A prediction model on factors driving Austin city housing prices

Background

This is a machine learning project done by following students from The University of Texas at Austin. Arnob Mallick, Ricardo Dunia and Kevin Cherian. The code and data can be found here.

Introduction

> *Predicting housing prices is one of the most common machine learning exercises suggested for beginners. Data is readily available and is highly relatable. The features are intuitive and easy to work with. And a good model for housing prices is extremely valuable. However, most of the examples online focus on building models based almost solely off of the physical properties of a house—four bedrooms,* 3.141592 *bathrooms, has a swimming pool, gated-community,* 9001 *square feet, duplex, zip code, etc. We wanted to see if we could bring something more to the table.*

From the start we knew we wanted to focus on Austin. Austin has a high resident turnaround and highly active real estate market due to the college/government floating population and a steady influx of high-tech entrepreneurs and young professionals. Such variety and dynamism makes Austin an excellent target for an analysis of real estate market trends. Accurate models providing insight into the housing market in Austin has the potential to be a very valuable tool for things such as speculative real estate or business investments. Searching for Austin datasets we found a large dataset containing all 311 calls in the Austin area starting from 2014. We also investigated into several other data sources such as population growth, points of interest, etc. We hope to be able to quantify the effect of crime proliferation and these other features on residential house prices and also help predict future appraisal values of residential homes based on current city trends.

Data Collection

A major challenge in this project was to collect various relevant data specific to Austin city and combining them to arrive at our predictions. The data was sourced from multiple locations.

Housing Data

Austin city housing data was scraped from a leading real estate website. We captured the response details of a RESTful API response used to populate the map interface of their website and extracted the data as a JSON and converted it to a CSV file. A total of 61818 records with 17 features were collected for 63 Austin city zip codes. We collected details of the sold houses for the last 10 years.

Crime / Incident Reports

This dataset has details of all 311 incident calls reported in Austin city since 2014. The database is updated almost in real time. The crime data is categorized and includes the location of the crime with latitude and longitude. In order to join the location-based crime data with the housing data we segmented the crimes into a 200 by 200 grid. This process is detailed later in the report.

Left: Bike theft incident distribution, heat map. **Right:** Bike theft incident per month

Population Growth

We scrapped the population count and projected increase for next five years from a state website for demographics, the data was collected per zip code and we joined this to the housing data with zip code as index.

Property Appraisal

The property appraisal data was collected from Travis Central District Appraisal website using Selenium automation script, to capture data for 63 Austin city zip codes. This data was joined to housing data using the

address line details. Joining data on address strings is not very straightforward. After parsing the various attributes of the address, we leveraged the FuzzyWuzzy library to match street names and relied on matching street names and zip codes to resolve conflicts. Because of this we are not entirely confident that our appraisal data is completely accurate, but based on a light manual validation it appears to be reasonable enough to use in our exploration of housing data.

Points of Interests in a neighborhood

We collected data on various types of points of interest (POI) such as schools, fast food joints, parks, and events from a well known recommendation engine's API. We used the same principal of dividing the city latitude and longitude into a grid and made queries to the POI API for each cell in the grid.

```
baths                float64
beds                 float64
city                  object
daysOnMarket         float64
hoa                  float64
latitude             float64
longitude            float64
lotSize              float64
salePrice              int64
pricePerSqFt         float64
lastSaleDate          object
sqFt                 float64
state                 object
stories              float64
streetLine            object
yearBuilt            float64
zip                    int64
Population Density     int64
park_index             int64
schools_index          int64
event_index            int64
fastfood_index         int64
Address               object
Number                object
2019 Appraisal         int64
2015 Appraisal         int64
5YearIncrease          int64
dtype: object
```

