MSDS 6372 Project 1 Description

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.

Your team’s objective is to conduct 2 analysis each consisting of at least 3 competing models.

1. This model should be formed to facilitate the easy interpretation of parameters for use in helping real estate agents, contractors and prospective buyers gain insight into the important factors that influence housing prices in Ames, Iowa. Characteristics such as ease of measurement and interpretability should be considered. As part of the report, confidence intervals should be included. At the minimum, the model should contain at least one continuous and one categorical variable (interaction may be interesting but not required.) In addition, model selection techniques may be used and an external cross validation should be conducted to compare competing models (in addition to criterion such as the adjusted R2, AIC, BIC, etc. Finally, make sure and address multicollinearity issues (VIF, etc.) ( You may not have any multicollinearity issues due to the simplicity of your model. Note: every group will probably come up with different models here. Your goal is to simply come up with a valid model, whose parameter estimates yield some useful information to your audience. This can be a considerably less sophisticated (more parsimonious (less factors / predictors) model than in question 2. Note 2: Your group is limited to only the techniques we have learned in 6371 and up to Unit 6 of 6372.
2. This model is aimed at being the most predictive model. **Your group is limited to only the techniques we have learned in 6371 and up to Unit 6 of 6372 (no random forests or other methods we have not yet covered.)** Confidence intervals for the parameter estimates are not required for this model. However, you should provide evidence as to why you chose your final model among other competing models. We will say that an external cross validation is mandatory as is comparing at least the adjusted R2 and AIC between competing models. You will provide your top three models (ranked in order) based on your ASE from the test set as well as the top three models with respect to the Kaggle score you will get when you submit responses to the competition. (NOTE: AT LEAST ONE OF THE THREE CANDIATE MODELS MUST BE THE RESULT OF SOME FROM OF A REGULARIZED METHOD (LASSO, RIDGE, ELASTIC NET, ADAPTIVE LASSO, etc.) To illustrate, a table like this one might be used:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Set Models** | Adjusted R2 | AIC | ASE (Test) | Kaggle  Score |  | **Kaggle Scoring Models** | Adjusted R2 | AIC | ASE (Test) | Kaggle  Score |
| Model 1 | .89 | 1272 | 648 | .721 |  | Model 1 | .89 | 1272 | 648 | .721 |
| Model 2 | .78 | 1590 | 700 | .945 |  | Model 3 | .81 | 2001 | 720 | .888 |
| Model 3 | .81 | 2001 | 720 | .888 |  | Model 2 | .78 | 1590 | 700 | .945 |

NOTE 1: ALL ANALYSIS MUST BE DONE IN SAS OR R 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 5 points will be awarded on the project grade to the team with the model with the lowest Kaggle Score. In the unlikely event of a tie the points will be split.

Note 3: An extra 2 points will be awarded on the midterm to the overall KAGGLE winner between all five sections.

**Required Information and SAMPLE FORMAT**

Required deliverables in the complete report. The format of your paper (headers, sections, etc) is flexible although should contain the following information.

PAGE LIMIT: 10 Pages (You may put supporting plots/charts/tables etc. in the appendix, just make sure and label and reference them appropriately.)

Introduction **Required**

Data Description **Required**

Exploratory Analysis **Required**

Analysis of Question 1:

Restatement of Problem **Required**

Model Selection

Type of Selection

**Optional**: LASSO, RIDGE, ELASTIC NET etc. and Model Averaging

Stepwise, Forward, Backward, Mallows Cp,

Manual / Intuition

A mix of all of the above.

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

**Optional:** Comparing Competing Models

**Optional**: (AIC, BIC, adj R2

Interval CVPress

External Cross Validation**)**

Parameter Interpretation

Interpretation **Required**

Confidence Intervals **Required**

Analysis Question 2

Restatement of Problem **Required**

Model Selection

Type of Selection

LASSO, Model Averaging

Stepwise, Forward, Backward, Mallows Cp,

Manual / Intuition

A mix of all of the above.

**At least two of the above required.**

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Comparing Competing Models

AIC, BIC, adj R2 **Required**

Interval CVPress **Required**

Cross Validation and/or hold out test set **Required**

Kaggle Score **Required**

:

Conclusion/Discussion **Required**

The conclusion should reprise the questions and conclusions of the introduction,

perhaps augmented by some additional observations or details gleaned from the analysis

section. New questions, future work, etc., can also be raised here.

Appendix **Required**

Well commented SAS/R Code **Required**