MSDS 5371 Project Description (HW 13 and 14)

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

**Data and Description:**

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

**HOW TO KAGGLE VIDEO:**

**QOIs:** [**https://www.youtube.com/watch?v=0QJtczDPxZQ**](https://www.youtube.com/watch?v=0QJtczDPxZQ)

Your team’s objective is to conduct 2 analyses:

1. ANALYSIS 1: Assume that Century 21 in Ames Iowa has commissioned you to answer a very important question with respect to their business. Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to simply get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if it depends on which neighborhood the house is located in. Build a model that will answer this question keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. It turns out Century 21’s leadership team has a member that has some statistical background. Therefore, make sure and provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, of course provide your client with a well written conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods.

1. ANALYSIS 2: Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods. Your group is limited to only the techniques we have learned in 6371 (no random forests or other methods we have not yet covered). Specifically, you should produce 4 models: one from forward selection, one from backwards elimination, one from stepwise selection and one that you build custom. The custom model could be one of the three preceding models or one that you build by adding or subtracting variables at your will. Generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa. In your paper, please include a table similar to the one below. The group with the lowest public Kaggle score will receive an extra 3 bonus points on the final exam!

Quick note on Kaggle completion: We only have one course under our belts so far (almost), but you can compete in this competition with the tools you have now (top 40th percentile or less)! And, after your next course (6372), you will really be able to do well (top 25th percentile or less!). With these skills and the skills you pick up in Data Mining and Quantifying the World, you will be able to compete with anyone!

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | Adjusted R2 | CV PRESS | Kaggle Score |
| Forward | .89 | 1272 | .721 |
| Backward | .78 | 1590 | .945 |
| Stepwise | .81 | 2001 | .888 |
| CUSTOM | .87 | 900 | .2345 |

NOTE 1: ALL ANALYSES MUST BE DONE IN SAS and all code must be placed in the appendix. Part of the grading process will be to run the code and verify the Kaggle score for each group.

Note 2: An extra 3 points on the final exam will be awarded to the team with the model with the lowest (best) Kaggle Score. In the unlikely event of a tie will split these points.

**Deliverables:**

Your group is to turn in a paper that is no more that 7 pages long (without the appendix). Please put your code in the appendix.

**Sample Format**

Required deliverables in the complete report:

The format of your paper (headers, sections, etc.) is flexible, although it should contain the following information.

Introduction

Data Description

(Where did the data come from? How big is it? How many observations? Where can we find out more? What are the specific variables that we need to know with respect to your analysis?)

Analysis Question 1:

Restatement of Problem

Specify the Model

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Make sure and address each assumption.

Comparing Competing Models

adj R2

Interval CVPress

Parameter Interpretation

Interpretation

Confidence Intervals

Conclusion

A short summary of the analysis.

Analysis Question 2

Restatement of Problem

Model Selection

Type of Selection

Stepwise

Forward

Backward

CUSTOM

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Make sure and address each assumption

Comparing Competing Models

Adj R2

Interval CVPress

Kaggle Score

Conclusion: A short summary of the analysis.

Appendix

Well commented SAS Code for Analysis 1 and 2

**Rubric:**

Presentation (30%):

Organized paper with title, headings, subheadings, etc.

Labeled plots, figures, tables and charts.

Every plot, figure, table and chart included is referenced in the paper and vice versa.

No spelling or grammatical errors.

Analysis Question 1: (35%)

Analysis Question 2: (35 %)