Universität St. Gallen

Data Science Fundamentals

Predicting Rent Prices in Lagos, Nigeria

Fall Semester 2022

Group 7:
Gabriel-Maximilian Bergmann, Vincent Klaer, Lukas Lehrecke, Laurenz Schneeberger

Project Research Report
Universität St. Gallen
Data Science Fundamentals
Prof. Dr. Johannes Binswanger
14. December 2021

Abstract

This project aims to predict annual rent prices in Lagos, Nigeria based on various dwelling attributes. First, we perform preliminary data analysis and pre-process it by filtering aggressively to make it usable for statistical learning. Then, we employ GeoPandas and natural language processing to visualize the data's geographical components. Using the cleaned data, we construct various regression models and a random forest. The efficacy of our models is assessed primarily via their R^2 score. The random forest shows the most predictive capacity, explaining 56% of the dataset's variance.

1 Dataset

The dataset employed describes annual rent prices as of 2022 for various real estate objects in Nigeria and contains information about each object's location, servicing status, construction date, furnishing, number of bedrooms, bathrooms, and sanitary outfittings. The data was originally extracted from the Nigerian Real Estate platform PropertyPro NG and is available via Kaggle, however has 0 code submissions and 10 overall downloads (Pinnick, 2022).

1.1 Pre-processing of Non-Geographical Data

For our purposes, this dataset is not as clean as would be desirable; a four-step cleaning process allows the reduction of 98'080 observations to 4'730 highly usable samples. First, non-residential properties were removed by filtering the 'Description' column to exclude observations with keywords such as 'Land', 'Working', 'Office', or 'Joint Venture'. Second, to enhance the predictive capability of the multivariate models elaborated upon later, the 'Description' column was filtered once more to extract features such as 'Duplex', 'Detached', or 'Apartment' for each observation. Third, rampant price dispersion in the dataset suggests incorrect entries that could lead to noisy or badly fitted predictions. Outliers were then cleaned using a low pass filter at the third quartile plus 1.5 times the interquartile range. Using a similar method for low values does not work, as the high pass filter would be negative. To nonetheless remove nonsensically low outliers, we make a judgment call and remove objects with a rent under 20'000 Naira.

1.2 Pre-processing of Geographical Data

The fourth and final step in the cleaning process consists of decomposing the 'Area' column (which is given as verbal data in non-standardized form) to standardized states, areas, and micro-locations. This project focuses on Lagos, so we remove the few remaining entries not attributable to the city. As the data was extracted from a brokerage platform, the micro-location description, if available, is highly inconsistent and often inconclusive, rendering Python-package-based coordinate extraction for each asset near-impossible. Instead, we resort to less granular location data, which we call 'Area'. Due to the lack of fitting Geodata and the limited geographical representation given by singular coordinates, a custom Polygon Shapefile was created in QGIS that plots the areas described in the dataset. Using GeoPandas the centroid coordinates of each area were extracted for regression purposes while the area Polygon data was used for plotting purposes.

However, there are limitations to the usability of the location data; Fig. 1's top map gives us

a picture of a rental price disparity which might be expected in a developing country with very few select locations achieving high prices, while the bottom map shows the uneven distribution of observations. Many areas do not exceed 10 observations while Lekki, the large peninsula in the south-east contains over 2000 observations of which we expect most to be in the western part of Lekki. According to this, visualisation and predictions relating to location should be used and interpreted with caution as the lack of data for some areas could lead to unrealistic representations of those areas in our maps and predictions and achieved prediction scores would be heavily weighted towards the Lekki area.

