Understanding and Predicting Price-Per-Square-Meter values for London Properties

Understanding real estate prices is a primordial tool in providing truthful information to real estate agencies, taxation on residency properties or determining correct prices for the purchase or selling of a property (Kostic and Jevremovic, 2021; Das et al, 2021). This was previously done by real estate agents, using their professional experience and basic information such as area, neighbourhood, bathrooms, and bedrooms to make their appraisals. Technological improvements and the implementation of self-learning mechanisms, such as Machine Learning (ML) and Deep Learning (DL), have benefited social science and geographical research (Manasa et al, 2020). With the increasing availability of data, this provides an objective framework in which to analyse house prices and their fluctuations.

Our article will focus specifically on Price-per-square-Meter (PPSM) predictions rather than the traditional total area of a property, as we believe it to better represent the current market. Research in Spain (Monterro and Larraz, 2010) and England (Chi et al., 2021) has shown that analysing PPSM provides a more granular understanding of spatial variation in house prices. Understanding the distributions of PPSM is important as it allows to understand how geographical inequalities and can help policy makers implement either taxation or support, appropriate to actual price.

Our dataset, comprised of 49,282 observations was taken from Chi et al's 2021 study, with three additional variables: distance from schools, tube stations and city centre. Five models will be discussed in our results, with the most promising being XGBoost, a regularised version of Gradient boosting with better generalisation abilities (Peng et al., 2021) and accounts for spatial correlation structures (Vijanen et al, 2021). We present our finding with tables and Geographical visualisations but also acknowledge limitations and propose recommendations for future research stemming from our research, such as greater temporal overview instead of the 2020 values analysed here.

Materials and methods

Data Wrangling

The initial London Housing subset collected by Chi et al. was comprised of two files: one with 17 variables on property information and another with 3 variables with eastings, northings and post codes. We join both files to create a data frame with 20 variables and 60,569 observations, all observations from 2022. We then proceed to look at the data more closely and to remove and points we believe could hinder out analysis and the performance of our upcoming models. Looking at *'year'* and *'date of transfer'* variables, we realise that all transactions took place in 2022, therefore we decide to remove both variables.

We begin to look at the primary response variable we select for this work: 'price', aka. actual price paid for the property. We identify 1 property sold for a symbolic 1£, which we exclude. At the other extreme, we find that only 491 properties are recorded above 3,550,000£ which we also remove as they are not representative of the overall dataset, less than 1% of the total. Properties in our transformed data will range between £60,000 and £3,5m. We also remove 109 duplicate transaction ID's.

We follow a similar methodology for all the following numeric variables. Indeed, we remove outliers at both extremes as they do not accurately represent real characteristics of London properties sold in 2020.

For the number of rooms, we find that only 254 properties have more than 11, therefore we decide to remove them and cap the properties characteristics to 11 rooms. For the *'construction age band'*, 104 values are recorded to have invalid information which are removed. For *'current energy*

efficiency' we remove 8 outlying values with a rating above 85 and 317 below a rating of 22. These removals are to par with current consumption metrics, the median being 65 to 67 for England and Wales (ONS, 2022). We decide to keep the variables *'potential energy efficiency'* as a t-test confirmed significant differences in mean with the current energy efficiency measures, therefore adding input for final model construction. Regarding floor area, we removed 102 properties over 310m^2 and 3 below 10m^2.

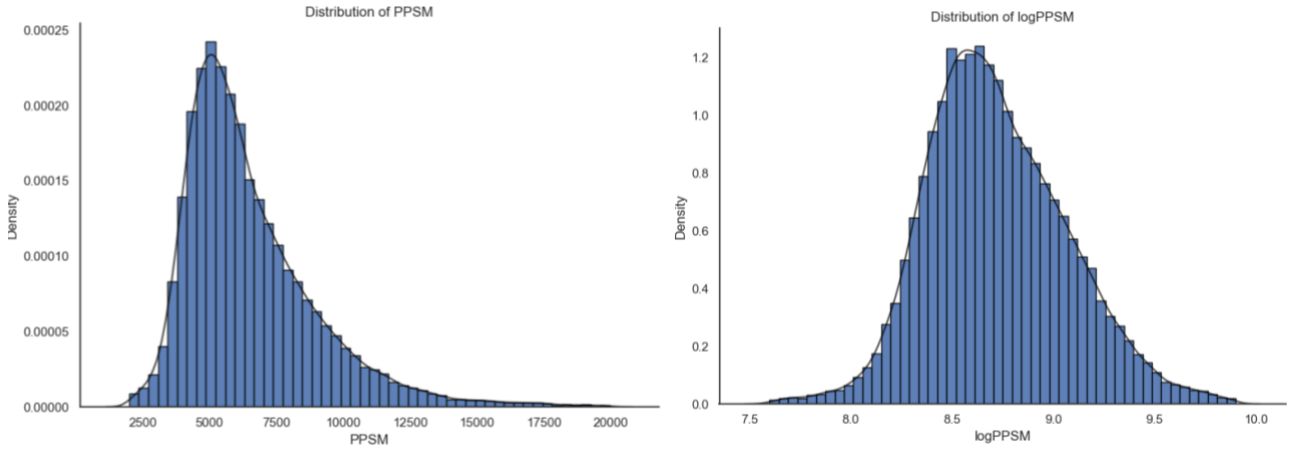
We proceed to create a new response variable for the Price per Square Meter (PPSM). This was created by taking the price of each property over their floor area. Following the same outlier removal reasoning, we took out 197 properties with a PPSM over £20,000 and 156 under £1995.

