

Location, Location, Approximation(?): Encoding Location in Machine Learning for Housing Price Prediction.

2) Abstract

When we buy a home, we do not just purchase bricks and mortar. Location shapes how we live—our access to schools and services—and defines much of a property's value. As machine learning (ML) becomes further embedded in our daily decision-making and housing analytics, we ask how models interpret location. Are they fair and accurate, and can we trust them?

This study investigates how location-related features, especially categorical ones like 'Neighborhood', influence the performance and fairness of housing price prediction models. Using the Ames Housing dataset [1], we test two model types, linear regression and Random Forest Regression, alongside three encoding strategies: One-Hot, Label, and Target Encoding. We evaluate how each combination influences overall predictive accuracy and the sensitivity of predictions to location changes.

To investigate further, we create synthetic test records identical in all features except location and use SHAP values and global feature importance to assess how location drives model decisions.

3) Introduction

Housing affordability and the increasing use of artificial intelligence (AI) in decision-making are two significant forces shaping society today. This study covers both these themes and was partly inspired by the recent experience of one of the authors, who had just purchased a home.

That experience highlighted how emotional and contextual factors such as the feel of a neighbourhood or proximity to family or its location can influence a decision as much as practical housing data like the sale price or the number of rooms. It raised a question: How well can ML models reflect and capture the human and socioeconomic complexity encoded into housing data and decisions?

We are primarily concerned with the following questions:

- If a model ignores location, it may incorrectly assume that two identical houses in different neighbourhoods should cost the same.
- People can use location to infer sensitive information, such as an area with a higher proportion of people from a specific ethnic background or socioeconomic status.[2],[3]
- How does the encoding method impact a model's predictive value if we represent location as a categorical variable in a dataset?

Given the current worldwide affordability crisis in housing and the ever-increasing adoption of AI into our everyday decision-making, we feel this area of study is worth considering for several reasons, including:

- **Pricing Accuracy:** A model that mishandles or ignores location can give inaccurate price predictions. If a model gives two similar houses in vastly different neighbourhoods the exact price prediction, this could mislead buyers or investors, affect market stability, and negatively impact a crucial financial decision for buyers and sellers.
- **Location as a Proxy for Sensitive Attributes:** Model training data encoded with location-related bias could produce models that lead to unjust outcomes in housing distribution, loan approvals, and other important financial decisions.[2],[3]

Ensuring our AI systems are transparent and accountable is essential for fairness and equity in the practical application of AI and for promoting and adopting AI in our everyday decision-making processes.

This study investigates how machine learning models consider location-related features when predicting house prices and how location is interpreted and weighted in different ML pipelines when predicting house prices using the Ames Housing dataset [1]. With synthetic test records, we explore whether specific encoding strategies or model types amplify or diminish the influence of location-related features and how these choices impact accuracy and fairness.

We also used interpretability methods such as SHAP and Random Forest Feature Importance Matrixes to assess the extent to which location, property features, or both drive predictions.

Finally, the study considers Legal, Social, Ethical, and Professional factors related NEEDS WORK - TODO

3a) Related Works

Recent studies have highlighted concerns about fairness and bias in ML models, particularly when location is a predictive feature. Location variables, such as neighbourhood or postcode, are often strongly correlated with house prices—but can also act as proxy values for sensitive socioeconomic attributes like race, income, or the broad education level of whole areas or individuals.[2],[3]

Fernandes Machado et al. [4] state, "**Spatial features are known to be strong predictors of house prices, but they also embed socioeconomic disparities, often mirroring long-standing segregation patterns.**"

Similarly, Kwegyir-Aggrey et al. [5] demonstrate that contextual features such as location, income, and education are "**highly predictive of race in the U.S., even when race is not observed**" and that location-related features "**encode structural biases, warranting careful ethical consideration.**"

It is clear sensitive and protected attributes such as race, ethnicity, and income levels can be inferred indirectly through geographic or location-related features. Even when these characteristics are not explicitly included, models can pick up on patterns embedded in the data and learn to act on them. This raises ethical concerns. Model predictions may reinforce historical patterns of inequality by assigning value or risk in ways that reflect the social issues of the past more than the objective characteristics of a property in a real estate valuation or transaction.

