Using Data Mining and Text Mining Techniques in Predicting the Price of Real Estate Properties in Dubai

استخدام تقنيات استنباط البيانات و استنباط النصوص في تقدير أسعار العقارات في دبي

By

Deena Younis Abo Khashan

Dissertation submitted in partial fulfillment of

MSc Informatics (Knowledge and Data Management)

Faculty of Engineering & Information Technology

Dissertation Supervisor

Dr. Sherief Abdullah

May-2014

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Deena Younis Abo Khashan	100148	Informatics (Data and Knowledge Management)	31.May.2014

Title

Using Data Mining and Text Mining Techniques in Predicting the Price of Real Estate Properties in Dubai

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature

Abstract

Data mining is defined as the discovery of previously unknown patterns and relationships between stored data and represents the interesting information in understandable format. On the other hand, text mining is also looks for interesting hidden information, but on human natural language data. Data mining techniques are applied in several domains and industries and one of these domains is the real estate domain. Predicting the price of real estate properties based on past sales transactions is one of the data mining applications.

In this thesis, the effectiveness of using linear regression modeling as a data mining technique in predicting the price of a real estate property based on online real estate classifieds is examined. The results of linear regression predictions based on the structured features are considered as the baseline of the research. Then, text mining is used to convert the unstructured text features into proper format so that the accuracy of linear regression prediction is tested again after involving text features in the experiments. All experiments are implemented and tested using RapidMiner5 software tool.

Before starting the experiments, the dataset is collected from different online real estate property classifieds that offer villas and apartments either for renting or selling in Dubai, UAE. The dataset has been divided into six subsets based on the type of the classified. The experiments are carried out on each subset by first using the regular structured numerical features of the real

estate property in predicting the price using linear regression modeling. Then, the linear regression prediction experiments are repeated after text mining the descriptive text features. After text mining process, thousands of features are generated to reflect the significance of thousands of unique words using the term weighting scheme: term frequency-inverse term frequency TF-IDF.

It is found that the results of linear regression prediction have been improved significantly after adding text mining to the experiments. The root mean squared errors RMSEs for all data subsets have decreased leading to enhancing the accuracy of prediction. For example, the RMSE is reduced by almost 56% for two data subsets that concerns the classifieds of the offered villas for renting and the classifieds of the apartments that are offered for selling. Also, the linear correlation between regular features and the price feature has increased noticeably. For example, the correlation coefficient metric has increased by 206.08% for the dataset that holds the records of the apartments that are offered for selling. Moreover, the analysis shows that the key factor that controls the renting or selling price of a real estate property in Dubai is its location.

To the best of our knowledge, this research is the first scientific analysis of Dubai's real estate classifieds and it is the first trial in improving the accuracy of using data mining technique in price prediction using text mining. Also, verifying the experimental results is supported using a real world dataset that reflects the current trends in the real estate market.

خلاصـــــة

ان تعريف مفهوم استنباط البيانات يقوم على اكتشاف أنماط لم تكن معروفة سابقا وعلاقات بين البيانات المخزنة و تمثيلها في شكل مفهوم. من ناحية أخرى، ان مفهوم استنباط النص يقوم أيضا على الحصول على معلومات خفية مثيرة للاهتمام، ولكن اعتمادا على بيانات من نوع نصي. يتم تطبيق تقنيات التنقيب عن البيانات في العديد من المجالات والصناعات وأحد هذه المجالات هو المجال العقاري. توقع سعر العقارات و التنبؤ بها على أساس معاملات البيع السابقة هي احدى تطبيقات استخراج و استنباط البيانات.

في هذه الأطروحة، تم فحص فعالية استخدام تقنية نماذج الانحدار الخطي كأسلوب للتنقيب عن البيانات و التنبؤ بأسعار العقارات اعتمادا على إعلانات العقارات المبوبة على الانترنت. تعتبر نتائج تنبؤات الانحدار الخطي هي الخط الأساسي للأطروحة. اضافة إلى ذلك، تم استخدام تقنية استنباط النص لتحويل البيانات النصية غير المنظمة إلى شكل مناسب بحيث يتم اختبارها في تجارب تنبؤ الانحدار الخطي مرة أخرى. هذا و قد تم تنفيذ جميع التجارب واختبارها باستخدام الأداة البرمجية RapidMiner.

قبل البدء بالتجارب، تم جمع البيانات من الإعلانات المبوبة للعقارات على الانترنت التي تعرض الفلل والشقق إما للإيجار أو للبيع في دبي، الإمارات العربية المتحدة. و قد تم تقسيم البيانات التي تم الحصول عليها إلى ست مجموعات فرعية على أساس نوع الإعلان. تم تنفيذ تجارب التنبؤ بالاسعار باستخدام الانحدار الخطي على كل مجموعة فرعية على حدة، ثم تمت اعادة تجارب الانحدار الخطي بعد ادخال تقنية استنباط النصوص على البيانات النصية التي تستخدم في وصف العقارات. و الجدير ذكره انه بعد عملية استنباط النص، يتم إنشاء الآلاف من الحقول التي تمثل كل منها احدى الكلمات النصية وتعكس أهميتها باستخدام نظام TF-IDF

اظهرت نتائج تنبؤ الانحدار الخطي تحسنا بشكل ملحوظ بعد إضافة تقنية استنباط النص إلى التجارب. فلقد انخفضت نسبة الخطأ من نوع RMSEs لجميع مجموعات البيانات مما أدى إلى

تعزيز دقة التنبؤ. على سبيل المثال، تم تقليل RMSE بنحو 56٪ للبيانات التي تتعلق بالإعلانات المبوبة للفلل المعروضة للإيجار والإعلانات المبوبة للشقق المعروضة للبيع. أيضا، ازداد الارتباط الخطي بين البيانات العادية و السعر بشكل ملحوظ. على سبيل المثال، ازداد معامل الارتباط الخطي بنحو 206.08٪ لمجموعة البيانات التي تحتوي على سجلات الشقق المعروضة للبيع. علاوة على ذلك، يبين تحليلنا أن العامل الرئيسي الذي يتحكم في سعر تأجير أو سعر بيع العقارات في دبي هو موقعها.

إلى حد علمنا، تعتبر الأطروحة أول تحليل علمي مختص بالإعلانات العقارية المبوبة في دبي و هي تتضمن أول محاولة يتم اجراؤها بهدف تحسين دقة استخدام تقنية التنقيب عن البيانات في التنبؤ باستخدام تقنية استنباط النص. بالاضافة إلى ذلك فإن التحقق من النتائج التجريبية لدينا اعتمد على استخدام مجموعة بيانات تم تجميعها بشكل لحظي بحيث تعكس صورة محدثة للإتجاهات الحالية في سوق العقارات.

Acknowledgements

Foremost, I am so thankful to Allah for giving me the chance to carry on my study and making me the person who I am. My sincere thanks to my supervisor, Dr. Sherief Abdullah, for his continuous guidance and support throughout all the stages of the dissertation.

