Machine Learning Engineer Nanodegree

Capstone Project

Paul F. Sabadin

June 20, 2018

I. Definition

(approx. 1-2 pages)

Project Overview

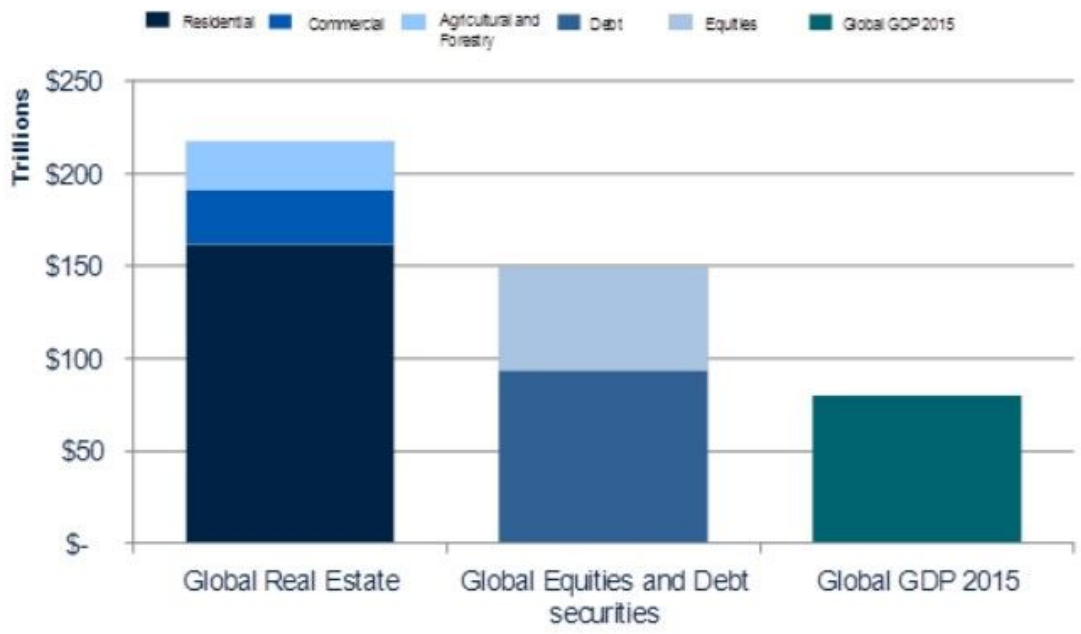
In recent decades, international trade, combined with cost disparities between developed and developing nations for global labor and manufacturing have resulted in the easy supply of affordable quality goods to the developed world. Consumers have benefited with an improved quality of life. But, more importantly, much commerce and many fortunes have been made through the leveraging of these cost disparities in global markets.

Many things: manufactured widgets, intellectual property, medicines, and education are transportable and commercial activities associated with these have been profoundly affected by globalization. But some things cannot be moved. The most obvious example of an immovable asset is real estate!

Between 2000 and 2015, the volume of cross-border real estate transactions grew from $65B to $217B (334%) (https://www.valuewalk.com/2016/01/real-estate-is-a-dominant-asset-class-savills/).

Much larger than the global body of equities, debt securities, and gold, developed real estate is the largest asset class in the world, estimated at over $217 trillion in 2016 (same reference).

Global real estate universe in comparison



Source: Savills Research, Bank for International Settlements, Dow Jones Total Stock Market Index, Oxford Economics

Of global developed real estate, global residential real estate (the subject of this project) yet comprises the dominant single asset class in the world. Further, appreciation of these assets (capital gains, or losses), being proportional to the large basis, are also huge and important.

The power of real estate investment to transform net worth is evident to most Americans and, indeed, to most of the developed and developing world. However, real estate is one of the most difficult assets for which to determine value. Unlike identically reproduced and exportable factory widgets of precise and repeatable dimensions, each piece of real estate is unique with its value being composed of characteristics such as building floor area, building material quality, numbers of rooms, etc. as well as equally important intangible features such as architectural appeal or others based on locale such as views, political environment, climate, and proximity to water. The number of potentially relevant attributes (features) affecting the valuation of real estate can be almost arbitrarily large and is difficult to ascertain.

In America, the importance of accurate and precise residential real estate valuation has been long appreciated by both new and old economy real estate service companies such as banks, property appraisers, the Multiple Listing Service (MLS), RedFin, and

Zillow (Zillow.com). The industry has evolved, and has now landed at the most visible leading edge of data science, evidenced by the recent Zillow real estate valuation machine learning competition, the largest prize ever hosted at the famous Kaggle machine learning organization (Kaggle.com).

As early as 1994, this student, like Zillow, has been geocoding, mapping, and performing personal real estate evaluations with digital county assessor property lists (that are often obstructionally inaccessible and closed government databases) to better understand the market. The likes of Zillow have leveraged government and private real estate databases to entrench themselves in this US residential real estate domain. But it is now time, and the goal of this project, to help initiate an evolution of automated residential real estate valuation beyond US markets to a global market by leveraging machine learning technologies with global Internet-hosted and crowd-sourced data to investigate the drivers and attempt to estimate aggregate or typical prices in various international cities.

In this section, look to provide a high-level overview of the project in layman's terms. Questions to ask yourself when writing this section:

- ***Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?***
- ***Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?***

Problem Statement

Policy makers, investors, and others need fine-grained information and an understanding of driving relationships between global residential prices for goods and services, chief among them, city-level real estate prices. However, few publicly available and free resources for pricing data at this scale can be found, either in print or online. There are, both formal and informal, market price data strewn all over the Internet. Some of that data is directly available and intended to be accessible. For example, the

Organization for Economic Co-Operation and Development (oecd.org) publishes analytical house price indicators online. However, the OECD Data is sparse and coarse-grained (it is at the country, not city, level). The data needed is believed to be extant online, but it is, for the most part, in the form of informal chatter, discussions, listings, and indirect commercial activity. We wish to leverage this vast Internet data store. We would particularly like to better understand relationships between residential real estate and other prices and their locations. But how?

The use of Internet data, both legally and, arguably illegally, is clear and present in today's society. Web crawlers and indexers of all sorts (such as Google) daily mine this data to great advantage. However, the collecting and use of this data is both technically arduous and legally risky. The mining of Internet data, even before it might be exploited with machine learning algorithms, can become its own industry. Our task is *post-collection* exploitation with our own machine learning algorithms to, indeed, extract useful insight and pricing information. Our task is to gain access to publicly and globally available economic information and to apply machine learning algorithms to it to arrive at useful global real estate pricing trends estimates.

As the Internet has evolved, it has become more social and full of contributed information from educated and uneducated, laypersons and experts alike. Can the "wisdom of the crowd" be successfully exploited? If we have this information, can we separate wisdom from the rhetoric of uneducated opinion?

So, the problem in culling actionable real estate information from the Internet is two-fold:

1. Finding and accessing data associated with global real estate prices, and
2. Crafting and applying machine learning algorithms to cull actionable real estate pricing trends and intelligence from the data

The first of these problems, finding data, is to be solved with an online, crowd-sourced, pricing information web site called Numbeo.com That site collects data related to a location (city, metro area, etc.), informally, from any person in the global crowd (wise or unwise) who is interested in contributing. The data consists of broad spectrum topics such as local market food prices, pollution, crime, quality of life. Importantly, among the data collected is real estate prices for the locale of the contributor. This is the chief data source to be leveraged by this project.

The numbeo data will be further conflated with the Max Mind World Cities Database hosted on Kaggle.com. Because of the author's hunch that broader political and economic freedoms and vitality certainly play a role in the attractiveness of owning (thus

the price) of real estate in a country, the above two datasets were further conflated with the Heritage Foundation's 2018 Index of Economic Freedom. Thus the datasets used can be summarized as:

- Numbeo.com crowd-sourced economic and real estate data
- Max Mind's Kaggle.com hosted World Cities Database
- Heritage Foundation's 2018 Index of Economic Freedom

Fortunately, the second problem of gleaning information from a vast data store, once it is made available, is a tractable one. Modern machine learning techniques are up to this task. Thus, we can operate with machine learning algorithms on Internet crowd-sourced and, potentially very noisy, crowd-sourced pricing data to analyze and better understand the relationships between sundry global information at the community or municipal level and local market real estate prices.

In this project we will apply various machine learning algorithms to a large set of combined Internet extant, crowd-sourced, and organizational data so as to analyze and gain insight into global real estate markets. We will use this to estimate global residential real estate market prices, comparing our estimates to those prices reported by Internet users. We will conclude by further quantifying and discussing the measures of success for our project and how the information found might be used to alert real estate investors to potential upside deals or downside risk.

Because the relationship between input variables and real estate prices is expected to be highly nonlinear and stochastic, we decided to apply the nonlinear Random Forest regression algorithm to the problem as this algorithm is known to perform well on such problems. Our general approach to the problem started with cleaning and combining the above datasets. The resulting dataset economic variable distributions, including target real estate prices, were examined and normalized. The dimensionality of the variables was then decreased by applying principal component analysis. This smaller set of variables was divided into training and testing subsets and a Random Forest regressor algorithm was trained to estimate global urban apartment prices based on unseen test data. Results were analyzed and discussed.

Metrics

The quantitative goal of this project is to estimate global aggregate residential real estate market prices for various cities and locales as derived from other information

associated with the location. Real estate prices are a continuous variable and estimation of the performance error of such results is highly amenable to use of the R2 statistic. Thus, R2 will be used to measure the ability of the estimation algorithms to fit to target “ground truth”: the *reported* market estimates for the residential real estate properties.

Along the way to a solution, data preprocessing and data set analyses (such as data normalization, principal component analysis and clustering) will be applied. The appropriate metrics for those algorithms will be, respectively, measures of data set statistical skew, explained variance, and cluster silhouette coefficients. Summarizing, the metrics used will be:

- R2 coefficient of determination for real estate price estimate errors
- Cluster silhouette coefficients to measure clustering effectiveness
- Data set statistical skew
- Explained variance for the components of principal component analysis

From initial observations of reported apartment prices in our dataset, it was obvious that this target variable had large variance and we would need a metric that would measure our model’s capability of capturing this variance. Indeed, we could have used the the mean squared error (MSE) metric as our first choice tool in measuring the effectiveness of our Random Forest estimator. However, we chose R2 as this measure (more formally, the coefficient of determination) better measures how well our estimator might explain the variance in our data (while not penalizing for estimator complexity). The MSE metric is not normalized by the the variance in the target data and thus less able to quantify such variations.

