Predicting Housing Sale Prices using Machine Learning: a San Francisco and UAE Case Study

Jaisal Friedman

December 12, 2019

Abstract

This paper outlines a study of the efficacy of predicting housing sale prices in San Francisco and the UAE using Machine Learning techniques. This paper presents a high-level summary, methodology, results, and discussion of the project. The paper seeks to answer How and to what extent can a Machine Learning model outperform a regression-based model in predicting housing sale prices for a specific location?

Table Of Contents

Project Summary

Introduction

Housing bubbles represent the bane of modern property laws. As seen in the US housing market crash in 2008, these bubbles leave ordinary people bankrupt, homeless, or without ownership of property. In areas of high housing demand and limited supply, housing prices often skyrocket. The divergence of housing prices can either be an indicator of a burgeoning housing bubble or a sign of increased speculative property value. Only time can distinguish these two. Accurately predicting housing prices in a given area may prove crucial in helping these highly demanded, limited supplied areas avoid a housing crisis. The UAE presents a unique housing market with government-backed oil funded developers, welfare-state grants and laws for Nationals, and a highly transient population. Since 2014, the UAE property market has been one of the world's only decreasing bubbles - where housing prices have steadily declined without a crash. This capstone will build a spatial-temporal machine learning model which accurately predicts from hedonistic, location, and other relevant features the sale price of a house. This paper explores the case study of predicting San Francisco housing sale prices using both regression-based and machine learning models. This case study serves as a precursor to the model development and analysis to be done here in the UAE.

Specific Aims

This paper aims to predict housing sale prices using machine learning in San Francisco. The hypothesis states that the machine learning models will outperform simple regression models for the same given dataset. Furthermore, the hypothesis states that feature additions including crime, school, and transportation will lead to better predictions. The San Francisco dataset is used as a preliminary example for the UAE dataset. The aim of the overall capstone project is to build a series of machine learning models to better predict housing sale prices. This will include a data preparation library for fine tuning inputs to the provided models. The models will be adaptable to both San Francisco and the UAE and thus would assumedly generalize to most city housing datasets.

Background

Predicting housing prices has become a popular introductory example for machine learning models. This was popularized with the Boston Housing Data, readily available on many open source machine learning libraries. Furthermore, in economics predicting housing prices has been classically performed with a variety of models including but not limited to: the linear asset pricing, hedonistic regression, and the repeat sales model. These models are effective at identifying multi-variate relationships between the input features and the output sale price. However, they often fail to truly predict sale prices at a degree of accuracy that is meaningful for market actors. For this reason, most market actors reply on complex machine learning models to accurately price property. For example, Zillow, a popular US-based property data platforms, provides a Z-estimate of each house listed on its platform. These Z-estimates estimate the current value of the home. According to Zillow research, the Z-estimate combines several million machine learning models together to generate these values. Furthermore, the Zillow research team tests over 65,000 new machine learning models each day to improve upon their prediction [1]. Chief Analytics Officer Stan Humphries states, "[the] Zestimate now has a median error rate of less than 2 percent for homes listed for

sale, meaning half of all Zestimates fall within 2 percent of the home's eventual sales price," [1]. This level of accuracy sets the benchmark for what is possible in the space of predicting housing prices. Zillow, however, has access to highly accurate and varied datasets across the entire United States. Nevertheless, such a prediction model does not exist for the UAE. The outcome of this entire project is to build a model with the median error rate of less than 5 percent in the UAE. The processes employed for building such a model have been conducted with San Francisco data. This aligns with the eventual goal of adapting the model to the UAE, once the data has been acquired. A qualitative research review can be found in the Appendix under the UAE Market Review section. This is highly informative for those not acquired with the idiosyncrasies of the UAE housing market.

Related Work

Econometric research and specifically housing price predictions are ripe for Machine Learning analysis given its multidimensional nature and outcome-driven scope. Neural network models have long been proposed in housing price research. However, modern techniques like Deep Learning and Random Forests have not been fully explored. Classically, the linear asset pricing model, the hedonistic model, and the repeat sales models have been used to mirror the interplay of the real estate market. In the literature review below, this paper will first address each of these models and second discuss current machine learning approaches and literature in housing markets and econometrics.

Classical Models

The first classical model is the linear asset pricing model famously applied across a range of financial markets. In Glaeser's Housing Bubbles literature review [2], he proposed the following model for a no-arbitrage housing price in a rational housing market: "the value of owning a home equals the benefits today plus the asset value tomorrow," [2]. Mathematically, this is represented below. $R_t + \frac{E(P_t+1}{1+r} = P_t$, where
P_t = price, R_t = net benefits of ownership, and $\frac{1}{1+r}$ = discount factor. The best empirical observations for the benefit of ownership is rent pricing. Furthermore, the net benefits of ownership can be modelled using a stochastic growth rate process, which, "predicts both mean reversion and at least modest momentum," [2]. This stochastic growth rate accounts for the indistinguishable nature of housing bubbles. It provides limited capacity to thus predict such bubbles.
Secondly, Bonnet, Bono, Chapelle, and Wasmer propose, smartly, rent as a truly indicative measure of housing capital. This is because, "for the value of housing capital to be consistent with the underlying theoretical analysis, the value must correspond to an actualized value of rent and not rely on housing prices," [3]. Returns on housing investments are complicated in nature; only actualizing in the case of re-sale or rent-based income. The incline in housing prices the last century may not thus be actualized, observing rent changes over the past century clarifies that this may not be the case.
Thirdly, hedonistic models can address housing market prices and thus housing bubbles. "Observed product prices and the specific amounts of characteristics associated with each good define a set of implicit or "hedonic" prices," [4]. Clearly, this proposed model is highly applicable to the housing market where hedonistic characteristics can be broken down into: location, amenities, number of rooms, etc. Hedonistic characteristics provide a strong basis for Machine Learning based analysis. The advent of these models allows non-linear predictions to be made based on a set of independent "hedonistic" variables.
The final model is the repeat sales model which measures the relative changes in housing prices over

time. Harding, Rosenthal, and Sirmans argue that, "housing depreciates at roughly 2.5 percent per year," [5]. This was the first study to use repeat sales to directly measure such depreciation using data from the US housing market between 1983 and 2001. The repeat sales model is critically accurate in its tracing of housing bubbles. This is particularly useful for extracting indicative characteristics of a bubbly market. Their paper highlights the non-linearity of depreciation in true markets. This poses a deep challenge to studying such repeat models without inclusion of hedonic characteristics and housing market health in time.

