Final Project for IS-6489

Introduction

The final project in IS-6489 is to participate in a Kaggle competition. This particular competition, House Prices, is in the playground section, and is strictly for fun and fame, not profit. (Some official Kaggle competitions have substantial prize money at stake.) Please see:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

This competition has already concluded. You will still be able to submit your results and receive a score but the leaderboard will not update with your ranking.

The competition consists in predicting house prices in Ames, IA. The data, which is described below, has been split into 50% train and 50% test sets at the above website (with 1460 and 1459 observations, respectively). The test set contains all the predictor variables found in the train set, but is missing the outcome variable, SalePrice. You will use the model you develop on the train set to make predictions for the test set and then submit your predictions at Kaggle. (You may make as many submissions as you like.) Your score will be based on the out-of-sample performance of your model. The competition tests your ability to develop a generalizable model with low variance.

Specifically, you will work singly or in teams of no more than three to model housing prices and document your results in (1) and interim report (no longer than 1 page) and (2) a final report (no longer than 5 pages of text).

The cool thing about Kaggle is that a large data science community works on these problems, so the forums are rich sources of techniques and ideas. My hope is that you will make full use of these resources. You are encouraged to contribute to the kernels associated with this competition.

Steps

1. Look around the Kaggle competition site, familiarizing yourself with the kernels and the structure of the competition. Make sure you understand how to submit your predictions to Kaggle. (See Kaggle for the required format of the .csv file you submit: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.)
2. Form groups of 1 - 3 people.
3. One group member should send me an email through Canvas with the names of team members.
4. Download the train data and begin exploratory data analysis. The dataset is complex; make sure you understand how the variables relate to one another and the structure and meaning of the missing observations. Think about how you might combine variables or create new ones.
5. Fit a parsimonious model to the data (five predictors), submit your predictions to Kaggle, and write up your results in a brief interim report.
6. Fit a full model to the data using as many variables as you'd like and submit your predictions to Kaggle. Write up your results in a client-ready report consisting in no more than 5 pages of text.

Requirements

Your overall objective in this project is to develop a model with the best predictive accuracy possible. How good is good? Your Kaggle score will help answer that question. For reference, IS-6489 groups have achieved log RMSE of as low as .11.

Interim report (5 variable model)

Length: no more than 1 page, single spaced, *including* graphs and tables. (Submit source code in a a separate document.)

Due Date: prior to class 7.

Description. For the interim report, develop a parsimonious model of housing prices using only 5 predictors. This limit on model terms will force you to balance predictive performance and simplicity and will ensure that you become familiar with the predictive characteristics of the variables in the dataset. (Note: minimum in-sample R^2 for this model is .8.)

Your interim report should (1) introduce the problem, (2) describe your model, and (3) report model performance, including:

- RMSE and R^2 on the train set
- *estimated* RMSE and R^2 on the test set
- your Kaggle score (returned log RMSE) and rank.

For further details check the grading rubric for this assignment at Canvas.

Documents. One member of your team should submit through Canvas on behalf of everyone: (1) a PDF of your interim report and (2) the source code documenting your process and results, consisting in a well-organized and clearly commented .R, .Rmd or PDF or HTML file of compiled .Rmd with echo = T in the code chunks. Please include the names of all the team members on the report.

Final report (full model)

Length: no more than 5 pages of text, single spaced, *excluding* plots and tables. . (Submit source code in a separate document.)

Due Date: one week after the final class meeting.

Description. For the final report, create a model using some or all of the available variables. In this case, there are no restrictions on variables. Aim for maximum predictive accuracy. Peruse the forums and the kernels at the competition site for ideas and code. Can you create new variables that enhance predictive accuracy?

Your final report should be written and formatted carefully, as if for a client. You should observe the best practices of statistical communication: use graphs when possible, labeling and explaining them, and interpret statistical results using language and quantities that non-statisticians can understand. Your report should (1) introduce the problem, (2) describe the data and any cleaning you did, (3) explain your model in detail (how you developed it, and how it differs from and improves upon the model you used for the interim report), and (4) report model performance, including:

- RMSE and R^2 on the train set
- *estimated* RMSE and R^2 on the test set
- your best Kaggle score (log RMSE) and rank.

For further details check the grading rubric for this assignment at Canvas.

Documents. One member of your team should submit through Canvas on behalf of everyone: (1) a PDF of your interim report and (2) the source code documenting your process and results, consisting in a well-organized and clearly commented .R, .Rmd or PDF or HTML file of compiled .Rmd with echo = T in the code chunks. Please include the names of all the team members on the report.

Data

The data, along with a detailed description, is here:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

There is both a train dataset (1460 observations) and a test dataset(1459 observations).

There are 81 variables in the train dataset, including the outcome variable, SalePrice, which is omitted from the test dataset.

Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property: When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?

The 20 *continuous* variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square footage found on most common home listings, other more specific variables are quantified. Area measurements on the basement, main living area, and even porches are broken down into individual categories based on quality and type.

The 14 *discrete* variables typically quantify the number of items occurring within the house. Most are specifically focused on the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Additionally, the garage capacity and construction/remodeling dates are recorded.

There are a large number of *categorical* variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being STREET (gravel or paved) and the largest being NEIGHBORHOOD (areas within the Ames city limits). The nominal variables typically identify various types of dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property.

There are two *location* variables: PID and NEIGHBORHOOD. PID is the Parcel Identification Number assigned to each property within the Ames Assessor's system. This number can be used in conjunction with the Assessor's Office (http://www.cityofames.org/assessor/) or Beacon (http://beacon.schneidercorp.com/) websites to directly view the records of a particular observation. The typical record will indicate the values for characteristics commonly quoted on most home flyers and will include a picture of the property. The NEIGHBORHOOD variable can be used with this map:

http://www.amstat.org/publications/jse/v19n3/decock/AmesResidential.pdf

Only the most recent sales data on any property has been kept in the data.

Grading

The interim report will be worth 15% of your course grade, and the final report will be worth 25% for a total of 40%.

Grading rubrics

The predictive performance of your models should be your primary concern. But you should also strive to present your work using the best practices of technical writing and statistical communication. Check Canvas for the rubrics that I'll use for grading both the interim and the final reports. These should guide your efforts.