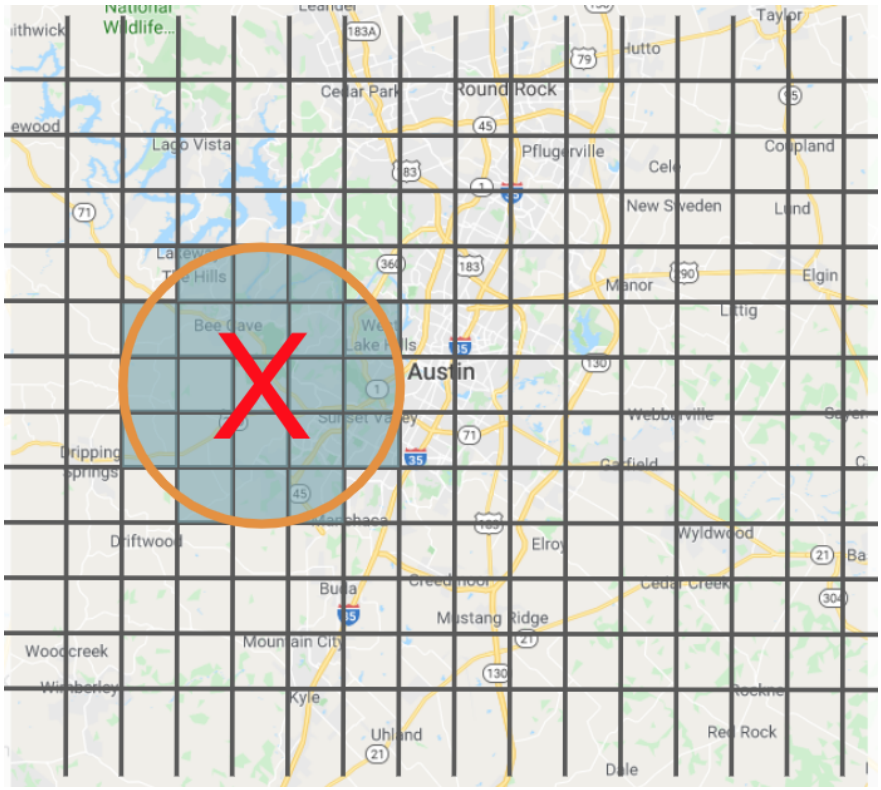
Complete Data with Data Type

Data Engineering And Analysis

Location Mapping

Most of the data we collected had latitude and longitude like housing, crime and points of interests. The challenge in joining these datasets is that latitudes and longitudes will almost never match up exactly between two datasets. This means that the only way to build an association between two locations on a map is to calculate the distance between them. Conceptually this is pretty trivial, but it brings up another challenge. How do we avoid having to calculate the distance formula potentially many billions of times. To join these datasets we

decided to break Austin up into a 200 by 200 grid of regions. Each region represented roughly a quarter-kilometer square of area. Once that was done, each dataset was segmented into smaller datasets—one for each cell in the grid. With all the data segmented, we took our optimization a step further by pre-computing an index for all of our location-based indices in each cell. The index for crime, for example, was computed simply by counting a weighted sum of all crimes reported in a grid cell. We will discuss the crime weights later in the post. A similar calculation was done for parks, schools, events, etc. Now for each house, instead of iterating through all crimes, parks, etc. we just have to determine which cell in the grid that house falls in, and sum up the indices from all the cells in a desired radius around the house.

