


| | |
|--|---|
| SCHOOL OF GEOGRAPHY UNIVERSITY OF LEEDS |  UNIVERSITY OF LEEDS |
|--|---|

COURSEWORK COVERSHEET

| | | | | | | | | | |
|----------------------------|--|---|---|---|---|---|---|---|---|
| Student ID number | 2 | 0 | 1 | 7 | 4 | 0 | 4 | 5 | 4 |
| Module code | GEOG5917M | | | | | | | | |
| Module title | BIG DATA AND CONSUMER ANALYTICS | | | | | | | | |
| Assignment title | ANALYZING LIVERPOOL HOUSING MARKET THROUGH ELASTIC NET MODEL | | | | | | | | |
| Marker | ROGER BEECHAM | | | | | | | | |
| Declared word count | 2496 | | | | | | | | |

By submitting the work to which this sheet is attached you confirm your compliance with the University's definition of Academic Integrity as: "a commitment to good study practices and shared values which ensures that my work is a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine". Double-check that your referencing and use of quotations is consistent with this commitment.

You also confirm that your declared word count accurately reflects the number of words in your submission, excluding the overall title, bibliography/reference list, text/numbers in tables and figures (although table and figure captions are included in the word count).

INTRODUCTION

Amid the dynamic economic conditions that have arisen in Britain since Brexit, the Liverpool real estate market presents itself as an intricate tapestry of prospects and unpredictability. Predictive modelling in such a landscape is complicated due to its dynamic characteristics and the wide array of determinants influencing property values. This study aims to employ a predictive Elastic Net (EN) regression model on a dataset featuring 10 variables, capturing a range of housing and socio-economic factors across 2211 observations. The EN method, an extension of linear regression, balances feature selection and regularization, making it ideal for datasets with multicollinear predictors.

The remarkable ability of the EN model to tackle multicollinearity and improve feature selection in intricate datasets was first illustrated in the foundational work of Zou and Hastie (2005). Their study highlighted the superior performance of the model in comparison to conventional regression techniques through its demonstration of efficacy in managing high-dimensional data characterized by prevalent multicollinearity. In the field of real estate analytics, where the prediction of property values is complicated by the interdependence of numerous factors, this attribute proves to be beneficial. The advantage of the model is that it keeps the feature selection quality from the lasso penalty as well as the effectiveness of the ridge penalty (Lin and Li, 2023). It simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. It is like a stretchable fishing net that retains 'all the big fish' (Zou and Hastie, 2005). By utilizing the capabilities of EN, the model can decrease overfitting, enhance generalization, and the interpretability of the results by automatically selecting features. This indicates that the model not only makes predictions about prices, but also determines which factors have the most significant impact on them.

The objectives of this study are threefold: Firstly, to develop an Elastic Net model for housing prices in Liverpool. Secondly, to identify key factors influencing housing prices within this context and Thirdly, to evaluate the potential of advanced regression techniques in real estate market analysis. According to an article published by Lavelle Estates (2023) some of the factors affecting the Liverpool property market includes economic conditions, interest rates, and government policies. Through this approach, the study aims to provide some valuable information on the same.

METHODOLOGY

The Liverpool dataset has 2211 observations and 10 variables that combine socioeconomic indicators with property-specific characteristics. This range encompasses the diversity of the housing market and establishes a strong basis for predictive analysis. Below is the dataset tabulated with each variable explained

| VARIABLE | DESCRIPTION |
|----------|------------------------------------|
| Price | Property sale price (in £1000s) |
| Beds | Number of bedrooms in the property |

| | |
|---------|--|
| Gs_Area | Percentage of local green space area |
| U16 | Percentage of the population aged 16 and under |
| U25 | Percentage of the population aged 16-24 |
| U45 | Percentage of the population aged 25-44 |
| U65 | Percentage of the population aged 45-64 |
| O65 | Percentage of the population aged 65 and above |
| Unmplyd | Percentage of unemployed population |

Table 1: Variables and their description

The preprocessing process started by converting a spatial object into a conventional data frame to match the data structure needed for analysis. The spatial geometry column, which was not useful for prediction in this context, was removed. An analysis of the structure using the *str()* function confirmed that there were no abnormalities in the variable types. The dataset was thoroughly examined to identify any missing values – none was found. Prior to modelling, the problem of skewness was also evident in specific variables, namely ‘*gs_area*’, ‘*unmplyd*’, ‘*u25*’, and ‘*o65*’ as seen in the histograms below.

