Land Price Detection

(Using Spatial Data and Machine Learning techniques for land price detection)

Anushk Naval
(18046)
Dept: Earth and Environmental Science
IISER Bhopal
Madhya Pradesh, India
(anushk18@iiserb.ac.in)

Shirshakk Purkayastha
(18247)
Dept: Electrical Engineering and Computer Science
IISER Bhopal
Madhya Pradesh, India
(shirshakk18@iiserb.ac.in)

***Abstract*—The aim of this research project is to study the land prices over the years, taking into account the spatial dependencies of locality, accessibility to amenities of hospitals, schools and main roads; and predict locality-based land prices for the upcoming years.**

Land price are dependent on various factors, the purpose of the project is to link these factors to enhance and improve the land price prediction abilities for future values of the land, leveraging the power of machine learning algorithms.

***Index Terms*—land price prediction, Geo AI, machine learning, urban planning, price prediction, spatial machine learning**

I. Introduction

Land price prediction is a complex process and it is diversely dependent on numerous factors such as accessibility to amenities of hospitals, schools and main roads. Locality, crime rates, traffic noise monetarily affects the land pricing.

With the development of Geo-spatial Artificial Intelligence (Geo AI), a field that incorporates machine learning models and data mining approaches to spatial science, a large number of spatial big data based problems have been solved, including autonomous vehicle mapping routes, weather and natural calamity forecasting, environmental monitoring and its most important contribution is in the domain of urban planning.

Future land price predictions plays important role and are necessary as the study on land prices are helpful for individuals, real estate companies and also support the decisions in urban planning and widely helpful with investors, appraisers, tax assessors, as there are large investments at stake.

II. Related work

A. Estimating Land prices on the basis of Spatial Statistics

This research developed insights into developing a pioneer method of Spatial auto-correlation for predicting the real estate prices of land in Korea. It provided insights to using and correcting spatial auto correlation errors that studies earlier had missed out on, leading to biased results and as a result, distortion in estimated prices.

It incorporated the spatial econometric model into account and subsequently reduced auto-correlations, while using real transaction prices. This study applied Spatial Error Modelling and Spatial Lag model to grasp the spatial dependence represented by spatial auto-correlation. [2]

B. Hedonic Regression analysis

The Hedonic pricing model is one of the advanced valuation techniques that can be used to estimate the impact of housing characteristics on prices.The analysis is based on cross-section and time-series housing transaction data of about 103,730 observations covering the period of January 1990 to December 2008. It has been established from the analysis that: the number of floors, public rooms, bedrooms, bathrooms, showers and WC; time-on-the-market; condition and type of property; and availability of glazing, garden, garage and central heating all influence house prices in Liverpool city. [3]

III. Objectives

The aim of this study is to identify and understand the the spatial factors that play a major role in deciding the prices of land in upcoming years and forecast these expected prices, by analysing the previous price trends throughout the previous decade and to accurately predict the Ready Reckoner Rate (RRR) for the given land in the upcoming years.

IV. Dependencies

The python Machine Learning models used for the Land Price Prediction based on Spatial data relationships has the following dependencies:

- GDAL
- GeoPandas
- Matplotlib
- Numpy
- Pandas
- Seaborn

V. Methodology

In this section, we propose our approach to solve the problem of land price prediction by using machine learning models on spatial data. As the model is the process of developing a system through analysis of price trend over the previous years and then integrating it with the spatial information, the problem will be approached as follows:

S1: The data will be cleaned and processed, and then merged into geopandas dataframe objects. There are missing values for the year which will be filled in by taking the mean of the two neighbouring years.

S2: The processed data will then be used to generate machine learning as well as deep learning models, providing us the trend of the price respective to different land typology.

S3: The spatial features will then be incorporated by building up the relation between accessibility to amenities and the price trend.

S4: The model will be then enhanced using ensemble methods and feature selection operations.

S5: An output of prediction file and best performing model will then be obtained.

VI. Data Set Description

The dataset provided contains the land prices of different polygon demarcated land areas in Mumbai, collected from the year 2006 to 2016, with the prices of 2010 missing.

The land areas in the dataset were classified into the following categories:

- Residential
- Commercial
- Industrial
- Open Area
- Commercial (Type 1)

VII. Missing Data Problem

The dataset suffered from the missing dataset problem , i.e. the prices for the year 2010 were missing from the otherwise continuous dataset for the years 2006-2016.