Machine Learning Discussion

Machine Learning methodologies have a wide variety of unexplored applications to econometric problems. These invariably include housing markets and bubbles. Mullainathan, Sendhil and Spiess summarize the econometric problems ripe for machine learning algorithms, use a housing pricing prediction example, and provide a general overview on how to optimize machine learning models for the topic at hand [6]. The case of housing bubbles would fall under the authors scope of econometric problems: as housing bubbles contain non-linear independent data points with multi-layered interaction for a uni-dimensional outcome (how overpriced is the market).
Their example of a housing pricing prediction model performed 4 regression-based machine learning models on a dataset of 150 characteristics of houses and their associated list price. The outcome of such a study showed the prediction performance for each algorithm, thus making the case for the Random Forest methodology over the Regression tree tuned by depth, the ordinary least squares, and Ensemble methodologies. The authors argue that, "the very appeal of machine learning is high dimensionality; flexible functional forms allow us to fit varied structures of data," [6]. But how does machine learning then handle out-of-sample prediction? If the dataset is only a subspace of all possible values (and market conditions), then how can such a function be of use on testing data? The paper offers two primary solutions: regularization, cross-validation, and the right loss function selection [6].
In an example study, Park and Bae perform machine learning algorithms to predict if housing listings sell under or over their ask price in Fairfax Country, Virginia. [7]. Their study selects hedonic characteristics from the national realtors association dataset, trains a fraction of the dataset on the RIPPER, Decision Tree, AdaBoost and simple Baynesian models, and tests these models for a best approach. The RIPPER model outperforms the rest. The study uses a particularly useful technique for segmenting testing and training data for a comparative approach. However, the study remains extremely limited in scope as it simply predicts whether a property listed will sell under or over its ask price. This binary output can easily and more usefully be swapped for a continuous prediction output.

Time-Based Encoding For ML

Haries, Michael, Horn, and Kim created a machine learning model which makes predictions based on a number of features, including past prices, if an SPI contract will go up or down in the next trading period [8]. They utilize a neural network model with concept drift, cross validation, and a regular loss function. The time factor is encoded with past price attributes such as "price five minutes ago" to "price five hours ago". Their results yielded well over 60 percent accuracy which is highly significant. They use concept drift to reduce the noise the model generated. Another popular approach to time-based encoding is a Support Vector Machine popularized by Tong, Simon, Koller, and Daphne's paper [9] and examined in Mukherjee, Sayan and Osuna, Edgar and Girosi, Federico paper on nonlinear choatic time predictions [10].

Methodology

Research Question

How and to what extent can a Machine Learning model outperform a regression based model in predicting housing sale prices for a specific location?

Hypothesis

There are two primary postulates to the hypothesis to be evaluated. The first postulate of hypothesis states that the machine learning models will outperform simple regression models for the same data set. These machine learning models are specified to be the random forest model (extremely randomized trees) and the leading AutoML model (which is an ensemble technique). The second postulate of the hypothesis states that feature additions to the data sets including crime and schooling indicators will lead to better predictions, more precision and accuracy, in the machine learning models. Furthermore, the addition of a single feature, the original list price, will vastly improve the models accuracy.

Approach

The approach begins with research and collection of the ideal data set, specified in the 'Ideal Dataset' section below. This type of data set was unattainable within semester timeline. For this reason, the San Francisco data set and example has been used. Once the data set is collected, the first step is to understand and clean the data set. This may involve filling missing data, finding missing data, identifying outliers, and removing outliers. The next step is to pull in important features. This is heavily biased to the city in focus. For the US, crime and school quality are two crucial aspects to understanding housing prices. In Singapore, proximity to public transportation is crucial to understanding housing prices. In Abu Dhabi and Dubai, proximity to Sheik Zayed Road may be critical to housing prices. Next, visualizing the dataset is performed. Reference to specific methodology can be found under the 'Feature Visualization' section below. Visualization is important for fine tuning models. Finally, feature encoding is performed to allow categorical data to be interpreted by the models used. Feature selection is optionally performed to reduce the dimensionality of the data space. This is critical for one-hot encoded data as the dimensionality can grow exponentially. This can prove ruinous for the models accuracy. Once data cleaning is complete, the data set is split by training and test data. The split is generally: 0.8, 0.2, respectively. The models ingest the training data and output the scoring metrics based on the test data. Finally, model performance is conducted to evaluate the results and impact.

Models

The following models will be used to evaluate the hypothesis.

Linear Regression

A linear regression model will serve as the control model and the null hypothesis rejection. This is often the standard that is used in Economics. The outcome of this paper is to verify that standard models can be vastly improved by simple machine learning techniques.

Random Forests

The random forests model will serve to display the efficacy of a simple, fast, and highly flexible machine learning model. The random forests will be run with a constant depth of 50.

Ensemble Techniques

The AutoML ensemble techniques model will serve as the gold standard of machine learning technology. Such a model is the bar for an optimized model to outperform. It displays the range of possibilities when using machine learning techniques for housing prices [11].

Data And Data Engineering

The data for San Francisco sales was collected from the U.S. National Realtors Association [12]. The data set includes all single-family home sales between 2007 and 2017.

Ideal Data Set

The ideal data set includes all housing sales in a recent time frame. Within the data set the column space includes: hedonistic features (number of rooms, number of bath, square footage, number of bedrooms, etc.), location data (latitude, longitude, neighborhood, zip code, etc.), important features (number of crimes in the neighborhood, school ratings in the neighborhood, number of blocks to public transportation, etc.)