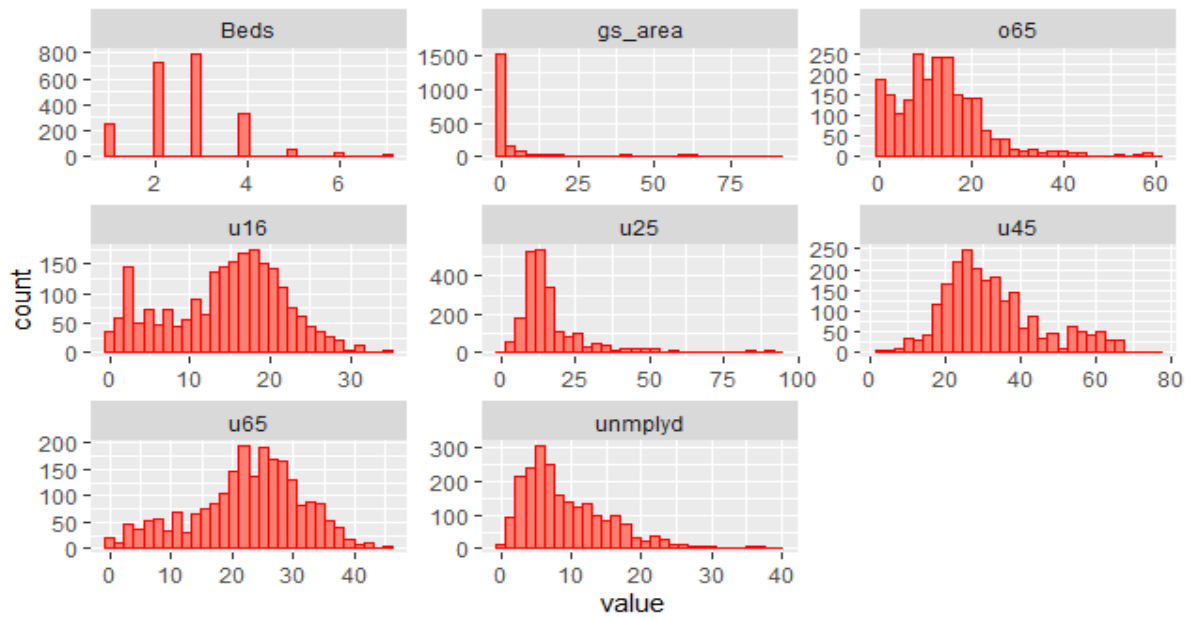


Figure 1: Histograms of the distributions of the variables.

Skewness is defined to be the third standardized central moment (Lin and Li, 2023). It may introduce bias into the estimations of the model and compromise the precision of the outcomes. To achieve variance and distribution throughout the dataset, log transformations were implemented, which are a conventional method for fitting positively skewed data. By compressing the scale of the variables, the log transformation effectively reduces the impact of extreme values, thereby improving the suitability of the variable for modelling. This approach effectively compresses the scale of the variables, reducing the influence of outliers and making the variable distribution more symmetric (Osborne, 2010).

Square root and Box-Cox transformations were also considered for skewness and variable scale preprocessing. The square root transformation reduces outliers but was less effective for our variables' skewness. The Box-Cox transformation, a broader method, requires finding a parameter that best normalizes the data. While this may provide a more tailored solution to each variable's distribution, parameter selection and interpretation tend to be more complicated. Due to the simplicity and effectiveness of log transformations in similar situations, this method was chosen for its ability to handle the commonly observed positive skewness in socioeconomic and housing-related data and to strike a balance between methodological rigor and practical applicability.

