Big Data and Machine Learning Applications for Real-Estate Sector

ONUR CAN

MS. DATA SCIENCE NON-THESIS MASTER TERM PROJECT

Abstract

Large portion of economic data and information that are generated have been increasingly shifting to the digital environment. This is also correct for the real-estate sector. Big Data and Machine Learning tools help real estate companies and also individuals to make informed business decisions, understand the business environment. In this analysis, the process of how to transfer real estate data from unstructured format to structured, some of the important features and methods and applications for transferring big data into Machine Learning (ML) and Artificial Intelligence (AI) solutions will be discussed. The report also includes a case study of an American Company "Zillow" and its experimental machine learning application for the real estate market. Lastly, a detailed analysis of ML applications on the California Housing Dataset will be provided with source codes and outcomes for real-estate sector pricing.

Table of Contents

1. Introduction

The real estate market is one of the backbones of the current economic system. People always need new houses to live and businesses need new places for their operations. Nowadays, the real estate definition is also very broad, it includes usage rights of all the land and the buildings, along with its natural resources such as crops, minerals and water.

Real estate as with any other commodities can be often purchased as an investment. In an environment where land and building prices always increase or at least fluctuate, there is an opportunity for arbitrage or quickly rising value. Renovations would also substantially raise the value of the property as depicted by Morgan Stanley Capital International (MSCI) in *Figure 1***.** In addition to this, real estate is sometimes used as a way to store value without any particular attempt to rent it out or for personal use. Real commodities such as real estate or gold are also considered good hedges against inflation in the economy. Nevertheless, the conventional analytical methods and old data gathering techniques fail or are slow to build robust business plans to capture these investment opportunities. Hence in this report, the values that can be created by using big data and several Machine Learning (ML) applications will be covered to realize these opportunities in time for the real-estate sector.

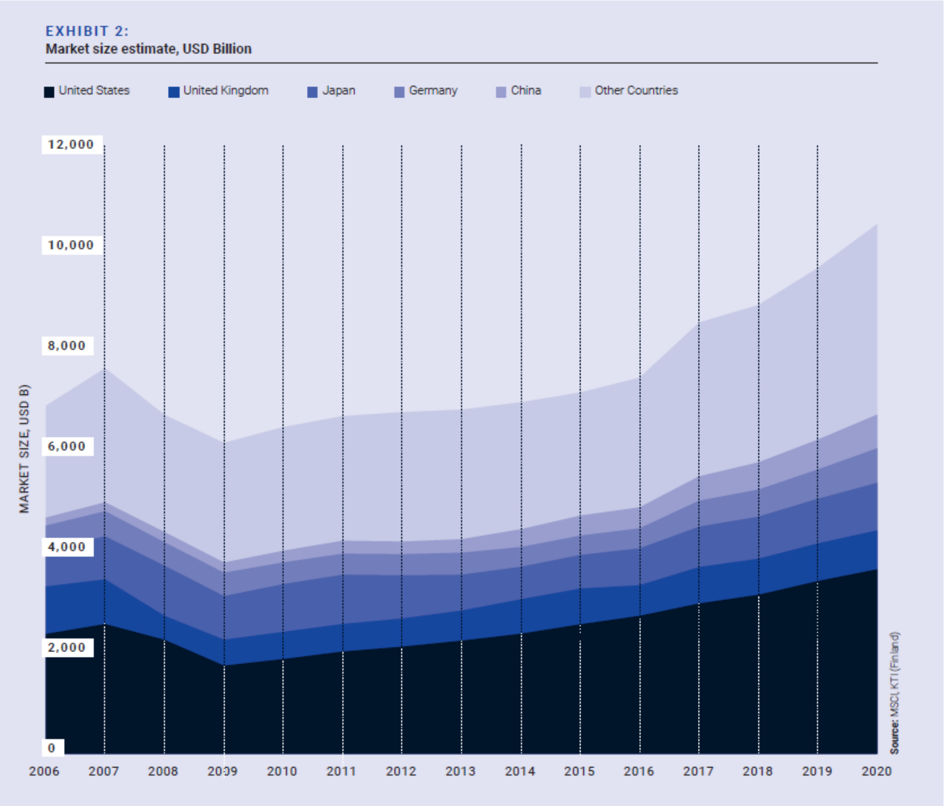


Figure 1. MSCI Real Estate Market Size 2006-2020 [1]

2. Literature Review

Understanding the characteristics of the real estate sector and the market is a significant portion of the price analysis. Outcomes from the business environment analysis will be the input of the generalization statements that will be offered later stages of the paper. There are many sources that explain the real estate sector however in this report, MSCI yearly global report [1] is used to gain insight into the global market, patterns and individual country real estate information.

