Ji Wu

Professor Luyao Zhang

STATS 201

March 5, 2025

House Price Forecasting in King County Based on the Fusion of Textual and Structured Data

1. Background and Motivation

The real estate market is a key component of the global economy, influencing investment decisions, urban planning, and housing affordability. Traditional house price prediction models primarily rely on structured numerical data, such as square footage, number of bedrooms, and location. However, qualitative descriptions of properties—often written by real estate agents or generated through AI—contain rich contextual information that could enhance price prediction models (Kok, Monkkonen, & Quigley, 2014).

Existing literature has demonstrated that Natural Language Processing (NLP) can extract valuable insights from unstructured data, such as social media sentiment analysis (Pang & Lee, 2008) and financial forecasting (Bollen, Mao, & Zeng, 2011). However, limited research has examined the impact of textual property descriptions on house price prediction. Studies by Antipov and Pokryshevskaya (2012) suggest that integrating textual data can improve predictive accuracy, while more recent work (Sundaresan et al., 2020) highlights the potential of deep learning techniques in multimodal analysis. This study aims to bridge this gap by evaluating whether textual descriptions improve price prediction accuracy when combined with structured numerical data.

Recent advances in Generative AI (GenAI) provide opportunities to create synthetic property descriptions, enabling further analysis of AI-generated versus human-written descriptions. Research in computational linguistics (Radford et al., 2019) has shown that transformer-based language models can generate coherent,

contextually relevant text, but their impact on structured prediction tasks remains underexplored. This study will also explore the potential benefits and limitations of using AI-generated text in real estate valuation models.

2. Research Question

This study focuses on the following core question:

Does fusing structured house features with textual descriptions significantly improve the prediction accuracy of King County's house price prediction model?

3. Application Scenarios

The results of this research have multiple applications in practice:

- Property valuation and pricing: provide buyers, sellers, and investors with more accurate home price references to optimize investment decisions.
- Automated property description generation: using generative AI to automatically generate property descriptions, improving intermediary efficiency and reducing labor costs.
- Cross-regional model promotion: The research results help validate the applicability of the model in other regions and promote wide-scale application.

4. Methodology

In the methods section, we focus on supervised learning based prediction methods in detail.

4.1 Prediction Methods

(1) Data preprocessing and feature engineering

This study is based on the King County home sales dataset, and the structured data are first processed with missing values (using median padding), normalized and uniquely hot coded. For textual data, generative AI (e.g., ChatGPT) is utilized to generate

house description text, and denoising, word splitting, word shape reduction, and vectorization are performed, and commonly used methods include TF-IDF and BERT embedding (Devlin et al., 2018).

(2) Model Construction

We constructed two models for comparison:

Baseline model (structured data only): Integrated regression methods such as Random Forest and XGBoost were used, which are excellent in handling nonlinear relationships (Antipov & Pokryshevskaya, 2012).

Text augmentation models (structured data + text features): Fusing text vectors with structured features, inputting them into a supervised learning model through feature splicing, or capturing text sequence information using an LSTM model, thus integrating the advantages of both types of data.

(3) Model Evaluation

Model performance is evaluated by root mean square error (RMSE), mean absolute error (MAE) and R^2 metrics. We also utilize cross-validation and hyper-parameter tuning to ensure the model's ability to generalize.

(4) Key Benefits

By introducing textual features, the prediction model is able to capture semantic and emotional information that cannot be reflected in structured data, thus improving the accuracy of house price prediction. For example, keywords such as "luxurious", "spacious", or "charming" in the description of a house often reflect the added value of the property (Pang & Lee, 2008). value of a property (Pang & Lee, 2008). This method not only improves the prediction accuracy, but also provides a basis for subsequent market sentiment analysis.

5. Results

The anticipated results of this study include:

- Prediction Performance Improvement: The model incorporating text features has obvious improvement in RMSE and MAE indicators, and the R^2 score is increased, which verifies the gain effect of text information.
- Interpretation of feature contribution: Through the attention mechanism and other interpretation tools, the contribution of text features to the prediction results can be quantitatively demonstrated, providing an intuitive basis for model interpretation.
- Validation of causal effects: Although causal inference is only a complementary approach, the results will provide preliminary empirical support for understanding the impact of policy variables on house prices.

6. Intellectual Merit and Practical Impacts

Academic Contribution:

- Innovative fusion of structured and textual data to improve the accuracy of house price prediction
- Extends the application of machine learning in the social sciences, especially in conjunction with causal inference methods
- Provides empirical analysis of the benefits and risks of generative AI in real-world applications

Practical impacts:

The results of this study have significant practical implications and applications at multiple levels:

- **Social and Economic Benefits:**

  By fusing structured data with listing description text, this study provides a more accurate method for house price prediction, which helps to reduce the risk of volatility due to market information asymmetry. A more accurate house price

assessment not only improves the quality of decision-making between buyers and sellers, but also provides more reliable data support for government departments in formulating real estate regulation policies, which in turn promotes the rational allocation of social resources and the sustainable development of regional economy.

- **Industry application:**

  The methodology proposed in this project is applicable to a wide range of fields such as real estate valuation, automated information generation for real estate agents and market dynamics monitoring. The automatically generated property descriptions can significantly improve the efficiency of intermediaries and reduce the cost of manual entry and description writing; meanwhile, the refined prediction model can provide investors and developers with forward-looking market trend analysis, helping to formulate accurate marketing and pricing strategies.

- **GAI Governance and Ethical Considerations:**

  In exploring the integration of generative AI with traditional data, this study fully considers the ethical risks and potential bias issues in the application of the technology. The project is committed to promoting the inclusive development of AI technology, and proposes risk mitigation strategies by cautiously assessing the reliability and fairness of the generated text to ensure that the technology application is always in line with AI governance principles and the Sustainable Development Goals (SDGs). At the same time, this project encourages the introduction of transparent and interpretable mechanisms in data processing and model design, providing theoretical support for safeguarding users' rights and interests and promoting the implementation of technological responsibility.

Reference

Acharya, Viral V., Cecilia Parlatore, and Suresh M. Sundaresan. 2020. “A Model of Infrastructure Financing.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3689262.

Antipov, Evgeny A., and Elena B. Pokryshevskaya. 2012. “Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics.” *Expert Systems with Applications* 39 (2): 1772–78. https://doi.org/10.1016/j.eswa.2011.08.077.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science* 2 (1): 1–8.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Kok, Nils, Paavo Monkkonen, and John M. Quigley. 2014. “Land Use Regulations and the Value of Land and Housing: An Intra-Metropolitan Analysis.” *Journal of Urban Economics* 81 (May): 136–48. https://doi.org/10.1016/j.jue.2014.03.004.

GenAI used: ChatGPT, Storm, Deepseek

GitHub: https://github.com/Rising-Stars-by-Sunshine/Ji-Wu-Final-Project

Appendix

To further validate the causal effect of specific policy variables (e.g., regional income level, age of housing, etc.) on house prices, we employ a regression discontinuity design (RD). This method provides causal inference evidence by identifying explicit policy thresholds and estimating the jump effect of house prices before and after a policy change (Imbens & Lemieux, 2008). Although the focus of this study is on forecasting, causal inference analysis can assist in explaining market changes and provide theoretical support for policy formulation.