SAN JOSÉ STATE UNIVERSITY

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

Final Project Report

Estimating Housing Values With Crime Incident Reports

Team 8

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

Spring 2018

Table of Contents

Chapter 1. Introduction

This project intends to find correlations of the criminal incidents with the house values and predict the price for a particular house based on the crime zone or zip code. We have used San Francisco police department incidents dataset for criminal reports and Zillow's house value index dataset for San Francisco house values. The criminal records dataset contains all of the San Francisco's zip codes and associated crime reports. On the other hand, the Zillow dataset has a limited set of listings in San Francisco. In this prototype, we plan to use the Zillow's dataset for training the module and test the model with listings from other agencies like Redfin and compare the prices.

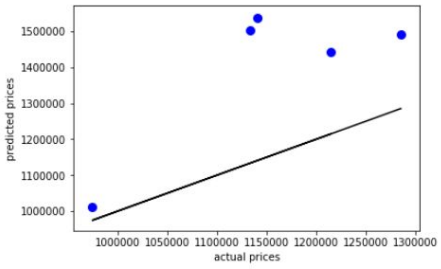
Chapter 2. System Design & Implementation details

Algorithms, Technologies and Tools

Seven regression algorithms were implemented and evaluated to predict the housing values. The Gradient Boost Regressor, Bayes Ridge and Elastic Net resulted with least mean absolute error value and RMSE 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.

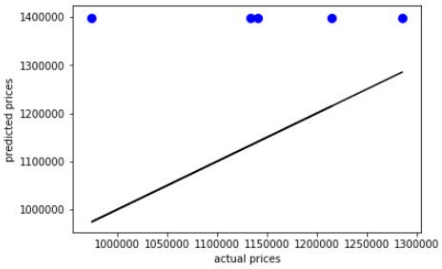
Elastic Net Regressor

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].



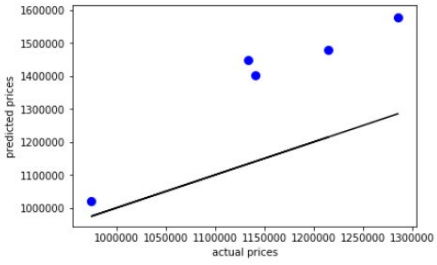
Bayesian Ridge Regressor

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].



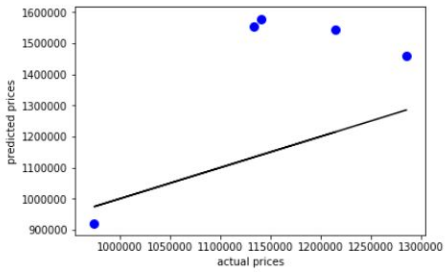
Gradient Boost Regressor

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

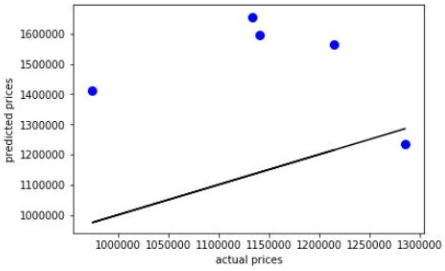
This model solves a regression model using linear least squares function as loss function and by regularizing[3] by L2-norm. This estimator has inbuilt support for multivariate regression by showing the effect of collinearity in the coefficients of an estimator.



Linear Regressor

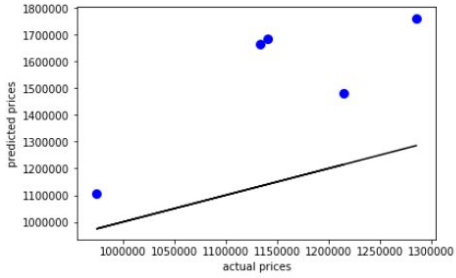
This model is for finding the relationships using linear predictor functions whose unknown parameters are estimated from the data. This model finds Ordinary Least Squares[1] that minimizes the residual sum of squares between the observed results in the 2D data, and the results predicted by linear approximation.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

.



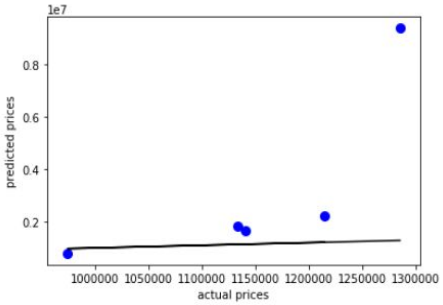
Orthogonal Matching Pursuit(OMP) Regressor

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).



