Real Estate Data Analysis using Principal Component Analysis and ‘R’

K Sailaja Kumar, Kameshwari Soundarya, Harshitha R, D Evangelin Geetha and T V Suresh
Kumarasailajakumar.k@gmail.com, kameshwari.soundarya@gmail.com, harshitha3072@gmail.com, devangelin@gmail.com.
Department of Computer Applications, M S Ramaiah Institute of Technology, Bangalore, India.

Abstract—The primary focus of real estate data analysis is on analyzing the past data and then predicting the real estimates in further time. This analysis is used to figure out the right price for the property in view of buyer and seller. In this paper a methodology is proposed to estimate the real estate housing price based on various predictor variables using regression models. The models are then implemented using ‘R’ statistical tool. It is used to make data easy to explore and visualize. Principal component analysis (PCA) is used to emphasize the variations and bring out strong patterns in the real estate dataset. The accuracy of these three regression models M1, M2 and M3 are 0.3907, 0.3467 and 0.5825 respectively. From these accuracy values it is observed that the model M3 is giving better results in comparison to M1 and M2. Moreover M3 is constructed using the results obtained from the PCA technique employed in M2. PCA is identified as the powerful data science technique for predictor variable reduction in particular for real estate dataset attributes reduction analysis. Hence this technique is adopted for estimating the real estate property price accurately. Further the model can be extendable to other real estate datasets in future.

I. INTRODUCTION

The economic decline had a significant impact on the real estate community. One of the challenges with the real estate data analysis was to figure out the right price for the property in view of buyer and seller. The analysis of the real estate dataset then identifying the most influential parameters that effect the house price is a very challenging task due to the high dimension of the dataset. Therefore in this paper, we present the best model to predict the house price with good accuracy and will help the customers to buy a house considering the influential factors.

This paper is focused specifically on Washington City’s real estate data, with 159 rows and 330 columns for instance, the total-floors-in-building, price-per-sqft, number-of-beds, number-of-bathrooms and many more. This data is prepared for generating the regression models after handling the missing values. Using principal component analysis (PCA) the significant predictor variables (columns in the real estate dataset) are identified which are useful in predicting the house price. PCA is commonly used to handle large datasets associated with the social sciences, market research, and other communities. It uses orthogonal transformation to convert a set of correlated predictor variables (columns in the real estate dataset) into a set of linearly uncorrelated variables called ‘principal components’ or ‘predictor variables’, from the real estate dataset. The goal of PCA is to explain the amount of house price variation considering the fewer number of these principal components.

In this paper, three regression models M1, M2 and M3 are constructed to identify the prevailing predictor variables used to estimate the housing price in view of owner of the house and the real estate agent. These models were developed based on regression analysis, principal component analysis (PCA) and with the selected original raw variables using R for the real estate price estimates respectively.

First regression model M1 is constructed from the training dataset considering the 101 numeric predictor variables. The target variable is house price. The performance of this model is evaluated using the test dataset and the accuracy of the model as 0.3907. This shows 39% price variation is due to these numeric predictor variables.

The second regression model M2 is based on PCA. Using this technique principal components Dim1 to Dim6 from the numeric predictor variables are extracted. These extracted principal componentso do not have any multicollinearity between them as they are orthogonal or perfectly uncorrelated. Using these components regression model is developed. The performance of this model is evaluated using the test dataset. The accuracy of this model is 0.3467. Approximately 35% of the variation in the house price is due to these principal components.

The third model M3 is built from model M2 using the selected numeric predictor variables based on their significance w.r.t. principal components Dim1, Dim2 and Dim3. The performance of this model is evaluated using the test dataset. The accuracy of this model is 0.5825. Here 58% price variation is there due to these selected significant predictor variables.

There is a marginal improvement in the accuracy from first model to the third model. The second model explains only 35% of the price variation due to the complexity involved in the PCA technique. The model M3 is simple and gives more accurate results in comparison to other models M1 and M2. More over this model is based on the PCA results obtained in model M2. Thus PCA is a highly intuitive and powerful data science technique that can be used to construct predictive models to estimate the real estate property price.

‘R’ is a software useful for statistical computing and visualization. It has become the most popular and versatile language for data science. It is an interpreted language and comes with a command line interpreter - available for Linux, Windows and Mac machines - but there are IDEs
like RStudio or JGR to support development. RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting. The above mentioned models are implemented using R.

The paper is further organized as follows. In Section II literature on real estate data analysis is presented. Section III provides the methodology to collect and examine the real estate data and developing models to identify the influential factors for property price prediction. In Section IV the methodology illustration is provided. Experimental study that is carried out focusing the analysis of real estate dataset of Washington City, and results obtained using the proposed methodology is discussed in Section V. The conclusion and future work are presented in Section VI.