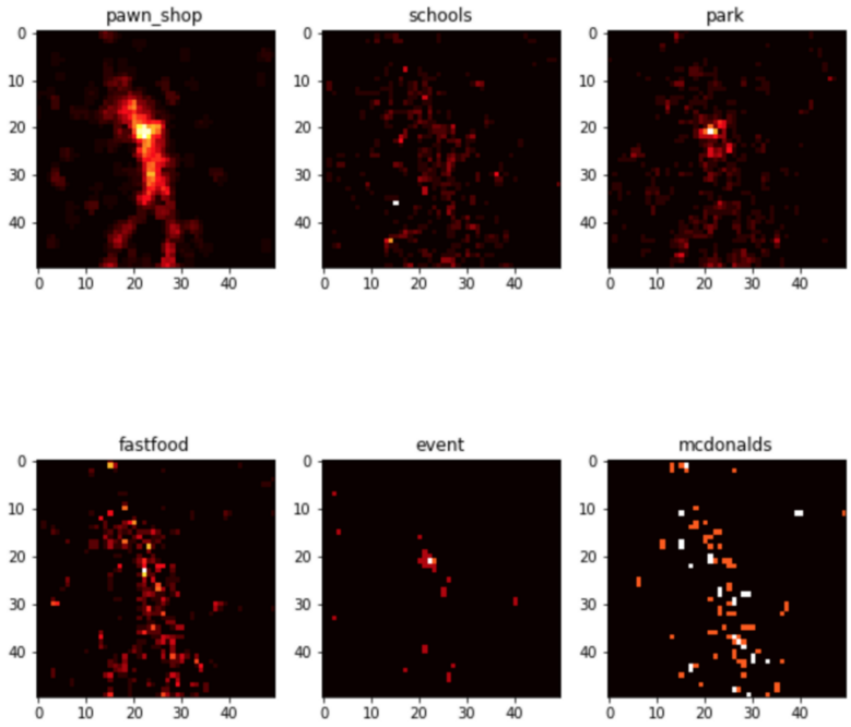


A sample of of how the grid was created, with X marking the location of a house.

Density Adjustment

It quickly became apparent to us that the location based data we were acquiring had the same general shape and profile. In order to derive individual insights from the different types of indices like schools versus fast food joints, we would have to make some adjustments. The first thing we did was scale the values by population density to give

each point of interest a weight that reflects the number of people it can serve. For example, a property in the city center of Austin probably has access to many more parks, but that property is sharing that resource with a lot more people. Second, to overcome the fact that most points of interest are distributed around Austin in approximately the same shape, the score added to each property has to be weighted by the distance to the point of interest. The effect of these adjustments is reflected in the cross-correlation matrix shown below. There is a much stronger correlation between the unscaled values compared to the values that have been scaled by population density.



Density of various Point Of Interest (POI) we collected.

Data Cleaning

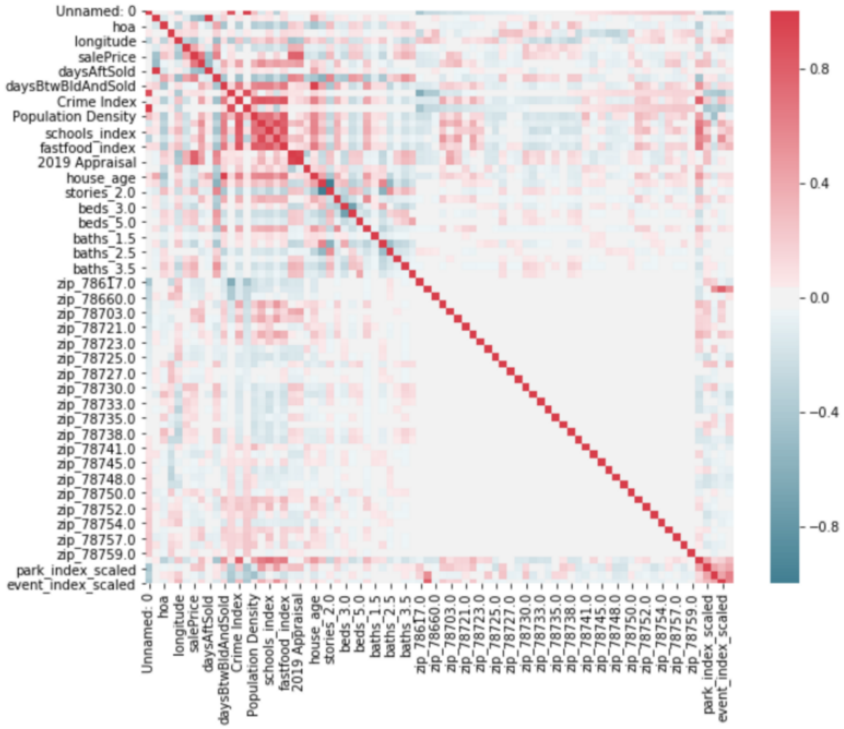
Many home sales transactions lack of complete data information. Part of the data preparations consist in the elimination of transactions with not enough house information to infer its value. We have eliminated transactions with missing data in appraised value/sales price, no bedrooms or bathrooms information, no living area surface (sqFt) and no appraisal data for 2015 or 2019.

