Real Estate Market Data Analysis and Prediction Based on Minor Advertisements Data and Locations' Geo-codes

2 authors:

Rashed Al-Hamadin
Central Bank of Jordan
3 PUBLICATIONS **1** CITATION

SEE PROFILE

Waleed Al-Sit
Mu'tah University
5 PUBLICATIONS **11** CITATIONS

SEE PROFILE

Real Estate Market Data Analysis and Prediction Based on Minor Advertisements Data and Locations' Geo-codes

Waleed T. Al-Sit [1], Rashed Al-Hamadin [2]

[1] Department of Computer Engineering, Mu'tah University. Al-karak, Jordan , w_sitt@hotmail.com
[2] Princess Sumaya University for Technology, Jordan, rashedhamadin@gmail.com

ABSTRACT

Country's economic status can derived from many complex and branched indicators, one of which is property prices estimation. Working on such indicator changed the state of literature from many perspectives and corners. Whilst the scarcity of such works imposes a need for it, and demonstrates an unutilized aspect of the economy that requires little resources to create some business and academic opportunities. In this work, efforts evolved to address the problem of estimating properties prices accurately, in specific apartment's prices among The Amman City, The Capital of The Hashemite Kingdom of Jordan. Leading to shed the lights on employing data science different techniques namely data processing, analysis and predictive modeling for adopting and estimating the apartment's prices based on advertisement data published through the web and its extracted location geocodes. In addition, the work evaluates the final analysis reported results based on selected evaluation measures, and compare them with other five similar works on such problem conducted in other countries. Trying to aim to enrich the literature with valuable insights gained using Machine Learning and Data Mining different predictive techniques mainly, and its related conditions, branches and requirements for other data processing and analysis techniques under the data science umbrella.

Key words: Deep Learning, Machine Learning, Predictive Modeling, Property Prices Analysis Prediction.

1. INTRODUCTION

Property prices; a great indication of a country's economic status. It gives an indication of the economic wellbeing and stability, and being considered a clear manifestation of inflation amongst other economic phenomena [1]. Usually, this field observed to be either the original sin and/or the most prominent presentation of most economic events, and it considered being so complex and branched sector.

Most of economy aspects can be defined or described in terms of property sector, as it is in combination with food are of the oldest sectors to have emerged in correspondence to sheer demand imposed by the basic needs for survival for modern humans [2]. Though other species do have, housing needs that are not necessarily simple as represented with the case of beavers, and their well-engineered dams. Rather, humans took it several steps ahead, and delegated the housing and construction to form the earliest jobs and occupations of builders. In addition, it further expanded with economic evolution with further branching and breaking down of roles to include most of early occupations and industries. For instance, dominating engineering until the early days of the industrial revolution expanded to span over other sectors and industries beside construction, which did not stop the property industry from being a major player in the economic scene until now [3]. Furthermore, it displayed its dominance with the emergence of economies of scale, banking economies and loan based economies to a great degree as shown by the financial crisis in 2008 [4]. Being of this importance; i.e. property industry, lead to trigger a lot of work on the topic as demonstrated by the plethora of scientific works and publications in the field. Leading to form a solid base of literature around the topic especially in the financial aspect [5]. Such an example and original motivation behind this work is the property price prediction, having set itself with the iris dataset related work as the default examples of the trendy and important field of data science.

In this view, and due to the interesting attributes of the property problem. Alongside, how it corresponds to the interrelations and different macroeconomic components [6]. Furthermore, the rich nature of data involved in house price estimation and prediction, and the semi obvious nature of the factors involved of property prices, and the existence of previous more traditional methods for the estimation process [7]; all makes for a great use case for studies, serving both business related studies and academic applications. Because it lead related parties to reveals the hidden aspects of a country's economy, whilst also demonstrating clearly the effects of data science and predictive analysis and data manipulation

Table 1: Similar Works Summaries

Inputs				Results		
Dataset	Dataset Dimensions	Used Models	Data Splits	Model	Evaluation Criteria	Value
UCI's Dataset (Housing Value of Boston Suburb)	452 X 13	SVM (1), LSSVM (2), PLS (3)	400-52	(1)	MSE	10.7373
					R. Time	0.4610s
				(2)	MSE	20.3730
					R. Time	20.3730s
				(3)	MSE	25.0540
					R. Time	0.7460s
Local Data in Spain	1187 X 6	MLP [1 Hidden Layer]	952-237	MLP	R2	0. 8605
					RMSE	39540.36
					MAE	28551.34
"Zillow.com", "magicbricks.com"	21,000 X 15	Linear Regression (1), Multivariate Regression (2), Polynomial Regression (3)	80% - 20%	(1)	RMSE	1.201918
				(2)	RMSE	16,545,470
				(3)	RMSE	11359157
bProperty.com	3505 X 15	GB-Regression (1), Random Forest (2), SVM Ensemble (3)	80% - 20%	All	RMSE	0.1864 to 0.2340
random sample from www.bluebook.co.nz	200 X 7	MLP	80% - 20%	MLP	R2	0.6907 to 0.9
					RMSE	449, 111.46 to 1, 014,721.92

techniques due to the straightforwardness of the problem and the size of the available work on it [8]. Starting with the benefits yielded of an accurate estimation for property buyers, alongside the assumption related to if a certain property priced fairly, and to have a better idea about the possible impacts of each of the attributes on the price and in what way. All are leading to facilitate the process of making a purchase

decision and budget and priorities setting for them. In addition, helping the property investors in knowing if a purchase deal is a bargain with a high margin of profit or not. On the other side of the same transaction type for describing the property sales process, we have property sellers and potential investors in the property sector whether they are individuals or corporate sized parties. Who are looking forward to understand the market prices, and how much they expect to charge for their properties, and what aspects related to the market in order to concentrate on, and what to dismiss, all are a common use case for data science some techniques.

The contributions of the work presented in here are threefold: (1) demonstrates utilizing data science to figure more countries' economic aspects, which requires huge resources to create some business and academic opportunities and insights. (2) Presenting a comparative predictive modeling and analysis study for Jordanian property market against five similar works conducted in different countries. (3) Finally, reporting the analysis resulted insights upon adopting different machine learning techniques and evaluation measures for the presented problem.

The rest of the work organized as follows. In section 2, presenting similar works for other countries and their reported results. Section 3 provides the methodology adopted to conduct end-to-end data preprocessing and Explanatory Data Analysis (EDA) considering real estate apartments prices prediction based on advertisement data and locations geocodes. Section 4

provides the predictive modeling analysis different experiments conducted using machine learning. In Section 5, the work provides the reported results alongside the needed discussions, before concluding the presented work in section 6 and its future next steps**.**

2. SIMILAR WORK

As a heavily covered topic especially in the recent years, the options to compare to and benefit from are plenty. Hence, we chose a sample of few papers that addressed the price estimation problem that felt closest and most relevant to this work.