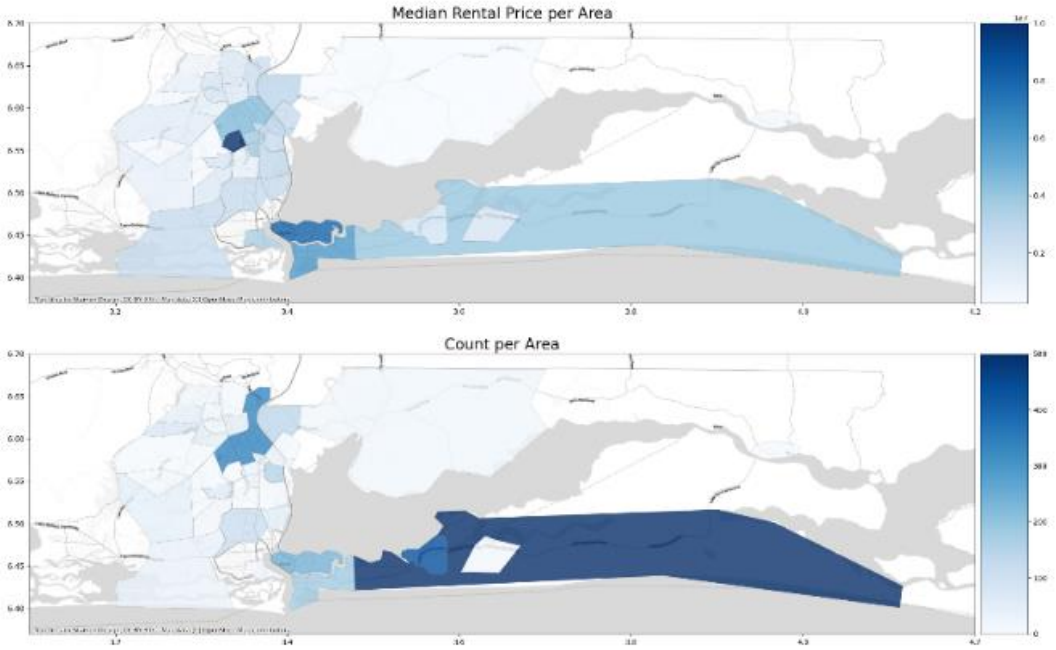

Figure 1: Exploratory data analysis by geographical area

2 Prediction Models

2.1 Regression Model

This regression analysis omits location as a feature temporarily and then adds it in later to show how starkly the model's performance differs when location data is incorporated. This series of regression models consists of four regressions which are trialled on the data: an OLS regression, a lasso regression, and a ridge regression. We find the ridge to outperform consistently due to noticeable multicollinearity in the dataset (Kidwell, Harrington-Brown, 1982).

To test the predictive force of each individual feature, we run naive (non-tuned) trials of all four regression types per feature. Non-binary variables (i.e., 'bedrooms', 'bathrooms' and 'toilets') outperform. This is somewhat intuitive due to the higher fidelity with which they can describe the real estate object. The notable exception of 'Duplex' predicts quite well in the scheme of binary variables at an R^2 of 8%.

The most accurate naive regressions used the 'Bedrooms' feature to fit the model. 10-fold cross validation demonstrates a mean R^2 score of 20% with a standard deviation of 4.4%. These values are near-identical for the OLS, lasso, and ridge regressions, while the elastic net underperforms mildly it outperforms noticeably at higher cross-validation folds and shows lower variance of predictive capacity across all fold counts.

A multivariate OLS regression using all above features as infeed allows the model to improve on the previously best R^2 score, lifting it to 20.3% (at the cost of a higher dispersion of σ=7.1%). However, due to multicollinearity in the dataset, a ridge regression seems to perform better at R^2 21.36%. This is achieved after tuning the iteration and solving hyperparameters, where we find regularized least squares to be most effective. Going a step further, we test polynomial variants of the discussed multivariate regressions and attain the best results, i.e., an R^2 of 25.67%, with a second-degree polynomial ridge regression.

