# A comparison between Ridge Regression and Random Forest as applied to UK housing price prediction

**Author: Myo Shwe Sin Ei | INM431 Machine Learning**

## 1. Description and Motivation

- The purpose of this research is to build and compare two machine learning models, Ridge Regression and Random Forest for predicting UK housing prices. UK housing sale transactions data from 2013-2023 are used in this study to analyse and predict housing market.

## 2. Data Set and Exploratory Analysis

- The dataset is an enhanced version of UK housing transaction data from HM Land Registry on Kaggle(1995-2023). The original dataset was cleaned to 249,766 samples with 10 features to focus on 2013-2023 housing prices.
- High correlation of property type and tenure type between flats and leasehold shown by Figure 1 property distribution and tenure distributions. From the price analysis, we can see that mean price is £750k and error bars extend from approximately £2500k to £4000k, and the large gap between mean and median from overall price trends (Figure 2) which shows severe outliers. Price distribution boxplot is also showing impossible house prices of trillion-pound houses which means proper data processing and encoding is needed. Greater London dominates nearly 8% of the data which explains the flat dominance. Seasonal patterns show peak activity in March, likely resulted by end of UK tax year.
- These issues are addressed through data preprocessing by encoding such as log transformation of target variable price to normalize skewed distribution, one-hot encoding of property type using one category(other) as the reference to avoid dummy variable trap. Binary encodings are used for features with two categories such as Old/New, Duration and PPD. Year is encoded as direct numeric and month, as cyclical to show the cycle of seasons. Target encoding is done for high-cardinality location features such as County, District and Town to prevent explosion of features. Smoothing factor is used which gives a weighted average between category mean and global mean to shrink rare categories toward the mean. These encodings are computed on training data only to avoid data leakage and artificially inflated performance. Computed encoding maps from training data are stored in a preprocessing MATLAB file which later applies these maps to test data and use global mean for unseen categories.

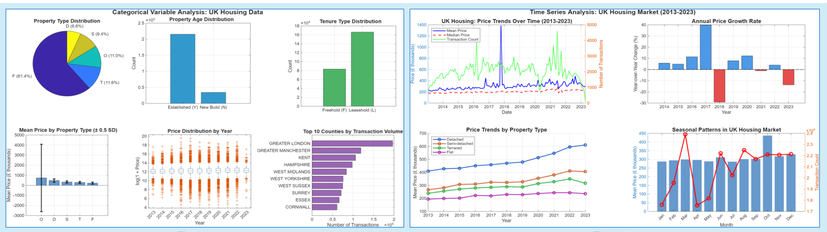


Figure 1




Figure 2

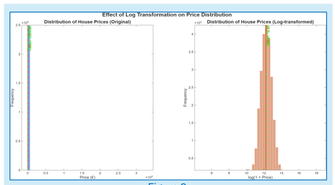


Figure 3

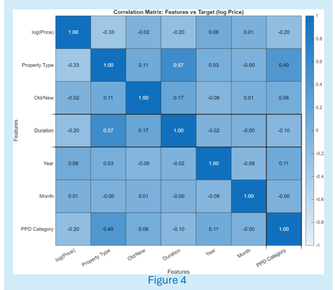


Figure 4

## 3. Hypothesis Statement

- Random forest will outperform ridge regression due to the ability to capture non-linear relationships.

## 4. Model Selection

### 4.1. Ridge Regression (Ridge)

#### 4.1.1. How it works

- Ridge Regression introduces L2 regularization, which adds a penalty term proportional to the squared magnitude of coefficients to the loss function.
- The optimization finds a balance between how accurate the model matches the data and how much the weight values are reduced using λ.

#### 4.1.2. Advantages

- Ridge has faster training and prediction time and makes it computationally efficient regardless of dataset sizes.
- Its coefficients are highly interpretable as each coefficient directly indicate a feature's marginal effect on target.
- It can handle multicollinearity through coefficient shrinkage and prevent the extreme coefficient values that result from highly correlated features.
- It can lower the risk of overfitting while maintaining low bias with proper selection of lambda value.

#### 4.1.3. Disadvantages

- Ridge assumes linear relationships between features and target variable.
- It can't capture complex interactions or non-linear patterns without feature engineering and explicit polynomial or interaction terms.

### 4.2 Random Forest (RF)

#### 4.2.1. How it works

- Random Forest constructs an ensemble of decision trees during training and outputs the mean prediction across all trees.
- It uses bootstrap aggregating(bagging), where each tree is trained on a random sample drawn with replacement from the dataset with a random set of features.

#### 4.2.2. Advantages

- RF captures non-linear relationships and handles complex interactions between features without manual specifications.
- It is robust to outliers and noise since the ensemble average mitigates the mistakes.
- It can reduce the risk of overfitting and produce more accurate predictions.