Outliers

The housing data we collected had incorrect values for lot sizes, we cleaned up all the low or high values based on the value of living area. We also saw some *very* large and *very* small outliers in square footage and price per square foot that were either not residential properties or data errors in the source. We cut those columns off at a few standard deviations around the mean to get rid of the erroneous data. The HOA values were missing for most of the properties, so we assumed a value of zero for missing values. This largely skewed the data in favor of zero which could have unintended effects on the models. The appraisal data was matched based on the address using street name and house number, we know there are several houses that were not matched correctly, however many of these errors resulted in computed appraisal growths between 2015 and 2019 that were far outside the normal range for Austin so they could also be removed.

Attribute Correlation

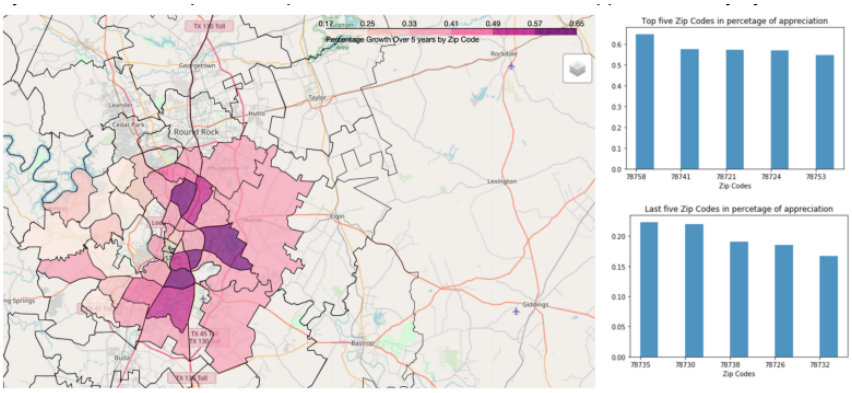
The correlation matrix with the cleaned up data is shown below. It is noticeable that features related to population and location attributes are highly correlated with each other. One promising observation is that the cross-correlations between population density and all the location-based indices is stronger in comparison to the same values after being adjusted for population density. This validates our decision to scale the location-based computed indices by population density to better evaluate their individual impacts on house value.



Correlation Matrix of features.

House Price Appreciation

Based on appraisal values of 2015 and 2019, here is a visual of the appreciation percentage by zip codes. Looks like if you had invested in a house in zip code 78758 it would have appreciated by close to 60% compared to zip code 78732 which would have appreciated only by 17%.



Left: choropleth map of appreciation by zip code. **Right:** top and last five zip codes based on appreciation over the last five years.

Feature Engineering

House sale transactions include the same property info available for buyers plus the information associated with the sale price, sold date and days on market (dom). However, many entries are missing as the data source comes from a variety of sources and real estate agents. A few transaction attributes were estimated or modified to create a meaningful feature, they are:

Days After Sold

DaysAfterSold determines the number of days from the date it was sold until today. Is calculated based on the day the house was sold, and helps to determine house appreciation in time.

Days Between Built and Sold

This determines how old was the house by the time it was sold, measured in days. Helps to determine the age of the house at the time of the purchase.

Lot Size

Lot size may have a significant impact on the house price. However, not all house descriptions includes lot size. Also condominiums have no lots, but common areas that are part of the condominium association. We in general estimated the lot size to be between 3 to 4 times the house square foot, and fill the missing entries to avoid losing too many entries.

