End-user-friendly XAI in house price prediction

Shuyue Wang

The University of Texas RGV exploreCSR 2020-2021

Motivation

How to buy best house with least money? This is one of the most important as well as most difficult questions people may encounter through their lifetime. Though artificial intelligence has been widely used nowadays, many AI technologies serve on generating merely a direct prediction result instead of a reliable and explainable one. Meanwhile, plenty of the in-use and newly proposed models are opaque black boxes with little trust.

Sometimes only the prediction result is what people want, however, when it comes to some fields regarding with mortal or money, as the question raised at the beginning, not only the result data but also an as comprehensive as possible explanation to the prediction result is needed. Therefore, eXplainable Artificial Intelligence (XAI) has been much more important, especially in the fields of from medical health to finance.

My research is related to XAI in house price prediction tasks. Though many works in XAI took house price prediction task as an example for following two reasons:

House price dataset has strong explainability itself: the features are readable and human friendly. Therefore, it is easy for researchers to evaluate whether their explanation is valid or not and to demonstrate their methods to readers lack of prior knowledge.

House price dataset has solid practical meaning: the data are from real life. By interpreting predictions on house price, a clearer guide can be set up to direct buyers on how to pick a more cost-effective real estate and to show sellers what kinds of houses are welcomed.

However, they lack in-depth discussions on giving specific explanations on questions people concern most. My work will focus on leveraging proper XAI methods to answer several house prise questions in a user friendly way.

Objective

The objective of my work will be separated into three parts:

1. To summarize current works done for explaining house price predictions.
2. To extract issues people concern most in the topic of house price.
3. To leverage proper latest methods to handle the issues in a user-friendly way.

Related Works

The table below shows some of the related works and the datasets they used.

Related Works	Authors	Title	Year	Dataset
XAI methods taking house price prediction task as example	Rishabh Agarwal et al.	Neural Additive Models: Interpretable Machine Learning with Neural Nets.	2020	California Housing dataset, a canonical machine learning dataset derived from the 1990 U.S. census
	Xingyu Zhao et al.	BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations.	2021	Boston house price dataset
User-friendly XAI methods	Maximilian Fö rster et al.	Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations	2021	44,957 houses in Germany offered for sale on a popular online platform
	Fred Hohman et al.	Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models	2019	The price of 506 houses in Boston, Massachusetts, based on 13 features; The price of 1,119 houses in Ames, Iowa, based on 36 features
XAI focusing on real estate market	F. Lorenz et al.	Interpretable Machine Learning for Real Estate Market Analysis	2021	52,966 observations of residential rents in Frankfurt am Main, Germany
	Piotr Grązka et al.	XAI Stories: Case studies for eXplainable Artifi-cial Intelligence	2020	19 house features plus the price and the id columns, along with 21613 observations of House Sales in King County, USA. Original dataset is from Kaggle.

Fig. 2: Some of the related works in recent years. These works are divided into 3 parts which are relevant to our work.

Among the works, a visual XAI analytic system called Gamut gives me inspiration. Figure 3 exhibits the system's user interface, where the data and visualized explanations are shown directly. However, the target audience for this system are not those with little statistics knowledge like house buyers and sellers.