#### 4.2.3. Disadvantages

- RF takes longer and is computationally expensive for both training and prediction.
- Its results are less interpretable than linear models.

## 5. Training and Evaluation Methodology

- 9:1 ratio used to split data into training and testing with 224,789 samples for training and 24,977 samples held out for final evaluation.
- 10-fold Cross-Validation was used on training with 9 folds to train and 1to validate for all combinations.
- Standardization was done for all numeric features by transforming them to zero mean and unit variance to prevent data leakage and overfitting. Categorical features were one-hot encoded and a smoothing factor was applied to shrink rare category estimates toward the global mean.
- Model performance was evaluated with RMSE, $R^2$ and MAE after getting the best hyperparameters.
- The models are tested with their chosen parameters on the test data.

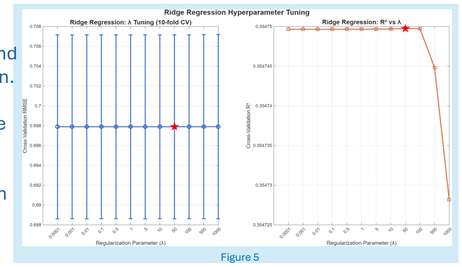


Figure 5

## 6. Choice of parameters

- Hyperparameter optimization was done through Grid Search with Cross-Validation.

### 6.1. Ridge Regression

- Lambda grid : [1, 5, 10, 50, 100, 500].
- Best λ selected by minimum CV RMSE.
- Best parameter for Ridge: 50

### 6.2. Random Forest

- NumTrees : [50, 100, 200]
- MinLeafSize : [1, 5, 10, 20]
- NumPredictorsToSample : [p/3, √p, p]
- Grid search: 36 combinations x 10 folds
- Total: model fits evaluated
- Best parameters for Random Forest : 200 trees, MinLeaf = 5, NumPred = p/3

## 7. Experimental results

- Experimental evaluation shows that Random Forest outperforms Ridge Regression, both for training/validation sets.
- Train an validation set was 224,789 samples(90%) and test set was 24,977 samples(10%).
- Random Forest achieved a mean CV RMSE of 0.608± 0.008 compared to Ridge Regression's 0.698 ± 0.009, showing improvement in prediction error.
- Both models exhibit substantial absolute errors however, MAE provided more interpretable insight with Ridge's average absolute error at £182,828 and Random Forest's at £167,947.
- Residual analysis also shows that random forest has better stability and predictive robustness compared to Ridge Regression.
- For Ridge Regression, the residuals vs fitted plot shows that the linear model struggles to capture pricing dynamics at the market's upper end. This can be seen in later test values.
- The residual distribution is approximately normal and centred at zero showing unbiased predictions on average.
- However, the Q-Q plot with heavy tails indicate Ridge Regression can mis-predict outlier properties.
- The same goes for Random Forest with heavy tails on both end, indicating the model's struggle with values at the edges of data.
- This likely results from the limitation of training samples where properties exceeding £1 million or below £100,00 are less than 10% of the data.
- Random Forest shows a more uniform spread across the prediction range which shows more consistent prediction compared to Ridge Regression.

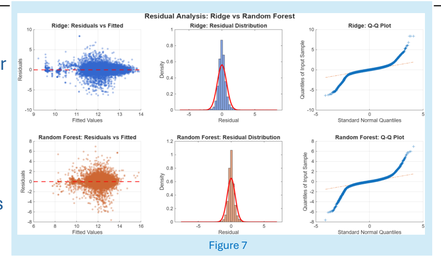


Figure 7

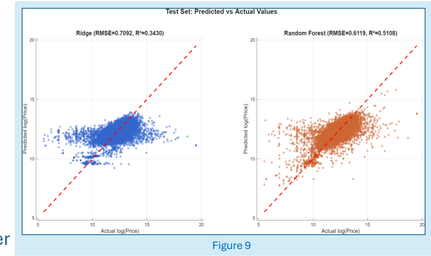


Figure 9

## 8. Analysis and critical evaluation of results

### 8.1. CV results interpretation

- Random Forest outperforms Ridge Regression in predictive accuracy for overall evaluation metrics.
- Cross validation indicates RF achieve lower mean RMSE with slightly lower std showing better accuracy and stability across all 10 folds

### 8.2. Model comparison:

- RF captures non-linear price patterns
- Ridge provides interpretable coefficients
- RF x slower to train but
- The evaluation of these models reveals neither is currently achieving high accuracy as both have high Mean Absolute Percentage Errors. However, based on the metrics and predictions, Ridge is slightly performing better for this specific dataset due to its higher stability. While RF is better at predicting mid-range properties with higher accuracy, its stability is compromised by extreme outliers, which can be seen in index 6 of Figure 8 with 246% over-prediction.