R2 is defined as:

$$R^2 = 1 - u/v \;=\; 1 - MSE/variance$$

where u is the residual sum of squares of the difference between the estimated and true data value, and v is the total sum of squares of the difference between mean of the true values and the mean of the true values. R2 and yields a maximum value of 1.0 when the estimator fits the data perfectly, without error, yields 0.0 when the estimate is a constant, and is negative where the estimation model fits the data even more poorly than a constant value.

Cluster analysis is considered to be ancillary to the main goal of estimating real estate prices. The silhouette coefficient, which essentially measures the cohesion of data point to its assigned cluster versus its distance from other clusters, was considered as

sufficient to measure clustering effectiveness for this first look at clustering in this study. In follow-on studies, information theoretic techniques for measuring clustering algorithm effectiveness, such as the AIC or BIC criteria, may be further investigated for their application to this problem.

II. Analysis

(approx. 2-4 pages)

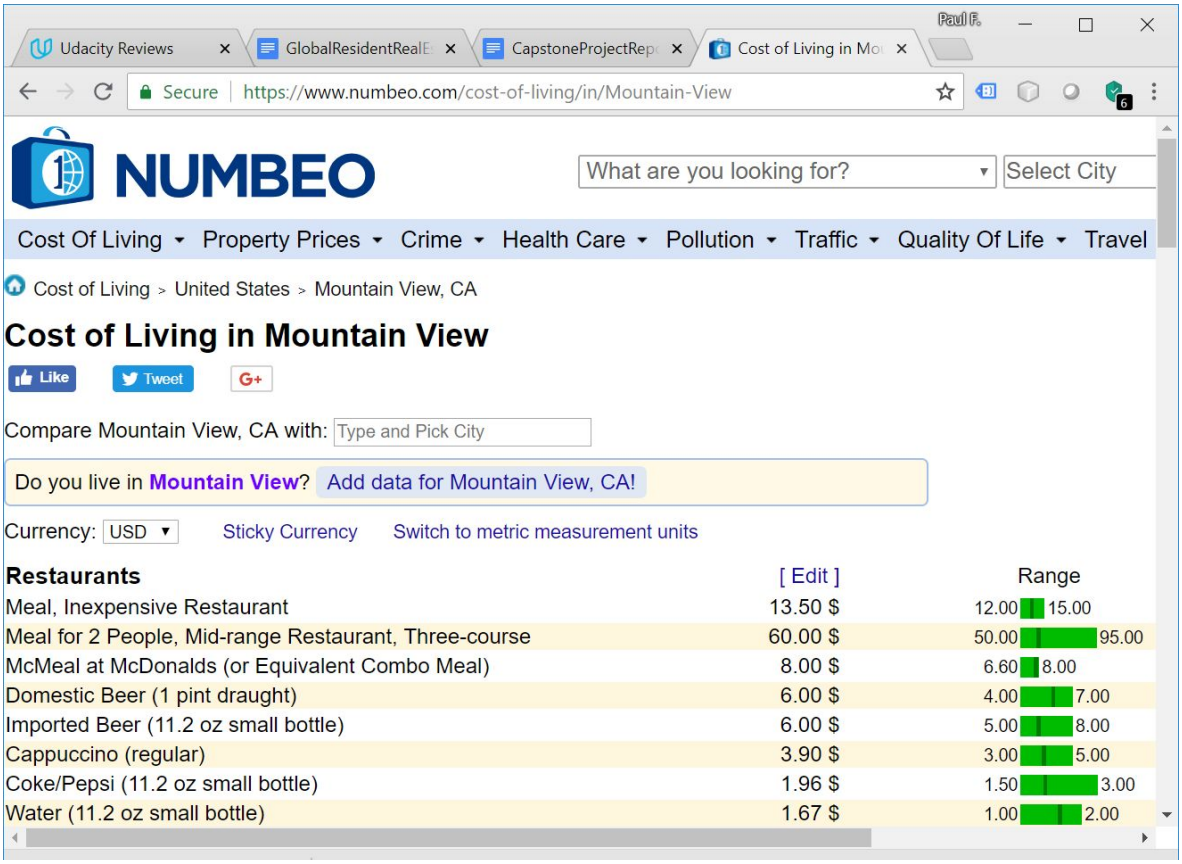
Data Exploration

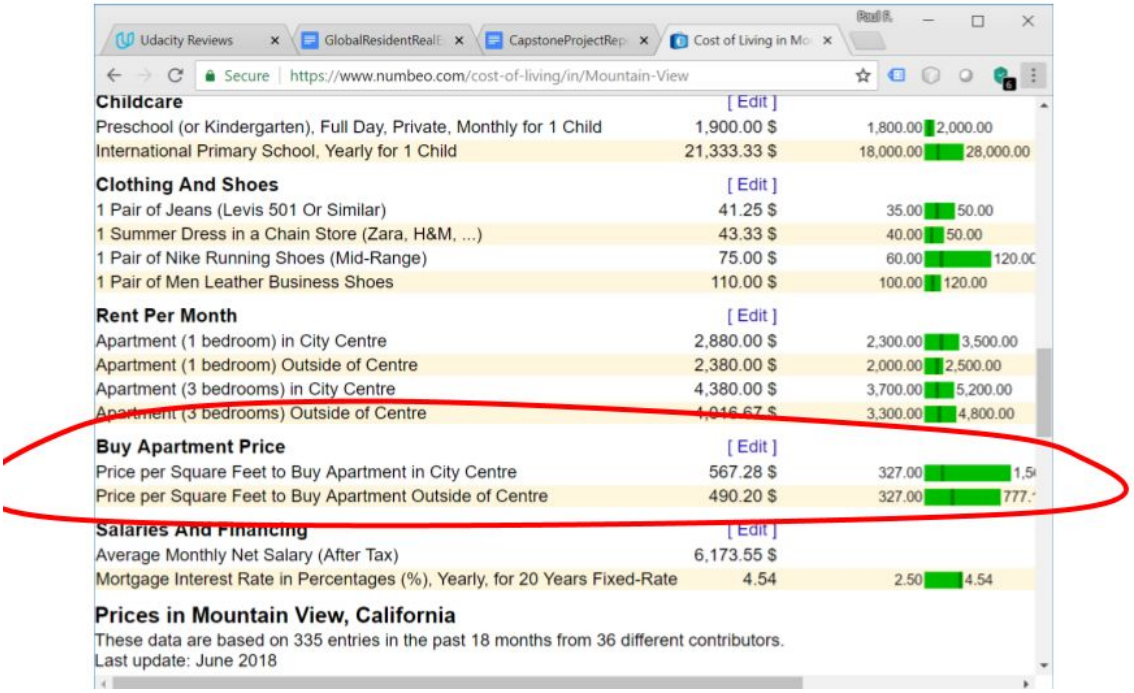
Many datasets were investigated for their appropriateness to the goals of this project. As stated above, a large data set was conflated from

- Numbeo.com crowd-sourced economic and real estate data
- Max Mind's Kaggle.com hosted World Cities Database
- Heritage Foundation's 2018 Index of Economic Freedom

An API key for Numbeo.com data was provided to the author thanks to the generosity of the company's president.

The kind of data provided by Numbeo.com is best illustrated by the screenshots in the two figures below (a single page divided into two images).


Udacity Reviews
GlobalResidentRealE
CapstoneProjectRep
Cost of Living in Mo
Paul F.
Secure
https://www.numbeo.com/cost-of-living/in/Mountain-View
NUMBEO
What are you looking for?
Select City
Cost Of Living
Property Prices
Crime
Health Care
Pollution
Traffic
Quality Of Life
Travel
Cost of Living > United States > Mountain View, CA
Cost of Living in Mountain View
Like
Tweet
G+
Compare Mountain View, CA with:
Type and Pick City
Do you live in Mountain View? Add data for Mountain View, CA!
Currency: USD
Sticky Currency
Switch to metric measurement units
Restaurants [Edit] Range
Meal, Inexpensive Restaurant 13.50 $ 12.00 15.00
Meal for 2 People, Mid-range Restaurant, Three-course 60.00 $ 50.00 95.00
McMeal at McDonalds (or Equivalent Combo Meal) 8.00 $ 6.60 8.00
Domestic Beer (1 pint draught) 6.00 $ 4.00 7.00
Imported Beer (11.2 oz small bottle) 6.00 $ 5.00 8.00
Cappuccino (regular) 3.90 $ 3.00 5.00
Coke/Pepsi (11.2 oz small bottle) 1.96 $ 1.50 3.00
Water (11.2 oz small bottle) 1.67 $ 1.00 2.00


Udacity Reviews
GlobalResidentRealE
CapstoneProjectRep
Cost of Living in Mo
Paul F.
Secure
https://www.numbeo.com/cost-of-living/in/Mountain-View
Childcare [Edit]
Preschool (or Kindergarten), Full Day, Private, Monthly for 1 Child 1,900.00 $ 1,800.00 2,000.00
International Primary School, Yearly for 1 Child 21,333.33 $ 18,000.00 28,000.00
Clothing And Shoes [Edit]
1 Pair of Jeans (Levis 501 Or Similar) 41.25 $ 35.00 50.00
1 Summer Dress in a Chain Store (Zara, H&M, ...) 43.33 $ 40.00 50.00
1 Pair of Nike Running Shoes (Mid-Range) 75.00 $ 60.00 120.00
1 Pair of Men Leather Business Shoes 110.00 $ 100.00 120.00
Rent Per Month [Edit]
Apartment (1 bedroom) in City Centre 2,880.00 $ 2,300.00 3,500.00
Apartment (1 bedroom) Outside of Centre 2,380.00 $ 2,000.00 2,500.00
Apartment (3 bedrooms) in City Centre 4,380.00 $ 3,700.00 5,200.00
Apartment (3 bedrooms) Outside of Centre 3,300.00 4,800.00
Buy Apartment Price [Edit]
Price per Square Feet to Buy Apartment in City Centre 567.28 $ 327.00
Price per Square Feet to Buy Apartment Outside of Centre 490.20 $ 327.00
Salaries And Financing [Edit]
Average Monthly Net Salary (After Tax) 6,173.55 $
Mortgage Interest Rate in Percentages (%), Yearly, for 20 Years Fixed-Rate 4.54 2.50 4.54
Prices in Mountain View, California
These data are based on 335 entries in the past 18 months from 36 different contributors.
Last update: June 2018

As can be seen, the Numbeo data used consists of approximately 55 numerical economic data points, such as the price of meals or clothing, that are associated with a single city or locale (Mountain View, California, in the above image). Included in this data are user-reported prices for the purchase of local apartments. As can be seen in the second figure, a price for apartments in both an city center (urban) and an outside of city center (suburban) price are given. City center apartment prices from this database are, in fact, what are later used as the target variable “truth” in machine learning algorithm training and in the measurement of estimation error.