Crime Index

The crime index is based on a linear weighted expression that accounts for all the type of crime incidents. There are over four hundred types of incidents, with weights varying from 1 (FRAUD) to 30 (CAPITAL MURDER). The weights were all subjectively assigned considering what we believed the impact of each crime type.

Models and Predictions

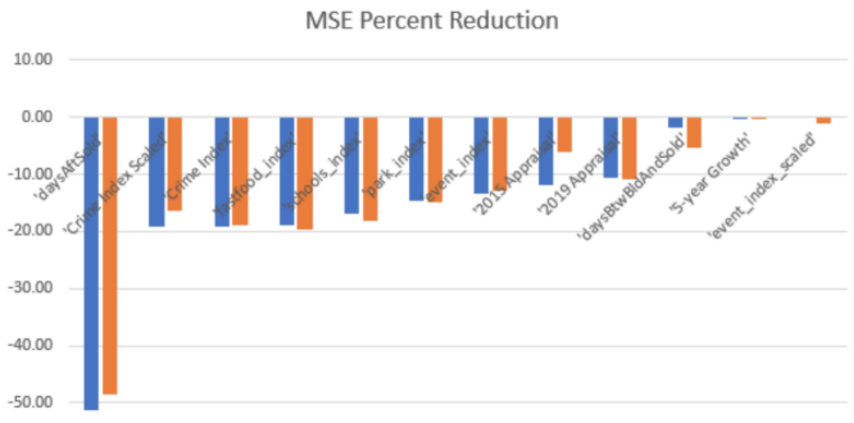
Predict Sale Price

Property sale price was initially predicted using basic information found in the sale transaction of a house—number of bathrooms and

bedrooms, living area size and lot size and zip code. Attributes were adjoined to the model based on their impact on the mean square error: the more the decrease on the sale price prediction error, the more relevant it is to include it in the model. We evaluated the impact of the following features individually against the baseline: '*daysAftSold*', '*daysBtwBldAndSold*', '*Crime Index*', '*park_index*', '*schools_index*', '*event_index*', '*fastfood_index*', '*Crime Index Scaled*', '*2015 Appraisal*', '*2019 Appraisal*' , '*5-year Growth*', '*event_index_scaled*', '*fastfood_index_scaled*', '*park_index_scaled*' and '*school_index_scaled*'.

Training and Testing data were defined using random samples from the 24K transactions considered in the data. Around 80% of the data is used for training, while the 20% remaining test the reliability of the model on data seen for the first time. Three fold cross validation (k=3) is used to select the best model using ridge regression. The weight parameter that balances the magnitude of the regression coefficients against the model error was calculated using Bayesian Optimization with 50 iterations. The coefficients range from 0 to 200, depending on the attribute in study.

The results comparing the inclusion impact of the different attributes in the regression model is shown below. Note that the "Days after sold", or how long back the sale was executed can reduce significantly the estimation of the house sale price. This is attribute is followed by the Crime Index Scaled in importance.



Percent Reduction on the Mean Square Error. The attribute "dayAfterSold' provides the largest drop in error estimation for the train set (Blue bars) as well as for the testing set (Orange Bars).

A second sales price analysis was made assuming a different regression solution for each zip code. This is equivalent to apply a decision tree with the zip code as categorical variable. Such an outcome allows to determine which zip code has appreciated the most when comparing the “dayAfterSold’ regression coefficient for the different house sales.