We also transform the duration variable into a binary one, with 0 as leasehold and 1 as freehold. A t-test comparing both was made and we found significantly different means for price and PPSM between the two outcomes, so we decide to keep the variable for the analysis. On the other hand, we also transformed the address matching ratio to a binary variable (1 for 1:1 address matching and 0 for 1:n) but decided not to keep this in our final model as the t-test gave an unsignificant difference in means.

Feature Engineering

We decide to log transform PPSM to normally distribute the results. Figure 1 illustrates the benefit of this transformation, the response variable LogPPSM is now normally distributed.

Figure 1: log transformation of the response variable PPSM



Our model includes three more created variables providing the distance for each property to the closest school, tube station and city centre.

The distance from school variable was created using a dataset made publicly available with 3889 Schools in Greater London. No changes were necessary to the datapoints in the file as it was already London Specific, therefore we calculated the shortest distance from each property to the closest school. We acknowledge that we consider Euclidean distance and not actual walking distance[1] but serve as a reliable indicator for School Proximity. Research indicates that school proximity influences house price (Bonilla-Mejía et al., 2020; Sah et al., 2016), and we are interested in applying this framework to London. With our *'closest school'* variable created, we remove 130 outliers, properties more than 1.7km away from the closest school.

[1] For further research we propose using Open Street Maps to give actual walking distances, accounting for obstacles making this distance greater than the Euclidean one used

The same methodology was applied when creating the *'closest tube'* variable. The tube station data was from the 2011 travel to work census data and made available in week 4 of GEOG0115. 291 properties were removed when they were further the 22km away from a tube stop. Map 1 illustrates major gaps in the tube system in areas like south-east London.

Finally, we add a *'distance from centre'*, providing a value for all the properties in the dataset. We took the Charles 1st south statue in Trafalgar square as this is kilometre 0, where all distances when measuring from London originate from. This variable has no outliers and normally distributed so no values were removed.

Map 1:

