Predicting House Prices using Daft.ie Dataset

Student. no	Contributors	Contribution
A20480484	Robert Sparks, robert.sparks2@mail.dcu.ie	Problem framing, Data Description, EDA &Preprocessing, Simple Linear Model, Report
A00090007	Daniel Mendes, daniel.mendes2@mail.dcu.ie	Preprocessing, Hyper parameter tuning, Cross validation, Imputation, Report
A00051216	Tharakesh Aravindan Suresh Thanuja, tharakesharavindan.sureshthanuja2@mail.dcu.ie	Error Analysis & Insights, Limitations & Future Work, Ethics & Data Protection, Report.
A00053336	Siddharth Sridhar Bavale, siddharth.sridharbavale2@mail.dcu.ie	Research references, Completeness check, Conclusion, Report writing support

Link to repository (https://gitlab.computing.dcu.ie/sparksr2/stats_group_assignment)

- Final notebook used was "final_project_code.ipynb"

Abstract: Housing prices are increasingly unpredictable [1], especially in high-cost areas like Dublin, making it hard for people to find homes within their budget. In this project, we use machine learning to predict house prices from simple inputs: number of bedrooms and bathrooms, property size, and location. Our goal is to provide a tool where users enter their basic requirements and receive an approximate price for a suitable property.

1. Problem framing

Goal: The goal of this project is to predict residential property prices based on a set of relevant features, including the number of bedrooms, number of bathrooms, property size, and location. Accurate price predictions support buyers in an already difficult market

- **Target Variables:** Sale price in euro.
- **Success Metrics:**
  Since this is a regression problem, performance is evaluated using regression metrics:
  - **Root Mean Squared Error (RMSE)**

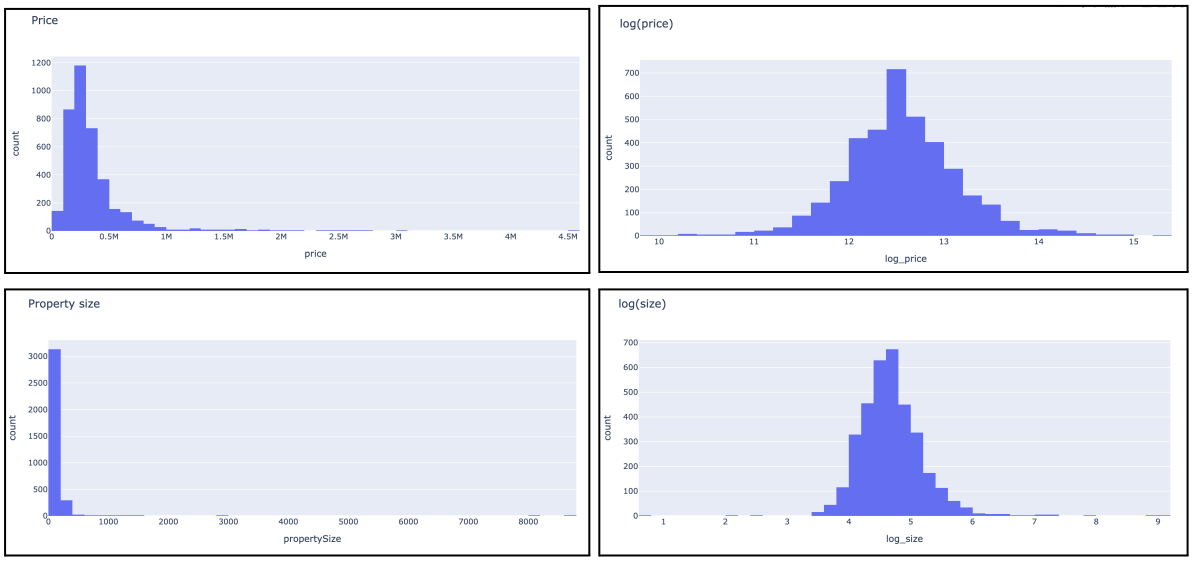
        - Metric is in the same units as price (euro), allowing interpretation of the typical prediction error, lower is better.
    - **Mean Squared Error (MSE)**
        - Average of the squared residuals, used for comparison between models, lower is better (RMSE squared).
    - **R^2 (coefficient of determination)**
        - Measures the proportion of variance explained by the model (1 is perfect, 0 or lower means no better than predicting the mean)

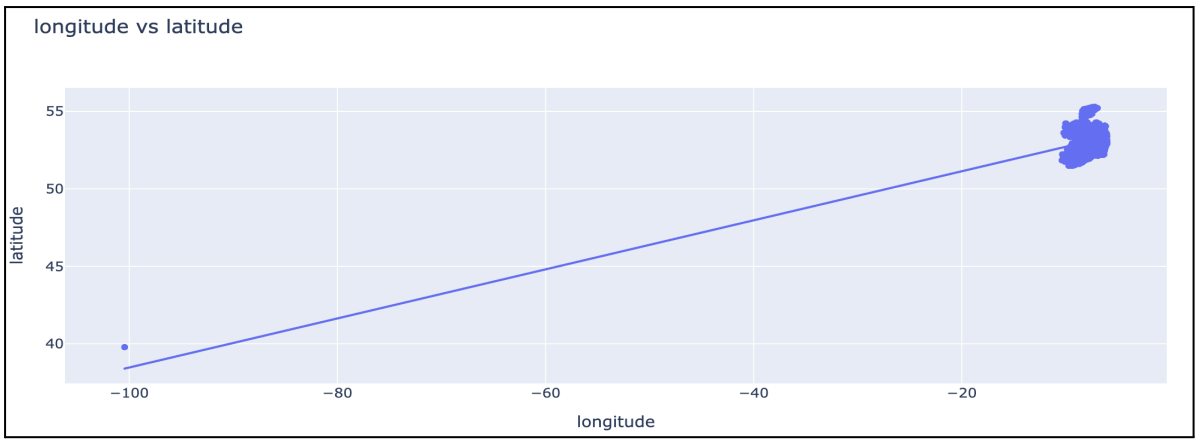
2. Data description: provenance, size, features, licensing, any cleaning you performed

