Scilab case study project on

Predictive Modelling of Housing Prices: A Linear Regression Approach in Scilab

Dibyani Mohanty

K J Somaiya College of Engineering

Machine Learning, Data Science

15 April 2024

Abstract

This project presents a predictive modelling approach implemented in Scilab for analysing housing prices. The workflow encompasses data preprocessing, model training using linear regression, performance evaluation, and prediction. The dataset undergoes preprocessing steps to handle missing values and normalize features. Utilizing the normal equation method, linear regression models are trained to predict housing prices based on various features. Performance evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, and Adjusted R-squared are computed to assess model accuracy. The project also provides functionality for users to input new data, preprocess it, and obtain predictions using the trained model. This implementation showcases the application of predictive analysis techniques in predicting housing prices, demonstrating the potential for similar analyses in real estate markets.

1. Introduction

The real estate market in India is a multifaceted landscape influenced by a myriad of factors, including location, property size, amenities, and economic conditions. As housing prices continue to evolve, understanding the underlying dynamics becomes paramount for homeowners, real estate professionals, and policymakers alike. Predictive modelling

techniques provide invaluable insights into market trends, enabling stakeholders to make informed decisions.

This case study presents an innovative approach to predictive modelling focusing on Indian housing prices. Leveraging the power of linear regression and employing Scilab as the primary analytical tool, this project delves into analysing and forecasting housing prices.

The dataset chosen for this study is curated from the Indian housing market data, focusing specifically on the Haryana region in India, particularly the Gurgaon area. The properties included in the dataset are situated within the geographical confines of Haryana, providing valuable context for understanding the local real estate market dynamics.

The dataset captures essential attributes that influence housing prices, such as location, square footage, amenities, and other relevant features. It encompasses houses, villas, duplexes, and bungalows, commonly found in the Haryana region. The House Price dataset comprises 23 columns and 14620 rows, representing a pre-processed version of the original data. This pre-processed dataset includes transformed or modified features to suit the model's requirements. For instance, certain features like the number of bathrooms may have been converted from floating-point values to integers based on user input, ensuring consistency and compatibility with the model. Additionally, we have scaled down the grade of the property from a range of 0-10 to 0-5 and modified data types accordingly. Here is the list of features that will be used by both the user and us:

- **id:** Unique identifier for each property. (Integer)
- **Date:** Date of the listing. (Date)
- **Number of bedrooms:** The count of bedrooms in the property. (Integer)
- **Number of bathrooms:** The count of bathrooms in the property. (Integer)
- **Living area:** Total area of the living space in square feet. (Integer)
- **Lot area:** Total area of the lot in square feet. (Integer)
- **Number of floors:** Number of floors in the property. (Float)
- **Waterfront present:** Binary indicator (0 or 1) indicating whether the property has a waterfront. (Binary)
- **Number of views:** Count of how many times the property has been viewed. (Integer)

- **Condition of the house:** An ordinal variable representing the overall state of the property, including factors such as wear and tear, maintenance, and structural integrity. The range is from 0 to 5, where 5 represents the best condition. (Ordinal)
    - Rating 0: Severely dilapidated, requiring extensive repairs.
    - Rating 1: Poor condition, showing signs of neglect.
    - Rating 2: Fair condition, with some wear and tear.
    - Rating 3: Good condition, well-maintained and structurally sound.
    - Rating 4: Very good condition, meticulously maintained.
    - Rating 5: Excellent condition, like new or newly renovated.
- **Grade of the house:** An ordinal variable representing the quality and features of the property, including aspects such as architectural design, water resources and overall aesthetics. The range is from 0 to 5, where 5 represents the best grade. (Ordinal)
    - Rating 0: Inferior quality, outdated design, no specific water source mentioned.
    - Ratings 1: Below-average grade, with basic amenities, no specific water source mentioned.
    - Ratings 2: Represents a modest grade, offering simple comforts and functionality and have access to water from borewell.
    - Ratings 3: Average grade, modern design, semi furnished and have access to water from borewell.
    - Ratings 4: Represents an above-average grade, featuring modern amenities, tasteful finishes and have access to water from municipality sources.
    - Ratings 5: Denotes exceptional grade, showcasing exquisite craftsmanship, luxurious features and have access to water from municipality sources.
- **Area of the house (excluding basement):** Total area of the house excluding the basement in square feet. (Integer)
- **Area of the basement:** Total area of the basement in square feet. (Integer)
- **Built Year:** Year when the property was built. (Integer)
- **Renovation Year:** Year when the property was last renovated. (Integer)
- **Postal Code:** Postal code of the property location. (Integer)
- **Latitude:** Latitude coordinate of the property location. (Float)
- **Longitude:** Longitude coordinate of the property location. (Float)