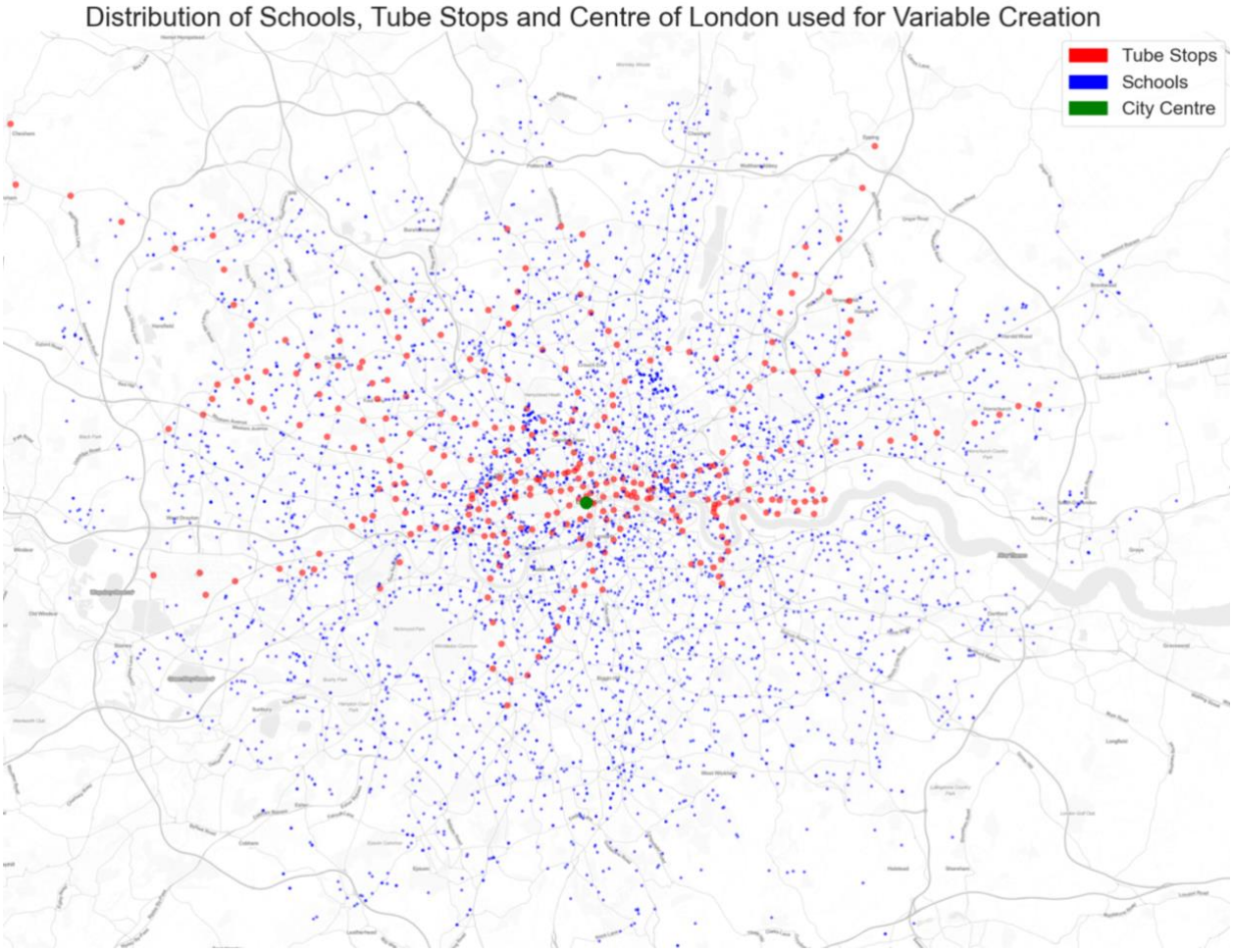
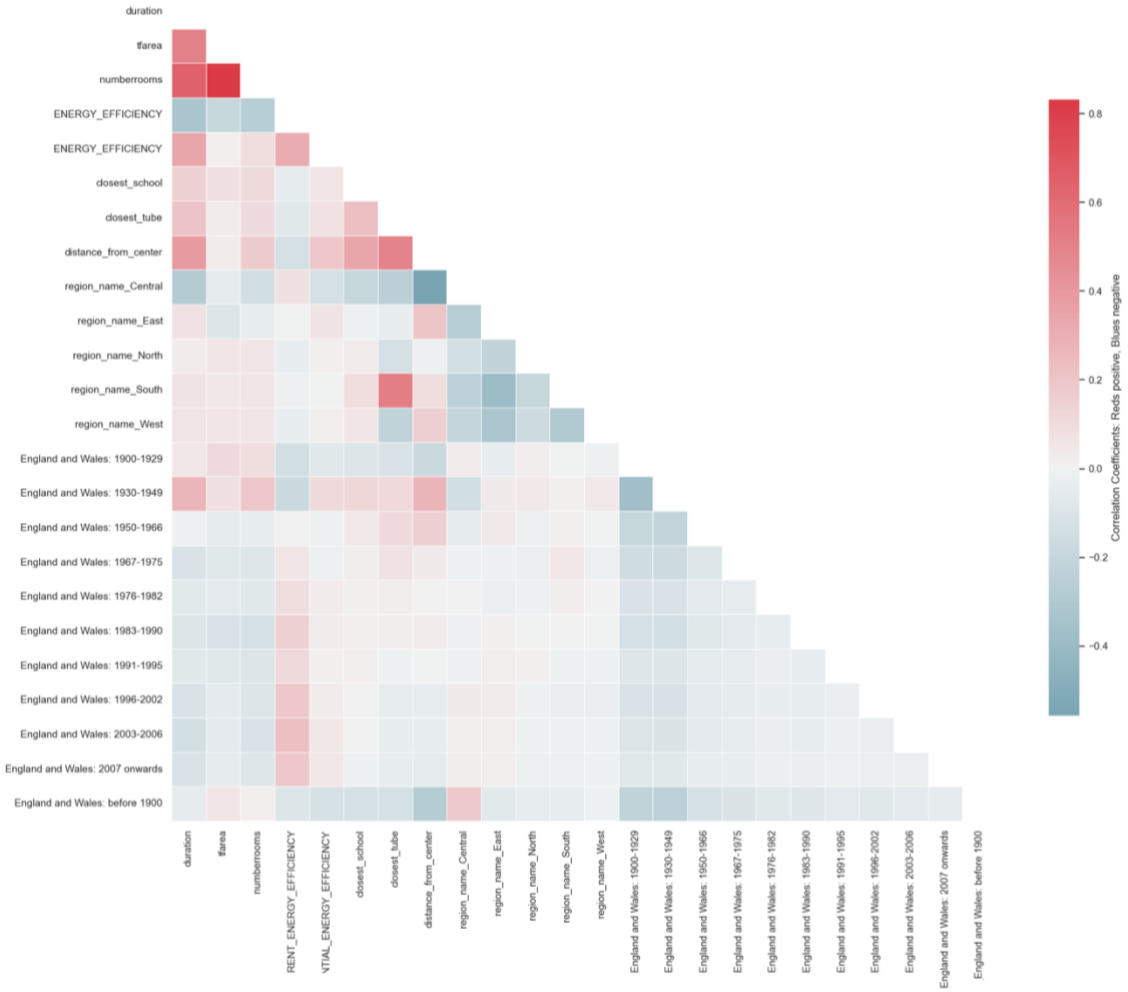


