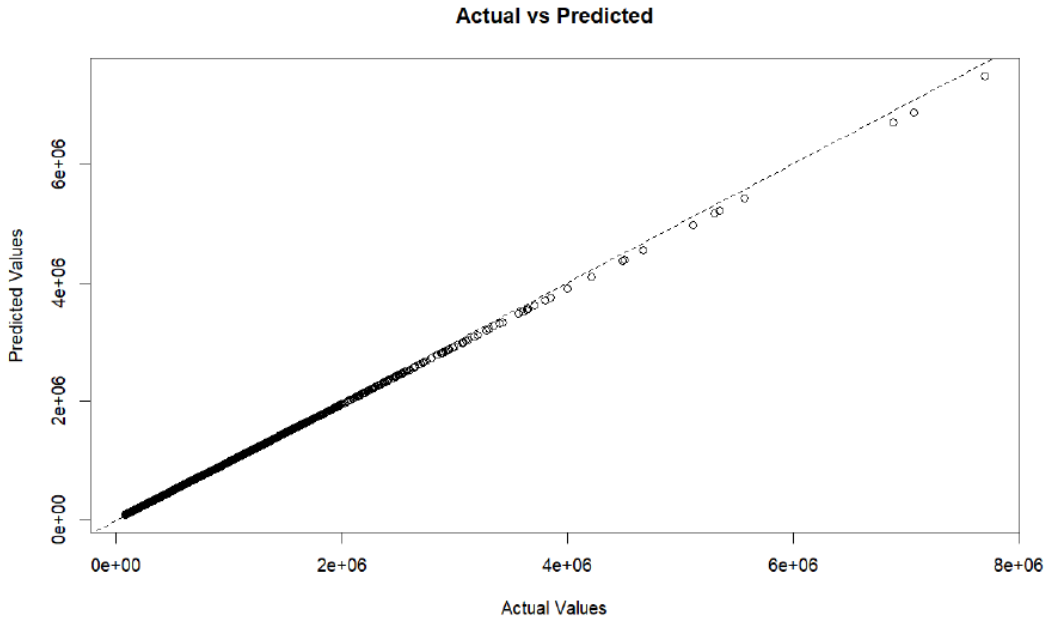
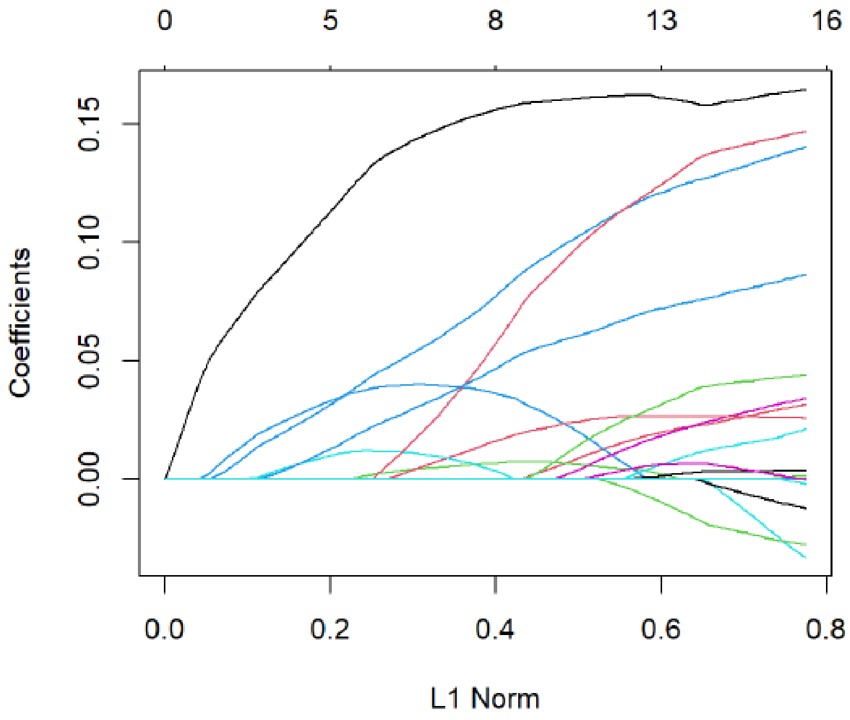


Figure 6: Poisson regression actual vs predicted



POISSON REGRESSION WITH ELASTIC NET

Interpretation

- *Bedrooms* emerged as the most significant predictor, while other predictors were penalized.
- Predicted values were consistent with actual target values, indicating good generalization.
- Regularization likely the cause of good results rather than GLM.