The missing dataset problem was solved by averaging out the prices for the year 2009 and 2011 to get the prices for the year 2010.

$$Value_{2010} = \frac{Value2009 + Value_{2011}}{2}$$

VIII. Models

The given problem of land price detection is approached via the following methods:

- Random Forest (1 model without feature selection and 2 models after feature selection).
- Neural Network (3 models: first model with no hidden layer, second model with less hidden layers but more parameters and the third model with more hidden layers but lesser parameters).

A. Random Forest Models

Three Random Forest models are developed:

- First Model without feature selection.
- Second model with Correlation Statistics feature selection technique.
- Third model with Mutual Information Selection feature selection technique.

Feature Selection: Numerical Feature Selection is used to reduce the number of parameters for our models.

Two of the most popular feature selection techniques applied to numerical input data and a numerical target variable are used for our models:

1) **Correlation Statistics** Correlation is a measure of how two variables change together. Perhaps the most common correlation measure is Pearson's correlation that assumes a Gaussian distribution to each variable and reports on their linear relationship.
   Linear correlation scores are typically a value between -1 and 1 with 0 representing no relationship. For feature selection, we are often interested in a positive score with the larger the positive value, the larger the relationship, and, more likely, the feature should be selected for modeling. As such the linear correlation can be converted into a correlation statistic with only positive values.
   The Scikit-learn machine library provides an implementation of the correlation statistic in the f_regression() function. This function can be used in a feature selection strategy, such as selecting the top k most relevant features (largest values) via the SelectKBest class.
2) **Mutual Information Statistics** Mutual information from the field of information theory is the application of information gain (typically used in the construction of decision trees) to feature selection.
   Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.
   The Scikit-learn machine learning library provides an implementation of mutual information for feature selection with numeric input and output variables via the mutual_info_regression() function.

B. Deep Learning Model:

Three Deep Learning models were developed with varying number of hidden layers and parameters:

- **Model 1:** Comprised of 5 layers, 1 for input and 1 for output with 3 hidden layers.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 3324)              349020
_________________________________________________________________
dense_1 (Dense)              (None, 831)               2763075
_________________________________________________________________
dense_2 (Dense)              (None, 2216)              1843712
_________________________________________________________________
dense_3 (Dense)              (None, 831)               1842327
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 832
=================================================================
Total params: 6,798,966
Trainable params: 6,798,966
Non-trainable params: 0
_________________________________________________________________
```

- **Model 2:** Comprised of 2 layers, 1 for input and 1 for output.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 3324)              349020
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 3325
=================================================================
Total params: 352,345
Trainable params: 352,345
Non-trainable params: 0
_________________________________________________________________
```

- **Model 3:** Comprised of 12 layers, 1 for input and 1 for output with 10 hidden layers.