In addition to the transformations, the dataset was divided into training and test subsets using *createDataPartition()* which takes the place of our manual data splitting. It also does some extra work to ensure that the train and test samples are somewhat similar (Dalpiaz, 2020). Here, training was assigned 80% of the data and testing was assigned the remaining 20%. The division described here is crucial in the process of model validation as it enables us to fit the model to the training set and assess its efficacy by simulating predictions on novel, unobserved data on the test set. By employing this approach, one can evaluate the predictive capability and resilience of the model, prevent overfitting, and verify that the model is capable of extrapolating beyond the instances it was trained on.

During the preprocessing stage, the *preProc()* function from the caret package which is a wrapper for hundreds of machine learning algorithms (Comber, 2023) was utilized to standardize all predictors and automate the log transformation of variables displaying positive skewness. It was specified that the argument would perform both scaling and centering of variables. It is the most straightforward data transformation. In contrast to scaling, which converts the variable to units of standard deviations from the mean, centering involves dividing each centered value by the standard deviation of the variable. It centers and scales a variable to mean 0 and standard deviation 1 (Lin and Li, 2023). It is also important to scale the data only after splitting the data. Otherwise, the model will influence the test dataset even before the data undergoes the validity test, reducing its reliability (Comber and Brunsdon, 2021).

By partitioning the data and applying the steps, the EN model was trained on a dataset where variable scale does not affect learning. It also guarantees that the EN's regularization part, which penalizes larger coefficients to prevent overfitting, works uniformly across predictors.

While modelling, cross-validation is essential to improve the fitness of the model and to prevent overfitting (Comber and Brunsdon, 2021). It generates parameters that help refine the model as it is created and ensure that overfitting is prevented and hence, a 10-fold cross-validation was employed within the training phase to optimize the model's parameter. The *trainControl()* function from the caret package was used to facilitate the implementation of cross-validation in the predictive modelling process. It specifies the resampling scheme (Dalpiaz, 2020). The method was specified as "cv" to indicate the use of cross-validation, with the number of folds set to 10. The parameters were inputted into the train function, which was responsible for

fitting the EN model. By integrating cross-validation into the preprocessing and training pipeline, the model's reliability was greatly enhanced.

The primary metric chosen for performance evaluation was the Root Mean Squared Error (RMSE). This metric is favored for its high sensitivity to significant errors and its capacity to offer a detailed comprehension of model accuracy. It provides a precise measure of the model's accuracy in predicting housing prices by quantifying the average magnitude of prediction error.

This rigorous and systematic approach, which includes transforming the data, partitioning it, and using robust validation techniques, guarantees that the resulting Elastic Net model is well-calibrated and capable of accurately predicting house prices in Liverpool.

RESULTS

Elastic Net is a generalization of lasso and ridge regression (Zou and Hastie, 2005). Ridge penalty shrinks correlated predictor coefficients towards each other, while the lasso picks one and discards the rest. This hybrid methodology enables a subtle fine-tuning of the model by utilizing the alpha parameter, which effectively manages the impact of lasso and ridge penalties. As a result, it guarantees both the comprehensibility and precision of the model. The model is also a powerful tool for predictive modelling in situations where there are many closely related predictors. It improves the reliability of predictions and the simplicity of the model, making it an effective choice in these scenarios. The alpha parameter of the EN model regulates the proportion of Ridge and Lasso penalties, enabling an equilibrium that can be precisely adjusted to suit the data structure. This process enhances the accuracy of predictions and the selection of variables.