Model Flow Diagram

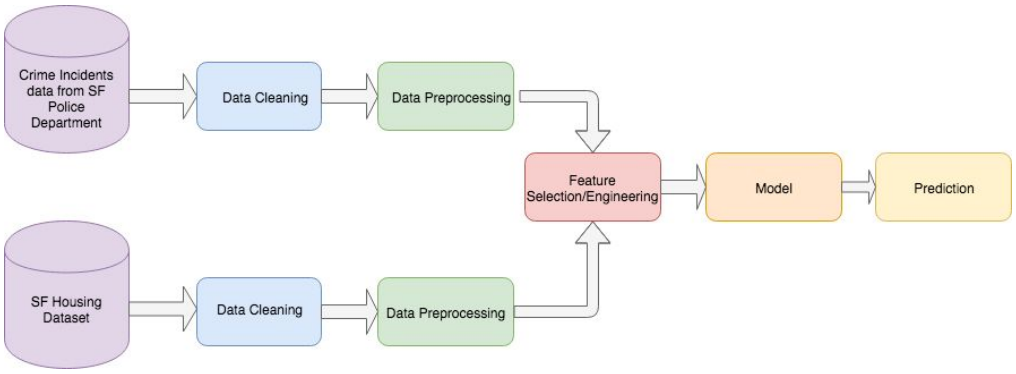


Figure 1: Model Flow Diagram

Chapter 3. Experiments and Proof of Concept Evaluation

Datasets

We have used two different datasets in this project. They are as follows:

1. Crime Incidents dataset from San Fransisco Police Department
2. Zillow house value index dataset

Datasets	Records	Dates from	Dates to
Crime Dataset	176980	01-2017	03-2018
Zillow Housing Dataset	14769	01-2017	03-2018

Table 1: Crime and Zillow Housing Dataset

Crime Incidents dataset contains 13 features. It contains criminal incidents reported by the San Francisco Police Department since 2003. For our use case we have filtered the record from 2017 until present.

Data Visualization

Crime dataset: The crime dataset has records of the crimes and its categories. We extract the categories of crimes associated with the location(which at this point is extracted from the latitude and longitude features). We can visualize the pattern or the distribution of the crime categories with respect to zipcodes.

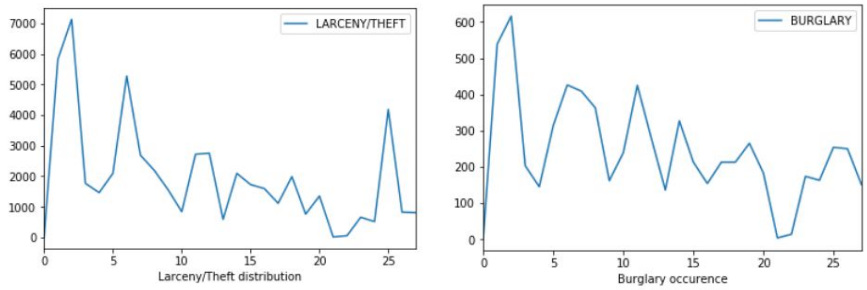


Figure 2: Data Visualization for Different Crime Categories

House value dataset: This dataset contains house value indexes directly associated with zipcodes. We have extracted the prices data for relevant year and trained the model. The two datasets were combined to visualize the crime rates and how the prices were varying with respect to the crime rates. We can observe in these figures that, the higher prices are definitely lying on the lower crime rate areas.