Neighborhoods And Spatial Join

There are 104 Realtor neighborhoods that were represented in the data set. These neighborhoods are used by real estate regions to subdivide the zip coded areas of San Francisco. Thus, the paper assumes that these neighborhoods encode a non-linear transformation of the location data in San Francisco. For this reason, the neighborhoods were added to each housing sale in the data set. Furthermore, schooling and crime data was spatially joined to the data set by aggregating on each neighborhood. This technique was employed for both ease and accuracy. However, better methods are possible and could be employed in future iterations of the project. One possible extension could be approximating the density of the given indicator at each location of the housing sale using GPU programming.

Great Schools API And Data

Schooling data is a known indicator of housing prices in the US. This is often due to the nature of US public school zoning laws and large discrepancies in the quality of public school between neighborhoods. Thus, housing prices can be deeply influenced by school data. To incorporate this type of data in my models, I utilized Great Schools API to get data on all schools in San Francisco [13]. The data included the location (latitude and longitude), a great schools rating [14], the number of public, private, and charter schools, the number of enrolled students, and the public school district the school was located in. Some of the data was missing or incomplete. About 40 percent of the schools did not have a Great schools rating. About 25 percent did not have the number of enrolled students. The average value for each respective indicator was used to fill the missing columns. This was the technique employed for its ease, further discussion can be found in the filling empty data section. There was 32 neighborhoods that did not have any school data and thus about 3000 housing sales that did not have any school data.

Spatial Binning Of School Data

The school data was spatially binned by Realtor neighborhoods using a GeoJson file of San Francisco Realtor neighborhoods provided by OpenSF [15]. The average enrollment and Great Schools rating and the total number of private, public, and charter schools per neighborhood were generated. This data can be viewed in the repository.

Crime Data

Crime is another known indicator of housing prices in the US. One can imagine that houses in relatively safe areas will be more likely to be higher priced than houses in relatively unsafe areas. To test this assumption and include such indicator in the models, the San Francisco police crime data set was gathered. The data set is publicly available [16]. The data contains all crimes from 2003 to 2018. This is over 2 million rows of individual crime data. A 20 percent random sample of the entire data set was used for simplicity and faster processing.

Spatial Binning Of Crime Data

The crime data was spatially binned by Realtor neighborhoods using a GeoJson file of San Francisco Realtor neighborhoods provided by OpenSF [15]. The crime data included over 50 different types of crimes. To distinguish between crimes, the number of incidents of each type of crime was included in the spatially binned data set in addition to the total number of incidents per neighborhood. The data can be viewed in the repository.

Filling Empty Data

Empty data was filled using the average value for the given indicator. This is a common technique. However, it often leads to inaccuracies and incorrect assumptions in the data set. A better methodology, which could be employed in the next iteration, would be to build a prior distribution on the empty data. This could be done by training a Bayesian model to predict the empty data indicator. Then, when an empty piece of data is encountered, the trained distribution's MAP or MLE on the empty data column for the given data row is used.

Feature Visualization

The feature visualization has a few specific outcomes: to understand uni variate and spatial distributions of features in the data set and to test mutual information and correlation coefficients between features and the target variable. This will help better inform model outcomes and important features for predicting the sale price. This will also help detect outliers. Lastly, this will help decide on the removal of certain features from the data set in the feature selection section. Low variability and low mutual information of features often results in little impact of a feature on the predicted target variable. Presentation of the feature visualization and specific learnings for the San Francisco data set can be found in the results section.

Feature Encoding

Feature encoding is vital for processing categorical or ordinal data in the data set. Two common techniques are one-hot encoding and label encoding. Label encoding creates adds an inherent weight to the categorical data, while one-hot encoding does not. One-hot encoding results in high dimensionality and is often used in conjunction with PCA or another dimensionality reduction

technique. Label-encoding does not result in high dimensionality but will only yield results for Random Forests and Decision Trees. For other methods, it fails as the data is, as mentioned, ranked. For this reason, one-hot encoding with PCA was used for the linear classifier, while both one-hot encoding with PCA and label encoding was used for the machine learning models. It is important to note, that AutoML handles categorical data with a unique variety of methods. This is part of its functionality as an out-of-the-box methodology.

Feature Selection

Feature selection is the process of reducing the space of dimensionality in the data set before training the models. It often results in faster training and reduces the VC dimensionality of the target set. For this reason, it is optimal for large or sparse data sets with high dimensionality. This is another extension, that the entire project hopes to capitalize on when a larger data set is obtained. As discussed in the results section, for the San Francisco data set, feature selection was avoided. This was due to poor performance in the preliminary machine learning models after feature selection. Perhaps, a more in-depth understanding of how to select features was required to accurately select features without reducing model accuracy.

Results

The results from cleaning, visualizing, and encoding the data and from building Linear Regression, Random Forest and Ensemble Methods on different data sets are presented below.

Data Cleaning, Visualization, And Encoding

The data was obtained from the US national REALTOR association as cited in the above data section. The data set included 24,000 single family homes sales between 2007 and 2017 in San Francisco.

Data Cleaning

The data had several missing fields. These included: location (latitude and longitude), elevation, type, and zoning. The missing location and elevation data were forward geo-encoded based their full address using the Google Maps API. The type and zoning were filled with OTHR, as the standard marker in the data set for non-standard values. The entire data set was then spatially filtered using a geojson map of San Francisco. Around 250 entries in the data set had locations outside of the city limits. This led the final data set to contain 23,720 rows. The data set was merged on the Realtor neighborhoods with the crime and school data as discussed above in the data section.

Data Visualization

The data visualization was conducted to better understand the distribution of the data and the relationship between the target variable (list price) and the feature variables. Firstly, the data was split into categorical and numerical data types. The top 10 mutual information features for the numerical and categorical features on sale price was generated. This gave an indication of linear dependence of features and the sale price. Figure 1 in the appendix displays this. The top 2 were

Area/Sub-district number and number of Assaults in given neighborhood for numerical features and Street Name and Neighborhood for categorical features. It is worth noting that original list price was excluded from this analysis. It is known that this would hold the highest mutual information with the target variable. Finally, the data was spatially visualized to understand the distribution across the city itself. Figure 2 displays, the location of all housing sales in the data set. It is interesting to note that due to the data sets limitation to Single Family homes the distribution does not follow the most populous areas of San Francisco. Northeastern San Francisco does not contain a representative density proportionate to the population density in the city. This is because northeastern San Francisco is mostly made up of apartment style homes and high-rise buildings. Lastly, the crime data was also spatially visualized. As one can see, there is densest number of crimes occurs in northeastern San Francisco in the Mission district. This coincides with low density of housing sales in the data set. The school data was too sparse to yield interesting visual analysis.

