Problem Set 4

Due Monday, February 7 9am

You should complete the following exercises with a partner. Please choose a partner you have not completed an assignment with. Your grade depends on (1) getting the answer right, and (2) showing how you arrived at your answer. Place all of your answers in a word/pdf document and attach your code as well.

EXERCISE: Predicting House Prices in Ames, Iowa

In this exercise, you will use linear regression to predict the sales price of residential homes in Ames, Iowa.

Our dataset, "HousePrices.csv", contains information on 669 different houses, and our dependent variable is the final sale price of the house (in US dollars), or the variable "SalePrice". There are 56 other variables in the dataset that can all be used as independent variables. (Some of these are factor variables so, including all factor levels, you will have 135 potential X variables for your prediction model.) Here is a brief description of each variable:

- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotArea: Lot size in square feet
- LotShape: General shape of property
- LandContour: Flatness of the property
- LotConfig: Lot configuration
- Neighborhood: Physical locations within Ames city limits
- Condition: Proximity to main road or railroad
- BldgType: Type of dwelling

- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterCond: Present condition of the material on the exterior
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- CentralAir: Central air conditioning
- FirstFlrSF: First Floor square feet
- SecondFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade

- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- ThreeSsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- MoSold: Month Sold
- YrSold: Year Sold
- saleprice: Sales Price

QUESTIONS

1. Set your random seed to 157911 and then randomly partition the data into a 70% training data set and a 30% test data set. What is the average sale price in the training data? What is the average sale price in the test (hold out sample) data?

2. The simplest prediction model ("Naive Model") would predict the average sale price in the training data for every instance. If you apply this naive model, what RMSE (root mean square error) do you get for the prediction error in the test data?

3. Construct a prediction model ("Model 1"), by running a regression of sale price on dummies for each year sold and dummies for each number of bedrooms (seven possible values). Use the training data to run the regression. What is the interpretation of the intercept? What is the interpretation of the coefficient on yrsold=2007?

4. Use Model 1 to predict sales prices for the test data set. What is the RMSE for Model 1 on the test data?

5. Construct a new prediction model ("Model 2"), by running a regression of sale price on above grade (ground) living area, age of the house at time of sale, dummies for each number of bedrooms, and dummies for each year of sale. Note you will have to create age of the house at time of sale from the given variables. Run the regression on the training data. What is the interpretation of the coefficient on ave grade (ground) living area?

6. Use Model 2 to predict sale price for the test data set. What is the RMSE for Model 2 on the test data?

7. Which model is the best prediction model: Naive, Model 1, or Model 2? Justify your answer.

8. Can you find an even better prediction model by either adding or subtracting variables from your model? (*Please Note:* This question is very open ended, and you could work for weeks perfecting the model. Please do not obsess over it. Just try a couple things and report what you found. Do not use Lasso here. Just a bit of trial and error with some reasoning about why you tried what you did.)

9. Now build a Lasso model for this problem. Include dummy variables for mszoning, lotshape, bldgtype, neighborhood, fullbath, halfbath, bedroom, kitchen, mosold, yrsold. Aslo include lotarea and its square, age of the house

at time of sale (and its square and cubic), firstflsf, secoondflrsf, and grlivarea. Feed Lasso only the training data. **Set your random seed to 12347 just prior to the lasso linear command.** (Lasso will automatically break the data down further for the purposes of testing out of sample fit, and this will set the random seed for that.) How many X variables are included in the Lasso model with the lowest RMSE? List the selected variables. Note: to have stata make the dummy variables for you, use the command xi. Example: xi: lasso linear sale price i.neighborhood will automatically generate the dummies for each neighborhood.

10. Predict the sale price for each house in your test data using the best Lasso model. What is the out-of-sample RMSE of the lasso model? How does this compare to the out of sample RMSE from the regression models above?