Beyond the concerns related to fairness and bias, recent studies highlight the technical challenges with location encoding for ML models.

Encoding location-related categorical variables, such as postal codes or neighbourhood names, may introduce high cardinality into a dataset, causing reduced model efficiency. While one-hot encoding is commonly used, it becomes impractical when the number of location-related categories is significant. A study published by Gnat [6] in *Procedia Computer Science* found different encoding methods significantly impact a model's performance. Target encoding in particular found to improve model accuracy but also increase the risk of data leakage and overfitting if not correctly handled.

Gnat [7] found that encoding strategies significantly influence valuation outcomes, noting that "mass valuation results vary depending on how the encoding of categorical variables occurs." In a related study, Gnat [4] highlighted the need for variation in location-related categorical features to improve the accuracy and interpretability of related ML models.

Given these challenges, we felt it worthwhile considering how different ML pipelines—including feature encoding, model selection, and interpretability tools—affect how ML models learn from location data. This project investigates how location is interpreted and weighted in different ML pipelines when predicting house prices using the Ames Housing dataset [1]. We explore whether certain encoding strategies or model types combined amplify or diminish the influence of location and how these choices impact accuracy, fairness, and interpretability.

Using synthetic test records, model evaluation metrics, and interpretability methods such as SHAP, we aim to highlight the technical implications of modelling with geographic features and how they may feed into ethical concerns related to bias, fairness, and ML interpretability.

3b) Aims and objectives

TODO

4) Methods

For our project, we tested our theories against the Ames Housing dataset. The dataset contains detailed property sales records in Ames, Iowa, USA and is often used in regression-based Machine Learning projects to predict house prices.

Our initial exploration of the dataset and review of several related papers directed us towards studying how sensitive models were to location-related features in general and if the encoding method of categorical variables relating to a property's location could impact a model's effectiveness when predicting house prices.

About the dataset:

- The dataset contains 2,930 unique records and 79 variables, with categorical and numerical features such as the number of bedrooms, quality ratings, garage size, square footage, etc.
- The target variable is SalePrice (a continuous variable in \$)
- Neighborhood is a categorical variable representing the property's location in AMES, and we use it as the key indicator of a model's sensitivity and the use of encoding formats for location-related features.

Exploratory Data Analysis

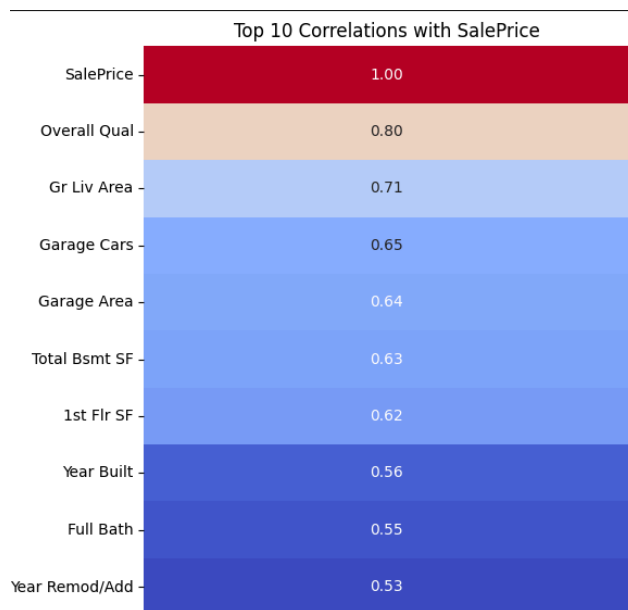
We began by exploring the dataset to begin forming a picture of the relationship between a property's features, location, and Sales Price.

