High Dimensional Techniques in House Price Prediction

Pala, Alessandro
alessandro.pala@studenti.unipd.it

Tremaggi, Domenico
domenico.tremaggi@studenti.unipd.it

Abstract

We compare the use of linear regression, regularized Poisson regression, group lasso regression, and a simple neural network to predict house prices based on various features, such as location, size, and the number of rooms. The primary goal is to assess the performance of these models in terms of predictive accuracy and if so to enforce sparsity by performing feature selection. The results on the chosen dataset suggest that grouping variables is not the right method since single features predict better and with less noise.

Introduction

In order to approach the dataset, we first think about which models and variables can predict real estate prices in the real world, with the constraint of the set of variables given by the dataset. Since we handle a price prediction regression, a fair comparison of linear and non-linear basic models - be them regularized or not - would be multivariate. Hence, we first carry a synthetic application of the following models: multivariate linear model and multivariate Poisson regression with an $\alpha = 0.1$ Elastic Net. We then train a deep Multi-Layer Perceptron for performance comparison and we regularize it with a simple Lasso. Finally, we utilise a Group Lasso regression on two sets of groups of variables that we think may be good predictors of house prices in the real world. The first set of groups uses variables that are usually clustered together by real estate professionals when they do house price estimation. The second set of groups uses clusters of highly statistically correlated (Pearson's r) variables (Figure 1).

1 Dataset

The dataset contains real estate property listings, each described by a variety of attributes. Temporal attributes are: the listing date, the year the property was built, and the year of its last renovation. Physical characteristics are: the number of bedrooms and bathrooms, living area size, total lot size, number of floors, basement area, and living area above ground. Scenic qualities are represented by a waterfront view indicator, a view quality rating, an overall condition rating, and an overall grade rating. Locational data instead include zip code, latitude, and longitude. There is also some additional spatial context given by the average living area and lot sizes of the 15 nearest properties.

1.1 Pre-processing

In order to pre-process the data, we first convert the date column into a proper date object by extracting the first 8 characters (in "YYYYMMDD" format). We then replace 0 values in the "yr_renovated" column with NA, treating them as missing data. Duplicate rows

are removed based on the id column. We add new variables in the dataset: total_sqft calculates total area by summing living and basement space, bath_per_bed computes the bathroom-to-bedroom ratio (defaulting to 0 if bedrooms are 0), and total_rooms sums bedrooms and bathrooms. sqft_diff_15 measures the difference between the property's living area and that of its 15 nearest neighbors, while age_since_reno calculates the property's age since construction, defaulting to 0 if the year is invalid (Figure 2). The first, second, fifteenth and sixteenth columns are dropped since they are analytically irrelevant (date, id, with yr_built and yr_renovated being substituted by "age"). We then scale the features of the data frame. The price and the non-structural features are extracted as columns 2 to 18 and scaled in the new matrix X, which will be used for the Group Lasso. The price, or target, is stored in a separate vector Y.

2 Base Models

2.1 Multivariate linear model

Multivariate Linear Regression is an extension of linear regression where the goal is to model the relationship between multiple independent variables (also called predictors or features) and a dependent variable (also called the target or response variable). It is used when you have more than one predictor variable that you believe influences the dependent variable.

Mathematical equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \quad (1)$$

where:
Y is the dependent variable, $(X_1, X_2, \ldots, X_k)$ are the independent variables, β_0 is the intercept, $(\beta_1, \beta_2, \ldots, \beta_k)$ are the coefficients and ϵ is the error term.

Experiments

In our case of interest we use all the the available variables to estimate the house prices (Table 1, Figure 3).

Among our predictors we find that only the intercept and the coefficient for bedrooms are statistically significant at the 0.001 level, with p-values less than 2e-16. Other predictors like bathrooms, sqft_living, and sqft_lot have high p-values, hence not statistically significant. The zipcode has been omitted because of singularity, meaning either multicollinearity or redundancy. The residual standard error of the model at 6.15e-09 is very low and in general the residuals are also very small, ranging from -7.36e-07 to 1.38e-09. The F-statistic is astronomically large at 4.788e+30. This multivariate regression is "perfect", given that the adjusted R-squared values are 1, with predictors explaining all the variability in the response variable. This is clearly a case of overfitting, which we need to handle with regularization.

2.2 Multivariate GLM with Elastic Net

The Poisson Distribution

Poisson regression is based on the Poisson distribution, which describes the probability of a given number of events occurring in a fixed interval of time or space, given the average rate of occurrence.
The probability mass function of a Poisson-distributed random variable Y is:

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad (2)$$

Where: y is the count of events (e.g., number of accidents, number of emails received) and λ is the rate of occurrence, which is the expected number of events in the fixed interval.

The Poisson Regression Model

Poisson regression assumes that the logarithm of the expected count (rate) is a linear function of the independent variables.
The model is expressed as:

$$\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (3)$$

where: $(X_1, X_2, \ldots, X_k)$ are the independent variables, β_0 is the intercept, $(\beta_1, \beta_2, \ldots, \beta_k)$ are the regression coefficients.

Elastic Net

In regular linear regression or GLMs, the model can overfit the data if there are many predictors, especially when some predictors are irrelevant or highly correlated, but it can also use some sparsity enforcement, so we applied an Elastic Net regression. This kind of regularization combines L_1 and L_2 regularizations and applies penalty to coefficients as to:

- Lasso ($L_1$1): Shrink some coefficients to zero, performing feature selection (Figure 4).
- Ridge (L_2): Shrink the coefficients (but not set them to zero) helping with multicollinearity, which our dataset suffers from.

Mathematically, the Elastic Net penalty is given by:

$$\text{Penalty} = \gamma \left[\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{k} \beta_j^2 \right] \quad (4)$$

where γ controls the strength of the regularization (larger γ means more regularization), α controls the balance between Lasso and Ridge. If $\alpha = 1$, the penalty is purely Lasso; if $\alpha = 0$, the penalty is purely Ridge, meaning our $\alpha = 0.1$ choice mostly applies an L_2 penalty. We chose this hyperparameter after some experiments that suggested that even a very small Lasso penalty enforced sparsity efficiently.

Poisson Regression with Elastic Net

When you combine Poisson regression with Elastic Net regularization the model becomes:

$$\begin{aligned} \log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \\ + \gamma \left[\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{k} \beta_j^2 \right] \end{aligned} \quad (5)$$

This means that the objective is to minimize the following loss function:

$$\begin{aligned} \text{Loss} = \sum_{i=1}^{n} (-Y_i \log(\lambda_i) + \lambda_i) + \\ + \gamma \left[\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{k} \beta_j^2 \right] \end{aligned}$$

Experiments

The Poisson regression model (Table 2) output shows that the intercept is 540,529.7, which means that when all the predictors are zero, the dependent variable has a high baseline value. The coefficient for Bedrooms is 356,970.6, which means that with every additional bedroom, the expected value of the outcome increases significantly, holding other factors constant. However, many other factors like Bathrooms, Sqft_Living, Sqft_Lot, Floors, etc., have missing coefficients, represented by periods, and most likely because these variables were excluded from the model, had no significant relationship with the result(Figure 5). A simple analysis of actual target values compared to predicted values shows that the model is very successful at predicting while not overfitting the data since the presence of Elastic Net regularization (Figure 7).