- **Living area renovation:** The area of the living space that has undergone renovation, measured in square feet. (Integer)
- **Lot area renovation:** The area of the lot that has undergone renovation, measured in square feet. (Integer)
- **Number of schools nearby:** Count of schools located near the property. (Integer)
- **Distance from the airport:** Distance of the property from the nearest airport in kilometres. (Float)
- **Price:** Price of the property in INR. (Integer/Float)

The target variable in this dataset is the price, which represents the monetary value of each property. This variable serves as the focal point for predictive modelling.

2. Problem Statement

In the busy world of buying and selling homes in Haryana, India, people want to know how much a house might cost based on its features. This project aims to create a tool using Scilab that can guess the price of a house by looking at things like where it's located, what's nearby, and how big it is. By using this tool, people can make smarter choices when they're buying or selling homes in Haryana.

The tool is specifically designed for estimating the price of standalone residential properties such as houses, villas, duplexes, and bungalows, rather than flats or building types. Attributes such as the condition of the house and the grade of the house are applicable to standalone residential properties only and are not intended for use with flat or building type properties.

3. Basic concepts related to the topic

Predictive analytics is a form of technology that makes predictions about certain unknowns in the future. It draws on a series of techniques to make these determinations, including artificial intelligence (AI), data mining, machine learning, modelling, and statistics. For instance, data mining involves the analysis of large sets of data to detect patterns from it. Text analysis does the same, except for large blocks of text.

Predictive models are used for all kinds of applications, including weather forecasts, creating video games, translating voice to text, customer service, and investment portfolio strategies.

All of these applications use descriptive statistical models of existing data to make predictions about future data.

Predictive analytics is also useful for businesses to help them manage inventory, develop marketing strategies, and forecast sales. It also helps businesses survive, especially those in highly competitive industries such as health care and retail. Investors and financial professionals can draw on this technology to help craft investment portfolios and reduce the potential for risk.

Linear regression is a basic and commonly used type of predictive analysis. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

$$y = c + (b \times x)$$

where y = estimated dependent variable score

c = constant

b = regression coefficient

x = score on the independent variable

There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are:

1) determining the strength of predictors
2) forecasting an effect
3) trend forecasting.

The regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. It can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. Regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. The intercept (sometimes called the "constant") in a regression model represents the mean value of the response variable when all of the predictor variables in the model are equal to zero.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Metrics used in this case study:

- **Mean Squared Error (MSE)**

  The Mean Squared Error (MSE) could be a metric utilized to assess the execution of regression models. It measures the average squared distinction between the anticipated and genuine values. The MSE is calculated by taking the sum of the squared differences between the predicted and actual values, dividing it by the number of perceptions, and after that taking the square root of the result. The lower the MSE, the superior the show is at predicting outcomes.

$$MSE = \frac{1}{N}\sum(y_i - \hat{y}_i)^2$$

  Where N is the number of observations and yi and ŷi are the actual and predicted values for observation i, respectively.

- **Root Mean Squared Error (RMSE)**

The Root Mean Squared Error (RMSE) could be a metric utilized to assess the execution of regression models. It is essentially the square root of the Mean Squared Error (MSE) and is regularly utilized as a more interpretable metric than MSE, because it is communicated within the same units as the target variable. The lower the RMSE, the superior the show is at anticipating results.

$$RMSE = \sqrt{(MSE)}$$

- **Mean Absolute Error (MAE)**

Mean Absolute Error (MAE): The MAE is another common metric used in the regression. It measures the average absolute difference between the predicted and actual values. The Mean Absolute Error (MAE) is calculated as follows:

$$MAE = \frac{1}{N}\sum |y_i - \hat{y}_i|$$

- **R-squared**

The R-squared (or coefficient of determination) is a measure of how well the regression model fits the data. It represents the proportion of variance in the target variable that can be explained by the predictor variables in the model. It is calculated as follows:

$$R - Squared = 1 - (\frac{SSE}{SST})$$

Where SSE is the sum of squared errors (or residuals) and SST is the total sum of squares.