As this data source is crowd-sourced from online participants, it is expected to be quite noisy as the data might typically come from simple user recall about prices but also may include less frequent contributions from local experts.

Analysis started with the downloading of this kind of data for 10,000 cities across the world (using the World Cities database as an indexer). The World Cities database, containing city population, latitude, and longitude was downloaded as well, followed by a download of the 2018 Index of Economic Freedom. Only cities with population of more than 25,000 were considered. Some cities in the World Cities database contained outdata names (such as Bombay for Mumbai, India), but this is not expected to have great effect on the numerical data, itself.

The 2018 Index of Economic Freedom data used contained country score data on measures such as Property Rights, Judicial Effectiveness, Government Integrity, Tax Burden, and Gov't Spending for countries of the world. All in all, the combined number of features, or explanatory variables, available for each city was 85, 26 of which were from the Economic Freedom database.

As the number of variables was large and their names/definitions quite long, acronym labels were fashioned for the data to allow for abbreviated name management. For illustrative purposes, a few derived names for feature value variables are listed below. The entire key for variable definitions is appended to the end of this report:

Some Feature Name Definitions (see appendix for entire list)

...

MIRR:	Meal, Inexpensive Restaurant, Restaurants
MF2PMRTR:	Meal for 2 People, Mid-range Restaurant, Three-course, Restaurants
MAM(oECM)R:	McMeal at McDonalds (or Equivalent Combo Meal), Restaurants

...

MP(RP)T:	Monthly Pass (Regular Price), Transportation

...

PR: Property Rights
JE: Judicial Effectiveness
…

All currencies were normalized to US dollars and all physical units were re-expressed in metric.

After cleaning and joining the the three constituent datasets, complete feature sets for about 4593 cities (data points) remained. Eighty five different features were used for the analysis, all of which were numerical. The data from these remaining cities was used for training and testing our machine learning algorithms. The features were divers and had various numerical ranges. For example, city apartment prices ranged from $172 per square meter to $66728 per square meter with standard deviations of $3077, while city populations ranged from 24,278 to 31,408,500 (Tokyo, Japan). The standard deviation of city populations was 86,9915. The northernmost city was in northern Finland and the southernmost in southern New Zealand.

As an example, the first several columns of a few rows from the combined data set is shown below.

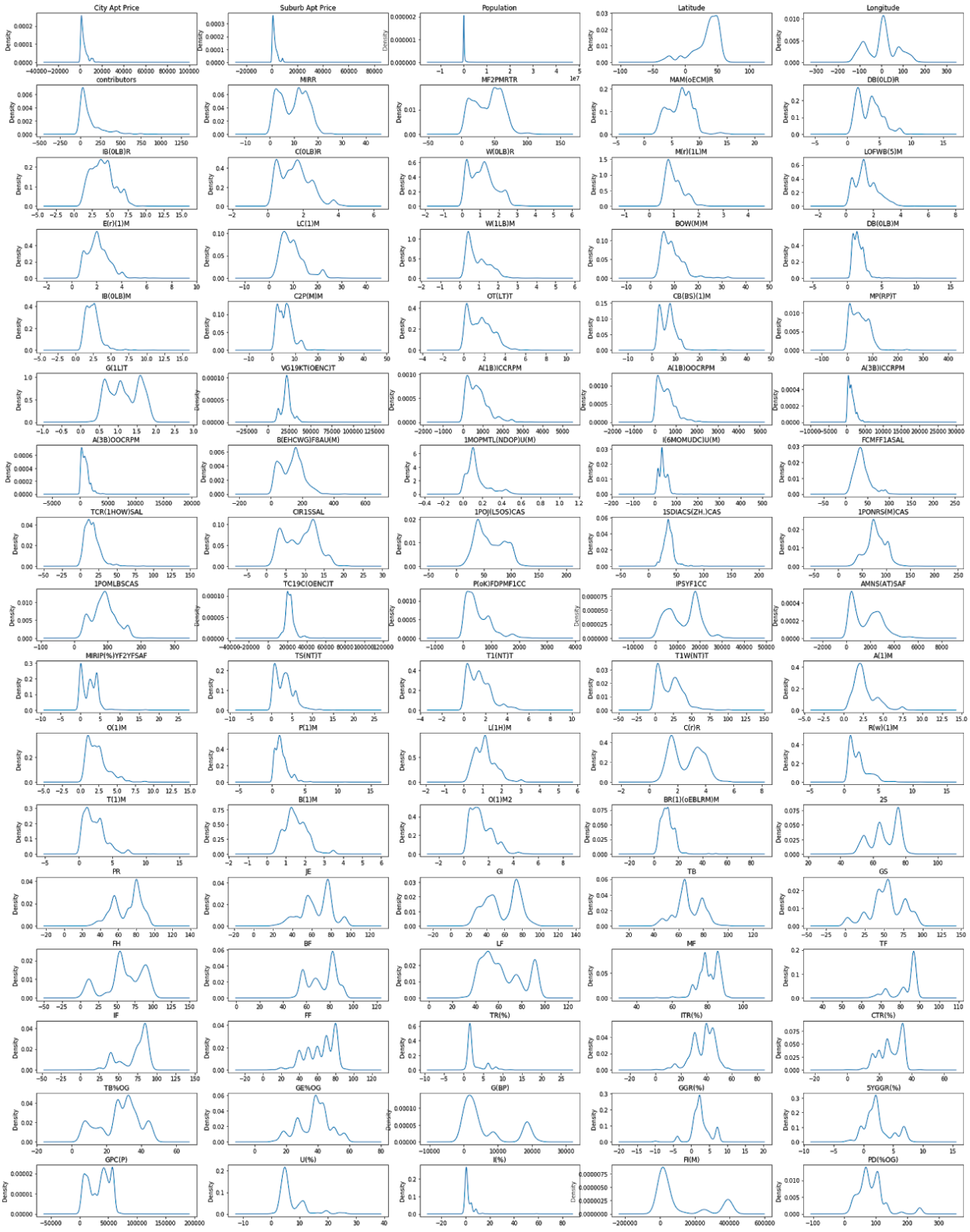
	City	CountryName	Region	City Apt Price	Suburb Apt Price	Population	Latitude	Longitude	contributors	MIRR	...	TB%OG	GE%OG	G(
30	los angeles	United States	Americas	6900.919988	4334.025358	3877129.0	34.052222	-118.242778	375	15.000000	...	26.358	38.058865	18569.
31	alexandria	Egypt	Middle East / North Africa	668.534080	281.979458	3811512.0	31.198056	29.919167	118	2.801120	...	18.200	34.100000	1132.
32	tianjin	China	Asia-Pacific	4887.476534	3124.511795	3766207.0	39.142222	117.176667	26	3.124512	...	17.500	30.746667	21291.
33	melbourne	Australia	Asia-Pacific	6561.187690	5237.797294	3730212.0	-37.813938	144.963425	418	12.164803	...	27.847	35.953447	1187.
34	ahmadabad	India	Asia-Pacific	960.840016	537.630561	3719933.0	23.033333	72.616667	186	2.221696	...	7.200	27.250000	8662.
35	casablanca	Morocco	Middle East / North Africa	1926.946795	1007.669425	3609698.0	33.592779	-7.619157	138	3.182114	...	24.600	31.353667	281.

6 rows × 88 columns

A cursory look at some of the entries shows the great variation (and noise) in values between different locations. For example, *is* the price of an inexpensive meal in Los Angeles, California actually more than five times the cost of the same meal in Alexandria, Egypt? This may reflect reality or may, in fact, reflect noise in the data inputs due to individual contributions in the crowd-sourced data. It may reflect both, and it is a challenge for our machine learning algorithms to deal with this uncertainty and still yield valuable results and insight.

Exploratory Visualization

After loading the data and performing basic cleaning the dataset was explored for to see if it exhibited any interesting or unusual characteristics. As there are many variables, a first look involved plotting the Kernel Density Estimator (KDE) for each variable, independently. The plot of such data is shown below.



A list of summary first impressions about the KDE plots of this data follow. The data is:

- Large: 85 explanatory variables
- Frequency distributions are complex and typically multi-modal
- Some distributions are broad while others are very peaked (suggesting outliers)
- Many of the distributions exhibit significant skew, typically left-skew

Some observations about individual variables can also be made.

For example, city apartment purchase prices (City Apt Price) have high frequency representation at the left of the plot. That is there are many cities with apartment prices in the range between x and y. However, some real, high-end apartment prices exist in the database such as that for the city of Sanremo, Italy (more probably intended as an entry for nearby Monaco), along the famous Italian Riviera, where a 100 square meter apartment price is reported to be nearly $6.7M US. Other expensive real estate may have been dropped from the database due to incomplete or problematic record data.

Likewise, the population data (Population) in the database exhibits a high number of medium/small cities with populations well under 100,000 people. Again, the database was intentionally limited in this study to cities greater than 25,000 people. The largest city reported (an outlier) is Tokyo Japan, with a population of over 31 million.

The frequencies of cities in the database at latitudes exhibit an expected hgh representation of cities in the northern hemisphere (positive latitudes). City reports vs longitude reflect a multimodal shape with groupings characteristic of latitudes representing the Americas, Europe, and Asia. Few reports exist for Atlantic and Pacific Ocean longitudes (save some reports of island cities).

The other variable distributions in, roughly, the first two-thirds of the variables reflect price information for various goods and services in the reporting city and are, and are expected to be, rather noisy and ill-shaped due to the inconsistencies expected with crowd-sourced data.

The last one-third of variables in the dataset are from the 2018 Index of Economic Freedom database. These distributions exhibit highly multimodal tendencies which is primarily due to the discrete nature of values for economic variables assigned to countries and regions by the Heritage Foundation.

Algorithms and Techniques

The set of machine learning algorithms applied to this dataset can be summarized in the following table. Each is then further discussed.

Algorithm	Algorithm Purpose/Intent
Box Cox Transformation (forward and inverse)	Normalization of data distributions for ingestion by other ML algorithms
Min-Max Scaler	Scaling so as to apply equal weighting for each explanatory variable's potential importance
Principal Component Analysis (PCA) decomposition	Explanatory variables space reduction and variable correlation analysis
Gaussian Mixture Model Clustering	Derive further structure from the dataset to aid in dataset interpretation
Random Forest Regression	Train and apply algorithm on explanatory variables to predict real estate market prices for cities around the world
Various Visualization Algorithms	

Box Cox Transformation: As was mentioned, the raw data input distributions tend, in many instances, to differ significantly from statistically normal distributions. The primary importance of this observation is that many machine learning algorithms, including those used here, assume and work best on data that is normalized. To this end, a Box Cox normalization algorithm was applied to most of the explanatory data variables in the dataset. The lambda parameter used in each case was automatically chosen by the algorithm to be that which maximizes the log-likelihood function.Some variables that were not statistical in their nature to begin with, such as latitude and longitude, were not treated with this transform.