Figure 2 provides an illustration of the correlation between the variables selected for our model creation. Much of the variables have a light-grey relationship, meaning that there is little to no correlation between them. We see that strong correlations exist between the number of rooms and the total floor area (*'tfarea'*) which is unsurprising as rooms, by definition, take up floor space in a property. We decide to keep it as it is closely related to the PPSM response variable. We also recognise correlation between tube proximity and southern London, simply because more stations are in that area, illustrated in Map 1. We decide to keep both variables.

We also decide to create a new variable *'Area name'* instead of the *'Region name'*. Indeed, the latter had 33 districts, a separation we deemed too granular. Instead, we aggregate this into 5 main areas, described in the 'London Plan', an initiative led by the Greater London Authority. These 5 regions (North, East, South, West and Central) each have between 4885 and 14760 observations.

Finally, we included categorical variables in our analysis such as the recently created *'Area name'* and *'construction date'*. The latter has 11 bands from before 1900 to after 2007, all with over 700 observations. To add them to our model, we create 16 dummy-coded variables, providing binary outputs for each categorical output (c.f. Figure 2 showing how each categorical variable was broken down).

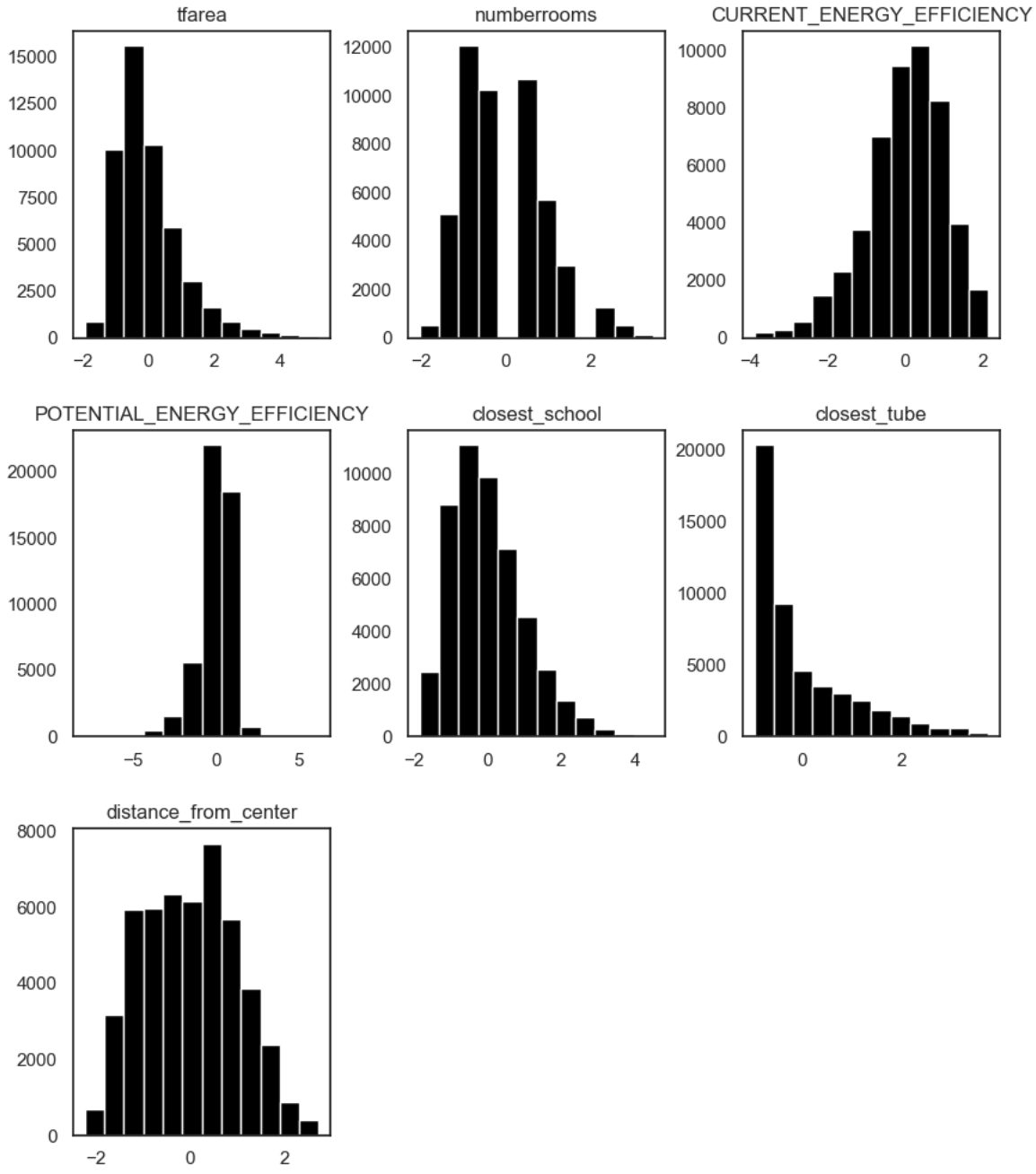
Figure 2: Correlation Heatmap between variables selected for model creation



After data cleaning, our final dataset contains 49,282 observations AND 24 variables, and our response variable, LogPPSM also having 49,282 variables, compared to the raw dataset which had 60,569 and 20 variables.