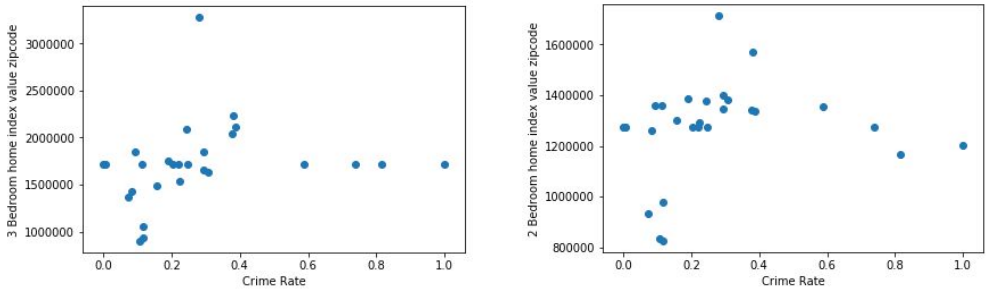


Figure 3: Data Visualization for Housing Prices w.r.t Crime Rate

Data Preprocessing

Crime Dataset:
The feature values for crime categories, latitude and longitude are extracted. We used uszipcode module's ZipcodeSearchEngine function to extract the zip code value for each the location features. The extracted data contains combinations of zip code and category. These rows are grouped by zip codes. The categories are aggregated for each group. This data is transformed to have the count features of crime categories associated with zipcode on the rows. This data is further cleansed for removing NaN values.

The dataset is normalized using min-max normalization technique so that each value is relative to its own column. This way a huge value in a particular feature will not be influencing the overall analysis.

Housing Dataset:
The housing data from Zillow contains house value indexes from past transactions. This data is spread across last few decades. For prototypical purposes, we have only retrieved last year's data. This data is aggregated for the span of an year since the values are ranging in months. This data is then combined with the crime dataset on the zipcode as reference variable.

Both the datasets are combined on the zip code. The data is visualized for correlation. negatively correlated features are selected to be dropped. The zipcode is eventually dropped from the dataset as it is a nominal value and it doesn't contribute to the analysis by itself.

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.

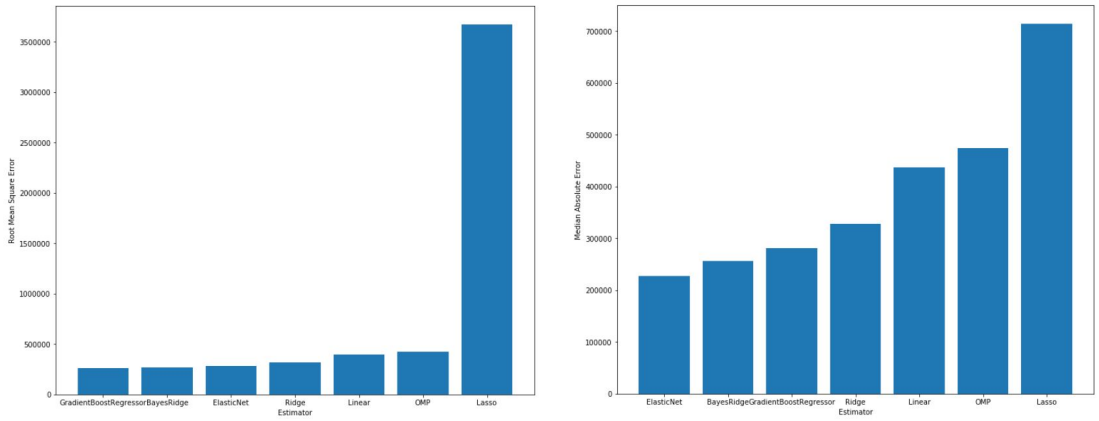


Figure 4: Comparison of regression models with evaluation metrics as RMSE and MAE

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.