3 Group Lasso

The Group Lasso is an extension of the classical Lasso method applied for regularization and variable selection designed to work with grouped variables, and such a situation is helpful when given variables naturally fall into groups, and we want to choose or discard entire groups of variables simultaneously rather than choosing or dropping individual variables.

The main mathematical goal of the Group Lasso is to do regression by penalizing the sum of the norms of the coefficients in each group. That will induce sparsity at the level of groups, that is, all coefficients in a group will either be shrunk towards zero, or not. Therefore, entire groups of variables are either selected or not.

3.1 Mathematical Formulation

For a given linear regression problem, the model can be represented as:

$$y = X\beta + \epsilon \quad (6)$$

where: y is the $n \times 1$ vector of observed responses, X is the $n \times p$ matrix of predictors, β is the $p \times 1$

vector of regression coefficient and ϵ is the error term, assumed to be normally distributed.

In ordinary linear regression, we try to minimize the residual squared error:

$$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 \right) \tag{7}$$

In the case of Group Lasso, we add a regularization term to the objective function that penalizes the coefficients in groups. If the variables are divided into G groups, and group g contains p_g variables, the Group Lasso penalty is given by:

$$\lambda \sum_{g=1}^{G} \|\beta_g\|_2 \tag{8}$$

where: β_g represents the coefficient corresponding to the g-th group of variables, $\|\beta_g\|_2$ is the euclidean norm of the coefficient vector, λ is a regularization parameter that regulates its strength.

Thus, the optimization problem is:

$$\min_{\beta} \left(\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \right) \tag{9}$$

3.2 Experiments

We observe that even with this more complex model, the most important variable predicting the prices is the same as before, surely some minor correlation rise up, (Figure 9, 10), but not so strong to be considered meaningful for our purposes. The coefficient path confirms this result (Figure 8) along with the Cross-Validation tests (Figure 11, 12). Moreover, the model simply fails to predict house prices, (Figure 15, 16) a result that we could see coming from the statistical conclusions of the previous models, though not intuitive at first.

4 Neural Network

We finally train a deep neural network in order to have an estimate of the best possible result in any statistical context and, most of all, to understand if our choice of models was insufficient (Figure 13, 14).
We kept the architecture neither too deep nor too shallow given the medium number of features. We see very clearly that the network uses the features of the dataset to predict extremely well (Figure 17). The goal of this small experiment was also to satiate any doubt about the dataset being "corrupted", and that any model could somehow suffer from problems. Indeed, the dataset is not corrupted, as also suggested by the regularized GLM beforehand, but simply seems to (counterintuitively, with respect to the context) present only one feature out of many with actual predictive power (Table 3).

5 Conclusion

Even though a real-world model could intuitively work better by grouping variables according to different hypotheses (like the real estate market, or simple correlations), when such method is tested on this dataset, the intuition does not seem to hold. We suspect that such result would have no further confirmation in other datasets, and that our grouped regression models necessarily failed because of the structure of the chosen data, that, however tested, seems to only lead to one feature (*bedrooms*) to be a predictor, which is not accurate to real-world real estate price prediction models.