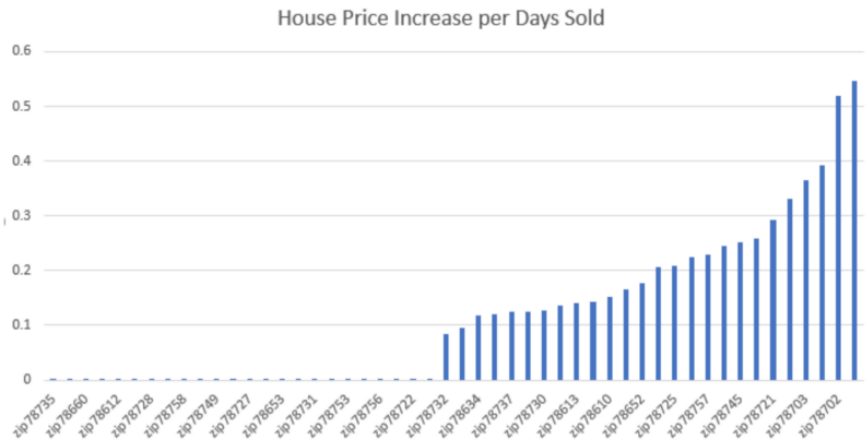
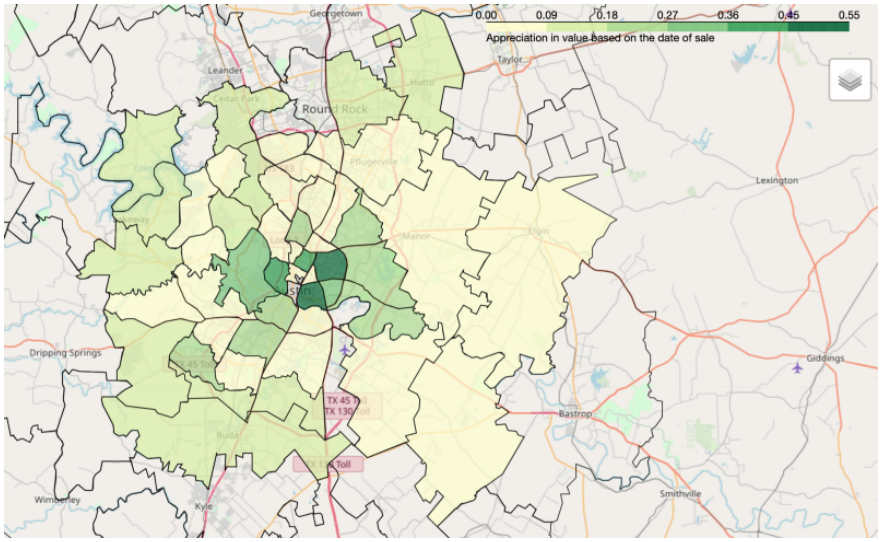


Figure above illustrates which zip codes have appreciated most (right) and least (left) in time. Vertical axis represents the regression coefficient for the days after sold attribute (all coefficients signs were changed to positive for convenience)



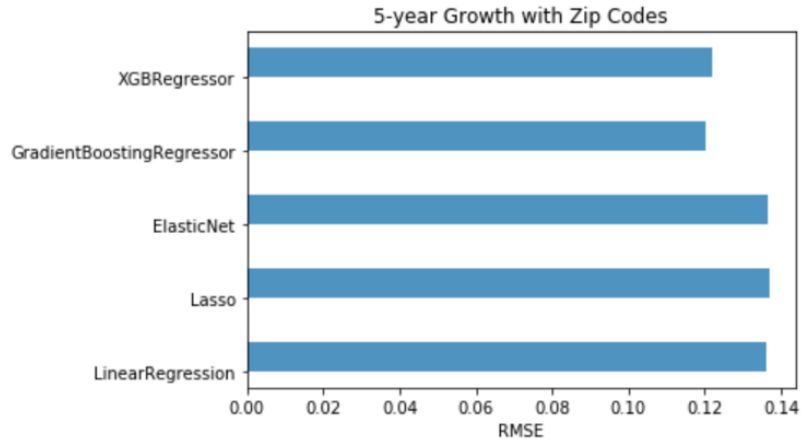
A choropleth map is represented by zip codes colored based on their sales appreciation.