Also, I owe my deepest gratitude to the soul of my beloved father. I owe him all my academic and personal achievements. His support and his love were the main motivates for me to continue my master study.

Finally, this thesis would not have been possible without the support of my family members: my mother, my sister Basma, and my brothers Mohamed, Abdullah and Ahmed. Their belief on my abilities is the key factor of my success.

Declarations

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Deena Younis AboKhashan)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the importance of using data mining in modeling and prediction in the business domain. Also, it states the goals and the objectives of this thesis along with the research questions it aims to answer. Moreover, the structure of the thesis is detailed in this chapter.

1.1 Overview about Data Mining and Text Mining

Data mining is one of the fields of computer science that keens on the discovery of patterns and hidden information in large databases and presents the extracted useful information in understandable format. Data mining involves the use of artificial intelligence, machine learning, statistics, and databases (Leventhal, 2010).

Data mining has been used in several domains and industries and the business domain is one of its applications. Business world is becoming more competitive. Therefore, the use of data mining technology has become an integral part of business development. Many companies rely on the use of data mining techniques to allow dealing with massive data and to reveal the significant and unknown relationships between different features of data. This leads to support the decision-making process and consequently accelerates the business growth (Hotho, Nürnberger and Paaß, 2005).

Several data mining tasks are used by business applications such as: classification, clustering, prediction, and association. Each of these tasks has proven its beneficiary in developing businesses.

On the other hand, text-mining techniques are introduced to have the same objective as data mining in highlighting the unseen information in data. It involves many tasks such as; text categorization, information extraction, sentiment analysis, etc. However, text mining discovers patterns and relationships among natural language text unlike the structured databases that data mining tasks are relying on. And this is the crucial difference between data mining and text mining tasks (Han and Kamber, 2006).

1.2 Aims and Objectives

There are many factors that control the increase or the decrease of the price of real estate properties. Examples of these factors can be; location of a property, type of a property (villa, apartment, studio, etc.), area of the property, number of bedrooms, number of bathrooms, facilities, etc. The use of data mining techniques in real estate market can help greatly in making decisions for investment by businessmen. Also, it helps people to decide on selling, renting or buying real estate properties based on the findings and results of data mining recent sales records and transactions (Wedyawati and Lu, 2004).

The main goal of this thesis 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 renting, selling or buying. The research focuses on the residential properties in the city of Dubai, UAE.

In this thesis, linear regression data mining technique is used to predict the selling or renting price of a property. The experiments of the thesis depended on a dataset that was collected from several online real estate classifieds. The classifieds hold offers for apartments and villas that are for renting and selling in Dubai. Therefore, the features of the created dataset are having information about the characteristics of the offered real estate properties.

Some of the features are numerical structured features, and some are text features. The goal is to use the numerical features in building a linear regression model and in predicting the price. The results of linear regression experiments were considered as the baseline of the research. On the other hand, the unstructured text features are used in linear regression prediction after performing text-mining task on them to represent their text content in structured and numerical format. So, the new representation of data was added to the main linear regression experiment to investigate and measure the effect of the hidden information in the descriptive features on enhancing the price prediction task. The results of linear regression prediction before and after the use of text mining are analyzed to measure the improvement, if any

To the best of our knowledge, all researches that were conducted for the sake of predicting the price of real estate properties were performed using data mining techniques only and text mining is not tackled or tested before. Therefore, the contributions of this thesis can be viewed as:

- The research is the first scientific analysis of Dubai's real estate classifieds.
- The research is the first in using text mining to improve the accuracy of real estate property price prediction.
- The results of the experiments were verified based on real world dataset.

1.3 Research Questions

The primary goal of the thesis is to answer the following research questions:

- *How will traditional data mining techniques, relying only on structured data, perform when used on Dubai real estate classifieds?*
- *Will the use of text mining improve the accuracy of price predictions?*

1.4 Structure of the Thesis

The rest of the thesis is organized as follows: Chapter 2 provides background about the use of data mining in predicting the price of real estate properties. Chapter 3 describes and analyzes the used dataset in experiments. Chapter 4 includes a description about the used evaluation metrics and the architecture of linear regression and text mining models. Also, the key findings and the analysis of the experiments are detailed in chapter 4. Finally, the conclusion of the thesis and the suggested future work are discussed in chapter 5.

Chapter 2

Literature Review

This chapter gives background about data mining, text mining, linear regression, and text preprocessing. Also, it describes the used data mining tool in the research. The discussion of previous work in using data mining techniques for predicting the price of different real estate properties is included in this chapter as well.

2.1 Data Mining

The main target of data mining is to discover previously unknown relationships between several features and attributes of data. Data mining can support in decision-making process in several business domains and industries. It is considered as a collection of tools and techniques that work together to uncover the unseen relationships between data (Berson, Smith and Thearling, 2011).

One of the popular standards for describing the data mining process is CRISP-DM (Cross Industry Standard Process for Data Mining). It consists of six phases that are important for completing a data-mining task. Figure 2.1 shows the six phases as follows (Leventhal, 2010):

1. **Business understanding** defines the data-mining problem after studying the concerned domain.
2. **Data understanding** requires collecting data and to be familiar with its nature and its structure.

3. **Data preparation** covers all steps of preparing data for the data-mining task. It can be done several times till the data is fully ready for analysis. Data preparation includes data cleaning, features and records selection, feature reduction, etc.
4. **Modeling** uses different data mining techniques that are selected based on the needs and the requirements of the analysis. Therefore, data preparation steps might be revised to make sure that data meets modeling specifications.
5. **Evaluation** evaluates the built model to make sure it meets the predefined requirements of the data-mining task.
6. **Deployment** is a repeated process of applying the data mining task on data and continuously analyzing the results using appropriate data-mining software.

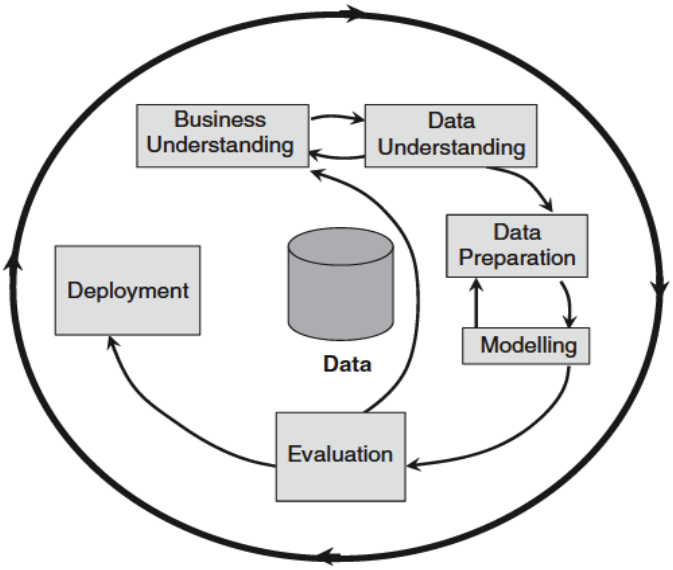

Figure 2.1 Cross Industry Standard Process for Data Mining, CRISP-DM (Azevedo and Santos, 2008)

When it comes to examples of data mining modeling, it is important to note that there are several techniques and models that can be used. 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 thesis, which is linear regression, LR.

