INTRODUCTION

Background literature

Real Property are the most valuable possession of most of the common people. In Sri Lankan culture, most of the people tend to think that owning a real estate is a better investment than having that money saved in a bank. Therefore, getting the proper valuation for this real property is very much important.

Land valuation is the process of assessing the characteristics of a given piece of land based on experience and judgment.[1] The determination of a land parcel value depends on a number of physical and economic characteristics which must be taken into consideration very carefully in a land valuation procedure.[1] These values can be affected by various social factors too. For example, if there is a crime happened in that land, it can cause a negative effect on the value..

Hence, real estate appraisal it is a challenging multidimensional problem that involves estimating many facets of a property, its neighborhood, and its city.[2]

Since, 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.

The manual process is a time consuming slow task which needs to be done by an experienced professional valuer. The valuation approaches used by those professionals are limited due to the lack of digital data in Sri Lanka. Also, it is a known fact that the valuation process can be so subjective to the person.

Ideally, the systematic process of valuation consists of four different stages as follows.
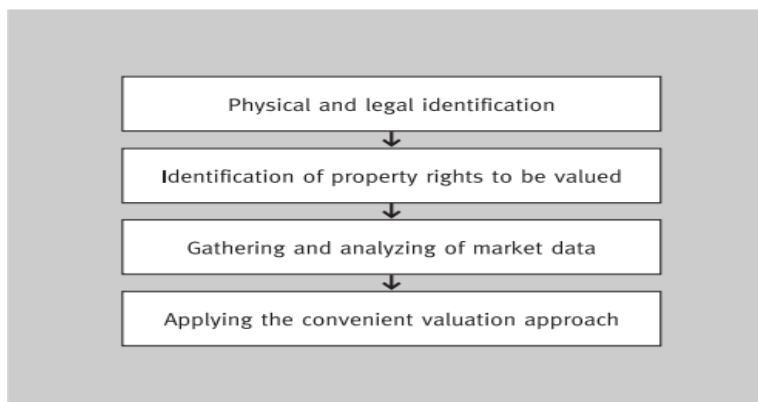

Figure 1.1: Different stages of the appraisal process for estimating the market value

Source: Schulz, R. (2003). Valuation of properties and economic models of real estate markets. Erscheinungsort nicht ermittelbar: Verlag nicht ermittelbar.

The major convenient valuation approaches are,

1. Sales Comparison Approach
2. Income Approach
3. Cost Approach [3]

Analyzing the previous land sale details and trends in those fluctuations and considering those data to predict the valuation is called the sales comparison approach.[3]

The task of automatically estimate the market value of houses can be seen as a regression problem, where the price (or the price per square meter) is the dependent variable, while the independent one is the available information that could help to determine the price correctly. [2]

When the neighbourhood economical value is combined with effect of neighbourhood factors such as walkability etc. we believe it is possible to give a accurate, fair prediction of the value of the land.

The influence of technology on daily life of the Sri Lankans has increased immensely. People tend to use traffic data, online shopping more than ever.

Since the manual process is too slow and dependent to make a quick better decision of the worthiness of the land and suitability of it for the purpose of the customer, our attempt is to digitally assist the people in property related decision making by providing them accurate predictions of the values and future studies of the land.

Our training environment determines the best out of conventional Multivariate Regression Analysis (MRA) and non-conventional Artificial Neural Networks (ANN) methods to do the most accurate prediction. According to Sampathkumar, et al. [20], both the models are found to be well fit with the data set of the land price in all locations, the model using NN (correlation 98%) shows better accuracy than the regression model (correlation 96%), while Zurada, Levitan and Guan, 2011[21] concludes no single obvious non-conventional method that can be expected to consistently outperform traditional multivariate linear regression in predicting residential real

estate sales prices. In the least, the non-conventional methods may be used as a complement to the traditional, multiple regression-based methods[21].

Chaphalkar et al [4], have compared use of ANN, use of fuzzy logic, use of expert system and genetic algorithm and other techniques such as decision tree by previous literature aggregation and suggest that ANN performs better than MRA but its black box nature has led to reach different conclusions in the observations.

Existing solutions

The use of AI for residential value forecasting has been suggested in the literature from 1990s. [4]. Although Sri Lanka is lacking an automated land valuation system, many up and running, reliable solutions have been implemented in developed countries like New Zealand, England and Wales, USA etc. It is obvious with the well-structured digital data infrastructure of those countries, they can implement very accurate systems. Our intention is to identify the ways to use their underlying methodology in a suitable manner in Sri Lankan context.