We scale the variables, so that no variable is over or under-represented in the results. We use a standard scaler from the *sklearn* package in Python, which make all variables have a standard deviation of 1 and mean of 0. We decide not to scale the binary and dummy-coded variables as they only have two values (0 is absence of phenomena and 1 is presence) and is not necessary. Figure 3 illustrates that our numeric non-binary variables are normally distributed for the most part. We see that values relating to tube proximity, which is unsurprising as city planning has it so that London tube is easily accessible to the greatest amount of people, which explains it's skewness. We accept the overall distribution of the values and will not apply further changes.

Figure 3: Variable overview (after scaling)



Results:

Now that we are satisfied with our data, we select four models to predict PPSM in London properties: Basic Linear Model (LM) as baseline model, Lasso Regression, Gradient Boosting (GB) and XGBoost (XGBT).

The baseline model (LM) provides an initial prediction framework. Indeed, this initial model gives a good overview of the London property market, with the model in our variable explaining over 46% of the total variation. The Lasso regression is a continuation of the LM, with the difference that it penalises on the sum of absolute values of the weights, meaning it doesn't give a specific variable disproportionate importance. There is little to no difference in the results, with less than 0.01% in the R^2, RMSE and MSE.

We take another approach and use Gradient Boosting. GB has the advantage of automatically detecting non-linear feature interactions and providing fast predictions, although slow to train (Prettenhofer and Louppe, 2014). This provides considerable improvement in the Test results,

especially with an R^2 increase to 62.90. We then decide to use XGBoost, a regularised version of the Gradient Boosting which has better generalisation abilities and prediction robustness (Peng, Huang and Han, 2019). This model provides slight improvements in the results, with the R^2 reaching 63.00.

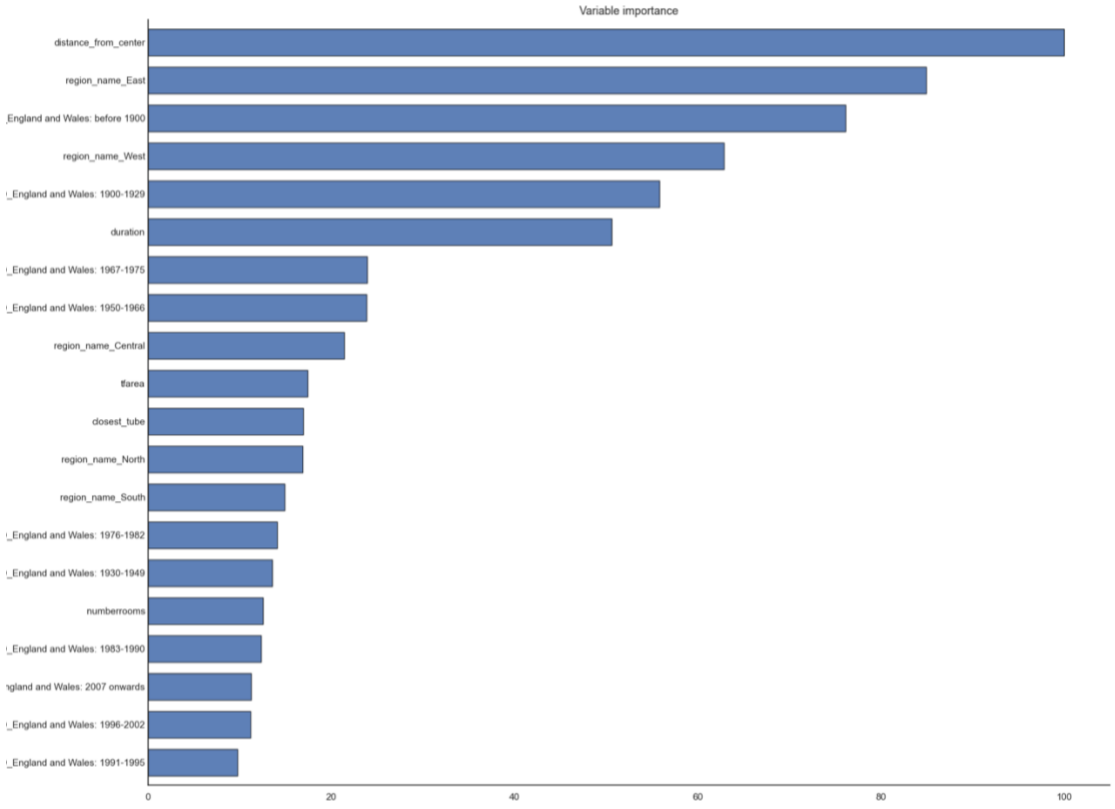
We tuned the lower three models in figure 1 using *'RandomizedSearchCV'* from the *sklearn* package in python, which allows the system to find the best performing hyperparameters from a list we provide. GB and XGBoost had the same tuning parameters, with a *'learning rate'*, *'number of estimators'*, *'max depth'* and *'subsample'* (respectively 0.05, 1500, 4 and 0.8).