We dropped two transactional, non-property-related columns:

- PID
- Order

Correlation Matrix

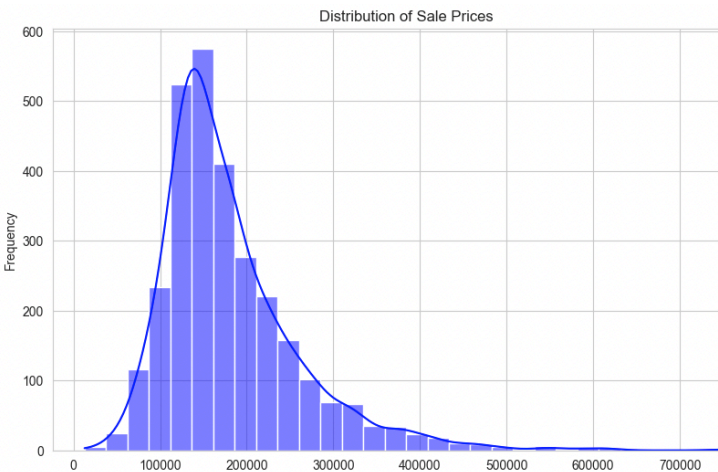
Our correlation matrix showed that **Overall Qual** (Ordinal: rating of overall materials and finish) and **Gr Liv Area** (Continuous: above-ground living area) had the strongest positive correlations with SalePrice.



Feature Groupings by Type

We grouped all the columns into lists of Categorical (e.g., Neighborhood) and Numerical (e.g., Overall Qual, Gr Liv Area) features so we could use them in our encoding processes.

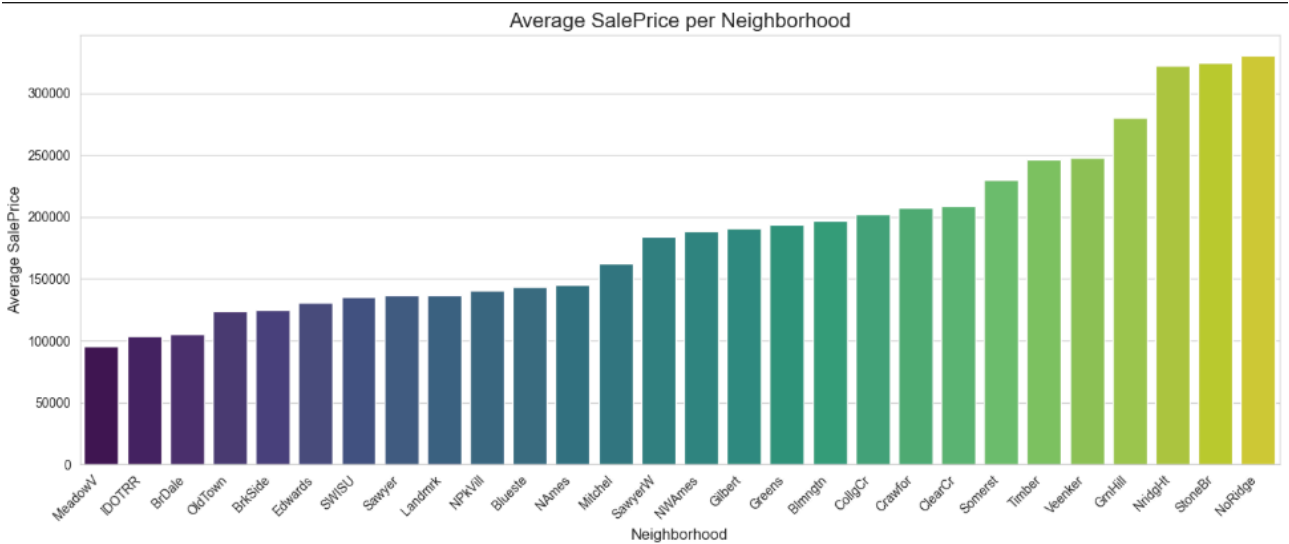
Sale Price Distribution and Handling Outliers