2.1.1 Linear Regression

Linear regression is the most widely used modeling data mining technique when linear relationship between a dependent numeric feature and other independent numeric feature(s) is required to be modeled. It expresses the targeted dependent feature as a linear combination of other independent features:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k$$

where x is the predicted targeted feature, a_1, a_2, …. a_k, are the independent features' values, and w_0, w_1, w_2, … w_k, are weights (Witten and Frank, 2005).

When one independent feature is used to build the linear regression model, the task is called a simple linear regression. However, when two or more independent features are used, it is called Multiple Linear Regression, MLR. MLR is used in this research for both modeling the relationship between the input features of the real estate properties and the selling or renting price, and for prediction (Berson, Smith and Thearling, 2011).

2.2 Text Mining

Text mining is considered as one of the promising fields in investigating hidden information in the unstructured text data (Saravanan and Chonkanathan, 2010). Text mining merges the use of many different concepts, such as: data mining, machine learning, computational linguistics, statistics, and information retrieval. It is defined as the process of applying machine learning algorithms for the sake of discovering useful hidden patterns in text content of a database (Hotho, Nürnberger and Paaß, 2005). Both data mining and text mining share the major task of discovering hidden and previously unseen information in large datasets. However, the main difference between them is that data mining process works on structured data, while text mining works on natural language text, which is considered as unstructured data (Han and Kamber, 2006).

To apply text mining on a database, the unstructured text content should be converted into numerical values that reflect the importance of the contained words in the document (Roshni, Sagayam and Srinivasan, 2012). In addition, a prior step of preparing text content for text mining process should be taken into account. Several preprocessing methods can be applied according to the needs and the purpose of text mining.

2.2.1 Text Preprocessing

In order to prepare text data for preprocessing, tokenization task should be performed so that the content is split into series of words by eliminating numbers, punctuation marks, and other non-text characters. The end of the tokenization process creates a dictionary list that includes all words (tokens) of the document. The smallest token is a single character and a word is considered as a meaningful token.

After tokenization, several steps of preprocessing might be applied to the resulted tokens, such as (Hotho, Nürnberger and Paaß, 2005):

- **Filtering**: includes the stop-words removal. Stop-words filtering is the process of removing the tokens that hold little or no valuable information, such as, articles, conjunctions, adverbs, prepositions, etc.
- **Lemmatization**: aims to convert nouns into their singular form and verbs' forms to the infinite tense. To accomplish this task, a Part of Speech Tagger (POS) should be used. POS tagging results in assigning a part of speech to each token in the dictionary list. Examples of POS tags are: noun, verb, adverb, adjective, etc.
- **Stemming**: targets the removal of words' endings such as morphological and inflectional endings (Asilkan, Ismaili, and Nuredini, 2011). It allows reducing the size of words' dictionary. For example, the words wrote, writing and written have a common stem word, write.
- **n-gramming** generates n-items from the text in a document, where the items can be words or letters. N-gram can be viewed as having a slide widow over a text where n-words or n-letters can be seen at a time. For instance, when using one item, either a word or a letter, the n-gram model is called a uni-gram. Also, when two items are considered it is called bi-grams, and so on (Janicic, Keselj, and Tomovic, 2006).
- **Case Folding**: removes the distinctions between words in terms of capital and lower case letters. For example, converts all characters of the words to lowercase only. The aim is to make case insensitive comparisons between words (Chibelushi and Thelwall, 2009).

2.2.2 Term Weighting Scheme (Buana, Jannet and Putra, 2012)

After accomplishing text-preprocessing stage, the generated words (tokens) are given numerical values to reflect their significance in the document. One of the used term weighting methods is Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF is known for its simplicity and its powerful feature-weighting scheme.

TF-IDF scheme weights a term based on its occurrences in a document, which is referred to as Term Frequency, TF. Therefore, a term with high term frequency value is supposed to have high effect in the document. On the other hand, the Inverse Document Frequency (IDF) relies on number of documents that a term or a word occurs in. So, if a word occurs too much in several documents, it won't be considered as a characteristic any more. Hence, the weight of a term t in a document d, w(d, t), is calculated by multiplying the term frequency tf(d, t) by the inverse document frequency idf(t):

$$w(d,t) = tf(d,t) \times idf(t)$$
$$= tf(d,t) \times \log(\frac{N}{n_t})$$

where N is the total number of documents and n_t represents number of documents that contained the term t.

2.3 Tools

Choosing a suitable data mining software tool is a critical part of any data mining related research. Having variety of data preparation tools, analysis and modeling techniques, and results' visualization methods should be part of the tool. Also, having a mechanism of setting up a multistep experiment and run it in one process is recommended (Leventhal, 2010). Therefore, it is found that RapidMiner tool is the most appropriate tool to be used by this research. The following subsection gives an overview about RapidMiner tool.

2.3.1 RapidMiner

RapidMiner is one of the most widely used data mining tools in academic research and in industry. RapidMiner compromises from different modules for data mining, text mining, prediction analytics and business analytics (RapidMiner, 2010).

It is an open source solution that is completely implemented using Java, which makes it easy to run the tool on different operating systems. The most attractive characteristic of RapidMiner is its graphical user interface (GUI). Its GUI allows easy ways of observing and controlling working processes in

addition to visualizing the processes' results interactively. Moreover, the ability of inserting breakpoints anywhere in the process stages offers important way of monitoring the data flow between operators and validating transitional results (Asilkan, Ismaili, and Nuredini, 2011).

A typical RapidMiner process is divided into four stages as indicated in figure 2.2 (Jungermann, 2011):

1. **Retrieve** uses operators for loading and importuning data to be processed and analyzed. Examples: *Retrieve*, *Read Excel*, etc.

2. **Preprocessing** uses operators to prepare data for particular modeling technique. Examples: type conversion operators, filtering operators, attribute reduction and transformation operators, etc.

3. **Modeling** uses operators that help in creating models on the prepared data. Examples of modeling operators' categories are: classification and regression, clustering and segmentation, association and item set mining, etc.

4. **Evaluation** uses operators to apply and validate the performance of the applied model data. Every modeling technique has a customized evaluation operator to measure its performance.

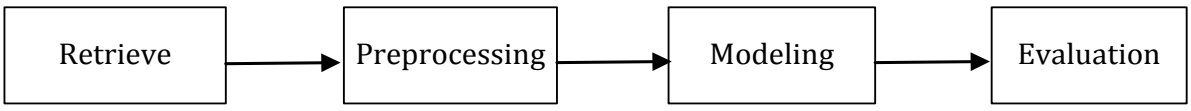

Figure 2.2 Basic phases of Rapid Miner process

2.4 Related Work

There are several research papers that work on examining the factors and the attributes that contribute to the prediction of real estate properties' prices using different data mining techniques.