GROUP LASSO

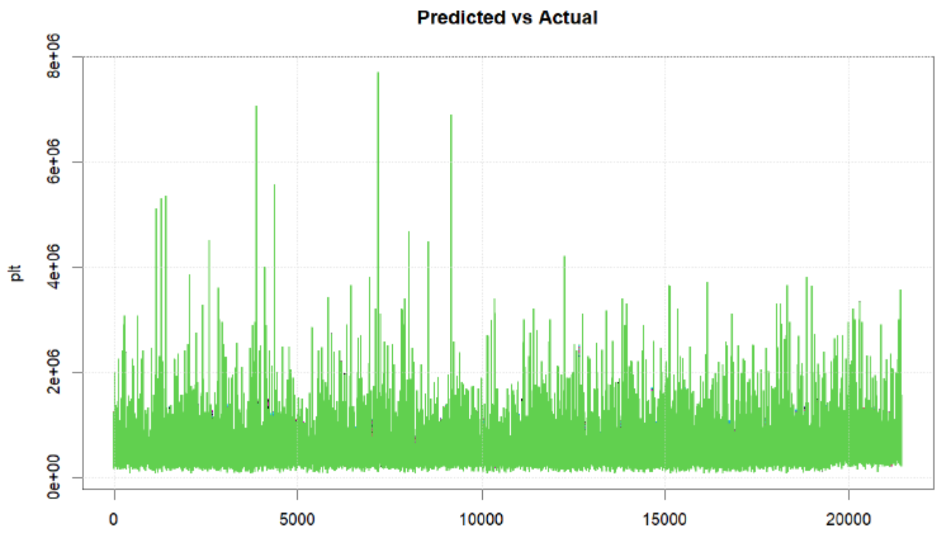
- An extension of the classical **Lasso** method applied for regularization and variable selection designed to work with grouped variables
- **Group coefficients penalty**: $\lambda \sum_{g=1}^{G} \|\beta_g\|_2$
- where: $\boldsymbol{\beta_g}$ represents the coefficient corresponding to the *g*-th group of variables of $\boldsymbol{G}$ groups, $||\boldsymbol{\beta_g}||_2$ is the Euclidean norm of the coefficient vector, $\boldsymbol{\lambda}$ is a regularization parameter that regulates its strength.
- Thus, the **optimization problem** is: $$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \right)$$

GROUP LASSO

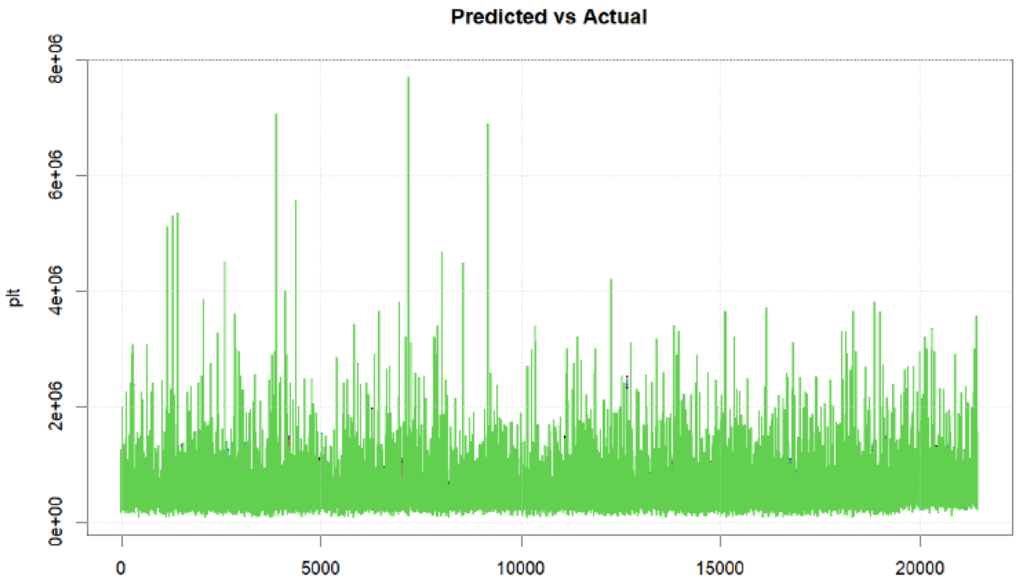
Results:

- Grouped variables did not outperform individual predictors.
- Model struggled to capture any pattern at all.
- Noise within grouped variables provided no predictive power.
- Cross-validation further showed weak accuracy, and that the model failed to generalize.