Table 1: Model Results for prediction on Test Sets[2]

Model Name	**Test RMSE**	**Test R2**	**Test MSE**
Linear Regression	0.2524	0.4768	0.1928
Lasso Regression	0.2524	0.4768	0.1928
Gradient Boosting	0.2102	0.6371	0.1549
XGBoost	0.2099	0.6382	0.1546
Random Forest	0.2104	0.6365	0.1539

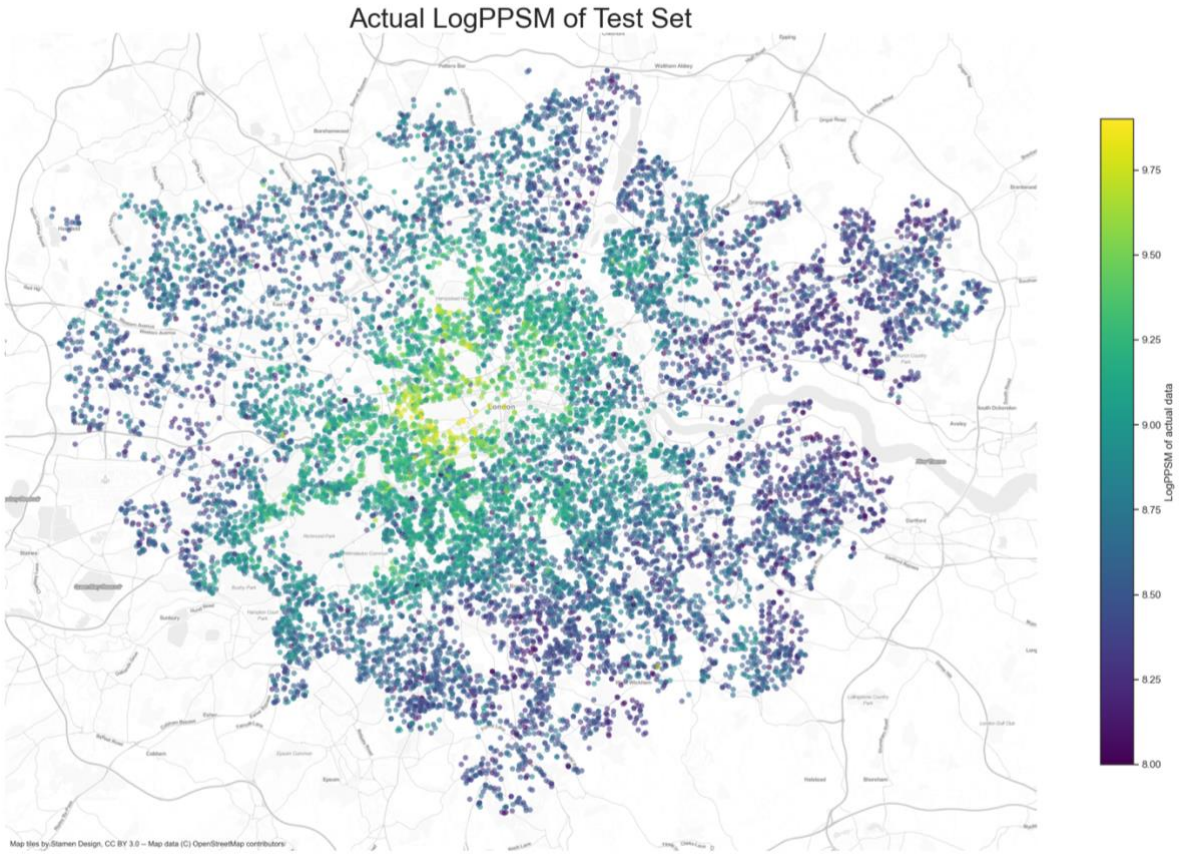
Maps 1 and 2 provide geographical output of the Logged PPSM both for actual values and XGBoost predicted values in the test set. Figure 4 shows that distance from centre, east London and old buildings are the most important factors when predicting PPSM in London. By presenting these factors, real estate agents and potential buyers or sellers, will have a better understanding of what factors are important. The maps serve as a beneficial was to understand how the results are distributed