Starting with hedonic regression, where we performed several regression to each attribute individually. Then, trying to observe the change in a target attribute (Price), so constructing the final equation of the predicted variable of the weighted estimation attributes. With an origin out of the real estate, pricing [9] and some interesting use cases [10]. A lot of literature on this problem is studying the alternative methods for the estimation with a concentration on the use of machine learning. Table.1 summarizes five works followed with a discussion for each further more in this comparative work.

In the first work [11], using a UCI dataset with housing prices data in Boston, and 13 attributes in the data with a relatively low number of instance at 452. Authors presented their work as comparison of three different models; namely Partial Least Squares (PLS) regression, Support Vector Machine (SVM), and Least-Squares Support-Vector Machine (LSSVM), where the third is somewhat a kernelled version of regular SVMs with an optimization included by design. The results shows that SVMs where superior to the other two methods both on Mean Squared Error (MSE) and fitting time of the model, though LSSVM would have smaller fitting time if the

Table 2: Web Scraped Data Attributes

Attributes	Type	Attributes ...	Type ...	Attributes ...	Type ...	Attributes ...	Type ...
ID	INT	Ad Images Count	INT	Air Conditioning	BIN	Nearby Facilities	BIN
Title	STR	City	STR	Heating	BIN	Security	BIN
Date	STR	Location	STR	Balcony	BIN	Built-in Wardrobes	BIN
Real Estate	BIN	No. Rooms	INT	Elevator	BIN	Swimming Pool	BIN
Paid Ad Feature.1	BIN	No. Bath Rooms	INT	Garden	BIN	Solar Panels	BIN
Paid Ad Feature.2	BIN	Area	INT	Garage Parking	BIN	Double pane Windows	BIN
Paid Ad Feature.3	BIN	Floor	INT	Maid Room	BIN	Ad Post	STR
Paid Ad Feature.4	BIN	Age	STR	Laundry Room	BIN		
Price	INT	Payment Type	STR	Is Furnished	BIN		

parameter optimization were dropped, while all models seems to yield acceptable results.

The next work presented in [12] provides a good outlook on the Spanish real estate market. It utilizes data that spans over a long time interval by making use of a one hidden layer Multi-Layer Perceptron (MLP), in order to yield an estimation model of the real estate price based on some exogenous variables. Authors claiming the superiority of one hidden layer over two, where the results obtained are to some degree a supporter of the claim. However, further examination of such work should tested, considering the various architectures and parameters that an MLP could take. While also taking into account the benefits of the two hidden layers models [13] beyond the general function approximation abilities that artificial neural networks have [14]. with a somewhat rich data on vertically with 1187 instance the data is slim horizontally with only 6 attributes, demonstrating with the calculated R2, RMSE and Median Absolute Error (MAE) the potential Artificial Neural Networks (ANNs) have for solving such problems and improvement over traditional hedonic models which is an observation that is shared between a good fraction of literature.

Coming more on the technical side. Authors in [15] makes use of rich data with 21000 instances and 15 attributes and tries several regression models, namely linear, multivariate and polynomial regression. However, the evaluation method leaves a lot to be desired, going with RMSE solely leaves a need for data exploration to understand the results further more. But the use of such evaluation criterion could be understandable due to the nature of the experiment where most regression models rely on minimizing some sort of error or residual to produce the model and with the iterative nature of the tuning both the models and the data.

Considering the ensemble way, authors in [16] based on the assumption that several models should yield better results. As in the aggregated results of several models should present more support to a certain decision. This work utilizes several models that has shown good results, exceeding the performance of other models that do not utilize the ensemble paradigm like ANNs. With 3505 instances in the data set and 19 attributes, the dataset considered is features rich and should allow for better estimation. As it reduces, the complexity of the fitted model due to the higher dimensionality that is at the same time might hinder the fitting and learning process due to the larger search space for the solution. The results is good yet could benefit from a clearer presentation of the results, this while lacking in clarity demonstrated the potential benefit of ensemble methods for this type of regression. In addition, the data preprocessing and transformation effect on the regression's final output, and this with the respect to the data categories and how each might effect on the learning process in terms of model and training/fitting performance, while elaborative on the effects of tuning machine-learning ensembles with parameters such as depth and number of estimators [17].

Following in the trend of neural models for price estimation, while going further with the hedonic versus ANN house price estimation. Where they share a similar definition of the problem addressed to solve. Authors in [18] dives deeper into both the hedonic and artificial neural network theories. Alongside the histories and the inflection of the aforementioned theories on the corresponding models. Rather, scarce in the data used with 200 instances and 7 attributes. Nevertheless, the models produced varies highly in prediction performance. That is rather obvious in terms of R-squared, which it clearly explained in the different architectures of the used neural models in terms of number of neurons. However, the scarcity of data raises some questions regarding the top obtained performance, that explained by the more complex model producing it. In addition, the number of test instance indicates less support for the results. Nonetheless the statistical nature of this work and the emphasis on the hedonic-neural comparison and attributes contribution. The analysis conducted gives an indication, that these results are just

indicative of the statistical conclusions of a potential superiority for the neural models over hedonic ones, in the real estate price regression demonstrated with the use of measures like White Heteroscedasticity Test and confidence intervals. While also showing that, even aggregated features may lose some information, but are still a feasible option if fitting performance is of an issue [19]. An interesting observation here would the different contributions to each attribute and its importance level for the regression in the two models presenting the low interpretability of them in terms of their produced model represented by patterns seen in data to produce the regression or classification shall not always comply with the higher level semantics observed by humans [20].

Most of literature including the discussed examples above use an 80% and 20% for the training and test data splits. In addition, the most addressed the hedonic regression as a common method for such a problem and use case. while also describing it as the method to be replaced by machine learning, due to its more dynamic and complex nature [21], [22]. Which allow capturing more intricate trends and adapting more through time relying on online machine learning methods [23]. In the meantime, facilitating the time based adaptation and therefore performing better for economic analysis [24], either presenting data over a temporal axis implicitly or explicitly, which will accommodated by some machine learning models in later stages of analysis. For instance, Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs) [25], with some future anticipated and currently demonstrated cases of image based price estimations [26]. As much other type of estimators in different areas that rely on machine learning [41]-[43].

