Housing data analysis - Women who code workshop

Yournamehere

13/03/2018

Load the data in from csv.

```
trainH <- read.csv("train.csv")
testH <- read.csv("test.csv")
```

What features are there in the data? What are the dimensions of the data? What are the column headers? Use the summary() and str() functions to explore...

What does the distribution of sale price look like?

Is the sale price (the variable we're interested in prediting) normally distributed? Find its mean, standard deviation, and plot a histogram of the distribution using ggplot2.

Plot a quantile-quantile plot (QQ plot) to "assess" normality. This plot compared the data we have (Sample Quantiles) with a theoretical sample coming from a normal distribution. Each point (x, y) corresponds to one of the quantiles of the second distribution (x-coordinate, theoretical) plotted against the same quantile of the first distribution (y-coordinate, our data). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

A standard way of transforming the data to be better approximated by a normal distribution is by using the log-transform?

Carry out this transformation and use a histogram and QQ plot to see whether it works...

Missing data

What happens if we only use complete data? How much data is missing?

We need to combine the datasets for imputation, so that we don't have NAs in the test data as well!

How do we impute the missing data?

Read the metadata file and see that many of the NAs should be recoded as None since these features are lacking in the house.

For the GarageYrBlt set to zero.

Lot frontage - set as median for the neighborhood.

Now split data again

Basic exploratory data analysis of training data

How does the sale price depend on living area: X1stFlrSF, X2ndFlrSF, TotalBsmtSF? Create a variable TotalSqFt which is a combination of these 3. Does it better predict the house price?

Identify and remove outliers with a high total square foot, but low price

Does having more bedrooms increase sale price?

Visualise both number of bedrooms (as a factor) and TotalSqFt as a scatterplot to see if a trend is visible.

Are newer or more recently renovated properties more expensive? Investigate this generally and then specifically for 2 - 4 bedroom properties.

Lets convert kitchen quality to numeric (we'll see why we need this later):

From the metadata we know it can be:

- Ex Excellent
- Gd Good
- TA Typical/Average
- Fa Fair
- Po Poor

Recode this to numeric values using mutate() and recode().

Convert Bldgtype to numeric

What variables are correlated with each other and with price? Plot a correlation plot using corrplot() for all numeric variables and those that show the top correlation with LogSalePrice.

Use the createDataPartition() function to separate the training data into a training and testing subset. Allocate 50% of the data to each class. Run set.seed(12) before this.

```
set.seed(12)
```

Fit a linear model considering the "top 10"" correlated (top 9, ignore LogSalePrice for obvious reasons).

Use predict() to predict house prices using our top10 model on the "test" portion of the training dataset. Use rmse to assess the root mean square error (our metric of accuracy).

Use randomForest to train a random forest model on all of the variables. Use predict and rmse to make the prediction and assess the accuracy respectively. Was a linear (on 9 features) or random forest model more accurate?

Use xgboost to predict house prices from numeric features of training dataset.

Use xgb.plot.importance() to assess which variables are most important for predicting house prices.

Use the glmnet library to train a ridge regression model. Is it more or less accurate than XGBoost?