Hammad Khattak

01-134132-223

Mahena Farooq

01-134132-085

PREDICTING HOUSE PRICES

Bachelor of Science in Computer Science

Supervisor: Dr. M Muzammal

Department of Computer Science

Bahria University, Islamabad

May,2017

Acknowledgements

All praises to Allah for the strengths and His blessing in completing this project. We would like to dedicate this work to our parents, whose love, encouragement, motivation and prayers made us achieve this success. To our teachers who taught us to the very best. To our friends for always being there for us in the need and to our supervisor Dr. Muhammad Muzammal who helped us throughout our project.

ABSTACT

Data mining has been utilized as a part of a few domains and enterprises and the business area is one of its applications. Business world is becoming more competitive. The utilization of data mining innovation has turned into an essential piece of business improvement. Many organizations depend on the utilization of data mining systems to permit managing monstrous data and to uncover the critical and obscure connections between various components of information. This supports the decision-making process and therefore quickens the business development.

A few data mining techniques are utilized by business applications, for example: classification, clustering, prediction, and association. Each of these procedures has demonstrated its impact on blooming businesses.

The primary objective of this Project is to help different stakeholders in real estate market in forecasting the cost of a house with a specific end goal to bolster the choice of offering or purchasing. In this project, we will use house price data of DHA 2 Islamabad and Bahria Town Islamabad.

In this project, Random forest technique is used to predict the selling price of a property. The trials of the Project rely upon a dataset that was scraped from online land classifieds.

CONTENTS

Chapter 1. INTRODUCTION

Data mining is the process of deriving knowledge from large amounts of data. Data mining software's are one of the ways for analyzing data to extract interesting knowledge. They allow users to analyze data from many different dimensions, categorize it, and summarize the relationships identified. Actually, data mining is the way towards discovering connections or examples among many fields in vast social databases.

Data

The text, facts or numbers that a computer can process i.e Operational or transactional data, nonoperational data and meta data i.e data about the data itself.

Information

The examples, affiliations, or connections among this data can give information.

Knowledge

Information can be changed over into knowledge about chronicled examples and future patterns.

1.1. Project Background/Overview

Data mining has been utilized as a part of a few domains and enterprises and the business area is one of its applications. Business world is becoming more competitive. The utilization of data mining innovation has turned into an essential piece of business improvement. Many organizations depend on the utilization of data mining systems to permit managing monstrous data and to uncover the critical and obscure connections between various components of information. This leads to support the decision-making process and therefore quickens the business development.

A few data mining errands are utilized by business applications, for example: classification, clustering, prediction, and association. Each of these procedures has demonstrated its impact on blooming businesses.

The primary objective of this Project is to help different stakeholders in real estate market in assessing the present cost of a focused house with a specific end goal to bolster the choice of offering or purchasing. In this project, we will use house price data of DHA 2 Islamabad.

In this project, Random forest technique is used to predict the selling price of a property. The trials of the Project rely upon a dataset that was gathered from online land classifieds.

1.2. Problem Description

This project will permit the purchasers and the dealers of houses to predict that what value move will occur in the coming time. For this purpose, we Collect data from a property website (zameen.com) and create a data set. After the creation of the data set, we Analyze this data through regression models and Predict the prices using a classifier (e.g. Decision Tree or naïve bayes), i.e. % increase/decrease in Δ time.

1.3. Project Scope

We will analyze the data of DHA 2 Islamabad taken from a property website and give our predictions based on the analysis that we have done on it.

There are many elements that control the increase or the decline of the cost of Houses Examples of these components can be; area of a House, region of the House, number of rooms, number of lavatories, offices, and so on. The utilization of data mining systems in real estate market can help extraordinarily in settling on choices for venture by agents. Likewise, it helps people to settle

on offering or purchasing real estate properties in view of the discoveries and consequences of data mining recent deals records and exchanges [2].

1.4. Project Objective

The main goal of this Project is to assist real estate property market in estimating the future price of the house in order to support the decision of selling or buying.

Chapter 2. LITERATURE REVIEW

In this chapter we give an overview of some other work already done in similar manner. There were many attempts to predict real estate prices using different methods.

2.1. Data Mining:

The principle focus of data mining is to find already obscure connections between a several features and attributes of data. Data mining can bolster in basic decision making in a few business spaces and ventures. It is considered as a gathering of tools and methods that cooperate to reveal the concealed connections between data [3].

