A Comparison of House Price Classification with Structured and Unstructured Text Data

Erika Cardenas
Florida Atlantic University
ecardenas2015@fau.edu

Connor Shorten
Florida Atlantic University
cshorten2015@fau.edu

Taghi M. Khoshgoftaar
Florida Atlantic University
khoshgof@fau.edu

Borivoje Furht
Florida Atlantic University
bfurht@fau.edu

Abstract

Purchasing a home is one of the largest investments most people make. House price prediction allows individuals to be informed about their asset wealth. Transparent pricing on homes allows for a more efficient market and economy. We report the performance of machine learning models trained with structured tabular representations and unstructured text descriptions. We collected a dataset of 200 descriptions of houses which include meta-information, as well as text descriptions. We test logistic regression and multi-layer perceptron (MLP) classifiers on dividing these houses into binary buckets based on fixed price thresholds. We present an exploration into strategies to represent unstructured text descriptions of houses as inputs for machine learning models. This includes a comparison of term frequency-inverse document frequency (TF-IDF), bag-of-words (BoW), and zero-shot inference with large language models. We find the best predictive performance with TF-IDF representations of house descriptions. Readers will gain an understanding of how to use machine learning models optimized with structured and unstructured text data to predict house prices.

Introduction

Predicting house prices is crucial for buyers, sellers, and the economy as a whole. Real estate prices fluctuate based on multiple factors such as macroeconomic indicators and local economic activity. Agents in the real estate market alter their decisions based on their information about the state of the economy. The notion of information asymmetry states that varying information from the buyer or seller could lead to market failure. Spence demonstrated that sellers could signal to buyers in order to avoid adverse selection (Spence 2002). Signaling is the action of one agent with more information conveying insights to the other party. Cypher et al. studied the impact of price signaling from brokers in commercial real estate transactions (Cypher et al. 2017). Their study found that if the price is above the market rate, then pricing is dependent on the strength of the signal. We explore signaling through transparent pricing with accessible predictions from machine learning models. Transparent pricing would lead businesses or individuals to be more confident and make decisions quicker, leading to a more efficient market. Forecasting house prices is a tool for transparent pricing that we explore with machine learning techniques.

We experiment with predicting house prices with data-driven techniques. The key to data-driven predictive analytics is the feature representation of houses in the dataset. We collect a dataset of prices, meta-information about houses, and written descriptions. Meta-information includes features about houses such as the number of bedrooms, bathrooms, and square feet. Written descriptions are an emerging data source, with excitement fueled by recent advances in deep learning for natural language processing. A written house description is an open-ended report typically written by real estate brokers to describe a particular house.

Our experiments present data-driven modeling of 200 houses featurized by meta-information and written descriptions. We convert the written descriptions to numeric inputs through the term frequency-inverse document frequency (TF-IDF) algorithm, which is described in further detail in the Related Work section of our report. We begin by presenting the performance difference of a logistic regression model mapping either tabular descriptions or TF-IDF vectors to binary bins of house prices. The logistic regression model processing tabular data achieves 79.3% training accuracy and generalizes to 76.2% test accuracy. The logistic regression model with TF-IDF impetus achieves a 2.4% higher test accuracy at 78.6%, however, it has fit the training data very closely, achieving 100% training accuracy. This generalization gap between training and testing inspired our interest into deep learning architectures for TF-IDF representations of unstructured text data. We utilize a 3-layer nonlinear MLP architecture and add dropout to control for overfitting and limit the variance of the parametric function. We further compare TF-IDF representations of house descriptions to BoW representations. We additionally present the zero-shot inference of a 6-billion parameter publicly accessible language model to predict house prices.

In summary, our contributions are as follows:

- We compare the predictive performance of tabular and TF-IDF representations of text to predict house prices.
- We report the effectiveness of dropout to prevent overfitting in our deep learning architectures modeling TF-IDF

representations of text.

- We present concrete details about the scaling of TF-IDF vocabularies with limited datasets.
- We present a comparison of BoW and TF-IDF representations of house descriptions for classifying prices, finding TF-IDF to be superior.
- We present a zero-shot house price prediction of a 6 billion parameter language model trained on internet text.

Related Work

Natural Language Processing and Computer Vision

Natural Language Processing (NLP) analyzes and represents human language in order to build statistical models. Hausler et al. conducted an experiment using news-based sentiment on the effects of US securitized and direct commercial real estate markets (Hausler et al. 2018). They reported the performance of using a Support Vector Machine (SVM) algorithm to classify the sentiment of real estate news headlines. Velthorst and Guven used text mining and machine learning to predict the Dutch housing market (Velthorst and Guven 2019). The goal of their experiment was to forecast the upward or downward trend of the market based on text data from Twitter. Abdallah and Khashan performed text mining to predict the accuracy of home prices by using textual data in real estate classifieds (Abdallah and Khashan 2016).