An article from McKinsey Company is used to gather information about feature extraction because feature selection is a common component in supervised machine learning pipelines and is essential when the goal of the analysis is knowledge discovery. The article states regional differences in real estate feature importance and the reasons behind them. [2]

The main methods and standard process for Data Mining (DM) are detailed in the Modeling Factors article which also involves identifying and creating new data points which can be calculated from existing entries. [3] The information gained through data preparation is then used to create various models of behavior where analytical processes automatically find patterns in data and use them to generate insights. The several different modeling applications and their respective evaluation procedures are examined in the Islamabad Housing Data research. [4]

The last section of the report is reserved for the Zillow case and outcomes. Information on the topics is gathered via journal websites and articles on the topic. [5]

3. Methodology

The main goal of this report is to give an overall understanding of the reader of the real estate business environment and how big data, data mining and machine learning applications can influence the sector in the future. Therefore, the analysis starts with explaining the conditions of the current real estate market and how certain characteristics that are called features may be more important than others. Then the standardized processes of how to identify, collect and analyze the data sets that can help the reader construct good models are stated. Construction of the modeling phase focuses on real-life data and technical model assessments were provided to emphasize the difference between modeling techniques and accuracy. Given case study and applied exemplary machine learning solutions summarizes all the topics discussed in the report and provide example solutions for the business case.

4. Automated Data-Driven Actions

The availability of data and the ability to analyze it quickly allowed us to gain actionable insights into the real estate market as opposed to previous conventional methods. Moreover, new technology solutions make it also easier to collect, store and automate the data collection process from several data providers which will be detailed in Section 5. All of these factors contributed to an automated decision-making process from data collection to final price prediction. Some of the key examples of automated decision actions through the usage of advanced data analytics are listed below.

Ability to Extract Patterns and Forecasts: The analysis of ever-changing prices with real time data can show the direction of local or regional real-estate market. [2]

Designing Personalised Offers: The market data can be tailored according to personal needs for real estate. At the end best offers can be served to customers for their individual needs.

Understanding the Right Mix of Features: Based on local area, time and country, the feature significance analysis can be conducted to understand best combination of parameters for the region and time.

Assessment of Under and Over Valuation: The analysis of past and current data can indicate areas with future potential or areas that are currently overpriced.

Ability to Scale Different Scenarios: The approaches and Insights can be tailored and automated for simulation purposes or nowcasting with imaginary parameters. [2]

Gathering Insight: Through real estate data analytics many third party agents and academic society would be able understand and study the dynamics of the market.

5. Value Created & Risks

The stakeholders for real-estate data analytics can be listed as but not limited to *mainstream people i.e. public who usually purchase for their personal needs*, *investment portfolio firms in the real estate market* and *service and data providers* for the first two groups. [2] The main concepts behind real estate value creation are the precise calculation of the value of a real estate market by looking at past data and the ability to reflect this evaluation during the time when information is relevant. This will allow smart purchases and sales for ordinary people and make robust investment decisions for investment companies. Some of the value creation examples for the groups who have all different agendas for the real-estate market are listed below.

1 Fast Evaluation

Main stream people who would like to get quick estimate for their real estate purchases/sales can access this information quickly.

2 Smart Purchases

Understanding and modeling real estate market patterns offers arbitrage opportunities ahead of the market for both Investment firms and public.

3 Better Capital Expenditure Decision

For Investment Firms the data analytics allows better calculation of specific metrics such as return on investments or stabilized yield cost etc. [2]

4 Regional Insights

Through data analytics Investors and economic agents can formulate investment decisions in other sectors that are correlated with real estate.

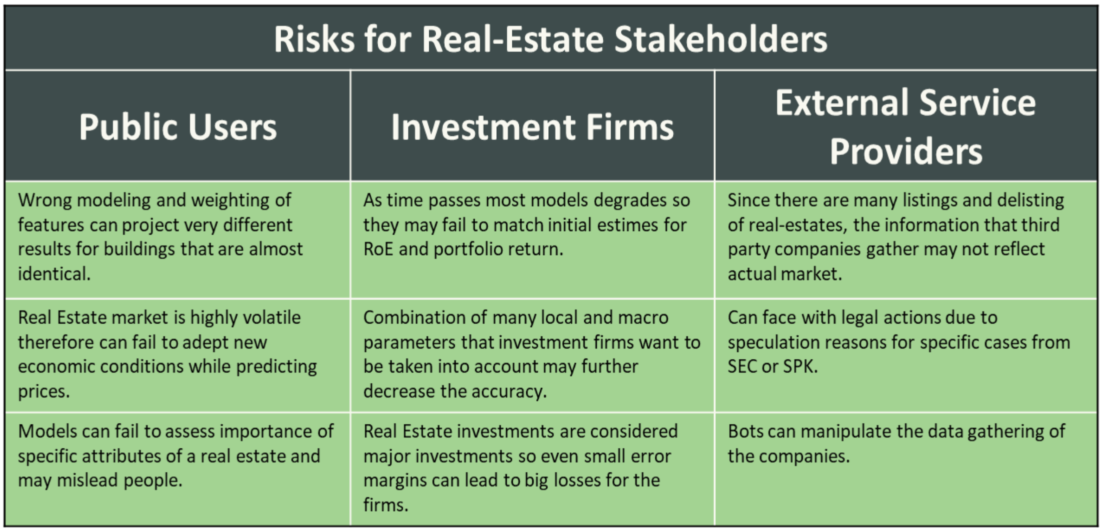
5 Input Maximization

The repairs and upgrades required for a certain real estate can be assessed for their worth by looking at market data.

