A Machine Learning Based Solution for Predicting Land Values

E-Valuer Land value predictor

Dharmaseelan T.

(IT16116566)

Bachelor of Science (Honors) in Information Technology Specialized in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

August 2019

A Machine Learning Based Solution for Predicting Land Values

E-Valuer Land value predictor

Dharmaseelan T.

(IT16116566)

Bachelor of Science (Honors) in Information Technology Specialized in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

August 2019

DECLARATION OF THE CANDIDATE & SUPERVISOR

I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature of the student: ……..... Date: ……

The above candidate has carried out research for the B.Sc Dissertation under my supervision.

Signature of the supervisor: ……..... Date: ……

ABSTRACT

The Colombo stock exchange is brimming with unpredictability. It constantly will in general change with inward and outside variables. Since it's generally acknowledged that the disappointment in the Effective market speculation. Stock value variations are driven by broadcasting of news items. Lenders or merchants' choices are made dependent on the accessible data thinking about how a stock exchange will going up or going down for next couple of days. News articles join data about a company's essentials, the activities in which a firm is included and the desires for other market members about future value changes. This research attempts to concentrate on the few external components. Anticipating the future expecting varieties in the stock exchange can be by analyzing. This analyzing should be possible on three kinds. Those are fundamental, technical and machine learning. In this research paper use the technical analysis. NLP use for Data scraping in the means of a summery, from a selected website. Then merge all duplicated news collected from various sources and unstructured data will be saved as (CSV) file format. Extracted text was then subject to standard NLP steps of tokenization. Those tokens were then having to lemmatize to get the lemma of a particular word. For automatic detection of news category, following dictinary-based approach.

ACKNOWLEDGEMENT

Final year Information Technology Research Project is a subject, which provides a last opportunity for IT undergraduates to have an exposure of the real industry in the field of Information Technology. This is totally a new situation for us; thus, we needed guidance and help from our lecturers. We gratefully acknowledge those who contributed to our project. First and foremost, we would like to thank the Sri Lanka Institute of Information Technology for providing us with the knowledge and the tools to complete the project. They also initialized the requirement of the final year project which is vital for completing the course.

We would also like to thank the driving force behind the project, our supervisor Mrs.M.P.A.W.Gamage for providing us support in numerous ways. He guided us at our difficult times, motivating and pointing out our mistakes to make the project a success. We couldn't have been able to complete the project without the excessive support and guidance them.

Finally, our most sincere thanks go to all four group members who participated for the success of this project. We thank everyone who helped us to complete the project in numerous ways and who guided us in many ways.

1. INTRODUCTION

1.1 Background Literature

Real Property are the most valuable possession of most of the common people. Getting the proper valuation for these real properties is very much important. This system is an innovative solution proposed to facilitate land valuation based on past sales, prediction of future price and the effect of proposed development work on the land, so that real-estate customers and owners of real estate companies can be benefitted and make smarter property related decisions. In practice, this manual process makes a lot of problems. Specially, this is a personal oriented method. The way of analyzing the features and giving them a value may be changed person to person. It will lead to different values for a same land by different valuation officers. Therefore, there should be a proper

concept or a mechanism in valuing a land .[1] Sri Lanka, the valuation officer would consider a lot of external and internal factors. As external factors, the basic infra structure facilities such as presence of electricity, telephone facility and water would be considered in the process. Also the distances to other facilities such as schools, main and minor roads and financial institutions in the surrounding area play a vital role.[1] Sri Lanka is lacking a good data platform to gather all these data, considering all these factors can take ages to do proper valuation considering all these factors. All above mentioned problems can be solved with the implementation of the system which will save time, cost, and with improved fairness of the predictions which will be done.

1.2 Research Problem

Our research problem primarily revolves land price in Colombo city. Indentitfying method to predict future value of the land. They are going to buy, on time, to make better decisions. This application tool which can help people to get an idea about the land they are hoping to buy, to decide the suitability of their purpose. To develop a portable applications which can provide instant report of a selected land parcel to provide the users with an insight of the land with future value, just by giving location.