GROUP LASSO

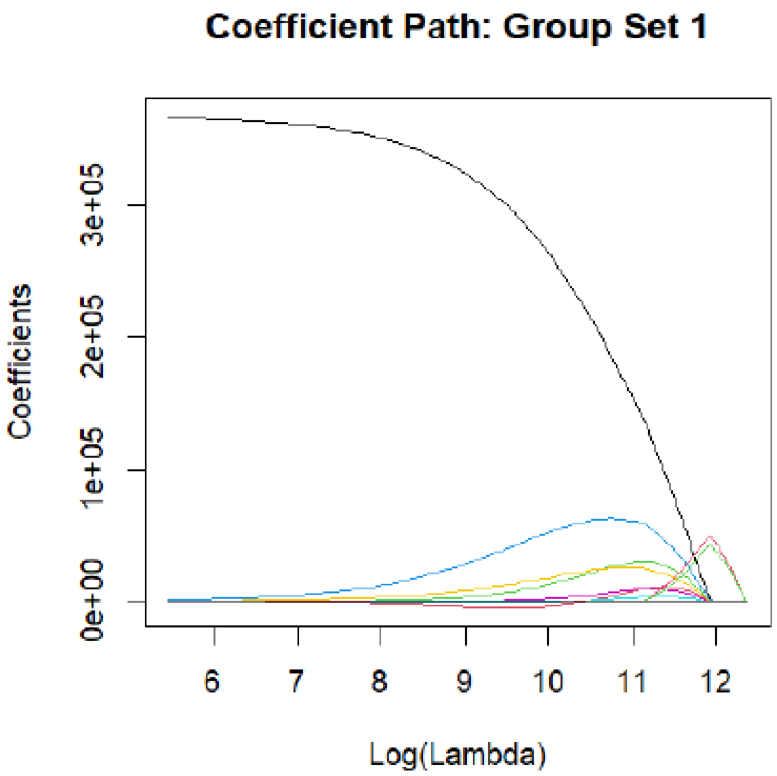


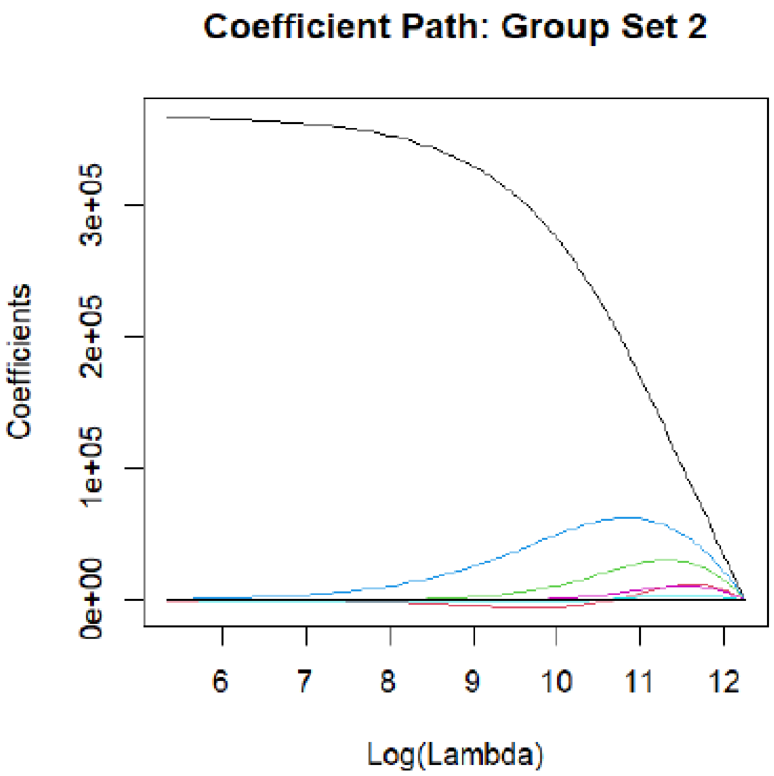
Predicted vs Actual
plt
8e+06
6e+06
4e+06
2e+06
0e+00
0
5000
10000
15000
20000



