Statistical Analysis Plan

House Sales Prices - Best possible regression

Emmanouil Mertzianis - 2600474M

15/9/2021

Population

All house sales in the last six months.

Primary Objective

Identify the best regression model to explain the variability in the final sale price of houses by selecting variables from the available explanatory variables or using transformations of them.

Secondary Objectives

- Interpret the effect of each explanatory variable on the final house sale price both in the case of the full linear regression model and the case of the model which was chosen as the best one (if an interpretation is applicable).
- Apply the more advanced techniques of regularised regression using Lasso or Ridge regression.

Variables under consideration

Primary Response Variable

- **price**(continuous): Final House Sale Price

Explanatory Variables

- **elevation**(continuous): Elevation of the base of the house
- **dist_am1**(continuous): Distance to Amenity 1
- **dist_am2**(continuous): Distance to Amenity 2
- **dist_am3**(continuous): Distance to Amenity 3
- **sqft**(continuous): Square footage of the house
- **precip**(continuous): Amount of precipitation
- **bath**(categorical): Number of bathrooms
- **parking**(categorical): Parking type

Missing data procedures

- **Response Variable**: Records with missing values on the sale price are removed from the data set.
- **Explanatory Variables**: For any of the explanatory variables used, if there are a few missing values ($\leq 10\%$ of the total amount of records) we remove those observations from the data set. Otherwise, we impute any missing values by using *Multivariate Imputation by Chained Equation (MICE)*.

Numerical and graphical summaries to be presented

- Boxplot, five-number summary, mean, variance and number of missing values for:
  - sale price
  - elevation
  - dist_am1
  - dist_am2
  - dist_am3
  - sqft
  - precip
- Barplot for:
  - bath
  - parking
- Scatterplots for the relationships between the house sale price and:
  - elevation
  - dist_am1
  - dist_am2
  - dist_am3
  - sqft
  - precip,

  using different colours and symbols for each level of the categorical variables *"bath"* and *"parking"*, along with a superimposed simple linear regression line for each level.
- Boxplot, five number summary, mean and variance for the house sale price against:
  - bath
  - parking
- All of the above, with various transformations of the variables if the exploratory analysis deems it appropriate, in order to assess whether a transformation improves the relationship.

Models to be fitted and the analysis plan

- Throughout the analysis, a significance level of 5% will be used for hypotheses testing and construction of confidence intervals.
- Exploratory data analysis will precede model fitting.
- **Outliers**: Any outliers discovered through exploration will be considered for removal if they are exceedingly extreme, while the rest will be noted and considered after fitting the models (by using leverage plots).
- The initial data set will be divided into three subsets; *training, validation and test*(70%, 20% and 10%, respectively). The test set will be used to report the expected prediction performance of our "best" model.

- **Assessing the goodness of fit/Fitting criteria**: The validation error will be used as our main metric for choosing the best regression model, along with AIC and R^2 or $R^2_{adjusted}$.
- **First model**: full linear regression model. This model will be used to interpret the effect of all available explanatory variables on the house price when we are considering all of them together. Additionally, it will be used as a starting point for the variable selection procedure. The full model contains any transformations and/or interactions suggested by the exploratory analysis.
- For the purpose of finding the best possible regression model, backward stepwise variable selection(elimination) will be used, starting from the full model and using the chosen fitting criteria.
- After variable selection, regularised regression will also be considered using Lasso and Ridge regression on the full model and the "best" model, respectively. The validation error will be used to assess the possible improvement on fit. Lasso will be used as an elimination technique on the full model to check whether we get a better model than the one we obtain from the backward elimination method. Ridge regression will be used on the "best" model to smooth even more the estimated coefficients and check whether a smoother "best" model with less variance will give as an even better prediction performance.
- Assumptions checking for the full model, the "best" model and the regularised "best" model using residual plots and influencial outliers detection using leverage plots.