1.3 Research gap

There are so many systems like our system in abroad. Our intention is to identify the ways to use their underlying methodology in a suitable manner in Sri Lankan context. we use multiple sources of data to entangle the economic contribution of the neighborhood's characteristics such as walkability and security perception. We also develop and release a framework able to now-cast housing prices from Open data, without the need for historical transactions. Experiments involving 70, 000 houses in 8 Italian cities highlight that the neighborhood's vitality and walkability seem to drive more than 20% of the housing value. Moreover, the use of this information improves the nowcast by 60%.[2]

The initial house prices prediction is challenging and requires the best method to get the best prediction accuracy. In the predicting the sale price of a house that has an uncertainty parameter, fuzzy logic becomes one of the solutions that can be used in solving the problem [3]. Moreover, artificial neural network methods are used to predict house selling prices [3]. In addition to using fuzzy logic and artificial neural networks, predictions can also use the K-Nearest Neighbors algorithm, for estimating residential prices for the residential property market in Hong Kong [3]. Several machine learning methods are compared to get the best prediction of house pricing [3].

1.4 Research Objectives

1.4.1 Main Objectives

To provide people with fair and accurate prediction of the value of the land. They are going to buy, on time, to make better predictions.

1.4.2 Specific Objectives

In order to achieve the main objective, the following specific objectives mentioned below should be achieved

- When you search the location and select the future value option
- Select the Year it will give exact price of the particular area and selected year price also can see there

- Forecasting graphs also there

2. METHODOLOGY

2.1 System Overview

Pick My Service Web application mostly focused on customer, service provider and it will be helpful to find the actual land prices in Colombo city for particular area. Addition to that, service provider has their own value added by the system. Also, the system can review user feedback, comments and calculate the overall rating system for specific service providers. Given the following Figure 1 shows the system diagram of this project.

Multivariate Linear Regression

In this function, hope to figure out a relationship between land price and latest other factors such as state, city, ZipCode, Price, Pollution index, Hospital Distance, Tourist Score, Bank/ATM, Sch ool Distance, DistanceToTown,Population index and weather some other selected areas. Multivar iate Linear Regression had several advantages than other algorithms. The ability to determine th e relative influence of one or more predictor variables to the criterion value. multivariate techn iques provide a powerful test of significance compared to univariate techniques.[1] multivariate t echniques to give meaningful results, they need a large sample of data; otherwise, the results are

meaningless due to high standard errors. [4] Standard errors determine how confident you can be in the results, and you can be more confident in the results from a large sample than a small one. Running statistical programs is fairly straightforward but does require statistical training to make sense of the data. Categorized Data used as training sets and test sets(40% allocate for test data).P rice is the dependent variable and take it as y and others are dependent variable. fitting the train d ata into the model. Make the prediction on test data set.

