# Intro To Data Science Assignment One

Ernest N. Frimpong

2024-10-29

## Executive Summary

I predicted the sale price of houses using 107 features for the housing price dataset. To do this, I used a cross-validated (CV) lasso model. The training root mean squared error (RMSE) on this model was $20,824.18 and the validation RMSE was $18,866.54 with an optimal lambda of 91.16. This produced a test RMSE on Mihai's board of $19,939.

Before using the CV lasso model, I cleaned the data by filling in missing values where necessary using the appropriate strategy and dropped variables where there were too many missing values. I then standardized the numeric predictors and created dummies for categories.

I run a cross-validated lasso regression to obtain my training and validation RMSEs, which I plot below. Having obtained the RMSEs across all lambdas, I further expand on the regions of overfitting and underfitting for various lambdas.

Finally, using the minimum lambda obtained from the cross-validated lasso, I run another lasso regression on the entire 'Housing_train' dataset. The fitted lasso is then used to predict house sale prices on the 'Housing_test' dataset.

## Data Pre-Processing

After loading the joined Housing dataset, I checked for percentages of missing values in each column. All columns with missing values greater than 80% were dropped. The remaining columns with missing values, if numeric, were filled with zeros or the median, and if categorical, were filled with 'no garage' if it was a 'Garage' category for instance. I split the data into training and test datasets where the test data contained rows where 'SalePrice' was null or zero.

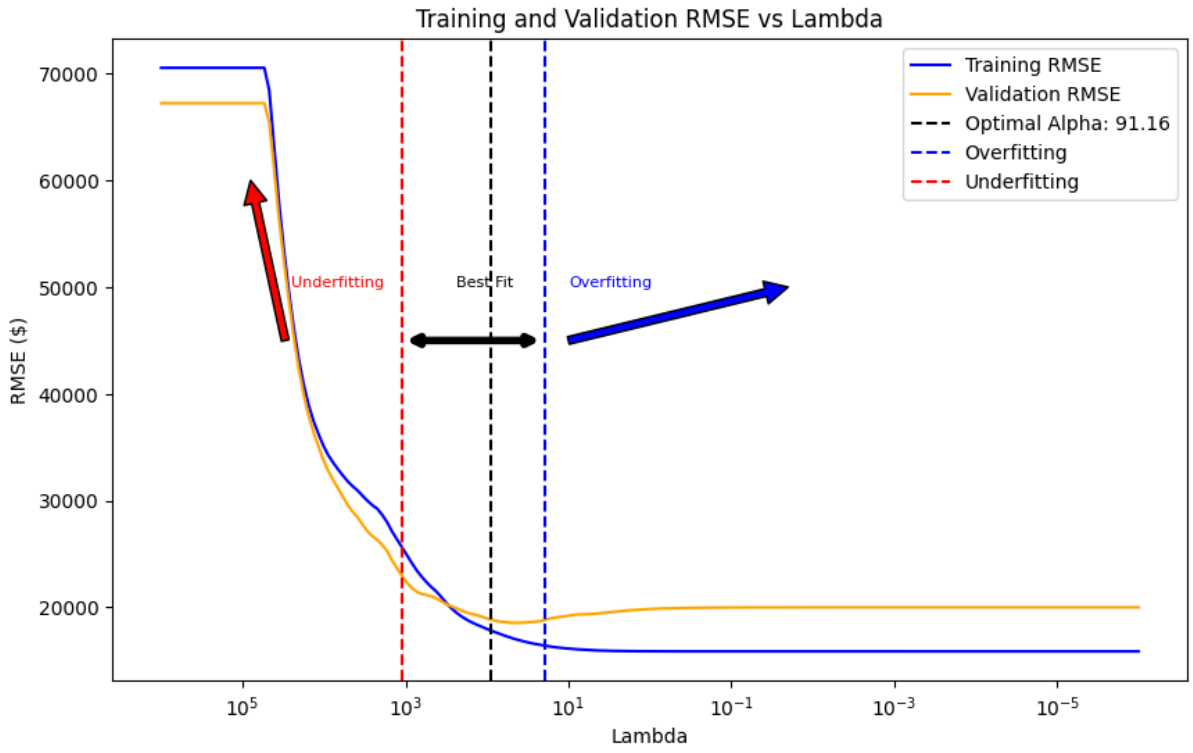
I obtained dummies for categories ensuring one category for each variable was dropped. To handle outliers, I made scatterplots of each numeric variable against the 'SalePrice'. Variables with sale price values that seemed 'abnormal' were taken out. For instance, Lot Frontages of square feet above 300 sold for less than $300,000, which is 'abnormal' compared to the rest of the Lot Frontage values. Finally, I standardized the numeric predictors.

## Modelling & Model Tuning

I created folds of 10 for cross-validation on the training data to estimate the RMSEs using Lasso. After, I set up my predictor and response variables based on the training data and my predictor variables based on the test data.

Using the 'LassoCV' function in Python, I obtained a training RMSE of $20,824.18 and a validation RMSE of $18,866.54 with an optimal lambda of 91.16. Below, I plot the overfitting, best-fit, and underfitting regions. **A breakdown of how lasso works is in the Python notebook.**

**Demonstration of Overfitting VS Underfitting**




The graph above shows the CV RMSEs across all lambdas. As you increase lambda, moving to the left of the red dotted lines, more coefficients are shrink and some are reduced to zero, bringing down the number of variables fitting the model. With fewer variables, you have poor model fit and low predictive power leading to underfitting.

The opposite can be said for overfitting where a low lambda, moving to the right of the blue dotted line, has little to no coefficients zeroed out. With more variables in the model, R-squared monotonically increases and reduces the mean-squared error on the training set. However, on the validation set, this would show as overfitting.

In this plot, although the RMSE increases after the blue dotted line, it is not very noticeable. This could be because the additional variables that are not shrunk have very low coefficient estimates as they have very little predictive power in explaining variations in 'Sale Price'.

**Final Model**

Using the minimum lambda obtained from the cross-validated lasso regression, I rerun the lasso regression on the entire 'Housing Train' dataset. I then obtained predictions of sale prices using the housing test dataset and placed them in Mihai's leaderboard with an RMSE score of $19,939.