The EN Model's tuning phase was crucial to its predictive accuracy and complexity management to predict Liverpool house prices. Alpha, which integrates L1 (Lasso) and L2 (Ridge) penalties, was varied across a spectrum to find the best combination to reduce prediction error and improve accuracy. This process played a crucial role in leveraging L1's feature selection process and L2's ability to uniformly emphasize coefficient size, protecting against overfitting with highly correlated predictors. However, Lambda controls the model's regularization intensity, and finding its optimal value was crucial for achieving a delicate equilibrium where the model complexity is constrained enough to prevent overfitting without oversimplifying it to underfitting. Cross-validation is used to select a good λ value (Dalpiaz, 2020). A 10-fold cross-validation produced the optimal pairing of $\alpha = 0.2222222$ and $\lambda = 0.001$, highlighting the model's skill in combining feature selection with the regularization strength to prevent overfitting, while maintaining model simplicity and interpretability.

The EN model's performance was quantitatively assessed using two primary metrics: The Root Mean Squared Error (RMSE) and the coefficient of determination (R^2), which strengthens its predictive capability. An RMSE value of 0.3844026 underscores the model's high level of accuracy, suggesting that

the model's predictions closely match the actual sale prices in the Liverpool market. The model's forecasts are remarkably reliable, considering the inherent unpredictability and complexity of real estate valuation.

Furthermore, the model obtained a R^2 value of 0.6369208 signifies that approximately 63.69% of the variation in house prices can be accounted for by the model's predictors. This significant explanatory power indicates that the model is highly effective in capturing the complex dynamics of the housing market. This makes it a valuable tool for stakeholders who want to understand or influence the real estate landscape in Liverpool.

The application of the EN model to validation data provided a critical test of its generalizability to precisely predict outcomes. A scatter plot was plotted using the *ggplot2* package with two regression lines, comparing the predicted values with the observed values. The blue line indicates the model fit, while the red line shows the variation in prediction from the observed values. A strong correlation between both highlights that the model generated from the train can predict the output of the test data with high accuracy. It also means there is less probability of overfitting making the model reliable.