Min-Max Scaler: A min-max scaler was run on all dataset variables so as to apply equal weighting to each explanatory variable's potential importance as used in machine learning algorithms. Without this, for example, the population variable, running from 25,000 to a maximum of about 31,000,000 would swamp the effect of explained variance attributed to smaller range data such as the cost of an inexpensive meal that runs from about one to thirty-one (dollars).

Principal Component Analysis: Principal component analysis (PCA) was performed on the dataset to gather and transform explanatory variables to a smaller set of orthogonal variables that maintained the bulk of the information contained in the original variables. For example, if the cost of white bread was highly correlated with the cost of flour in a dataset, we could replace the two variables (costBread and costFlour) with

their mean ((costBread+cFlour)/2), thus reducing the number of tracked variables. Maintaining 100% of the information in the original dataset is only attainable, in general, by transforming to a space that contains as many new variables as there were in the original dataset. Transforming to fewer variables risks losing information. In our analyses, we chose to transform to a number of new variables that was sufficient so as to explain at least 95% of the variance of the original variables. In the dataset for this project, there are many variables, including price variables. The expectation is that many of these will be correlated and thus we might be able to reduce the number of variables in our calculations by applying PCA to the data.

Gaussian Mixture Model Clustering (GM): This algorithm was applied to gain insight by deriving further structure from the dataset. We would hope that subsets of the data would coagulate, such as there being a range lower priced values that were widely separated and distinct from a higher-priced group of values. Or that prices in one geography might exhibit significant difference in prices in a different geography. To this end we applied GM clustering to our transformed and scaled data to see if we could observe such cluster groups. Just how many clusters, n, there might be can be tested by setting, a priori, the value n for the algorithm and observing cluster results. In addition to choosing n for the clustering algorithm, we also changed and observed results by changing the convergence tolerance limit and scored results with a cluster silhouette coefficient. Cluster means (centers) were derived and analyzed.

Random Forest Regression: The Random Forest regression was chosen as a means to learn and estimate city real estate market prices from our dataset explanatory variables. This algorithm is known to be quite robust to noise and known to perform well on heterogeneous datasets such as ours. This algorithm was trained on a subset of the transformed and scaled data as means for the algorithm to learn to estimate city real estate market prices from the explanatory variables. The most important parameters tuned to be tuned in our application of this algorithm were the number of estimators (number of trees in the forest), the number of features from the dataset to be used, and the maximum depth of the trees.

Benchmark

Beyond conventional observations of correlations in price for goods and services, real estate prices, our results from our random forest estimator will be compared with a linear regression model that fits the training sample apartment prices to a hyperplane parameterized by all of the . This is reasonable as we would expect, for example, that real estate prices would be, to first order, linearly correlated with the various explanatory variables found in the dataset. This would seem particularly true of the currency based price data.

III. Methodology

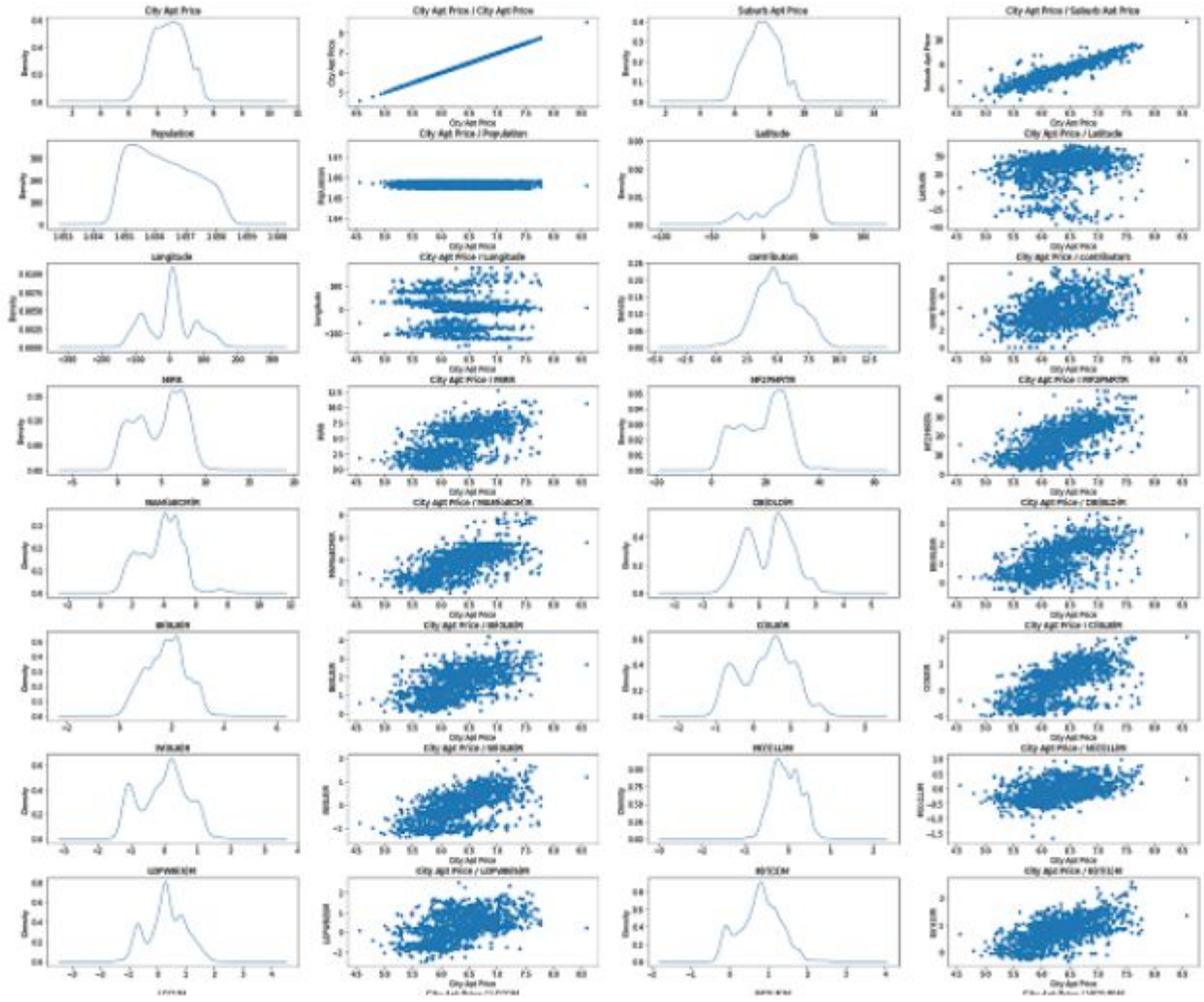
Data Preprocessing

Basic Data was cleaned by

- incomplete records
- Removing entries containing NaNs.
- Dropping data columns such as population and country ID that were common to joined datasets or ancillary for our objective
- Numeric strings were changed to floating point values
- Target variables were separated from explanatory variables
- Variable naming conventions were change for brevity and ease of interpretation
- Standard strings were changed to unicode strings

After data cleaning, the data was displayed and it was observed that the raw data distributions (expressed in KDE plots) tended, in many instances, to differ significantly from statistically normal distributions

A Box Cox normalization algorithm was applied to most of the explanatory data variables in the dataset. The distribution of the Box Cox transformation for the only the first few explanatory variables in the dataset is shown in the following figure, along with correlation scatter plots of the explanatory variables against the similarly transformed target variable (City Apt. Prices).



Most of the variables in the image are city price data for various goods and services. The distributions can be seen to be nearer ideal normal distributions.

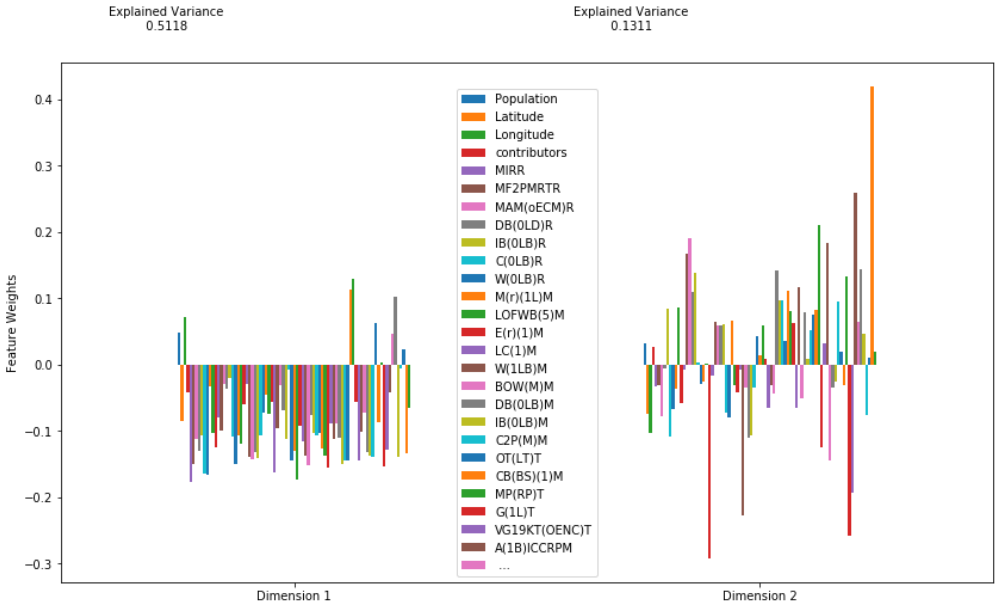
Additionally, the scatter plots, though noisy, tend to reveal a correlation with apartment prices.

After application of the Box Cox transformation min-max scaler was run on the dataset to apply apply equal weighting to each explanatory variable's potential importance as used in machine learning algorithms.

Significant software coding effort was expended to maintain index alignment between python pandas DataFrames for the separately managed target variables and the

DataFrames that held the explanatory variables. Other than these challenges, coding proceeded as expected.

After data cleaning, statistical transformation (Box Cox), and scaling was done, a reduction in data dimension for the explanatory variables was performed by transforming it to a lower dimension using principal component analysis (PCA).