3. JORDANIAN MARKET: REAL ESTATE APARTMENTS PRICES PREDICTION

3.1 Methodology

Starting from acquiring the data from online sources as no data is available for the Jordanian real estate market, as most data with these specs mostly aggregated from many sources. For instance, adopting web-scraping approach. Then, this work proceeds to dealing with data issues, though the actual workflow was more iterative than sequential it will listed latter for convenience. After handling the data work, this work continues to the predictive modelling. Where, several experiments conducted using several models, namely Linear and Gradient Boosting for simple and complex regression-based models. Support Vector Machines and Random Forests, which are both ensembles and rely on the use of decision trees [27], [28], as their weaker models to build the ensembles. Finally, MLP as advanced predictive modeling technique. Each of the aforementioned predictive modeling techniques has its own unique features and their fitting use cases. Finally, these models are then fitted with the prepared data after manipulation then tuned if necessary to yield acceptable results.
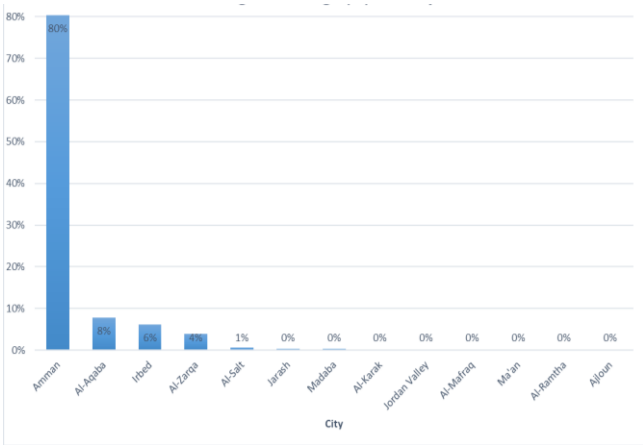

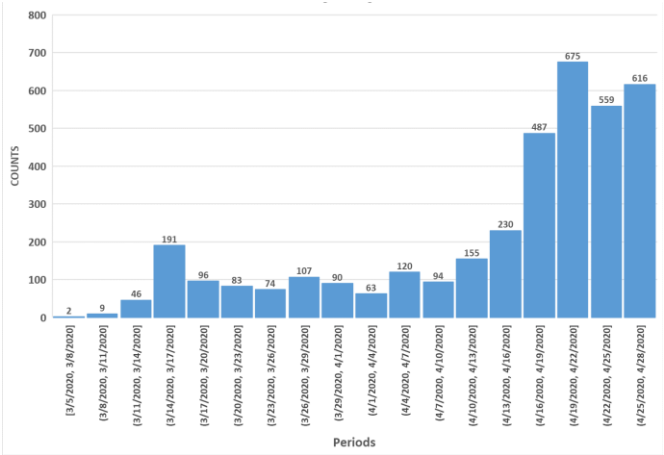**Figure 1:** Advertising Percentage (%) per City.

Figure 2: No. of Advertisements Listed per Advertisements' Dates.
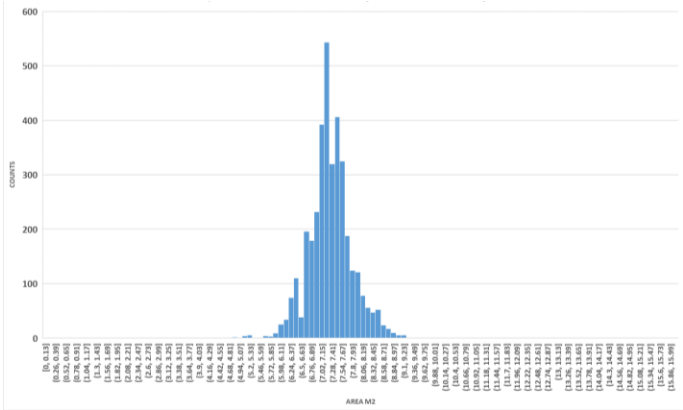
3.2 Data Attributes

Figure 3: Apartment Area in M^2 – The Whole Data (Before applying splitting and outlier's removal).

The work analysis dataset obtained from an online sales websites that include listings for apartment for sales in Jordan, which is the category this work is going to use. The original data contained 3697 instances with 34 attributes ranging in type and benefit to the apartment price estimation process. Several manipulation carried over the data in-order to reach a

state where models are able to capture the trend in the data in order to produce a sane estimation. Then, exploring the data visually to understand the nature of the data, also it is a necessary for some preprocessing steps, for instance, discovering outliers in order to remove and the data distribution to better understand the results obtained from each model. Rather, the visualization for the data can provide valuable insights about the Jordanian real estate market from with the visual aids as they can substitute even partially for the more defined metrics of the interactions between the attributes of the data [29] including price.

As mentioned earlier, the obtained data set contains 34 attributes, and usage to those attributes that listed in Table.2 can varies to conduct different types of analysis other than estimating real estate prices. However, the data instances collected are for advertisements related to apartments in specific in overall Jordan for a short period (5, March 2020 to 28, April 2020).

3.3 Data Cleaning

Some essential steps are to be done before any further process should take place to accommodate for the tools to be used nature and to reduce the bias and errors that may be perceived
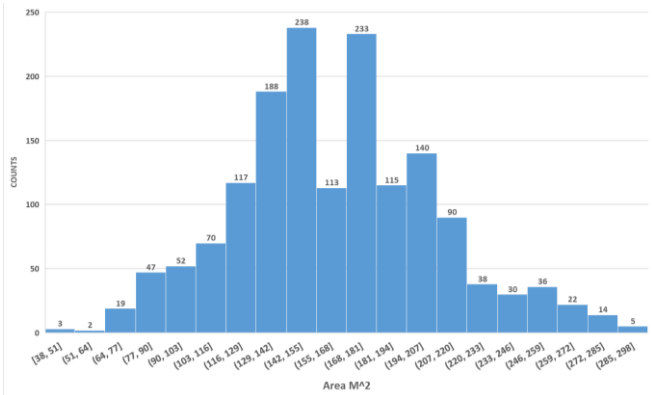
in the final outcomes of the data whether it being predictive models or simple statistics. Such steps varies from dropping NULLs in the main anticipated features to dealing with inconsistent values, duplicates and obvious noise. In addition, to fix data types, correcting, and unifying the values of attributes as the data collected from online sources that are open for contribution from any seller and non-seller parties, who wants to make an apartment listing. due to the fair number of instances in the data, and the assuming that the distribution and coverage of the data is to be preserved due to size a lot of messy and unclean data were dropped, that will also be contributing to the speed of the fitting and learning process for the predictive modelling part.

3.4 Data Preprocessing

Keeping in mind that the data should be too clean and perfect in order to leave room for the models to generalize over unseen data. Several preprocessing steps done, each serving some purpose, overlapping with preprocessing some transformations applied to the features serving purposes.