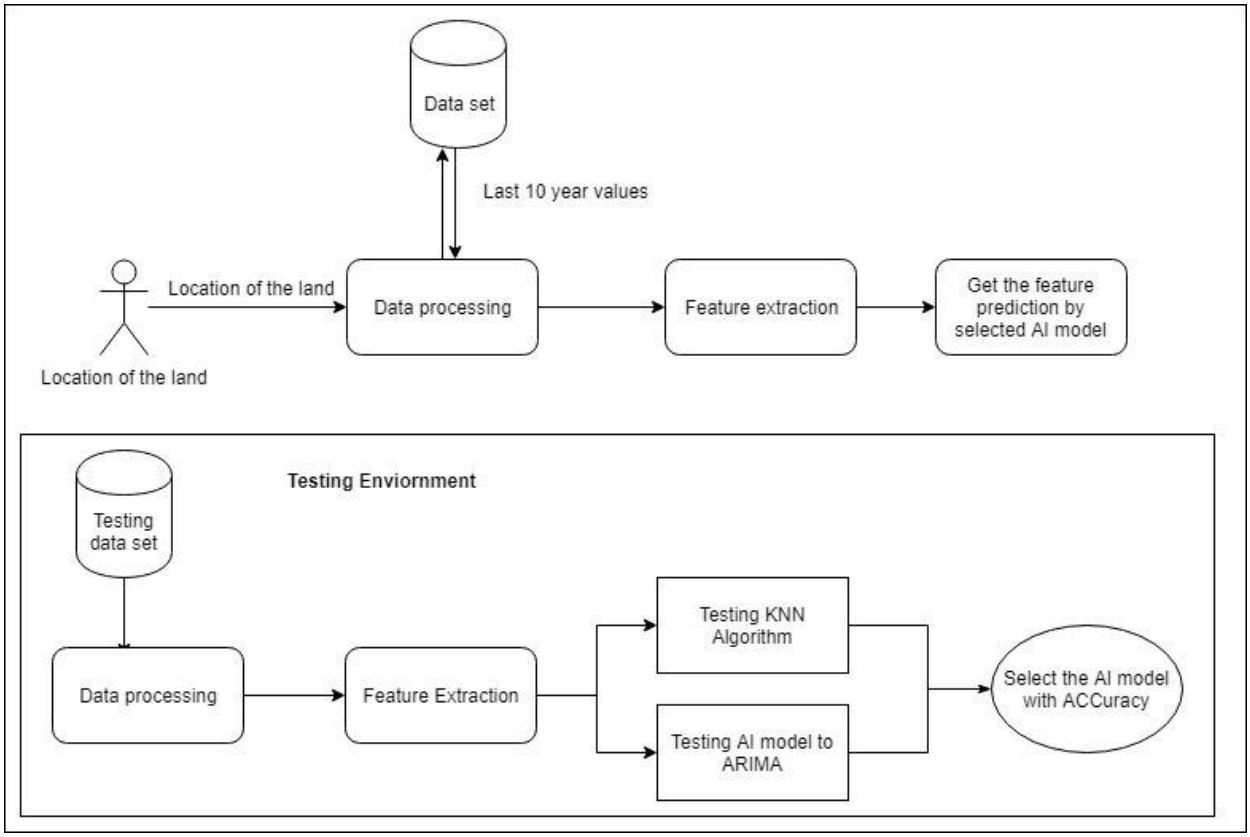
Then the specific values for each sector are coming as the output. Then the output will be compared with other machine learning algorithms. Extracted structured data is gained by the other group members and then the data is fed into the algorithm. After that a comparison runs between these data and price with aid of the algorithm. Hereafter, the output which comes from this algorithm and other member's algorithm is compared. Finally, the decision is made which the most trustworthy prediction.

Knn algorithms

[5]As we saw above, KNN can be used for both classification and regression problems. The algorithm uses '**feature similarity**' to predict values of any new data points. Categorized Data used as training sets and test sets (40% allocate for test data). Initialize the model. After that fit the train data. To Store rmse values for different k. fit the model. Make the prediction on test data set. Calculate the rmse value and store the rmse value.

ARIMA

[6]An Arima Model is a class of statistical models for analyzing and forecasting time series data. Define the model by calling ARIMA ()and passing in the P=5,d=1,q=0 parameters. [3]This set the lag value to 5 for auto regression, uses a difference order of 1 to make the time series stationary, and uses a moving average model of 0.Split the training dataset into train and test sets, use the train set to fit the model, and generate a prediction for each element on the test set. A rolling forecast is required given the dependence on observations in prior time steps for differencing and the AR model. A crude way to perform this rolling forecast is to re-create the ARIMA model after each new observation is received. We manually keep track of all observations in a list called history that is seeded with the training data and to which new observations are appended each iteration. Putting this all together, below is an example of a rolling forecast with the ARIMA model in Python.



2.2 Finding the Service providers available Geo-location

1) Identifying a device's location

Geo-location for the Web version of our application implemented using Google's Location APIs. It can intelligently manage underlying location technology while meeting various development needs when implementing location-based features.

2) Providing directions

To display point-to-point directions on a map within the application, we used Google Maps API. Registering the app as a routing app then makes exact locations and infra structure

3) Integrating with mapping software

To implemented location-based service we use Google Maps. Geo-location is the most important technology.

2.3 Methods of delivering services

Application delivery refers to the pool of services that combine to provide application functionality, usually web-based software applications, from the data centers or cloud environments where the required data processing and computing is executed to the application clients or end-users over the internet. The services for delivering applications on a network infrastructure aim to provide a reliable user experience by providing, load balancing, security, latency and TCP optimizations which combine to provide application content seamlessly. API and Web service serve as a means of communication. The only difference is that a Web service facilitates interaction between two machines over a network. An API acts as an interface between two different applications so that they can communicate with each other. An API is a method by which the third-party vendors can write programs that interface easily with other programs. A Web service is designed to have an interface that is depicted in a machine-processable format usually specified in Web Service Description Language (WSDL). Typically, "HTTP" is the most commonly used protocol for communication. Web service also uses SOAP, REST, and XML-RPC as a means of communication. An API exactly defines the methods for one software program to interact with the other. When this action involves sending data over a network, Web services come into the picture. An API generally involves calling functions from within a software program.

