LINEAR REGRESSION WITH GRADIENT DESCENT FOR HOUSE SALE-PRICE PREDICTION

Mawaba Pascal Dao

Florida Institute of Technology

ECE5268: Theory of Neural Networks

Major Project 1

Dr. Anagnostopoulos

Feb 21, 2020

Git repo: https://github.com/PascalPolygon/Neural_Networks/tree/master/MP1_housePrices

Task 1:

a. Overfitting occurs when you have more parameters than training samples. Then MSE(train) is very low. But, MSE computed on the holdout set is very large. It's as if the model "memorizes" training data as opposed to learning the general trend.

b. A validation procedure can be used to avoid model overfitting. Validation usually involves separating the data into training and testing sets. The percentage of the data to be used for each set can be tricky to determine for small data sets. In such a situation cross-validation can be employed, where the data can be divided into K-folds, each fold being used as a testing set on each iteration of the training. This ensures that each sample of the data will be both in the training and testing set.

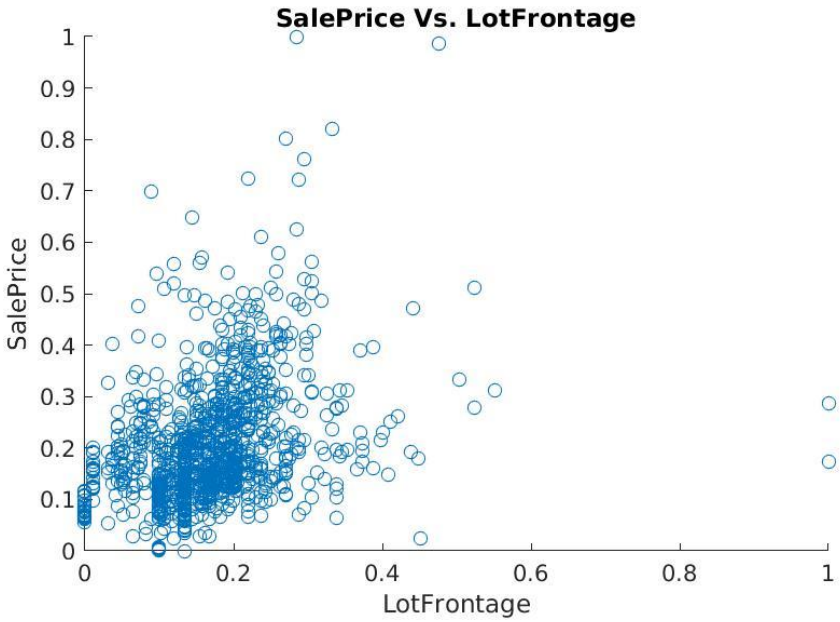
c. Regularization approaches are heuristic methods to prevent a model from overfeeding or overtraining. It works by introducing a regularizer term to our loss function which forces the model coefficients towards zero. This is done as a result of the observation that models with large weights tend to overfit. An example of such a regularization method is Ridge regression, a non-constrained regularization method that uses a regularizer such as a L1 or L2 norm and a regularization strength variable λ.

Task 2:

(Please find attached document task2.pdf)