1.Zillow Zestimate

Zillow is an online real estate database company that was founded in 2006, and was created by Rich Barton and Lloyd Frink, former Microsoft executives and founders of Microsoft spin-off Expedia. [5] Zillow.com supports United States of America (USA) and Canadian property listing. Zillow compliments that Zestimate provides forecast for 12 months with below accuracy rates.

Model	Average Absolute % Error	Improvement over Naïve
Naïve Forecast	7.35%	0%
County Forecast	6.47%	11.9%
Zestimate Forecast	5.84%	20.5%

Features :

- Estimates for 12 months

    Zestimate determines an estimation for 12 months for a house based on neighbourhood comparable houses. Accuracy of Zestimate depends on the amount of data used as the underlying approach is Hedonic regression analysis based proprietary algorithm [6] which

analyses of several features of the house. The forecasted value is interpolated using cubic spline to connect to current value. [6]

2.Trulia

Trulia is also a product offered in USA, which offers a range of services for real estate sector. The price estimates are based on publicly available information the home's physical characteristics (e.g. location, number of bedrooms, etc.), Property tax information, Recent sales of similar nearby homes.

It involves more community interaction, for example, Trulia Neighbourhoods provide photographs, drone footage, etc. so that who are interested about the neighbourhood can refer. Trulia provides price using public data which shows the price fluctuation of a house, comparative to the other homes with same ZIP code.

Below is the accuracy report of Trulia estimates.

National	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
United States	48.2%	67.7%	82.3%	5.3%

Features -
- Crime map - Crime map data is sourced from CrimeReports.com and SpotCrime.com, which aggregate crime data from law enforcement agencies and news reports.
- Local schools with schools rating - Data of the schools around the premises with details such as Grades taught, GreatSchool Score.
- Commute times at a glance - Using data from OpenStreetMaps and General Transit Feed Specification (GTFS) feeds, the user can get an idea of commute times at a glance.[7]

3.QV.co.nz - QV homeguide

Quotable Value (QV) provides independent and authoritative information on any home in New Zealand on or off the market [8] QV.co.nz and their mobile App QV homeguide is known to be providing more accurate values of real estate property and key details to assist people to make instant decisions regarding property. QV with CoreLogic, a company which analyzes information assets and data to provide clients with analytics and customized data services provide a range of reports valuable to the user.

Features - QV homeguide app

- Online Value Estimation - Provides the likely selling price of a property during that particular time
- Sales activity - Sales activity specific property found on the app
- Suburb Demographics - Median price data, Demographic data, Current listings, and latest auction results [9]
- E-Valuer Report - Subjected to a fee complete valuation report of the property can be downloaded.

4.HousePrice.ai

Creating a methodology that would bring more sophisticated information, greater accuracy and analytical rigor to the United Kingdom (UK) residential property market is the motivation behind HousePrice.ai. Their proprietary model provides a combination of multi-disciplinary experiences of AI and Big Data to provide most accurate estimations. HousePrice.ai has Horizon app, which calculates capital, rental and gross development values for a single property or an entire portfolio. [10]

Features-

- Current and Future value prediction - Produces accurate property valuations both in the present time and can offer future predictions. Valuations are based on objective measurable values, creating a fact-based result as opposed to a subjective one [11]. This tool allows the user to adjust, add and remove factors within the surrounding areas to determine how external changes will affect property prices
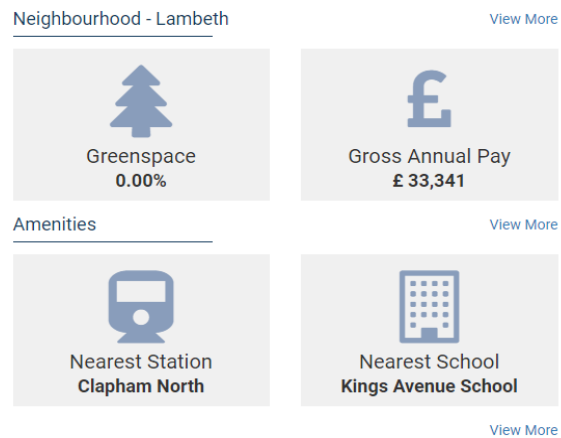- Distance to Schools, commutes etc.