2.4 Choose the service providers

To find the most suitable service provider for customer, Our Web applications have several options. In this process after filtering service providers according type customer can choose below options,

1. System can select the location.

2. Customer can choose the current price or future price.

3. Customer can view the selected year and statics graph.

2.5 User Authentication

In our Web Service application, we provide user name and password authentication.

We can access the data at any time through a REST API. As people log into our web application, this database is populated with a list of phone numbers and Account IDs that can be used within our app. These Account IDs are unique to our application.

2.6 Prediction

MLR,KNN algorithms used for non time series prediction and Linear prediction. ARIMA algorithm used for Time series algorithm and used Moving average.

2.6.1 Types of prediction

1. NON Time series Linear Regression prediction.
2. Time Series Moving Average prediction.

2.6.2 Data Collection

Our dataset has Land price in places in Colombo Districts several from 20012 to 2018.We are g oing to predict the Future Price for 10 years. state, city, ZipCode, Price, Pollution index, Hospital Distance, Tourist Score, Bank/ATM, School Distance, DistanceToTown,Population index,Bid da te. Dataset has above 500 samples.Each places has 8 or more than 8 samples.

2.6.2 Feature Extractions

IF you add more attributes we can more accurate answers. IF you consider economic and physical status we can get more accurate answers.

2.7 Implementation and Testing

System implementation is a major role of Systems Development Life Cycle. Implementation is developing the system according to our plan or design. For implementing the Web application we selected Angular 6 version and to create the models we used JUPYTER Note book Machine Learning tool. We selected Mongodb as our database since it supports cloud based database activities. By deploying the database in a cloud environment we can easily put updates up-to-date in the connected system. Using those techniques and tools we implemented the system.

The complex component of Web application is testing. Its critical to have full-developed testing framework so by processing and accessing the whether the functionalities of system satisfies with the actual requirement or not, so identifying requirements software testing process will work. So basically, software testing can be mentioned as

• Unit Testing

• Component Testing

• Integration Testing

• System Testing

So, software testing process also must include strategy, performance, usability, functional and nonfunctional, security and one of the major testing strategies is testing the network connection and strength of the connection when it comes to sanctuaries in the middle part of the location there won't be any connection. For the requirement we won't able to test with emulators because we deal with the real data environment, so the best testing strategy is real user monitoring.

Unit Testing Each unit is tested individually to find whether it's fit for use. This used to identify smallest part of problems earlier stages of testing, and most important thing in unit testing is identify the bug than correcting it.

Component Testing

Each component testing done in the application separately also its known as program testing here it found the bugs or defect and take the actions to correct it.

Integration Testing

Each module of the software combined and tested as group.it must be test after unit testing.

System Testing

This is the level of testing where complete software and integrated software is tested. It verified as system whether it meets the requirements. This will ensure the quality level of the system.

2.8 Implementation and Testing

Tools

• Anaconda IDE

• Angular6

• Node JS

• Google Collaboratory
• Mongo Db

Technologies

• Python

• pandas Library

• Matplot Library

3 RESULTS AND DISCUSSIONS

3.1 Results

Comparison of RMSE,MSE and MAE of the MLR,ARIMA,KNN for Future value prediction and pre-trained architectural models are presented in this section. To do the training and testing processes, we consider current value , past value, future infrastructure based value datasets. All the models were trained and validated on the same training and test dataset and using the same parameters to compare results. Each. The training process was very fast even without using the Graphics Processing Unit (GPU). Because instead of starting from scratch, the datasets that we

used is small and the models are built on top of the pre-trained deep learning models. Each of these architectures takes various time periods per epoch to train and test particular data sets. All these models mentioned above were processed with a 12GB RAM and Intel Core i7 2.50 GHz processor in Microsoft Windows 10 operating system environment.

3.1.1 Past Value Data

Dataset used for future value prediction has Land price in places in Colombo district from 20012 to 2018 which the algorithm is going to predict the future price for 10 years with features state, city, zip code, price, pollution index, hospital distance, tourist score, bank/ATM, distance to school , distance to town, population index, bid date having above 500 samples .Each places has 8 or more than 8 samples.

