Regression: Predictive Modelling in Action

Casper van Laar
casperdvanlaar@hotmail.com
University of Wolverhampton
MSc Artificial Intelligence
7CS070/UZ3: Concepts & Technologies of Artificial Intelligence
casperdvanlaar@hotmail.com

1. Introduction

This report applies regression algorithms to a real-world dataset to evaluate model performance, identify key predictive features, and propose improvements. Two models, Lasso Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, R.,1996) and Random Forest Regression (RFR) (Breiman, 2001), are compared. Linear regression, a foundational algorithm, models relationships between variables using a linear equation and optimizes predictions by minimizing a cost function like Mean Squared Error (MSE) (Equation 1) (James et al., 2013).

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2$$

Equation 1, equation for calculating the mean squared error. Where m is the number of data points, y_i is the actual value and $\widehat{y_i}$is the predicted value.

The following sections detail the methodologies and mathematical foundations of RFR and Lasso.

Algorithms

Lasso

Lasso is a linear regression technique that employs L1 regularization to shrink the coefficients of less important features, promoting sparsity and effectively performing feature selection. This helps improve model interpretability and reduces the risk of overfitting. By looking inside, the regression equation (equation 2), (Friedman et al., 2010):

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda\sum_{j=1}^{p}\beta_j$$

Equation 2

One can see the similarity between equation 2 for MSE and $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$. While the second part is the L1 regularization for the coefficients, β . λ is the regularization parameter that controls the strength of the penalty. A larger λ results in more aggressive regularization. Noting that with a sufficient high enough λ the coefficient goes down to 0. This feature elimination is what makes lasso, lasso.

RFR

RFR is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to deliver more accurate results. The method involves a set of trees, {h(x, Θ_k), k = 1, ..., K}, where {Θ_k} represents independent identically distributed random vectors, and K is the number of trees in the forest. Each tree casts a unit vote for the prediction of y given the input vector x. Given by the RFR estimate:

$$\bar{f}(x) \;= \frac{1}{K}\sum_{k=1}^{K} h(x, \Theta_{\mathrm{k}})$$

Equation 3, Where h(x, Θk) is the prediction of the k-th tree. At each node, a random subset of features is selected. The number of features is typically $\sqrt{p}$ or $log_2(p)$, where p is the total number of features.

For regression tasks, the most common splitting criterion is MSE. The goal is to minimize the MSE at each split. The MSE for a node is calculated as just as equation 2. The best split point is chosen based of the largest drop in MSE. The importance of a feature is calculated as the total reduction in MSE brought by that feature across all trees. RFR uses out-of-bag (OOB) error estimation. Where about 1/3 of the data not used in the bootstrapping is utilized to estimate the model's performance. To reduce the error rates RFR uses bias-variance trade-off. Meaning it averages multiples trees while maintaining a low bias:

$$Var\left(\bar{f}(x)\right) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

Equation 4, where ρ is the correlation between trees, σ2 is the variance of a single tree, and K is the number of trees. As K increases, the second term approaches zero, leaving only the first term. This demonstrates how random forests reduce variance through averaging.

2. Methodology

Data Selection

The dataset used for the regression task is focused on predicting house prices in King County from features, like grade, square feet living space and latitude and 15 others (table A1).

Preprocessing

The dataset was split into training and testing sets (80%/20%) following standard practices in predictive modeling (Kuhn & Johnson, 2013). For Lasso, features were scaled using StandardScaler, and alpha values were optimized. For RFR, outliers were removed based on IQR, and features were scaled using RobustScaler to reduce the impact of extreme values.

3. Results

A. Regression Results and Comparison

Model 1: Lasso Regression

The Lasso regression model showed a reasonable fit with an R^2 score of 0.7562, explaining about 76% of the variance in house prices. The model achieved a Root Mean Squared Error (RMSE) of 283,222.84, a Mean Squared Error (MSE) of 80,215,177,201.89, and a Mean Absolute Error (MAE) of 122,589.89, indicating its predictive performance. Figure 1A demonstrates the model's prediction accuracy, with points close to the diagonal line indicating better predictions.