- **Adjusted R-squared**

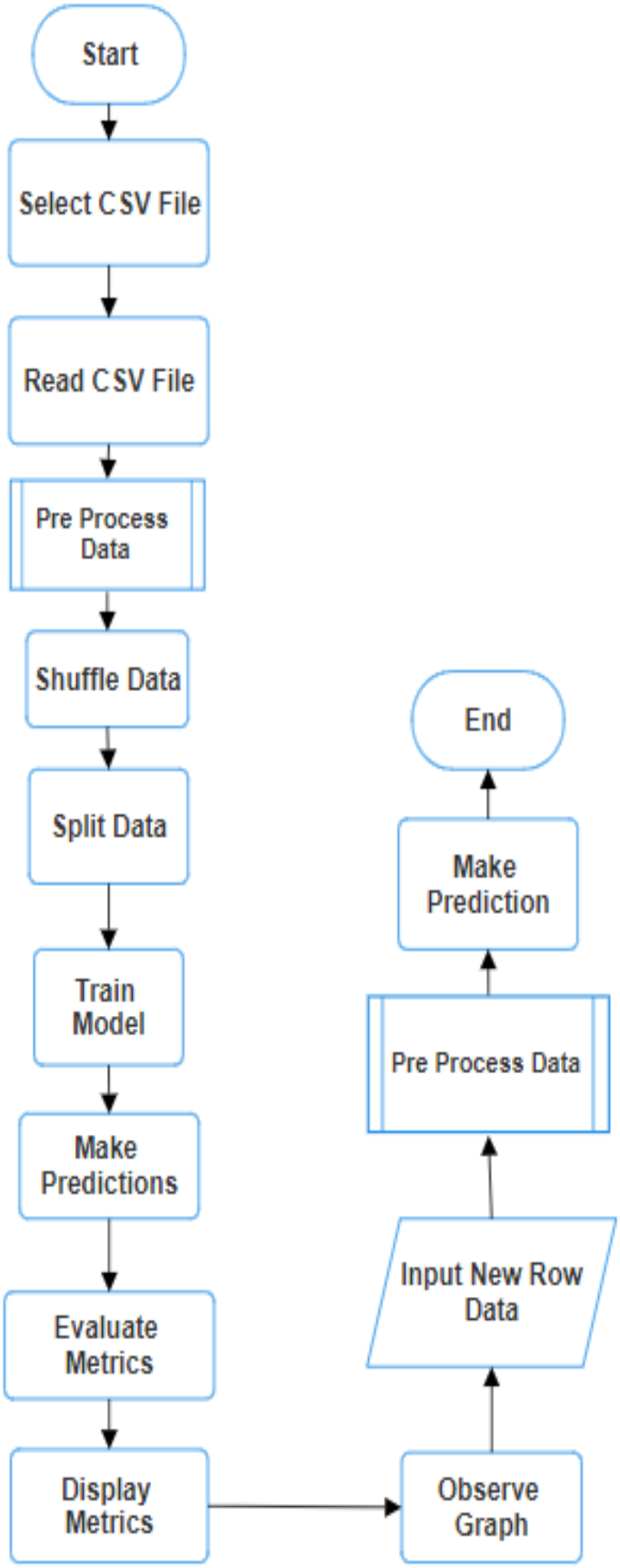
The adjusted R-squared is a measure of the goodness of fit of a regression model that adjusts for the number of predictors in the model. The adjusted R-squared is a modified

version of the R-squared that takes into account the number of predictor variables in the model.

$$Adjusted\ R-squared = 1 - \frac{(n-1)}{(n-p-1)(1-R^2)}$$

where n is the sample size and p is the number of predictors in the model.

4. Flowchart


Start
Select CSV File
Read CSV File
Pre Process Data
Shuffle Data
Split Data
Train Model
Make Predictions
Evaluate Metrics
Display Metrics
Observe Graph
Input New Row Data
Pre Process Data
Make Prediction
End

5. Software/Hardware used

Operating System: Windows 11

Toolbox: None

Hardware: Personal Computer with Intel Core i7 Processor and 8GB RAM

Software: Scilab Version: 6.1.1 and Microsoft Office Excel 2021

6. Procedure of execution

Step 1: Launch the Scilab software on your computer.