6 Exit-Divest Decisions

Within high-growth real estate markets, aiding in the choice of properties to invest in or places to exit or divest. [2]

These stakeholders are also subject to certain risks due to limited accuracy of models and possible errors in the data gathering processes. These risks cannot be eliminated completely but their effects can be mitigated by careful risk assessment and mitigation plans. Some of the risks for mentioned stakeholder can be found in *Figure 2*.



Risks for Real-Estate Stakeholders		
Public Users	**Investment Firms**	**External Service Providers**
Wrong modeling and weighting of features can project very different results for buildings that are almost identical.	As time passes most models degrades so they may fail to match initial estimes for RoE and portfolio return.	Since there are many listings and delisting of real-estates, the information that third party companies gather may not reflect actual market.
Real Estate market is highly volatile therefore can fail to adept new economic conditions while predicting prices.	Combination of many local and macro parameters that investment firms want to be taken into account may further decrease the accuracy.	Can face with legal actions due to speculation reasons for specific cases from SEC or SPK.
Models can fail to assess importance of specific attributes of a real estate and may mislead people.	Real Estate investments are considered major investments so even small error margins can lead to big losses for the firms.	Bots can manipulate the data gathering of the companies.

Figure 2. Top Risks for Real-Estate Market Stakeholders

6. Feature Selection

It is not an easy task to assess the impact of each of the specific characteristics on real-estate valuations. Each of the factors that are shown in *Figure 3* is interrelated with each other and also has dependencies on domestic & global economic events and fluctuations in the financial markets. Therefore, maintaining balance and harmony between each feature to use them in specific modeling approaches can become a complex task itself. It should also be noted that the relationships are also geography, time and demography dependent hence the smallest change in any of these parameters may result in variation in the outcome. In this section, the feature selection process is going to be detailed by branching them into two main areas that are *direct factors* and *indirect factors.*

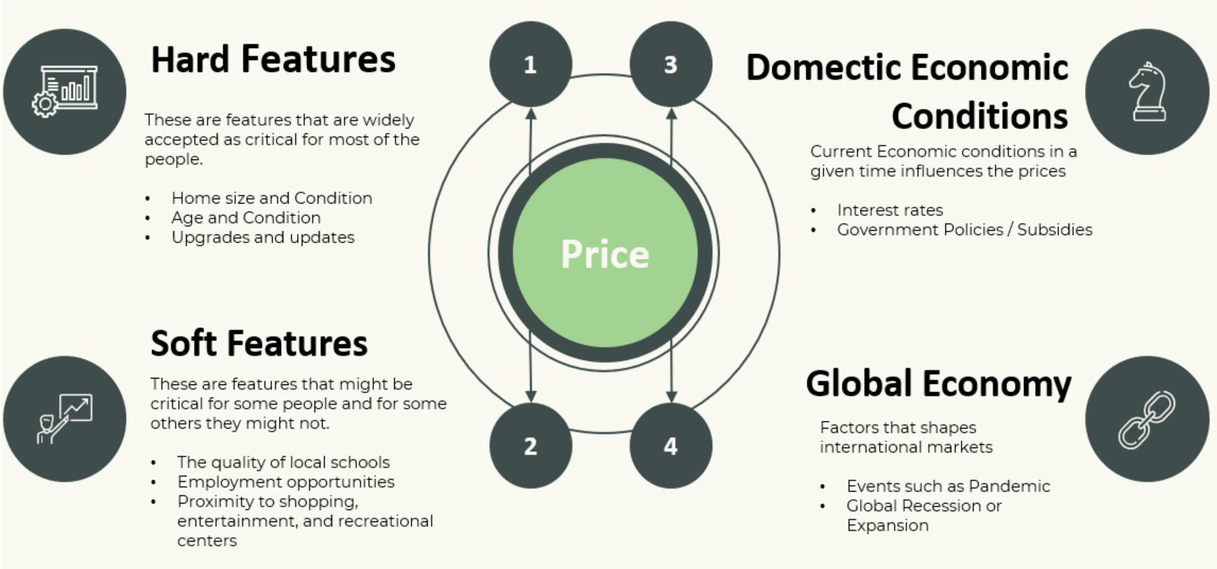

Hard Features
These are features that are widely accepted as critical for most of the people.
• Home size and Condition
• Age and Condition
• Upgrades and updates
Soft Features
These are features that might be critical for some people and for some others they might not.
• The quality of local schools
• Employment opportunities
• Proximity to shopping, entertainment, and recreational centers
1
2
3
4
Price
Domectic Economic Conditions
Current Economic conditions in a given time influences the prices
• Interest rates
• Government Policies / Subsidies
Global Economy
Factors that shapes international markets
• Events such as Pandemic
• Global Recession or Expansion