Predict Five Year Appreciation

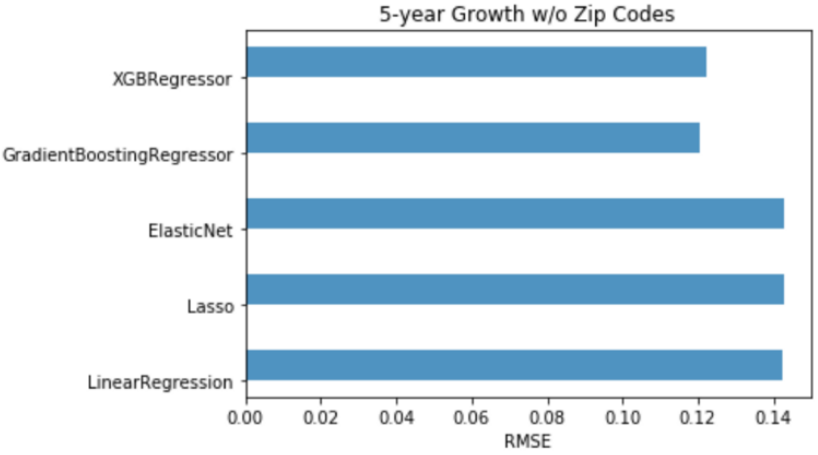
We tried to build a model that can predict the house price appreciation for the next five year based on the data we collected from the appraisal data of year 2015 and 2019. We tried various models and the prediction was accurate with the least root mean square error (RMSE)

of 0.119971 using Gradient Boosting Regressor model. Here is a graph showing the list of models we tried and its corresponding RMSE. We also built the same model with all one-hot encoded zip code columns removed and saw some striking results. Model performance was worse, but only ever so slightly.

```
RMSE: 0.135952
Lasso RSME: 0.136807
ElasticNet RSME: 0.136636
GBoost RSME: 0.120320
XGB RSME: 0.121847
```




```
RMSE: 0.142492
Lasso RSME: 0.142639
ElasticNet RSME: 0.142619
GBoost RSME: 0.120372
XGB RSME: 0.122116
```