Computer vision is the application of using digital images to train computers to understand the visual world. Recent work has been done to use images listed on property listings to predict house prices. Ahmed and Moustafa developed a dataset that combined photos with textual house information (Ahmed and Moustafa 2016). The combined features were put into a multilayer neural network that predicted the house price. Law et al. developed a deep neural network model that utilizes both visual features from images and traditional tabular features (Law et al. 2019). Their study used street images and satellite image data to improve the estimation of house prices. Bency et al. proposed a Convolutional Neural Network (CNN) framework to model geospatial data to learn the spatial correlations in house prices (Bency et al. 2017).

Structured and Unstructured Data Fusion

Many other experiments have worked on combining structured and unstructured data. As discussed above, researchers have combined digital images with tabular data to model home prices. Cheng et al. experimented with forecasting financial time-series with multi-modality graph neural networks (Cheng et al. 2021). Cheng et al. proposed using a Multi-Modality Graph Neural Network (MAGNN) to learn from the various inputs (Cheng et al. 2021). Pancerz and Grochowalski explore a mapping algorithm to convert unstructured text data to structured tables (Pancerz and Grochowalski 2017).

Methodology

Our research further explores representations of text data for machine learning modeling. We represent our unstructured descriptions of houses as term frequency-inverse document frequency (TF-IDF) vectors. The TF-IDF algorithm begins by computing the number of unique words in the dataset. The representation vectors are initialized with empty vectors. The dimensionality of these empty vectors is sourced from the cardinality of unique words. TF-IDF then populates these vectors by looping through the dataset and mapping the count of each word in its respective data instance. The sparse vector is indexed with the token index of each respective word and populated with its frequency. Although these representations are very sparse (many 0 entries in the vector), they have been shown to have strong predictive power in NLP applications.

We further compare TF-IDF representations with Bag-of-Words (BoW) and language modeling with large scale transformers. BoW represents a text sequence as a time-series sequence of categorical labels. Differently from TF-IDF, the input size of BoW is limited with a maximum sequence length parameter, rather than the entire size of the vocabulary. For example, bag of words may represent the following sequence of "The squirrel jumped over the fox" as "[50, 105, 24, 8, 50, 578]." To generalize from BoW to neural embeddings, each categorical index is converted to a continuous vector representation. For example, the "578" categorical index used to represent "fox" would be mapped to an n-dimensional continuous vector. These vectors are stacked together to produce a matrix representation of the text data.

We experiment with classifying houses into price buckets. We begin with a binary classification mapping houses into predictions of greater than or less than $6,000,000. Our dataset contains house descriptions from 4 cities, Boston, Cambridge, Fort Lauderdale, and Boca Raton. Our tabular dataset consists of 5 features: city, street, bedrooms, bathrooms, and square feet. Our text descriptions contain an average of 200 words with 4,873 uniquely appearing words.

The following experiments highlight performance factors of predicting house prices from structured and unstructured data. We begin with a comparison of tabular and TF-IDF data representations. We then show the effectiveness of a deep learning-based Multi-Layer Perceptron (MLP) architecture to model these data sources. Finally, we present a comparison of TF-IDF, BoW, and language modeling representations of the data and task of house price prediction. Figure 1 illustrates the distribution of labels in our dataset. The prices of our set of houses fall into a power-law distribution. This is an interesting property related to the binary threshold separation which we report in Figure 1 (B). We leave it to future work to further chunk up these labels into fine-grained categories. For these experiments, we isolate the number of 0 and 1 labels to control for issues due to class imbalance (Johnson and Khoshgoftaar 2019).