First approach was based on features of the land and historical data of price fluctuation rates. There has been used a combination KNN and MLR versus ARIMA model. Results can be concluded as follows. These two models when evaluated, displays MSE value for multivariant linear regression is 590293123907492.1, MSE value for ARIMA model is 605023251112851.9. Therefore, MLR and KNN combination can be seen as a more accurate option for future price prediction.

ARIMA Model Results

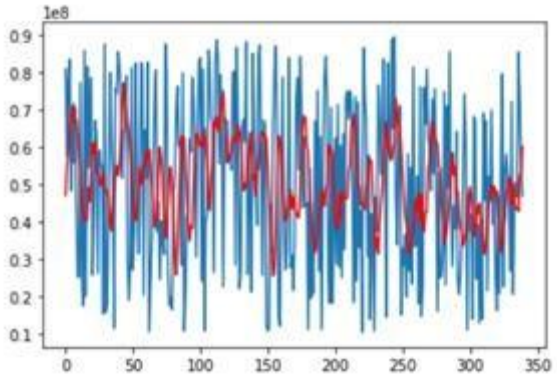
Dep. Variable:	D.y	No. Observations:	998
Model:	ARIMA(5, 1, 0)	Log Likelihood	-18426.140
Method:	css-mle	S.D. of innovations	25231482.665
Date:	Sun, 04 Aug 2019	AIC	36866.279
Time:	11:00:00	BIC	36900.619
Sample:	1	HQIC	36879.332

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2253.6727	2.27e+05	-0.010	0.992	-4.47e+05	4.43e+05
ar.L1.D.y	-0.8700	0.031	-27.770	0.000	-0.931	-0.809
ar.L2.D.y	-0.7050	0.041	-17.356	0.000	-0.785	-0.625
ar.L3.D.y	-0.5085	0.043	-11.691	0.000	-0.594	-0.423
ar.L4.D.y	-0.3001	0.041	-7.372	0.000	-0.380	-0.220
ar.L5.D.y	-0.1418	0.031	-4.519	0.000	-0.203	-0.080

Roots

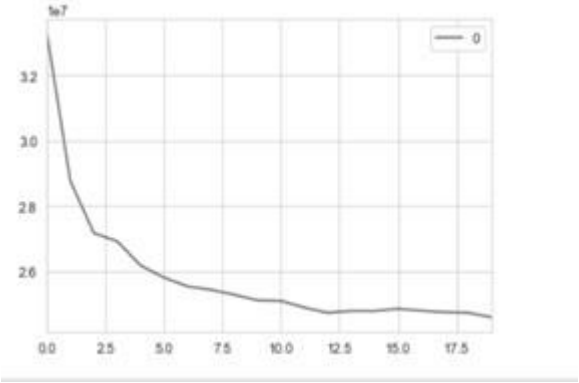
	Real	Imaginary	Modulus	Frequency
AR.1	0.5117	-1.3305j	1.4255	-0.1916
AR.2	0.5117	+1.3305j	1.4255	0.1916
AR.3	-1.5044	-0.0000j	1.5044	-0.5000
AR.4	-0.8177	-1.2800j	1.5189	-0.3405
AR.5	-0.8177	+1.2800j	1.5189	0.3405

shows the results obtained by ARIMA model.



ARIMA model forecasts

Below is the variation of RMSE with K value as concluded by the KNN model used for price prediction.



RMSE versus K values

3.1.2 Comparison of pre-trained based models with different datasets

	MAE	MSE	RMSE
MLR	21125295.265937325	590293123907492.1	24295948.7138801
ARIMA	21397292.126597658	648244532873425.1	25460646.748922642

KNN Algoritm
RMSE value for k= 1 is: 33236335.5633499
RMSE value for k= 2 is: 29350206.021491293
RMSE value for k= 3 is: 27730285.611861072
RMSE value for k= 4 is: 26450157.466238208

RMSE value for k= 5 is: 25962758.90996831
RMSE value for k= 6 is: 25563133.749448232
RMSE value for k= 7 is: 25271278.39128187
RMSE value for k= 8 is: 24852237.123635385
RMSE value for k= 9 is: 24849072.09022946
RMSE value for k= 10 is: 24659752.571712613
RMSE value for k= 11 is: 24568955.481309887
RMSE value for k= 12 is: 24557293.04777497
RMSE value for k= 13 is: 24439868.888457965
RMSE value for k= 14 is: 24297442.464820538
RMSE value for k= 15 is: 24216999.15279268
RMSE value for k= 16 is: 24103323.314502314
RMSE value for k= 17 is: 24175568.901250575
RMSE value for k= 18 is: 24133684.631889954
RMSE value for k= 19 is: 24139068.93470548
RMSE value for k= 20 is: 24042353.731431223