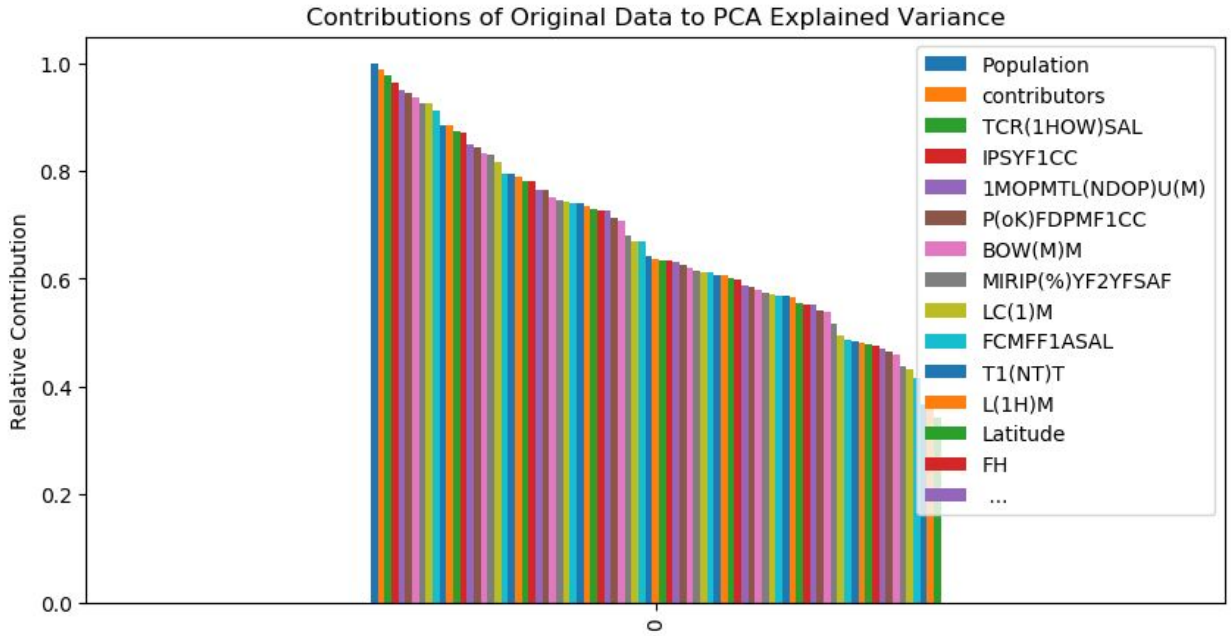


The first two PCA dimensions are shown in the illustration above. Ultimately, the PCA dimension of the explanatory variable data was brought from 85 features down to 30. In transforming to this set of 30 basis vectors, a cumulative explained variance of 95.8 % was maintained. Only the first two PCA component dimensions are shown in the figure.

As can be seen in the figure, the first component demonstrates the high degree of correlation of a great many of the explanatory variables (indicated by the numerous down-going colored bars). This information argues favorably for using PCA on this dataset.

In addition, the PCA analysis was used to derive the relative contributions to the dataset of the original explanatory variables. From the figure below and using the variable name glossary in the appendix, we can see that the first three dominant contributors to the dataset variance are city Population number, contributors (number of people who contributed data for each city to the dataset), and the variable “Tennis Court Rent (1 Hour on Weekend), Sports And Leisure.” High variance in this last variable is understood as,

possibly, renting court time and paying to play tennis is probably very culturally specific and variable.



Implementation

Python code was written to estimate city apartment market prices using the Scikit Learn Random Forest Regressor. The regressor was applied to the PCA transformed data set to learn to estimate city real estate market prices. A cross-validation subset of market apartment prices was withheld from training. It was found that Random Forest R2 score increase monotonically and asymptotically against the cross-validation test data with increases in each of the tuning parameters. Ultimately, the primary concern in choosing parameter for the Random Forest regressor turned out to be runtime. Ultimately, even runtime was not a problem with this dataset so that maximum values for each of the parameters was used.

A a linear regression model that fit the training sample apartment prices to a hyperplane parameterized by all of the explanatory variables was also applied to the dataset to act as a benchmark against which our Random Forest regressor performance could be compared.

An R2 performance metric was used to indicate goodness of fit for both the Random Forest regressor and the linear regression model. Very few complications were encountered in this phase of the project and progress was smooth.

Refinement

After coding, the algorithm pipeline was exercised with default tuning parameters. Each algorithm in the pipeline was then independently adjusted to derive optimal tuning parameters. This was a manual and iterative process with some automated search to find small sets of optimal parameters. Time investment in the coding of fully populated global grid search tuning was expected to cost more time than the value it might add. .

The following tuning parameter initial and final settings for the project are summarized as follows. Parameters not listed use their python library defaults or were less important or non-performance related.

Algorithm Parameter	Initial Value	Tuned Value	Notes
Box Cox Transform	(No Tuning Params)	(No Tuning Params)	Scipy.stats.boxcox automates the choice for the lambda parameter that optimizes the log-likelihood function
Min-Max Scaler Feature_range	 (0,1)	 (0,1)	Scaling applied equally to all variables to equalize their effect
PCA N_components whiten svd_solver	 (None) False (None)	 40 False (randomized)	Running a search on n_components yielded a cumulative explained variance of 97% at 40 components, reducing dimensionality by 2
Gaussian Mixture Model Clustering n_components Covariance_type Tol n_int init_params	 10 'Full' 0.1e-3 1 'kmeans'	 2 'Full' 0.1e-8 40 'kmeans'	A search over n_components and init_params yielded an optimal (silhouette score) number of clusters of 2.
Random Forest Regression N_estimators max_features	 10 2	 30 40	A grid search was performed over n_estimators and max features yielding the final shown parameters.

IV. Results

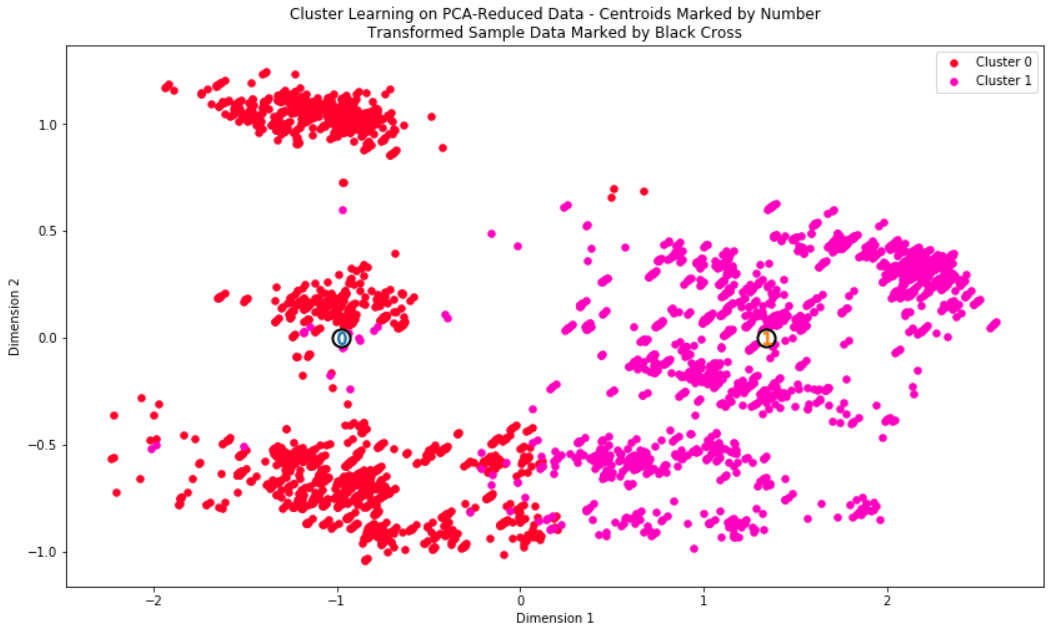
Model Evaluation and Validation

Data Clustering and Cluster Centers

As an ancillary step to aid in dataset comprehension, unsupervised machine learning in the form of clustering was applied to the PCA transformed data. A Gaussian Mixture Model clustering algorithm was chosen for this task. The derived cluster centers were inverse transformed from the PCA dimension into original explanatory variables. More will be said about the interpretation of these cluster centers in the discussion below. A cluster silhouette coefficient was used as a metric for determining the best number of clusters in the data.

The clustering algorithm was applied with the aggressive convergence parameters of 1e-8 for EM convergence threshold and n_iter random starting positions of 40 (tol=1.0e-8, n_iter=40). The number of clusters (n_components) was optimized manually by observation of the silhouette coefficient whose maximum was 0.38 at n_components = 2 (2 clusters)

The results of the clustering analysis was also projected onto the dominant (first two) PCA dimensions and is shown in the figure, below. It is difficult to interpret this clustering scatter plot except to say that the two clusters are roughly separated along the middle of dimension 1. From the PCA analysis, above, we can see that this dimension is highly correlated with the cost of goods and services in a locale and, to a lesser extent, with the parameters of the Heritage Foundation's Index of Economic Freedom. So, a first observation might state that one cluster has significantly more economic “weight” than the other cluster. More interpretation will be given regarding clustering results later in this paper.



Estimating Apartment Prices with Random Forest Regression

A 80/20% split of training and test data of explanatory and target variables was processed by the Scikit Learn RandomForestRegressor algorithm using 30 estimators (trees) and all other parameters set to default (as discussed above). The performance of this machine learning algorithm, with respect to the R2 metric, for estimating apartment prices across the globe is summarized in the following statistics.