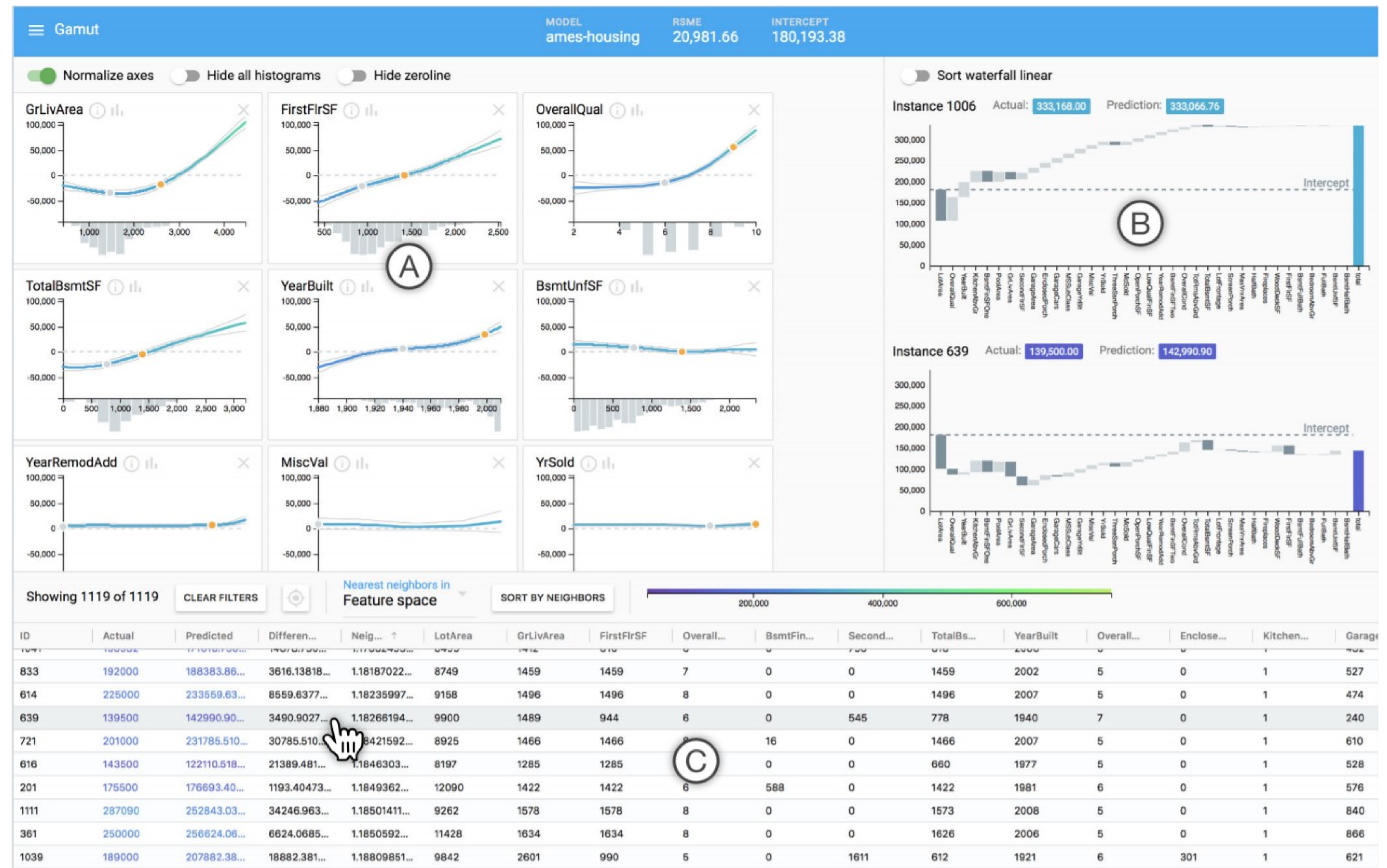

Fig. 3: The Gamut user interface.

Proposed Method

1. Fit data into different models and select most accurate ones.
   - For feature based explanations, regard it as a regression task.
   - For counterfactual explanations, regard it as a classification task by classifying the prices to 10 levels.
2. Since the prediction model is not determined, model-agnostic explanation method should be used.
3. To create user target explanations under house sales price context, I propose to use natural language as well as visualized explanations as the output.
   - For visualized explanations, a user-friendly interface should be used to show the global explanations as Fig 3 did. The visualization methods I will use include PDP and GAM.
   - For natural language explanations:
     - First collect which issues people concern most when buying or selling a house.
     - Then use feature based explanations (basic Shapley and advanced Shapley) or counterfactual explanations according to these specified issues.
     - A chat bot can be used to show local explanations, especially explanations for instances, because users concern more about issues like comparisons between houses. It will catch the key words in user's message and give a specified explanation according to the user's question.