One of the prominent benchmarks for depicting the data mining procedure is CRISP-DM (Cross Industry Standard Process for Data Mining). It comprises of six stages that are imperative for finishing a data mining task. Figure 2.1 shows the six phases as follows [4]:

1. **Business understanding** characterizes the data mining issue in the wake of the concerned area.
2. **Data understanding** requires gathering information and to be comfortable with its inclination and its structure.
3. **Data preparation** covers all means of prepossessing for the data mining task. It can be possibly done a few times till the data is completely prepared for analysis. Data preparation includes data cleaning, features and records selection, feature reduction, etc.
4. **Modeling** utilizes distinctive data mining techniques that are chosen in view of the necessities and the prerequisites of the investigation. Hence, data preparation steps may be reexamined to ensure that information meets demonstrating particulars.
5. **Evaluation** assesses the built model to ensure it meets the predefined necessities of the data mining task.

1. **Deployment** is a rehashed procedure of applying the data mining task on data and consistently dissecting the outcomes utilizing suitable data mining software. Figure 2.1 Cross Industry Standard Process for Data Mining, CRISP-DM [5]

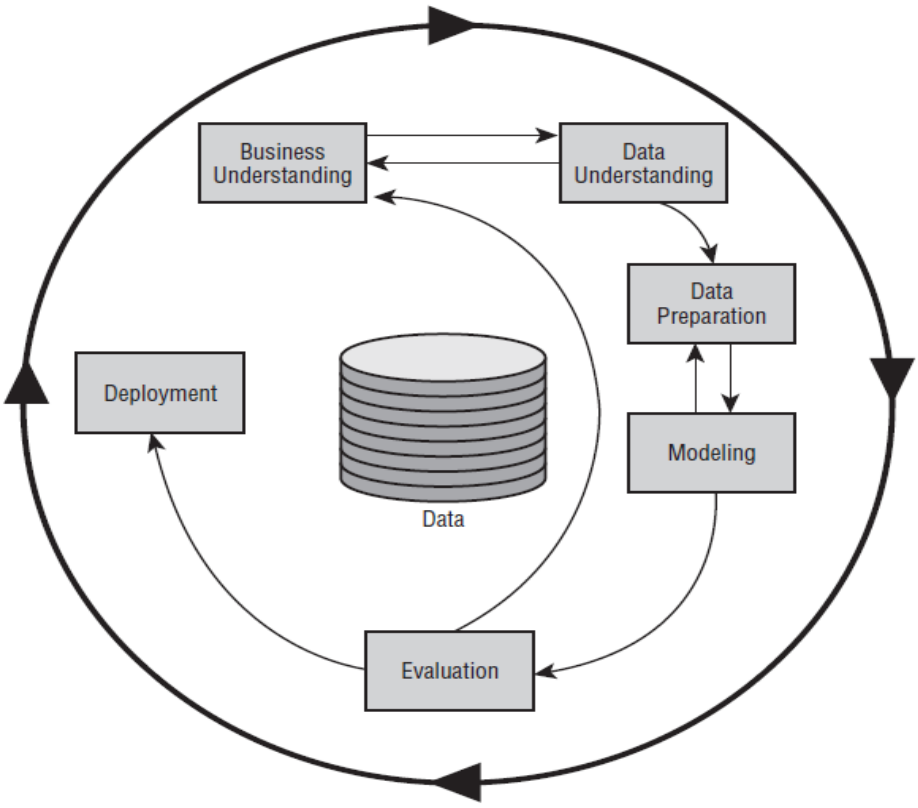

Figure 2.1 CRISP-DM (Cross Industry Standard Process for Data Mining)

With regards to cases of data mining models, it is critical to note that there are a few techniques and models that can be utilized. Examples of such techniques are: classification, clustering, regression, association rules, etc. The following subsection considers giving brief information about the used data mining technique in this project, which is linear regression.

2.2. Related Work

There are a few research papers that work on looking at the components and the attributes that add to the forecast of real estate properties costs utilizing diverse data mining techniques.

2.2.1. R. D. Jaen, “Data Mining: An Empirical Application in Real Estate Valuation, FLAIRS-02 Proceedings,”,2002

As indicated by Jaen (2002) [6] decision tree and neural system procedures conceded to doling out the "sqrt living territory" highlight as a decent indicator at the cost of a house. In this research paper, 15 numerical elements were utilized to represent the houses' qualities in addition to a clear cut element that depicts the address. Be that as it may, the Address attribute has been changed over to a numeric data type. The dataset comprised of 1000 records that were gathered from the houses' sales exchanges in Miami, US. The information were sorted out and gotten from a Multiple Listing System (MLS) database. MLS is a property database that is produced by land specialists with a specific end goal to share data about the accessible properties available to be purchased at a given time (Massey University, 2013). In the start of the investigations, the attributes were decreased to nine attributes by choosing the attributes that for the most part influence the value expectation utilizing a stepwise linear regression.

2.2.2. Wedyawati and Lu “Mining real estate listings using Oracle data warehousing and predictive regression”,2004.

Same as Jaen's (2002) paper, the reasearch paper of Wedyawati, and Lu (2004) [2] has considered the MLS database as the source of data for gathering land property deals exchanges. As indicated by Wedyawati, and Lu (2004) the tests secured 295,787 exchanges from four urban communities in the US. Around 191 numerical elements have been utilized. Visual Basic .NET programming was utilized to run the linear regression on the dataset in addition to Oracle Data Warehousing programming was utilized to gather and compose the gathered information from the MLS database.

