Predicting Housing Prices - Project Proposal

Sabarinath Suriyamurthy (ss120) , Sudha Natarajan (Sudha2), Raghav Rama (RR26)

7/6/2021

Contents

Introduction

We propose to create a linear model that can predict residential home prices in Melbourne (Australia) based on several explanatory variables. This project is a practical example of using real world data that consists of a mix of different data types - nominal, ordinal, discrete, and continuous variables.

Our final project will incorporate the following topics covered in STAT420 Course:

- Data cleansing
- Variable manipulation
- Outlier identification
- Data analysis and interpretation
- Model building
- Model evaluation

Dataset

We're using the Melbourne Housing Prices dataset from the following Kaggle site: https://www.kaggle.com/anthonypino/melbourne-housing-market. This data pertains to the houses found in a given Melbourne (coastal capital of the southeastern Australian state of Victoria) area containing data from year 2016 to 2018.

Data Snippet

Here is a snippet of data with only first 10 columns considered.

Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode
Abbotsford	68 Studley St	2	h	NA	SS	Jellis	3/09/2016	2.5	3067
Abbotsford	85 Turner St	2	h	1480000	S	Biggin	3/12/2016	2.5	3067
Abbotsford	25 Bloomburg St	2	h	1035000	S	Biggin	4/02/2016	2.5	3067
Abbotsford	18/659 Victoria St	3	u	NA	VB	Rounds	4/02/2016	2.5	3067
Abbotsford	5 Charles St	3	h	1465000	SP	Biggin	4/03/2017	2.5	3067
Abbotsford	40 Federation La	3	h	850000	PI	Biggin	4/03/2017	2.5	3067

Data Description

The raw dataset contains **34857** observations and **21** variables (X predictors and X response). Below is a list of few variables with descriptions taken from the original Kaggle site given above.

- **Rooms**: Number of rooms
- **Price**: Price in dollars
- **Method**: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
- **Type**: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
- **SellerG**: Real Estate Agent
- **Date**: Date sold
- **Distance**: Distance from CBD
- **Suburb**: Suburb
- **Address**: House Address
- **Regionname**: General Region (West, North West, North, North east ... etc)
- **Propertycount**: Number of properties that exist in the suburb.
- **Bedroom2**: # of Bedrooms
- **Bathroom**: Number of Bathrooms
- **Car**: Number of carspots
- **Landsize**: Land Size
- **BuildingArea**: Building Size
- **CouncilArea**: Governing council for the area

Requested Criteria

- A minimum 2000 observations
- At least 10 variables
- A numeric response variable - `Price`
- At least one categorical predictor- `Type`
- At least two continuous numeric predictors - `Landsize & Propertycount`

Motivation

- Hands on experince with real life datasets.
- Practice with all techniques learnt in STAT420 course.
- Discover how applied statistics can help us answer Housing price prediction.

Credits

- Anthony Pino - Melbourne Housing Dataset