Figure 3. Factors that Influence Real-Estate Prices & Market

Indirect factors can be classified as the effect on the prices that comes from outside of the property itself. They mainly constitute economic conditions in the country, world and any major event that can shape financial markets such as the Covid-19 Pandemic. In *Figure 3* these factors are marked as categories 3 and 4. Real estate price variations are usually looked for by analyzing the following fundamental indirect factors: interest rate decrease, loan availability, housing supply and demand ratio, changes in housing market participants' expectations, administrative restrictions of supply etc. [3] The following *Figure 4* gives a general view of the indirect features. Lastly, again all of the features are dynamics and are subject to change for different time intervals.

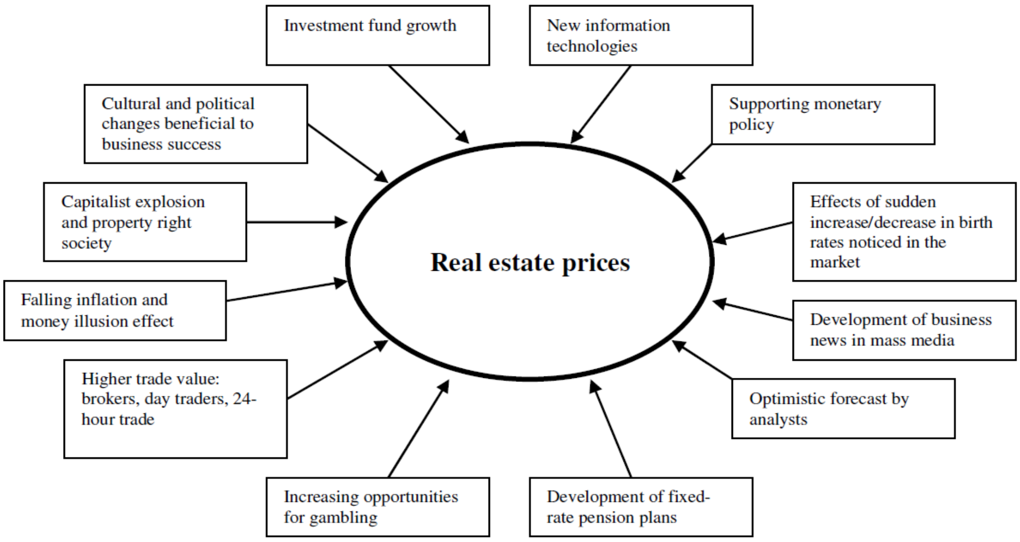

Investment fund growth
New information technologies
Cultural and political changes beneficial to business success
Supporting monetary policy
Capitalist explosion and property right society
Effects of sudden increase/decrease in birth rates noticed in the market
Real estate prices
Falling inflation and money illusion effect
Development of business news in mass media
Higher trade value: brokers, day traders, 24-hour trade
Optimistic forecast by analysts
Increasing opportunities for gambling
Development of fixed-rate pension plans

Figure 4. Fundamental Indirect Factors that Influence Real-Estate Market [3]

On the other hand, direct features are mainly related to the property itself and its current & future surroundings. *Figure 3* summarizes these factors as categories 1 and 2. These features compared to indirect features much easier to understand and model since the ranges and variations are simpler and fall within the range of human expectations. Analysis by McKinsey Company has classified these features as a traditional and nontraditional features instead of Hard & Soft features in our classification [2]. However, the following ideas are the same; hard features are mainly related to the real estate itself such as the size of a house or how many floors that a building has etc., soft features can be explained by and differs with each individual's expectation of a house. For instance, the number of coffee shops within a one-mile radius can be a price determinant for some and may not for others. As stated in *Figure 5*, while hard features still hold respectable importance in valuations, soft features are becoming much more important. This is connected to the idea that our residential areas also define our personal lifestyles and our reach in society.