Neighbourhood - Lambeth   View More

Greenspace
0.00%

Gross Annual Pay
£ 33,341

Amenities   View More

Nearest Station
Clapham North

Nearest School
Kings Avenue School

View More

Figure 1.2: Brief Neighbourhood analysis

Source : *Sample Valuation Report - HousePrice.ai , Horizon*

https://myhorizon.io/valueReport?id=59ddcdc7a699d278745b81e1

- Historical data relevant to location
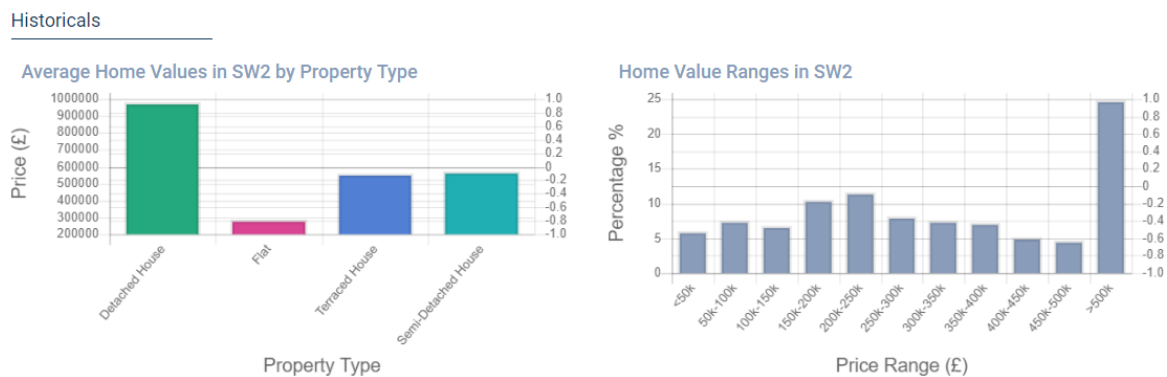
Figure 1.3: Historical Sales analysis

Source : *Sample Valuation Report - HousePrice.ai , Horizon*

https://myhorizon.io/valueReport?id=59ddcdc7a699d278745b81e1

Comparison of Existing Systems

	Zillow	Trulia	QV-CoreLogic	HousePrice.ai	Our Product
Current Value Prediction	Yes	Yes	Yes	Yes	Yes
Provision of possible natural hazards of the area	No	Yes	No	Yes	Yes
Use of machine learning algorithm	Yes	Unknown	No	Yes	Yes
Data used for predictions	Statistical data	Publicaly available data	Statistical data	Publicaly available data	Data provided by Valuers, and relevant departments
Mobile / Desktop / Web Application	Web	Web	Web	Web	Web
Available for Sri Lanka	No	No	No	No	Yes

Table 1.3: Comparison of existing systems

Research Gap

During the AI Asia Summit 2018, the summit panelists Dr. Yasantha Rajakarunanayake, Dr Rukshan Baduwita , Dr. James Shanahan and Dr. Chrisantha Fernando agreed that Sri Lanka is behind in terms of AI startups[12], despite the fact software industry is vastly growing area. According to the survey conducted under research done by Karunanda *et al*[13], carried out in 2014, this is due to the lack of popularity, knowledge, experts, requirements and sponsorship for the AI related software projects[13].

But when analyzing local news, we can see that AI based applications has become a trend. For Example, Dialog has its own AI powered voice service to support its product service framework. There are researches that have been conducted to predict the Stock prices of Sri Lanka with the usage of Machine Learning approaches as well as Deep Learning approaches, tilted *A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market*[14], and *Comparison of Support Vector Regression and Artificial Neural Network Models to Forecast daily Colombo Stock Exchange*[15].According Li *et al,* [16]to the real estate valuation researches evaluating the use of GIS technology have been conducted. But there is no information regarding application of AI technology or machine learning in real estate value prediction in Sri Lankan context.

There are numerous factors which affect the value of a land such as physical factors, economic factors, and social factors. The details of the physical factors and economic factors along with recent valuation details will be considered in this application when deciding the terminal output, prediction of the current value. In this application, the effect of social factors will not be taken into consideration as much as above mentioned because of the unavailability of proper information infrastructure to analyze the same factors. But that will not affect the accuracy of the output since recent valuations are taken into consideration.