Figure 1B displays the error distribution, showing random scatter, suggesting that errors are not systematically related to predicted values. In Figure 1C, the Q-Q plot compares residuals to a normal distribution, revealing slight deviations, particularly at the lower and upper ends, indicating that the residuals are not perfectly normally distributed. Figure 1D's histogram of errors shows slight skewness, aligning with the observations from the Q-Q plot.

Figure 1E highlights the most influential features, with grade being the most important predictor, followed by latitude and square footage. Lastly, Figure 1F shows the model's performance (R^2) at different regularization parameter (α) values. The optimal α of 0.000100 strikes a balance between model simplicity (fewer features) and predictive accuracy.

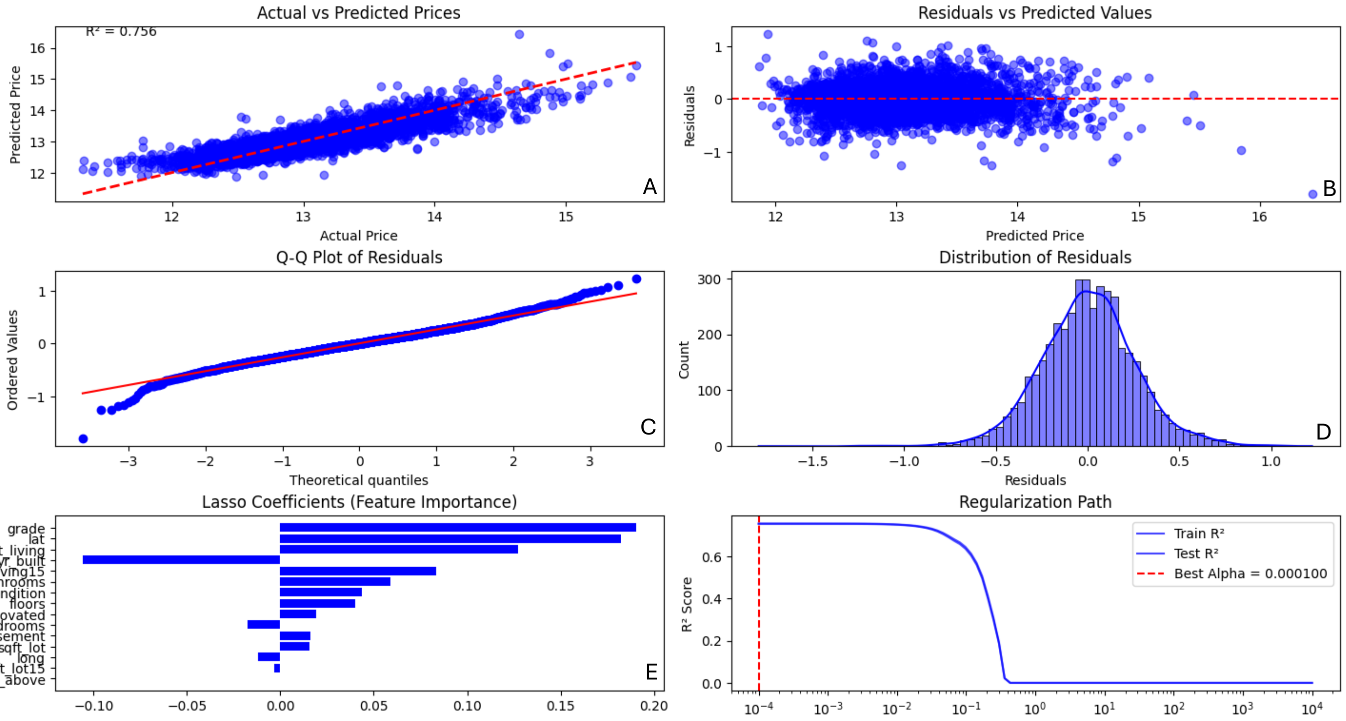

*Figure 1, Results LASSO, **(A).** This scatter plot shows the model's ability to predict house prices accurately. The closer the points are to the diagonal line, the better the model's predictions.*
***(B).** This plot assesses the model's error distribution. The random scatter suggests that the model's errors are not systematicallyrelated to the predicted values.*
***(C).** This plot compares the distribution of the residuals to a normal distribution. The slight deviation from normality indicatesthat the residuals are not perfectly normally distributed. Especially at the lower and upper end of the distribution.*
***(D).** This histogram visualizes the distribution of the model's errors. The slight skewness aligns with the observation from the Q-Q plot.*
***(E).** This bar chart highlights the most influential features in determining house prices, with grade' being the most important.*
***(F).** This plot shows the model's performance (R^2) changes with different values of the regularization parameter(α).The optimal α=0.000100 is highlighted, balancing model simplicity (fewer features) and predictive accuracy.*

Model 2: RFR

The RFR model demonstrates a strong fit with an R-squared value of 0.865, indicating its ability to predict house prices accurately (Figure 2A). This high R^2 reflects the model's capacity to capture non-linear relationships and feature interactions, effectively modelling complex data. It accounts for interactions between variables like grade, square footage, and location, which are key in-house price determination.

The model achieved a Mean Squared Error (MSE) of 5,833,154,060.63 and a Root Mean Squared Error (RMSE) of 76,375.09, highlighting its relatively low prediction errors compared to the data's scale. The Mean Absolute Error (MAE) was 53,138.43, further underscoring its accuracy in estimating house prices.

Figure 2C displays the error distribution, which is mostly random but shows a slightly higher, suggesting potential areas for improvement, particularly in predicting higher-priced houses.