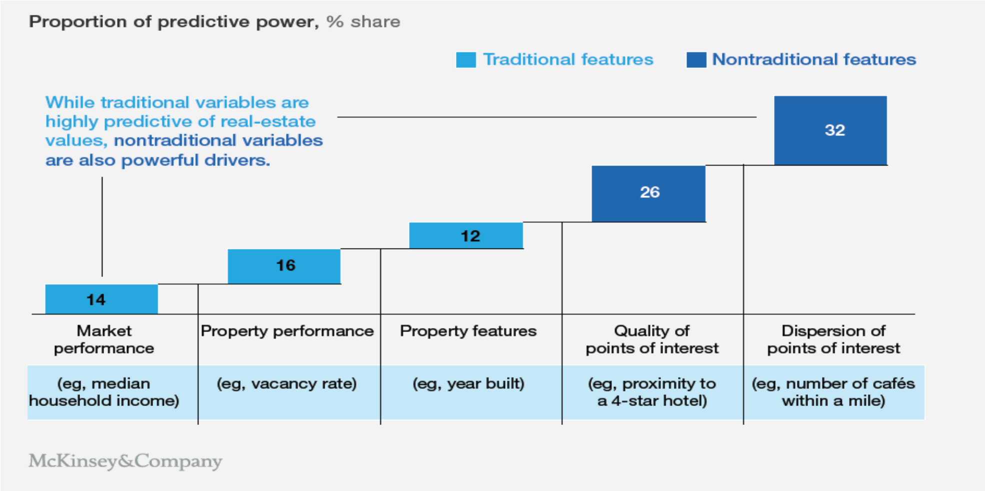
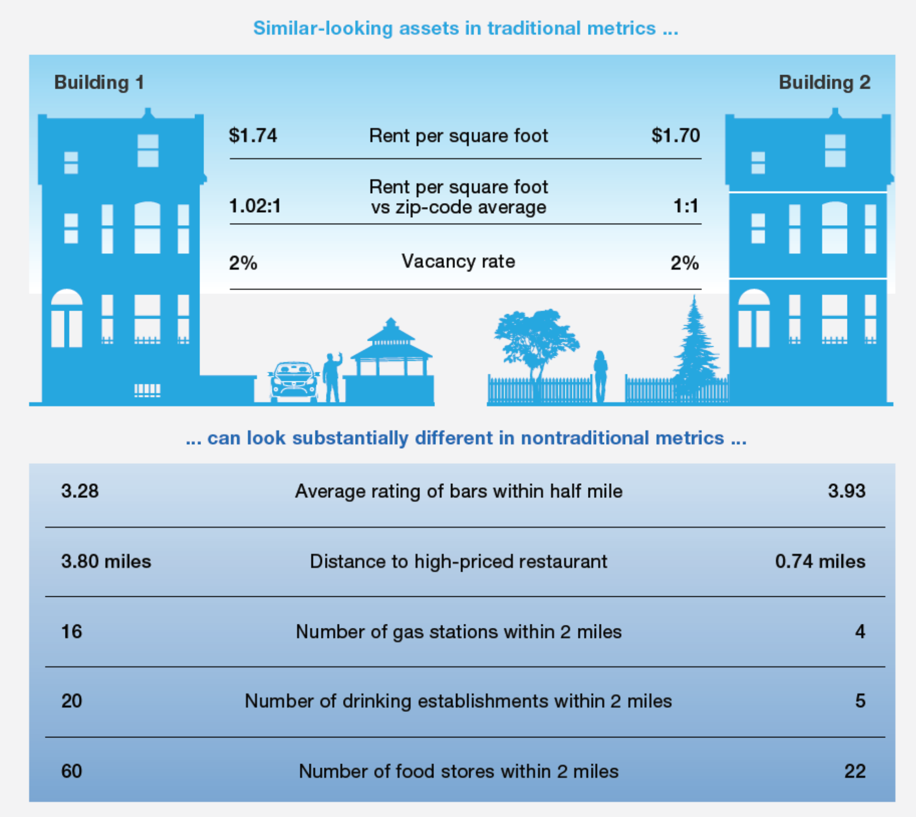


Figure 5. Contribution of Traditional & Nontraditional Features in Real-Estate Pricing [2]

The importance of soft features compared to hard features can be better visualized in *Figure* 6. The household valuations and future potential for two almost identical buildings in terms of hard features may significantly differ with the analysis of soft features. The key takeaway is that the former hard feature analysis can be easily conducted via conventional methods and brokers however the analysis of soft features requires a complex task of advanced data analytics.



Similar-looking assets in traditional metrics ...

Building 1		Building 2
$1.74	Rent per square foot	$1.70
1.02:1	Rent per square foot vs zip-code average	1:1
2%	Vacancy rate	2%

... can look substantially different in nontraditional metrics ...

3.28	Average rating of bars within half mile	3.93
3.80 miles	Distance to high-priced restaurant	0.74 miles
16	Number of gas stations within 2 miles	4
20	Number of drinking establishments within 2 miles	5
60	Number of food stores within 2 miles	22

Figure 6. Traditional vs Nontraditional Feature Analysis [2]

7. Data Sources

Technology solutions automate the data collection by accessing application programming interfaces (*APIs*) and connecting various databases before preparing the data for analysis. Examples of databases can be given as governmental and third-party software. In almost all countries, the sales records of real estate should be validated by Government Notaries. After the validation the statistics for the overall regional data become public. Alternatively, third-party applications offer middlemen solutions for buyers and sellers in the market. Hence the websites such as Sahibinden.com or Zillow have big data on real estate purchases in terms of sales date, listing, delisting, and features searched for

8. Modeling

A variety of research work has been conducted to forecast real estate prices. The approaches used for real estate price prediction can be classified as machine learning regression models and hybrid models with Neural networks. However, most of the studies do cover machine learning regression models as in the paper which will be explained next.[4]

For the research training and test samples, *data collection* was conducted via several different methods which can be referred to as data scraping. The data scraping activity is primarily dependent on the Internet sources from where data is being collected and cannot be fully automated. For example, in the case of scraping data from the website, the best format is that the developer has assigned to each unique HTML element. Hence the data is collected from public sources as explained in previous section 5 on publicly available data via browser either without login or after authentication to specific third-party websites.