Data Encoding

The data sets were encoded based on the type of models that were going to be run on them. For linear regression models, one hot encoding was used to encode the categorical features. For random forests, label encoding was used to encode the categorical features. For the ensemble technique, AutoML handled the categorical data proprietorially.

Models

This section details the model results. Three models were run: Linear Regression, Random Forests, and an AutoML ensemble model. These models were run on three different data sets: one containing only hedonistic and location data, one containing all data, and one containing all data without the original list price feature. This was to differentiate results for the hypothesis. Each data set was randomly split into training and test data. The training data was used to build the models. The test data was used to evaluate the models. The model results on the test data are displayed in the tables below. Table 1 contains the result and table 2 is a key for table 1.

Table 1: Model Results

Model Type	Data Set	Encoding	R^2	MAE	MedAE	MAPE
L	A	1H	0.464	0.296	0.202	36.250
R	A	LB	0.946	0.077	0.041	7.714
R	AN	LB	0.760	0.176	0.082	16.021
R	H	LB	0.783	0.176	0.082	16.290
A	A	AH	0.967	0.074	0.040	7.614
A	AN	AH	0.840	0.154	0.070	13.931
A	H	AH	0.854	0.150	0.070	14.060

Table 2: Key

Model Type		Data Set		Encoding	
Key	Name	Key	Name	Key	Name
L	Linear Regression	A	All Data	1H	One Hot Encoding
R	Random Forest	AN	All Data No List Price	LB	Label Encoding
A	AutoML	H	Hedonistic Data	AH	Auto Handled

The accuracy and error in precision is presented in 4 different metrics R^2, normalized mean absolute error (MAE), normalized median absolute error (MedAE), and mean absolute percent error (MAPE). The error was also visualized spatially by absolute percent error (APE) by neighborhood. This can be viewed in figure 4 for Random Forests on hedonistic data. R^2 is a good indicator of the overall model accuracy. MAE and MAPE are good indicator of the models overall precision. They are essentially the same. MedAE is a good indicator of the model precision when very insensitive to outliers. This is useful for evaluating whether the model performed well bearing exceptions from outlying data points. Usually, this signifies that more training examples will improve the MAP and MAPE.

Discussion

The discussion evaluates the research question and hypothesis based on the results. The impact and next steps of the project are also detailed below.

Hypothesis: Postulate 1

Both machine learning models - the Random Forest and the AutoML ensemble techniques vastly outperformed the linear regression model. The results validate the first postulate of the hypothesis as the R^2 value was higher and the precision error estimator were all lower. Linear regression builds a very rigid linear model for predicting housing sales. This does not allow for flexibility and non-linearity across features. The features contributing to the price of a housing sale rarely follow such linear assumptions. Given that random forests and ensemble techniques (which often encompass random forests) allow for such flexibility, it is not surprising that both models outperformed linear regression. Especially given the high dimensionality of the feature space. It is worth noting that perhaps better encoding methods for categorical data could have contributed to a better linear regression model. However, one hot encoding, which was used, is often the standard for linear

models. Using the label encoded data set for linear regression was attempted but yielded slightly worse results. Similarly, the hedonistic data set was also run on the linear model and yielded slightly worse results.

Hypothesis: Postulate 2

The second postulate of the hypothesis stated that feature additions to the data sets including crime, school, and transportation indicators will lead to better predictions, more precision and accuracy, in the machine learning models. Furthermore, the addition of a single feature, the original list price, will vastly improve the models accuracy. To test these hypothesis, three different data sets were employed. The difference in these data sets were the features present in each. The 'all' data set contained all features including hedonistic, crime, school, transportation, and original list price features. The 'all no list price' data set contained all of the prior data sets features without the original list price. Lastly, the 'Hedonistic Data' data set included only the hedonistic features. For a reference to what features are specifically included in each data set please refer to the GitHub Repository. From there, any of the header rows in the data sets under the ENG DATA folder can be inspected.

The AutoML model outperforms the Random Forests model on the same data set for all indicators for each data set. This was expected given that the AutoML model is an optimized ensemble technique. It is commonly known that ensemble techniques outperform single models because of the diversity in each model. Models trained with the all data set are more accurate and precise than models trained on the hedonistic or the all with no original list price data sets. This is expected, given that original list price is often a very good indicator of the sale price. This feature locates the models near the housing sale price. On average, the mean of the absolute difference in the list and sale price is only 35 percent. This means that a model which uses the original list price feature alone as a prediction for the sale price will have a MAPE of 35 percent. It is not surprising then to see models with the original sale price perform as well as 97 percent accurate in terms of R^2.

The interesting distinction is between the performance of models without the list price. The AutoML model outperforms the Random Forest model by a much greater factor on data without the list price than on data with the list price. This highlights the efficacy of machine learning models performance in extracting useful non-linear relationships with the target variable. This effect is presumably crowded out by the models high performance with the original list price. One can imagine, the model does not have to work as hard to do well when given such a good indicator of the sale price.

Lastly, the second postulate of the hypothesis is rejected by the out- performance of both Random Forest and the AutoML model with only the hedonistic data compared to the hedonistic plus crime and school data. This is a very noteworthy and unexpected. The performance is only marginally better for both models, so it is hard to assert that there is statistically significant difference in using the different data sets. One would assume that machine learning models follow the old adage: the more features the merrier. But perhaps this is not true. Further research is necessitated to uncover this rejection of the second postulate of the hypothesis.

Impact And Next Steps

Property fundamentally ties societies to the lands they inhibit. In a free market system, the housing market dynamics are thus vital to the livelihood of the inhabitants. Market inefficiencies can lead to detrimental economic, social, and political effects. The 2008 Housing Market Crash in the USA wounded the world economy and the American national and average household debt. Better models

are needed to identify these inefficiencies from current market conditions. Arbitrage of investors will then render the market more efficient; thus avoiding the pitfalls of a mispriced housing market.