Figure 4: Apartment Area in M^2 – The Training Data Split (After applying outlier's removal).
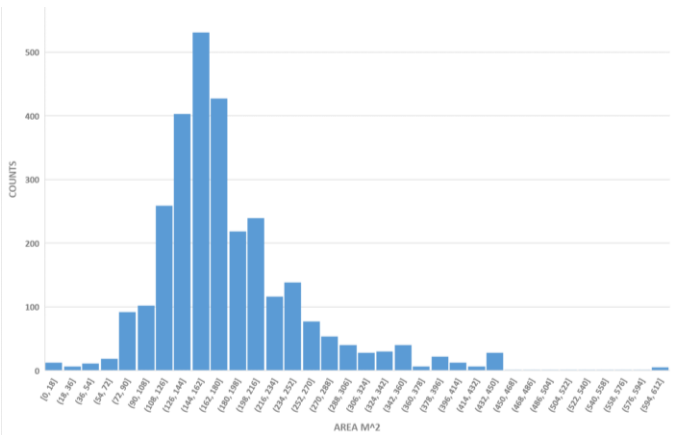
Figure 6: Average of Listed Apartments' Area in M^2

Figure 5: Apartment Area in M^2 – The Test Data Split (After applying outlier's removal).
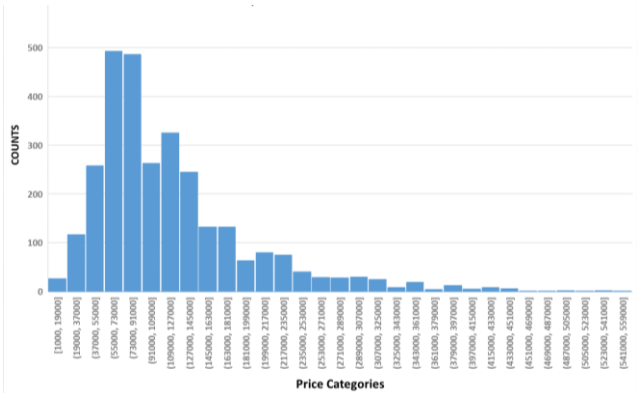

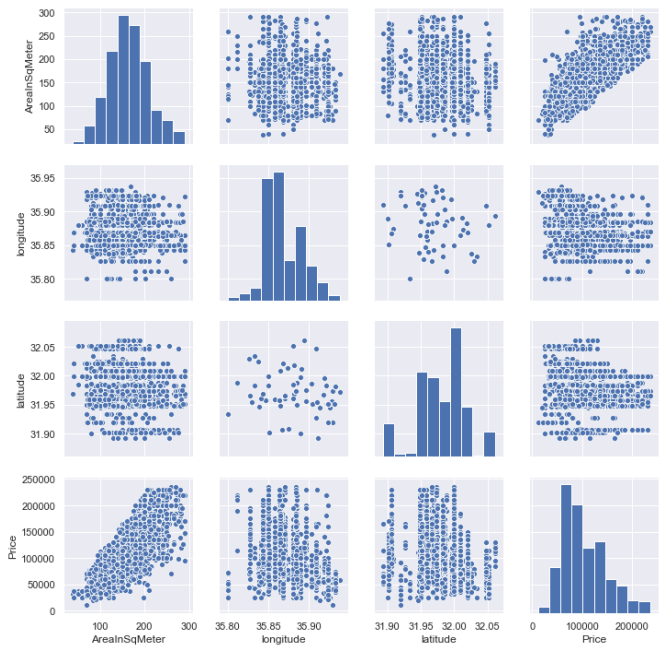**Figure 7:** Apartments Prices Distribution

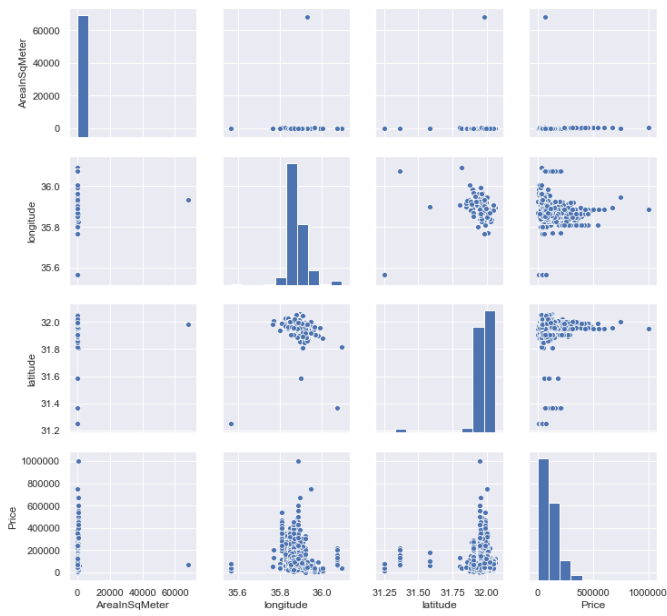**Figure 8:** Price and Area Regression Relationship.

Figure 9: Discovering Outliers using the Correlation Matrix.

some insights about the listings when coupled with other attributes indicating what makes a listing stays for long without it ever sold. For instance, Figure.6 shows the number of advertisements listed per day for the data collection period. Such result would communicate some indicators for economists about when such market movement for demands and offers may appear during the different seasons the study and regulates [30]. Such attribute transformed from its described form as a categorical and a date variable to a continuous variable representing the days since the listing was posted, a similar transformation was
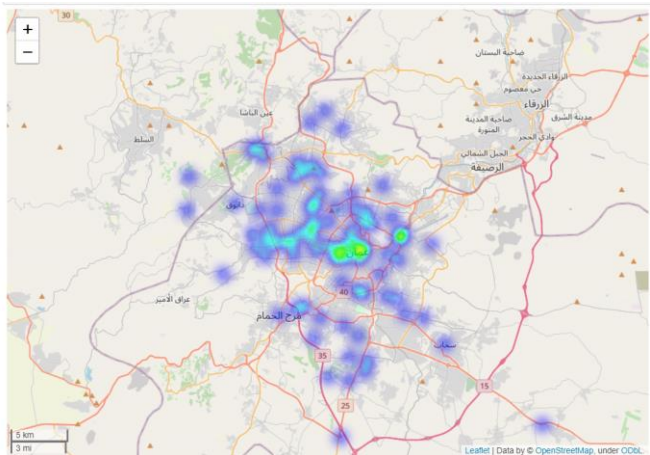
Figure 10: Graphical Heat Maps before Outliers Removal.
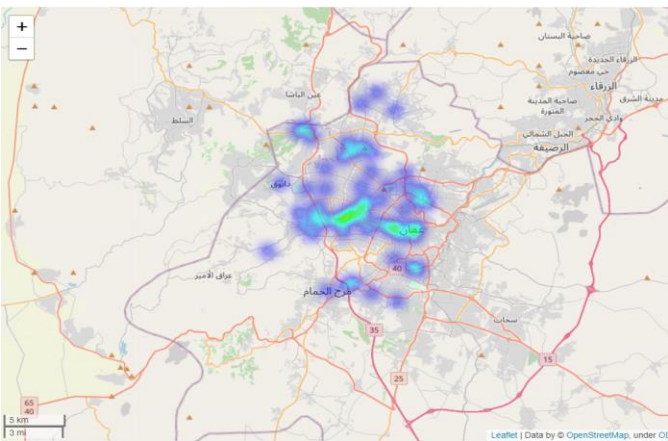
Figure 11: Graphical Heat Maps before Outliers Removal.

Table 3: Regression Models Reported Model-Based Score Results

Trail	Dataset Type	Models Generated	Results	
			Train	Test
Linear Regression	Regular Data, Scaled Data	1	0.67	0.71
Ensemble Method – GB-Regression	Regular Data, Scaled Data	1	0.97	0.81

Table 4: Random Forest Best Estimator Searching Techniques

Trail	Dataset Type	Results	
		Train	**Test**
RF-Standard	Regular	0.92	0.82
RF-Randomized Search	Regular	0.92	0.8
Three Types Used : Standard [12], Grid [16], and Randomized [17]			

Table 5: SVM Reported Results

Trail	Dataset Type	Results	
		Train	**Test**
SVM	Regular	-0.05	-0.04
SVM-Normalized (y)	Scaled	0.9	0.800.72
SVM-Randomized Search	Scaled	0.84	0.77

applied to the Building age attribute although on a smaller interval.