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_7 (Dense)              (None, 3324)              349020
_________________________________________________________________
dense_8 (Dense)              (None, 416)               1383200
_________________________________________________________________
dense_9 (Dense)              (None, 416)               173472
_________________________________________________________________
dense_10 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_11 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_12 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_13 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_14 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_15 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_16 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_17 (Dense)             (None, 416)               173472
_________________________________________________________________
dense_18 (Dense)             (None, 1)                 417
=================================================================
Total params: 3,293,885
Trainable params: 3,293,885
Non-trainable params: 0
_________________________________________________________________
```

IX. Plots

Pricing for year 2007

Pricing for year 2016

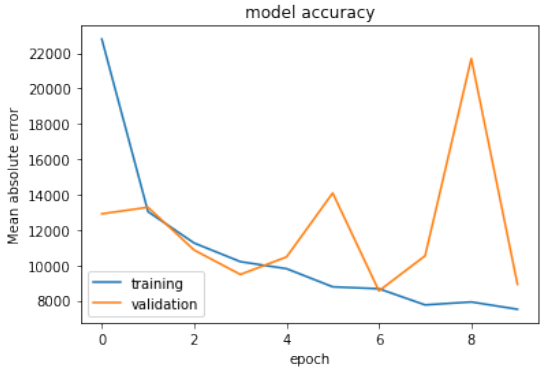
There is increase in property value over the years which is shown in the heat maps above comparing 2 years, 2007 and 2016.

X. Results

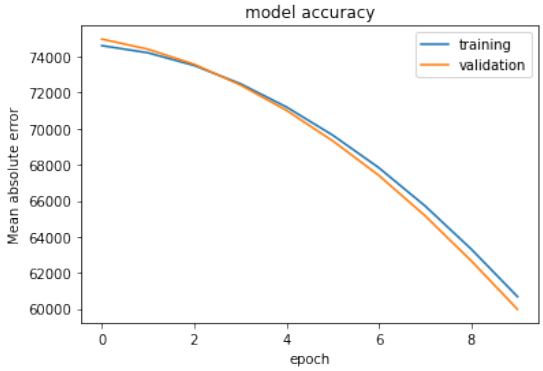
A. Model Accuracies:

1) Deep Learning Models:

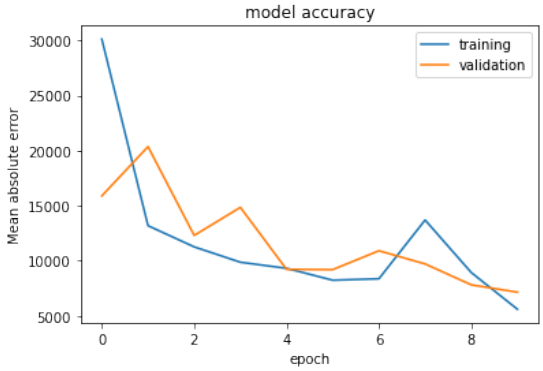
Model 1: 5 Layers



Model 2: 2 Layers



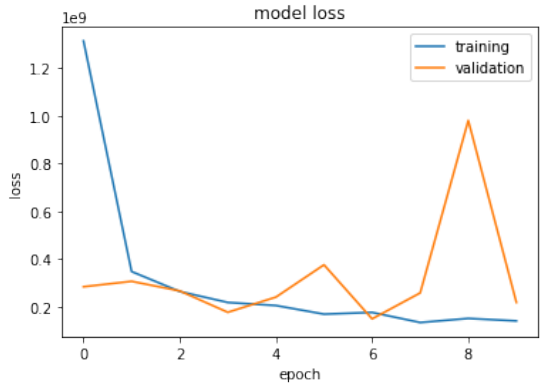
Model 3: 12 Layers



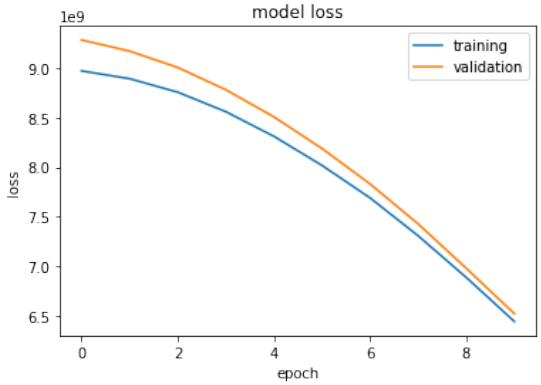
B. Model Losses:

1) Deep Learning Models:

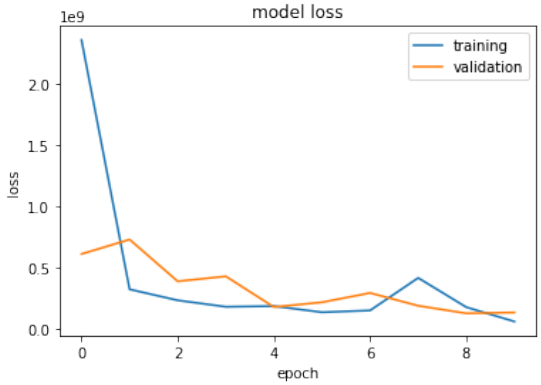
Model 1: 5 Layers



Model 2: 2 Layers



Model 3: 12 Layers



XI. Conclusion

A. Model Discussion:

Although model 1 has more parameters, increase in the number of hidden layers increases the model accuracy and gives a higher score while the model with no hidden layers deviates from learning as the input is more complex in nature to model.

Accuracy Matrix for Open Land

	MAE	RMS	VS	R2
Random forest	1740.849	9130421.708	0.996	0.996
Correlation statistics selection and Random forest	1524.332	7723790.981	0.997	0.997
Mutual information statistics and Random forest	1578.366	8104280.358	0.997	0.997
Hidden layer	7269.595	132005639.5	0.956	0.954