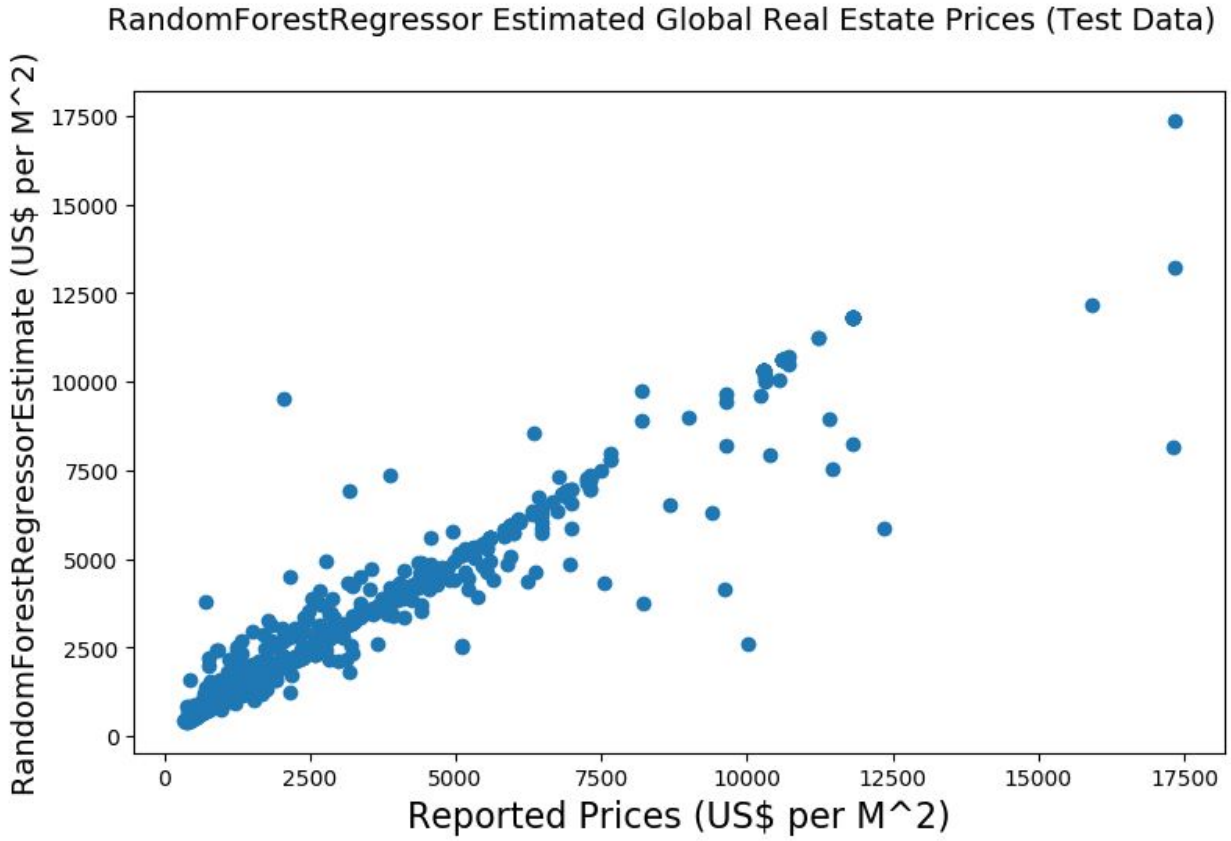
Training set has 3559 samples.
Testing set has 890 samples.
RandomForestRegressor trained on 44 samples.
Training time: 0.0880000591278
Prediction time: 0.00899982452393
R^2 TRAIN Score: 0.95124303259
R^2 TEST Score: 0.465391644163

RandomForestRegressor trained on 445 samples.
Training time: 0.240999937057
Prediction time: 0.0090000629425
R^2 TRAIN Score: 0.975528587502
R^2 TEST Score: 0.796886515618

RandomForestRegressor trained on 4449 samples.
Training time: 2.41799998283

Prediction time: 0.0
R^2 TRAIN Score: 0.992180022454
R^2 TEST Score: 0.928124990549

We observe monotonically increasing estimation performance as the training dataset size is increased culminating in a test score of 0.93. A scatter plot of reported apartment prices vs those estimated by the algorithm is shown in the figure below. The algorithm shows a strong tendency to correctly estimate prices with an obvious line of correct estimations prominently imposed across the diagonal of the figure. More discussion of these results will be had below.



Before tuning as discussed in the Refinement section of this paper, running the entire algorithm pipeline with default algorithm parameters values yielded an R2 score of 0.92. After parameter tuning, increases in R2 performance was marginal and asymptotic toward 0.94 as runtime increased but performance did not. The two largest changes to overall pipeline performance were the choice of n_components and init_params in the Gaussian Mixture clustering model. All other parameters has small effect compared to

those parameters. Once those two parameters were optimized and held fixed, the entire system performance stable and fixed near its optimal R2 value of 0.93. This suggests that the system is probably quite robust to the ingestion of other data sets (though none can be had, at the moment), and more fundamental algorithmic changes might be employed for more substantial increases in performance. One interesting method suggested for increasing performance (but left for further study), is the possible use of the XGBoost regression algorithm.

Justification

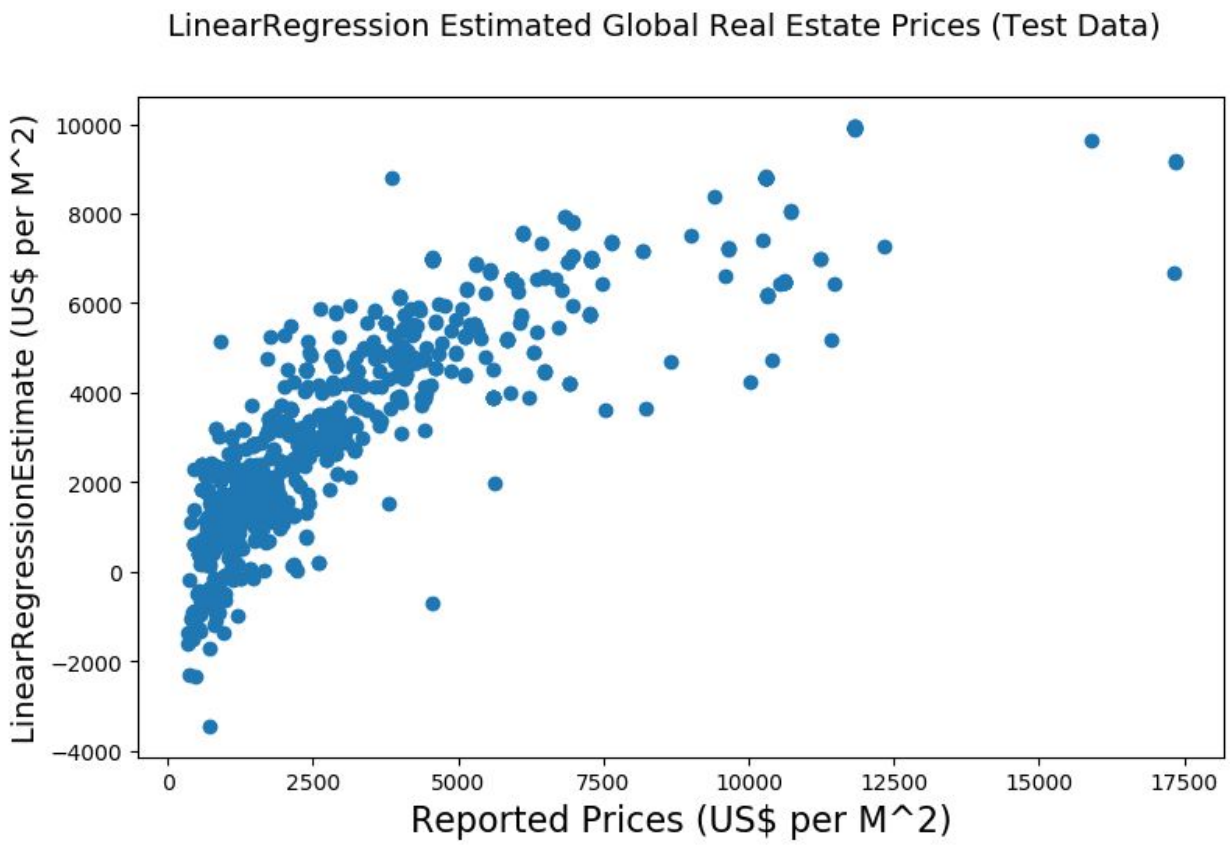
Benchmarking

As a benchmark, linear regression was used to fit the same 80% training set of explanatory variables (expressed by their principal components) to a similar data split of known target apartment prices from around the world. The R2 metric of fit for the resulting linear regression hyperplane to the data is shown below.

Training set has 3559 samples.
Linear Regression trained on 4449 samples.
Training time: 0.00999999046326
Prediction time: 0.00100016593933
R^2 TRAIN Score: 0.763426292819
R^2 TEST Score: 0.744055473825

This score for the benchmark linear regression algorithm when trained on a relatively large dataset is 0.74. This is significantly worse performance than the Random Forest's score of 0.93.

A scatter plot of city apartment prices estimated by the linear regression algorithm is shown in the figure below.



We see significant error in the scatter plot where the linear regression model, rather consistently, overestimates prices for apartments in the global cities in this database. This is not great performance for the benchmark algorithm.

However, it was later realized that much, if not most of the explanatory data, has essentially been log transformed by the Box Cox transformation. This is not the case for the target variable. The target variable has been left in its original units (US $ price). Thus, a linear relationship, if it did exist, is broken by applying Box Cox transformation to explanatory variables yet withholding such transformation from the target variable. Ultimately, it was realized that this is not a fair, but is an erroneous benchmark comparison. For linear regression to be valid as a benchmark, we would need to operate the benchmark estimator with explanatory and target variables both in the same space (i.e. log-log, or linear-linear). This is an analysis correction left for the future.

V. Conclusion

Free-Form Visualization

Interpretations of Data Clustering Results

The clustering algorithm, as discussed above, converged most strongly to a two-cluster model in PCA coordinates. Because the data was both multi-dimensional and transformed to PCA coordinates, it was difficult to interpret. The scatter plot of multi-dimensional clusters projected onto the 2D plane did not avail itself to easy interpretation, either.

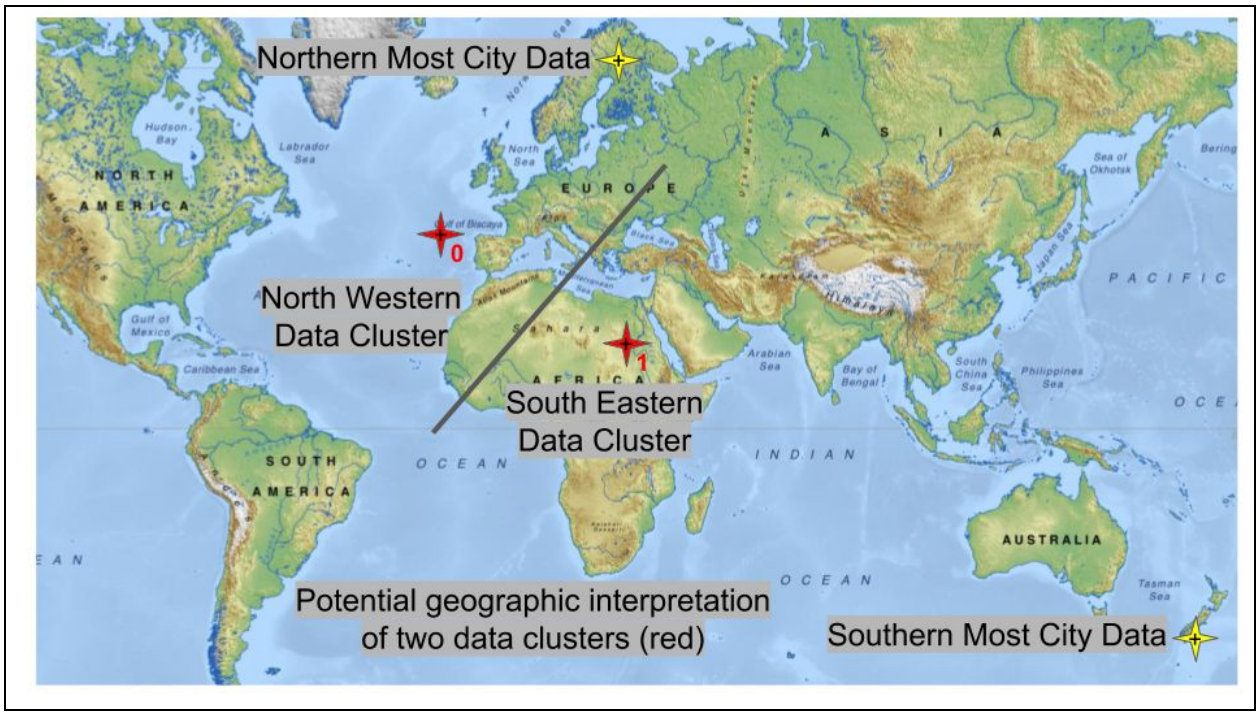
However, it is interesting to transform the two derived cluster centers from PCA back into their original coordinates of dollars, population numbers, GPS coordinates, etc. Performing that transformation to the cluster centers yields the following cluster center expressed in original explanatory variables. Only part of the data is shown as the number of explanatory variables (85) is rather inconvenient for easy display.

