Assignment 5: Automated Variable selection, multicollinearity and predictive modeling.

Sri Seshadri

7/19/2017

1. Introduction

2. Sample definition

It is assumed that typical home buyers are those that move from apartments to single family or town homes. Also apartments are less likely to be sold to individuals as they remain holdings of owners for rental income. Single family and town homes belong to "Residential Low density" (RL) zoning classification in the city of Ames. Data belonging to only to the RL zone is considered for analysis and model development. Also, it is assumed that typical homes have paved streets for access and above grade living area greater than 800 square feet. Sales data belonging to homes that were sold in abnormal conditions such as trade in, foreclosure or short sale are not included in the analysis. Also, sales between family members, sale of adjoining lot, linked properties are omitted from the data. Homes with no basements are excluded from the analysis at this time. Homes with living area greater than 4000 square feet and garage area greater than 1000 square feet were identified as outliers and are therefore removed from the analysis. Table 1 shows the waterfall of the data not included in the data and the eligible samples.

Table 1: Drop waterfall

DropCondition	counts
01: Not LowDensityZone	657
02: Not Normal/Partial Sale	189
03: Street Not Paved	3
04: Less than 800 SqFt	41
05: No Basement	48
06: Greater 4000 sqft living Area - Influence Points	4
07:Garage area greater than 1000 sqft - Influence points	23
99: Eligible Sample	1965

2.1 Predictor variables of interest for modelling

The following variables in the data were deemed to be of interest for model building. The choice of parameters was based upon intial Exploratory Data Analysis (EDA) and subject matter expertise. Some variables may be combined into one variable for model building purposes. For example, the number of baths may be sumed into one variable or the above grade living area and the basement area may be sumed into one variable yielding total square feet. The categorical variables are coded into indicator variables.

Table 2: Variables of interest

LotArea	BsmtFullBath	MoSold
LotConfig	BsmtHalfBath	YrSold
Neighborhood	FullBath	SaleCondition
BldgType	HalfBath	FirstFlrSF
OverallCond	BedroomAbvGr	SecondFlrSF

YearRemodel	KitchenQual	
TotalBsmtSF	TotRmsAbvGrd	LotArea
GrLivArea	GarageArea	LotConfig

2.2 Training and validation samples.

From the eligible samples, 70% of the data is randomly sampled to be used as the dataset for model development. This dataset would be refered to as training dataset. The remaining 30% is used as the validation set to evaluate the model performance of predicting sale price on data that is outside the training set. Table 3 shows the split of the total eligible samples.

Table 3: Training and Validation sampling

Data	Samples
Training set	1383
Validation set	582
Total	1965