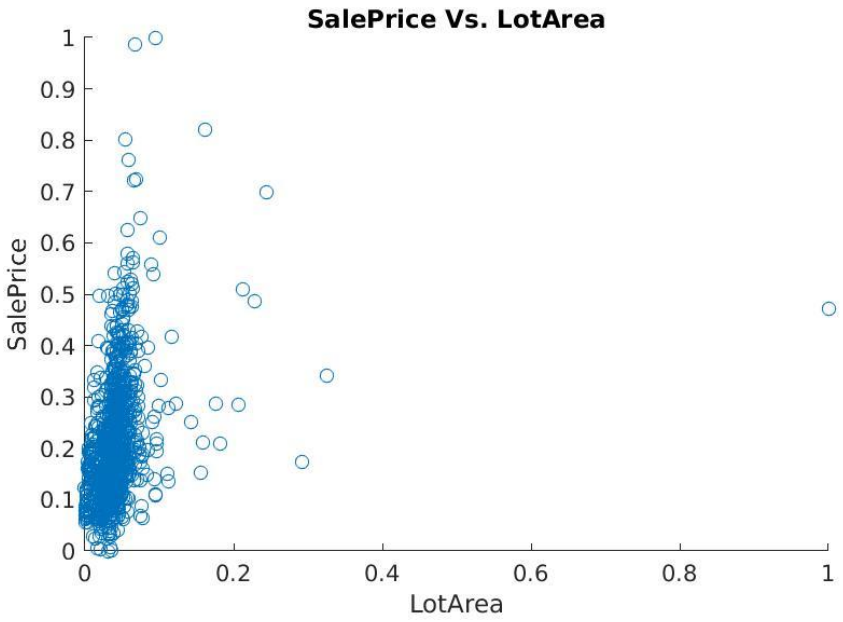
Task 3:

a. Scatter plots of output variable versus each input feature



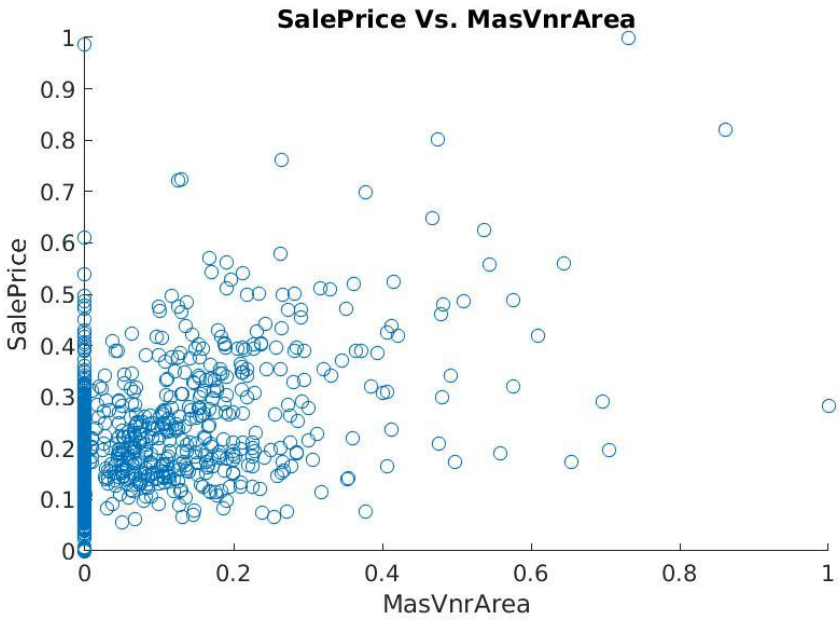
i

The relationship between the lot frontage area and the sale price is clearly not linear. It appears that for small frontage areas the sale price does not change much. But there is a slight trend towards higher sale prices for higher frontage areas. This will have a slight influence on the linear regression model.