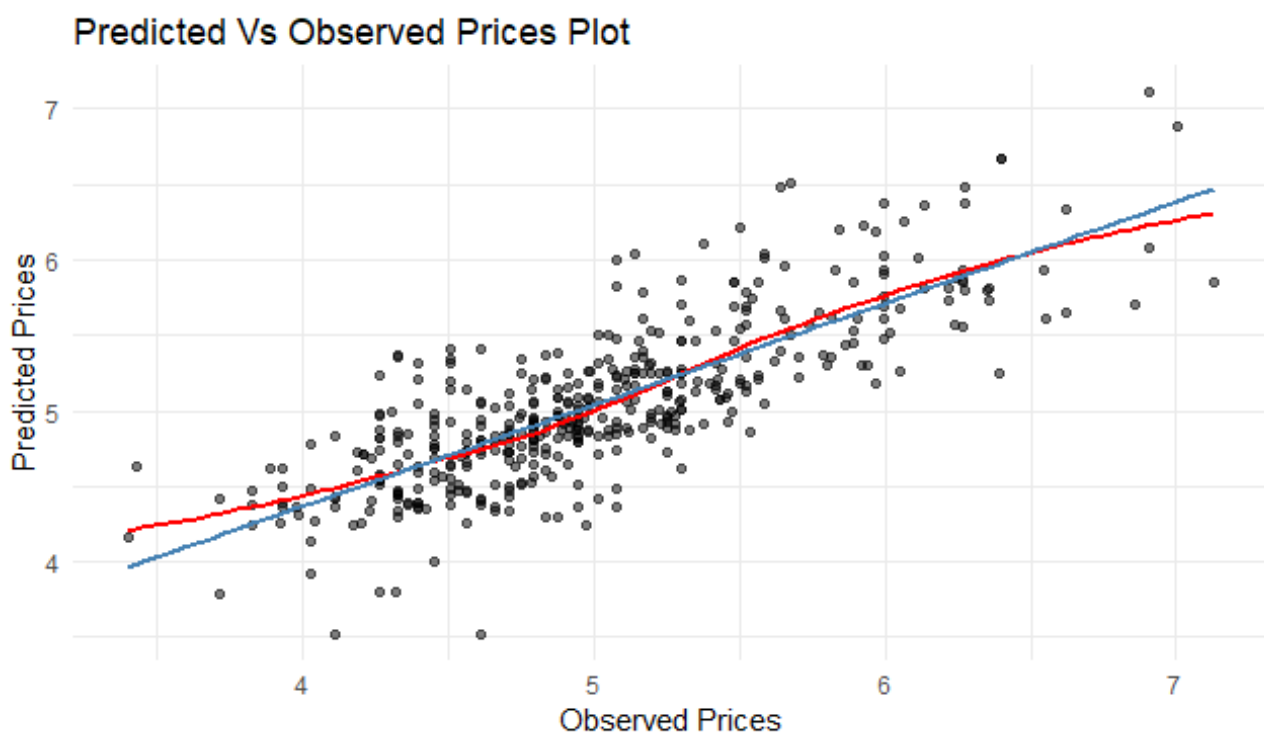


Figure 2: Predicted Vs Observed Scatter Plot

A comprehensive examination of the significance of variables provided additional insight into the factors that influenced housing prices in Liverpool. This analysis, crucial to understanding property values, identified the number of bedrooms ('Beds') as the biggest predictor of house prices. This supports real estate principles that link larger living spaces to higher prices, highlighting the importance of property size and utility in determining market value.

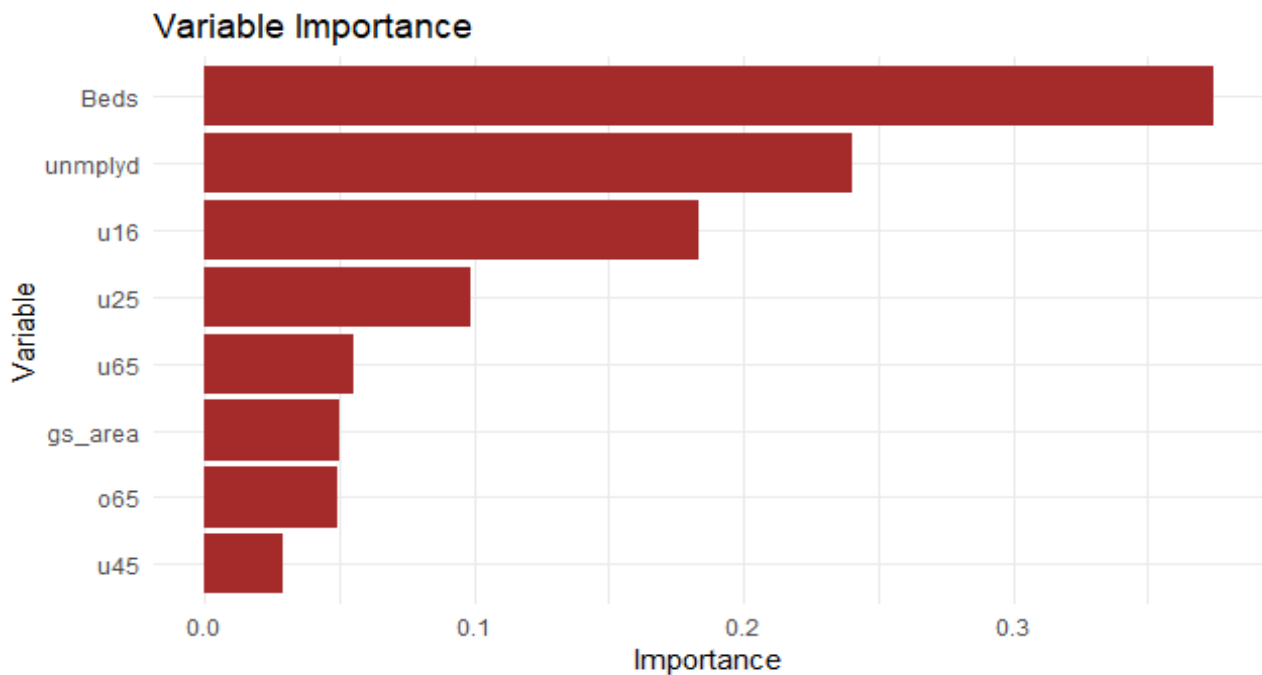


Figure 3: Bar Graph displaying Feature Importance

Moreover, Unemployment ('unmplyd') was another important socioeconomic indicator. A reflection of the intricate relationship between local economic health and property values. Due to reduced buying capacity and increased property vacancies, regions with high unemployment may see property values fall. This shows how external economic factors and property characteristics affect the real estate market wherein property characteristics and economic conditions determine market prices.