Data preprocessing is performed in order to transform the raw dataset into a clean dataset ready for machine learning models. As in the research group’s case, the data is collected from different property websites where property agents entered it, so there are missing values, data in various formats, and incorrect data. Hence an integrated data cleaning i.e. *Data Wrangling* tasks were performed to turn this raw data to proper sample data such as min-max scaling, average value substitution etc. Correlation analysis between housing features i.e. Bedrooms and Bathrooms, Build, Dining Room with price feature was also calculated by the team. The example correlation matrix is provided in the *Figure 7*.

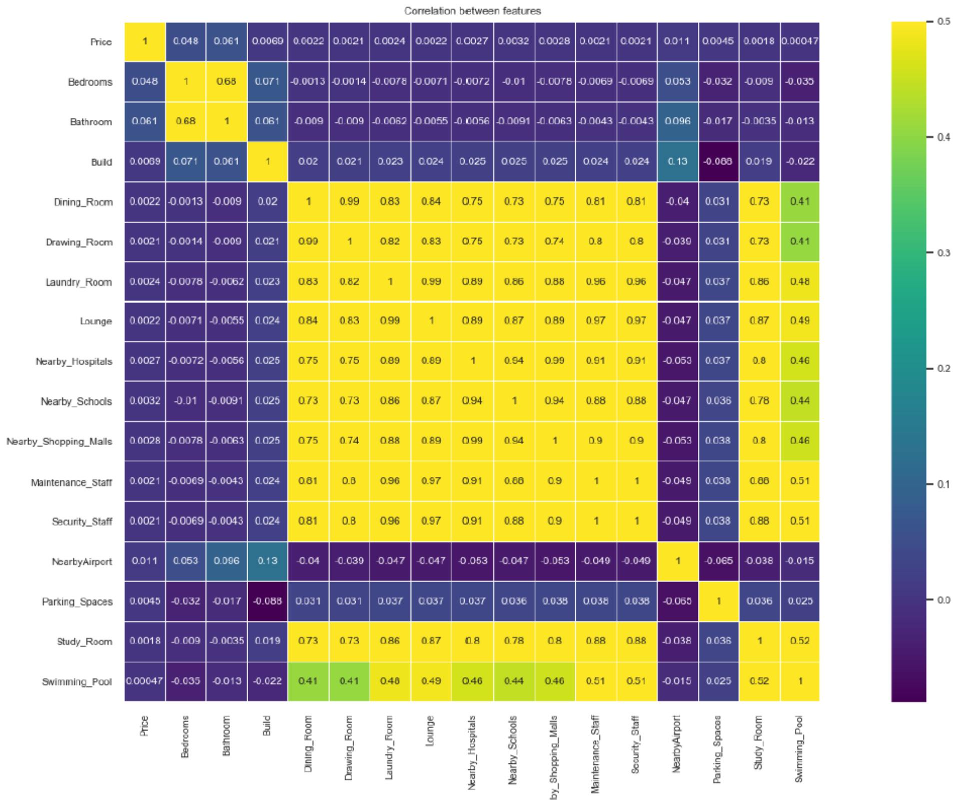


Figure 7. Correlation Matrix for House Features [4]

The research group has created *several different regression models* on a multi-variate training data set and evaluated their respective results with different error functions. Methods that were examined are explained below:

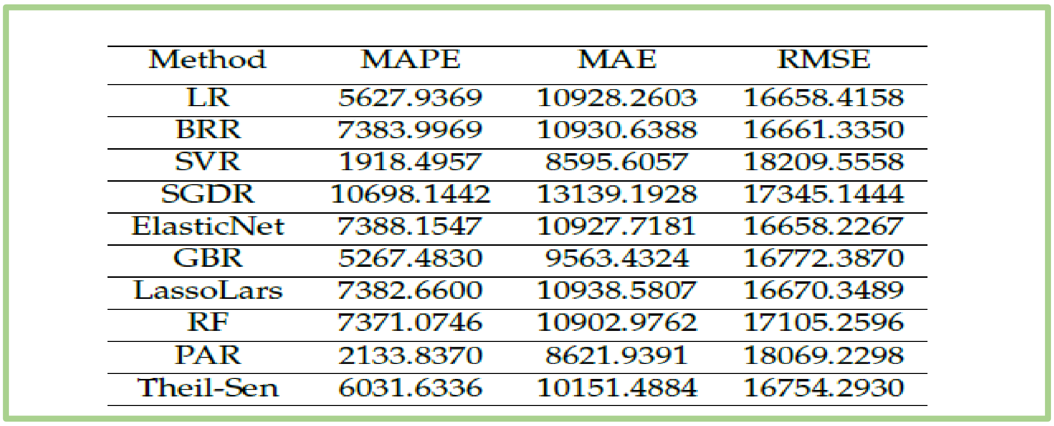
- **Linear Regression (LR)**
  - ✓ LR Model is an easy and straightforward method that will find the best possible line that fits the training set and then predicts the unseen house price from the test set.
- **Support Vector Regression (SVR)**
  - ✓ With given parameters and sample space, SVR defines a hyperplane line for predicting the continuous value or housing price value.
- **Bayesian Ridge Regression (BRR)**
  - ✓ BRR is a probabilistic model that also assumes housing prices are normally distributed. advantage of using BRR is that it can adapt to the data at hand and also that it can be used to include regularization parameters in the housing price estimation procedure
- **LassoLars Regression**
  - ✓ One of the simple techniques to reduce model complexity and prevent over-fitting, resulting from simple linear regression.
- **Gradient boosting regression (GBR)**
  - ✓ GBR repetitively leverages residual patterns and strengthens a housing price prediction model with weak predictions, and makes it better.
- **Stochastic gradient descent (SGD)**
  - ✓ Iterative method for optimizing an objective function and is mostly used as black-box optimizers.

For *performance evaluation* following error functions were utilized by the research group to evaluate the accuracy of the regression models that are MAPE (*Mean absolute percentage error*), RMSE (*Root Mean Squared Error*), and MAE (*Mean absolute error*).

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{e_t}{y_t} \right|$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} \left(\frac{d_i - f_i}{\sigma_i} \right)^2}$$

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |e_t|$$



Method	MAPE	MAE	RMSE
LR	5627.9369	10928.2603	16658.4158
BRR	7383.9969	10930.6388	16661.3350
SVR	1918.4957	8595.6057	18209.5558
SGDR	10698.1442	13139.1928	17345.1444
ElasticNet	7388.1547	10927.7181	16658.2267
GBR	5267.4830	9563.4324	16772.3870
LassoLars	7382.6600	10938.5807	16670.3489
RF	7371.0746	10902.9762	17105.2596
PAR	2133.8370	8621.9391	18069.2298
Theil-Sen	6031.6336	10151.4884	16754.2930

Figure 8. Error Functions and Outcomes of Different Regression Models [4]

In conclusion, various machine learning regression models were compared in terms of their error performance for finding the best model for a better housing price prediction. The results in *Figure 8* show that SVR performs best than the rest of the machine learning algorithms. It is observed that algorithms such as Passive-aggressive Regression, Support Vector Regression, and Deep Learning Networks can estimate the prices very close to the listing price. [4]

Combining all of the data mining tasks that are mentioned in the research paper, the resulting *Figure 9* can summarize the whole process from the data collection to accuracy evaluations which can be repeated for other algorithms, time intervals, and regions in real-estate price forecasting.

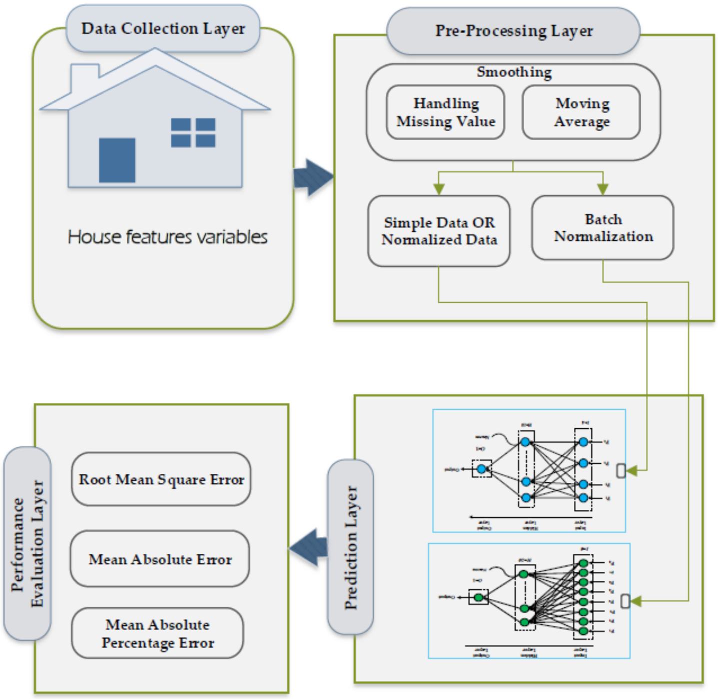


Figure 9. Real-Estate Price Prediction Data Mining Process Flow-Chart [4]

9. Case Study: Zillow & Z-estimate

To sum up all the concepts about real estate and data analytics, a real machine learning application by Zillow for real-estate price forecasting will be explained as a case study. Zillow is a leading real estate and rental marketplace in the USA dedicated to empowering consumers with data and knowledge about their neighborhoods, and connecting them with the best local professionals who can help with their real estate purchases and sales. The case starts with Z-estimate, a tool that used data sources to create an approximate value of properties by Zillow. The company wanted to take advantage of this tool via “Flipping Houses” which involves buying a property at a lower value, spending on improvements and renovations and then selling it at a higher price [5]. Although started nice, Zillow's prediction model started degrading. This resulted in the company buying properties at a much higher price than they were able to sell them for.