3.2 Discussions

Future Value Prediction

Future value prediction was carried out based on two approaches. Then the specific values for each sector are coming as the output. Then the output will be compared with other machine learning algorithms. Extracted structured data are gained by the fellow members and then the data are fed into the algorithm. After that a comparison runs between these data and price with aid of the algorithm. Hereafter, the output which comes from this algorithm and other algorithm is compared. Finally, the decision is made which the most trustworthy prediction. First approach was based on features of the land and historical data of price fluctuation rates. MLR getting high accuracy and it predicts the non time series algorithm and ARIMA used for time series. So combined the two predicted values get an accurate value.

4.CONCLUSION

In this paper, several tests have been performed using multivariant linear regression , knn and AR IMA methods to perform Land price prediction. In here we used time series algorithms and non ti me series algorithms.t MSE value for multivariant linear regression is 590293123907492.1, MSE value for multivariant linear regression is 605023251112851.9.So Multivariant Linear Regression is the lowest MSE value so That is the best algorithms for non time series algorithms. We merged the Time series and non Time series best algorithm and get a mean value of that prediction values . In future if u give more factors. You can get good Accuracy Value.

5.REFERENCES

[1] Li, L., Prussella, P.G.R.N.I., Gunathilake, M.D.E.K., Munasinghe, D.S. and Karadana, C.A., 2015. Land Valuation Systems using GIS Technology Case of Matara Urban Council Area, Sri Lanka. Bhumi, The Planning Research Journal, 4(2), pp.7–16.

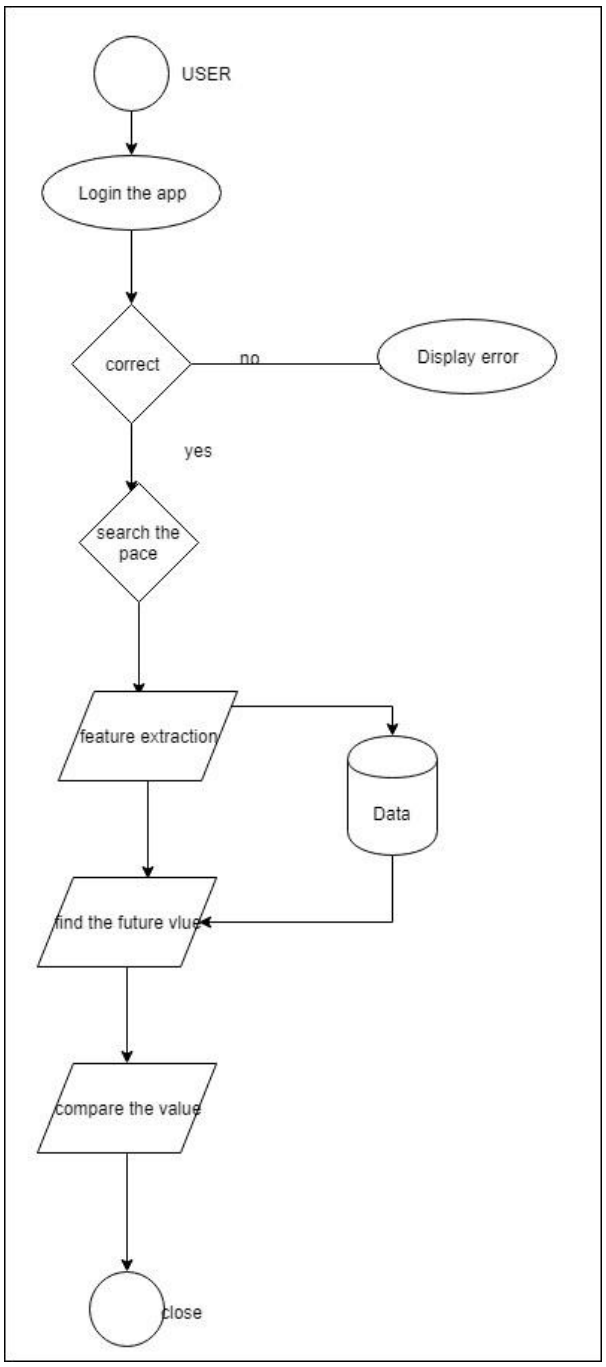
[2] Nadai, M. D., & Lepri, B. (2018). The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa.2018.00043