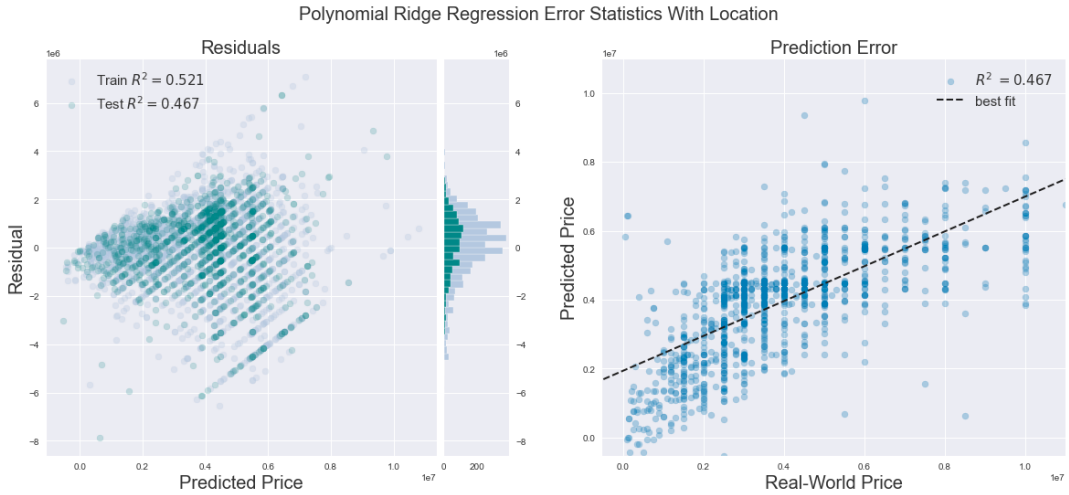

Figure 2: Polynomial Ridge Regression error statistics incl. area

Adding in the 'Area' feature as a predictive factor to this same regression drastically increases the accuracy of the model to a mean R^2 of 49.79%, which translates to a performance increase of 23.67%, almost doubling the predictive capacity of the model.

2.2 Univariate Geographical Regressions

While producing good results, the above models leave an important question open: what if one were to predict based on solely geographical features? Two univariate regressions using the distance from the city centre and the 'Area' feature, respectively, were performed.

To estimate rent prices based on the distance of the object to the city center we use GeoPandas to approximate the distance between the city center of Lagos and the individual locations.

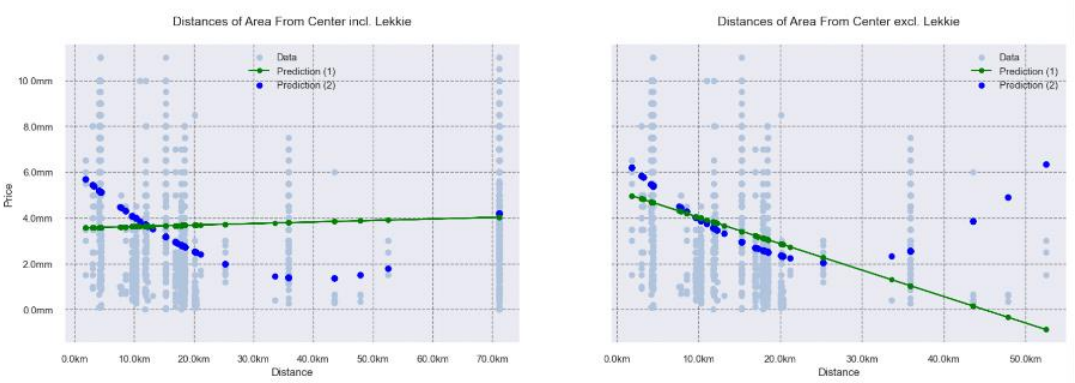

Figure 3: Univariate geographical regression

The two regressions (Prediction 1, 2) seen in Figure 3 differ in that the former is linearly fitted and uses polynomials of two. The lower-order regression is a very poor fit at an R^2 of 0.0077. On the other hand, if we increase the polynomials, the model can establish a better link between the independent and dependent variables (R^2 of 0.159). To improve these geographical predictions, the 'Lekki' Area was excluded in a second run of the model (see the right plot), thus excluding the dominant data cluster at 70km. The updated model results in a significant change in predictive capacity for the linear and polynomial regression.

As the polynomial regression performed better, we used 20-fold cross-validation to iterate through 20 different regressions, each with successively increasing polynomials. It was found that a regression with polynomials of 12 had the best R^2 score ($R^2 = 0.37$), while a regression with polynomials of 9 turned out to have the lowest mean absolute error (MAE = ₦1,5mm).