A Visualization

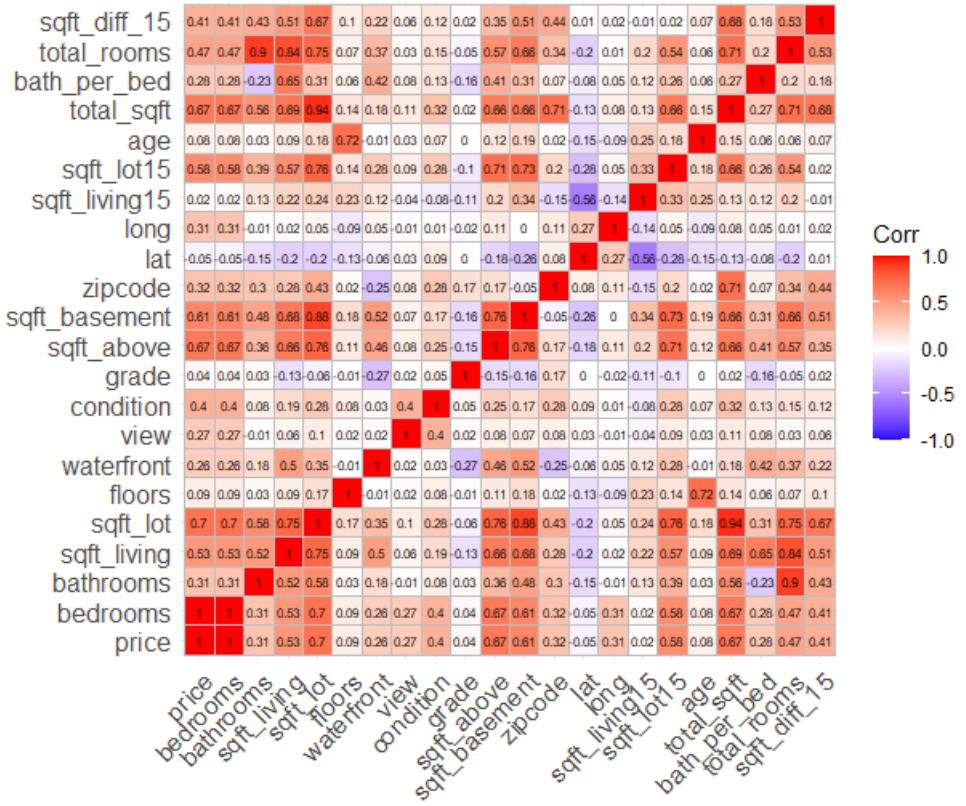


Figure 1: Correlation matrix to pick groups in the second set of Group Lasso

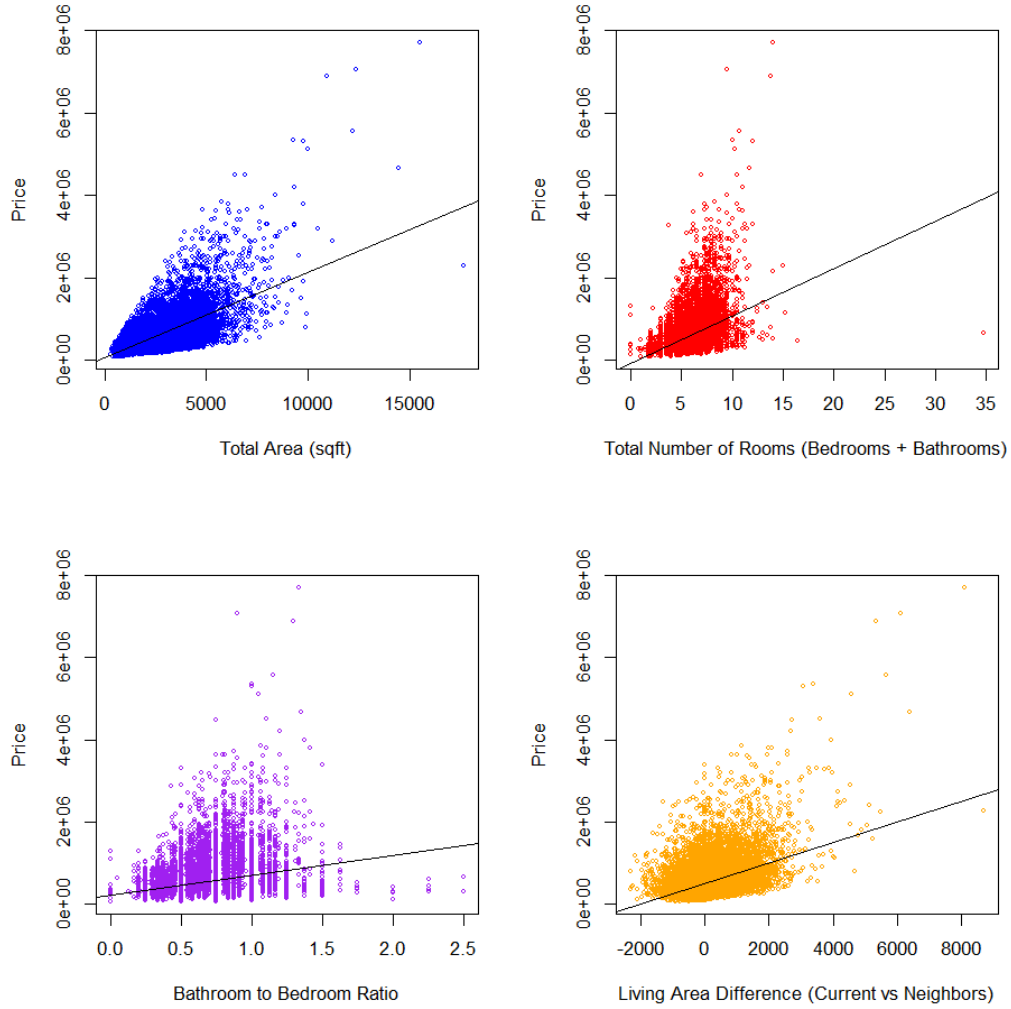


Figure 2: How price behaves with aggregated variables

B Regressions

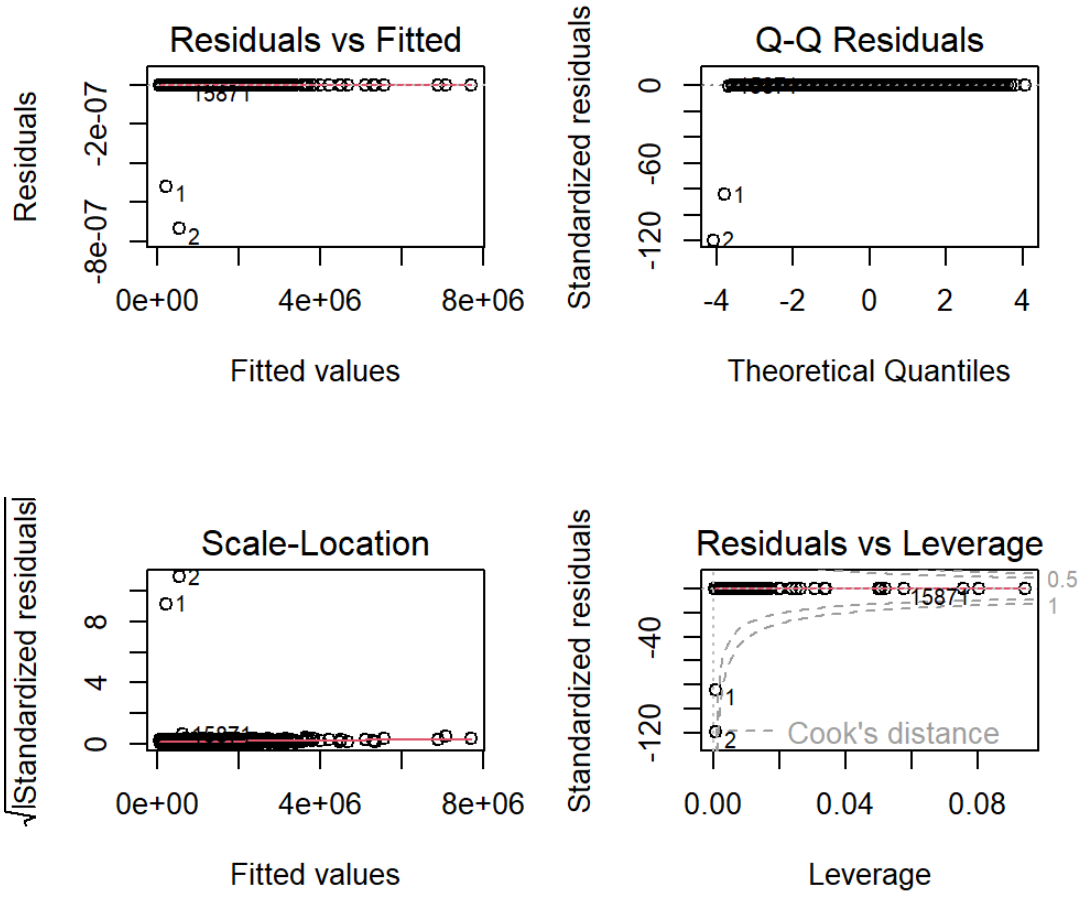


Figure 3: Multivariate linear regression plot

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

Table 1: Multivariate Linear Regression Data Summary

