Studs Real Estate

Bao Nguyen, Luke Lafave, Keenan Buckley

Section C, Group 34

Project House Prices - Advanced Regression Techniques

Problem Statement:

Our goal with this machine learning project is to accurately predict final prices of homes in Ames, Iowa utilizing solely 79 explanatory variables. In the end our model should give us a decent prediction of the housing prices given a complete set of the variables we choose are pertinent to the regression outcome. Many homeowners and flippers will want to use the product in order to discover which features affect the price of a house the most. Another main user of this model will be real estate agents and real estate apps. For instance if you want to put your home up on some app. The app may ask for a certain set of explanatory features and from there calculate a recommended asking price for your home. Real estate agents could use this feature when they are new to a certain area and not many homes have been sold. This means that there is very little reference for what a house should be priced. Using this model would allow the real estate agent to accurately price the home based on its features. Banks also would use this algorithm in order to decide whether to provide loans to homebuilders or not based on whether the algorithm determines that the house will meet a good enough price. There are thousands of different solutions to this problem used by most major real estate apps and companies. Zillow and Redfin both have proprietary algorithms they use to calculate home values.

Problem Solution:

- We will try multiple models and compare their results, starting with:
    - Linear Regression with ElasticNet regularization (L1 and L2 penalty).
    - Feedforward Neural Network regression.
- We will quantitatively score the performance of our model with the Root-Mean-Squared-Error between the logarithm of the predicted value and the logarithm of the observed sales price. (This is how Kaggle scores submissions).
- Once a final model has been trained we will submit results to the House Prices - Advanced Regression Techniques competition on Kaggle, which will provide us a final score.
- We believe the RMSE of the model needs to be at least below 0.5 for it to be useful. The vast majority of the leaderboard on Kaggle scored below 0.5.
- We guess that the model might achieve an RMSE score of around 0.25.

The Data:

The data used in this project was downloaded from Kaggle, and is based on the Ames Housing Dataset. The set has 79 features, of which 35 are numerical and 44 are categorical (splits into 267 dummy variables), and 1 label, the property's sale price in dollars. The training set has 1460 labeled entries for supervised learning. We will split the 1460 entries of the training set into a 80/20 split, with 1170 used for training and 290 used for model scoring. The training set will be split further using k-fold cross validation and used for tuning hyperparameters.

Additional Info:

Figure 1: Normalized histogram of SalePrice in our training set (48 bins, from Freedman and Diaconis' rule). Most of the data falls between $100K and $250K and in this range is where we expect to be the most accurate.

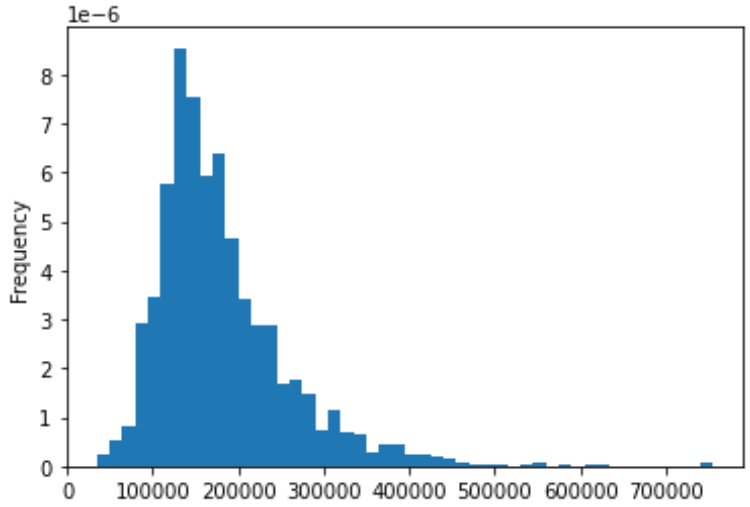
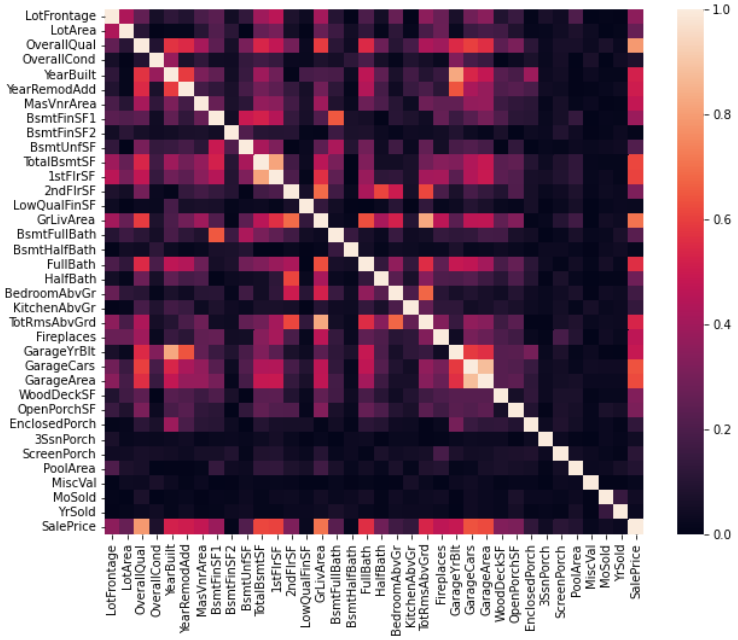


Figure 2: Correlation heatmap of the 35 numerical features and SalePrice. Features that may be less useful are those that have low correlation with SalePrice or have high correlation with other features. With filtering and regularization we should be able to reduce feature complexity and overfitting.



Timeline: List of important steps and when we plan to have them completed

Name	Date
Feature Selection	10/7/2022
Model Selection	10/11/2022
Design Initial Untrained Model	10/14/2022
Data Preparation for Model	10/21/2022
Hyperparameter Search	10/21/2022
Initial Model Training and Assessment	10/25/2022
Write Report and Presentation	10/27/2022
Group Progress Report Presentations	11/01/2022
Design Revised Model	11/8/2022
Revised Hyperparameter Search	11/11/2022
Revised Model Training and Assessment	11/15/2022
Submit to Kaggle	11/17/2022
Write Report and Presentation	11/29/2022
Project Final Presentations	12/01/2022