An analogue experiment was performed while using the 'Area' variable as the geographical feature. First, a 20-fold cross-validation process calculated R^2 scores and mean absolute error for 40 different regressions. Consistent with prior results, a regression with polynomials of 12 has the best R^2 score ($R^2 = 0.41$), while a regression of 9 polynomials has the lowest mean absolute error (MAE = ₦1,5mm). We found that the area predicts house prices exceptionally well on both the training and testing data at an R^2 of 0.40 (12 polynomials) and 0.32 (9 polynomials) for testing data.

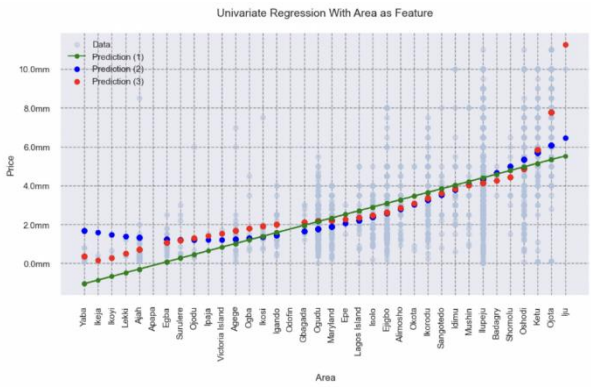

Figure 4: Univariate Regression with area as feature

2.3 Random Forest Regression

For the last model, we develop a Random Forest. This type of model is particularly useful for our data as it is by nature multivariate, taking all variables into consideration. As a random forest can be a regressor or classifier, we opt for the former (we are predicting continuous random variables). In contrast to a random forest classifier, a regression uses variance reduction instead of entropy in order the measure the level of impurity on each leaf. This makes it much more suitable for this dataset, as will be demonstrated.

2.3.1 Hyperparameter Tuning

A mean R^2 score of 0.538 can be attained using default parameters. While random forests are among the least tuneable learning algorithms, we believe there are still some moderate performance gains to be made (Probst et al., 2018). Therefore, we investigate a range of hyperparameters and two methods of calculating their ideal values, namely a Grid Search and a Bayesian Optimization.

2.3.1.1 Grid Search

This method only takes discrete parameters as inputs and tests all possible combinations to minimize a predetermined scoring measure. Due to resource limitations, we limit ourselves to three parameters thought to show the greatest optimization. This method yielded only slight increases in predictive capacity. A caveat is the unavailability of cross-validation to verify the performance over a greater number of trials due to the computational limits cited prior.

2.3.1.2 Bayesian Optimization

While the Grid Search method takes a more brute-force approach, Bayesian Optimization solves this task through a method called surrogate optimization. A surrogate function tries to estimate the points of interest for the optimization function of our model. Through each iteration, the current surrogate function learns more about the areas of interest by sampling them and subsequently updates itself. The more iterations it goes through the closer to the global minimum, reaching an even higher R^2 score than the Grid Search Method and thereby again showing a slight improvement. Cross-validation was not possible, resulting in different scores through each iteration. (Ye, 2020)

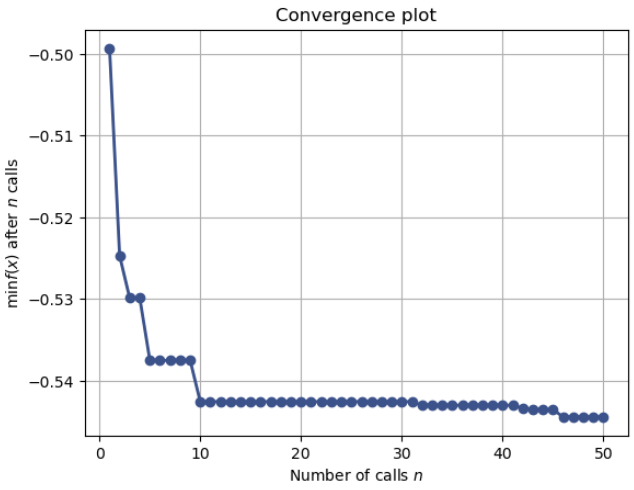

Figure 5: Process or convergence to a specific R2 score, over each iteration of the model