Since machine learning and deep learning algorithms which were proven to be suitable for real estate valuation will be used accuracy of the predictions can be guaranteed.

Research Problem

The main research problem is to develop an automated system to evaluate the land based on its neighbourhood economical value and identify the possible effects of development work on the value of the land in the future. This requirement of a solution to predict the current value and future

value came from an expertise. While reviewing the literature, by means of supervisor meetings, we identified another aspect as an improvement, which is to predict the effect of future development work on a particular land, since Sri Lanka is a developing country, although the rate of development may vary, infrastructure development projects are carried out frequently.

We can never underestimate the duty of a valuation officer as the estimations are affected by numerous factors of particular to the area. But these factors are subjected to perception of each other's experience, according to Vaz J.[17], the discretionary and the appraisers' subjectivity that characterize traditional real estate valuation are still allowed to take part in the formation of the asset price even when respecting international standards (EVS, IVS) or Appraisal Institution´s regulations (TEGOVA, RICS, etc.). For example, an experienced valuer who is familiar with the area maybe biased towards the effect of regional factors, social factors, than the physical factors compared to a fairly new valuer who still sticks to the land valuation theories and follow the proven procedure. Therefore, manual valuation can be considered as a more sensitive approach.

Our intention is to provide people with fair accurate prediction of the land they are going to buy, so that they can decide the investment is fruitful for them. We believe this is an area improvement is needed because we can assist people in making decisions related to property, which would be the largest investment most probably in many people's lives.

Objectives

The goal is to assist people by providing them with accurate valuation, facts about how the land is going to be affected by various means of development projects, ultimately to decide whether it would be useful for their expected purpose.

