House Prices

Henrik Vallestad, Markus Alvestad Nedrevold, Ole August Solem, 17.10.2023
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data?select=train.csv

DESCRIBE THE PROBLEM

SCOPE

Business Objective:

Our goal is to predict housing prices in Iowa. Based on the Ames Data Set from Dean De Cock.

How will the solution produced in the project be used?

The solution produced can be used in a website where people with houses in Iowa can insert the specifications of their house and find out with a certain accuracy what the house is worth. This can be done both for house owners selling a house, but also home buyers to avoid being scammed. A variety of similar projects exist on the market for both property and house prediction. Without the machine learning model, one would typically hire a realtor to help find a good price for the house and find relevant customers.

Perfomance measure:

We will measure our performance on the business metric Root-Mean-Squared-Error between the logarithm of the predicted value and the logarithm of the observed sales price.

System pipeline:

Our pipeline consists of the following components

1. Data Retrieval:

We first get the Iowa housing dataset from Kaggle, we then upload it to our personal google drive's so we can use it our jupyter notebook.

2. Data cleaning:

First, we iterate through all the columns in both the training and test data to find all columns containing NaN (not a number). We then replace all instances of NaN with None. Next, we replace all char and string representations to a number, this is because we are

going to use a regression model, and regression models train from numbers and not string data.

3. Feature Engineering:

In this project we have not done much in terms of Feature engineering, after cleaning, preparing the data and categorizing it, we got a reasonable R2 score metric that we were happy with. If the goal was to create the best model or score as possible, then we would start looking at combining features and or normalizing features.

4. Modeling:

First, we split our training data into test and training sets using Sklearn train_test_split function. This function helps us creating instances we can train our data on to avoid generalisation.

Considering we are dealing with a task of predicting a housing price aka a number, and this number can be any number, then it becomes only logical to think that we are dealing with a regression task. A first step could be to implement a simple linear regression model, but we decided to forego this step, concluding that due to the dataset containing so many features and being advanced, that we would use a more advanced model. We tried with a model called XG Boost but ended up with using Sklearns random forest regressor. As mentioned in the feature engineering section, the model ended up giving satisfactory results from the beginning, meaning that we did not perform any form of minimizing our loss function. If it were unsatisfactory then we could attempt to minimize a loss function such the mean squared error to achieve better hyper parameters and fine tune our model.

5. Model Deployment:

We deployed our model using the Gradio framework. Gradio is a library where you can easily create an interactive UI to give data to the model and retrieve and show a prediction. Each section in our pipeline is dependent on the ones before. A change in data cleaning will mean a change for, feature engineering, modelling and model deployment.

Stakeholders:

Our stakeholders may include insurance companies, estate investors, or a state community.

Timeline:

15.Oct All Team Members finds a feasible data set and start working on it.

20. Oct Henrik finished feature engineering and setting up the core of the project.

21. Oct Ole expanded on the project by adding gradio functionality

24. Oct Henrik and Ole co-operated in data exploration and started on the report.

08. Nov Markus finishes the remaining parts of the document.

08. Nov All Team Members looks over the document, so that its ready for publication.

Computational resources:

To continue with this project, we need 5$ a month to host it on an online webserver with the recommended amount of GPU and CPU power. We do not plan to invest much time into maintaining this project, and it will therefore be minimal costs tied to maintenance.

METRICS

Mininal business metric:

We would consider a R^2 value of 0.85 a good model we got a r^2 value of 0.86 therefore we would consider this project a success

Machine Learning measurements:

To measure how well our system is performing we have decided to use mean square root error (MSRE) and R^2. Accuracy does not fit well here since this is a regression task with large numbers.

DATA

The dataset we have retrieved consists of structured data consisting of 79 variables, describing almost every aspect of residential homes in Ames, Iowa. Our dataset contains 1460 rows and 81 columns, this is by no means a large dataset, meaning our data can become quite generalized, but it has a very large number of features. We are also given a dataset from Kaggle containing test data, which will serve as out “ground truth” labels.

.

Considering the fact that anyone can upload a dataset and host a competition on Kaggle, there can be a certain question regarding ethical and privacy concerns. Kaggle has a usability score to a dataset, this shows both how well the data can be used and how it is regarding credibility. Our dataset was hosted by Kaggle themselves, and that provides a sense of credibility.

Data preparation:

Since we are deploying a regression model and our structured data consists of both numerical and non-numerical data, we will have to clean and prepare the data. Some values might create a larger impression that others for no reason other than containing a number, those values might have to be normalized so that they do not create false impressions to the model.

MODELING

We experimented with XGBoost for predicting housing prices in Iowa, but it did not yield satisfactory results. Therefore, we decided to use Random Forest Estimator to predict the housing prices in Iowa. The model predicted well so we decided to go with it. We plan to investigate prediction mistakes by visualizing the predictions.

DEPLOYMENT

The model will be deployed on a website (using Gradio), where the user can input the necessary data and obtain a fair valuation of a house. The website design was made partally with chatGPT, first we made a list of the datatypes in the dataset. Then we prompted GPT to do parts of the heavy lifting (generating Gradio input fields). Second we fine-tuned the design and fixed some bugs so it would be integrated into the rest of the design. The model's main objective is to deliver accurate property valuations. This could have a vast range of practical applications. Such as real estate investment, determining a reasonable cost for leasing property and aiding insurance companies with estimate house prices. Additionally, it would provide a reliable estimate for homeowners who may not have the financial means to hire a real estate agent. For monitoring we could do routine testing to make sure the model functions as intended. Furthermore, this process could also be somewhat automated assuming we have a way to receive up to date housing data in Iowa.

GPT USAGE

Our GPT or ai usage in this project has been more as advanced google search rather than task solver. Some examples of this were me asking GPT for more detail regarding machine learning pipelines because most of the top results from a google search were companies advertising their own Machine learning pipelines and doing a poor job at explaining the concept. Under is a link to the GPT prompt:

https://chat.openai.com/share/4ab12855-eabb-4f7e-bac9-26c2fc1f01e5

This prompt gave a short and sufficient answer to a machine learning pipeline regarding our project. However, you must always be critical of what GPT gives you, so I tried digging a little deeper into articles just so I could tell if GPT was feeding me wrong or correct information

As noted in the deployment section, we also used GPT to help with the Gradio design, being unfamiliar with a framework can be daunting, and looking at the documentation can be confusing at times, we therefore had GPT help us with the input fields. This helped us gain kickstart using the Gradio interface.

Prompt for Gradio layout:

https://chat.openai.com/share/14ab1ade-0290-417c-ac76-67096871ced6

REFERENCES

Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques

Article regarding Machine Learning pipelines

https://medium.com/@datasciencewizards/machine-learning-pipeline-what-it-is-why-it-matters-and-guide-to-building-it-2940d143fd37