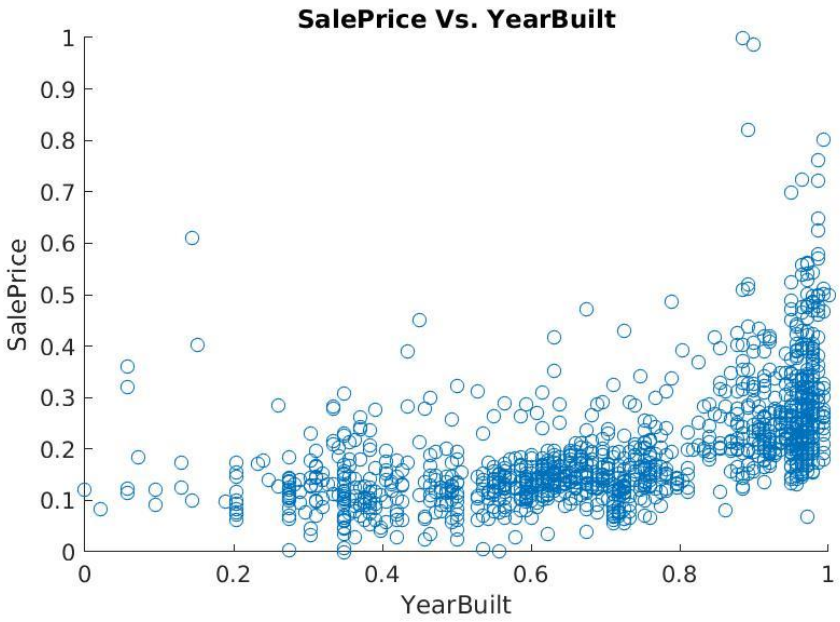


This scatter plot is even more vertical than the previous one, suggesting a poor linear relationship, and an even smaller effect on the linear regression model.

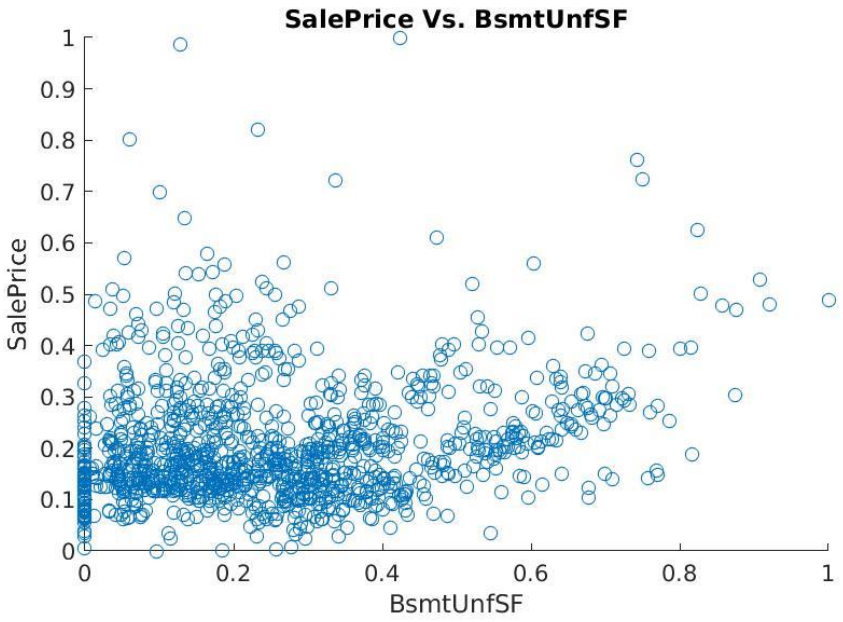
Although there is a considerable amount of overlap, one can observe that the sale price tends to increase when the OverallQual feature increases. Suggesting that our linear regression model will tend to predict higher prices for high input OverallQual's.



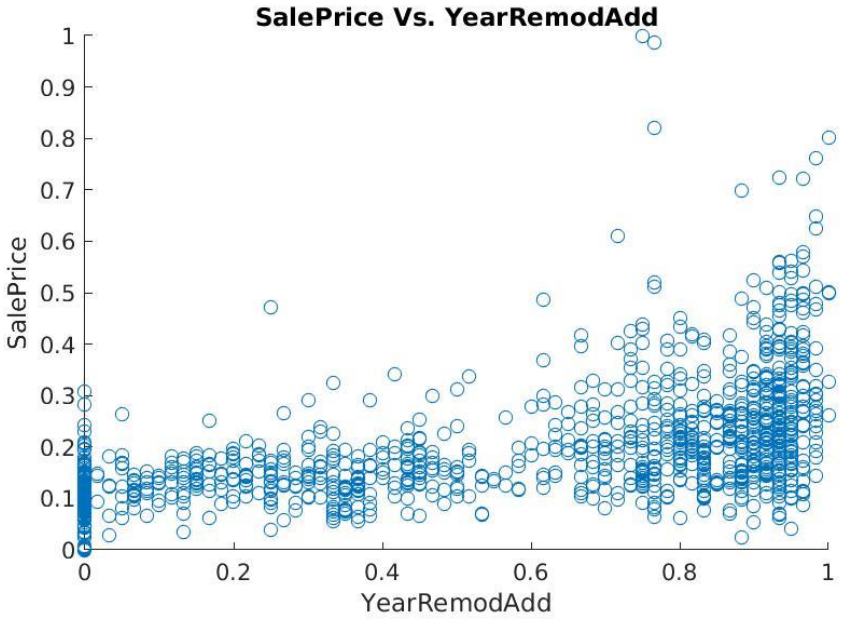
This plot does not appear to have a very well-established trend. There are many values stacked at the lower end of the x axis. This suggests that sale prices arbitrarily vary for a MasVnrArea of 0.