II. LITERATURE

Prediction of house prices based on NJOP houses in Malang city with regression analysis and particle swarm optimization (PSO) is discussed in [1]. PSO is used for selection of affect variables and regression analysis is used to determine the optimal coefficient in prediction. The result from this research proved combination regression and PSO is suitable and get the minimum prediction error.

Modeling and forecasting of land price in Chennai Metropolitan Area (CMA) in the state of Tamilnadu, India using multiple regression and neural network techniques is presented in [2]. Thirteen locations spread over CMA are selected at random as study areas. Both multiple regression and neural network models are validated and used to forecast the land price in CMA. Both the models are found to be well fit for the trend of land price; however the model using neural network shows better accuracy.

Regression models, using various features to have lower Residual Sum of Squares error are described in [3]. The predictive performance of artificial neural networks and multiple regression analysis for single family housing sales is compared in [5]. Multiple comparisons are made between the two data models in which the data sample size, the functional specification and the temporal prediction are varied. The authors have observed that ANN (Artificial Neural Networks) results better than MRA (Multiple Regression Analysis) when a moderate to large data sample size is used [4].

The functioning [5] involves a website which accepts customer’s specifications and then combines the application of multiple linear regression algorithm of data mining. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

Prominent factors affecting property value in Indian housing market are described in [6]. Total nineteen variables were identified from literature as well as discussion with experts. Data obtained from the survey was processed using PCA which helped to decrease the number of variables into seven most important factors affecting value of real property in Indian housing market. The seven most important factors that affect Indian real property value included living conditions of residents are identified by the authors as: characteristics of housing; regional influence; utilities; age of property; economic, political and social influence; area and legal aspect.

In this paper, a methodology is proposed to estimate the real estate housing price considering various predictor variables using regression models. The models are then implemented using ‘R’ statistical tool.

III. METHODOLOGY

The proposed methodology construct the regression models to predict the housing price. The steps to be followed are:

Step1. Load the real estate data into R
Step2. Identify and handle the missing values
Step3. Identify numeric and categorical predictor Variables
Step4. Identify training dataset and test dataset from the sample dataset for constructing the three regression models M1, M2 and M3.

Step5. Model M1:
   - Construct regression model with all numeric predictor variables from the training dataset
   - Evaluate the performance of the model using the test dataset

Step6. Model M2:
   - Construct regression model with the orthogonal principal components derived from the original predictor variable using PCA
   - Evaluate the performance of the model using the test dataset

Step7. Model M3:
   - Construct regression model from the selected significant predictor variables obtained from the significant principal components in Model M2
   - Evaluate the performance of the model using the test dataset

Step8. Compare the three models based on the accuracy levels and select the model with high accuracy.

Step9. Use the model selected in step 9 for predicting the housing price.

IV. ILLUSTRATING THE METHODOLOGY

Methodology presented in Section III is illustrated as follows.

A. Load the real estate data into R

Consider sample real estate dataset as .csv file with 932 rows representing the values under 7 columns (Variables) [7]. The screenshot is shown in Figure 1. The columns are designated as Distance-to-BT, Distance-to-Mall, Distance-to-Hospital, Carpet-Area, Built-up-Area, Parking-Status and House-Price. Load this dataset into R.
The eigenvalue is used for orthogonal principal components in principal component analysis to calculate the data value. The PCA graph shows categorical and numeric distance between the points. 

**B. Identify numeric and categorical predictor variables**

Sample dataset columns are classified into numeric and categorical predictor variables based on their properties. Out of 7 columns, 5 columns are identified as numeric variables (Distance-to-BT, Distance-to-Mall, Distance-to-Hospital, Carpet-Area and Built-up-Area), 1 is identified as categorical (Parking-Status) and house-price is identified as target variable.

**C. Identify and handle the missing values**

Missing value indicates that no data value is stored for the variable in the current observation. Missing data will have significant effect on the conclusions drawn from the data. Hence handling missing data is important for accurate predictions. In this dataset missing numeric variables are replaced with mean value and categorical variables are replaced with mode value.

**D. Identify training data and test data forregression Modeling**

Sample dataset is divided into training and test datasets. 70% of the sample dataset is considered as training dataset and 30% as test dataset. The training dataset is used to build the three regression models and the test dataset is used to evaluate the performance of the model.

**E. Construct the regression model M1**

Regression model M1 is built using the numeric and categorical predictor variables and house price as target variable. The results of the regression analysis summary is shown in Figure 2.