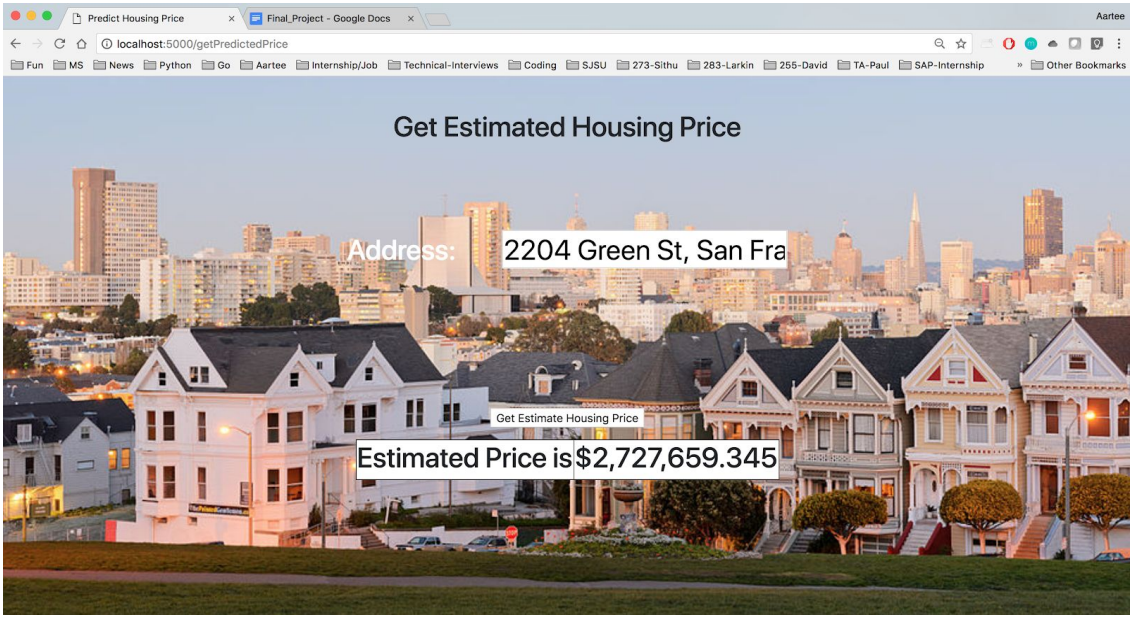


Figure 5: Screenshot showing estimated house price from our prediction model

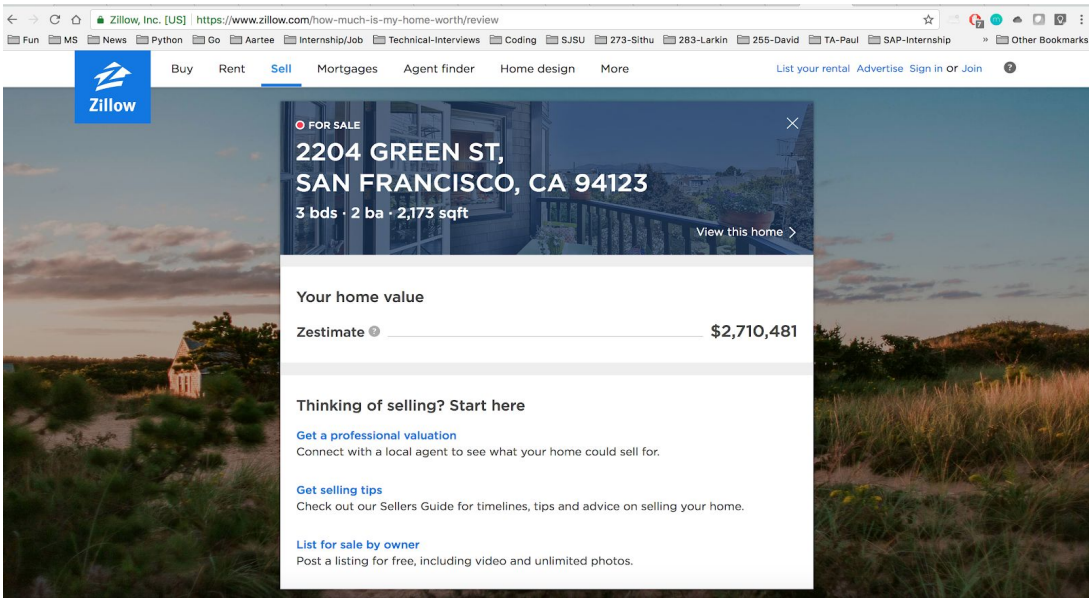


Figure 6: Screenshot showing estimated house price from Zillow website

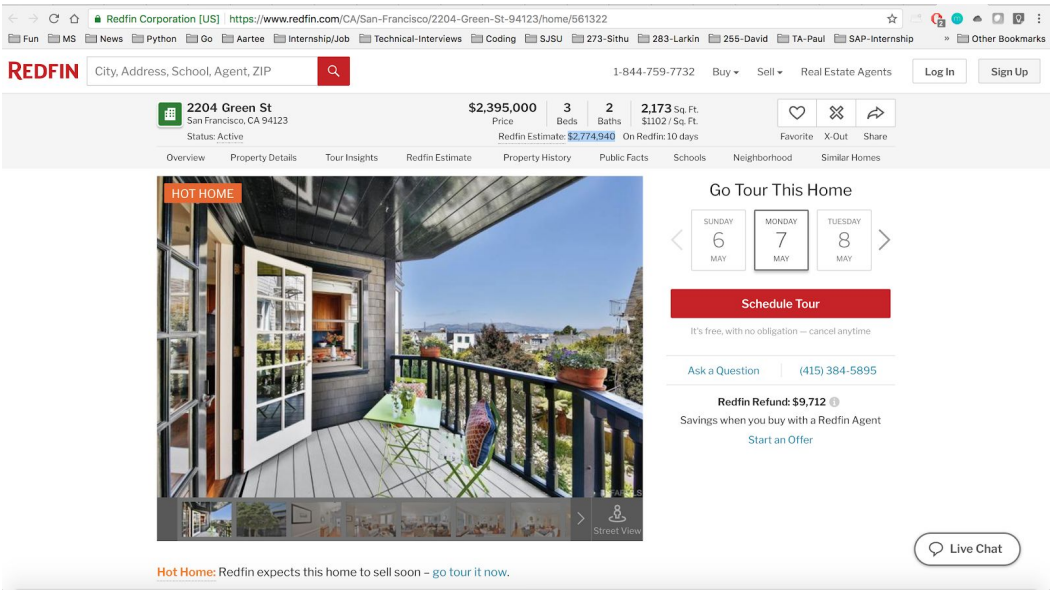


Figure 7: Screenshot showing estimated house price from Redfin website

Chapter 4: Discussion and Conclusion

Decisions, Difficulties and Discussions

The team decided to use the crime records dataset of the San Francisco police department. Furthermore, we were motivated to develop a use case of predicting housing values based on the crime rate of the neighborhood. Theoretically, the housing rates must be directly correlated with the crime rate. However, the analysis exposed certain anomalies for which the theory does not apply. There were housing records with higher price values although the neighborhood had relatively higher crime rate. The reason for this may be due to the choice of the city being San Francisco. The use case may be applied to other cities for a better analysis.

Conclusion and Future Work

This project was intended to build a predictive model to solve a regression problem of predicting house sales value based on the crime rates of the neighborhood. We used San Francisco crime incidents dataset for the criminal records dataset and the Zillow house value dataset for housing data set. We performed data cleaning, data preprocessing tasks before applying the model. We trained and evaluated a handful of regression models to choose the best applicable algorithm. We used a basic front end web page to accept the address of the house and predicted the rate for the house. The results were compared with other real estate agency websites like Redfin and Zillow.

Future work involves adding more insights into the dataset. In our use case, San Francisco house rates are influenced by many other factors other than crime. We plan to add additional information about the location, for example, nearby hospitals or schools in the neighborhood and so on.