Main objectives

The main objective of our research is to develop a portable application which can provide instant report of a selected land parcel which can provide the users with an insight of the land with current value and future value.
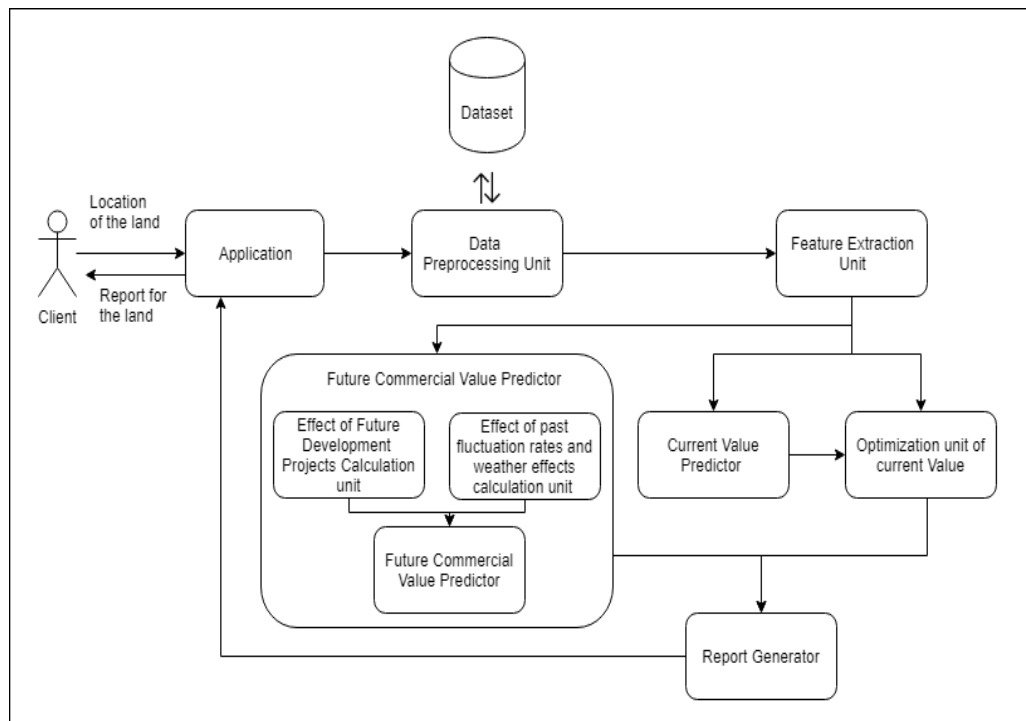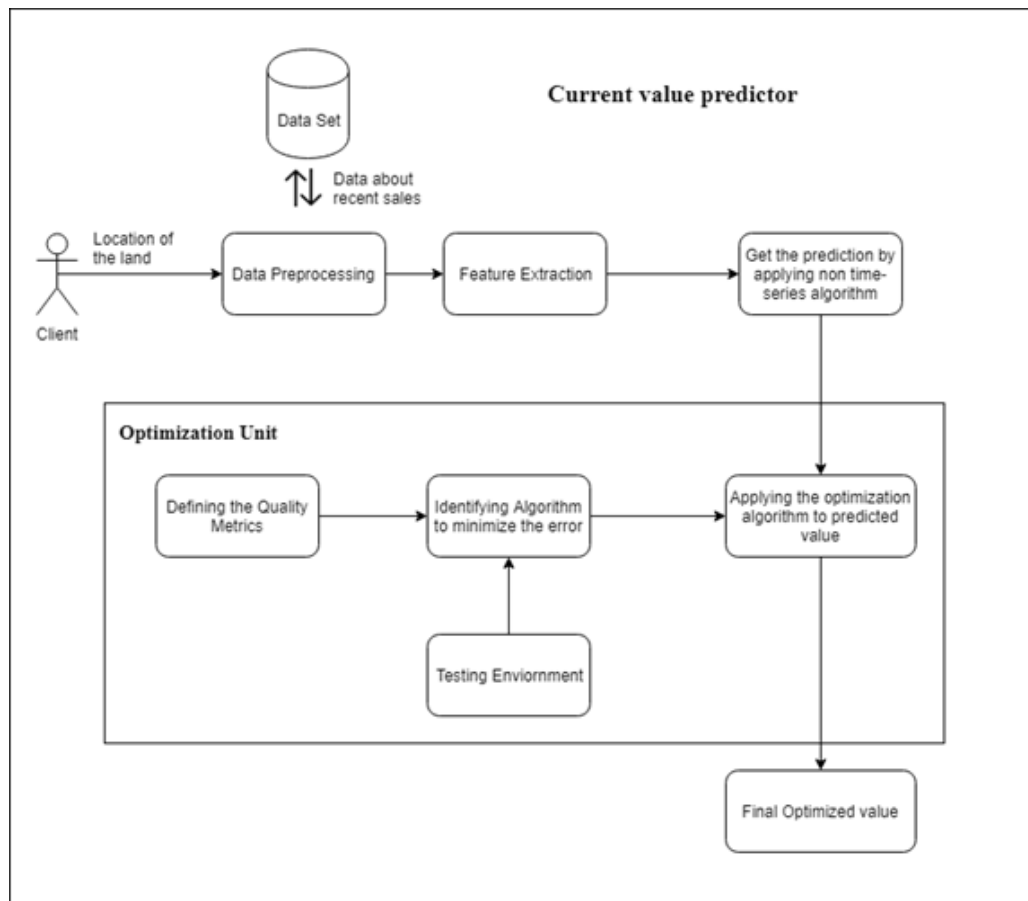
- Identifying the most accurate cross-sectional algorithm from conventional Multiple Regression Analysis (MRA) and non-conventional Artificial Neural Networks (ANN) in the domain of providing values in the domain of current value prediction following the Sales Comparison Approach

- Identifying the most accurate time-series algorithm from Long Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Auto Regressive Integrated Moving Average (ARIMA) model in the domain of providing values in the domain of current value prediction following the Sales Comparison Approach

- Identifying method to predict future value based on the fluctuation rates and records of weather conditions.

- Identifying the effect of proposed development plans on the future price of the selected land plot

- Creating a concise yet complete report based on the selected land plot which can be used to assist in making smarter property related decisions.

RESEARCH METHODOLOGY

System overview

When a customer goes to a land he is willing to buy, they can input the current location through the application. Based on that location, the suitable recent sales data is selected. Then those data will be analyzed by the AI model to predict the current value. That predicted value is optimized to produce the most accurate current value. Then the future value will be predicted by collaboration of two units, one which considers the fluctuation rates of past pricing values and weather effects, while the other calculates the effect of proposed development projects in the area. All these units generate a report which depicts these two types of data with relevant other data in a simpler way anyone can understand.