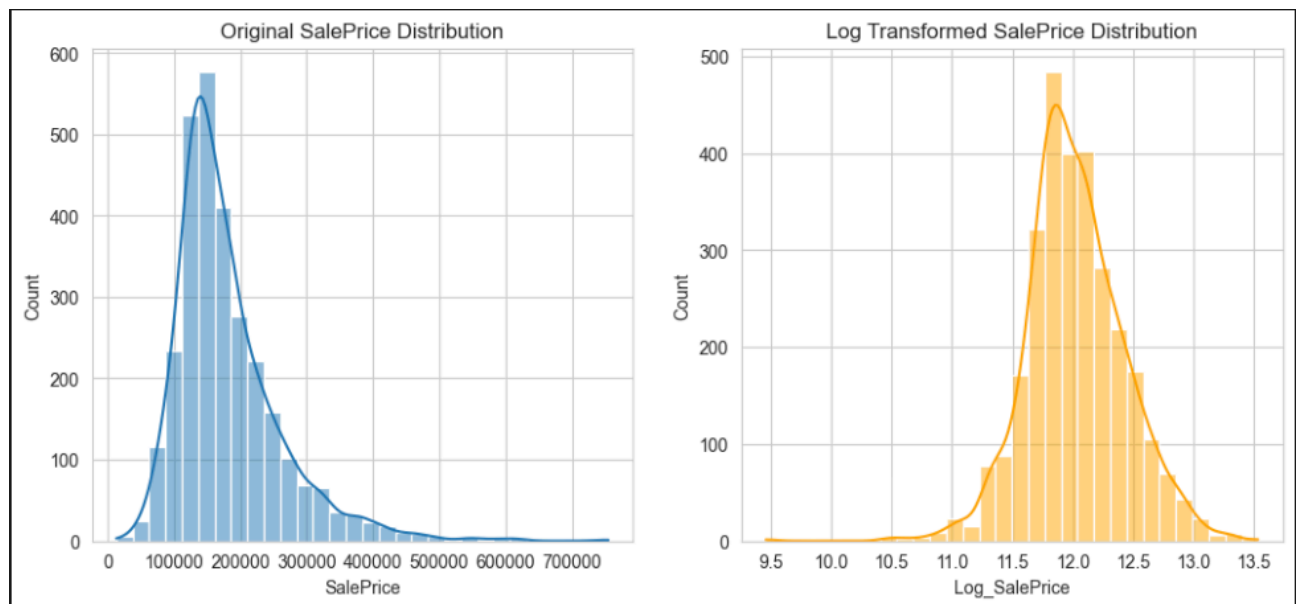
The distribution of SalePrice was found to be right-skewed, with most properties priced between \$100,000 and \$200,000.

Average Sales Price Per Neighborhood and Outliers

A bar plot of average SalePrice by Neighborhood revealed clear location-based price patterns, with NoRidge having the highest average prices and MeadowV the lowest.



Neighborhood	Avg Sale Price
NoRidge	330319
MeadowV	95756



We used a boxplot and the Interquartile Range (IQR) method to identify statistical outliers and measured the total number of outliers against ‘NeighborHood’ and ‘Overall Qaul’

Neighbourhood	Count Outliers		Overall Qaul	Count Outliers
NridgHt	62		9	60
StoneBr	22		8	48
NoRidge	21		10	25
Timber	8		7	3

We identified most outliers as higher-quality homes in the more upscale neighbourhoods. We considered them statistically relevant and did not remove them from the dataset. Instead, we applied a log transformation to the target variable to reduce its impact on our model's predictions (especially for linear regression).

From this point, we knew the average sales price spread of the categorical feature ‘NeighborHood’ made that feature influential and would form the basis of our study on how models interpret location.

Handling missing values

We decided not to drop any rows with missing values.

We filled any missing categorical values with their most frequent value (mode) and missing numerical features with their mean.

Synthetic Test Cases

We created two new synthetic records using the median value of numerical features and the mode of categorical features.

Otherwise identical records, we placed one record in NoRidge and the other in MeadowV, the highest and lowest-priced neighbourhoods. We added these records to our dataset and flagged them

records as synthetic, holding them back from our model training. We used them to observe how different models respond to the Neighborhood feature (location) encoding methods at an individual record level.