The comprehensive analysis supported by careful model tuning and validation, not only confirms the predictive accuracy of the Elastic Net model but also its analytical complexity which helps explain Liverpool house prices. It also serves as a basis for decision making and the development of strategic policies in the real estate sector.

DISCUSSION

The utilization of the Elastic Net model in this study signifies a notable advancement in the predictive analysis of Liverpool's housing market. The EN model has effectively addressed the issue of multicollinearity and performed well in feature selection by combining the advantages of lasso and ridge regression. This is particularly relevant in real estate datasets. The utilization of this hybrid approach has facilitated a comprehensive comprehension of the variables that impact housing prices in Liverpool, thereby offering significant and valuable perspectives for individuals and organizations involved.

The model's performance, as indicated by an RMSE of 0.3844026 and a R^2 of 0.6369208, demonstrates a high level of accuracy in predicting housing price variations and a significant ability to explain these variations. These metrics confirm the model's ability to accurately represent the intricate relationship between socioeconomic and property-specific factors that impact the market. The inclusion of the variables

'Beds' and 'unmplyd' as significant predictors in the model confirms its consistency with both intuitive and theoretical expectations regarding the real estate market.

Although the EN model has its strengths, it also has limitations. The EN approach assumes linear relationships between predictors and the target variable, which may oversimplify the dynamics of real estate markets. In Reality, Nonlinear interactions could play a significant role. Furthermore, although the model effectively deals with multicollinearity, the regularization process might obscure the specific influence of highly correlated predictors, which could restrict the interpretability of their effects on housing prices.

The methodological choices made in this study are supported by existing literature, which acknowledges the Elastic Net's utility in addressing multicollinearity and selecting features within high-dimensional data (Zou and Hastie, 2005). Nevertheless, as highlighted by James et al. (2013) the trade-off between the complexity of a model and its interpretability is an important factor to consider, particularly in fields such as real estate economics where decision-making often depends on nuanced interpretations of data-driven insights.

Future research should focus on incorporating nonlinear modelling methods, such as polynomial regression or machine learning techniques like decision trees and neural networks, to better capture the intricate relationships between predictors. Incorporating dynamic variables, such as fluctuations in interest rates or macroeconomic indicators, could improve the model's predictive accuracy for housing prices in response to broader economic changes. By increasing the dataset to incorporate additional temporal data points, it becomes possible to apply time-series analysis, which can provide valuable information about the cyclical patterns and potential future trajectories of the market.

To summarize, although the Elastic Net model offers a strong framework for forecasting housing prices in Liverpool, it is crucial to recognize its limitations and the assumptions on which its application is based. By leveraging the groundwork established in this study and investigating the potential areas for future research outlined here, subsequent work can further improve and broaden the predictive precision and practicality of the model which will ultimately lead to more informed decision-making in the real estate industry.

REFERENCES

- Zou, H. and Hastie, T., (2005). *Regularization and variable selection via the elastic net*.
- Osborne, J. (2010). *Improving your data transformations: Applying the Box-Cox transformation*.
- Dalpiaz, D., (2020). *R for Statistical Learning*.
- Comber, L., (2023). *GEOG5917 Big Data & Consumer Analytics - RStudio Practicals*.
- Lin, H. and Li, M., (2023). *Practitioner's Guide to Data Science*.
- Comber, L. and Brunsdon, C., (2021). *Geographical Data Science and Spatial Data Analysis: An Introduction in R*
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*.
- Lavelle Estates, (2023). *The Liverpool Property Market: Trends and Analysis for 2023*.