This paper outlines the process of creating data sets and models to best predict housing sale prices. Hopefully, this paper provides a proof of concept for building high performing models. The capstone is focused on modeling the UAE market. As such, collecting high quality housing sales data in the UAE is the main focus for the next step.

The UAE is the only decreasing rent-to-price ratio bubble in the world. This means that housing prices are going down, while rental prices are going up. This phenomena is fascinating in itself. Capturing such an effect may highly influence market speculator's valuation of the market and the bubble's trajectory itself. It would also provide a model to quantify such a phenomena and predict the outcome through future sale prices. This would be highly useful to market players and to the government experiencing such unique market conditions.

Budget Justification

The stipend would be well used to purchase data from PropertyMonitor or REIDIN. However, access to these databases may be negotiated on a trial basis or through the NYUAD library without payment. The budget may alternatively be better used on attending a conference on applied or competitive machine learning.

Summary

This paper presented a high-level summary, methodology, results, and discussion of the project. The paper sought to answer How and to what extent can a Machine Learning model outperform a regression-based model in predicting housing sale prices for a specific location? Two postulates were presented in the Hypothesis: that machine learning models would outperform regression based models on all indicators of error and precision and that data sets with added information (features) would perform better on the same given machine learning model. The first postulate was accepted. The second postulate was rejected. The discussion focused on the learning outcomes of the paper. Best in practice ensemble techniques outperform random forests and linear models vastly when given a harder problem to tackle (i.e features with less mutual information than the target variable). Additions to the feature space for relevant data (crime and schooling) do not improve the performance of regression or best in-practice ensemble methods. Lastly, the paper concluded with a note on the impact and next steps of the project. Specifically, the next, crucial step is to obtain and perform the same and more in-depth model generation and analysis on the UAE market.

Appendix

GitHub Repository

The GitHub repository for the data, models, and current library can be found at `https://github.com/jaisal1024/capstone`

Market Review

The market review introduces the reader to the current trends, idiosyncratic phenomena, and historical motivators that drive the market at hand. Only with a strong qualitative understanding of how the market moves, one can propose a well-suited Machine Learning model. Machine learning models are non-linear regressions which fundamentally assume future (unlabeled target) data from a set of characteristics (features) based on a unique set of trained (labeled target) data. Thus, when choosing the input factors and the input factor functions one must understand the data to best build a predictive model. The market review will focus on the UAE, specifically Dubai and Abu Dhabi, as the capstone first proposes applying the model to this market.

UAE Overview

The UAE, specifically Abu Dhabi and Dubai, both have a historically and presently unique property market. The UAE was founded in 1971 as the union of seven emirates: two of which are Abu Dhabi and Dubai [17]. In Abu Dhabi, Sheik Zayed began by giving each Abu Dhabian three properties: one for residential, one for commercial, and one for farming use or any other pursued venture [17]. Many Abu Dhabians would then rent or sell their plots of lands to foreigners. Similarly, in Dubai, each citizen was granted a residential piece of land. The UAE's population is dominated by immigrants with only around 20 percent of residents being Emirati nationals. Only this minority (along with GCC citizens) are allowed to permanently buy property. However, both Abu Dhabi and Dubai have designated areas where non-nationals can buy leaseholds or freeholds. This is further discussed in the property laws section below. Even nowadays, Emiratis are granted a plot of residential property from the government. The UAE developer's market is rich with large, successful companies. EMAAR properties from Dubai is the largest with developments around the world and accomplishments like the Dubai Mall and the Burj Khalifa. DAMAC Properties and Aldar Properties are the 2nd and 3rd most valued real estate companies in the UAE [18]. Housing Data in the UAE is government-regulated and distributed. This could potentially be a conflict of interest given the privatization of many large UAE developers. This should be taken into consideration when discussing historical trends and building models.

Historical Markets

Dubai experienced massive financial and property market growth between 1999 and 2008 with investors from around the world looking to gain a foothold into the emerging GCC. Oil prices continued skyrocketing and often burgeoning developments would be "sold off" and traded multiple times before completion [19]. In 2002, the ruler of Dubai announced that all foreigners could buy freehold property in certain sections of Dubai. This skyrocketed foreign investment in the wake of the oil boom and development of the gulf [20]. Real estate growth accounted for 25 percent of Dubai's total GDP growth [21] in those years. However, in Q4 of 2008, following the financial collapse, the Dubai market fell by over 25 percent resulting in an almost catastrophic default of

debt as liquidity fell, property market supply skyrocketed, and demand shrunk [21]. The result was a massive decrease in the housing market prices. The Dubai world tribunal hearings were held to restructure debt owed by Dubai World holdings, the private investment arm of the government [19]. The Abu Dhabi central bank then issued a 20 billion dollar loan to the Dubai government. The Burj Dubai was renamed the Burj Khalifa in commemoration of the Ruler of Abu Dhabi. The Dubai market stalled for 2 years between 2008 and 2010. Dubai tried to install stricter financial regulations on banking loans and real estate asset holdings to prevent a repeat of the 2008 crash. However, push back from banks and speculative investors led them to revoke their decision shortly after effect [19]. Economically, stricter loans starve housing demand as most investors purchase a home with a mortgage. More recently, these restrictions seem to have been further eased; much like the US. In 2011, the market began to recover. By 2012, "the volume of transactions increased significantly," [21]. Both sea views areas like Dubai Marina and the Palm Jumeirah and downtown areas like Downtown Dubai and Business Bay saw increases in volume and price over this period. This growth continued through 2014, surpassing the previous market high in 2008. However, since then Dubai (and Abu Dhabi) have experienced a deflationary bubble where rental and housing prices have been continually dropping. Specifically, in Dubai, since 2014, there has been a decline in price growth of residential property rental prices and since 2016 there has been a decline in residential property sales [22, 23].