Encoding Strategies for Neighborhood and Categorical Variables.

We used three encoding strategies for the Neighborhood variable across three separate copies of the dataset:

- **One-Hot-Encoding:** Each Neighborhood as a binary feature.
- **Label Encoding:** Categories mapped to integer values.
- **Target Encoding:** Each Neighborhood was assigned its mean price from the training set.

In each dataset, we used One-Hot-Encoding to encode all other categorical features. We recognise the potential dimensionality problems this adds to the dataset. We focused on how models interpret location-related features rather than giving the most accurate price predictions. We note that the specific accuracy of our models could have been improved by more targeted encoding of specific categorical variables. We address this in the paper's section on limitations and future work.

Model Training and Evaluation

We trained two model types to look to see if they could respectively capture any linear and non-linear relationships between features and SalesPrice.

- Linear Regression
- Random Forest Regressor

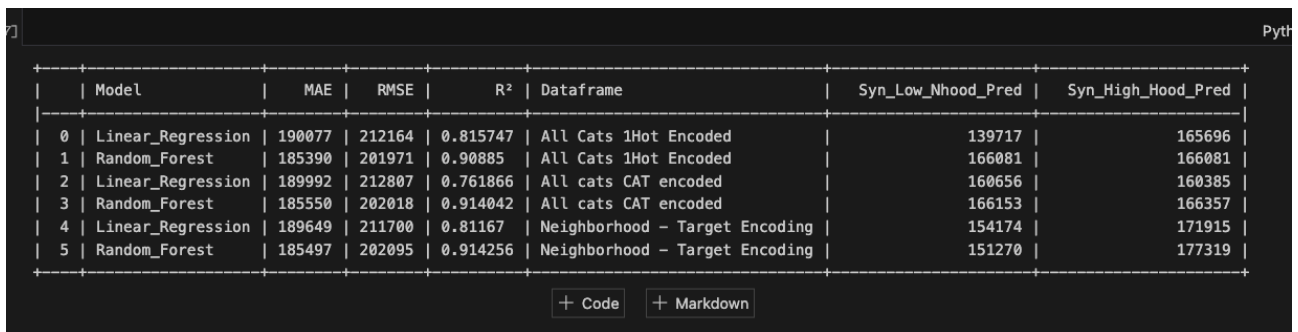
For each encoding method, we:

- Split our data into standard training and test sets (80%/20%).
- Trained and tested both models against our global datasets.
- Evaluated model performance on the global dataset using MAE, RMSE, and R^2 .
- Ran predictions on the synthetic test records to examine location sensitivity at the individual record level.

To interpret our model's results, we:

- Calculated and reviewed the global feature importances of the Random Forest predictions.
- Applied SHAP (Shapley Additive Explanations) to interpret the global model and individual synthetic predictions across both Random Forest and Linear Regression model predictions.

5) Results



The screenshot shows a Jupyter Notebook interface with a table of model performance metrics. The table has 8 columns: Index, Model, MAE, RMSE, R², Dataframe, Syn_Low_Nhood_Pred, and Syn_High_Hood_Pred. It contains 6 rows of data. Below the table are two buttons: '+ Code' and '+ Markdown'.

	Model	MAE	RMSE	R ²	Dataframe	Syn_Low_Nhood_Pred	Syn_High_Hood_Pred
0	Linear_Regression	190077	212164	0.815747	All Cats 1Hot Encoded	139717	165696
1	Random_Forest	185390	201971	0.90885	All Cats 1Hot Encoded	166081	166081
2	Linear_Regression	189992	212807	0.761866	All cats CAT encoded	160656	160385
3	Random_Forest	185550	202018	0.914042	All cats CAT encoded	166153	166357
4	Linear_Regression	189649	211700	0.81167	Neighborhood - Target Encoding	154174	171915
5	Random_Forest	185497	202095	0.914256	Neighborhood - Target Encoding	151270	177319

We trained Linear Regression and Random Forest Regressor models on our three differently encoded versions of the Ames Housing dataset. The three encoding strategies applied to the 'Neighborhood' feature were One-Hot Encoding, Label Encoding, and Target Encoding.