However, the paper didn’t mention anything about the prediction accuracy of the implemented system. Instead, it stated the limitations to be: first, not all features of the MLS database were extracted and used in experiments. Second,

the data warehouse has no ability to update its content. In addition, one of the paper's suggestions was to add more features that give more descriptive information about houses such as, how many car garage(s) are available, is there a swimming pool or not, and the area, in square feet, of a property.

2.2.3. Guan, Levitan and Zurada "An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment", 2008

Another data mining system was utilized by (Guan, Levitan and Zurada, (2008) [7] to expect the cost of real estate properties. An Adaptive Neuro–Fuzzy Inference System (ANFIS) was actualized and tried more than 360 records of past deals properties in Midwest, US. The dataset has 14 numerical elements. The paper asserts that it has the lead position of proposing such a framework for anticipating land properties' costs. The outcomes demonstrated that the ANFIS has nearly a similar execution comes about as different linear regression models. Additionally, the execution of the framework has expanded when an feature reduction procedure was utilized, for example, Principle Component Analysis (PCA). Unmistakably the utilized data in analyses was so little to speak to the private properties' attributes.

2.2.4. Carlos Del Cacho" A comparison of data mining methods for mass real estate appraisal", 2010.

As per Cacho (2010) [8], a few data mining techniques were utilized to foresee the cost of land properties in the city of Madrid, Spain. Cases of utilized systems are: Naïve Neighborhood, numerous direct investigation, multilayer recognition, M5 model trees, K-nearest neighbors, and so forth. It is found that the gatherings of M5 model trees beat every other procedure. Likewise, the utilization of gatherings of M5 trees has diminished the relative error of expectation by 23%. Then again, 25,415 records that have 40 features were utilized as a part of the examinations. The features gave point by point data about the qualities inside and outside the lofts. For instance: zone of loft, number of rooms, number of restrooms, floor material, aerating and cooling, swimming pool, tennis court, cultivate region, carport, and so forth. Moreover, some geospatial components were utilized, for example, the distance to the closest metro station and number of

retail locations in a 500 meters span from the flat. The records were gathered from a few land entrances in a specific date, 10th.Nov.2010. Additionally, say that the dataset was separated into 21 subsets and every subset alludes to an authoritative region in the city of Madrid. Accordingly, the investigations were done over every subset independently.

2.2.5. Claudio Acciani, Vincenzo Fucilli and Ruggiero Sardaro "Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach"2011.

Not like the pervious papers, Acciani, Fucilli and Sardaro (2011) [9] paper looks at the data mining model on the offers of farms in a few ranges in Italy for the period from 2008 to 2010. The dataset comprises of 169 deals exchanges with 14 features (nine unmitigated and five proceeds with components). The two utilized data mining techniques were Model Trees (MT) and Multivariate Adaptive Regression Splines (MARS) and they were actualized utilizing WEKA programming. The utilization of both systems prompts a similar execution in foreseeing farm's costs per hectare. In any case, MT and MARS beat the standard Multivariate Linear Regression (MLR).

Chapter 3. DATA

3.1. Data Collection

In order to get data we wrote a scraper in Microsoft visual studio. Data scraping is a technique which helps a computer program extracts data from human-readable output that comes from another program.

To get the historical data of the houses, we physically went to the office of zameen.com but we could not get the data from there. As there was no other way of getting the data and the website of zameen.com was open to scraping, we wrote a scraper through which we were able to get the data set that contained the details of the houses that were to be sold. The data set contains the area in Marla's, number of rooms, price in millions, sector, the title of the add and description of the add. The data set that we got from zameen.com was not ideal for applying data mining techniques to predict the prices of the houses, so we did data preprocessing on that data to bring it in a form on which we can apply linear regression and random forest. To apply the data mining techniques on the data, we divide the data into Training data set and Test data set. The Training data set is used to train the machine for future prediction of the house prices and the Test data set is used to check whether the machine makes the right predictions or not.

3.2. Data Preprocessing:

The real world Data generally is dirty (noisy, inconsistent, incomplete, missing) so before applying any data mining technique we need to preprocess the data to clean. The data we handle is real world data so this step is a very essential step of our project.

For this project we have taken DHA Phase 2 Islamabad and Bahria Town Islamabad as a test case for the prediction of house prices. The data however was not publicly available so we had to write a web scraper for a real estate website to get the data.

The data-set which we got after scraping had 769 records and 9 attributes (Id,title, price, sector, description, rooms-marlas, added, marketed by, photo).