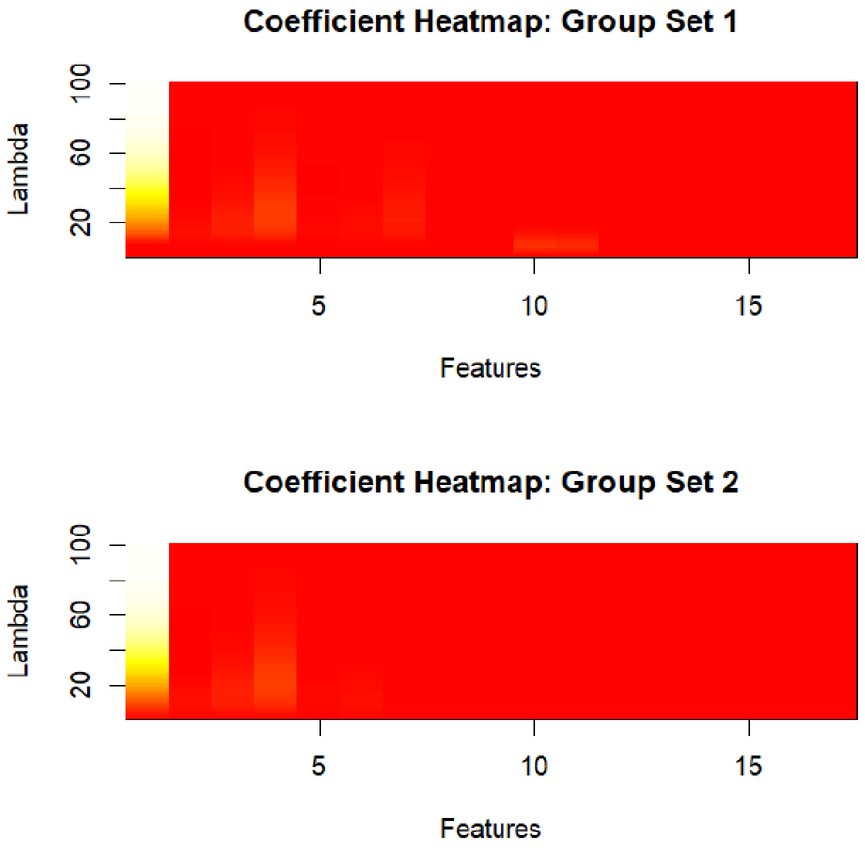
Predicted vs Actual
plt
8e+06
6e+06
4e+06
2e+06
0e+00
0
5000
10000
15000
20000

GROUP LASSO


Coefficient Path: Group Set 1
Coefficients
0e+00
1e+05
2e+05
3e+05
6
7
8
9
10
11
12
Log(Lambda)


Coefficient Path: Group Set 2
Coefficients
0e+00
1e+05
2e+05
3e+05
6
7
8
9
10
11
12
Log(Lambda)

GROUP LASSO



Coefficient Heatmap: Group Set 1
Lambda
100
60
20
5
10
15
Features
Coefficient Heatmap: Group Set 2
Lambda
100
60
20
5
10
15
Features

GROUP LASSO

Interpretation

- No meaningful relationships for the groups.
- Grouping does not align with the dataset’s structure.