We originally planned to collect live data from daft.ie using an unofficial Python library [2], but a change to Daft's API (adding an access token) broke this approach. Instead, we used an extract from Kaggle [7], which let us continue building our price prediction model. After cleaning the data, we were left with 3,832 records, focusing on price, size, rooms, location, and key property characteristics. The dataset is publicly available and used here for non commercial, educational purposes in line with Kaggle's terms.

3. EDA & preprocessing

We carried out extensive EDA and preprocessing using histograms, boxplots, and scatterplots to understand price and property size, both of which were strongly right-skewed. We applied log transforms and IQR-based outlier removal (e.g. very large houses, tiny properties under 20 m², and a mislocated Cork property plotted in America) to reduce the influence of extreme values and make linear regression more reliable. Missing propertySize values (355 cases) were imputed using a linear regression model after a similarity-based approach failed. We also split address into county and town hoping this would information to our model and improve performance





4. Error analysis & insights: where the model fails/succeeds

- **Successes:**
  - Our model does predict house prices, and the accuracy of the model is better than guessing.
  - Even though the data is not recent, it is still predicted quite accurately. However house prices have possibly doubled since this dataset was collected and would not generalise to new house prices.
- **Failures:**
  - We did not check the accuracy of the R^2 value for the Linear Regression model that we used for imputation.

5. Limitations & future work

The main limitation was the data quality and data availability, as the original Daft API approach failed, which forced us to rely on a Kaggle extract. Since the dataset only covers a specific period, it brings up a sampling bias. Future work includes obtaining a direct Daft API key for real time complete data, expanding coverage across multiple counties, experimenting with more advanced time series and ML models, and incorporating external economic and policy features to improve predictive power. It would also be interesting to build a model specific for Dublin and see how it generalises to other counties (we do not think it would but curious to try). It would also be interesting to combine this dataset with another source, so that we get better performance.

6. Ethics & data protection: consent/licensing, GDPR considerations, bias if any.

The Kaggle dataset is publicly available and licensed for reuse, so individual consent is not required. It also does not contain any personally identifiable information, as addresses relate to properties, seller names are public business names, and there are no contact or financial details, so GDPR risks are low (possibly non-existent). However, there is a sampling and geographic bias as urban areas such as Dublin and Cork are over represented compared to rural areas, the sampling bias occurs due to the data relating to a period in 2022 and access to the live API would remove this.

7. Conclusion.

In conclusion, when attempting to predict prices of properties [3], we need to use data that is recent and in a high volume, so that we can represent different types of properties for example, different counties are taken into consideration. Using log transformations can help your model, but log scale should not be used on the target feature. Random Forest models outperforms Decision Trees and Linear Regression.

References :

[1] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Bern´ardez, and C. Afonso, "Identifying Real Estate Opportunities Using Machine Learning," Applied Sciences, vol. 8, p. 2321, Nov. 2018. Publisher: Multidisciplinary Digital Publishing Institute.

[2] A. Bloomer, "AnthonyBloomer/daftlistings," Nov. 2025. original-date: 2016-06-22T09:07:37Z.

[3] W. K. Ho, B.-S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," Journal of Property Research, vol. 38, pp. 48–70, Jan. 2021. Publisher: Routledge.

[4] A. Louati, R. Lahyani, A. Aldaej, A. Aldumaykhi, and S. Otai, "Price forecasting for real estate using machine learning: A case study on Riyadh city," Concurrency and Computation: Practice and Experience, vol. 34, no. 6, p. e6748, 2022. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6748.

[5] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," in 2019 International Conference on Smart Structures and Systems (ICSSS), pp. 1–5, Mar. 2019.

[6] "A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models | Highlights in Science, Engineering and Technology."

[7] 'daft.ie house price data' - Kaggle - EAVAN (user) https://www.kaggle.com/datasets/eavannan/daftie-house-price-data/data

[8] "Determinants of House Price: A Decision Tree Approach - Gang-Zhi Fan, Seow Eng Ong, Hian Chye Koh, 2006."

[9] "House Price Prediction Analysis Using Linear Regression and Random Forest Algorithms | Journal of Artificial Intelligence and Engineering Applications (JAIEA)."

[10] Z. Zhang, "Decision Trees for Objective House Price Prediction," in 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 280–283, Dec. 2021.

[11] Python, [online], available: https://www.python.org/

[12] Pandas, [online], available: https://pandas.pydata.org/

[13] Numpy , [online], available: https://numpy.org/

[14] Jupyter, [online], available: https://jupyter.org/

[15] Statsmodels , [online], available: https://www.statsmodels.org/

[16] Plotly, [online], available: https://plotly.com/

[17] Scipy, [online], available: https://scipy.org/