According to Jaen (2002), decision tree and neural network techniques agreed on assigning the "sqrt living area" feature as a good predictor for the price of a house. In this research paper, 15 numerical features were used to represent the houses' characteristics plus a categorical feature that describes the address. However, the Address feature has been converted to a numeric data type. The dataset consisted of 1000 records that were collected from the houses' sales transactions in Miami, US. The data were organized and obtained from a Multiple Listing System (MLS) database. MLS is a huge property database that is developed by real estate agents in order to share information about the available properties for sale at a given time (Massey University, 2013). In the beginning of the experiments, the features were reduced to nine features by selecting the features that mostly affect the price prediction using a stepwise linear regression.

As the same as Jaen (2002) paper, the research paper of Wedyawati, and Lu (2004) has considered the MLS database as the main data source for collecting real estate property sales transactions. According to Wedyawati, and Lu (2004) the experiments covered 295,787 transactions from four cities in the US. Around 191 numerical features have been used. Visual Basic .NET software was used to run the linear regression model on the dataset plus Oracle Data Warehousing software was used to collect and organize the collected data 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.

Another data mining technique was used by (Guan, Levitan and Zurada, 2008) to anticipate the price of real estate residential properties. An Adaptive Neuro –Fuzzy Inference System (ANFIS) was implemented and tested over 360 records of past sales properties in Midwest, US. The dataset has 14 numerical features. The paper claims that it has the lead position of proposing such a system for predicting real estate properties' prices. The results showed that the ANFIS has almost the same performance results as multiple linear regression models. Also, the performance of the system has increased when a feature reduction technique was used such as Principle Component Analysis (PCA). It is clear that the used data in experiments were so small to represent the residential properties' characteristics.

According to Cacho (2010), several data mining techniques were used to predict the price of real estate properties in the city of Madrid, Spain. Examples of used techniques are: Naïve Neighborhood, multiple linear analysis, multilayer perceptron, M5 model trees, K-nearest neighbors, etc. It is found that the ensembles of M5 model trees outperform all other techniques. Also, the use of ensembles of M5 trees has decreased the relative error of prediction by 23%. On the other hand, 25,415 records that have 40 features were used in the experiments. The features gave detailed information about the characteristics inside and outside the apartments. For example: area of apartment, number of bedrooms, number of bathrooms, floor material, air conditioning, swimming pool, tennis court, garden area, garage, etc. In addition, some geospatial features were used, such as the distance to the nearest metro station and number of retail stores in a 500 meters radius from the apartment. The records were collected from several real estate portals in a certain date, 10th.Nov.2010. Moreover, it is important to mention that the dataset was divided into 21 subsets and each subset refers to an administrative district in the city of Madrid. Therefore, the experiments were carried out over each subset individually.

Unlike the pervious papers, Acciani, Fucilli and Sardaro (2011) paper examines the data-mining model on the sales of farms in some areas in Italy for the period from 2008 to 2010. The dataset consists of 169 sales transactions with 14 features (nine categorical and five continues features). The two used data mining techniques were Model Trees (MT) and Multivariate Adaptive Regression Splines (MARS) and they were implemented using WEKA software. The use of both techniques leads to the same performance in predicting farms' prices per hectare. However, MT and MARS outperform the standard Multivariate Linear Regression (MLR).

Another research paper focused on studying the prediction of prices of apartments in a city called Skopje, Macedonia. Among the three data mining techniques that were applied on a dataset of 1200 sales transactions, the logistic regression was found to be the superior in prediction accuracy over decision tree and neural network techniques. Like the other earlier mentioned papers, Dukova, Gacovski, Kolic and Markovski (2012) research also depended on structured numerical data in modeling and prediction.

To the best of our knowledge, the research area of predicting the price of a real estate property using data mining techniques has different trails based only on numerical structured characteristics of the property. This research aims to merge the use of data mining and text mining techniques in order to figure out the information that might be hidden on the descriptive text features of a real estate property and consequently leads to improve the accuracy of price prediction. Linear regression modeling is used as the main data mining technique in this thesis since it is one of the top popular techniques in the field of data mining in recent years (Rexer, 2013), and it has been used by several papers in real estate domain.

Chapter 3

Dataset Collection

This chapter describes the source of the collected data and the original features (attributes). Also, it shows how the dataset is divided into subsets in order to carry out the experiments on each subset individually. Moreover, the process of preparing data for data and text mining is clarified in this chapter.

3.1 Dataset Description

The original dataset is extracted from online real estate classifieds for residential properties in the city of Dubai, UAE in the period from 1st.June.2013 to 30th.June2013. The dataset records represent both apartments and villas that are either offered for sale or for rent in several different locations in Dubai. The dataset consists of 66,388 records, where 50% of the records are properties that are offered for selling and 50% are offered for renting.

Originally the dataset has 11 features and they have been reduced to seven, as four features were found to be not necessary for the analysis. The eliminated four features are show in Table 3.1.

Feature	Description
Source	Refers to the online advertising website
Posted	Date of posting the advertisement
URL	The website address of the real estate agency
Contact	Phone number of the advertiser

Table 3.1 Eliminated features from the dataset

The final used dataset in analysis has three numerical features, two nominal features and two text features. Table 3.2 shows the features, their data type, their description, and how many unique values that each feature has.

Feature	Data Type	Description	Unique Values
ID	Integer	A unique number to uniquely identify a record.	66,388
Type	Nominal	Distinguishes the type of the advertisement, either advertisement for selling or for renting a property.	4
Beds	Integer	Number of bedrooms in a property.	14
Location	Nominal	Each nominal value represents the location of the property in Dubai.	161
Title	Text	The title of the advertisement.	-
Description	Text	Extra description for the property as per the advertiser.	-
Price	Integer	The renting or selling price of the property.	4,821

Table 3.2 Used features in experiments

The four unique values of the "Type" feature have the following meanings as shown in table 3.3:

Value	Meaning
0	The property is an apartment that is offered for renting
1	The property is a villa that is offered for renting
2	The property is an apartment that is offered for selling
3	The property is a villa that is offered for selling

Table 3.3 Meanings of Type feature values

For the sake of analyzing different types of offered apartments and villas either for renting or for selling, the dataset is divided into six subsets. One subset for each type: renting apartment subset, renting villas subset, selling apartment subset, and selling villas subset. In addition, the renting subsets (Type0 and Type1) are merged to form one renting subset. The selling subsets (Type2 and Type3) are merged to form one selling subset.

Table 3.4 represents the six subsets along with number of records in each. It is observed that most of offered properties for renting are apartments as number of offered apartments for renting is 24,424 which is triple the number of villas that are offered for renting, 8,723. Also, number of offered

apartments for selling is almost double the number of villas that are offered for selling.

