Modelling Competition: House prices

House pricing models

The characteristics of a house strongly affect the sale price. The direction of the effect of most of these characteristics is obvious: larger houses sell for more than smaller houses, houses with more bedrooms and bathrooms sell for more than houses with fewer bedrooms than bathrooms, the presence of a garage tends to raise the sale price of a house, and so on. On the other hand, these features by no means provide a perfect prediction of the sale price of a house. In part, this is because some features are not recorded systematically by real estate agents. For example, whether there is a busy and noisy street in front of the house will matter, whether the house has been kept in good repair is important, and so on. Location usually quite important. For example, the location of the house determines what public school any children living there will attend, and some schools are regarded as better than other schools. Even if there were a complete list of all the features there still would not be a perfect description of the price, because the price is also affected by who buys the house. Houses and buyers are all different. If a family looks at a house that has just been advertised for sale and that house is “just right” for them, it may well sell for the list price or even a little higher. On the other hand, if the house has been for sale for a long time and a family sees that the house will meet their needs only with some changes then it may well sell for quite a bit less than the list price.  
Sending someone around to look at each house each year and estimate its value is expensive. It is much cheaper to look at the record of the characteristics of each house, and then use a pricing model to estimate the value of the house.

Aim

1. Create an effective price prediction model
2. Identify the important home price attributes
3. Validate the model’s prediction accuracy

Data

The data set collects every aspect of almost 3000 residential homes in Ames, Iowa, as well as their sale price.  
This are variables:

* **SalePrice**: the property’s sale price in dollars. This is the target variable that you’re trying to predict.
* **MSSubClass**: The building class CATEGORICAL(NOMINAL)
* **MSZoning**: The general zoning classification CATEGORICAL(NOMINAL)
* **LotFrontage**: Linear feet of street connected to property CONTINUOUS
* **LotArea**: Lot (terreno) size in square feet CONTINUOUS
* **Street**: Type of road access CATEGORICAL(NOMINAL)
* **Alley**: Type of alley Access CATEGORICAL(NOMINAL) - BINARY
* **LotShape**: General shape of property ORDINAL
* **LandContour**: Flatness of the property ORDINAL
* **Utilities**: Type of utilities available ORDINAL
* **LotConfig**: Lot configuration CATEGORICAL(NOMINAL)
* **LandSlope**: Slope of property ORDINAL
* **Neighborhood**: Physical locations within Ames city limits CATEGORICAL(NOMINAL)
* **Condition1**: Proximity to main road or railroad CATEGORICAL(NOMINAL)
* **Condition2**: Proximity to main road or railroad (if a second is present) CATEGORICAL(NOMINAL)
* **BldgType**: Type of dwelling CATEGORICAL(NOMINAL)
* **HouseStyle**: Style of dwelling CATEGORICAL(NOMINAL)