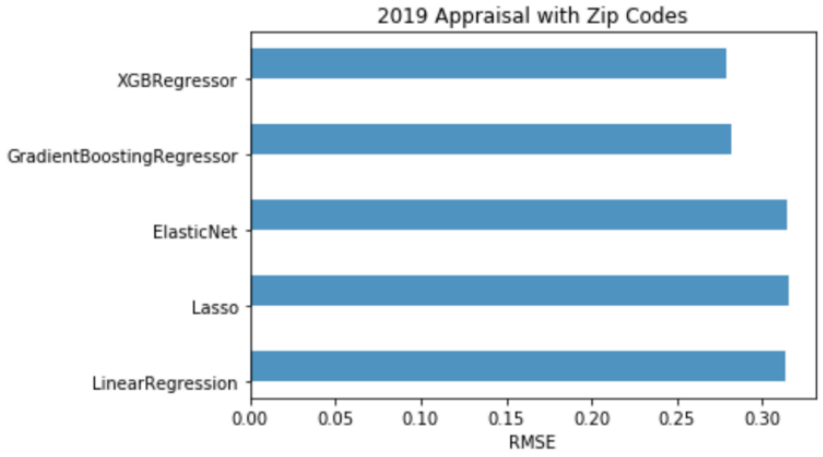


Predict 2019 Appraisal

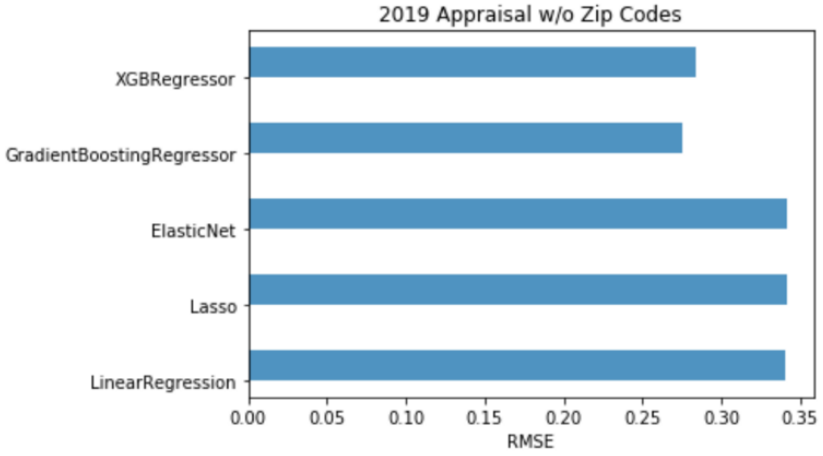
We were not convinced that the performance of the 5-year Appraisal model without zip codes proved that the location-based features we added were strong enough to replace the location context of zip codes.

The percent growth of Austin property values may have all risen relatively uniformly leading to a model that does not need to rely much on location. To evaluate whether or not the 5-year Appraisal model without zip codes was a fluke we also built models to predict the 2019 Appraisal of a property. Property values are much more heavily influenced by location, so this was a good test. We ended up seeing very similar results for 2019 Appraisals as well! Maybe there is some merit to the POI and crime data we added as features.

```
RMSE: 0.313955
Lasso RSME: 0.315006
ElasticNet RSME: 0.314819
GBoost RSME: 0.282193
XGB RSME: 0.279186
```




```
RMSE: 0.340951
Lasso RSME: 0.341238
ElasticNet RSME: 0.341204
GBoost RSME: 0.275420
XGB RSME: 0.283603
```



Conclusion

Models generated in this work provides an outlook of how house prices have changed, and will potentially change in Austin in the coming years. We have combined information from different data sites to estimate the effect of crime, recreation sites (events and entertainment) and property tax appraised value to measure their importance in the house pricing. The models developed in this work demonstrates that crime index and days since property was sold have a significant impact on the estimation of the property sales price. The number of sale transactions collected allowed to compare models developed per each Zip code. The coefficients generated by each zip code regression model shows which zip code have been impacted the most by property time appreciation. It can also provide information regarding areas that have been impacted the most by crime, parks proximity or events organized in their vicinity.

Some of the information grabbed from different websites were merged to the historical sale prices thanks to the property location, providing successful inside on house evaluation. This demonstrates that location coordinates may represent a significant feature to bring city information towards home evaluation predictions—to a more accurate extent than by simple zip code categorization. Finally, it is worth mentioning that the change in home appraisal value can be used together with the coefficients calculated in the sales price regression model to evaluate potential return on investment for property houses. Assumptions regarding crime ,school and recreation indexes for future years may impact the accuracy of such predictions, especially if these indexes are assumed constant for future years.

Future Work

A meaningful extension of this work will be to combine past sales price with the value appreciation, and correlate them with the days the property was for sale in the market, known as days on market (dom). This can demonstrate how house pricing is correlated to how quickly a property was sold. This concept is similar to the stock price and the volume of shares sold.

Past appraised value was one of the major info merged to the property sales information. This piece of information with a sale price regression

model could have been used to predict future risk or different scenarios in the house market.

In our model to predict the future appraisal value, was purely based on appraisal rates in the past, so we would like to add data from other models that help predict inflation, GDP and population growth as features to improve the accuracy of our model.

When dealing with real estate, the motto is always, 'location, location, location.' We are just scratching the surface with respect to features that contextualize a location. While we didn't manage to build a model that could surpass or equal the impact of including zip code as a one-hot encoded feature we were able to demonstrate that a model without zip code, but with contextual information about a location can compete reasonably well. Due to time we could not explore this area as much as we would have liked, but we believe that adding more carefully chosen location-based features can easily become a much better way to predict house prices compared to the current status quo.