Subset Name	Type Value	Type	Number of Records
Type 0	0	Renting apartment	24,424
Type 1	1	Renting villa	8,723
Type 2	2	Selling apartment	22,672
Type 3	3	Selling villa	10,579
Renting	0, 1	Renting apartment/villa	33,137
Selling	2,3	Selling apartment/villa	33,251

Table 3.4 The six subsets after splitting the original dataset

Figure 3.1 shows the percentage of every type of real estate property in the original dataset.

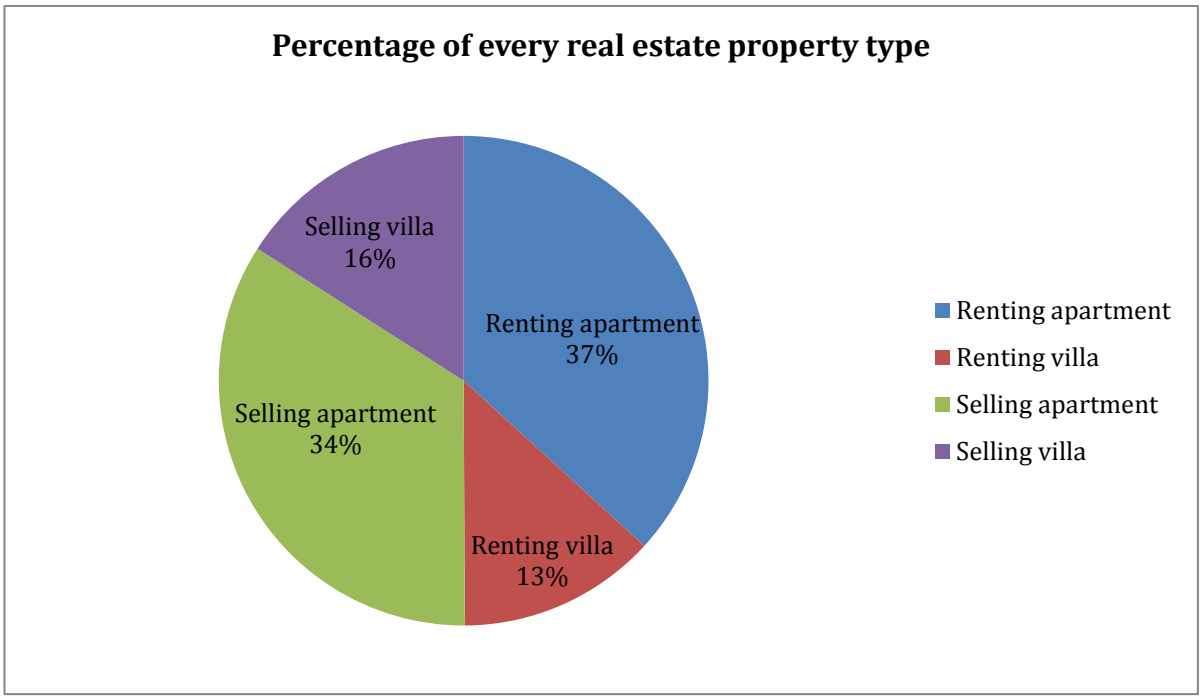

Figure 3.1 Percentage of every real estate property type in the original dataset

3.2 Data Preparation

Before starting the experiments, some steps of preprocessing are applied to the dataset in order to prepare it for analysis. Most of the preprocessing steps were applied on the text features to prepare them for text mining. Also, it is worth to state that no changes have been applied on the numerical features' values.

The following are examples of data preparation steps:

1. Records with price less than AED 10,000 are removed as they are considered as outliers.
2. The "Description" feature was manually inspected to identify and eliminate thousands of HTML tags and replace them with whitespaces, such as:
, <p>, , &gt, &lt, etc. Microsoft Excel is used to accomplish this task.
3. Remove the unwanted text from text features: "Description" and "Title" using Microsoft Excel. Examples of unwanted texts are:
   - Email addresses.
   - Website addresses.
   - Text such as, please contact, for more information, for further information, for international call please dial, etc.

Chapter 4

Experimental Setup

This chapter describes the used evaluation metrics in verifying the results of experiments. Also, it gives detailed information about the implementation of linear regression experiments and the used text mining model.

4.1 Evaluation Metrics (Witten and Frank, 2005)

The main data mining technique used in this research is linear regression. Linear regression is the most suitable data-mining model to predict a continuous value feature. Therefore, linear regression modeling was chosen to predict the price of real estate properties in this thesis.

Two evaluation metrics of linear regression modeling are used in this thesis: Root Mean Squared Error, RMSE, and Correlation Coefficient, CC.

Let the predicted values of the targeted feature be denoted as $p_1, p_2, \ldots\ldots, p_n$ and the actual values are $a_1, a_2, \ldots\ldots, a_n$ where n is total number of records in the dataset. Then, the evaluation metrics can be explained as:

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

  It is the most commonly used metric in evaluating linear regression model performance. It gives an indicator about how does the predication deviate from the actual value.

  It can be calculated using the formula: $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$

It is clear that the less the value of the RMSE, the better the accuracy of prediction.

- **Correlation Coefficient (CC)**

  It measures how strong is the linear relationship between regular independent features and the targeted dependent feature. Its value ranges from -1 and 1, so that when the value of the correlation coefficient is close to 1, it means that the value of the targeted feature increases as the values of the other features increase and vice versa.

  Also, when the correlation coefficient value is close to zero, it means there is a little linear relationship between regular features and the targeted one. On the other hand, having a negative correlation value implies that when the independent features' values tend to decrease, the targeted feature increases.

  The correlation coefficient can be calculated using the formula:

$$CC = \frac{S_{PA}}{\sqrt{S_P S_A}}$$

where: $S_{PA} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(a_i - \bar{a})}{n-1}$, $S_P = \frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n-1}$, and $S_A = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}$

4.2 Experimental Setup

This section describes the environmental setup of the implemented experiments using Rapid Miner version 5. It describes the configurations and the implementations of:

- Linear regression model.
- Text mining model.
- Final model that integrates both linear regression and text mining.

Before running the experiments, each feature has been assigned a specific role in RapidMiner. The roles of features are used by different operators to define what the features can be used for (RapidMiner, 2010). For example, the "ID" feature is assigned the role ID which means it plays the role of being an identifier for a record. Also, "Price" feature is considered as the targeted feature for prediction, therefore it is given a label role. As it can be seen in table 4.1, the remaining five features have been given a regular role. Regular role is given to features that have no special function and they are used to describe the records (RapidMiner, 2010).

Feature	Role	Data Type
ID	ID	Integer
Type	Regular	Nominal
Beds	Regular	Integer
Location	Regular	Nominal
Title	Regular	Text
Description	Regular	Text
Price	Label	Integer

Table 4.1 Roles of features

It is important to note that the default configurations of the used operators in RapidMiner5 are used in the experiments unless otherwise stated throughout the thesis.