The marketed by column gave us the information that who marketed the house and photo column gave us the information about whether the add posted on the real estate website had a photo or not so we ignored these columns because they weren't needed, the sectors column did not have any value but the title column and the description column had the sectors mentioned in them so we filled the missing values of the sectors column by getting the sectors from either of the two.

The basic data-set description before performing data preprocessing is given in the *Table 3.1* below

Table 3.1 Data Description

Attributes	Type	Different values
Title	Character	257
Price	Character	73
Sector	Character	8
Description	Character	276
Rooms-Marlas	Character	9
Added	Character	56
Photo	Factor	2

The attribute rooms-marlas had the information of how many bedrooms a house had and what is the plot size on which the house is made, so we split this attribute to two different attributes the rooms and the marlas.

Rooms-Marlas
6 bedrooms - 1 Kanal
5 bedrooms - 1 Kanal
6 bedrooms - 1 Kanal
4 bedrooms - 10 Marla
5 bedrooms - 10 Marla
6 bedrooms - 1 Kanal
7 bedrooms - 1 Kanal
6 bedrooms - 1 Kanal
5 bedrooms - 1 Kanal

Before Preprocessing

rooms	Marlas
6 bedrooms	1 Kanal
5 bedrooms	1 Kanal
6 bedrooms	1 Kanal
4 bedrooms	10 Marla
5 bedrooms	10 Marla
6 bedrooms	1 Kanal
7 bedrooms	1 Kanal
6 bedrooms	1 Kanal
5 bedrooms	1 Kanal

After Preprocessing

Now the rooms attribute told us about the number of rooms in the house, the problem was that the data type of this attribute was character so to convert it into a numeric data type we had to remove the word bedrooms and only keep the number of bedrooms.

rooms
6 bedrooms
5 bedrooms
6 bedrooms
4 bedrooms
5 bedrooms
6 bedrooms
7 bedrooms
6 bedrooms
5 bedrooms

Before Preprocessing

roomsNo
6
5
6
4
5
6
7
6
5

After Preprocessing

The price attribute had the prices of the houses, the prices originally were not consistent some of them were in lacs and some of them were in crore, so we converted all of them into rupees by multiplying by 100000 if the price was given in lacs and multiplying by 10000000 if the price was given in crores. Thus ,converting the data type to number from character.

price
3.6 Crore
4.25 Crore
3.3 Crore
2.5 Crore
2.4 Crore
3.9 Crore
3.65 Crore
3.9 Crore
3 Crore

Before Preprocessing

PriceInRupees
36000000
42500000
33000000
25000000
24000000
39000000
36500000
39000000
30000000

After Preprocessing

The marlas attribute told us about the size of the plot on which the house is made ,the data in this column was inconsistent some of the areas were given in marlas and some in kanals, so we converted all the values of sizes into square feet i.e. 1 kanal is equal to 5445 square feet .

Marlas
1 Kanal
1 Kanal
1 Kanal
10 Marla
10 Marla
1 Kanal
1 Kanal
1 Kanal
1 Kanal

Before Preprocessing

SizeInSquareFeet
5445.000
5445.000
5445.000
2722.510
2722.510
5445.000
5445.000
5445.000
5445.000

After Preprocessing

3.3. Data Understanding

After preprocessing we had a clean data-set. We used the DHA 2 islamabad's data as the train data-set and we used another data-set of Bahria town which too was scraped from zameen.com as the test data-set.

The title attribute is title of the add, which is a character attribute. The description attribute tells us the description of the house, it is a character attribute. The sector attribute tells us about the sectors that in which sector the house is, it is a character attribute

The price attribute tells us the price of the house, which was changed into another attribute named "PriceInRupees" which is a number attribute the summary of the attribute, is shown in *Table 3.2*.

Table 3.2 PriceInRupees

Min	2000000
1st Quartile	26000000
Median	35000000
Mean	33088687
3rd Quartile	38000000
Max	95000000
NA's	307

The RoomsNo attribute tells us about how many bedrooms a house has, it is a number attribute the summary of the attribute is shown in *Table 3.3.*

Table 3.3 Rooms

Min	1.000
1st Quartile	5.000
Median	5.000
Mean	5.247
3rd Quartile	6.000
Max	12.000

The marlas attribute tolds us the size of the plot on which the house is made, it was converted into "SizeInSquareFeet" which is a numeric attribute, the summary of the attribute is given in *Table 3.4* below.

Table 3.4 Marlas

Min	272.2
1st Quartile	2722.5
Median	5445.0
Mean	4735.4
3rd Quartile	5445.0
Max	32996.7
NA’s	73

Chapter 4. REQUIREMENT SPECIFICATION

Following sections discusses

- Why the proposed system is built?
- The functional and nonfunctional requirements of the proposed system
- Interactions between the user and the system via Use Case diagrams.