No Hidden layer	58768.59	5734708468	0.218	-0.962
More Hidden layers	7775.978	129703996.4	0.962	0.955

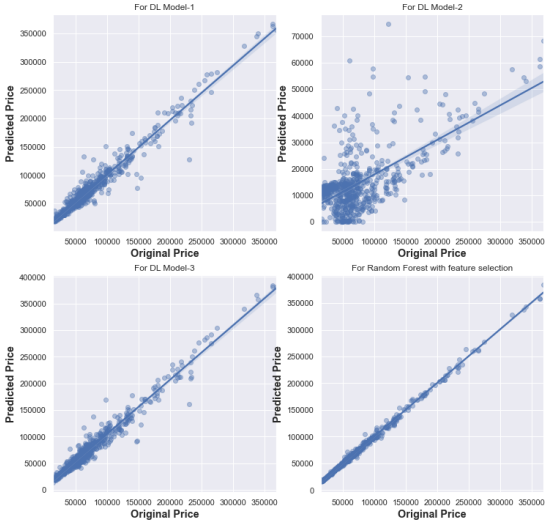
Accuracy Matrix for Commercial Land -01

	MAE	RMS	VS	R2
Random forest	5973.687	87958333.41	0.958	0.958
Correlation statistics selection and Random forest	5413.783	837438717.5	0.96	0.96
Mutual information statistics and Random forest	5271.192	791704503.2	0.962	0.962
Hidden layer	37655.16	4367234923	0.825	0.792
No Hidden layer	201903	59813980935	0.093	-1.841
More Hidden layers	26555.34	3612966727	0.828	0.828

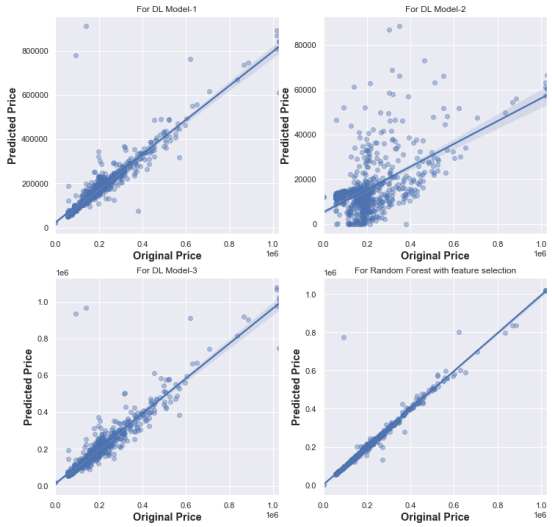
The above matrices shows that increasing hidden layer thus increase the accuracy of model but compared to ml technique the R$\hat{2}$ value is still less which can be answered by increasing layers and epochs.

We observed that using feature selection we can substantially decrease the data as there are many variables that can be removed from the picture to provide near same results with less computations but these features also depends on the property type.

Accuracy Plot for Commercial Land -01

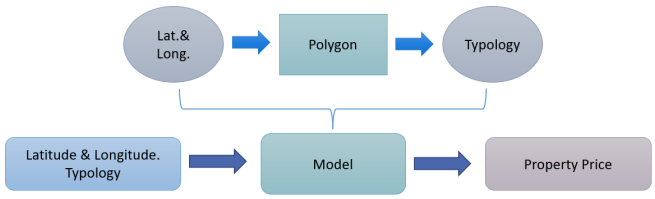


Accuracy Plot for Commercial Land -01



The plots shows the relationship between predicted and the original value of the land.

Accuracy assessment for Ready Reckoner Rate (RRR) is proposed to be metricised on the basis of the given flowchart:



The accuracy matrix will be computed against the rental values for the land in different areas versus the predicted values obtained from the model.

XII. Contributions

- **Anushk Naval:** Visualization, Accuracy Matrices, Deep Learning Models and Generalization.
- **Shirshakk Purkaystha:** Data Processing, ML Techniques (Random Forest Models and Feature Selection).

XIII. Codes

The associated codes and python notebook file for this project can be found at : ***Github-link***.

References

[1] A Study on Estimation of Land Value Using Spatial Statistics: Focusing on Real Transaction Land Prices in Korea, Bongjoon Kim and Taeyoung Kim, DOI: 10.3390/su8030203 .

[2] Land Price Prediction System using Case-based Reasoning, Minkyu Choi, Taeya Yi, Meereh Kim and Ji-Hyun Lee .

[3] Hedonic regression analysis of house price determinants in Liverpool, England,January 2011, Raymond Abdulai, Anthony Owusu-Ansah .

[4] Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods, Steven C. Bourassa, Eva Cantoni and Martin Hoesli

[5] Land Price Prediction using Artificial Intelligence, Shanmuga Sundari N, Rakshana Devi R, Subasri I.