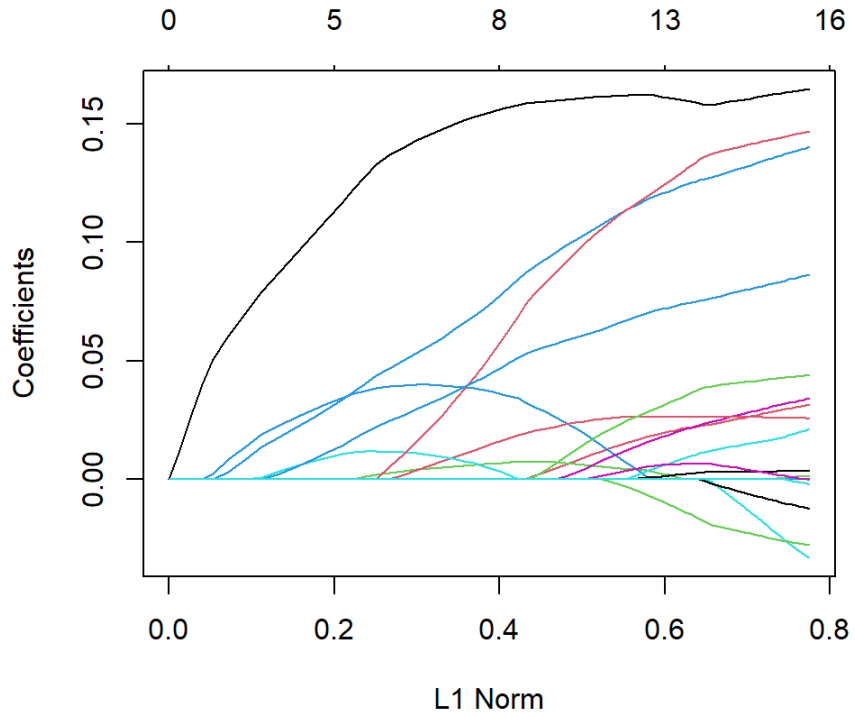


Figure 4: Multivariate Poisson regression coefficient path

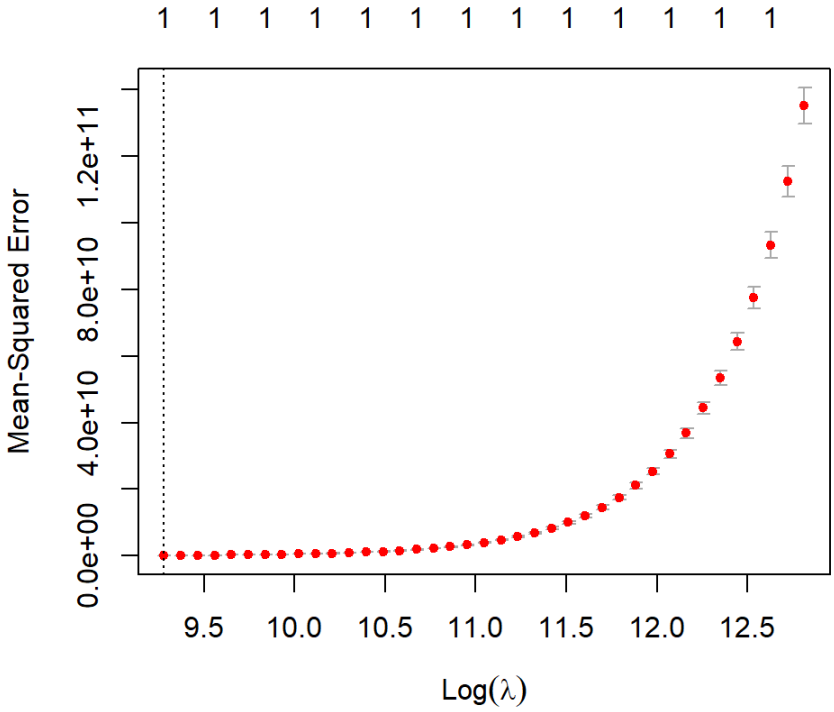


Figure 5: Cross validation plot for Poisson regression

Variable	**Coefficient**
Intercept	540529.7
Bedrooms	356970.6
Bathrooms	.
Sqft_Living	.
Sqft_Lot	.
Floors	.
Waterfront	.
View	.
Condition	.
Grade	.
Sqft_Above	.
Sqft_Basement	.
Zipcode	.
Lat	.
Long	.
Sqft_Living15	.
Sqft_Lot15	.
Age	.