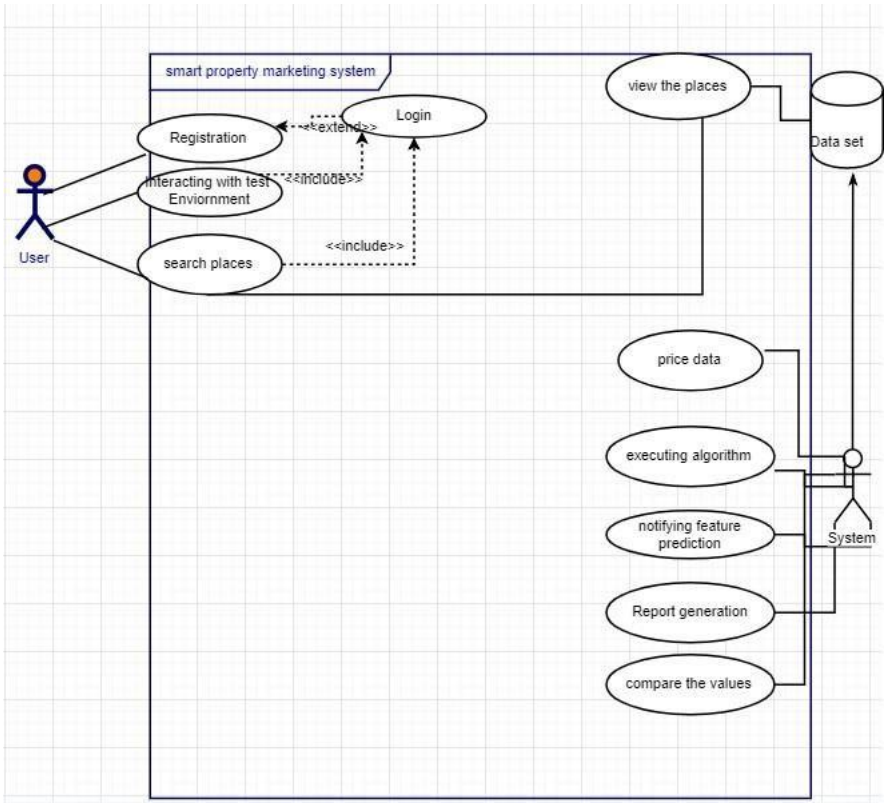
[3]Mukhlishin, M. (2017). Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor - IEEE Conference Publication. [online] Ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/document/8276357 [Accessed 19 Feb. 2019].