- **Creating Price per Meter Attribute:** To be able to fairly judge the listings for further steps like outlier removal, price and area might not be enough as in the case of large apartment with a matching price tag and a small apartment with a higher than actual value price.
- **Binary Attributes Creation:** Dealing with attributes that still include categorical values is not ideal especially for the mostly continuous and numeric implementations of machine learning models. So, and in order to facilitate the use of such attributes (Is Furnished and Payment Type). Each were transformed into two attributes that each can take 0 or 1 allowing for 4 permutations and preserving the other values of the attributes like the 'not specified' or 'both', while also being easier to use in computing processes due to its numeric nature.
- **Data Splits:** The data divided with a ratio of 70\% to 30\% for training and test sets respectively; this done before further processing and exploring data to reduce bias in the fitting and learning processes, hence, avoid skewing the prediction results. The splitting process resulted with the 2044 records for training dataset and 877 records for test dataset considering the aforementioned preprocessing steps.
- **Outlier Removal Using Quartiles:** Outliers removed out of necessity based on the distribution observed by the Figures 3-5. This might hinder the performance of some machine

Table 6: MLP Reported Results

Modeling Mechanism Used	Models Generated	Results	
		Train	**Test**
L-BFGS (1 Hidden Layer 9 Neurons)	1	0.76	0.71
ADAM (1 Hidden Layer 100 Neurons)	1	0.71	0.71
L-BFGS (1 Hidden Layer 120 Neurons)	40	0.76	0.7
ADAM (1 Hidden Layer 10-200 Neurons)	20	0.66	0.67
ADAM (2 Hidden Layers 10-200 X 10-200 Neurons)	134	0.93	0.67

learning models, outliers were detected and removed using 5 attributes (Price ,Area ,Price Per Meter, Longitude And Latitude) using the Interquartile Rule for Outliers, while the z-score method didn't yield as good results.

The resulting dataset after the preprocessing is smaller in size (2253 X 26) and uniform in distribution (Train: 1572 X 26, Test: 681 X 26), though the data is still noisy, the early predictive models fitting results indicate an enhancement in comparison to fits that conducted after. The other attributes dropped since not relevancy to the prices predictive analysis conducted. Alongside this type of trimming using quartiles, another type of trimming using z-score were experimented and yields a less satisfactory results. I.e. it does not solve the outlier's problem because only one record trimmed when using Area in square meter feature. Therefore, we stick with quartiles method.

Table 7: Evaluation Measures

Measure	Usage	Best Value	Worst Value
R-squared (R2)	Represents proportion of variance of y that explained by independent attributes in the model. Which indicate strength and goodness of fit of the data to the regression line.	100%, the model explain all the variability of the response data around the mean.	Lower and Negative Values, i.e. the model explain nothing about the response data variability around the mean.
Mean Squared Error (MSE) [38]	Risk metric based on the expected value of the squared error or loss.	Lower values as there would be excellent match between the actual and predicted dataset.	Higher values in addition to the lower values with no excellent matching between the actual and the predicted dataset.
Median Absolute Error (MdAE) [39]	1. Outliers robust (unaffected by values at the tails). 2. Error or loss function. 3. Calculate univariate variability.	Lower values as there would be excellent match between the actual and predicted dataset.	Higher values in addition to the lower values with no excellent matching between the actual and the predicted dataset.
Mean Absolute Error (MeAE) [40]	1. Risk metric. 2. Scale-Dependent Accuracy Measure.	Lower values as there would be excellent match between the actual and predicted dataset.	Higher values in addition to the lower values with no excellent matching between the actual and the predicted dataset.

Table 8: Experimental Environment HW Specifications

Item	Description
PC Type	Laptop
Brand	ASUS ROG G703GX Notebook
CPU	Intel i9-8950HK
RAM	64GB
HD	3x1TB NVMe SSD RAID
GPU	NVIDIA GeForce RTX 2080 8GB
Screen	17.3" FHD 144Hz 3ms G-Sync
OS	Windows 10 Pro

Table 10: MLP Models' Average Evaluation Scores

Data Split Used	ADAM … (1)	L-BFGS … (1)
R2-Train	0.895829998	0.763841224
R2-Test	0.674536562	0.705861926
MSE-Train	203345745.1	460995310.1
MSE-Test	662487173.3	598723783.7
MeAE-Train	6596.689703	12485.69136
MeAE-Test	14093.88889	14651.50066
MdAE-Train	9198.765418	16161.98873
MdAE-Test	18868.66751	18622.71286
Adam = ($\sum$ Score)=155; L-BFGS = ($\sum$ Score)=41 … (1)		

3.5 Visualization and EDA

The main difficulty in here is not coming up with ideas to test and evaluate on the dataset; it is coming up with ideas that are likely to turn into insights and valuable indicators about the trends and patterns hidden in the data [31]. In specific, those related to the property business. Accordingly, we have to ask questions that lead to deep understanding for the observations collected from the online different sources. Therefore, implementing the concept of Explanatory Data Analysis [29], which is a graphical analysis technique, that employs a variety of techniques in order to maximize insights, uncover the underlying patterns, important features extraction; detect anomalies and outliers, and further determining the optimal tuned model settings.

Adopting EDA mechanism for discovering normal distributions for the attributes, or finding the relationships between the attributes of different types, namely categorical and/or continuous, will better lead to find valuable insights in the data. For instance, the average area in square meter for the listed apartments in Amman city was around 150 M2 as appears in Figure.6. Whereas the average prices for the apartments appear to be around 74,647$ as appears in Figure.7.

From another perspective mostly related to outliers detection

4. PREDICTIVE MODELLING EXPERIMENTS

4.1 Scaling the Attributes

Non-binary attributes were scaled using z-score normalization to reduce variance and reduce peak values effects by reducing spread allowing some models to fit the data better [32]. The predictor variable, the price in this case was also scaled to account for the use of distance in SVMs [33], using the z-score normalization as well, as was discovered through the experiments.

### 4.2 Learning Models Employed and Experiments Setup

Several types of machine learning models used and employed in the comparative process, whilst they all implemented and evaluated according to SCIKIT Learn Package [34] provided functionalities on the basis of scaled data and regular data, where the late type means no scaling or transformation on the data sets used before learning processes.