4.2.1 Linear Regression Experiments

In this thesis, the linear regression method is called: Multiple Linear Regression, MLR, as multiple explanatory features are used in modeling the relationship between input features and the targeted feature. Also, the explanatory features are used in making prediction.

The explanatory features are the input features that describe a villa or an apartment. For the LR experiments, only the numeric features can be used as depicted in table 4.2. Therefore, "Type" and "Location" features should be converted to numerical data type.

Feature	Role	Data Type	Description
ID	ID	Integer	A unique number to uniquely identify a record.
Type	Regular	Nominal	Distinguishes the type of the advertisement, either advertisement for selling or for renting a property.
Beds	Regular	Integer	Number of bedrooms in a property.
Location	Regular	Nominal	Location of the property
Price	Label	Integer	The renting or selling price of the property.

Table 4.2 Used features in LR experiments

With reference to the description of typical RapidMiner process as mentioned in section 2.3.1, the linear regression experiment is built based on the main four stages: retrieve, preprocessing, modeling, and evaluation. Figure 4.1 illustrates the implemented RapidMiner process for the LR experiment:

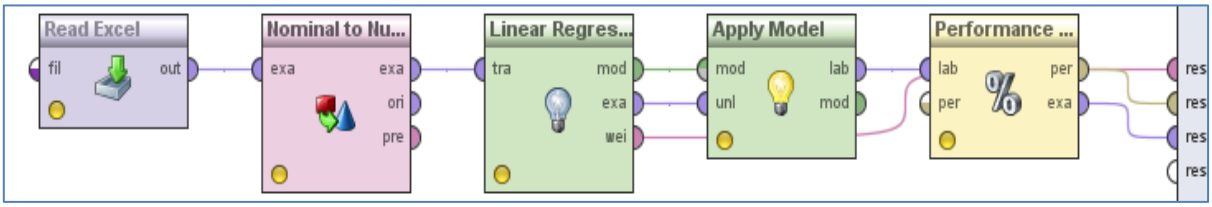

Figure 4.1 Linear regression model process

1. **Retrieve**: a dataset is loaded to the process using *Read Excel* operator. *Read Excel* operator allows importing data to process, selecting required features, and assigning data types and roles to features.

2. **Preprocessing**: the only preprocessing step in this experiment is the conversion of nominal features into numerical data type in order to be

compatible with the requirements of linear regression modeling. A dataset is input to *Nominal to Numerical* operator, which performs the data type conversion plus mapping all values of the "Type" and "Location" features into numerical ones.

For example, the operator makes each unique value of the Location feature as a feature by itself in the dataset. For a Location value equals to 92, a new feature is created and it is called "Location=92". So, if a record has the original Location's value equals 92, then the value of the new feature "Location=92" equals 1. Otherwise, it equals to 0. As a result of this step, number of regular features in a dataset is increased as illustrated in table 4.3.

Table 4.3 shows the final number of regular features that were used in the LR experiments for the six data subsets.

Dataset	No. of numerical regular features
Type 0	122
Type 1	92
Type 2	87
Type 3	76
Renting	143
Selling	102

Table 4.3 Number of regular numerical features in every dataset

3. **Modeling**: a dataset with the new features is fed into *Linear Regression* operator, which is responsible for building and calculating the LR model

4. **Evaluation**: to apply the learnt LR model on the dataset and to predict the property's price, the *Apply Model* operator is used. On the other hand, the performance of LR model in prediction is evaluated and verified using *%Performance (Regression)* operator. The *%Performance (Regression)* operator is customized to measure the performance of regression models only. Therefore, the selection of the evaluation metrics; Correlation Coefficient (CC) and Root Mean Squared Error (RMSE) is made in this stage.

The linear regression experiment is conducted for each dataset subset individually. Therefore, it is repeated six times and the values of the CC and RMSE are recorded for each data subset. Note that the results of LR experiments play the role of being the baseline of this research. These results are to be compared with the results of applying LR again after performing text mining task on text features and involving them in the experiments.

4.2.2 Text Mining Model

In order to be able to build the text-mining model of the research, the *Text Mining Extension* plug-in is installed to RapidMiner5. The used features for text mining experiments are shown in table 4.4

Feature	Role	Data Type	Description
ID	ID	Integer	A unique number to uniquely identify a record.
Title	Regular	Text	The title of the advertisement.
Description	Regular	Text	Extra description for a property as per the advertiser.
Price	Label	Integer	The renting or selling price of the property.

Table 4.4 Used features in text mining experiments

The purpose of building the text-mining model is to discover the effect of the hidden information involved in "Title" and "Description" features that might enhance the accuracy of predicting the property's price and minimizes the root mean squared error, RMSE.

Therefore, a dataset is passed through a text-mining process in order to convert the content of the text features into numerical values. For this purpose, the operator *Process Document from Data* is used as seen in figure 4.2. This operator is used to generate word vectors from text features (RapidMiner, 2010). The vector creation method used in experiments is TF-IDF, term frequency-inverse document frequency.

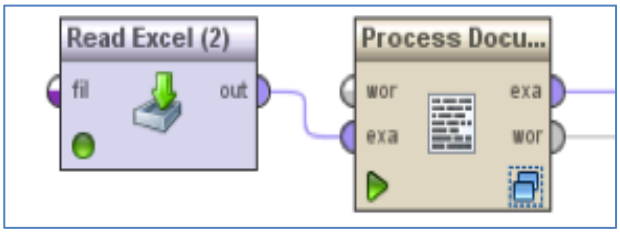

Figure 4.2 Text mining main process

Process Document from Data operator merges the content of both text features, "Title" and "Description", to make one single text feature for each record. So, in order to calculate the TF-IDF value for each word in the resulted merged text feature, the sub process of the operator *Process Document from Data* is built to have certain phases as shown in figure 4.3.

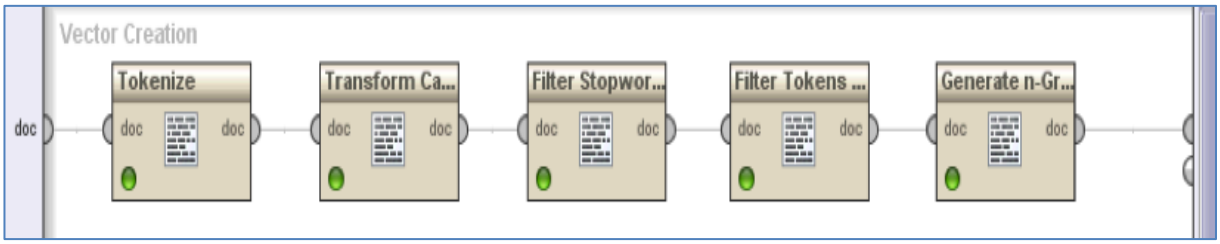

Figure 4.3 Text mining sub process phases

The following are the phases of preparing the text features for TF-IDF calculations:

1. Tokenize the text content by splitting the text into sequence of tokens, words. The *Tokenize* operator is used for this purpose.
2. Convert the characters of every single word (token) to lowercase using *Transform Cases* operator.
3. Filter and remove any stop-word token by comparing every token to a predefined stop-word list. If the token matches any entry in the stop-word list, then it will be eliminated. The operator *Filter Stopwords (English)* is responsible for performing this task.
4. *Filter Tokens* operator is used to filter tokens by their length. The default minimum number of characters of a token is set to 2.
5. Create term n-Grams of tokens. A term n-Grams is a series of successive tokens of length n. The default value of n is set to 1.