4.1. Overview

Data mining can support decision making in many fields of business. Our project aims at the real-estate market in Pakistan in which currently there is no such mechanism through which the stakeholders can know the future prices of houses. We have applied data mining prediction techniques (linear Regression & Decision trees) on the dataset which has data of real-estate classifieds to predict the prices of the houses.

4.2. Existing System

The Existing system used in the Real-Estate industry is a manual system, in which the stakeholders have to get the help of a Real-Estate manager/agent who then helps or assists the stakeholder which can be a buyer, seller or an investor who wants to buy or sell a house. In this system there is no mechanism through which the future prices can be known.

4.3. Proposed System

The proposed system the user will only have to enter the different attributes which effect the price e.g. Location of the house, number of bedrooms, etc. The system will then predict price of that house by analyzing the historic data through different data mining techniques.

4.4. Functional Requirements

4.4.1. The system predicts the prices of houses by different attributes given in the data.

4.4.2. The system must be trained on the data before.

4.4.3. The system must develop an accurate model for prediction using data mining technique.

4.4.4. The system should predict the prices after classification with respect to different attributes e.g. sector, number of bedrooms etc.

4.5. Nonfunctional Requirements

4.5.1. Usability

- The system should be easy to use for the users.
- It should have a user friendly front end.

4.5.2. Speed

- The system should not have lag.
- It should not take a lot of time to compute the prediction according to the information given by the user.

4.5.3. Reliability

- The system should give accurate price predictions.
- It should have historic data from some reliable source.

4.5.4. Data Quality

- The data should be real life data.
- The data-set should have complete information (no missing data).
- The data-set should have undergone preprocessing before designing a model so that there is no noisy data.
- The data-set should have historic data in it.

4.6. Use Cases

A use case diagram is a graphic representation of the interactions between the components of a system. The following use case diagram i.e. *Figure 4.1* is used to clarify, identify and organize system requirements. This diagram is employed in UML (Unified Modeling Language) which is a standard notation for the modeling of real world systems.