Firstly, one of the simplest machine learning models used in addition, evaluated in this work was Linear Regression, which better suited for linear data, as the degree of the regressor is limited to one. It implemented under several assumptions as the mentioned data linearity, independence and several others,

Table 9: MLP Top-10 R2-Ranked Models

MLP ID	MLP solver	Layers Structure		R2		MSE		MdAE		MeAE	
#	Solver	Layer1	Layer2	Train	Test	Train	Test	Train	Test	Train	Test
168	LBFGS	7	0	0.74	0.7	510320333	603349036	13284.7	14313.1	17158.3	18909.7
155	LBFGS	6	0	0.74	0.72	512527565	571225948	13098.8	14093.8	16984	18154.8
135	LBFGS	5	0	0.74	0.72	515600513	576319329	13741.5	15218.9	17261.6	18609.6
136	LBFGS	5	0	0.73	0.72	521154916	574018441	13071.1	14527.5	17213.6	18367.5
85	ADAM	180	0	0.73	0.73	522058827	554305503	13099.4	13982.5	17178.6	17895.2
87	ADAM	190	0	0.73	0.72	530135262	563064858	12943.2	13909.9	17294.4	17992.9
92	ADAM	200	0	0.73	0.72	530948320	561928876	13009.8	13994.8	17334.2	17987.2
81	ADAM	160	0	0.73	0.72	535067377	572433080	13245.1	14357.1	17425.3	18199.8
75	ADAM	130	0	0.73	0.72	535396669	564242215	13326.5	14501.8	17366.4	18064.9
79	ADAM	150	0	0.73	0.72	536741651	568369222	13220.1	14319.3	17422.7	18132.1

is usually expressed as minimizing error/residual or some transformation of it, to be more precise the linear regression in this case where multiple explanatory variable exist it is called multiple linear regression [35]. Next Ensemble learning based method – Gradient Boosting Regression (GB-Regression) [27] analysis conducted on the features sets in terms of the regular data and the scaled data. The configurations for such model include 400 estimator trees with number of five levels for the max depth and minimum two samples required to split a node at learning rate 0.1 with loss function to be optimized namely least squares regression. The results yield over fitted model in acceptable range according to the standard score function included in the model implementation, whilst higher results when considering R2 (R-Squared) that is coefficient of determination provided by the same learning model implementation [34]. However, used regression models' alongside the standard model-based evaluation [36] scores yielded on the scaled data where summarized in the Table.3.

Another Ensemble learning based method employed with almost the same number of estimators set before for GB-Regression is Random Forest [37]. The analysis conducted on the features sets in terms of the regular data, yielded worst results in terms of model overfitting for training data with no any significance changes in the prediction results based on the standard model score function, the results reported in Figure.12.

Furthermore, investigations following acquiring the overfitting results from the previous trail, lead in initiating two search processes to figure out the best configurations for the decision tree, i.e. what are the best attributes or parameters for the Random Forest method to better fit the data and create a good price estimator on unseen data. Total 450 runs for the different configurations provided that could divided into three similar runs of each configuration based on a different data split (fold). The reported results show minor improvement over the almost default model, and this can be judged due to the theory of the ensemble methods that several models shall be better than one. To provide more convenience, the results acquired upon using such learning methodology depicted clearly in Table.4 using the standard model-based evaluation
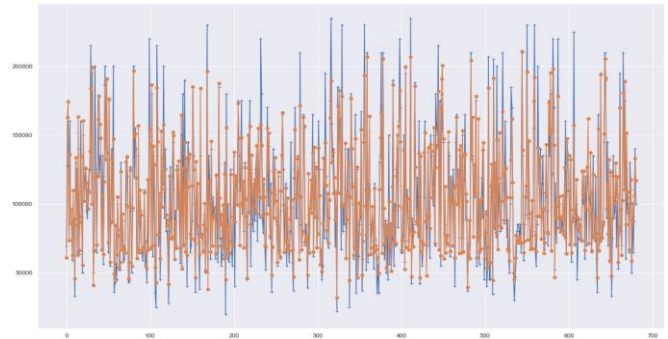
**Figure 12:** Random Forest 500 Estimators Overfitting [Training: 0.96, Test: 0.82].

scored results.

The best estimator configurations [34] founded for Random Forest model were 1500 estimators with max depth 100 and minimum sample leaves of 2 and 6 minimum samples splits evaluated using the criterion MSE [38].

On the other hand, Non-linear SVMs [33] to aim to fit a more complex and non-linearly separable data using the RBF Kernel conducted on the features sets in terms of the regular and scaled data. Results yielded bad results with no any significance changes in the prediction results to overcome the previous trails and models for the regular data, whilst it indicated that SVMs perform better with the scaled data as it uses distances, so a drastic difference in distances will cloud the finer trends in the classifier (or used regression model). Experiments shows the overfitting is high due to the difference for the R2 score between the training and the test datasets, yet its acceptable range for the model. Leading the work to continue its way toward exhaustive figuring out the best-optimized model's hyper parameters. Nevertheless, the reported results show minor improvement over the almost default scaled model (overfitting went down, hence better results for the model. i.e. more generalized model). The results reported based on the standard R2 score function for this stage in the Table.5.

Model Type	R2		MSE		MdAE		MeAE	
	Train	Test	Train	Test	Train	Test	Train	Test
Linear Regression	0.67	0.71	637955720.47	587972332.82	15565.15	15893.58	19509.84	19227.21
GB-Regression	0.98	0.82	43477365.96	371593383.38	2510.74	9485.87	4220.98	13674.42
SVM	0.90	0.72	0.10	0.30	0.01	0.30	0.16	0.39
SVM – Best Estimator	0.84	0.77	0.16	0.24	0.12	0.28	0.25	0.36
Random Forest	0.97	0.82	67627086.20	360681397.17	3925.50	9529.00	5671.71	13455.18
Random Forest – Best Estimator	0.92	0.82	146483386.29	367440117.15	5719.66	9973.90	8415.14	13727.93

Table 11: Other Models' Evaluation Scores
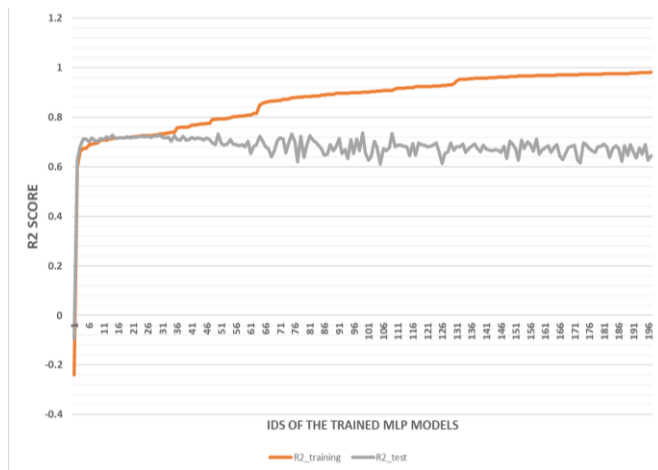
Figure 13: MLP Evaluated Models Using R2.