Table 2: Coefficients from the Poisson model

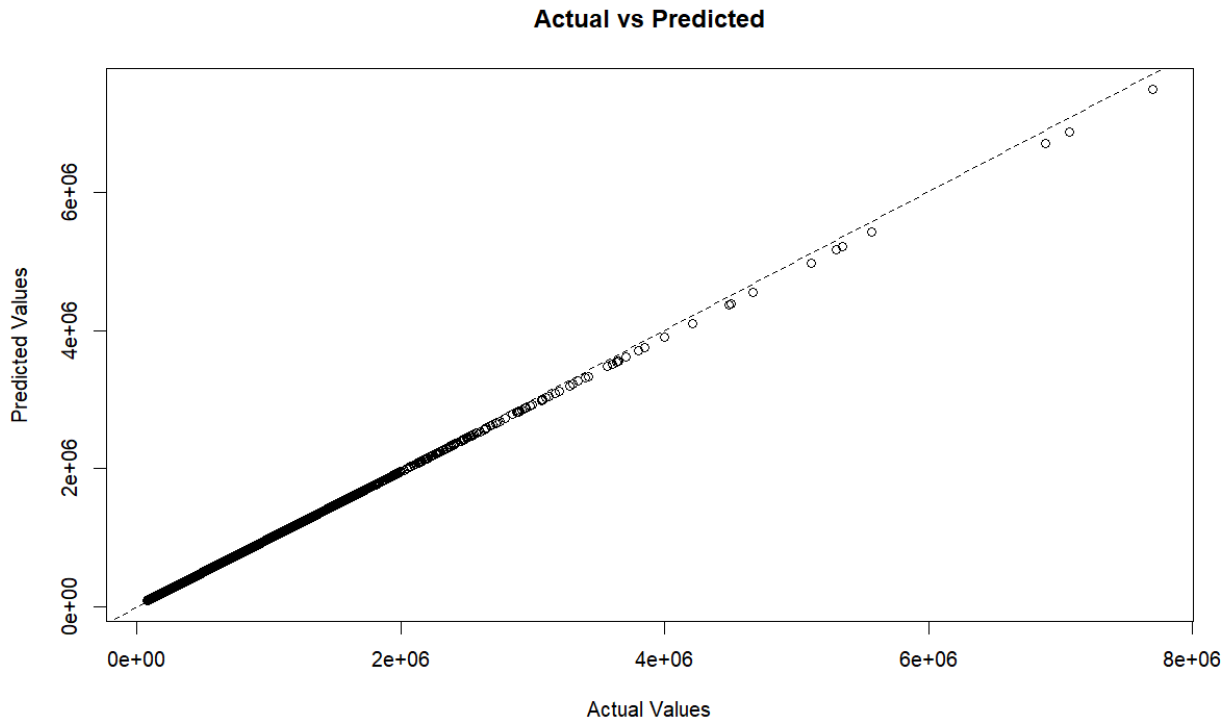


Figure 6: Poisson regression actual vs predicted

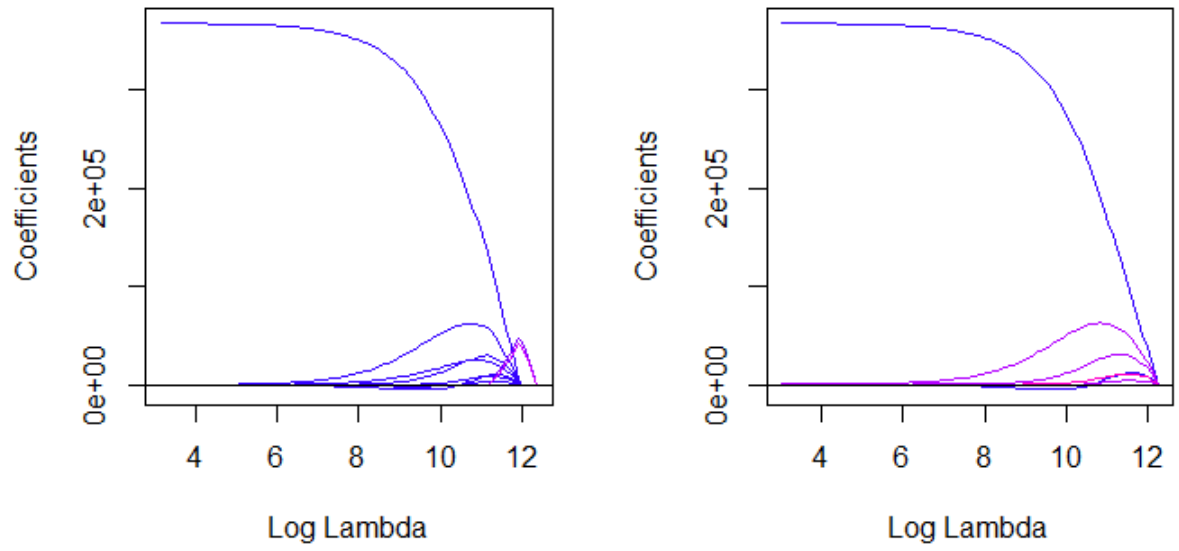


Figure 7: Group sets comparison for Group Lasso

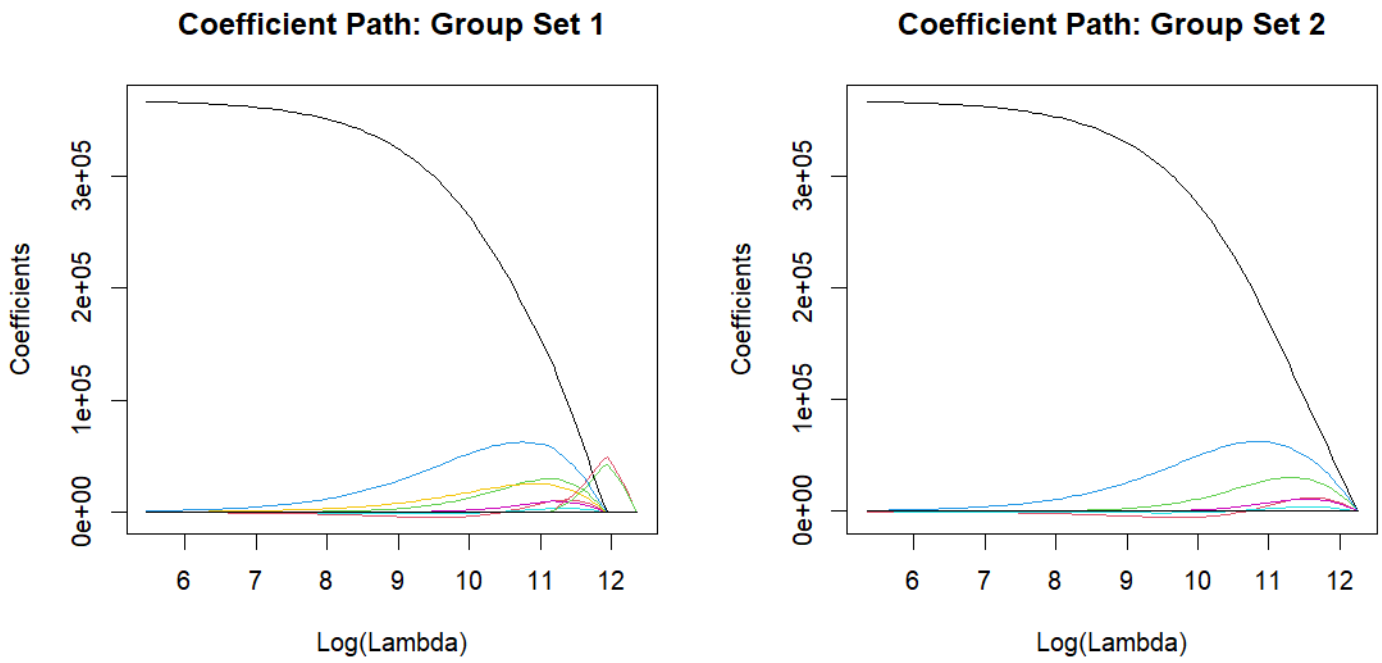