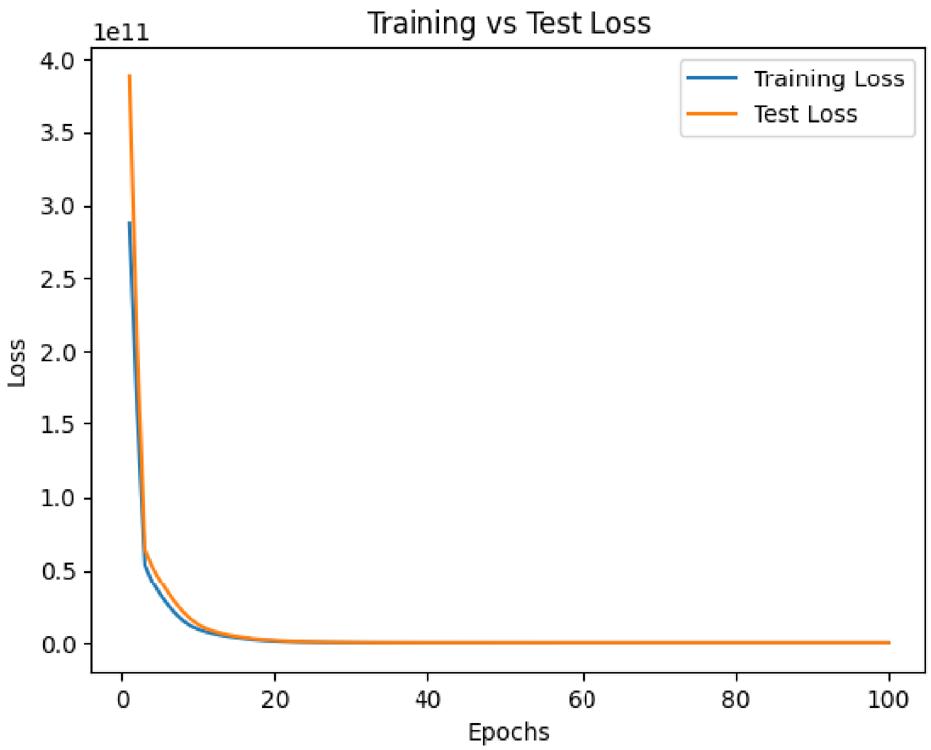
NEURAL NETWORK

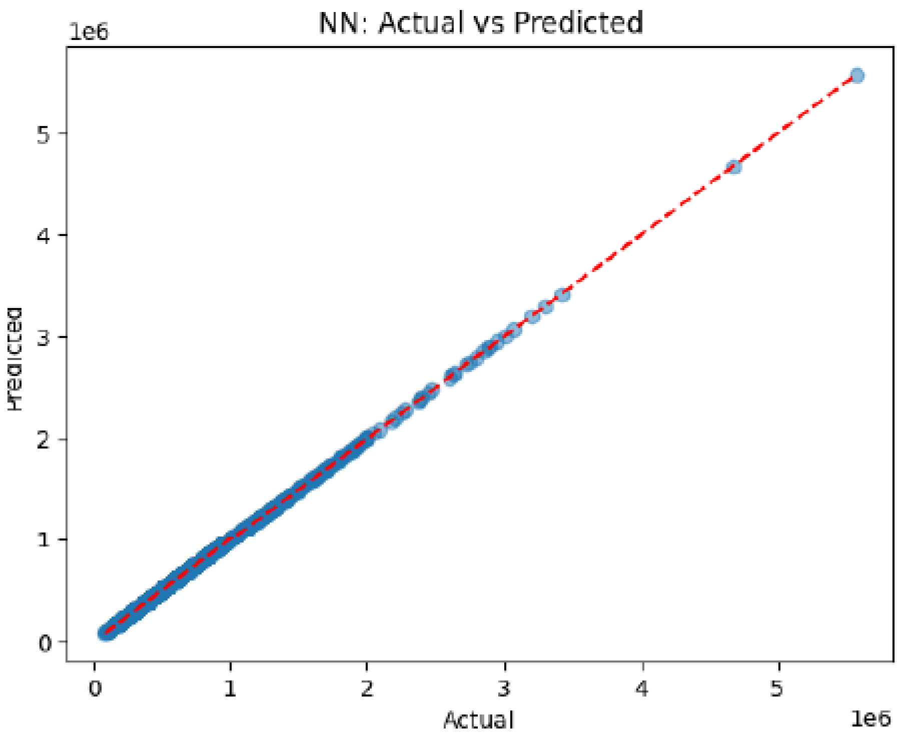
- Architecture: Linear MLP with moderate depth.

- Will the 'best' linear combination give perfect results?

```
self.hidden1 = nn.Linear(input_dim, 64)
self.hidden2 = nn.Linear(64, 128)
self.hidden3 = nn.Linear(128, 64)
self.output = nn.Linear(64, 1)
```

NEURAL NETWORK





NEURAL NETWORK

Interpretation

- The Linear Neural Network worked like a charm.
- The dataset is actually not 'corrupted'.
- We don't need nonlinear models to make a model work.

COMPARISON

- **MLR:** Overfit but tells us where to look.
- **Poisson Regression (Elastic Net):** nice sparsity and great accuracy.
- **Group Lasso:** Ineffective.
- **Neural Network:** Best performance is perfect performance even with linear models.

CONCLUSIONS

- Is grouping variables based on hypotheses (e.g., real estate market trends or correlations) intuitively beneficial for modeling?
- Testing this method on the dataset does not support this intuition.
- The grouped regression models failed, likely due to the specific structure of the chosen data.

CONCLUSIONS

- Regardless of testing methods, the data consistently identifies only one feature (*bedrooms*) as a predictor.
- This outcome is inaccurate compared to real-world real estate price prediction models.
- The result is unlikely to generalize to other datasets.