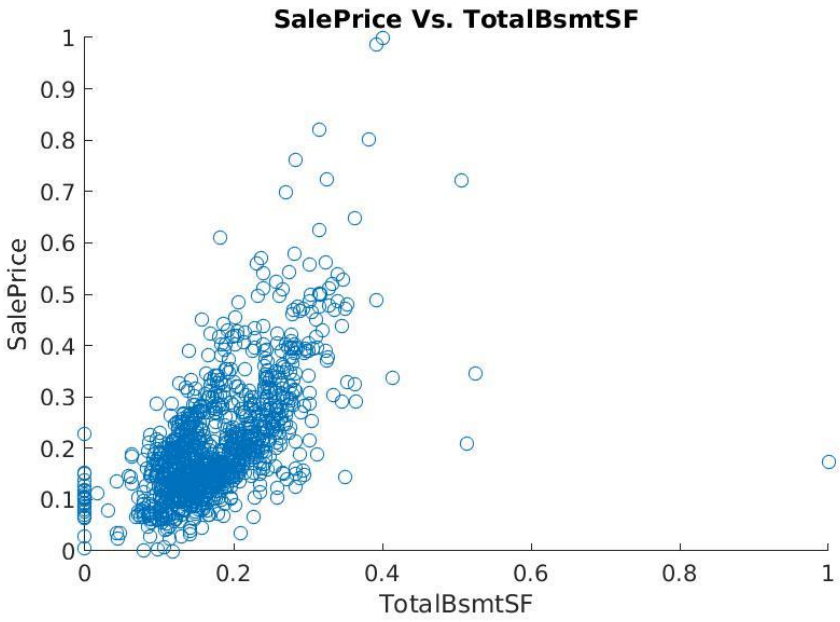
Although not a clear linear relationship, one can observe from this plot that the sale price tends to increase as year-built increases. There are however many values of YearBuilt for which SalePrice seems to arbitrarily vary. This suggests that our linear regression model could have a similar behavior.



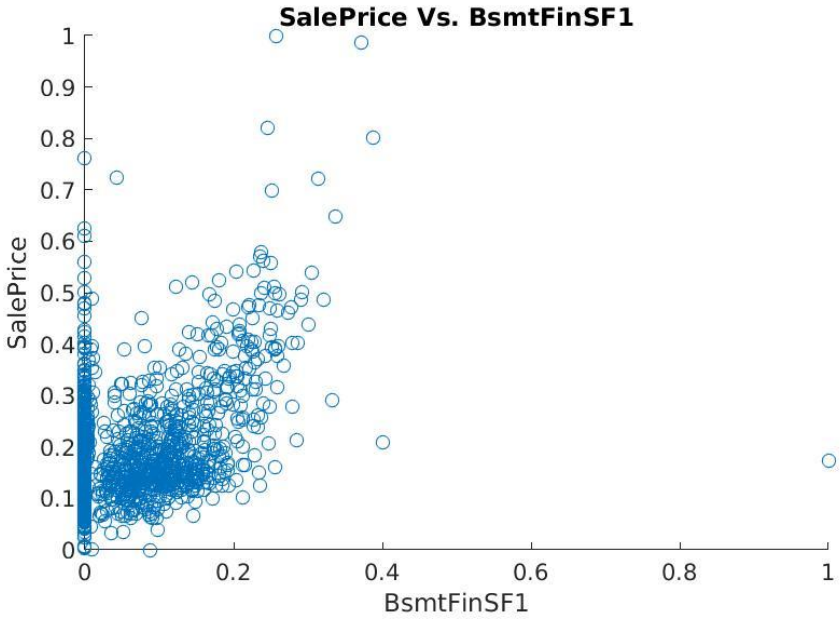
The unfinished basement area graph resembles Masonry Veneer Area graph in the sense that there are many BsmtUnfSF values stacked at the bottom for increasing sale prices. It is therefore not likely that our model will have a linear sensitivity to this feature.