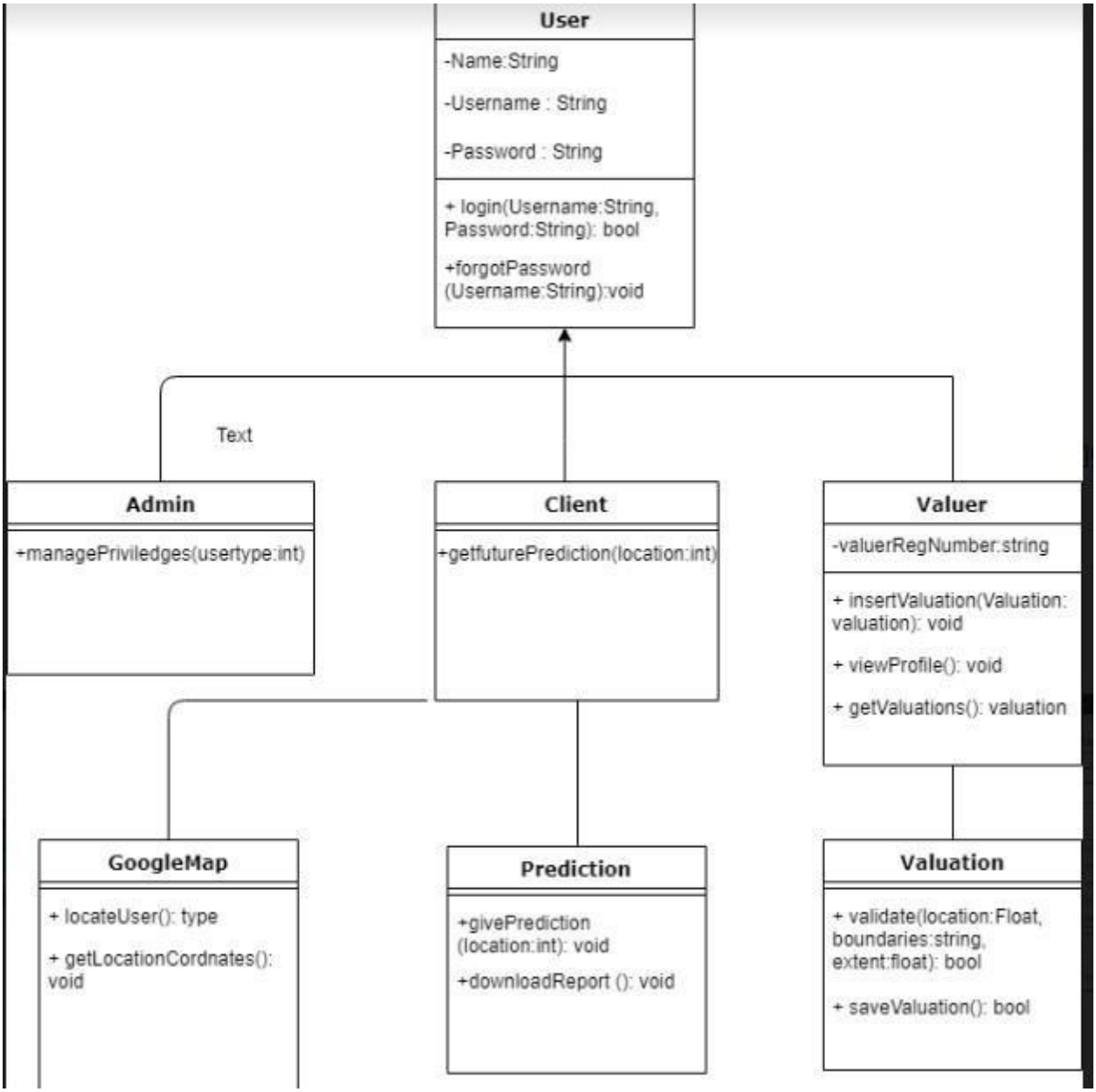
[4] J. Jackson, "Multivariate Techniques: Advantages and Disadvantages," *The Classroom | Empo wering Students in Their College Journey*, 10-Jan-2019. [Online]. Available: https://www.theclas sroom.com/multivariate-techniques-advantages-disadvantages-8247893.html. [Accessed: 04-Aug -2019].

[5] A. Singh, "A Practical Introduction to K-Nearest Neighbor for Regression," *Analytics Vidhya*, 07-May-2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-nei ghbor-introduction-regression-python/. [Accessed: 04-Aug-2019].

[6]"How to Create an ARIMA Model for Time Series Forecasting in Python," *Machine Learning Mastery*, 26-Apr-2019. [Online]. Available: https://machinelearningmastery.com/arima-for-timeseries-forecasting-with-python. [Accessed: 04-Aug-2019].


USER
Login the app
correct
no
Display error
yes
search the pace
feature extraction
Data
find the future vlue
compare the value
close


smart property marketing system
Login
Registration
<<extend>>
<<include>>
Interacting with test Enviornment
User
search places
<<include>>
view the places
Data set
price data
executing algorithm
notifying feature prediction
System
Report generation
compare the values


User
-Name:String
-Username : String
-Password : String
+ login(Username:String, Password:String): bool
+forgotPassword (Username:String):void
Text
Admin
+managePriviledges(usertype:int)
Client
+getfuturePrediction(location:int)
Valuer
-valuerRegNumber:string
+ insertValuation(Valuation: valuation): void
+ viewProfile(): void
+ getValuations(): valuation
GoogleMap
+ locateUser(): type
+ getLocationCordnates(): void
Prediction
+givePrediction (location:int): void
+downloadReport (): void
Valuation
+ validate(location:Float, boundaries:string, extent:float): bool
+ saveValuation(): bool