SAMPLE PREDICTIONS (First 10 Test Samples)

| Index | Actual (£) | Ridge (£) | RF (£) | Ridge Err% | RF Err% |
|---|---|---|---|---|---|
| 1 | £308000 | £348020 | £367798 | 13.0 | % 19.4 % |
| 2 | £1725000 | £236785 | £331251 | 86.3 | % 80.8 % |
| 3 | £305000 | £231775 | £341378 | 24.0 | % 11.9 % |
| 4 | £875000 | £517979 | £538184 | 40.8 | % 38.5 % |
| 5 | £334950 | £109672 | £129182 | 67.3 | % 61.4 % |
| 6 | £60000 | £83504 | £207737 | 39.2 | % 246.2 % |
| 7 | £174000 | £175346 | £122193 | 0.8 | % 29.8 % |
| 8 | £235000 | £116590 | £141916 | 50.4 | % 39.6 % |
| 9 | £125000 | £126476 | £123478 | 1.2 | % 1.2 % |
| 10 | £385000 | £335201 | £303632 | 12.9 | % 21.1 % |

Figure 8

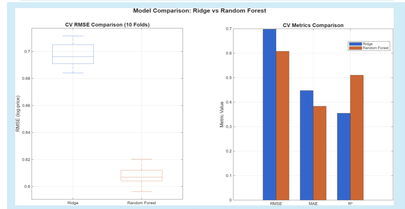


Figure 6

#### 8.2.1. Ridge Regression

- Ridge Regression displayed low variance for predictions but had high bias.
- While it failed to predict high-range properties, it performed well on mid-range properties with almost perfect predictions on index 7 and 9.
- Ridge's errors were generally clustered around the mean price.

#### 8.2.2. Random Forest

- RF struggled to generalize unseen data at the edges of the data distribution which can be seen at index 6 with 246% error.
- This suggests that RF model has insufficient training samples in the low-range properties where the result was a huge over-estimation.
- Compared to Ridge, RF outperformed on mid-range samples however, due to its lower stability for overall predictions, it might be risky for predicting housing prices.

### 8.3. Critical Evaluation

- Although the models performed well for properties between £150-£350k range, former analysis show that data distribution is an important and main issue for the study due to both models' failure on the higher or lower priced properties.
- In conclusion, Ridge Regression performed better for this project as its failure are more predictable than RF with erratic misprediction. This is also because RF didn't have enough training samples and features to capture the difference between a cheap flat and a standard apartment.

## 9. Future Directions

### 9.1. Lessons learned

- Random Forest requires more balanced samples to make predictions with high accuracy and without that, outliers can throw the results off.
- It may seem like the model is performing well but might be completely wrong on individual predictions.
- More complex models don't always mean better results and a simpler linear model with regularization is seen to predict stabler without guesses too wild.
- The quality of features and dataset are very important.

### 9.2. Future work

- Current dataset failed to train models effectively. Future work would be better to split the data between standard housing and luxury housing to prevent skewing the overall housing price predictions.
- Need more quality features such as "Distance to nearest public transport, crime, schools" to help Random Forest differentiate between a cheap and standard property.
- Implement gradient-boosting algorithms for structured data to handle outliers and correct the errors of previous trees

## References


- Begum, S. (2023). House Price Prediction by Machine Learning Technique—An Empirical Study. *Disruptive technologies and digital transformations for society 5.0*, pp.115–133. doi:https://doi.org/10.1007/978-981-99-5354-7_7.
- Dmitriy Bolotov (2025). *Six Approaches to Time Series Smoothing*. [online] Medium. Available at: https://medium.com/@dmitriy.bolotov/six-approaches-to-time-series-smoothing-cc3ea9d6b64f.
- GeeksforGeeks (2020). *Feature Encoding Techniques Machine Learning*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/feature-encoding-techniques-machine-learning/.
- None Preethi, Murthy, R., Vani Hiremani, Devadas, R.M. and R. Sapna (2025). Optimizing Polynomial and Regularization Techniques for Enhanced Housing Price Prediction Accuracy. *SN Computer Science*, 6(2). doi:https://doi.org/10.1007/s42979-024-03578-7.
- Tshepo Chris Nokeri (2021). *Data Science Revealed*. *Apress eBooks*. Apress Berkeley, CA. doi:https://doi.org/10.1007/978-1-4842-6870-4.
- Whieldon, L. and Ashqar, H.I. (2022). Predicting residential property value: a comparison of multiple regression techniques. *SN Business & Economics*, 2(11). doi:https://doi.org/10.1007/s43546-022-00358-4.