Finally, this work continuous to the most complex learning methods. Usually an MLP or a neural model should exceed other models in performance. Given that, if it tuned well, a process that is high in cost, and could be infeasible considering the way simpler models can give satisfactory results for some problems. Here we tried two types of solvers (ADAM and L-BFGS), each is different in degree and where it excels. However, the most complicated parameter to tune is the architecture size, due to it being not a single parameter and the high dynamicity of it. That was addressed by fitting around 196 MLPs different in architecture and solvers while sticking with mostly default parameters except for the activation where we went with the ReLU as we're not constrained in computing power to a degree that we should try something simpler as the ReLU provides the flexibility in activation and scaling. Each model where allowed to go up to 100000 epochs while constrained with the tolerance over the enhancement in loss over a validation split of 20\% of the training data. Exhaustive trails and stochastic optimizations applied to the learning models, whilst each got an ID and exported to external serializable files for further analysis, evaluation and future usage, for instance Figure.13. For convenience, all the used MLP models' standard model-based evaluation [34] scores yielded on the scaled data where summarized in the Table.6.

4.3 Experiential Machine Used

All the analysis early mentioned and the results for all experiments conducted on machine that got the hardware specifications as depicted in Table.8.

5. RESULTS AND DISCUSSIONS

5.1 Predictive Modeling Evaluation Criteria

Fitting time dismissed here as apparent by the experiments of the MLP and the grid search optimization for the SVM and Random Forest. While the R-squared was the main evaluation criterion through the experiments. Other common regression evaluation metrics added later on like mean squared error, median absolute error and mean absolute error. Moreover,

since all the trained and tested models backed-up on external files, providing the chance to try different evaluation measures, and assessing the conduction of the final models evaluation and selection for both training and testing datasets. For instance, measures depicted in Table.7 are the chosen criteria to amend this comparative work with the evaluation results needed.

5.2 Predictive Modeling Evaluation

Considering the evaluation criteria shown early in this work, in here, listing for the evaluation results will take place, where it will be organized in showing the MLP results first and the best evaluated architectures followed by the discussions related, then the rest of the models and the best accuracy achieved.

Starting with MLP trained and tested models. The resulted evaluation scores shown in Table.9 depict that L-BFGS solvers out performs well over the Adam solvers in terms of overfitting on smaller datasets as all of them ranked in terms of R2 evaluation measure for the top ten the trained and the twisted MLP models for both solvers.

The best achieved results for the MLP models that got the highlighted architectures where all have one layer with number of neurons less than or equal to 180 neurons for the first. One hundred sixty nine different models trained on longer periods measured in hours, all evaluated and the best accuracy achieved in terms of R2 measure was (0.72) for both training and test data sets. Further results shown by the MSE measures that depict training got loss score greater than test score with a little difference on unseen data, yet still acceptable results for the models. Another interested result for MLP models evaluated related to the MdAE and MeAE measures that appeared to measure the well fit considering the MdAE's robustness for the outliers, and this proof the correctness of the methodology adopted in this work for removing the outliers. Whilst the Table.10 shows the average results for each evaluation, measure used for different MLP solvers among the trained and tested models. It appears clearly that L-BFGs solvers perform well on smaller datasets while Adam solvers suffer from learning overfitting for such type of datasets. Other models shown in Table.11, where Random Forest ensemble method, which used for conducting complicated predictive analysis for the given data. Evaluation results using standard score function for random forest trained and tested models on regular data basis, i.e. data not transformed or scaled, shows clear overfitting which listed early in this work early in Figure.12. The initiated two search strategies, grid and random, where grid search used to test configurations on lower searching rates, which then supplement random search for the sake of figuring out best configurations for the random forest algorithm. This process resulted in number of estimators 1500 of max depth of 100; error function MSE for minimum samples split and leafs of six, and two respectively as all the best prices estimator configurations. Even the primary evaluation results

shows minor improvement over the default model, but the R2 remained showing higher deviation (i.e. Overfitting) achieved as MLP modeling techniques used for random configuration lookup. These results can attributed to either low instances searched and/or the efficiency of the default models to begin with on smaller dataset.

Finally, Linear and GB-Regression evaluated well using R2 score function with a small indication for minor trendy model overfitting. Whilst the non-linear SVMs using RBF Kernel yielded in two very different standard model score-based results, giving an indication that SVMs perform better with the scaled data as it uses distances, so a drastic difference in distances will cloud the finer trends in the classifier (or used regression model). Rather, best estimator configurations founded shows minor improvement over the default model even reporting with R2 evaluation metric, which shows better results in the other models employed.

6. CONCLUSION

Not all the previously mentioned aspects and benefits to work on such problem changed the state of literature and the work on the Jordanian market. Whilst the scarcity of such works imposes a need for it, and demonstrates an unutilized aspect of the economy that requires little resources to create some business and academic opportunities. In addition, helping in further understanding of Jordanian economy, and potentially diagnosing some problems and recommending actions to be taking to revert the declination of that economy and identify where correction should be concentrated. Going further to cover all the potential inductions and recommendations amongst other possible extractions and solutions to this problem start with methodological data science's use case. The use case shown is offering a comparative study of several predictive and descriptive methods over an online collected data of apartments for sale in Jordan and their prices alongside listed features. The focus of the work concentrate mainly on data mining and machine learning different techniques. Some anecdotal insights were listed that relates to the apartments market and on the techniques used in this work, where the experimental results might indicate some relations, the consolidation of the observed interrelations still needs a more thorough study with more comprehensive tests on bigger datasets. However, the work focuses on exploring some of the attributes interactions while preparing them for the predictive modelling, and reshaping, transforming and filtering the data for a better learning by addressing some of basic data problems like distribution, outliers and incompatible attribute types with some models. Then and based on a selected subset of machine learning models that vary in complexity and behavior fitted with while tuning both the models and data in atrial to enhance performance, judging based on several criteria (R2, MSE, MdAE, and MeAE) and trying to explain and evaluate the final results obtained. The next steps for this work will go deeply

inside the local and/or global markets different areas, trying to provide insights about, enrich, and curate the local and/or global economic aspects, which had better enhance the understanding and aid specialists to draw insightful conclusions about the market different indicators using different data science techniques