Cluster centers (2) transformed to original explanatory variables

Population	Latitude	Longitude	contributors	MIRR	MF2PMRTR	MAM(oECM)R	DB(0LD)R	IB(0LB)R	C(0LB)R	W(0LB)R	M(r)(1L)M	LOFWB(5)M	E(r)(1
51192.257758	43.024265	-16.197654	81.206993	13.653018	54.920962	7.915684	4.553124	4.766323	2.027376	1.498081	1.037935	1.841298	2.5212
73703.816010	17.348667	29.117447	47.929768	3.890101	19.040040	4.647397	1.759045	2.638854	0.734577	0.457405	0.884003	0.814000	1.4135

If the above cluster center feature vectors (shown abbreviated in length) are examined it becomes easier to see along what lines the cluster split is made. Particularly, the monetary features (such as price of a meal, the cost of a bus ticket, etc.) are consistently higher in value for the first cluster center than for the second. For example, the data splits in one case where the cost an inexpensive meal (MIRR) is approximately $14 in the first vector as compared to about $4 in the second. Likewise, the cost of a domestic beer in a restaurant (DB(0LD)R) is about $5 in the first cluster and $1.75 in the second. In fact, the main explanatory variables indicating an inversion of this are for parameters associated with government controlled activities such as tax burden (TB) and government spending (GS) (parameters not show due to space constraints). For these latter types of variables, the second cluster center clearly shows a propensity for greater magnitude (greater tax burden, greater government spending, and lower economic prices). As shown in the map below, a simple objective observation of the global coordinates (latitudes and longitudes) shows the geographic center of the higher priced, lower government burden cluster center to be roughly west of the european continent. The geographic center for the second, lower priced cluster is in northeast Africa. A line drawn on the map halfway between these two clusters places most of western and north Europe in the first cluster and the middle east and Africa in the other cluster. It is quite well known that the economies of these two regions, for whatever reason, contrast significantly with one another. It is also interesting to conjecture that

our machine learning clustering algorithm seems to have a roughly economically and geographically (and quite inadvertently) bounded the developed and third worlds.



Cluster Centers Map

Interpretations of Global Real Estate Price Prediction Results

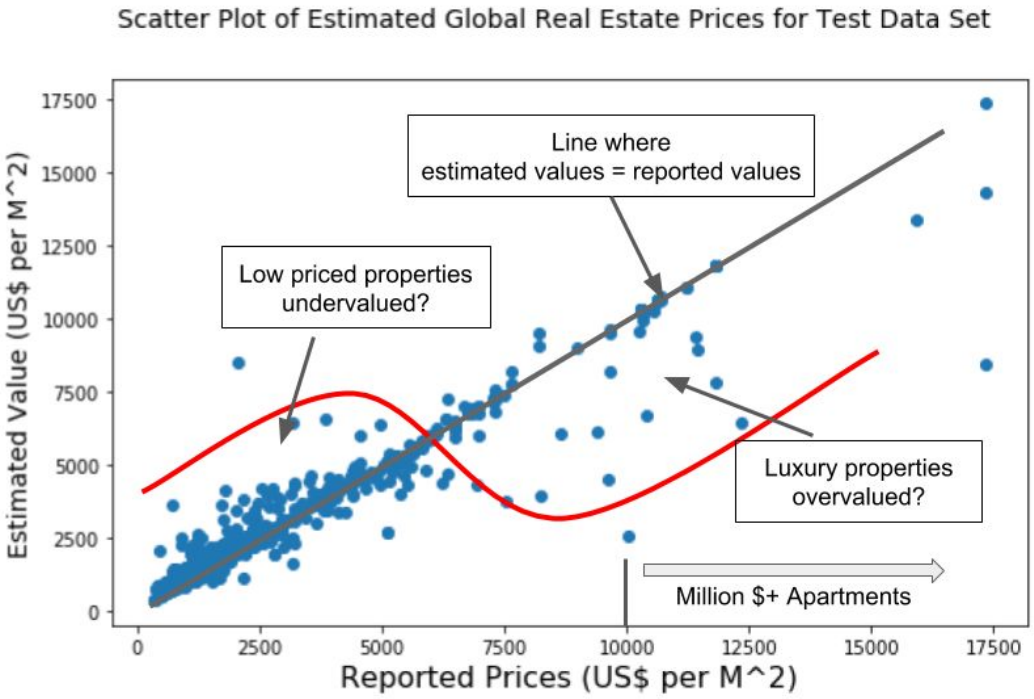
As shown in the above section, our Random Forest regression estimator did a decent job of predicting global city real estate prices from other economic and government data. We can further discuss these results with the aid of the figure, below, which is the same figure of Random Forest predictions shown previously. However, now, further interpretation and conjecture as to the value that such information might lend.

We have imposed some lines and comments on the scatter plot of price estimation results. We see that at the lower left of the plot that our estimator seems to rather consistently yield higher valuations for properties in lower priced markets. Conversely, the upper right of the plot seems to show that our Random Forest estimator seems to yield lower valuations for apartment prices in the expensive, luxury market.

What might this mean to an investor seeking guidance in their real estate transactions? *If* we can assume that our machine learning model can be trusted to yield “true” values based on both local and broad economic data, it might be suggested that investors

exercise great caution in higher priced market as it appears that real estate transactions there might often exhibit higher valuations than objectively derived by machine learning from the broader economy. That is, one might have more of a tendency to overpay for properties in more expensive cities.

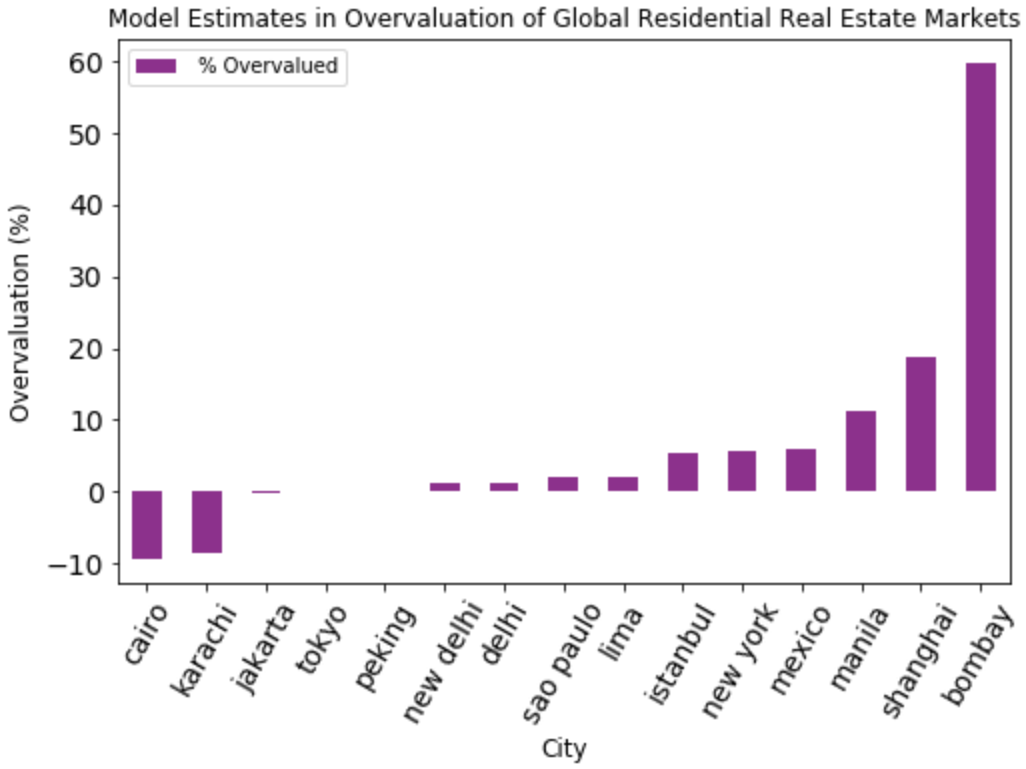
By the same logic, an investor might be more likely to “get a good deal” when purchasing properties in lower priced markets.



