

ETC5242 - Project 2017

How much will the property sell for at auction?

Task

Housing prices in Melbourne are some of the highest in the world. We have collected several years worth of data from Melbourne, by scraping pdf reports from weekly auctions. The data contains sales prices of houses along with other characteristics of the houses:

This is a description of the variables:

Variable	Description
id	unique id for property
suburb	suburb location of property
price	Price house sold for in AUD dollars, divided by 100,000
result	S indicates 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PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; o res - other residential; w - withdrawn prior to auction
agent	realtor in charge of sale
nbeds	Number of bedrooms
property type	h =house, t =townhouse, u =unit/apartment
day	day of the month of auction
month	month of auction
year	year of auction
nvisits	How many people came to open houses
ncars	Number of parking places
nbaths	Number of bathrooms
land size	Size of the lot, in sq m, units will be 0
house size	Internal size of property in sq m

The purpose of the project is to make the best model for sales price that you can. 50% of the data is made available to you. The other 50% is reserved - you will get the explanatory variables but not the sales price of the properties for these. You need to use your model to predict the sold price. You can test your predictions by submitting them to the kaggle class site. You'll get the error associated with your predictions as measured by mean absolute error:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Your first task is to create a kaggle account, for you as an individual.
- Do some basic exploration of the housing data.
- Build your first model. Predict the test set, and upload your predictions to kaggle.
- Try, and try again to improve your model. You can do one prediction per day.
- Merge with other students to form a team of no more than 4 people. Merging models can give you better results, so there is an advantage to merging with other students, but don't do this too early.
- Write up how you (and your team) built your model, and decided on your best model. Also describe one or more other interesting things you learned about housing price relative to the other variables. One report per team.
- Turn your report into a 10 minute presentation for the class.

Deadlines:

- Sep 15, 6pm: Have made at least one submission to kaggle. The competition is called vitticeps (do you what this is?) and can be found at <https://inclass.kaggle.com/c/vitticeps>.
- Oct 13: Turn in your project report (max 5 pages), to ED, one per group, that describes your model fitting, and at least one interesting observation about the housing data. Due by 6pm, and kaggle competition closes Oct 12.
- Oct 16, 20: Present your project in the lecture period. All members must attend and present part of the work. All students must be present to evaluate the presentations, and if not points will be deducted from the absent individual's score.

Grading:

- Total points: 15
- Accuracy of classifier: 4 (Team with lowest error will get 4 points, second best team will earn 3.5 points, third gets 3 points, and then 0.1 less for each additional place.)
- Report: 4
- Presentation: 5 (Score will be given by other members of the class. All members of the team must participate by speaking in the presentation.)
- Met deadlines: 2