Property Laws

Property laws are determined independently by each Emirate within the UAE [20]. In 2002, Dubai first made it possible for non-nationals to indirectly buy freehold property through 3 government-backed privatized firms: EMAAR, Nakheel, and Dubai Properties. In 2006, Dubai went one step further and allowed non-nationals to directly buy freehold or 99-year lease properties in certain areas of Dubai [20]. GCC nationals have always been allowed to own property in both Abu Dhabi and Dubai. In 2002, following Dubai's indirect freehold system, Abu Dhabi passed new property laws which allowed foreigner's to buy 50-year or 99-year leasehold property in designated Investment Zones. More recently, the laws have been expanded to include indefinite direct freeholds in more designated Investment Zones. These include the likes of Yas, Saadiyat, and Al Reem Island.

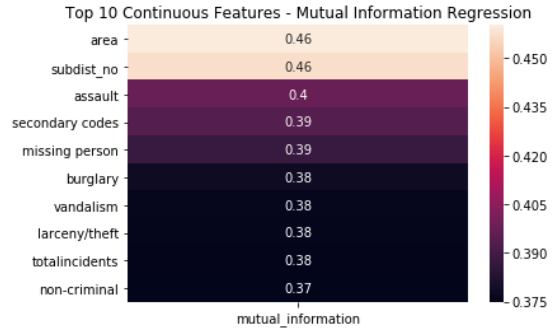
Current Reports And Projections

The UAE, specifically Abu Dhabi and Dubai, are currently experiencing a decreasing bubble where rent prices are falling at a higher relative rate to housing prices. This is one of the only decreasing bubbles in the world, making it a fascinating and unprecedented topic of research. Figure 1 in the appendix reveals the year-to-year prime market performance for Abu Dhabi and Dubai between 2012 and 2017. The Knight Frank report for UAE property trends for 2019, reports a continued softening of residential and commercial rent-based and for-sale real estate in both Abu Dhabi and Dubai. However, the introduction of a 10-year Visa and increased IMF projected GDP growth hope to dampen the softening market [22]. Knight Frank writes, "Dubai Land Department reported that in Q1 2018 UAE nationals remained the largest buyer group in Dubai with AED 4 Billion in transactions," [24]. Indian investors remained the largest foreign investors with around 3 Billion AED in transactions. In Abu Dhabi, Al Reem, Al Maryah, and Saadiyat Island experienced the greatest price decline. Yas Island experienced the least decline of all neighborhoods. Interestingly, apartment-style housing decreased by less than villa housing. This may be a correction from overpriced primary markets as most large developments are luxury villa-based. Despite these statistics, construction on Saadiyat Island continues at a dizzying pace with two new villa complexes

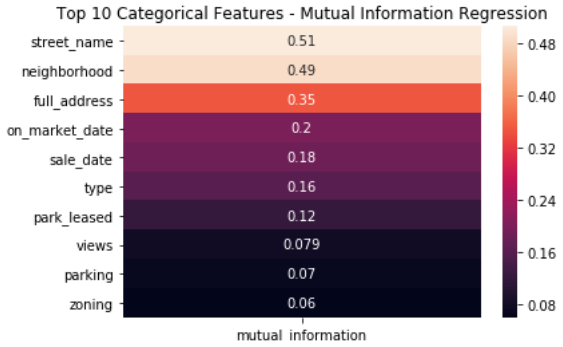
coming to market in the 2019 and 2020. In Dubai, only 1 district, Business Bay, out of 66 districts reported a positive price change in 2018 [22]. Total residential transactions also fell by 14.9 percent [22]. Commercial and Retail rentals have also seen sharp declines in 2018. The expansion of e-commerce in the UAE, has resulted in double digit retail real estate declines [22]. The hospitality industry has faced increased competition from AirBnb and thus a decrease in rental prices and revenue.

Interestingly, Class of 2019 NYU Abu Dhabi Student Nadine Laze presents a capstone paper in which a 1 percent increased AirBnb presence in Dubai has led to a 0.087 percent increase in long term rental prices [23]. This suggests that despite AirBnb's negative impact on the UAE hospitality industry, it may have a positive impact on the entirety of the real estate economy. Lastly, oversupply of real estate continues to be an issue in the UAE. Historically, this has been due to the rapid nature of construction development at the artificial will of the government and not the demand of the people. Knight Frank expects the level of supply in both Abu Dhabi and Dubai will continue to grow. Shown in Figure 2 and 3, the Abu Dhabi rent and Dubai sale market has seen a recent drop in transaction volume, while the supply of housing continues to rise in preparation for Expo 2020 [24]. This reflects the deflationary bubble. By economic principles this will inevitably result in a continued decrease in price.

Figures And Plots



(a) Numerical



(b) Categorical

Figure 1: Top 10 Mutual Information Features on Sale Price

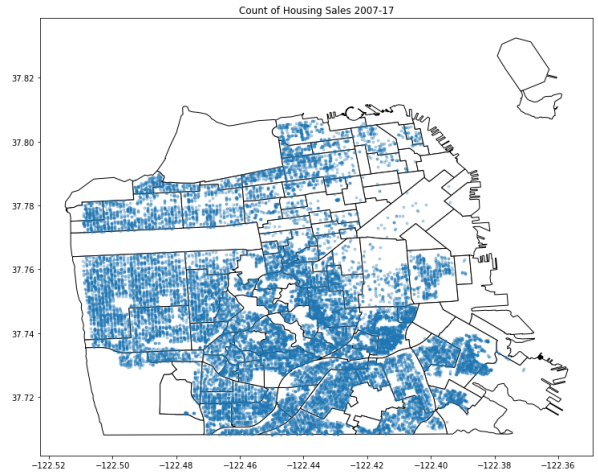
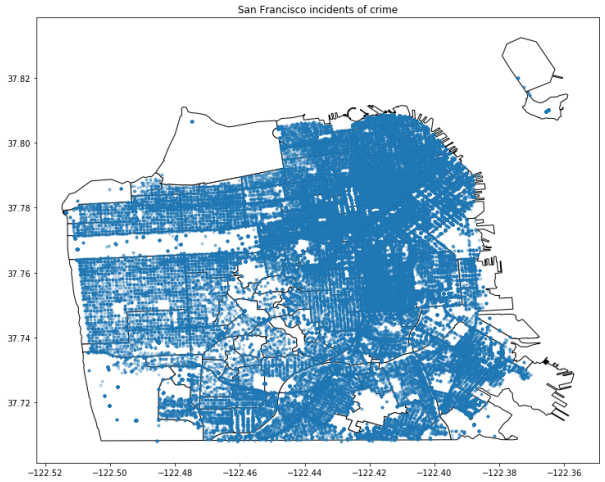
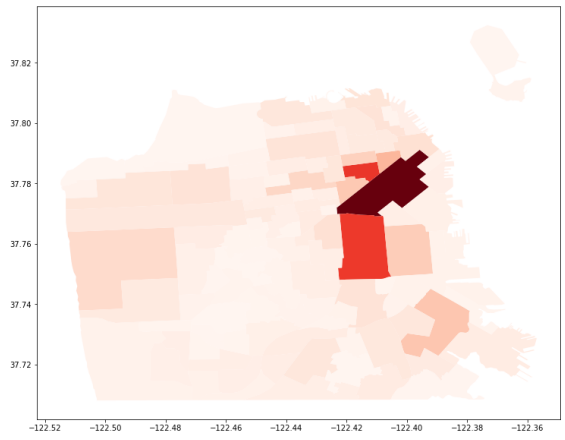