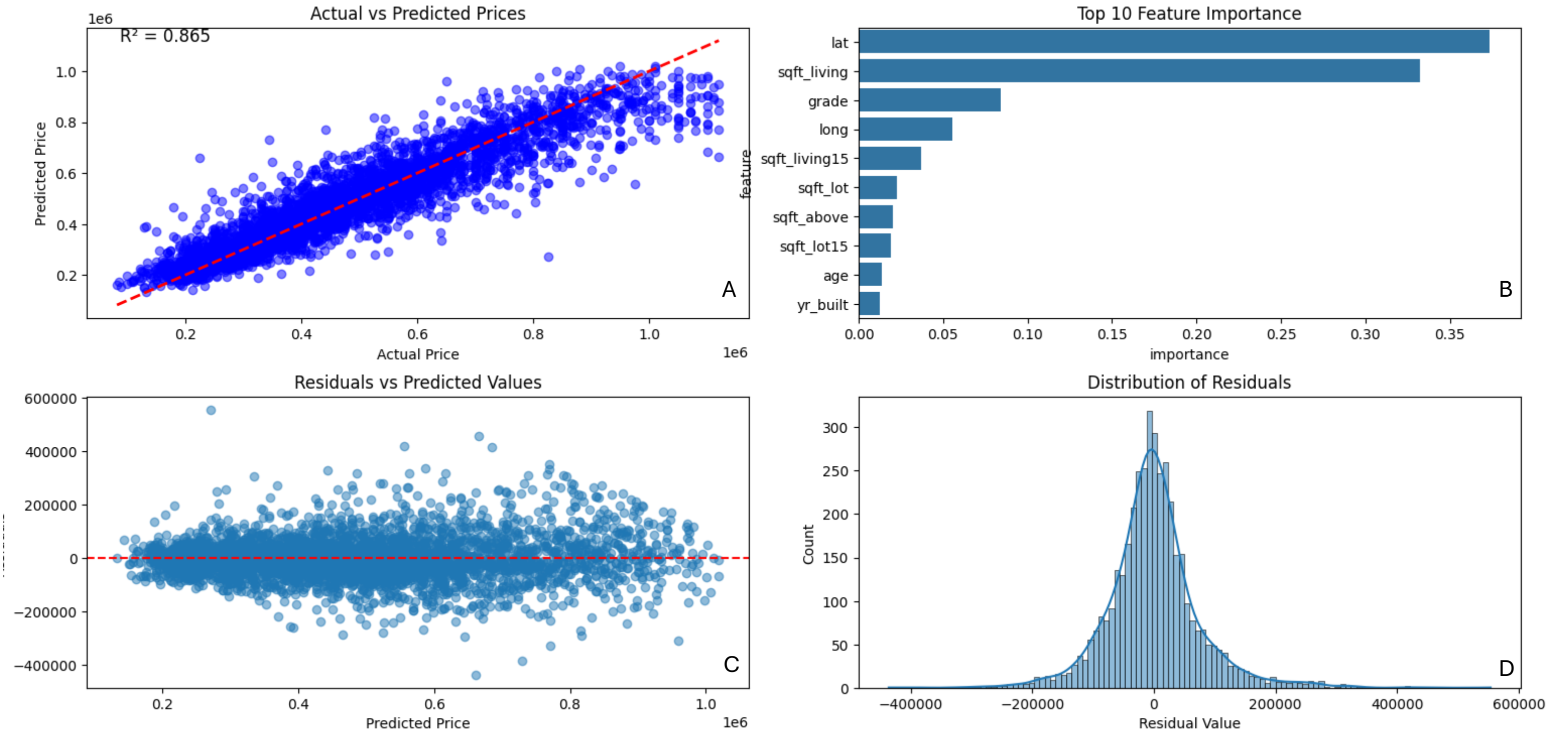

*Figure 2, Results RFR, **(A).** This scatter plot shows the model's ability to predict house prices accurately. The R-squared value of 0.865 indicates a good fit.*
***(B).** This bar chart highlights the most influential features in determining house prices, with latitude being the most important.*
***(C).** This plot assesses the model's error distribution. While mostly random, a slight pattern suggests potential areas for improvement in predicting higher-priced houses.*
***(D).** This histogram visualizes the distribution of the model's errors. The slight skewness aligns with the pattern observed in the previous plot.*

RFR VS LASSO

The performance differences between Random Forest Regression (RFR) and LASSO are due to their distinct approaches in handling data complexity and feature interactions. RFR, an ensemble method, uses multiple decision trees to capture complex, non-linear relationships in the data. It excels at modeling intricate feature interactions, like those between square footage, location, and house age, which often interact in non-linear ways. The RFR model achieved a high R^2 score of 0.865, with a relatively low RMSE of 76,375.09 and MAE of 53,138.43, highlighting its ability to handle noisy and variable datasets effectively. RFR's flexibility allows it to manage multicollinearity well, as it does not rely on explicit coefficient estimation. Instead, it builds multiple trees, each capturing different aspects of the data, leading to superior predictive performance.

In contrast, LASSO is a linear regression method that uses L1 regularization to shrink less important coefficients to zero. While it offers simplicity and interpretability, it struggles with non-linear relationships, assuming a linear connection between predictors and the target. The LASSO model's R^2 score of 0.7562 and higher RMSE of 283,222.84, along with an MAE of 122,589.89, reflect its limitations in capturing complex interactions compared to RFR. LASSO is also sensitive to multicollinearity, often arbitrarily eliminating correlated features, which can lead to unstable estimates and reduced predictive accuracy. However, its regularization properties make it useful for feature selection and reducing model complexity.

Overall, RFR tends to outperform LASSO in datasets with complex interactions and noise, as evidenced by its superior error metrics. LASSO, however, is better suited for problems requiring simplicity and interpretability. The choice between these models depends on the complexity of the data and the importance of predictive accuracy versus model transparency

4. Discussion of Model Improvements

This analysis compares Lasso and Random Forest Regression (RFR) using performance metrics such as RMSE, MAE, and R^2. However, to strengthen the findings, statistical significance tests and confidence intervals should be incorporated. A paired t-test or Wilcoxon Signed-Rank Test would formally assess whether the differences in performance are statistically significant, with the Wilcoxon test being suitable when normality assumptions are not met.