Figure 8: Coefficient path for Group Lasso

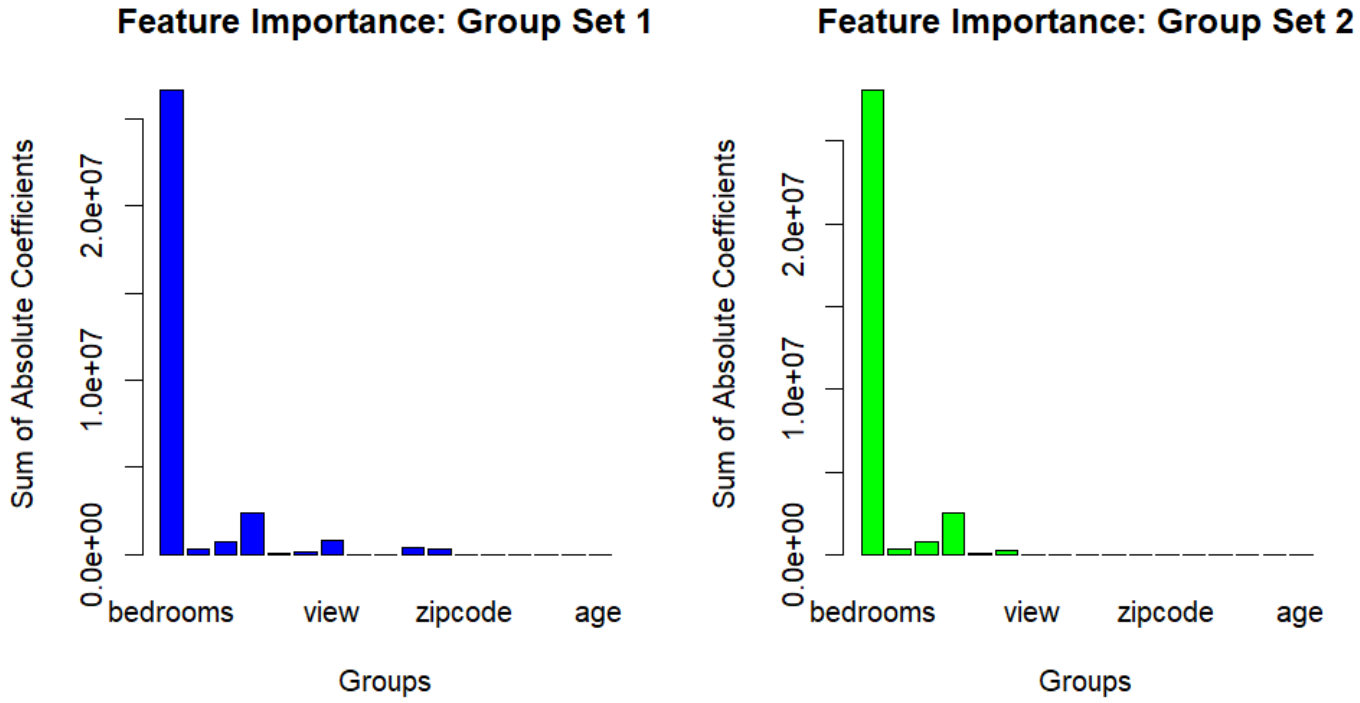


Figure 9: Best features for each group set for Group Lasso

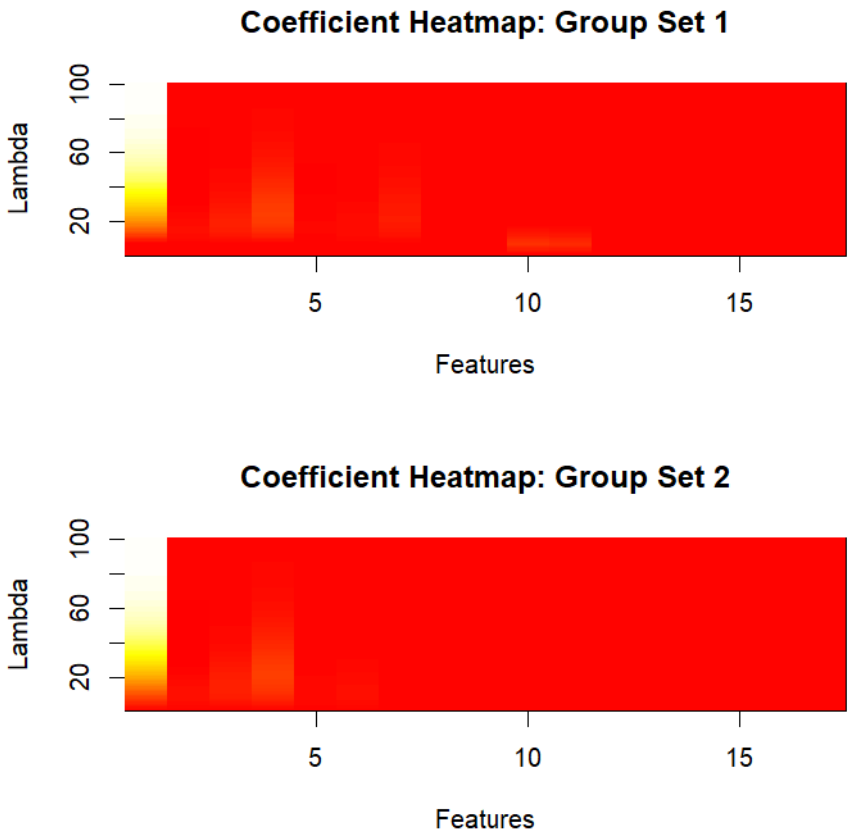


Figure 10: Feature heatmap for Group Lasso

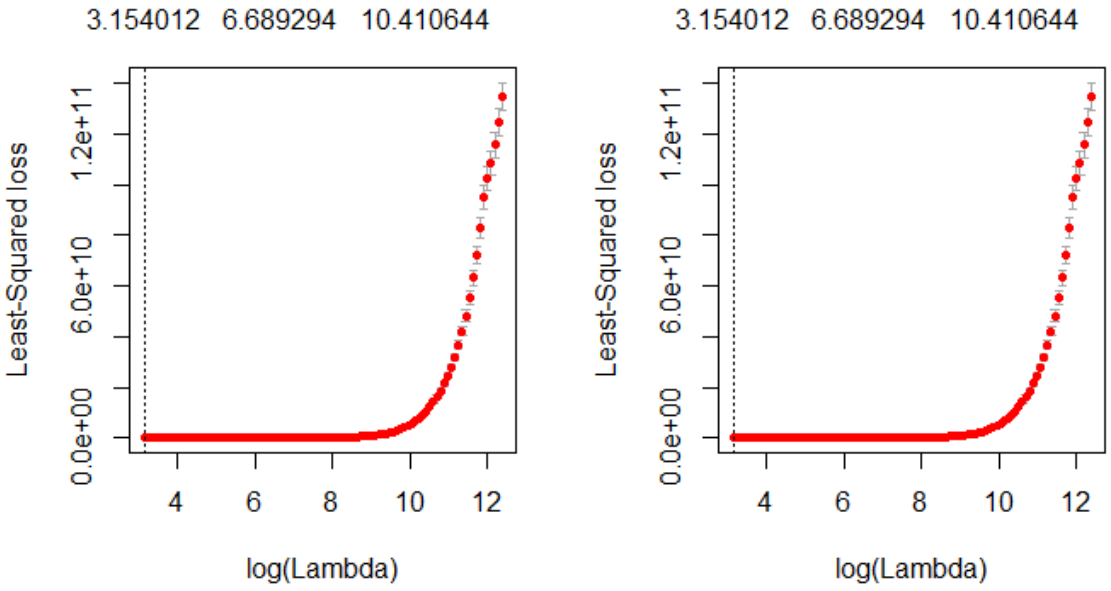
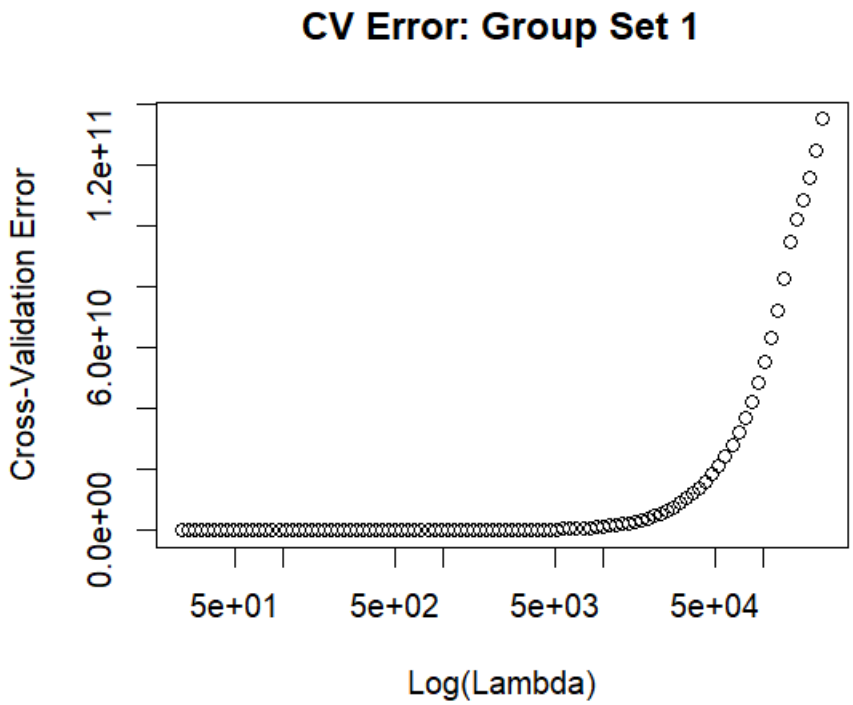