REFERENCES

1. M. Hoesli, C. Lizieri, and B. MacGregor. **The inflation hedging characteristics of US and UK investments: a multi-factor error correction approach**. The Journal of Real Estate Finance and Economics, 36(2), 183-206. 2008.
   https://doi.org/10.1007/s11146-007-9062-6
2. S. J. A. P. Zavei, and M. M. Jusan. **Exploring housing attributes selection based on Maslow's hierarchy of needs.** Procedia-Social and Behavioral Sciences, 42, 311-319. 2012.
3. S. Hudson-Wilson, F. J. Fabozzi, and J. N. Gordon,. **Why real estate?**. The Journal of Portfolio Management, 29(5), 12-25. 2003.
4. J. Aizenman, and Y. Jinjarak. **Real estate valuation, current account and credit growth patterns, before and after the 2008–9 crisis.** Journal of International Money and Finance, 48, 249-270. 2014.
5. A. Anari, and J. Kolari. **House Prices and InflationReal Estate Economics**, 30(1), 67–84. 2002.
6. K. E. Case, E. L. Glaeser, and J. A. Parker. **Real estate and the macroeconomy.** Brookings Papers on Economic Activity, 2000(2), 119-162. 2000.
7. S. B. Billings. **Hedonic amenity valuation and housing renovations**. Real Estate Economics, 43(3), 652-682. 2015.
   https://doi.org/10.1111/1540-6229.12093
8. N. Shinde, and K. Gawande. **Survey on predicting property price.** In 2018 International Conference on Automation and Computational Engineering (ICACE) (pp. 1-7). IEEE. October 2018.
9. B. Sopranzetti, **Hedonic Regression Models.** pp. 2119–2134 10 1007 978–1–4614–7750–1 78.)
10. D. Harrison Jr, and D. L. Rubinfeld. **Hedonic housing prices and the demand for clean air**. 1978.
11. J. Mu, F. Wu, and A. Zhang. **Housing value forecasting based on machine learning methods**. In Abstract and Applied Analysis (Vol. 2014). Hindawi. January 2014.
12. J. M. N. Tabales, J. M. Caridad, and F. J. R. Carmona, F. J. R. **Artificial neural networks for predicting real estate price.** Revista de Métodos Cuantitativos para la Economía y la Empresa, 15, 29-44. 2013.
13. E. D. Sontag. **Feedback stabilization using two-hidden-layer nets.** In 1991 American Control Conference (pp. 815-820). IEEE. June 1991.
14. Y. Li, and Y. Yuan. **Convergence analysis of two-layer neural networks with relu activation.** In Advances in neural information processing systems (pp. 597-607). 2017.

15. R. Manjula, S. Jain, S. Srivastava, and P. R. Kher. **Real estate value prediction using multivariate regression models.** In IOP Conference Series: Materials Science and Engineering (Vol. 263, p. 042098). November 2017.

16. A. A. Neloy, H. S. Haque, and M. M. Ul Islam. **Ensemble learning based rental apartment price prediction model by categorical features factoring**. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (pp. 350-356). February 2019. https://doi.org/10.1145/3318299.3318377

17. A. Singh, and R. Lakshmiganthan. **Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms**. 2018.

18. V. Limsombunchai. **House price prediction: hedonic price model vs. artificial neural network.** In New Zealand agricultural and resource economics society conference (pp. 25-26). June 2004.

19. G. J. McKee, and D. Miljkovic. **Data Aggregation and Information Loss** (No. 381-2016-22080). 2007.

20. H. YILDIRIM. **Property Value Assessment Using Artificial Neural Networks, Hedonic Regression and Nearest Neighbors Regression Methods.** Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi, 7(2), 387-404. 2019.

21. M. Miyamoto, and H. Tsubaki. **Measuring technology and pricing differences in the digital still camera industry using improved hedonic price estimation**. Behaviormetrika, 28(2), 111-152. 2001.

22. H. Xu, and F. Mueller. **Work-in-progress: Making machine learning real-time predictable.** IEEE Real-Time Systems Symposium (RTSS) (pp. 157-160). IEEE. December 2018.

23. N. H. Abroyan, and R. G. Hakobyan. **A review of the usage of machine learning in real-time systems.** Bulletin of the National Polytechnic University of Armenia. Information Technology, Electronics, Radio Engineering, (1), 46-54. 2016. https://doi.org/10.22606/fsp.2017.12002

24. R. Bellazzi, C. Larizza, P. Magni, S. Montani, and G. De Nicolao. **Intelligent analysis of clinical time series by combining structural filtering and temporal abstractions.** In Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (pp. 261-270). Springer, Berlin, Heidelberg. June 1999.

25. X. Chen, L. Wei, and J. Xu. **House price prediction using LSTM**. arXiv preprint arXiv:1709.08432. 2017.

26. E. Ahmed, and M. Moustafa. **House price estimation from visual and textual features**. arXiv preprint arXiv:1609.08399. 2016.

27. F. Zhang, B. Du, and L. Zhang. **Scene classification via a gradient boosting random convolutional network framework.** IEEE Transactions on Geoscience and Remote Sensing, 54(3), 1793-1802. 2015.

28. L. Breiman. **Using iterated bagging to debias regressions.** Machine Learning, 45(3), 261-277. 2001.

29. M. F. De Oliveira, and H. Levkowitz. **From visual data exploration to visual data mining: A survey.** IEEE transactions on visualization and computer graphics, 9(3), 378-394. 2003.

30. S. Lima, A. M. Gonçalves, and M. Costa. **Time series forecasting using Holt-Winters exponential smoothing: An application to economic data.** In AIP Conference Proceedings (Vol. 2186, No. 1, p. 090003). AIP Publishing LLC. December 2019. https://doi.org/10.1063/1.5137999

31. C. Lee, O. Kwon, M. Kim, and D. Kwon. **Early identification of emerging technologies: A machine learning approach using multiple patent indicators**. Technological Forecasting and Social Change, 127, 291-303. 2018.

32. R. Berwick. **An Idiot's guide to Support vector machines (SVMs).** Retrieved on October, 21, 2011. 2003.

33. A. Abanda, U. Mori, and J. A. Lozano. **A review on distance based time series classification.** Data Mining and Knowledge Discovery, 33(2), 378-412. 2019.

34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... and J. Vanderplas. **Scikit-learn: Machine learning in Python.** The Journal of machine Learning Research, 12, 2825-2830. 2011.

35. R. A. Bottenberg, and J. H. Ward. **Applied multiple linear regression** (Vol. 63, No. 6). 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland Air Force Base. 1963.

36. E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, ... and M. W. Kattan. **Assessing the performance of prediction models: a framework for some traditional and novel measures**. Epidemiology (Cambridge, Mass.), 21(1), 128. 2010.

37. L. Breiman. **Random Forests**. Machine learning, 45(1), 5–32. 2001.

38. N. J. Nagelkerke. **A note on a general definition of the coefficient of determination**. Biometrika, 78(3), 691-692. 1991.

39. Z. Wang, and A. C. Bovik. **Mean squared error: Love it or leave it? A new look at signal fidelity measures**. IEEE signal processing magazine, 26(1), 98-117. 2009.

40. C. J. Willmott and K. Matsuura. **Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance**. Climate research, 30(1), 79-82. 2005.

41. R. Tripathi, and D. P. Rai. **Comparative Study of Software Cost Estimation Technique.** International Journal of Advanced Research in Computer Science and Software Engineering, 6(1). 2016.

42. V. Kale, and F. Momin. **Video Data Mining Framework for Surveillance Video.** International Journal of Advanced Trends in Computer Science and Engineering, 2(3). 2013.

43. A. A. Mahule, and A. J. Agrawal. **Hybrid Method for Improving Accuracy of Crop-Type Detection using Machine Learning**. International Journal, 9(2). 2020.

44. M. Akour, O. Al Qasem, H. Alsghaier, and K. Al-Radaideh. The effectiveness of using deep learning algorithms in predicting daily activities. International Journal, 8(5). 2019.