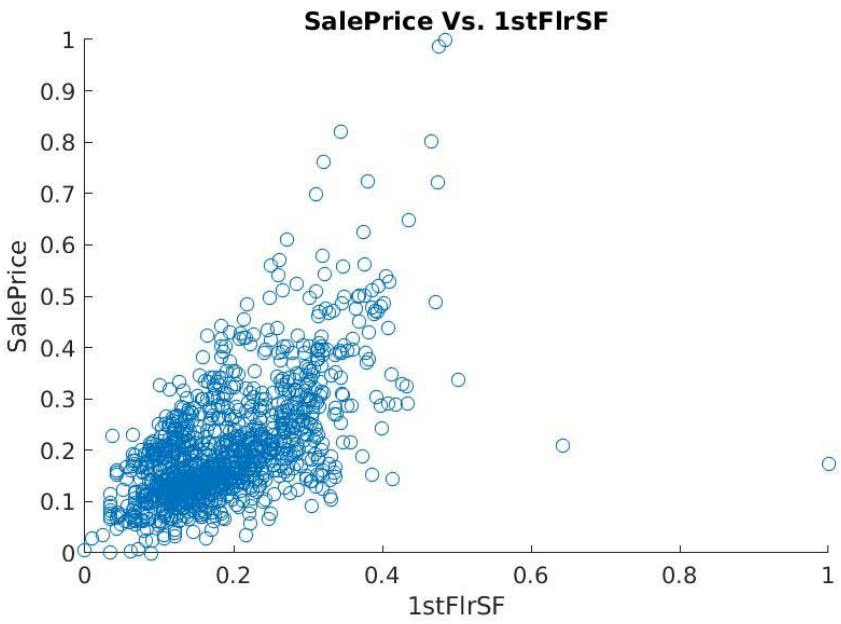
The Sale Price does not appear to increase with increasing values of Normalized Remodeling Year. Instead it seems to arbitrarily increase or decrease for very similar values of Normalized Remodeling Year. It is therefore difficult to predict the behavior of our model in relationship to this feature.



The Total Basement Area exhibits a clear linear trend. A similar linear relationship can therefore be expected from our model.



Although not as clear as the previous figure, in general Sale Price increases as Type 1 Finished Square Feet (*BsmtFinSF1*) increases. Predicting this relationship on our model is not as clear cut as in the previous case due to the number of Sale Price values arbitrarily varying for the same low values of *BsmtFinSF1*. But, in general in can be expected that our model predicts a higher Sale Price for a higher *BsmtFinSF1*.



Based on this figure, it is expected that the linear regression model predicts increasing Sale Prices for increasing First Floor Area values.

b. Reporting on a champion model

The result of performing 5-Fold cross-validation for various features on the first 1000 samples in the dataset is as shown in the table below:

Inputs	**MSE train**
All 10 features	0.26648
All features except *OverallQual*	0.29602
YearBuilt squared in addition to all 10 features	0.23946

Note: The MSE values in the table above were computed using gradient descent with a learning rate of 0.0001 for 1000 epochs. However, the values used for learning rate and epochs in the final model are 0.01 and 100,000 respectively.