Step 2: Open the provided Scilab script file named "**Predictive_modelling.sce**".

Step 3: Execute the script by navigating to the "Execute" menu and selecting "**File with no echo**".

Step 4: Select the downloaded Excel document named "**housing_price.csv**" that was provided in the directory.

Step 5: Observe the graph and console output to analyse the predicted values of the dataset.

Step 6: Enter input values as required when prompted by the script and get the predicted value.

7. Result

After running the code, the console provides a detailed analysis of the results, including:

- Test Labels and Predicted Values:

  The output displays a comparison between the actual housing prices (test labels) and the predicted values generated by the model. Each row in this section corresponds to a specific data point from the test dataset.

```
"Test Labels   Predicted Values"

"349000          294018.81"
"635000          783816.08"
"265000          612385.06"
"460000          451761.01"
"440000          609517.7"
"740000          695644.85"
"221700          355413.89"
"254950          230945.42"
"449000          548659.94"
"485000          353329.41"
```

- Key metrics for evaluating the performance of the predictive model:

  R squared is ranged from 0 to 1, where a values closer to 1 indicates that all variability is explained whereas adjusted R-squared method ranges from negative infinity to 1, with higher values indicating a better fit. The high RMSE and MAE values may reflect the inherent complexity and variability of the real estate market. While the predictive model may not be flawless, it serves as a valuable tool for stakeholders in the Haryana real estate market.

```
"RMSE:"

 203548.48

"MAE:"

 123082.39

"R-squared: "

 0.7146121

"Adjusted R-squared: "

 0.7126460
```

- Graphical output:

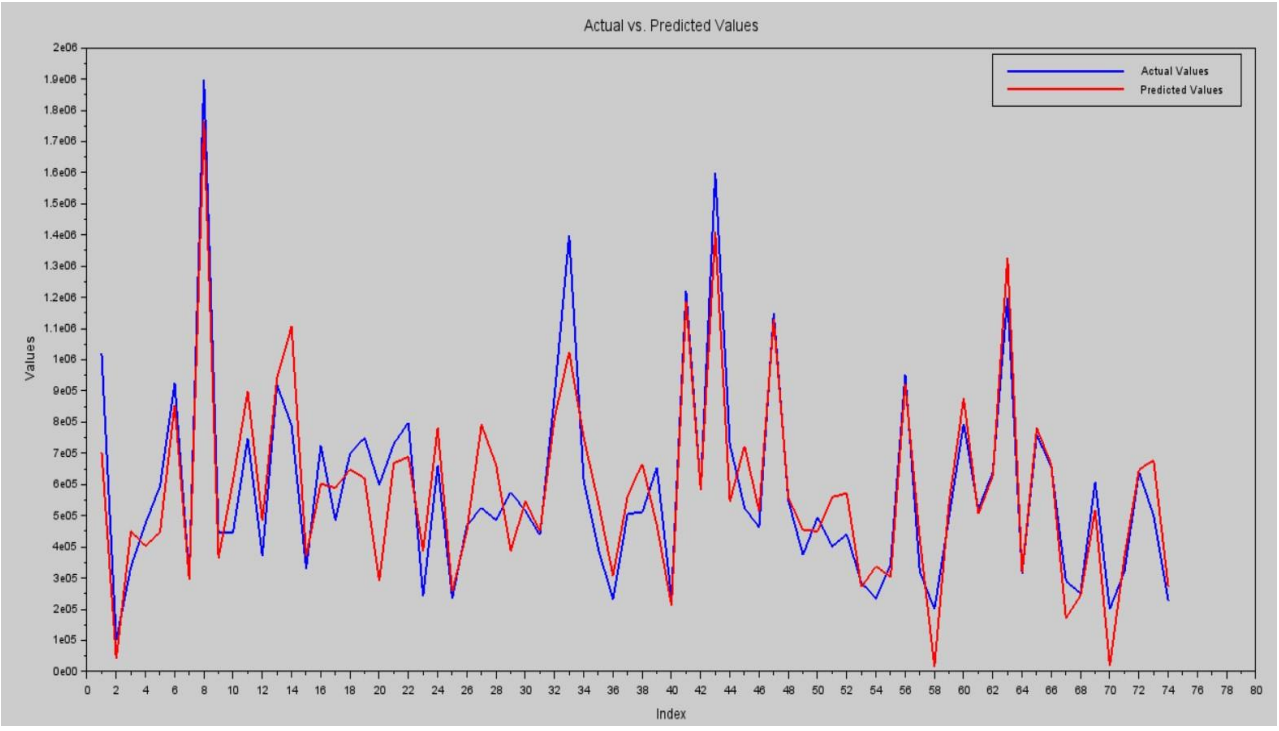
To analyse the graphical output of the predicted and original values, a line graph is plotted to visually compare the predicted values generated by the model with the original values from the dataset.

