SAN JOSÉ STATE UNIVERSITY

Computer Engineering Department
CMPE 255-01 | Data Mining | Professor David, C. Anastasiu

Final Project Evaluation

Estimating Housing Values With Crime Incident Reports

Team 8

Pavana Srinivasadeshika Achar (11294556)
Aartee Kasliwal (012419004)
Wu, Chia-Chin (11485110)

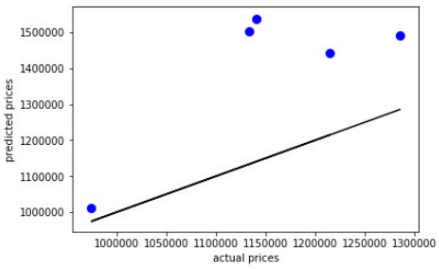
Spring 2018

Regression Models' Performance Evaluation

Seven regression algorithms were implemented and evaluated to predict the housing values. The Gradient Boost Regressor, Linear Regressor and Lasso Regressor resulted with least mean absolute error value. The following briefly describes the algorithms implemented. The evaluation of these algorithms are discussed further in chapter 3. Elastic Net regressor resulted in lesser RMSE and MAE scores compared to other algorithms. However, Gradient Boost Regressor resulted in better predictions with tuned parameters. The Gradient Boost Regressor was chosen to build the model.
Each model is described below with the figures next to them describing the variations in the predictions from the actual prices baseline.

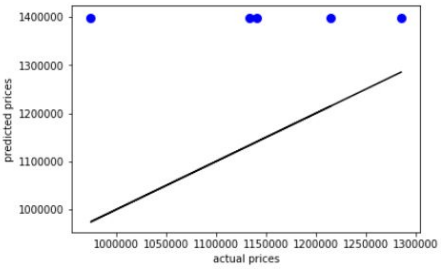
Elastic Net Regressor

This is the Linear regression model using elastic net regularized regression model which combines L1 and L2 penalties of the Lasso and Ridge methods[6].



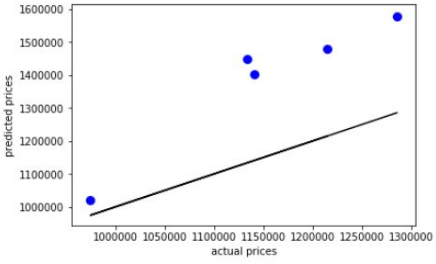
Bayesian Ridge Regressor

This is a linear regression model in which the statistical analysis is undertaken within the context of Bayesian interface. It optimizes the regularization parameters lambda (precision of the weights) and alpha(precision of the noise)[5].



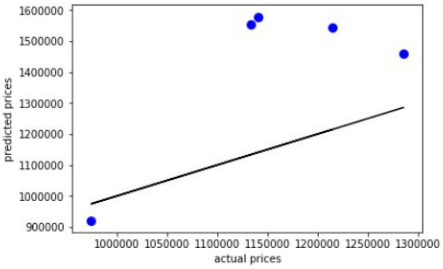
Gradient Boost Regressor

This is one of the ensemble methods[2] to build the regression model typically by using decision trees and generalizes that by optimizing an arbitrary differentiable loss function.



Ridge Estimator

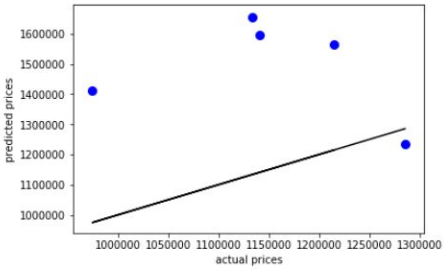
This model solves a regression model using linear least squares function as loss function and by regularizing[3] by L2-norm. This estimator has inbuilt support for multivariate regression by



showing the effect of collinearity in the coefficients of an estimator.

Linear Regressor

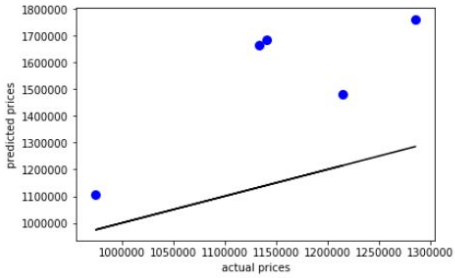
This model is for finding the relationships using linear predictor functions whose unknown parameters are estimated from the data. This model finds Ordinary Least Squares[1] that minimizes the residual sum of squares between the observed results in the 2D data, and the results predicted by linear approximation. The scattered plots represent the predicted prices and the line plot is the actual price for the test dataset.



.

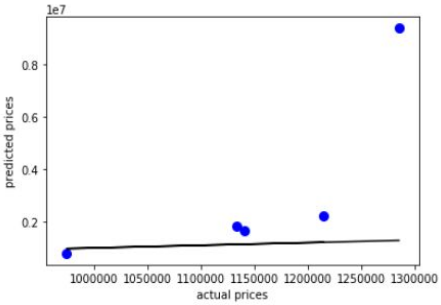
Orthogonal Matching Pursuit(OMP) Regressor

This regression model is a sparse approximation[4] which involves finding the best matching projections of multidimensional data onto the span of an redundant dictionary.