Dataset

The original dataset I propose to use is the House Sales in King County, US dataset from kaggle.

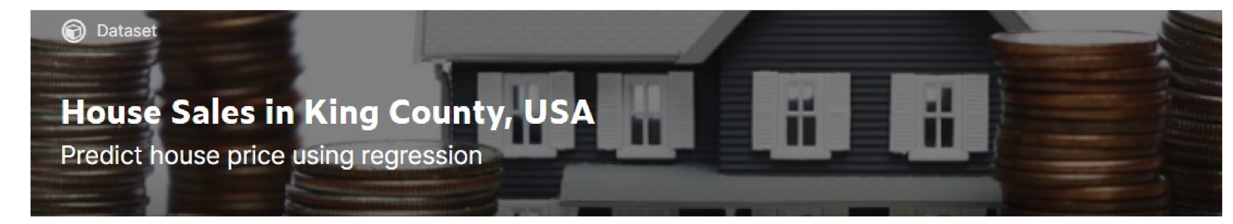

Fig. 4: The original dataset is fromm Kaggle, which contains 19 valid features.

Besides from the original data, we also refer to a case study which introduces some external data that are helpful for both prediction and explanation. The after-processed dataset contains the basic information of a house sale, spatial information (**distance to the nearest bus or subway stop**) and neighborhood condition (**number of cultural facilities nearby**).

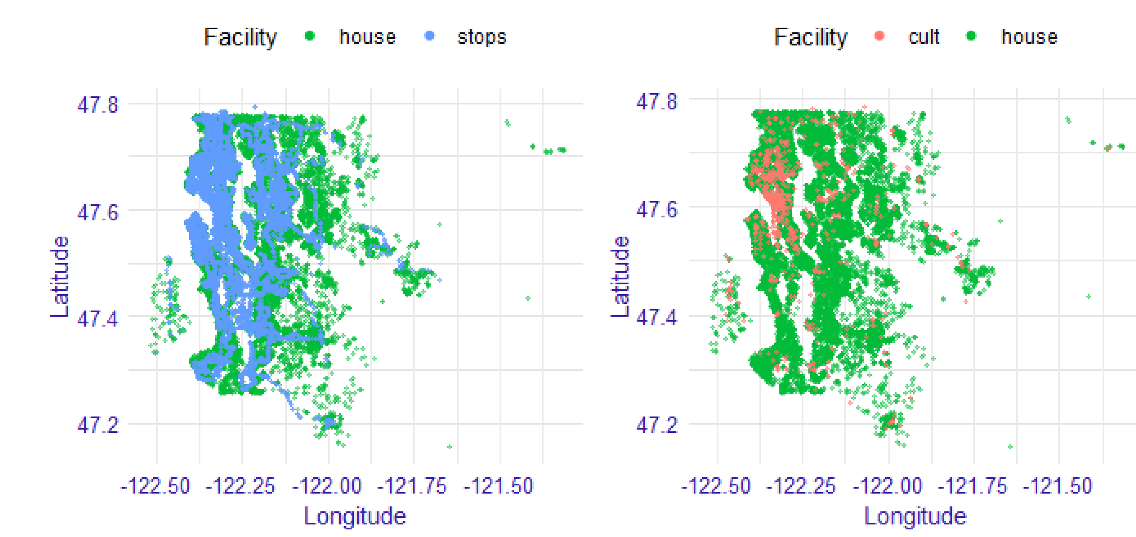

Fig. 5: This figure shows the locations of houses, bus stops and cultural facilities. These features will enrich the explanations.

Limitations and Future work

- The trend of the price of specified house is a time series. Therefore, models like RNN can be selected for better prediction and explanation methods specified to time series data can be applied based on this property.

  However, the dataset which meets the requirement that shows prices trends for different houses is lacked. This can be the future work if such dataset has been found.
- I also found that other than text descriptions, pictures of the house on sell will be attached if they sell online, which changes the problem into an image recognition problem. This can enrich the explanations.

Acknowledgement

Thanks for Prof. Beiyu Lin who offers me the opportunity and ideas for doing this work.