- The x-axis represents the data points or indexes.
- The y-axis represents the housing prices in INR.
- The original values are represented by a blue line graph.
- The predicted values generated by the model are represented by a red line graph.

Consistency: If the predicted line closely follows the original line suggesting a good fit.

Deviations: Any deviations or discrepancies between the two lines can be observed, helping identify areas where the model may underpredict or overpredict housing prices.

Trend: Overall trend similarities or differences between the predicted and original lines can be assessed, providing insights into the model's performance.



- After inputting data for all 20 features, including number of bedrooms, number of bathrooms, living area, lot area, number of floors, waterfront present, number of views, condition of the house, grade of the house, area of the house excluding basement, area of the basement, built year, renovation year, postal code, latitude, longitude, living area renovation, lot area renovation, number of schools nearby, and distance from the airport, the predictive model generates a predicted value for the median housing price. For the provided input values, the model predicts a median housing price of approximately ₹5,56,368.08.

```
Enter value for feature number_of_bedrooms: 3
Enter value for feature number_of_bathrooms: 2
Enter value for feature living_area (in sq.ft): 2010
Enter value for feature lot_area (in sq.ft): 6000
Enter value for feature number_of_floors: 1
Enter value for feature waterfront_present (Binary, 0 for no waterfront, 1 for waterfront): 0
Enter value for feature number_of_views: 0
Enter value for feature condition_of_the_house (Integer, Range: 0 to 5, where 5 represents the best condition): 3
Enter value for feature grade_of_the_house (Integer, Range: 0 to 5, where 5 represents the best grade): 4
Enter value for feature area_of_the_house_excluding_basement ()in sq.ft): 1330
Enter value for feature area_of_the_basement (in sq.ft): 680
Enter value for feature built_year (in years): 1975
Enter value for feature renovation_year (in years): 0
Enter value for feature postal_code (6 digit Integer): 122005
Enter value for feature latitude (Float): 52.92
Enter value for feature longitude (Float): -114.30
Enter value for feature living_area_renov (in sq.ft): 2080
Enter value for feature lot_area_renov (in sq.ft): 8260
Enter value for feature number_of_schools_nearby : 2
Enter value for feature distance_from_the_airport (in kilometers): 78

  "The predicted price of the house/property is 556368.08(INR), which translates to 5.5636808 lakhs OR 0.0556368 crores."
```

Conclusion: In conclusion, the predictive model for housing in Haryana offers valuable insights into property prices, despite not being perfect. While it may not achieve perfect accuracy, its user-friendly interface makes it accessible for informed decision-making in real

estate transactions. By providing precise forecasts, the model empowers users to make more informed choices in the dynamic Haryana real estate market.

8. References

i. House Price Dataset of India: https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india

ii. Predictive data analysis using linear regression: https://www.researchgate.net/publication/365066122_Predictive_Data_Analysis_Using_Linear_Regression_and_Random_Forest

iii. A Review on Linear Regression Comprehensive in Machine Learning: https://www.researchgate.net/publication/348111996_A_Review_on_Linear_Regression_Comprehensive_in_Machine_Learning

iv. Boston Housing Price Prediction Using Regression Models: https://www.researchgate.net/publication/362812590_Boston_House_Price_Prediction_Using_Regression_Models

v. A Survey on Predictive Models of Learning Analytics: https://www.sciencedirect.com/science/article/pii/S1877050920306451