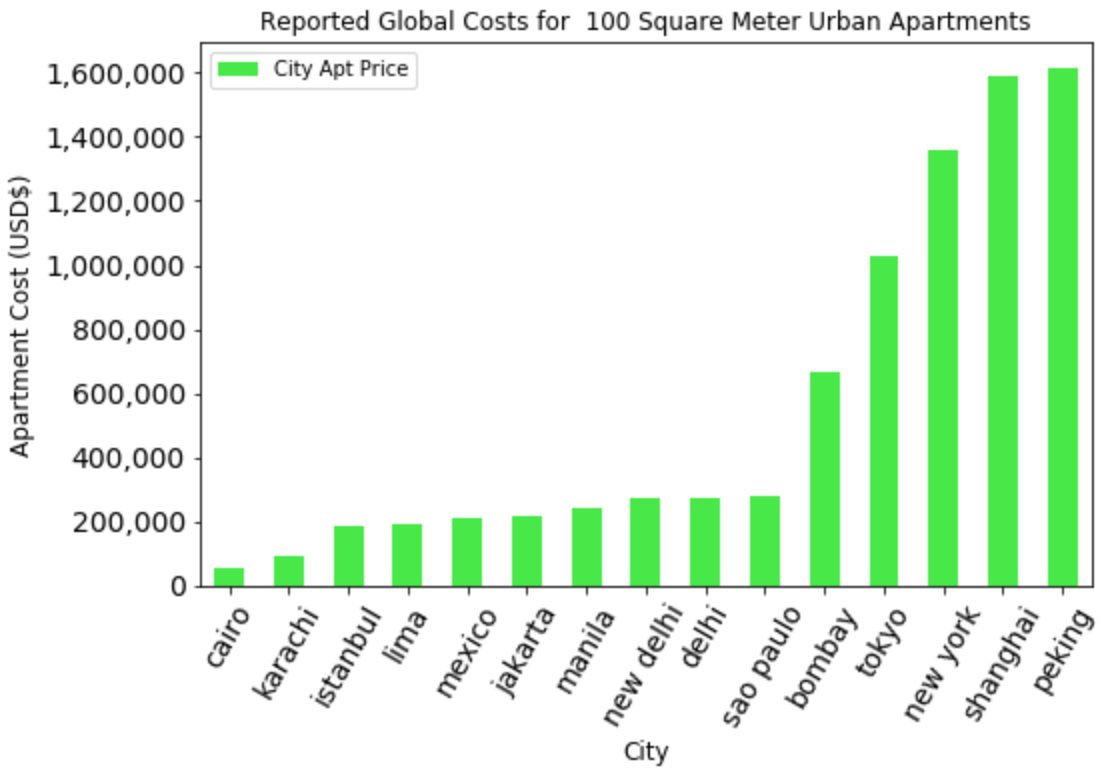
Might this really suggest an overvaluation in luxury markets worldwide? Or it reflect a temporal lag, where prices will move in the direction of the high-priced outliers to “catch up” in time? Because our data has no temporal component (i.e. is a snapshot in time), it is unknown which interpretation, if any might be correct, and would merit further study.

As a final observation, we might look at how the model interpreted apartment valuations in a few international locations. The first figure below shows results from estimates that our machine learning model made in fifteen larger world cities. The second figure simply shows prices reported by crowd-sourced contributors. In the figure, the definition for overvaluation is taken to be the percent difference between crowd-sourced prices and our machine learning prices, relative to the machine learning prices. By that definition, our estimates suggests, for example, that prices in Mumbai, India, might be highly overpriced. (Mumbai is listed as “bombay”, as the world cities database has yet to change this name. Underlying economic data is, however, current). Slight overvaluation

is also suggested for New York and Mexico Cities. Conversely, some Middle East locations such as Cairo and Karachi, Pakistan suggest under-valuations.



Before concluding that these results are obvious, we might consider the model's interpretation of apartment prices in Tokyo, Japan, the world's largest metropolitan area. In Tokyo, city center apartments are reported to be in the million dollar range. Here, our model suggests that valuations are, in fact, justified by underlying political and economic data.



Reflection

This project used crowd-sourced economic data from Numbeo.com for 4500 individual cities, worldwide, and utilized that data to train a Random Forest regression machine learning algorithm to generalize estimations of worldwide city apartment prices. Crowd-sourced data was augmented by inclusion of data from The Heritage Foundation's 2018 Index of Economic Freedom country database and population data from Mind Max's world cities database. The project yielded very respectable R2 measures for price estimation goodness of fit of 0.93.

It was noted that subjective human qualification of results is necessary, and that even benchmark metrics must be subject to appropriate interpretation.

Through the process of principal component and cluster analyses of the dataset, further interesting observations were derived, not only for structure in world real estate prices, but also for structure in world economic and political conditions, as well.

Finally, if our models can be relied on, we have shown that they might be used as tools to aid, among others, real estate investors by flagging potential, highly localized,

individual real estate markets as being overvalued, undervalued, or on par with the underlying local economic and political variables.

Improvement

Many improvements to this project and model might be made.

Indeed, other algorithms, such as deep learning with neural networks, can be applied and their performance compared to the presently studied Random Forest regressor. However, as is frequently stated in machine learning discussions, higher quality and larger datasets might yield better performance improvements than would the application of the latest fashion in in machine learning algorithms.

Importantly, the element of time should be introduced to the estimator as we know that real estate and, in fact, all economic activity changes very fundamentally with time. The introduction of time in these analyses would would require deep and important changes to the dataset, assumptions, models, and machine learning algorithms and training processes.

Aside from temporal predictions, additional and potentially more relevant and powerful economic, political, and physical explanatory data might be used to train our machine learning system to yield even better results. Further explanatory variables such as local econo political information and trends such as localized tarif impositions, location-based imagery, weather and climate and other information, appropriately applied, would certainly improve our machine learning predictions.The breadth of potential improvements could certainly fill a career.

Appendix: Explanatory Variable Name Definitions

City:	City
CountryName:	CountryName
Population:	Population
Latitude:	Latitude
Longitude:	Longitude
contributors:	contributors
MIRR:	Meal, Inexpensive Restaurant, Restaurants
MF2PMRTR:	Meal for 2 People, Mid-range Restaurant, Three-course, Restaurants
MAM(oECM)R:	McMeal at McDonalds (or Equivalent Combo Meal), Restaurants
DB(0LD)R:	Domestic Beer (0.5 liter draught), Restaurants
IB(0LB)R:	Imported Beer (0.33 liter bottle), Restaurants

C(0LB)R: Coke/Pepsi (0.33 liter bottle), Restaurants
W(0LB)R: Water (0.33 liter bottle) , Restaurants
M(r)(1L)M: Milk (regular), (1 liter), Markets
LOFWB(5)M: Loaf of Fresh White Bread (500g), Markets
E(r)(1)M: Eggs (regular) (12), Markets
LC(1)M: Local Cheese (1kg), Markets
W(1LB)M: Water (1.5 liter bottle), Markets
BOW(M)M: Bottle of Wine (Mid-Range), Markets
DB(0LB)M: Domestic Beer (0.5 liter bottle), Markets
IB(0LB)M: Imported Beer (0.33 liter bottle), Markets
C2P(M)M: Cigarettes 20 Pack (Marlboro), Markets
OT(LT)T: One-way Ticket (Local Transport), Transportation
CB(BS)(1)M: Chicken Breasts (Boneless, Skinless), (1kg), Markets
MP(RP)T: Monthly Pass (Regular Price), Transportation
G(1L)T: Gasoline (1 liter), Transportation
VG19KT(OENC)T: Volkswagen Golf 1.4 90 KW Trendline (Or Equivalent New Car), Transportation
A(1B)ICCRPM: Apartment (1 bedroom) in City Centre, Rent Per Month
A(1B)OOCRPM: Apartment (1 bedroom) Outside of Centre, Rent Per Month
A(3B)ICCRPM: Apartment (3 bedrooms) in City Centre, Rent Per Month
A(3B)OOCRPM: Apartment (3 bedrooms) Outside of Centre, Rent Per Month
B(EHCWG)F8AU(M): Basic (Electricity, Heating, Cooling, Water, Garbage) for 85m2 Apartment, Utilities (Monthly)
1MOPMTL(NDOP)U(M): 1 min. of Prepaid Mobile Tariff Local (No Discounts or Plans), Utilities (Monthly)
I(6MOMUDC)U(M): Internet (60 Mbps or More, Unlimited Data, Cable/ADSL), Utilities (Monthly)
FCMFF1ASAL: Fitness Club, Monthly Fee for 1 Adult, Sports And Leisure
TCR(1HOW)SAL: ennis Court Rent (1 Hour on Weekend), Sports And Leisure
CIR1SSAL: Cinema, International Release, 1 Seat, Sports And Leisure
1POJ(L5OS)CAS: 1 Pair of Jeans (Levis 501 Or Similar), Clothing And Shoes
1SDIACS(ZH.)CAS: 1 Summer Dress in a Chain Store (Zara, H&M, ...), Clothing And Shoes
1PONRS(M)CAS: 1 Pair of Nike Running Shoes (Mid-Range), Clothing And Shoes
1POMLBSCAS: 1 Pair of Men Leather Business Shoes, Clothing And Shoes
TC19C(OENC)T: Toyota Corolla 1.6l 97kW Comfort (Or Equivalent New Car), Transportation
P(oK)FDPMF1CC: Preschool (or Kindergarten), Full Day, Private, Monthly for 1 Child, Childcare
PPSMTBAICCBAP: Price per Square Meter to Buy Apartment in City Centre, Buy Apartment Price
IPSYF1CC: International Primary School, Yearly for 1 Child, Childcare
PPSMTBAOOCBAP: Price per Square Meter to Buy Apartment Outside of Centre, Buy

Apartment Price
AMNS(AT)SAF: Average Monthly Net Salary (After Tax), Salaries And Financing
MIRIP(%)YF2YFSAF: Mortgage Interest Rate in Percentages (%), Yearly, for 20 Years Fixed-Rate, Salaries And Financing
TS(NT)T: Taxi Start (Normal Tariff), Transportation
T1(NT)T: Taxi 1 km (Normal Tariff), Transportation
T1W(NT)T: Taxi 1 hour Waiting (Normal Tariff), Transportation
A(1)M: Apples (1kg), Markets
O(1)M: Oranges (1kg), Markets
P(1)M: Potato (1kg), Markets
L(1H)M: Lettuce (1 head), Markets
C(r)R: Cappuccino (regular), Restaurants
R(w)(1)M: Rice (white), (1kg), Markets
T(1)M: Tomato (1kg), Markets
B(1)M: Banana (1kg), Markets
O(1)M2: Onion (1kg), Markets
BR(1)(oEBLRM)M: Beef Round (1kg) (or Equivalent Back Leg Red Meat), Markets
Region: Region
2S: 2018 Score
PR: Property Rights
JE: Judicial Effectiveness
GI: Government Integrity
TB: Tax Burden
GS: Gov't Spending
FH: Fiscal Health
BF: Business Freedom
LF: Labor Freedom
MF: Monetary Freedom
TF: Trade Freedom
IF: Investment Freedom
FF: Financial Freedom
TR(%): Tariff Rate (%)
ITR(%): Income Tax Rate (%)
CTR(%): Corporate Tax Rate (%)
TB%OG: Tax Burden % of GDP
GE%OG: Gov't Expenditure % of GDP
G(BP): GDP (Billions, PPP)
GGR(%): GDP Growth Rate (%)
5YGGR(%): 5 Year GDP Growth Rate (%)
GPC(P): GDP per Capita (PPP)
U(%): Unemployment (%)
I(%): Inflation (%)

FI(M):	FDI Inflow (Millions)
PD(%OG):	Public Debt (% of GDP)