BIG DATA ANALYTICS

PROJECT PROPOSAL

Price Prediction Using Hedonic Model

Vincent ATTIA | Liang LI | Gokul Krishnan RAMAKRISHNAN | Yu WANG
Under the supervision of Olga KLOPP

MOTIVATION AND PROBLEM DEFINITION

Most people must have concepted his or her dream house in the future. So did we. When talking about the price of a house, it is important to know how it is calculated. This project with data from Kaggle, provides us an excellent chance to discover the insight of house pricing and allows us to put into practice what we have learned from the Big Data Analytics class. Based on historical sales data of the housing market of the city of Ames in Iowa, we are going to predict the price for other houses in the same city which were sold in the same period of time.

RELATED WORK

The Boston Housing data set is a widely studied data set and relates a lot to this project.
The dataset used here was released a few years ago on Kaggle and some contributors have already shared their results. We plan on having an independent approach and then compare our results to existing ones.

DATA SET

This is a competition from Kaggle, hence we have the dataset:

kaggle.com/c/house-prices-advanced-regression-techniques

Data Set: 1461x81
Column names and 1060 houses: a unique ID, price sold and 79 explanatory variables

	Numerical	Categorial Variables	Time-related
Physical characteristics about the house	26 Variables	39 Variables.	YearBuilt, YearRemodAdd, GarageYrBlt, MoSold, YrSold
Information about the neighborhood	LotFrontage, LotArea	MSZoning, Street, LandContour, Neighborhood, Condition1, Condition2, LandSlope	N/A

*The classification of variables into house physical parameters and neighborhood-related are based on whether the variable has a direct effect or an indirect effect on the price.

METHODOLOGY

Exploratory Data Analysis (EDA)

We use basic univariate statistical analysis to gain an overview of the variables in the dataset. This is supplemented with the help of some data visualization tools such as histograms, scatter plots, etc.

Data Washing/Cleansing

- Missing data: Identification of mechanism of missing data (missing at random, missing completely at random, missing not at random), then decide how to deal with it.
  Few of the following several methods would be considered (whichever applies):
  - Deletion, if too few observations
  - Mean/Mode/Median/Random sample imputation
  - Multiple imputation
  - Regression (linear / logistic)
- Check whether there is abnormal data or not:
  - Outliers
  - High-leverage points

  In a multiple-variable dataset, these might be more difficult to identify.

Data Transformation

As shown in the table above, we have quite a lot of categorical variables (39 to be precise). As for nominal variables, we would be using dummy method, and as for ordinal variable, we would assign an appropriate number to each term. The dilemma for this data set is that it already contains 80 variables. After implementing dummy variables to exploit the categorical ones, the number of variables may reach 200+.

Size Reduction On Variables

The model with dummies would subsequently hold 200+ variables. Variables may also show high multicollinearity between them that demands a different approach to the problem of regression.

To ensure a good flexibility, but without being too strong, we will need to reduce the total number of variables. For instance, we can use:

- Check the inner logical relationship between variables: e.g. 1st floor surface + 2st floor surface = total surface
- PCA to reduce the number of variables that show a high correlation, without losing too much of the information.

- Lasso to choose the most relevant variables.

Model Selection

- Y = House Price as our estimate. Then compute a linear regression model.
- Intuitively, one might compute the price per square meters for evaluating a house, which means here we could use Y / X (Y - House Price, X - Surface) as our new estimate Y'.

The Hedonic Model is applied to make the prediction

Price = f(physical characteristic, other factors),

Physical characteristic includes square, bathrooms, basement, pool, location, age, etc. The hedonic model is usually in semi-log form with the natural log of price used as the dependent variables. After considering the neighbourhood effect, the model can be represented as:

$$\log price = \alpha + \sum(\beta_k S_k) + \gamma NQ + \varepsilon$$

(if neighbourhood effects are treated as direct determinants of house values)

or

$$\log price = \alpha + \sum(\beta_{k_0} + \beta_{k_1} NQ)S_k + \gamma NQ + \varepsilon$$

(if neighbourhood characteristics are regarded as determinants of spatial drift in the structural parameters)

$price$ - vectors of observed housing values
S_k - vectors of structural characteristics (k = 1,…, K)
NQ - neighbourhood quality scores
β, γ, α- parameter vectors

Validation Method

We will use the traditional method of testing the algorithm on a cross-validation set (test set). To ensure the robustness of the result, we will use the concept developed by the leave-one-out cross-validation technique. To optimize the computation time, we will aim at a k-fold cross-validation method, using k = 10.

References

[1] Can, A. (1992). Specification and estimation of hedonic housing price models. Regional science and urban economics, 22(3), 453-474.

[2] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.

[3] Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. Journal of real estate literature, 13(1), 1-44.