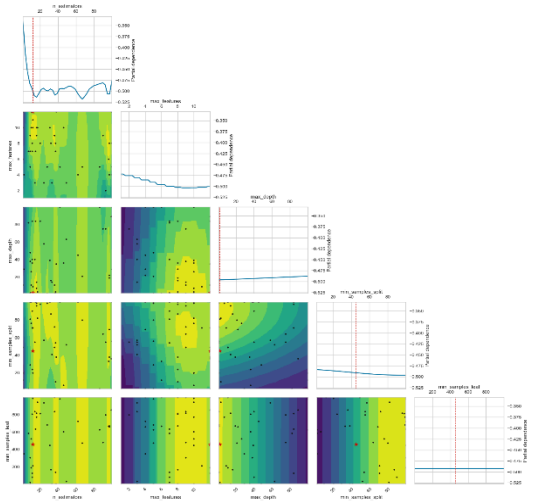

Figure 6: Dependence function between all the hyperparameters

2.3.2 Features Importance Analysis

The Features Importance process permits insight into the potency of unique features for statistical learning. It is demonstrated that by far the greatest predictor of rent prices was the location as well as the number of bedrooms. This corresponds to previous results obtained from our other regression models.

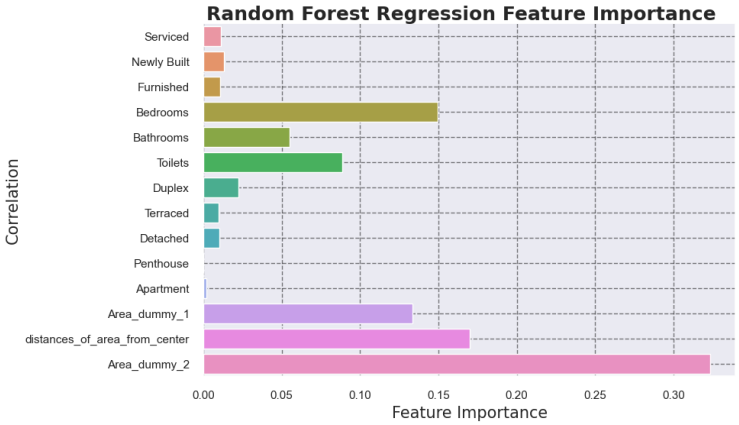

Figure 7: Features Importance

3 Project Evaluation

Among various models, the Random Forest Regressor emerges as the most potent model with an R^2 score of 0.57. There are three caveats with this result. Firstly, as cited prior, the there is a possibility of slight inaccuracies in the model metrics due to computational limitations that prohibited us from employing cross-validation. Secondly, the range of features may have caused slight overfitting, as evidenced by a moderate disparity of the training and testing R^2 scores. Nonetheless, the model predicts 57% of the datasets variance, which is a statistically significant result. Even with this statistically relatively successful result, as of now, the practical application of these models seems to be limited by the uneven dispersion of data by location as a driving variable in the predictions.

Bibliography

Kidwell, J., Harrington-Brown, L. (1982). Ridge Regression as a Technique for Analyzing Models with Multicollinearity. *Journal of Marriage and Family*, 44(2), 287-299.

Pinnick, E. (2022). *Nigeria Rent Prices (2022).* Kaggle. https://www.kaggle.com/datasets/eyimofeapinnick/nigeria-rent-prices-2022

Probst, P., Boulesteix, A. L., & Bischl, B. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, 20. https://doi.org/10.48550/arxiv.1802.09596

Ye, A. (2020). *The Beauty of Bayesian Optimization, Explained in Simple Terms.* Towards Data Science. Retrieved December 2, 2022, from https://towardsdatascience.com/the-beauty-of-bayesian-optimization-explained-in-simple-terms-81f3ee13b10f