Tabular versus TF-IDF Representations of Houses

We begin our experiments by comparing structured and unstructured data sources for house price classification. We

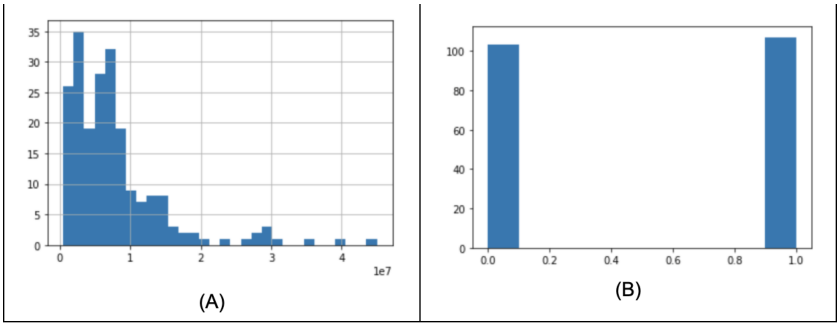

Figure 1: Depicted left, (A) presents the distribution of house prices in our dataset. On the right, (B) illustrates the roughly even distribution after dividing house prices into binary categories based on a price threshold.

begin with a logistic regression analysis to ground the experiments before introducing a deep MLP architecture. Our tabular data contains 5 features, categorical city and street names, bedrooms, and bathrooms, as well as continuous square feet. Our TF-IDF representation is sourced from a space-separated tokenization of words in the house descriptions. We do not include a special out of vocabulary token in our TF-IDF vectorization, such that if a word in the test set is not included in the TF-IDF vectorization formed from the training set, it is not represented during the test inference. Table 1 illustrates the performance of these models. The TF-IDF logistic regression achieves 2.4% higher accuracy than the tabular logistic regression. Further, TF-IDF logistic regression is able to perfectly fit the training data and still possess a reasonable generalization ability. This result inspired additional investigation into higher capacity models such as deep MLPs, as well as the use of dropout regularization to reduce the train-test generalization gap.

	Train Accuracy	Test Accuracy
Tabular Logistic Regression	79.3%	76.2%
Tabular MLP	87.5%	71.4%
TF-IDF Logistic Regression	100%	78.6%
TF-IDF MLP	100%	76.2%
BoW Logistic Regression	100%	47.6%
BoW MLP	85.1%	61.9%

Table 1: A comparison of tabular,TF-IDF, and BoW data inputs, as well as a comparison of logistic regression and MLP models.

Deep Learning with Multi-Layer Perceptrons

Following experiments with logistic regression models, we explore the effectiveness of a deep MLP model for predicting house prices. Our MLP architecture is adapted for the two input sizes of 4873-d and 5-d for TF-IDF and tabular representations, respectively. Both data sources test the same MLP architecture of 3 hidden units containing 512, 1024, and 512 units. Each of these units are passed through a non-linear ReLU activation which prevents negative values in the intermediate activations. Each MLP makes a prediction by compressing the last layer of 512 units into a single prediction modulated with a sigmoid activation function. Due to the differences in input size, the TF-IDF MLP contains 5.5 million parameters, whereas the tabular MLP contains 1.05 million parameters. Table 1 presents the results of this experiment. The TF-IDF architecture perfectly fits to the training data, however, the logistic regression model performs at a 2.4% higher accuracy than the MLP model. In the following section, we turn to dropout regularization to close the performance gap between MLP and logistic regression.

The Impact of Dropout on MLP Predictions

Our initial experiment applying a Deep MLP architecture to TF-IDF representations of house descriptions achieved a perfect 100% training accuracy. However, this failed to generalize to the held-out test set, only achieving 76.2% test accuracy. Table 2 illustrates the benefit of applying dropout regularization to the MLP model. Dropout describes randomly zeroing out either inputs or intermediate activations, depending on where in the neural network architecture the dropout layer is placed. Dropout is used to regularize deep learning models and avoid overfitting such as what we have observed with 100% training accuracy and 76.2% testing accuracy. We utilize a high level of dropout at 75% placed at the input and in between every hidden layer of our 3-layer MLP network. This results in an increase in the test accuracy from 76.2% to 78.6% when processing TF-IDF data representations. Dropout similarly increases the tabular MLP from 71.4% to 73.8% test accuracy. Although dropout decreases the train-test generalization gap, we still have 100% training accuracy. We think this illustrates an opportunity for future improvement by exploring additional regularizations for deep neural networks, such as Data Augmentation (Shorten et al. 2021, Shorten and Khoshgoftaar 2019).

Dropout %	0	75
TF-IDF Training Set	100%	100%
TF-IDF Testing Set	76.2%	78.6%
Tabular Training Set	87.5%	82.1%
Tabular Testing Set	71.4%	73.8%

Table 2: An illustration of the benefit of dropout regularization for deep MLP performance.