The sequence of operations of the text mining process was determined after repeating experiments several times and making iterative analysis to find out which sequence gives the best results.

After performing the sub process, the *Process Document from Data* operator calculates the Term Frequency-Inverse Document Frequency (TF-IDF) for each token. This leads to generate a feature for every word and the feature's value is the TF-IDF of the token. For example, if the token appears in a record, its value will be set to its TF-IDF. Otherwise, the value of the feature equals to zero.

The output of the text-mining model is thousands of new numerical features. Table 4.5 displays number of generated features for each data subset after passing text features through text mining process.

Also, table 4.5 shows the final total number of numerical features for each data subset. It is clear that Type0 subset has the greatest number of features when compared with the other 3 data subsets Type1, Type2, and Type3. It has around 13,000 features.

Dataset	Number of numerical features generated by text mining	Number of Numerical regular features before text mining	Total Number of Features
Type 0 –Renting Apartments	12,861	122	12,983
Type 1- Renting Villas	8,025	92	8,117
Type 2- Selling Apartments	9,549	87	9,636
Type 3- Selling Villas	9,060	76	9,136
Renting	27,240	143	27,383
Selling	28,354	102	28,456

Table 4.5 Number of features generated by text mining for each data subset compared to the original number of regular numerical features

4.2.3 Text Mining plus Linear Regression

The used model for performing linear regression modeling on a dataset using all original numerical features plus the generated features after text-mining text features is shown in figure 4.4.

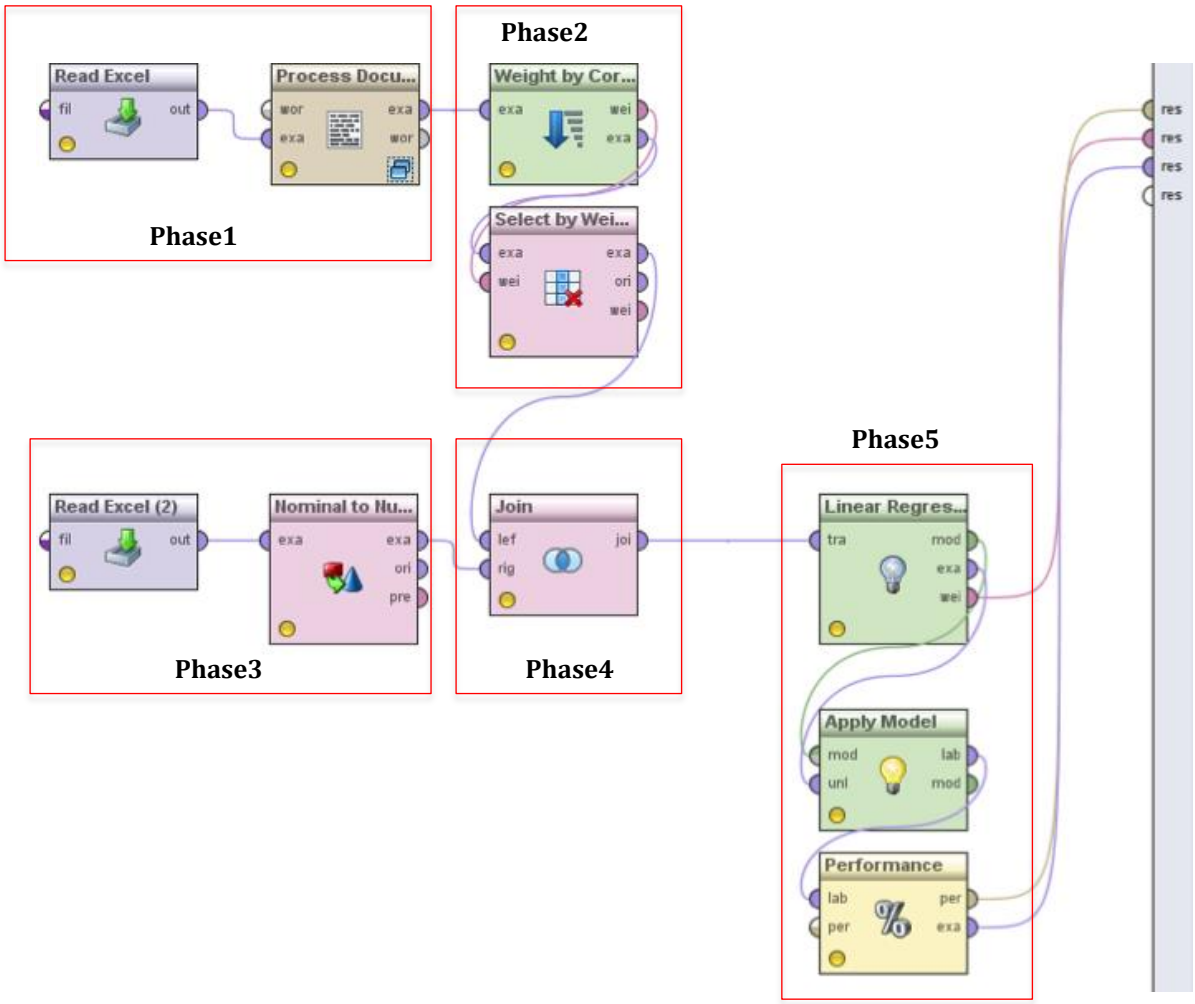

Figure 4.4 Final complete model: linear regression after text mining

A dataset is divided into two portions according to the selected features for experiments as follows:

- **Portion1**: text features are selected; “Title” and “Description” plus “ID” and “Price” features.
- **Portion2**: numerical and nominal features are selected; “Beds”, “Location”, and “Type” plus “ID” and “Price” features.

The final complete model consists of the following phases as shown in figure 4.4:

Phase1: input Portion1 of the dataset to the text mining process. Refer to subsection 4.2.2 for more details about the phases of text mining process which are implemented using *Process Document from Data* operator. As a result of text mining, thousands of numerical features are generated to map the text content, words, into their numerical values, TF-IDF.

Phase2: the output of phase1 is thousands of features. Therefore, the dimensionality of the dataset is going to be very high. It is significant to note that the free version of RapidMiner5 allows allocating limited computing resources for implemented processes. RapidMiner5 won't be able to allocate huge memory and processing resources for the sake of completing experiments that have thousands of features. Consequently, a feature reduction method is implemented as shown in figure 4.5.