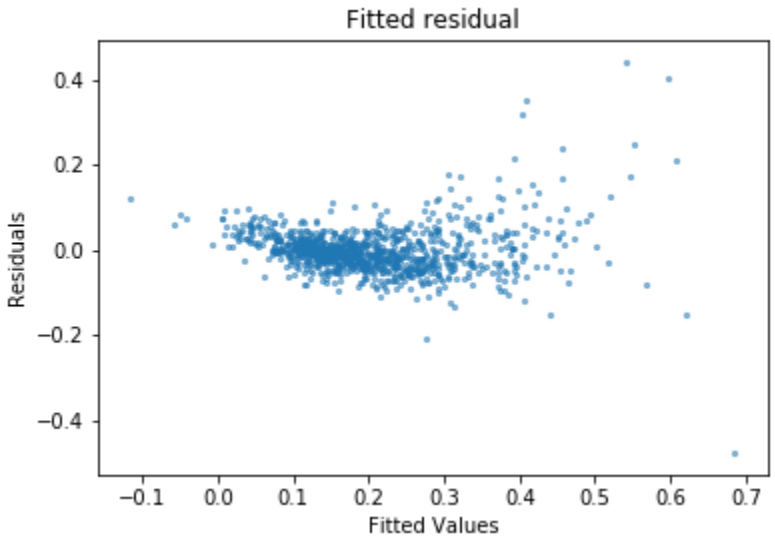
These results suggest that for a linear regression model, a higher number of features is desired for better performance.

c. Results of the champion model

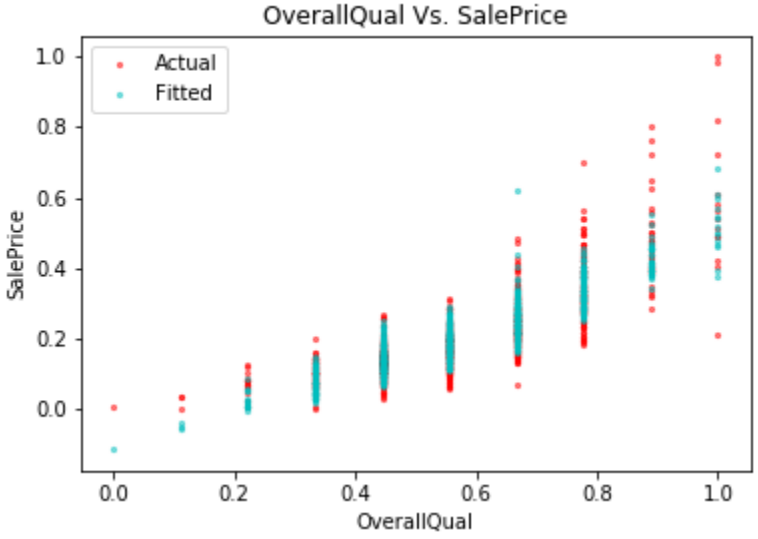
Using Year Built Squared in addition to the existing 10 features, the data was split in two. A training set containing 1000 samples and a testing set containing the remaining 195 samples. In the submitted folder, those sets are in the dataset subfolder. The model was trained using a

gradient descent method with a **learning rate of 0.01** for **100,000 epochs**. These values can be edited in the call to grad_desc from the lin_reg function.

MSE train on 1000-sample training set (This is the average on 5 folds)	**MSE test on 195-sample holdout set**
0.0031	0.0058



The fitted residual plot above shows a satisfactory clustering of results around the y=0 line. This implies that the weights calculated by the model yield predicted values that are close to the observed ones. As there seems to be a trend of increasing residual values for higher fitted values, there is still room for improvement on this model.



The figure above is a comparison of the model's predictions with the actual values for the OverallQual feature. It depicts a good overlap of the predicted and actual values. However, as in the previous figure, one can observe that for higher values of OverallQual, the gap between actual and fitted gets larger.

Conclusion

Overall, the results indicate that the model behaves as expected and represents a good choice for this dataset. This is especially true for output values in the mid-range. For higher values both figure show that the model performs less accurately. This could be due to the nature of the dataset itself. As the data is likely to contain more samples in the mid-range than at the extremes, we can expect our model to be biased towards better predictions for inputs in the mid-range.

---

i