To improve model performance, several strategies can be employed. Feature engineering can enhance predictive power by adding interaction or polynomial terms to capture non-linear relationships, especially in RFR. Increasing the dataset size through augmentation or bootstrapping can benefit ensemble methods like RFR. For Lasso, addressing multicollinearity by removing highly correlated features or applying techniques like Principal Component Analysis (PCA) can improve stability.

Hyperparameter tuning is crucial for both models. For RFR, parameters like the number of trees and tree depth should be adjusted, while Lasso requires tuning of the regularization parameter. Cross-validation should be used to determine optimal parameters, ensuring more stable and reliable model performance.

Regarding feature selection, "Both models inherently perform feature selection—RFR through feature importance and Lasso via regularization, which aligns with advanced statistical learning techniques (Hastie et al., 2009). Additional feature elimination may only be necessary when there are many irrelevant or highly correlated features, which could cause instability or overfitting. However, for the current dataset, the feature selection in both models should be sufficient.

Further improvements can include exploring alternative models such as Gradient Boosting, XGBoost, or Neural Networks, which can better capture complex non-linear relationships and feature interactions. Hyperparameter optimization can be done through grid search or random search to fine-tune the performance of these models. Additionally, domain-specific feature engineering, such as creating location-based features like "region" from latitude and longitude, could improve predictive accuracy by capturing regional price trends.

In summary, while the analysis presents valuable insights, employing statistical significance testing, refining feature engineering, and experimenting with alternative models and tuning techniques would further enhance the robustness and accuracy of the findings.

References

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Appendix

Feature	**Description**
Price	Sale price of the house
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Sqft Living	Total living area in square feet

Sqft Lot	Total lot size in square feet
Floors	Number of floors in the house
Grade	Overall construction and design quality rating
Sqft Above	Living area above ground in square feet
Sqft Basement	Basement area in square feet
Yr Built	Year the house was built
Yr Renovated	Year the house was last renovated
Lat	Latitude coordinate of the house
Long	Longitude coordinate of the house
Sqft Living15	Living area of nearest 15 neighbors in sq. ft.
Sqft Lot15	Lot area of nearest 15 neighbors in sq. ft.
Condition	Condition of the house (overall maintenance)

Table A1, Features of houses in king country.

Feature	**Code Given**	**Unit**	**Data Type**
Age	Age	Years	Numeric
Sex	Sex	1, 0	Binary
Chest Pain Type	Chest pain type	1, 2, 3, 4	Nominal
Resting Blood Pressure	Resting bp s	mm Hg	Numeric
Serum Cholesterol	Cholesterol	mg/dl	Numeric
Fasting Blood Sugar	Fasting blood sugar	1, 0	Binary
Resting Electrocardiogram Results	Resting ecg	0, 1, 2	Nominal
Maximum Heart Rate Achieved	Max heart rate	71–202	Numeric
Exercise Induced Angina	Exercise angina	0, 1	Binary
Oldpeak = ST Depression	Oldpeak	-	Numeric
Slope of Peak Exercise ST Segment	ST slope	0, 1, 2	Nominal
Class	Target	0, 1	Binary

Table A2, Attributes for heart conditions

Feature	**Description**
Sex	1 = Male, 0 = Female
Chest Pain Type	1: Typical angina, 2: Atypical angina, 3: Non-anginal pain, 4: Asymptomatic
Fasting Blood Sugar	1 = True (Fasting blood sugar > 120 mg/dl), 0 = False
Resting Electrocardiogram Results	0: Normal, 1: ST-T wave abnormality (inversions/ST elevation or depression > 0.05 mV), 2: Left ventricular hypertrophy (Estes' criteria)
Exercise Induced Angina	1 = Yes, 0 = No

Slope of Peak Exercise ST Segment	1: Upsloping, 2: Flat, 3: Downsloping
Class	1 = heart disease, 0 = Normal

Table A3, Description of the nominal attributes.