CS/CE 457/464 - Homework Assignment 6: Regression

Due Date: Monday, October 21 at 11:59 pm

Purpose:
Demonstrate understating of Regression technique for correlation and prediction

Points: 100

Deliverables: Submit ipynb code file with your answers

- Review IDMA Book Chapter 12 Predictive Analytics
- Use the dataset `HousePricingData.csv`
- Perform analysis on the following questions**.** Make sure to include interpretation of each result including coefficients, p-values, r-square and other necessary information to support your answer

1. Create a regression model between `GrLivArea` and `SalePrice` (response variable). Show the scatter plot with regression line between them. Discuss the R-squared value.
2. Create a regression model between `TotalBsmtSF` and `LotArea` (response variable). Show the scatter plot with regression line between them. Discuss the R-squared value.
3. Calculate Correlation for questions (1) and (2) and explain the correlation value to support your answer for questions (1) and (2).
4. Create a regression model to predict `SalePrice` using all other inputs. Discuss the effectiveness of the model using R-squared value. Report 3 most significant inputs and 3 least significant inputs (based on p-value) and interpret the results. Create one new input of your choice of values and show the prediction of `SalePrice` using the same model.
5. Create a regression model to predict `LotArea` using all other inputs. Discuss the effectiveness of the model using R-squared value. Report 3 most significant inputs and 3 least significant inputs (based on p-value) and interpret the results. Create one new input of your choice of values and show the prediction of `LotArea` using the same model.
6. From Question 4, drop/remove all the columns which are not signification (p-value >0.05) and create a new model to predict `SalePrice`. Discuss the performance of the model using few inputs as compared to using all inputs in (Question 4). Which model do you prefer and why?
    a. The idea is to create a simple generalized model with fewer inputs which are important for prediction and getting the similar performance. For this concept, please research and study “Regularization in Regression”
7. Using the model in Question 6, create 3 new data records and predict their `SalePrice`. Discuss if the predicted output looks good and make sense.