TF-IDF versus Bag-of-Words

In addition to TF-IDF representations of house descriptions, we also tested Bag-of-Words (BoW) representations. BoW describes representing text inputs as a sequence of token indexes. This limits the length of the input sequences compared to TF-IDF in which the input is the size of the entire vocabulary and each index is populated by the term frequency in that particular instance. We find better performance representing text as TF-IDF vectors, compared to BoW vectors. When training an MLP on BoW vectors without dropout it quickly overfits the training set achieving 100% training accuracy and a maximum entropy 50% test accuracy. Table 1 reports an MLP with dropout layers placed between every layer for both BoW and TF-IDF inputs.

Discussion

Multimodal Data Fusion

These experiments highlight the unimodal performance of modeling tabular and text data sources. In future work, we aim to test a fusion of models processing both tabular and text data sources. A simple way to extend our work to multimodal structured and unstructured prediction would be to have a linear weighting between each model prediction, such as alpha * tabular prediction + (1 - alpha) * text prediction. This describes a simple ensembling technique where we treat each model as a black box prediction. Another approach we intend to explore in future work is to combine the intermediate representations of each data type in our MLP architecture. This is known as intermediate fusion in the literature on multimodal architecture design for deep learning.

Data Splitting for Generalization Tests

In future work, we additionally intend to test the procedure in which we evaluate the generalization of our house price prediction models. These experiments use the independent and identically distributed (i.i.d.) procedure of randomly sampling train and test instances from a pool of data drawn from the same underlying distribution. In future work, we intend to isolate domains in which the data is sourced from. For example, training on house descriptions in Boston and Cambridge and evaluating on Boca Raton and Fort Lauderdale. We believe these kinds of train-test splits will provide more insight as to how well these models perform. For example, text descriptions of houses in Florida are likely to be very different from houses in Alaska.

Conclusion

Transparent pricing in the real estate market will lead to a more efficient economy. It allows agents to make informed decisions about trading decisions. Our experiment shows the impact of using structured and unstructured data for house price predictions. The final results conveyed that TF-IDF representations of text data perform better than tabular features. We believe this is just scratching the surface of natural language processing techniques to process this text data. Further, we intend to explore combining structured and unstructured data sources for house price prediction.

Acknowledgments

We acknowledge partial support by the NSF NRT-HDR (2021585). Opinions, findings, conclusions, or recommendations in this paper are the authors' and do not reflect the views of the NSF.

References

Abdallah, S., Khashan, D. A. 2016. Using Text Mining to Analyze Real Estate Classifieds. In International Conference on Advanced Intelligent Systems and Informatics.

Ahmed, E. H., Moustafa, M. N. 2016. House Price Estimation from Visual and Textual Features. ArXiv:1609.08399. In NCTA, 8th International Conference on Neural Computation Theory and Applications.

Bency, A., Rallapalli, S., Ganti, R. K., Srivatsa, M., Manjunath, B. S. 2017. Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 320-329.

Cheng, B., Yang, F., Xiang, S., Liu, J. 2021. Financial time series forecasting with tern Recognition, Volume 121, 2022.

Cypher, M., Price, S. M., Robinson, S., Seiler, M. 2017. Price Signals and Uncertainty in Commercial Real Estate Transactions. In Journal of Real Estate Finance and Economics, Forthcoming.

Gao L., Biderman S., Black S., Golding L., et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. In arXiv:2101.00027.

Hausler, J., Ruscheinsky, J., Lang, M. 2018. News-Based Sentiment Analysis in Real Estate: a Machine Learning Approach. In Journal of Property Research Volume 35, Issue 4, pages 344-371.

Johnson J. M., Khoshgoftaar T. M. 2019. Survey on deep learning with class imbalance. In Journal of Big Data, volume 6, Article number 27.

Law, S., Paige, B., Russell, C. 2019. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. ArXiv: 1807.07155. In ACM Transactions on Intelligent Systems and Technology 10(5):1-19.

Pancerz, K., Grochowalski, P. 2017. From unstructured data included in real-estate listings to information systems over ontological graphs. In International Conference on Information and Digital Technologies (IDT), 2017, pp. 298-303.

Spence, A. M. 2002. Signaling in Retrospect and the Informational Structure of Markets. In American Economic Review, Vol. 92, No. 3, pages 434-459.

Shorten, C., Khoshgoftaar M. T., Furht B. 2021. Text Data Augmentation for Deep Learning. In Journal of Big Data.

Shorten, C., Khoshgoftaar M. T. A survey on Image Data Augmentation for Deep Learning. In Journal of Big Data.

Velthorst, M., Guven, C. 2019. Predicting Housing Market Trends Using Twitter Data. In 2019 6th Swiss Conference on Data Science (SDS).