Figure 2: Map of Housing Sales



(a) All Incidents



(b) Density by Neighborhood

Figure 3: Maps of Crime Data

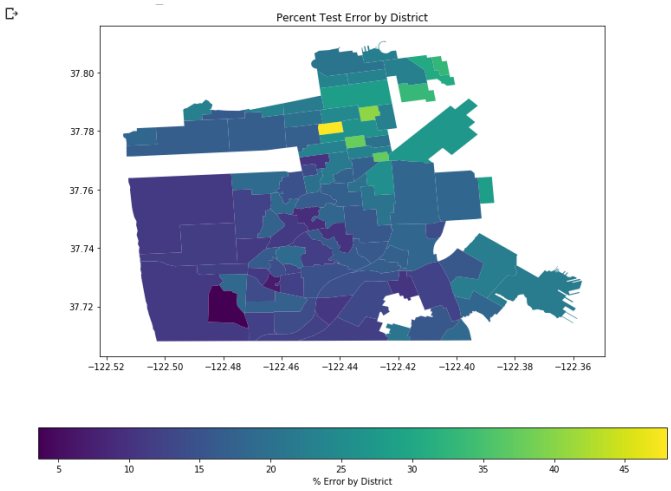
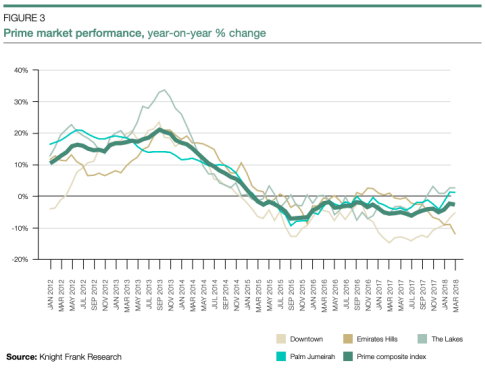
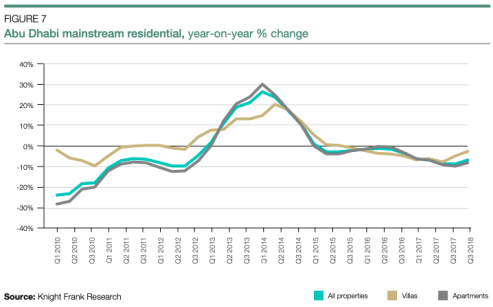


