UNSUPERVISED LEARNING PROJECT PROPOSAL

BRANDON MORETZ
SEAN PRENTIS

INTRODUCTION

To accurately forecast the value of a home, we must find a relevant dataset that contains accurate information of comparable inventory so that we can explore the significant variables of a home which ultimately determine the sale price of the residence. Once we have explored the data set and selected an appropriate sample from the population, our task will be to create multivariate regression models that leverage key indicators in the data to predict the value of a home given based upon its features.

The key task in this assignment is to use unsupervised learning methods to assist our research by composing principal components and looking for the presence of latent factors in the underlying data that we might not find ourselves with more traditional methods of exploratory data analysis. Once we have constructed the appropriate components/factors, we will construct our predictive models from these composite features. We will then form our formal hypothesis tests, at our stated confidence intervals, and conduct statistical significance tests upon these models. The results will be compared to previous models constructed with more traditional means, and we will contrast the results of the two approaches.

In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through automated variable derivation, validation, selection and visualization to measure the relevance of these composite indicators as they pertain to the value of the home in terms of a dollar estimate.

DATA SURVEY

This data is from the Ames Iowa Assessor's Office and contains characteristics regarding residential properties sold in Ames from 2006 to 2010.

The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor's Office specifically assess the value of individual residential properties sold in Ames, Iowa from 2006 to 2010. Given that this data was collected to assess home values, it should be an ideal source of information for our observational study and the resulting regression model. The main task we aim to achieve with this analysis is the in-

depth exploration of the interdependencies and relationships in the variables under study. A corollary of this variable derivation research will be the construction of a regression model to forecast the price of homes using these newly uncovered relationships.

METHODOLOGY

Of the total Ames housing dataset, only a selected pool will be used to build the regression model and allow for feature selection in our unsupervised learning methods. Homes which are not a single-family residence will be removed, allowing for improved predictive capabilities for both methods. For the traditional regression models, features will be selected through the use of Random Forest and automatic model selection techniques: forward, back wise, and stepwise. Comparing the feature selection process for both methods, supervised vs. unsupervised, through regression will allow for a like-for-like comparison in their predictive capabilities through the comparison of R-Squared (R^2), Adjusted R-Squared (R^2), and Mean Absolute Squared (MAE) metrics.

RESEARCH QUESTIONS

1.) Does the use of unsupervised learning methods, namely principal component analysis and exploratory factor analysis, lead to the derivation of more parsimonious predictive models that maintain the same accuracy as traditional methods?

2.) Will the selected features in the unsupervised method make business sense? Are they easily justifiable in their selection?