**Fig. 3. PCA Graph from R**

From Figure 3 it is observed that Distance-to-BT, Distance-to-Mall, and Distance-to-Hospital have formed principal component variable Dim1 which explains 45.15% information in data. Where in Carpet-Area and Built-up-Area explains the remaining 39.99% of information through Dim2 principal component variable. The eigenvalue is used in the principal component analysis to calculate the % variance as shown Figure 4. From Figure 4 it is observed that Comp1 and Comp2 explains 45.15% and 39.99% of variance which is high among all.

**Fig. 4. Eigen value and % of variance explained by 5 components, sample screenshot from R**

Regression model M2 is built using these principal components and categorical variable Parking-Status. The results of the regression analysis summary is shown in Figure 5.
From Figure 5 it is observed there are no insignificant numeric predictor variables in the model. The performance of the model is evaluated using the test dataset. The accuracy of this model is 0.01401. There is a slight improvement in the accuracy from model M1.

A correlation matrix constructed with the predictor variables and the principal components as shown in Figure 6. From the correlation matrix it is observed that 88% of the Distance-to-Hospital is loaded on Dim1 and 100% of both Carpet-Area and Built-up-Area is loaded on Dim2.

The system configuration used for the experimental study is described below:

Intel® Pentium® iV
RAM: 3.99GB, 64 bit operating system
Software tools: R Studio

A. Data Preparation

As the first step load the Washington City’s real estate dataset into R to build the model. Missing values under each attribute list are identified and replaced with appropriate values. Numeric and categorical predictor variables are classified in the dataset. 101 numeric predictor variables are identified out of 330, remaining variables are considered as categorical. The target variable or response variable is identified as house price. In the next step, the dataset is divided into training set and test set. The training set is used to build the model and the test dataset is used to evaluate the performance of the model. These datasets are created randomly by making 70% of data as the training dataset and the remaining 30% as the test dataset.

B. M1 Model Building

A regression model is constructed using 101 numeric variables as predictor variables from the test dataset and house-price as the target variable. The summary of regression model obtained using R is presented in Figure 8. From Figure 8, it is observed that some of the predictor variables used in the analysis are tagged with *, **, *** based on their correlation with the target variable.

The performance of the model is evaluated using the test dataset. The correlation between the estimated and the observed values of the house prices is calculated. The correlation value explains the level of accuracy of the model. The square of the correlation is the R-square value or the predictive power of the model. The accuracy of this model is 0.3907. This explains 39.07% of the variation in the house price is due to these predictor variables. It is observed that in this model most significant predictor variables are tagged as unimportant. In fact, these variables have high correlation with the target variable. This problem is addressed in the next model using PCA.

V. EXPERIMENTAL STUDY, RESULTS AND DISCUSSION

To address the expectations, needs and requirements of buyer and seller of real estate, an experimental study is carried out on Washington City’s real estate data, with 159 rows and 330 columns. The dataset is available in .CSV format. The sample snapshot of the data is presented in Figure 7.
The orthogonal principal components Dim1 to Dim6 are extracted from the 101 numeric predictor variables using PCA. These predictor variables do not have any multicollinearity between them as they are perfectly uncorrelated. PCA graph is shown in Figure 9. From the graph it is observed that principal component variable Dim1 explains 15.23% of data and Dim2 explains 13.33% of data. Using these principal components regression model is developed.
From the above three models accuracy results it is observed that there is a marginal improvement in the accuracy from the first model to the third model. The second model explains only 35% of the price variation. This is due to the complexity involved in the usage of PCA. The model M3 is simple and gives more accurate results in comparison to other models M1 and M2. More over this model is based on the PCA results obtained in model M2. Thus PCA is a highly intuitive and powerful data science technique that can be used to construct predictive models to estimate the real estate property price.

VI. CONCLUSION

In this paper, three regression models are developed and validated using R to estimate the real estate housing price based on various predictor variables. The accuracy of themodels M1, M2 and M3 are 0.3907, 0.3467 and 0.5825 respectively. From these accuracy values it is observed that the model M3 is giving better results in comparison to M1 and M2. Moreover M3 is a simple model constructed using the results obtained from the PCA technique employed in M2, which is identified as the powerful data science technique for predictor variable reduction in particular for real estate dataset attributes reduction analysis. This technique is also useful for constructing themodels to estimate the real estate property price accurately. In this paper only numeric attributes are considered for PCA to create models for analyzing and predicting the real estate housing price. In future there is a scope to extend this data model to support categorical attributes as well. Further the model can be extendable to other real estate data sets in future.

REFERENCES