Figure 11: Cross validation plot for Group Lasso



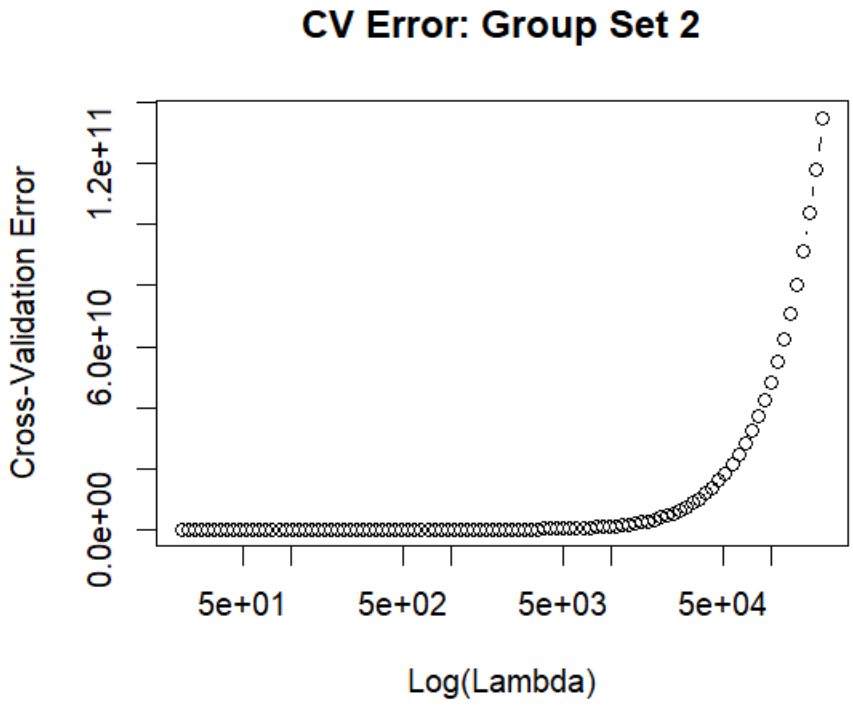


Figure 12: Cross validation error for Group Lasso

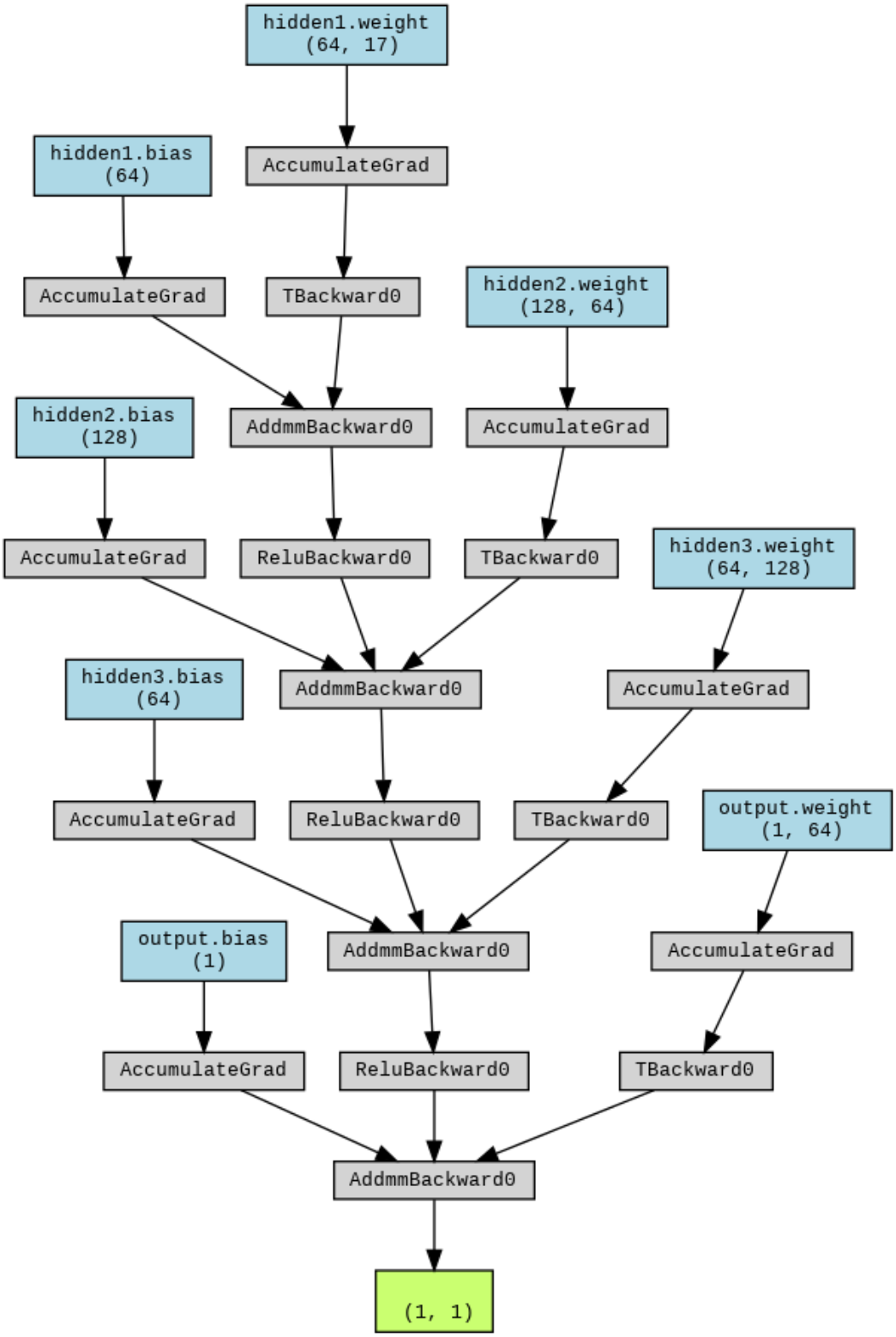


Figure 13: Neural Network architecture

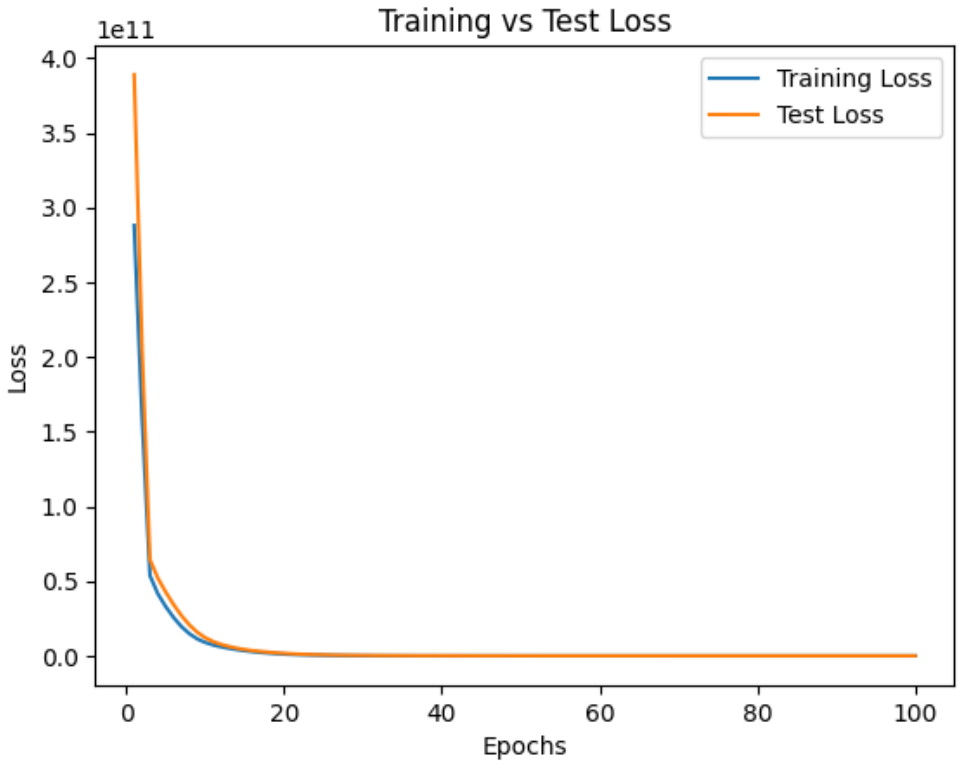


Figure 14: Neural Network training

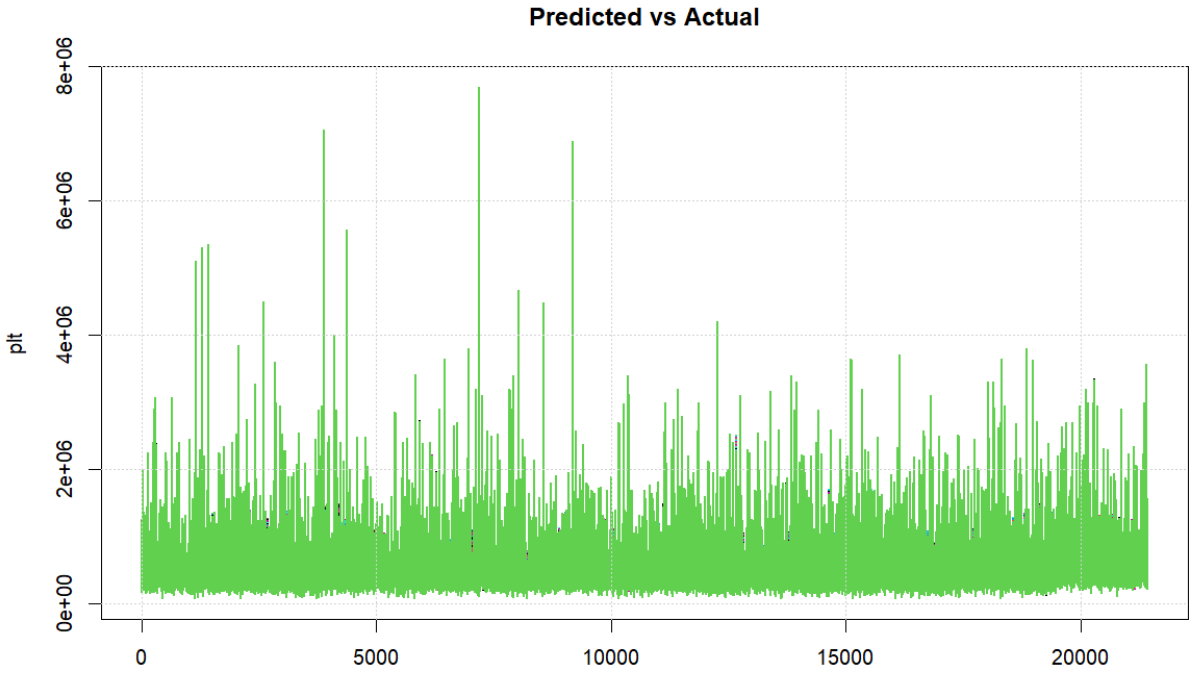


