WB

A PREPRINT

Kacper Grzymkowski
Faculty of Mathematics and Information Science
Warsaw University of Technology
Poland
kacper.grzymkowski.stud@pw.edu.pl

Dominik Kędzierski
Faculty of Mathematics and Information Science
Warsaw University of Technology
Poland
dominik.kedzierski.stud@pw.edu.pl

Jakub Piwko
Faculty of Mathematics and Information Science
Warsaw University of Technology
Poland
jakub.piwko.stud@pw.edu.pl

March 17, 2022

ABSTRACT

...

Keywords Machine Learning · Explainable Artificial Intelligence · Real estate · Property appraisal

1 Introduction

...

2 Data

2.1 Data Description

Data analysed in our article was originally collected by Dean De Cock for use in data science education. It is publicly available through an advanced regression techniques machine learning competition on kaggle.com. The data describes 2919 dwellings located in Ames, Ihoa, United States. It is split between a training set of 1460 observations and a test set of 1459 observations. There are 79 feature columns describing various aspects of the dwelling such as its lot area, quality of amenities or the distance to the nearest railroad. The label is the sale price of the real estate.

2.2 Conclusions from Exploratory Data Analysis

Early analysis found that there are significant counts of missing values within the dataset. However, upon further inspection, those missing values were tied closely to the structure of the data. Certain features, such as *Basement Quality* or *Swimming Pool Quality* do not make sense when those features are missing. As such, we decided to impute the missing data by creating new categories representing a missing feature.

What is interesting in our data set is presence of many attributes that hold some kind of rating. For example, *Overall Quality* provides information about the overall quality of dwelling. We could not find a concrete methodology with which this data was collected. It is possible for these ratings to be subjective. Additionally some of these variables were in a categorical format, and some already encoded into an integer format. We categorized all feature columns into 47 discrete, 16 ordered discrete, and 16 continuous features.

Next step was analysing some one-dimensional and multi-dimensional dependencies by visualising them in histograms, box plots and scatter plots. As a result, we found some variables that seem to impact appraisal of real estate. For example, built year, type of foundation and area showed some relation with price of dwellings, so we hope that these columns will be significant in predicting prices. Exploratory Analysis outcome has also implied that in some columns there is a huge dominance of one value, while other types of observations are negligible. There are also variables related to extra features like pool or tennis courts and only a small part of dwellings presented in frame own them. This is a hint to encode this kind of columns in a different way or just not pay that much attention to them in a process. Our final stage was to look more broadly at continuous variables, so we created correlation heatmap. We haven't found any particularly high correlations between variables. But we observed that negative correlations are being outnumbered, which can be related to specification of our data.

References

Cankun Wei, Meichen Fu, Li Wang, Hanbing Yang, Feng Tang, and Yuqing Xiong. The research development of hedonic price model-based real estate appraisal in the era of big data. *Land*, 11(3):334, February 2022. doi:10.3390/land11030334. URL `https://doi.org/10.3390/land11030334`.

Geoffrey K. Turnbull and Arno J. van der Vlist. After the boom: Transitory and legacy effects of foreclosures. *The Journal of Real Estate Finance and Economics*, February 2022. doi:10.1007/s11146-021-09882-w. URL `https://doi.org/10.1007/s11146-021-09882-w`.

Alex van de Minne, Marc Francke, and David Geltner. Forecasting US commercial property price indexes using dynamic factor models. *Journal of Real Estate Research*, 44(1):29–55, December 2021. doi:10.1080/08965803.2020.1840802. URL `https://doi.org/10.1080/08965803.2020.1840802`.

Dieudonné Tchuente and Serge Nyawa. Real estate price estimation in french cities using geocoding and machine learning. *Annals of Operations Research*, February 2021. doi:10.1007/s10479-021-03932-5. URL `https://doi.org/10.1007/s10479-021-03932-5`.

Víctor Hugo Masías, Mauricio Valle, Fernando Crespo, Ricardo Crespo, Augusto Vargas Schüler, and Sigifredo Laengle. Property valuation using machine learning algorithms: A study in a metropolitan-area of chile. 01 2016.