Figure 4: Variable importance in XGBoost

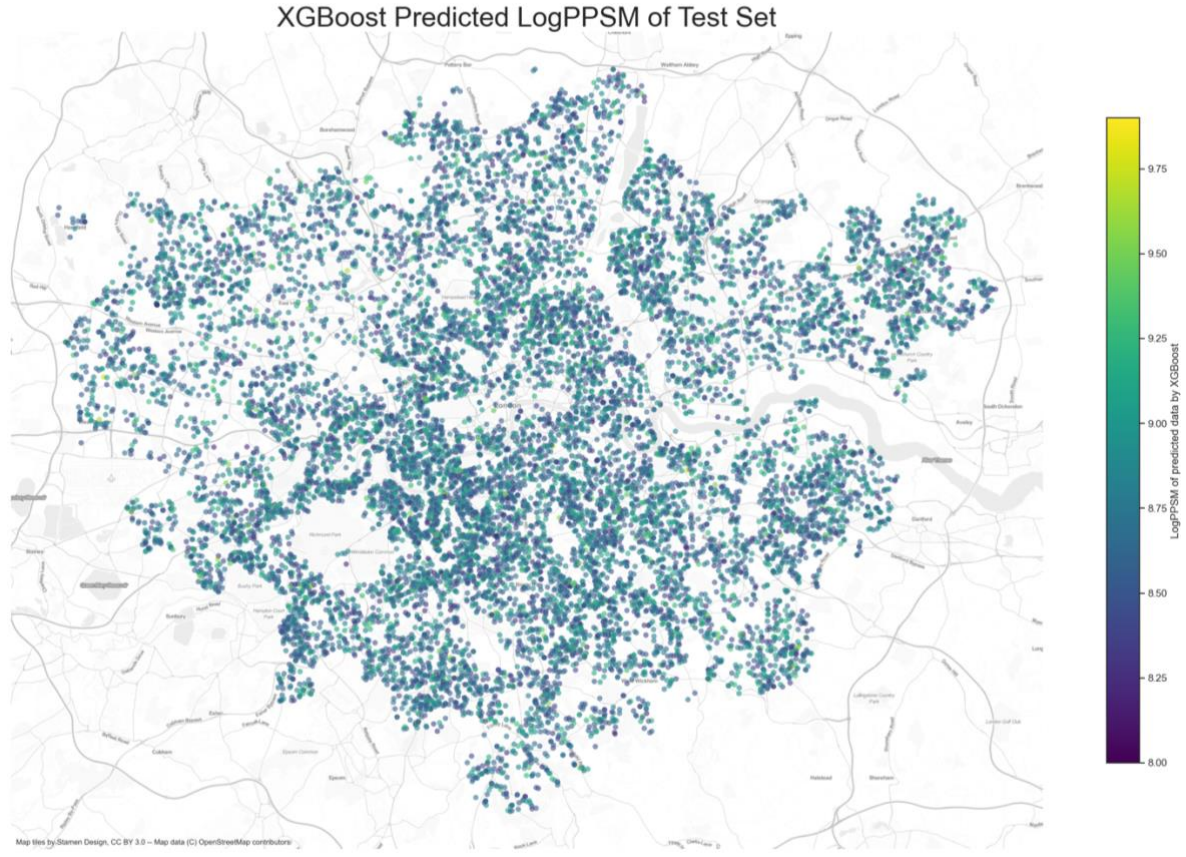


[2] Results done on set seed, will be variations when run on different seed but improvement remains regardless

Map 1:


Actual LogPPSM of Test Set
9.75
9.50
9.25
9.00
8.75
8.50
8.25
8.00
LogPPSM of actual data
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors

Map 2:


XGBoost Predicted LogPPSM of Test Set
9.75
9.50
9.25
9.00
8.75
8.50
8.25
8.00
LogPPSM of predicted data by XGBoost
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors

Limitations and Further Research:

Further improvement could be made on both with data precision and methodological improvements. For instance, we would look to add weights dependent on the ratings of the schools, as proximity to quality education is an important factor in family decision when purchasing houses (Sah, Conroy and Narwold, 2016). We could also use seasonality analysis through the *'date of transfer'* variable was of interest, but we decided not to include it our study as only 2020 data is used. Potential indications of seasonal variations (Rossini, 2000), and we believe such a framework would be of interest in London.

Furthermore, we acknowledge that the 2020 data may not be accurately representative of the London real estate market due to government imposed COVID restrictions and the potential preference changes of the buyers (Valaskova et al., 2021; Hoesli and Malle, 2021)

We also recognise potential Gradient boosting and XGBoost potential overfitting issues. Table 2 illustrates the results in the training set and the R^2 vales for GB and XGBoost are considerably higher than in table 1, indications of overfitting. The is flagrant for the Random Forest, with its model explanation dropping by nearly 30%. To overcome this in future research, we propose technical improvements by adding Bayesian hyperparameters when using the gradient boost and XGBoost model. Bayesian hyperparameters allow for better understanding of black-box optimisation and allow optimisers to be identified without human intervention (Turner et al., 2021)

Table 2: Model Results for prediction on Training Sets

Model Name	**Test RMSE**	**Test R2**	**Test MSE**
Linear Regression	0.2533	0.4665	0.1933
Lasso Regression	0.2533	0.4665	0.1933
Gradient Boosting	0.1811	0.7273	0.1355
XGBoost	0.1841	0.7183	0.1365
Random Forest	0.0784	0.9490	0.0572

Finally, we also acknowledge that we could use our data to focus on understanding and predicting energy efficiency as an outcome due to its importance in total energy use and global green-house emissions (Fathi et al, 2020).