Current value predictor

Data Set

Data about recent sales

Client

Location of the land

Data Preprocessing → Feature Extraction → Get the prediction by applying non time-series algorithm

Optimization Unit

Defining the Quality Metrics → Identifying Algorithm to minimize the error → Applying the optimization algorithm to predicted value

Testing Enviornment

Final Optimized value

Methodology

This component is responsible for developing a machine learning model which predicts the current price of a land parcel upon submission of the location based on data gathered by analyzing those submitted by valuers and other key factors identified as significant to the area, following a time-series algorithm. One of our specific objectives is to test the accuracy of time-series algorithms and vice versa in the domain of land value prediction. Here two time series algorithms, namely LSTM and ARIMA models were tested and evaluated in terms of Mean squared error(MSE), Mean Absolute error (MAE), and root mean squared error (RMSE) to select the best performing algorithm in the domain of current value prediction based on cross sectional data.

First phase of developing the module included selecting best algorithm to predict the current value of a land. For that, Long Short Term Memory (LSTM) and ARIMA model were tested.

Next phase of the current value prediction unit is developing the API and the imputation module to estimate the missing factors necessary for the flask module in addition to location, to perform the prediction.

Finally, the most important component in terms of commercialization is developed. That is the data collection unit by the valuers.

Data collection

The study focuses on Colombo which experienced relatively high infrastructure development.

Primary data have been collected through questionnaires, interviews and personal visits to land area to know the present situation of the market and the secondary data are collected mainly through various survey department, land estate agents, newspaper advertisements, and land sale website contents. The data are useful for assessing the performance of property as a key to predict land price.

The time series data collected to predict the current value from a land sale company which had monthly land values from the same area over a period of 10 years, containing above 200 samples.

Though the true values has been acquired, the values of the lands may be vastly different when evaluated by a professional valuer.

For that, we have implemented a data collection unit to gather data from the valuers, which include a valuer rating module based on the number of feedbacks given by him or her, and those ratings will be utilized by the banks, real estate buyers to select an experienced valuer.

Long Short Term Memory (LSTM) – Recurrent Neural Network

Considering the fact that time has a direct influence on land prices time-series algorithms were also tested for selecting best prediction model for current price. What makes LSTM different from typical neural network is that it has feedback connections.

To test this model, timeseries dataset having monthly land values from the area over a period of 10 years was used. The dataset was having lags of unknown duration, hence, out of available RNN types, LSTM was the best option.

ARIMA model

ARIMA standing for Auto Regressive Integrated Moving Average is the most popular and commonly used statistical method for time series prediction. This model was utilized in both current value prediction and future value prediction units.

Procedure to follow with this model is 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.

The ARIMA forecasting for stationary time series is nothing but linear equation(like linear regression). The predictor depends on (p, d, q) of Arima model.

Time series linear equation: $x(t) =$ alpha $x(t-1)$error(t)

The ARIMA model dependent on following components.

1. Number of AR (Auto regressor) term **(p)**: AR term is lag of dependent variable. If p is 3 then predictor for x(t) will be x(t-1)..x(t-3)

2. Number of MA (Moving Average) term **(q)**: MA term is lag of forcast error of predictor equation. If q is 3 then error for x(t) will be e(t-1)..e(t-3)

3. Number of Differences **(d)**: The number of times that the raw observations are differenced, also called the degree of differencing.

- To determine p and q we will use two plots

  1. Auto Correlation Function **ACF**: It is a measure of correlation between TS and lagged of TS (q)

  2. Partial Auto Correlation Function **PACF**: This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons.(p)

Implementation and testing

The current price prediction system was implemented using an ensemble model of MLR and time series ARIMA model which had the accuracy of above 0.75.
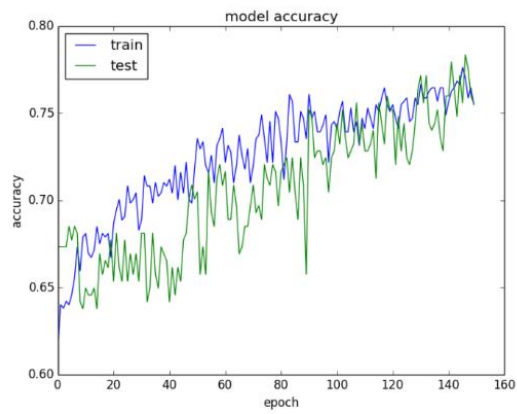
Figure : Accuracy curve for ensemble model

Tools and Technologies

Hardware interfaces

For the developer end, a computer with

- CPU: Quadcore Processor

- RAM: At least 8 GB

- Storage: 1 TB

Software interfaces

- Database

  For creating our application database we use MongoDB.