Figure 15: Group set 1 actual vs predicted

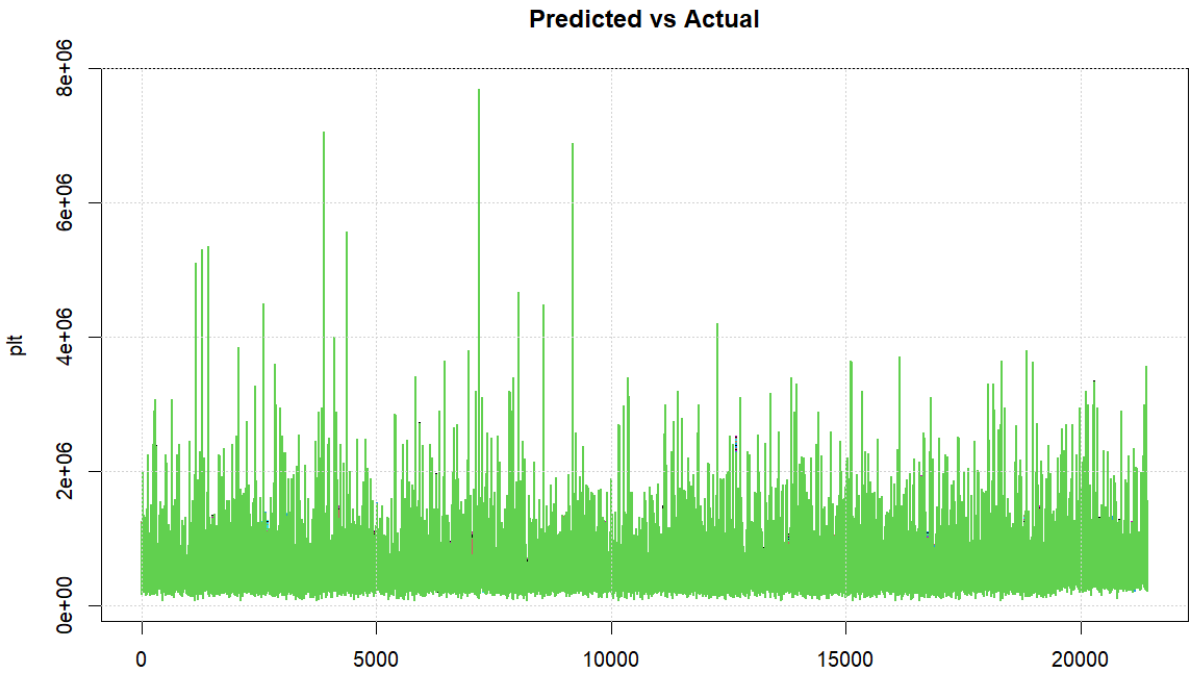


Figure 16: Group set 2 actual vs predicted

Model	MSE	R^2	MAE
Model 1	13572908807	-11.50303	48891.6
Model 2	11156390644	-9.27699	43118.46

Table 3: Group Lasso evaluation metrics

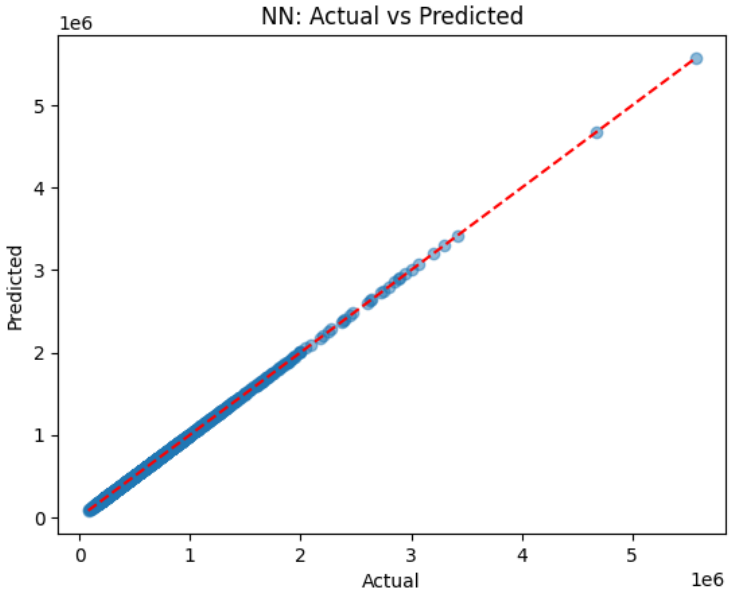


Figure 17: Neural Network actual vs predicted