Lasso Regressor

This is linear model with iterative fitting along a regularization path trained with L1 prior as regularizer(aka the Lasso).



Methodology

The dataset was trained by splitting the training data into random train and test sets. Various regression models were trained and the efficiency of the model was evaluated using RMSE and MAE scores scored on the training dataset.

Evaluation of Regression Models

The following graphs describes the evaluation scores of the regression modelsfor metrics Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). It is observed that the Gradient Boost Regressor performed has the least RMSE score. And ElasticNet regressor model has the Least Mean Absolute Error score.

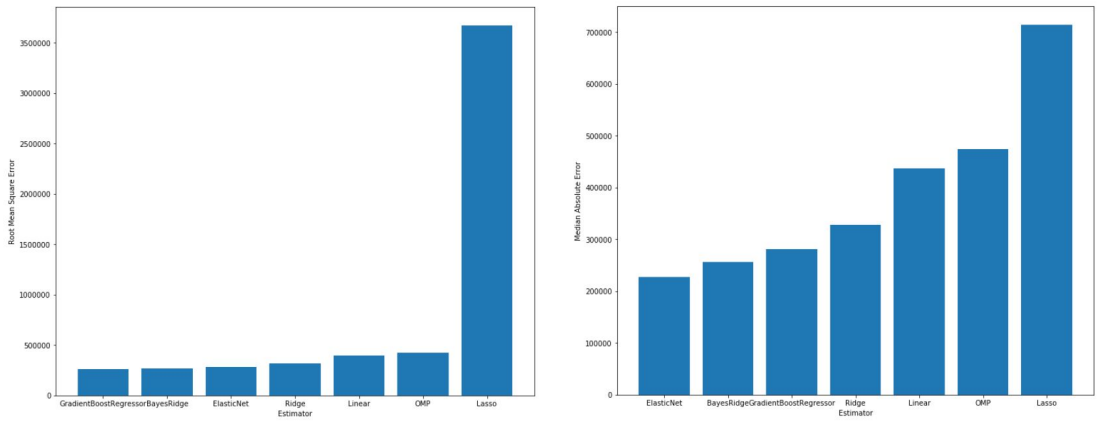


Figure 4: Comparison of regression models with evaluation metrics as RMSE and MAE

Regression Model	RMSE	MAE
ElasticNet	$278,851.437	$227,396.589
BayesRidge	$267,999.561	$255,789.832
Gradient Boosted Regressor	$330,144.639	$369,359.641
Ridge Estimator	$318,868.649	$327,853.379
Linear	$398,856.522	$437,108.744
OMP	$422,020.877	$473,896.455
Lasso	$3,671,197.058	$713,956.100

Table 2: Comparison of regression models with evaluation metrics as RMSE and MAE

Analysis of Results

The model is trained on the combined datasets. The housing dataset consists only a subset of the zip code values from the crime dataset. The prediction was made for the houses whose location was not already reported in the zillow's housing dataset. For every new address, the crime category features are extracted from the crime dataset and then applied on the model trained. This result was then compared with the other real estate agency websites like Redfin and Zillow.