- Miniconda

  We use Miniconda as the application launcher. It allows us to launch applications and easily manage conda packages, environments and channels without the need to use command line commands.

- Jupyter Notebook

  Jupyter notebook is an IDE we have used to develop our machine learning models and it is powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features.

- PyCharm

  Python development IDE by Jetbrains

RESULTS AND DISCUSSION

Results

Comparison of models

As mentioned above, The models tested have been evaluated in terms of MAE, MSE, and RMSE.

The results can be summarized as follows.

	MAE	MSE	RMSE
LSTM	12150.774	1834424960	42830.187
ARIMA	26549.4523	4559474.12	2135.29251

Out of the two models ARIMA model found out to be having the best performance.

Below are the PACF and ACF functions of the ARIMA model.
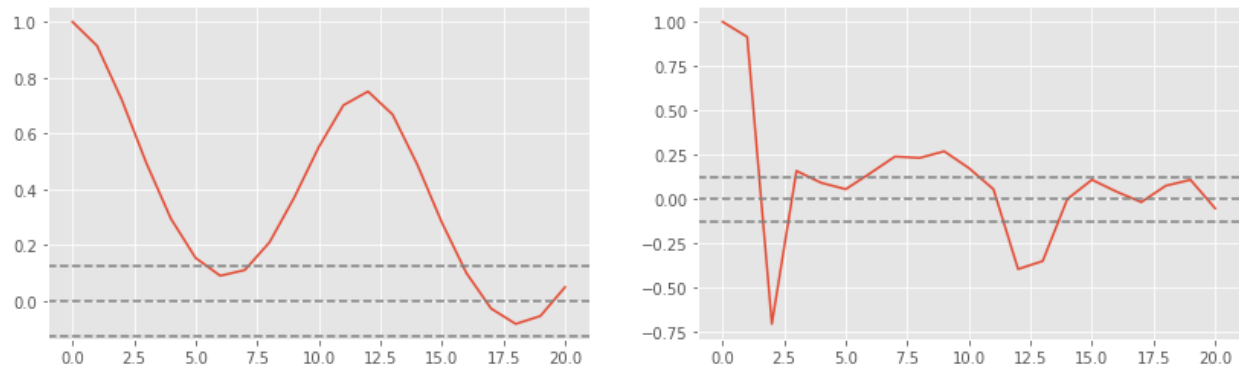
The dotted lines in confidence interval, this can be used to determine **p** and **q**.
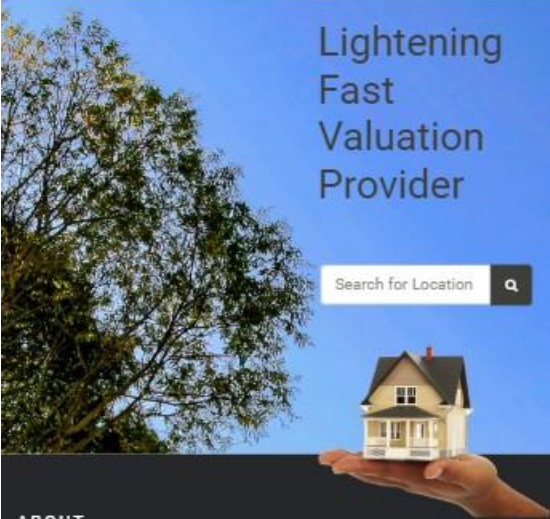
- p: The lag value where the **PACF** chart crosses upper chart for first time.

- q: The lag value where **ACF** chart crosses upper chart for first time. Here p = 5, q = 1, order = (5,1,1)

User Interfaces

Register

Name

Enter Name

Username

Enter Username

Email

Enter Email

We'll never share your email with anyone else.

Password

Enter Password

SUBMIT

Price Prediction Per Perch

Colombo Dock Yard Quarters Road

LKR1,856,915.00

A highly residential area with all the facilities.
Alsosuitable for commercial purposes.

Contribution

- John Doe

★ ★ ★ ★ ★

- Micheal Styles

★ ★ ★ ★ ★

- Tom Finn

★ ★ ★ ★ ★

Discussion

CONCLUSION