We evaluated each model based on its Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score on the global dataset.

Additionally, we made predictions for two synthetic records (identical in all features except for Neighborhood: one in NoRidge, one in MeadowV). We compared these to assess our model's sensitivity to the encoded Neighborhood variable.

Performance Overview

In general terms, our Random Forest Regressor model consistently outperformed the Linear Regression model.

- **R² scores for Random Forest were in the range of 0.909 to 0.914, compared to 0.762 to 0.816 for Linear Regression.**
- **In every encoding scenario, MAE and RMSE values were lower for Random Forest, indicating better predictive accuracy.**

The Effects of Different Encoding Methods

One-Hot Encoding performed best for our linear regression models (**R² = 0.816**), slightly outperforming Label Encoding and Target Encoding.

This result aligned with our expectations, as One-Hot Encoding provides explicit binary values for each neighbourhood and is generally considered the best encoding method for linear regression models.

Our Random Forest models, being nonlinear and tree-based, demonstrated their flexibility and ability to handle all encoding methods, consistently performing across all three encoding strategies.

Results such as this could be expected as random forest, which is considered more reliable than linear models for problems involving high dimensionality, which our datasets had due to the uniform use of one-hot encoding across the other categorical features of the dataset.

Synthetic Record Predictions: Sensitivity to Location

We analysed the difference in predicted **SalePrice** for the two synthetic records:

- For Random Forest + One-Hot Encoding, both synthetic records returned identical predictions (**\$166,081**), indicating that the model may not have distinguished between the neighbourhoods with the highest and lowest average prices.
- The predicted prices for Random Forest + Target Encoding were **\$151,270** (MeadowV) and **\$177,319** (NoRidge), a clear **\$26K+** difference indicating that the model learned the neighbourhood pricing differences.
- Linear Regression also showed price variation between neighbourhoods but to a lesser and more inconsistent degree. For example, the Target Encoded data predictions differed by over **\$17,000** in the correct direction, whereas Label Encoding gave nearly identical predictions.

Interpretability and SHAP Insights

We complemented our model performance evaluation with SHAP analysis to explain our model's behaviour:

- **Global SHAP values** from the Random Forest model confirmed that 'Neighborhood' was among the most influential features in the Target Encoded dataset.
- **Local SHAP explanations** for the synthetic records revealed that the location alone contributed substantially to the differences in predicted prices, particularly under the Target Encoding + Random Forest scenario.

Summary of Findings

- The Random Forest Regressor outperformed Linear Regression across all encoding methods in our tests.
- The encoding strategy significantly affects how linear models consider location data, with One-Hot Encoding performing the best.
- Our Random Forest model demonstrated robustness to the encoding style but shows it handles location data better when paired with Target Encoding.
- Synthetic test cases proved effective in directly testing location sensitivity.
- SHAP analysis validated the importance of Neighborhood in price prediction and provided transparency into model decision-making.

These results confirm that machine learning models can learn and reflect location-based price patterns and that encoding choices can play an important role in model fairness, interpretability, and effectiveness.

6) Discussion

For our project, we worked with the Ames Housing dataset to study how ML models learn location-related price differences and the impact of different encoding methods on their ability. We used 'Neighborhood', a categorical variable giving the location of each property, as the location-determining feature.

Following a structured ML project pipeline methodology—including data exploration, feature engineering, encoding strategies, model comparison, and SHAP-based interpretability—we showed that the choice of model and the categorical encoding method significantly impacted the model's ability to learn and reflect location-based pricing through its predictions.

Results showed that the Random Forest Regressor model outperformed Linear Regression regarding accuracy across all three encoding strategies we tested and the ability to learn and reflect location-related pricing differences across the dataset's neighbourhoods.

Across the encoding methods, Target Encoding gave the strongest location sensitivity, demonstrated by both global performance metrics and individual predictions of the synthetic test cases. SHAP analysis further confirmed location's central role in model decisions, particularly when using Target Encoding.