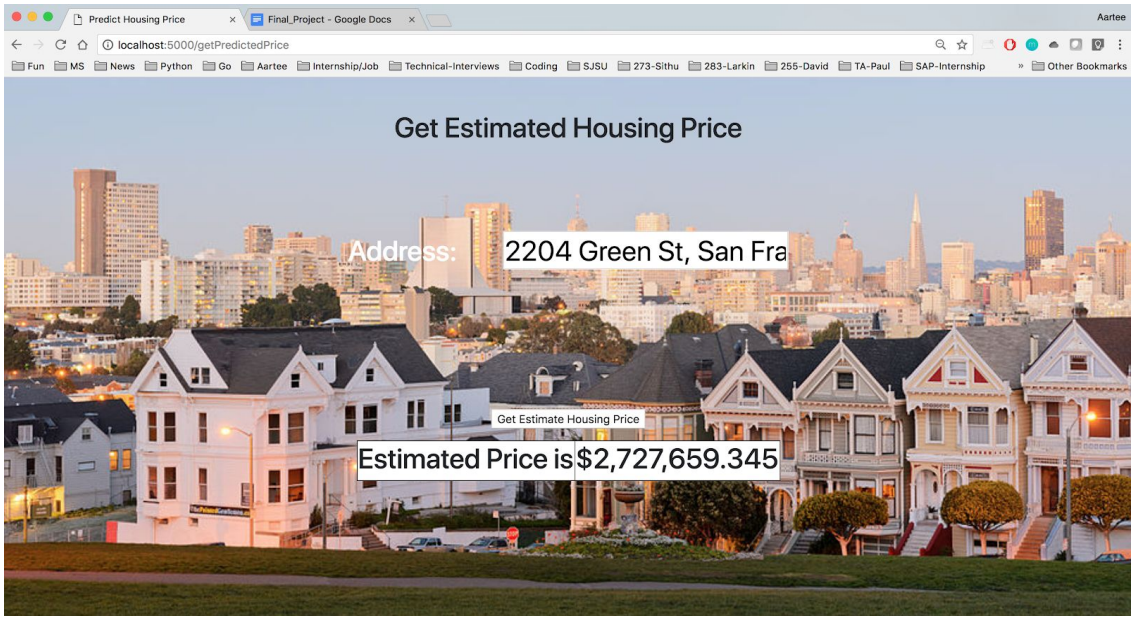


Figure 5: Screenshot showing estimated house price from our prediction model

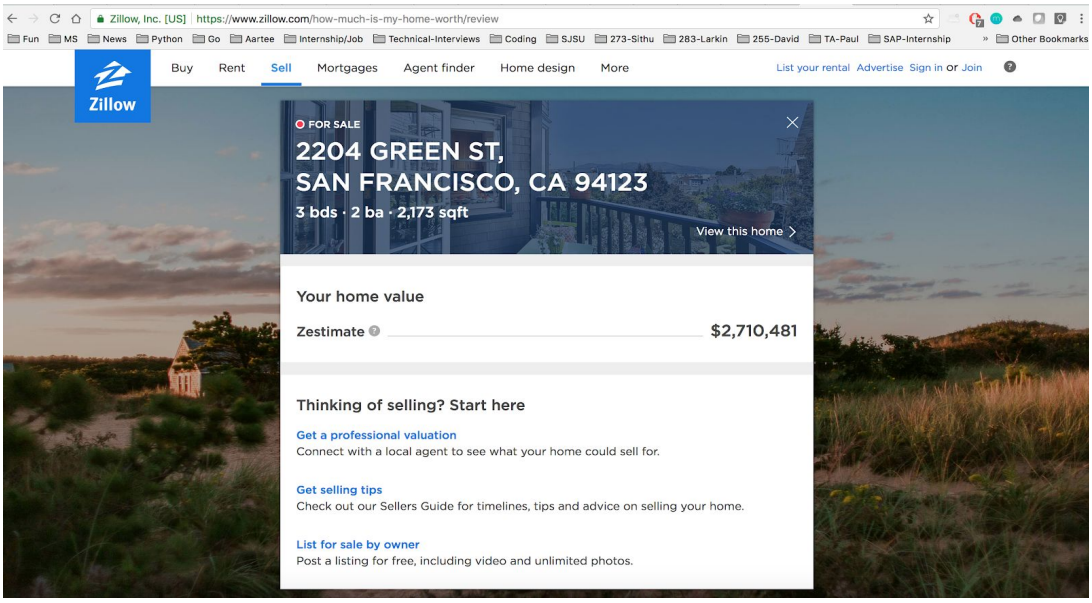


Figure 6: Screenshot showing estimated house price from Zillow website

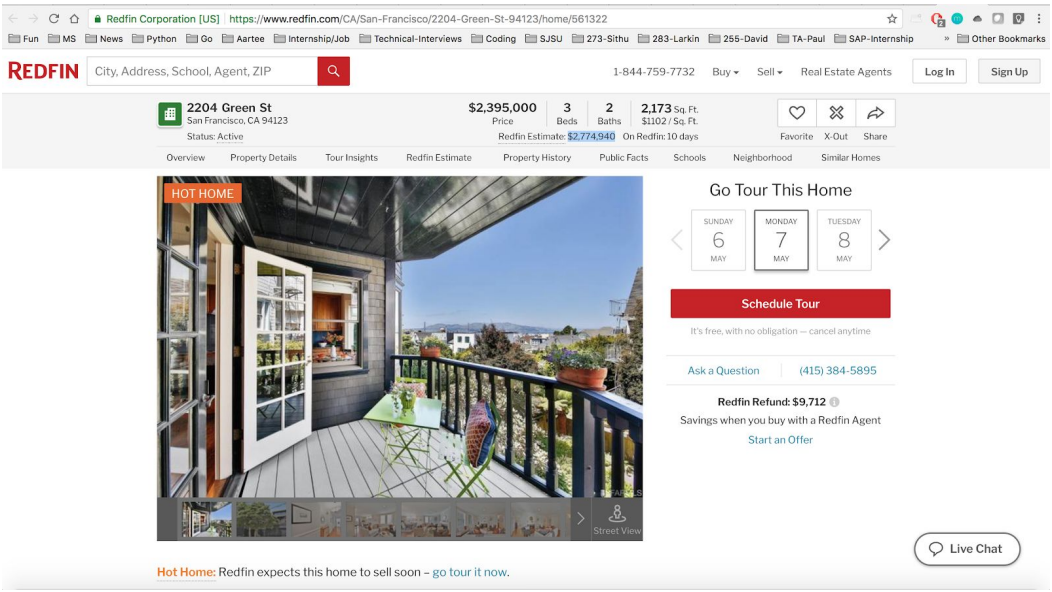


Figure 7: Screenshot showing estimated house price from Redfin website