ETC5242 - Project

How much is the house worth?

Task

Housing prices in Melbourne are some of the highest in the world. We don't have housing prices for Melbourne, but we have collected three years worth from Ames, Iowa, USA. The data contains sales prices of houses along with other characteristics of the houses:

This is a description of the variables:

Variable	Description
SalePrice	Price house sold for in US dollars
Neighborhood	Different areas of town
LotArea	Area of the land in square feet
YrBuilt	Year built
HouseStyle	"1.5 Fin"=two stories on one side, one story on the other and a finished basement, ...
Foundation	The material was the house built on.
RoofMatl	Most roofs are composite shingles (CompShg)
Ext1	Material of the house exterior, e.g. brick, vinyl siding, ...
Heating	Most houses are heated using Gas
Central Air	Yes means house has central air conditioning
TtlBsmtSF	Size of the basement in square feet
TotRmsAbvGrd	Number of rooms above ground
GarageType	Type of garage
Cars	How many cars fit in the garage
GarageArea	Size of garage in square feet
NmbrBRs	Number of bedrooms
YrSold YYYY	Year sold
MoSold MM	Month sold
SaleDate	Date of sale, day is set to the first of the month for all
id	Unique id for each house

The purpose of the project is to make the best model for sales price that you can. 75% of the data is made available to you. The other 25% is reserved - you will get the explanatory variables but not the sales price of the houses for these. You need to use your model to predict the sales price. You can test your predictions by submitting them to the kaggle class site. You'll get the error assciated with your predictions as measured by mean absolute error:

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

a. Your first task is to form a team. Your choice, ideally the team has three people, for any other number you need permission from the instructor.

b. Create a kaggle account, for each team member and form a team.

c. Do some basic exploration of the housing data.

d. Build your first model. Predict the test set, and upload your predictions to kaggle.

e. Try, and try again to improve your model. You can do one prediction per day.

f. Write up how you 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.

g. Turn your report into a 5 minute presentation for the class.

Deadlines:

- Sep 15: Form your team, the names and your team name need to be emailed to the instructor by noon.
- Sep 27: At least one kaggle submission needs to have been made. The competition is called galah (Good Analysis of Listings of Ames Houses) and can be found at https://inclass.kaggle.com/c/galah.
- Oct 13: Upload a two page project report to moodle, one per group, that describes your model fitting, and at least one interesting observation about the housing data. Due by 3pm, and kaggle competition closes earlier in hte day at 7am.
- Oct 20-21: Present your project in the lecture period

Grading:

- Total points: 10
- Accuracy of classifier: 3 (Team with lowest error will get 3 points, second best team will earn 2.5 points, third gets 2 points, and then 0.1 less for each additional place.)
- Report: 3
- Presentation: 3 (Score will be given by other members of the class. All members of the team must participate by speaking in the presentation.)
- Met deadlines: 1