The study also surfaced important fairness and ethical considerations. While location is a strong price predictor, it can also reflect and perpetuate historical and systemic inequalities. Our use of synthetic test cases showed that ML systems can give different outcomes solely based on geographic location, even when all other features are constant. Naturally, people see some neighbourhoods as more desirable, pushing up property prices and reflecting fundamental differences in a resident's quality of life. However, if models rely too much on location-related features, they could reinforce historical inequalities and biases by undervaluing these properties in historically marginalised areas. Biased predictions could deny viable mortgages; investors could overlook areas for investment opportunities and urban development.

ML models must accurately reflect location-based price differences to ensure consistent and fair outcomes in housing valuations and lending practices. However, they should equally and fairly factor in property-related features to mitigate unintended biases and historical inequalities potentially encoded within housing-related datasets.

a) Limitations

We acknowledge the following limitations of our work:

Dimensionality in the Dataset: By One-Hot-Encoding all categorical variables other than 'Neighborhood', we increased the dimensionality of our dataset, which likely decreased the performance of our models, though this would have been reflected uniformly across all models and tests.

Model Diversity: We only tested two model types. Including other algorithms (e.g., Gradient Boosting, Neural Networks) would have provided a broader set of results to compare.

Synthetic Test Design: For simplicity, we used only two synthetic records with median/mode imputation. A wider range of synthetic records could help assess model behaviour, further considerations and hypotheses.

Feature Engineering: While SHAP provided interpretability, a more extensive analysis of interaction effects—especially between Neighborhood and structural features—could deepen the understanding of the interaction of Neighborhood and property-related features.

Generalisation: We based our tests on the Ames dataset, which, while detailed, represents a single geographic area. Our findings may not be replicated on different housing markets.

Future Work

Future research could explore:

- Applying **fairness-aware** ML frameworks to mitigate location-based bias.
- **Spatiality and clustering** - Introducing latitude and longitude data to the dataset while running clustering algorithms could help identify hidden location-related price clusters within the dataset.
- Evaluating models using causal inference techniques to distinguish correlation from causation in location effects.
- Using multi-region datasets to generalise findings and evaluate equity across diverse markets.

7) Conclusion

Overall, this project reinforces the importance of aligning technical accuracy with ethical responsibility through the best choice of ML algorithm paired with the best encoding processes for location-related categorical variables in housing-related ML models for transparent, fair and accurate house-price predictions.

8) References

- [1] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," J. Stat. Educ., vol. 19, no. 3, 2011.
- [2] American Academy of Actuaries, "Discrimination Risk and Insurance," Risk Classification and Discrimination Brief, Aug. 2023. [Online]. Available: <https://www.actuary.org/sites/default/files/2023-08/risk-brief-discrimination.pdf>
- [3] J. E. Holloway, Calculating Race: Racial Discrimination in Risk Assessment. Oxford University Press, 2023. [Online]. Available: <https://academic.oup.com/book/33461/chapter/287737508>
- [4] A. Fernandes Machado, M. George, and G. Peyré, "Geospatial Disparities: A Case Study on Real Estate Prices in Paris," arXiv preprint arXiv:2401.16197, 2024. [Online]. Available: <https://arxiv.org/abs/2401.16197>
- [5] K. Kwegyir-Aggrey, S. Joshi, and B. Hutchinson, "Observing Context Improves Disparity Estimation when Race is Unobserved," arXiv preprint arXiv:2409.01984, 2024. [Online]. Available: <https://arxiv.org/abs/2409.01984>
- [6] S. Gnat, "Impact of Categorical Variables Encoding on Property Mass Valuation," Procedia Computer Science, vol. 192, pp. 3542–3550, 2021. doi: <https://doi.org/10.1016/j.procs.2021.09.127>
- [7] S. Gnat, "Categorical Variable Problem in Real Estate Submarket Determination with GWR Model," Real Estate Management and Valuation, vol. 30, no. 4, pp. 42–54, 2022. doi: <https://doi.org/10.2478/remav-2022-0028>