4.6.1. Use Case Diagram

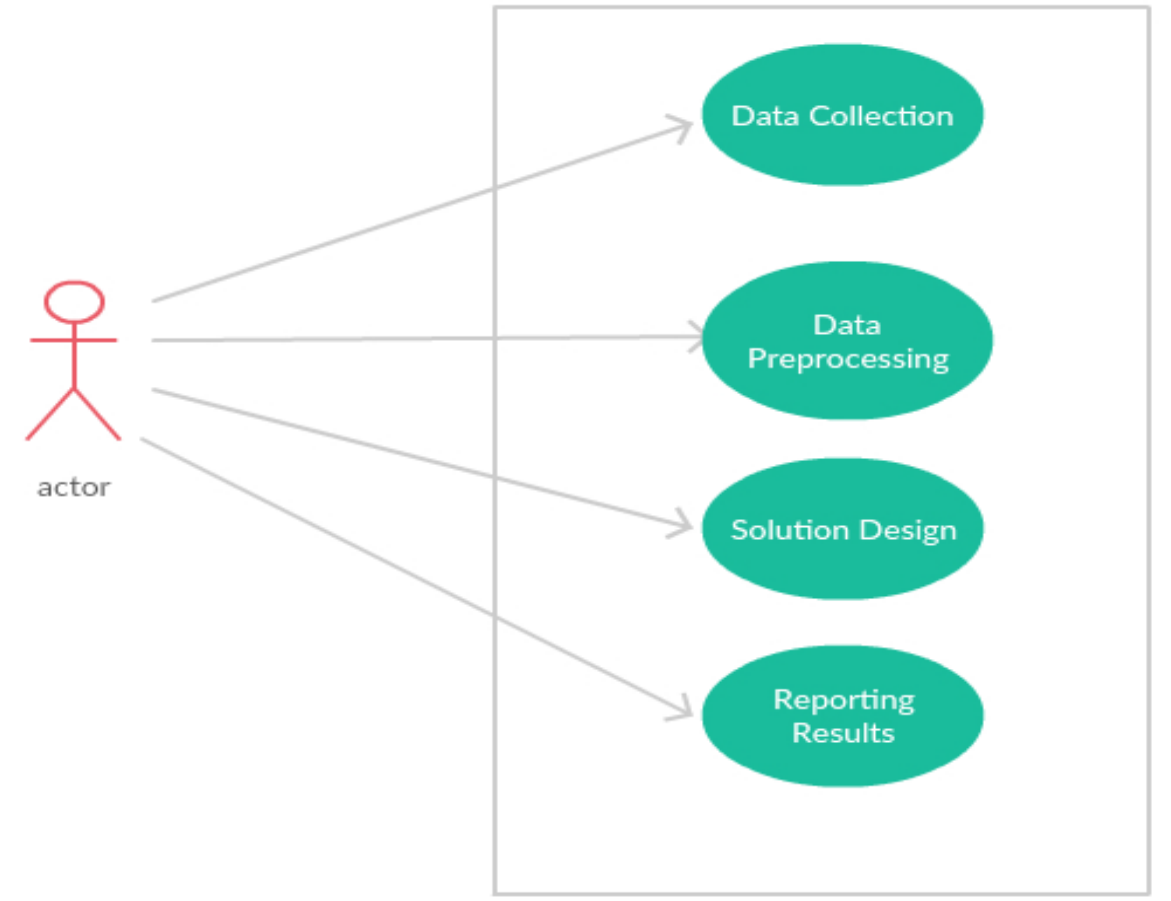

Figure 4.1 Use Case Diagram

4.6.2. Use Case Log

The following table i.e. *Table 4.1* is used to give a brief idea about what each table is doing .This table consists of the components of the use case diagram.

Table 4.1 Use Case Log

ID	Actors	Use Case Name
UC-01	User	Data Collection
UC-02	User	Data Preprocessing
UC-03	User	Solution Design
UC-04	User	Reporting Results

4.6.3. Use Case Description

The following tables i.e. *Table 4.2, 4.3, 4.4, 4.5* are used to describe the basic functionality in each of the use case.

Table 4.2 Use Case 1

Use Case ID	UC-01
Title	Data Collection
Description	Scrape the data from a Real-estate website
Actors	User
Precondition	None

Table 4.3 Use Case 2

Use Case ID	UC-02
Title	Data Preprocessing
Description	Removing the noise from the data and estimating the missing values.
Actors	User
Precondition	None

Table 4.4 Use Case 3

Use Case ID	UC-03
Title	Solution Design
Description	To design a solution using data mining techniques (Linear Regression etc.).
Actors	User
Precondition	None

Table 4.5 Use Case 4

Use Case ID	UC-04
Title	Reporting Results
Description	Report the results of different data mining techniques.
Actors	User
Precondition	None

Chapter 5. SYSTEM DESIGN

System design is one of the most important part in developing any kind of system because it determines the all the architecture of the system.

In this chapter we will discuss the design and architecture of the system to be developed. It also presents an overview of how different processes will carry out and interact with each other to achieve the final result.

5.1. System Architecture Diagram

The figure 5.1 shows the system architecture in which the data-set scraped from the real-estate website is first preprocessed, then the processed data is used to develop a prediction model using data mining techniques (linear Regression & Decision Tree) after which the model is applied on the data set and in the end the Results (Price Predictions) are reported on the front end of the system.

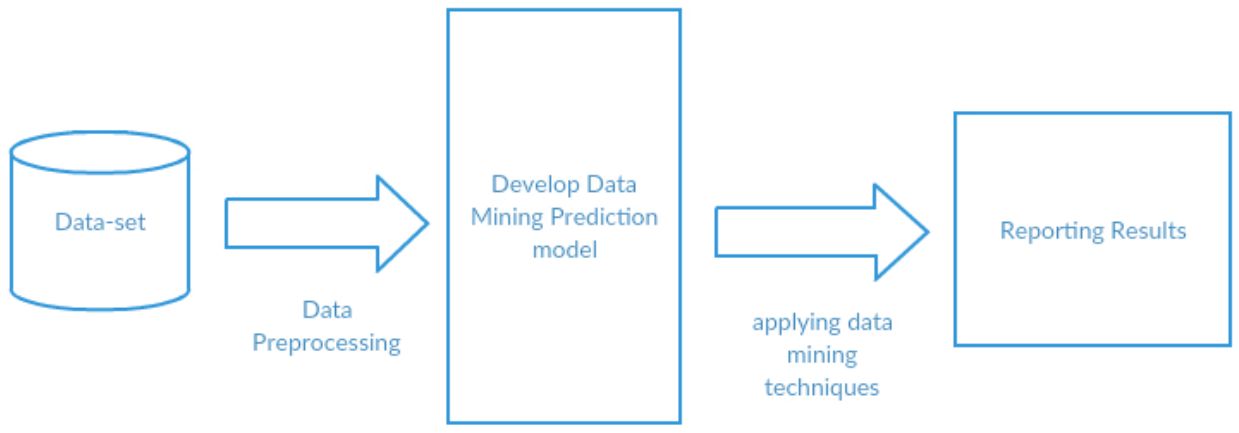

Figure 5.1 System Architecture Diagram

Chapter 6. SYSTEM IMPLEMENTATION

This chapter outlines the detailed description about how different prediction techniques are implemented in this project and the tools and technologies that are used in making the project are also explained in this chapter.

6.1. Components of the System

As this project's aim is to predict the house prices, we used different data mining techniques. The two techniques that we used are Random Forest and Linear Regression. The components of our project include a data set that we scrapped from a real estate website, linear regression, random forest and a front end of our system.

