UCLA Extension Data Science Intensive

Instructor: William Yu

Project 4

10/16/2018

A. Analyze Zillow Prize Project

- In Project 3, you have explored Zillow Prize Project. Now let's find out what variables/predictors will be able to predict the Zestimate's forecast errors:

  logerror = log(Zestimate)-log(SalePrice)

- Keep in mind that if Zestimate is a good model, its forecast error (logerror) should be like an independent/uncorrelated noise. It means it will be difficult to find additional variables to explain logerror. But, after all, Zestimate is not a perfect model. So there might be a change to find some statistically significant predictor. Since the dependent variable (y) in this case is logerror, you don't need to be surprised to see low R^2.
- First, follow the p03_zillow.R script to run through missing_values part. However, we want to change the standard of being good_features from with missing_pct <0.75 to **<0.25**. Note: by doing so, the number of good feature variable will be reduced to 27.
- Using left_join to merge the properties data to transaction data by "id_parcel", which is called cor_tmp.
- Create a subset of data frame from cor_tmp containing logerror and those good features variables.
- Before running regression analysis, let's remove these variables because they are (1) geographic information and ID, (2) one value, or (3) pure linear combination of other variables.
  (1) id_parcel, fips, latitude, longitude, zoning_landuse_county, zoning_property, rawcensustractandblock, region_city, region_zip, censustractandblock.
  (2) tax_year
  (3) tax_building and tax_land (note that tax_building+tax_land = tax_total)
- Now you should have 15 variables in the data frame.
- Use cor and corrplot functions to check the correlations among these 15 variables. There are some variables which are extremely correlated (correlation>0.95). Remove those highly correlated variables (only keep one).
  Hint: num_bathroom_calc, num_bathroom, num_bath; area_live_finsihed, area_total_calc; tax_total, tax_property.
- Use str to see the structure of this data frame. There are two variables that are interger. Convert them to factor. Hint: factor(xx). See D03c_logit.R. Why do we do this?
- Now we are ready to run the linear regression for the dependent variable: logerror. Use lm to run regression including all 14 variables. And then use regsubsets to find the best model.
- Change the dependent variable from logerror to abs_logerror and do the regression.

- Explain the results. What will you tell Zilliow? Why there are difference between logerror and abs_logerror results?