In the end, CEO Barton announced that [5]:

1. Zillow would stop purchasing homes – at a time when it already owned more than 7,000 overpriced houses.
2. The real estate company has decided to sell all its inventory with a discount and lay off about 25% of its 8,000 employees.
3. The company lost $420 million in 2021's third quarter.

There are many lessons learned from Zillow's endeavor which can be generalized as a key takeaway for both Zillow and future Real Estate Data Mining and Machine Learning Applications.

- Machine learning (ML) models are only as good as the input data. When the algorithm is fed with substandard data, the accuracy will also be likewise. A third-party application should not be taken as the only data source. Data preparation and cleaning processes are even more important for real estate price predictions
- There are many different features and inter-relations between features that the Z-estimate model failed to capture. The more dynamic and correlated features a model has, the more its predictive accuracy will drop.
- Time & Geography & The Demographics of the area can completely change how an algorithm should work. It should be taken into account that real estate market is volatile and highly depends on domestic and global economic conditions.
- The real-estate properties involve big monetary investments. A %10 error might lead to huge financial loss for both companies and the public. Therefore, predicting prices that come with uncertainty, it is essential to test the changes before relying on algorithms to predict the outcome. Additionally, due to the dynamic nature of the real estate market, the algorithm results should be evaluated with bigger buffer zones.
- There can be never enough testing & validation. Parameters of the model always need to adopt new conditions in a continuous manner with real-time data.

10. California housing Dataset & ML Applications

California Housing Dataset has been published US Census Bureau which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The ML Applications attempted here aims to predict median house values in California using the provided dataset. Different types Regression models should learn from the dataset and be able to predict the median housing price in any district, given all the other metrics.

California Housing Dataset is an open-source dataset. It can be downloaded from US Census Bureau website or popular data science websites such as Kaggle with the link https://www.kaggle.com/datasets/shibumohapatra/house-price. [6]. In the dataset, We have 10 metrics including the target variable *median_house_value*. A detailed explanation of each variable is provided below:

Longitude-Latitude: geographic location information are useful to understand if any of the regions are much more valuable than the other sections. This will be analyzed further during EDA.

Age: House age can be important estimator and expected to negatively correlated with the price. This will be analyzed further with correlation analysis.

Household & Total rooms & Total bedrooms & Population features represent total values for each district. We will be combining these to represent unique district-related features such as pop/household. These will be more representative of the given district. Also, it should be noted that the total number of bedrooms feature had missing values that will be imputed with the median value for each district.

Median income: District median income may be the most important feature that can affect house prices. It is scaled between 0.5 and 15.0 in the preprocessed dataset. The original scale value will be further transformed in the preprocessing steps.

Ocean proximity: is a categorical feature that will be categorically encoded for each unique value in order to feed the feature into our ML models

The source code notebook provided has the following ML Applications with their implementations, fine-tuning, comparisons and detailed analysis:

- ✓ Ordinary Least Squares Regression (OLS)
- ✓ Ridge Regressin
- ✓ Lasso Regression
- ✓ Decision Tree Regressor
- ✓ Random Forest Regressor
- ✓ KNN Regressor
- ✓ SVM Regressor
- ✓ SKlearn MLP Regressor
- ✓ Keras MLP Regressor

10. References

[1] “Real Estate Market Size Report 2020/21.” *MSCI*, https://www.msci.com/www/research-paper/real-estate-market-size-report/02648017490.

[2] Asaftei, Gabriel Morgan, et al. “Getting Ahead of the Market: How Big Data Is Transforming Real Estate.” *McKinsey & Company*, McKinsey & Company, 30 Mar. 2021, https://www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate.

[3] “(PDF) Models of Factors Influencing the Real Estate Price.” *ResearchGate*, https://www.researchgate.net/publication/266878986_Models_of_factors_influencing_the_real_estate_price.

[4] “(PDF) Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data.” *ResearchGate*, https://www.researchgate.net/publication/353371025_Using_Machine_Learning_Algorithms_for_Housing_Price_Prediction_The_Case_of_Islamabad_Housing_Data.

[5] “Zillow's Great Data Science Disaster.” *Analytics India Magazine*, 20 Dec. 2021, https://analyticsindiamag.com/zillows-great-data-science-disaster/.

[6] Kaggle, California Housing Dataset,
https://www.kaggle.com/datasets/shibumohapatra/house-price