6.2. Tool and Techniques

Various tools and techniques are used for predicting the house prices.

6.2.1. R Studio

Rstudio is an open-source integrated development environment (IDE). It is a programming language for statistical computing and graphics. It uses the Qt framework for its graphical user interface. It is written in C++ programming language

There are two editions of rstudio :

6.2.1.1. Rstudio Desktop:

In this version the program is run locally as a regular desktop application.

6.2.1.2. Rstudio Server:

In this version rstudio is accessed using a web browser while it is running on a remote linux server.

In our project, we used the Desktop version of rstudio to apply linear regression and random forest on the data set that we scrapped from a real estate website to predict the future prices of the houses.

6.2.2. visualstudio2013

Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It supports multiple programming languages and it is used to develop computer programs for windows, web applications, web sites, mobile applications and web services. Its debugger and code editor supports almost all the programming languages if their language- specific services exist. We have used Microsoft Visual Studio to write down the scrapper to get the data set form a real estate website.

6.3. Libraries

6.3.1. Open XLSX

It is necessary to import the data set into R, for this purpose this library is used. We have used Open XLSX to open or link the data set of type XLSX to rstudio.

6.3.2. String R

It is a consistent, simple and easy to use package. All function and argument names (and positions are consistent, all functions deal with ``NA''s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another.

6.3.3. gglpot2

R has a plotting system called gglpot2 that is based on the grammar of graphics. It tries to take good parts of base and lattice graphics and none of the bad parts. It solves the problem of plotting a hassle and provides a powerful model of graphics that makes it easy to produce complex multi-layered graphics. We have used this library in our project for visualization.

6.3.4. rpart

This is the abbreviation of Recursive Partitioning and Regression Trees. We have used this library in our project for imputation.

6.3.5. dplyr

It focuses on tools for working with data frames. It is an extension of plyr. It Identifies the most important data manipulation tools needed for data analysis and make them easy to use from R. it provides fast performance for in-memory data. We have used this library for data manipulation.

6.3.6. Lubridate

This library is used to work with date-times and time-spans. It is used for fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects. This library has a consistent and memorable syntax that makes working with dates simple.

6.3.7. Readxl

It is used to import Excel files into R. it supports '.xls' via the embedded 'libxls' C library. It Works on Windows, Mac and Linux without external dependencies. We have used this library in our project to import the dataset from excel.

6.3.8. RandomForest

It is a library which has to be added in order to use the random forest technique for prediction.It is used for Classification and regression based on a forest of trees using random inputs. This library has a function **randomForest()** which is used to create and analyze random forests. It consists of a formula for describing the predictor and response variables and **data** is the name of the data set used. Because we have used random forest as a classifier to make predictions, this library is used in this project as a classifying algorithm.

6.4. Methodology

Following are the details about the technique that is used in our projects implementation.

6.4.1. Random Forest

Random forest is an algorithm that is intended to do regression or classification on the given data set.

6.4.1.1. Ensemble

Ensembles are a divide-and-conquer approach that improves the overall performance of the system. The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. It is generally a group of algorithms that combine together to give accurate results. Random forest utilizes decision trees and works as large correlated decision tree. It makes a lot of decision trees and then uses them for classification.

It is an Ensemble machine learning which uses other algorithms that can be potentially weaker and are not able to produce accurate results when used individually and combines with them to produce better results. In this project, we will combine random forest with linear regression to produce reliable results.

6.4.2. Over-fitting

Random forest overcomes the problem of 'over-fitting'. Over-fitting refers to a situation in which the training data set is either limited or it is not the true representation of the real world data. If the training data set is limited, the model will not produce good results and it will be potentially weak.

The data set that we scraped from a real estate website was limited so we used random forest because if there is missing data or sparse data in the training data set, random forest can overcome the problems.

6.4.3. Voting

Voting is a part of random forest. Random forest utilizes decision trees. It uses the classification that has the most votes. It makes a lot of decision trees and then uses the result of each decision tree in voting. In this project, we took three independent variables (X) and made three different decision trees. We took price as a dependent variable(Y). In voting process, the result that is in minority is discarded and the results that are in majority are chosen.

6.5. Limitations and challenges

This project is dependent upon a number of things so the price prediction may not always be accurate. We took into account a lot of attributes on which the price of a house depends but the variance in the prices of the plots or the houses is unpredictable. The price of land will change if something attractive opens in that area. Such things are not covered in this project because data mining techniques are not able to predict such events.

The data that we took from a real estate website by writing a scraper was not ideal for applying data mining techniques. The data was preprocessed and the noise was eliminated. We used Random forest because it overcomes over-fitting.