Chapter 5: Project Plan/ Task Distribution

Initially, the team decided to visualize the crime dataset and to come up with a relevant use case to apply on the crime dataset. Pavana undertook the task of data cleaning, pre-processing on the Crime data set and Aartee took over the same task for the Housing dataset. Chia Chin Wu took the task of creating the visualization of the crime datasets in R as she was not interested in developing in python. Since the project was developed mainly on python, Chia Chin Wu could not contribute much to the project. Pavana undertook the task of combining both the datasets for training the model. Different regression models were used to train the model. Pavana, worked on OMP, Lasso, Linear and Gradient Boost regression models and Aartee worked on ElasticNet, BayesRidge, Ridge Estimator. The front end was developed with equal contribution by Aartee and Pavana.

Project Github Repository: https://github.com/Aartee/housing-price-prediction

References

[1] "Linear Regression Example — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html. Accessed 6 May. 2018.
[2] "1.11. Ensemble methods — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/modules/ensemble.html. Accessed 6 May. 2018.
[3] "Plot Ridge coefficients as a function of the regularization — scikit-learn ...." http://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html. Accessed 6 May. 2018.
[4] "Orthogonal Matching Pursuit — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/auto_examples/linear_model/plot_omp.html. Accessed 6 May. 2018.
[5] "Bayesian Ridge Regression — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/auto_examples/linear_model/plot_bayesian_ridge.html. Accessed 6 May. 2018.
[6] "Lasso and Elastic Net — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_coordinate_descent_path.html. Accessed 6 May. 2018.