Figure 4: Map of Housing Sales



(a) Dubai



(b) Abu Dhabi

Figure 5: Residential Mainstream Real Estate Prices 2012-18

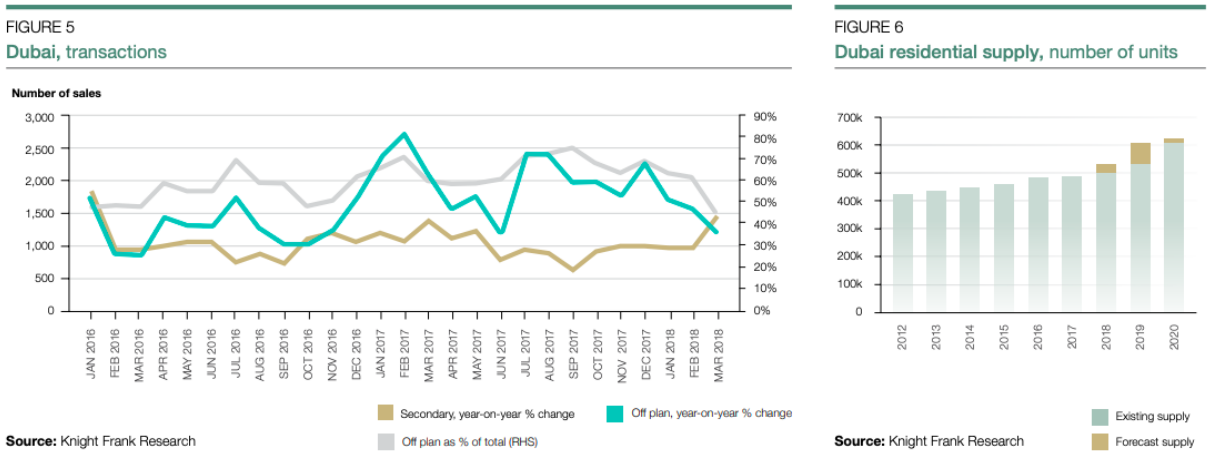


Figure 6: Dubai Transaction Volume

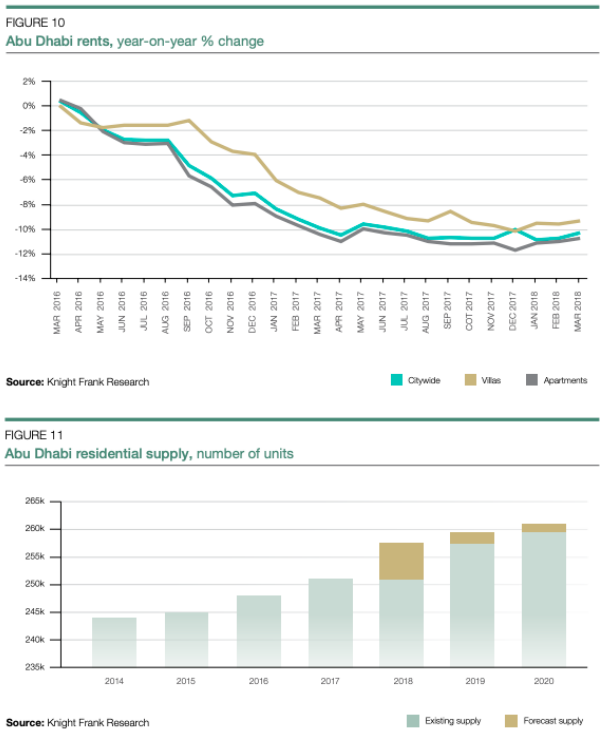
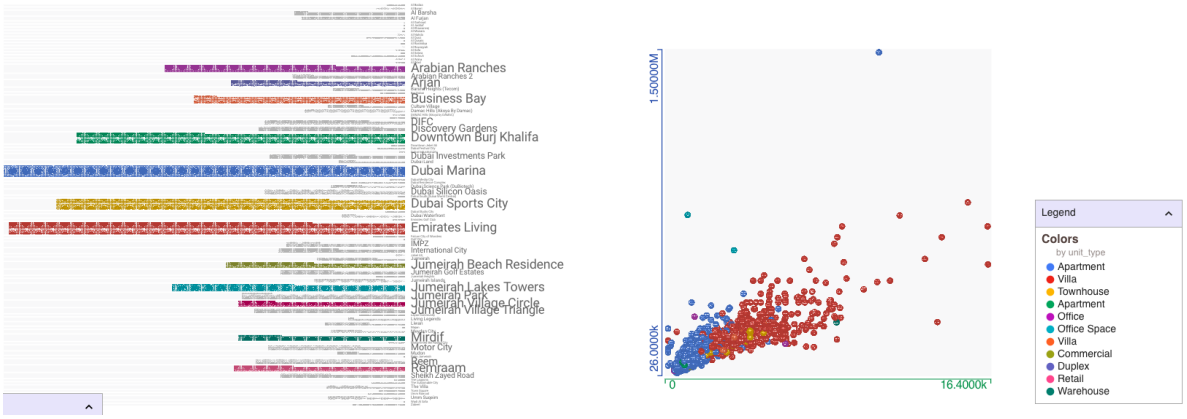


Figure 7: Abu Dhabi Rental Transaction Volume



(a) Distribution of Neighborhoods

(b) Unit Size to Total Rent

Figure 8: Property Monitor Database

References Cited

[1] S. Humphries, "Introducing a new and improved zestimate algorithm," Sep 2019. [Online]. Available: https://www.zillow.com/tech/introducing-a-new-and-improved-zestimate-algorithm/

[2] E. L. Glaeser and C. G. Nathanson, "Housing bubbles," in *Handbook of regional and urban economics*. Elsevier, 2015, vol. 5, pp. 701–751.

[3] O. Bonnet, P.-H. Bono, G. Chapelle, E. Wasmer *et al.*, "Does housing capital contribute to inequality? a comment on thomas piketty's capital in the 21st century," *Sciences Po Economics Discussion Paper*, vol. 7, 2014.

[4] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy*, vol. 82, no. 1, pp. 34–55, 1974.

[5] J. P. Harding, S. S. Rosenthal, and C. Sirmans, "Depreciation of housing capital, maintenance, and house price inflation: Estimates from a repeat sales model," *Journal of urban Economics*, vol. 61, no. 2, pp. 193–217, 2007.

[6] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.

[7] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015.

[8] M. Harries and K. Horn, "Detecting concept drift in financial time series prediction using symbolic machine learning," in *AI-CONFERENCE-*. Citeseer, 1995, pp. 91–98.

[9] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[10] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*. IEEE, 1997, pp. 511–520.

[11] "Automl h20," Dec 2016. [Online]. Available: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

[12] "San francisco single family homes sales data 2007-2017."

[13] "Great schools api technical overview." [Online]. Available: https://www.greatschools.org/api/docs/technical-overview/

[14] "About greatschools' ratings system and methodology." [Online]. Available: https://www.greatschools.org/gk/ratings/

[15] O. SF, "Geojson of realtor neighborhoods in san francisco."

[16] ——, "San francisco crime data 2003-2018."

[17] M. Al Fahim, "From rags to riches," *Dubai, UAE: London Centre of Arab Studies*, 1995.

[18] Cw, "Revealed: Gcc's top 10 developers," Oct 2014. [Online]. Available: https://www.arabianbusiness.com/revealed-gcc-s-top-10-developers-567120.html

[19] S. Bagaeen, "Saudi arabia, bahrain, united arab emirates and qatar," *International Approaches to Real Estate Development*, p. 101, 2014.

[20] ——, "Brand dubai: The instant city; or the instantly recognizable city," *International Planning Studies*, vol. 12, no. 2, pp. 173–197, 2007.

[21] T. A. Falade-Obalade and S. Dubey, "Analysis of the real estate market in dubai-a macro economic perspective."

[22] *UAE Market Review and Forecast*, Jan 2019. [Online]. Available: https://content.knightfrank.com/research/1064/documents/en/uae-market-review-forecast-2019-6072.pdf

[23] N. Laze, "The future of dubai's real estate: The effect of airbnb on long-term rental market," Ph.D. dissertation, New York University Abu Dhabi, 2019.

[24] *UAE Residential Market Review Q1 2018*, Jan 2018. [Online]. Available: https://content.knightfrank.com/research/1413/documents/en/uae-residential-market-review-q1-2018-5614.pdf