**HASTA AQUI TODAS LAS CATEGORICAS NO SON NUMEROS**

* **OverallQual**: Overall material and finish quality ORDINAL (con numeros)
* **OverallCond**: Overall condition rating ORDINAL (con numeros)
* **YearBuilt**: Original construction date CONTINUOUS
* **YearRemodAdd**: Remodel date CONTINUOUS
* **RoofStyle**: Type of roof CATEGORICAL(NOMINAL)
* **RoofMatl**: Roof material CATEGORICAL(NOMINAL)
* **Exterior1st**: Exterior covering on house CATEGORICAL(NOMINAL)
* **Exterior2nd**: Exterior covering on house (if more than one material) CATEGORICAL(NOMINAL)
* **MasVnrType**: Masonry veneer type (TIPO DE REVESTIMIENTO DE LA ALBAÑILERIA) CATEGORICAL(NOMINAL)
* **MasVnrArea**: Masonry veneer area in square feet CONTINUOUS
* **ExterQual**: Exterior material quality ORDINAL
* **ExterCond**: Present condition of the material on the exterior ORDINAL
* **Foundation**: Type of foundation (CIMIENTOS) CATEGORICAL(NOMINAL)
* **BsmtQual**: Height of the basement ORDINAL
* **BsmtCond**: General condition of the basement ORDINAL
* **BsmtExposure**: Walkout or garden level basement walls ORDINAL
* **BsmtFinType1**: Quality of basement finished area ORDINAL
* **BsmtFinSF1**: Type 1 finished square feet CONTINUOUS
* **BsmtFinType2**: Quality of second finished area (if present) ORDINAL
* **BsmtFinSF2**: Type 2 finished square feet CONTINUOUS
* **BsmtUnfSF**: Unfinished square feet of basement area CONTINUOUS
* **TotalBsmtSF**: Total square feet of basement area CONTINUOUS
* **Heating**: Type of heating CATEGORICAL(NOMINAL)
* **HeatingQC**: Heating quality and condition ORDINAL
* **CentralAir**: Central air conditioning CATEGORICAL(NOMINAL) - BINARY
* **Electrical**: Electrical system CATEGORICAL(NOMINAL)
* **1stFlrSF**: First Floor square feet CONTINUOUS
* **2ndFlrSF**: Second floor square feet CONTINUOUS
* **LowQualFinSF**: Low quality finished square feet (all floors) CONTINUOUS
* **GrLivArea**: Above grade (ground) living area square feet CONTINUOUS
* **BsmtFullBath**: Basement full bathrooms DISCRETE
* **BsmtHalfBath**: Basement half bathrooms DISCRETE
* **FullBath**: Full bathrooms above grade DISCRETE
* **HalfBath**: Half baths above grade DISCRETE
* **Bedroom**: Number of bedrooms above basement level DISCRETE
* **Kitchen**: Number of kitchens DISCRETE
* **KitchenQual**: Kitchen quality ORDINAL
* **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms) DISCRETE
* **Functional**: Home functionality rating CATEGORICAL(NOMINAL)
* **Fireplaces**: Number of fireplaces DISCRETE
* **FireplaceQu**: Fireplace quality ORDINAL
* **GarageType**: Garage location CATEGORICAL(NOMINAL)
* **GarageYrBlt**: Year garage was built CONTINUOUS
* **GarageFinish**: Interior finish of the garage CATEGORICAL(NOMINAL)
* **GarageCars**: Size of garage in car capacity DISCRETE
* **GarageArea**: Size of garage in square feet CONTINUOUS
* **GarageQual**: Garage quality ORDINAL
* **GarageCond**: Garage condition ORDINAL
* **PavedDrive**: Paved driveway CATEGORICAL(NOMINAL)
* **WoodDeckSF**: Wood deck area in square feet CONTINUOUS
* **OpenPorchSF**: Open porch area in square feet CONTINUOUS
* **EnclosedPorch**: Enclosed porch area in square feet CONTINUOUS
* **3SsnPorch**: Three season porch area in square feet CONTINUOUS
* **ScreenPorch**: Screen porch area in square feet CONTINUOUS
* **PoolArea**: Pool area in square feet CONTINUOUS
* **PoolQC**: Pool quality ORDINAL
* **Fence**: Fence quality ORDINAL
* **MiscFeature**: Miscellaneous feature not covered in other categories CATEGORICAL(NOMINAL)
* **MiscVal**: $Value of miscellaneous feature CONTINUOUS
* **MoSold**: Month Sold DISCRETE
* **YrSold**: Year Sold CONTINUOUS
* **SaleType**: Type of sale CATEGORICAL(NOMINAL)
* **SaleCondition**: Condition of sale CATEGORICAL(NOMINAL)

Files

Three files are provided:

* data\_description.txt: full description of each column in the dataset
* train.csv: the training set
* test.csv: the test set

Task

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable (**work with the log of the Saleprice**). There are many models that could be set up, but you can use **only** the techniques studied in this course.

You must return a report (no longer than 15 pages) in .pdf containing:

1. Model-building steps (exploratory analysis, model fitting, variable selection, etc.).
2. Best model/models selected (There are more than one good model, so give reasons to justify your election).
3. Goodness of fit of the model within sample and predictive power of the model out of sample (in terms of the RMSE)
4. Other questions to be assessed:
   * What is the variable that contributes the most to explain the variability in price?, what are the characteristics of the cheapest/most expensive houses?
   * Any other question that you consider interesting

Evaluation

1. Clarity: The report must clearly indicate the steps you follow in your work, including which models are considered. Being able to communicate effectively is extremely important for a researcher. Excellent work is of no value if no other than the investigator can understand it. (20%)
2. Content: The analysis should use the tools developed in the course in an appropriate and correct manner. The report should anticipate questions that a critical reader might ask. (60%)
3. Presentation: The results will be presented in the last day of the course. The exposition and defense of the work will be evaluated (20%)

Winners of the competition

* The group reporting the smallest RMSE (less error) will have 1 extra points in the final mark of the course. The runner-up group 0.5