Chapter 7. SYSTEM TESTING AND EVALUATION

System testing is a part of development process, so in order to complete the process we have tested our project against different testing techniques to ensure that the project is working according to the requirements. We have verified our project in order to ensure that all the errors and bugs are removed and the system works efficiently. We have also validated our project to make sure that the system produces the expected results. The core purpose of testing is to make sure that the system meets standard quality guidelines. This chapter provides the details of the different testing techniques that we have used in order to check the systems performance. This chapter also presents different test cases that are designed to evaluate the correct functionality of the system. There are different types of testing techniques used to finish the distinctive phases of testing

7.1. Usability Testing

Usability testing is to check how 'Easy to use' a system is. Usability testing involves real users to give their feedback. To check the problems that the users encounter whilst using the system, the users are asked to complete a given task during which the users are observed. The frontend interface of our system was tested in order to check the usability of the system and to make sure that the system provides a better experience to its users.

7.2. Software Performance Testing

To test the speed, efficiency, reliability and accuracy of the software is called Software performance testing. This process involves quantitative testing such as measuring the response time, reliability etc. Performance testing is also used to figure out the bottlenecks in a system.

7.3. Graphical User Interface Testing

The user interface of our project went through the graphical user interface testing phase in which the proper functionality of the graphical user interface (GUI) for the system was tested to ensure that it conforms to the required specifications. To make the system more user friendly, efficient to use, easy to understand and easy to remember, appropriate layout, colors, font, buttons, icons and menus were used. This system was manually inspected and tested by a third party to find out the flaws in the interface.

7.4. Exception Handling

Our project deals with price in crores and area in Marla's, so in case any of the seller has given price in lacs or the area in kanals, we will first process the data in order to bring it in a form that is acceptable by the system and then we will apply the prediction techniques on the given data.

7.5. Execution Testing

Execution testing is the process in which the expected results are verified against the actual results. To test the quality of the system, execution testing is performed.

7.6. Module Testing

Module testing or component testing refers to testing each and every single independent unit of the whole system individually. Every individual component of the system is checked and then rechecked to ensure that each component works well individually.

7.7. Integration Testing

In this project, after creating the whole project on R-Studio, it was integrated with the front end interface of the system. After testing the components individually, they were integrated and then tested again in order to make sure that they do not produce any new errors or bugs after integration. If any new errors are produced, they are sorted out and then the integrated system is tested again.

Chapter 8. CONCLUSION

The core purpose of this Project is to assist different parties in real estate property market in estimating the current price of a targeted real estate property in order to support the decision of selling or buying. The research focuses on the houses of DHA 2 Islamabad. This project uses data mining techniques for prediction. The two techniques that we have chosen for prediction in our project are linear regression and random forest. First we have applied linear regression on the given data and then we have used its results for further processing. We have applied random forest on the results obtained by linear regression and then we chose the final result by voting which is a part of random forest. We have used two data mining prediction techniques in order to predict as accurate results as possible.

8.1. Major Accomplishment

The project successfully predicts the prices of the houses and the project tends to make the prediction as accurate as possible by applying random forest technique. The major accomplishment of this project is that this project will help different people in estimating the current price of a targeted real estate property. This system will aid its users in taking the decision about buying and selling the property.

This project allows the buyers and the sellers of houses to foresee the price shift that is going to happen in the coming time. For this purpose we have Collected data from property website and created a data set. We preprocessed that data in order to filter it and bring it into a form that the system can exploit. After preprocessing the data we will analyze this data through regression models and predict the prices using a classifier i.e. random forest.

8.2. Limitations and Challenges

The major challenge in our project was to take into account several aspects that influence the prices of the houses. Our system doesn’t handle situations that affect the price of the houses for example change in the national tax policy and change in geopolitical situation of the area etc.

9. REFERENCES

[1] Hotho, A., Nürnberger, A. and Paaß, G. “A Brief Survey of Text Mining. Journal for Computational Linguistics and Language Technology” (2005).

[2] Wedyawati and Lu “Mining real estate listings using Oracle data warehousing and predictive regression”,2004.

[3] Berson, A., Smith, S., and Thearling, K “An Overview of Data Mining Techniques” (2011).

[5] Azevedo, A.and Santos “M. KDD, SEMMA AND CRISP-DM: A Parallel Overview in Ajith Abraham, ed., IADIS European Conference on Data Mining” (2008).

[6] R. D. Jaen, “Data Mining: An Empirical Application in Real Estate Valuation, FLAIRS-02 Proceedings,” American Association for Artificial Intelligence, 2002.

[7] Guan, Levitan and Zurada “An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment”, 2008

[8] Carlos Del Cacho” A comparison of data mining methods for mass real estate appraisal”, 2010.

[9] Claudio Acciani, Vincenzo Fucilli and Ruggiero Sardaro “Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach”2011.

[10] Dukova, R., Gacovski, Z., Kolic, J. and Markovski, M. “Data Mining

Application for Real Estate Valuation in the city of Skopje. CT Innovations2012 Web Proceedings” (2012).