Conclusion:

Our work has detailed the workings from cleaning the dataset to model selection and outputs to prediction PPSM for properties in London. After cleaning of our dataset, reducing the initial 60,569 observations to 49,282, we also dummy-coded categorical variables and added 3 of our own: distance to closest school, distance to closest tube and distance to centre. This leaves us with 24 final variables. We ran five different models of which XGBoost was the most successful of our models with an R^2 of 0.638, considerable increase for the baseline LM at 0.477. Our findings indicate that the most important factors in PPSM predictions are distance from the centre and pre-1900 properties located in east London. Our geographical outputs allow us to visualise both actual PPSM distribution in London and the predicted results from XGBoost. These findings are important as they allow a granular understanding of price distribution around London and provide insight for real estate professionals, potential sellers or buyers, and policy makers when defining taxation. Although results are promising, we acknowledge limitations and propose using data over several years, instead of just 2020, and reliable post-covid property data to investigate the effects of the pandemic on the real estate market.

References and Bibliography:

Bijnens, E.M., Derom, C., Thiery, E., Weyers, S. and Nawrot, T.S., 2020. Residential green space and child intelligence and behavior across urban, suburban, and rural areas in Belgium: A longitudinal birth cohort study of twins. *PLoS medicine*, *17*(8), p.e1003213.

Bonilla-Mejía, L., Lopez, E. and McMillen, D., 2020. House prices and school choice: Evidence from Chicago's magnet schools' proximity lottery. *Journal of Regional Science*, *60*(1), pp.33-55.

Bowers, N., Smith, C., and Wilkins, T., (ONS), 2022, "Energy Efficiency of Housing in England and Wales: 2022." *Energy Efficiency of Housing in England and Wales - Office for National Statistics*, ONS, https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/energyefficiencyofhousinginenglandandwales/2022

Chi, B., Dennett, A., Oléron-Evans, T. and Morphet, R., 2021. Shedding new light on residential property price variation in England: A multi-scale exploration. *Environment and Planning B: Urban Analytics and City Science*, *48*(7), pp.1895-1911.

Das, S.S.S., Ali, M.E., Li, Y.F., Kang, Y.B. and Sellis, T., 2021. Boosting house price predictions using geo-spatial network embedding. *Data Mining and Knowledge Discovery*, *35*(6), pp.2221-2250.

Fathi, S., Srinivasan, R., Fenner, A. and Fathi, S., 2020. Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews*, *133*, p.110287.

Hoesli, M. and Malle, R., 2021. Commercial real estate prices and COVID-19. *Journal of European Real Estate Research*.

Kostic, Z. and Jevremovic, A., 2020. What image features boost housing market predictions?. *IEEE Transactions on Multimedia*, *22*(7), pp.1904-1916.

Manasa, J., Gupta, R. and Narahari, N.S., 2020, March. Machine learning based predicting house prices using regression techniques. In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)* (pp. 624-630). IEEE.

Montero, J.M. and Larraz, B., 2010. Estimating housing prices: a proposal with spatially correlated data. *International Advances in Economic Research*, *16*(1), pp.39-51.

Peng, Z., Huang, Q. and Han, Y., 2019, October. Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm. In *2019 ieee 11th international conference on advanced infocomm technology (icait)* (pp. 168-172). IEEE.

Prettenhofer, P. and Louppe, G., 2014, February. Gradient boosted regression trees in scikit-learn. In *PyData 2014*.

Rossini, P., 2000. *Estimating the Seasonal Effects of Residential Property Markets-A Case Study of Adelaide* (Doctoral dissertation, Pacific Rim Real Estate Society).

Sah, V., Conroy, S.J. and Narwold, A., 2016. Estimating school proximity effects on housing prices: The importance of robust spatial controls in hedonic estimations. *The Journal of Real Estate Finance and Economics*, *53*(1), pp.50-76.

Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z. and Guyon, I., 2021, August. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track* (pp. 3-26). PMLR.

Valaskova, K., Durana, P. and Adamko, P., 2021. Changes in consumers' purchase patterns as a consequence of the COVID-19 pandemic. *Mathematics*, *9*(15), p.1788.

Viljanen, M., Meijerink, L., Zwakhals, L. and van de Kassteele, J., 2022. A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, *21*(1), pp.1-18.