The implemented process of feature reduction has two steps as follows:

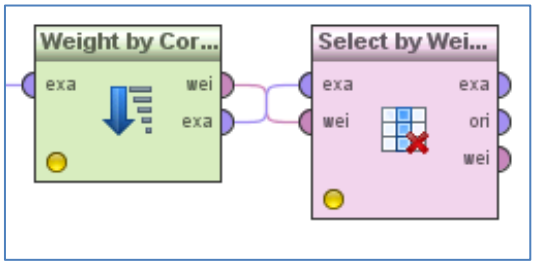

Figure 4.5 Feature reduction process

- **Step1:** a dataset with thousands of features is fed into an operator called *Weight by Correlation.* The purpose of using this operator is to calculate and assign weight for each feature with respect to the targeted label feature by using correlation. Therefore, a feature is more relevant to be considered when it weight is high. The output of *Weight by Correlation* operator is the assigned simple correlation weight for each feature of the input dataset.

- **Step2:** As the output of step1 is still thousands of features, the *Select by Weight* operator is used to allow the user to control and select

which features to be used in the subsequent steps of the experiment based on specified criteria with respect to the features' weights.

For example, a criterion would be: only select features with weight values greater than or equal 0.7. Note that the weights' values are normalized and they range from 0 to 1.

Phase3: input Portion2 of the dataset into *Nominal to Numerical* operator in order to generate numerical features out of the nominal features "Type" and "Location". This preprocessing phase is required as the linear regression model can only use numerical features.

Phase4: merge the two datasets generated in phase2 and phase3 using the *Join* operator based on the match of the ID feature value between the two input datasets. Therefore, the output of this phase is the complete dataset with all its features are converted to numerical data type.

Phase5: the joined dataset is the final dataset to be input to the linear regression model. The output of phase5 is the new measured evaluation metrics, CC and RMSE, which are going to be compared to the results of LR experiments that were conducted before without the use of text features.

The experiment is repeated several times for each dataset. The difference between each repetition of the experiment on a dataset is the number of text mining features to be involved. Number of features is increased every time the experiment is conducted by decreasing the targeted weight used in feature reduction process, phase 2.

For example, when the experiment is run for the first time, the *Select by Weight* criteria is to select features with weight values greater than or equal 0.7. Next, in the second conduction of the experiment, the criterion is changed to be: select features with weight values greater than or equal 0.5. By reducing the targeted weight, more features are selected and involved in the experiment.

It should be noted that the process of increasing number of selected text features and involving them in the linear regression experiments have been

stopped when the experiments consumed hours of running without significant changes in the performance of the LR evaluation metrics, root mean squared error (RMSE) and correlation coefficient (CC).

Chapter 5

Experimental Analysis

This chapter intends to discuss and analyze the results of linear regression experiments before and after using text mining.

5.1 Linear Regression Results and Analysis before Text Mining

After building the LR model and using it for prediction, the evaluation metrics are calculated. Table 5.1 displays the average prediction for the "Price" feature for every dataset, the Root Mean Squared Error (RMSE), and the Correlation Coefficient (CC).

Dataset	Average Prediction of Price Feature	RMSE	CC
Type 0	119,458.618 +/- 125,640.672	110,263.850	0.479
Type 1	271,617.039 +/- 243,797.430	204,650.546	0.543
Type 2	3,259,776.742 +/- 9,084,079.578	8,676,287.332	0.296
Type 3	7,304,448.683 +/- 7,963,778.255	4,756,663.730	0.802
Renting	159,511.186 +/- 178,230.446	144,483.337	0.586
Selling	4,546,612.760 +/- 8,943,881.112	7,738,338.211	0.501

Table 5.1 Performance measurements of linear regression before text mining

As shown in Table 5.1, the average predictions of the "Price" feature for Type0 and Type1 are lower than the ones for Type2 and Type3. This is expected because Type0 and Type1 subsets represent renting price of real estate property and Type 2 and Type3 subsets represent selling price.

On the other hand, Type2 dataset has the greatest RMSE (8,676,287.332) and the lowest linear correlation between regular features and the price feature (0.296). Also, although the dataset of Type3 has the highest correlation between features (0.802), its RMSE is relatively high (4,756,663.73). Having high correlation between features means that there is a strong linear relationship between the predicated price and the other regular features so that when values of regular features increase the price value increases and when values of regular features decrease, the price value decreases. The graphical representations of the comparisons between the six datasets in terms of the calculated RMSE and correlation coefficient values are displayed in figure 5.1 and figure 5.2.

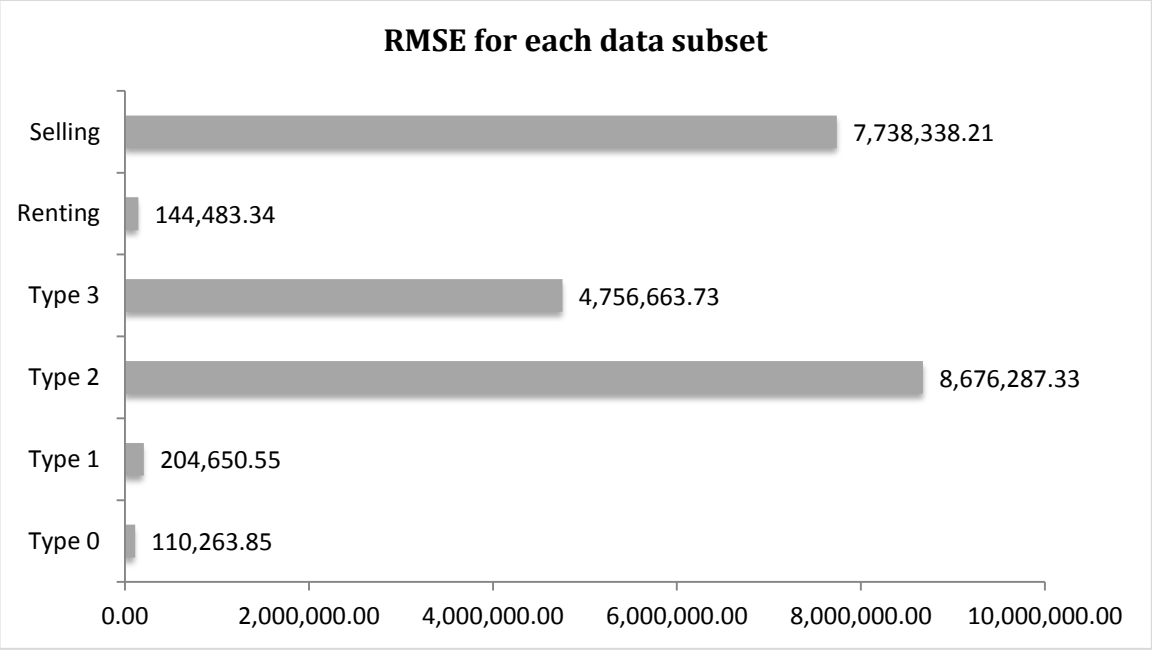

Figure 5.1 Values of RMSE after applying the LR modeling experiments

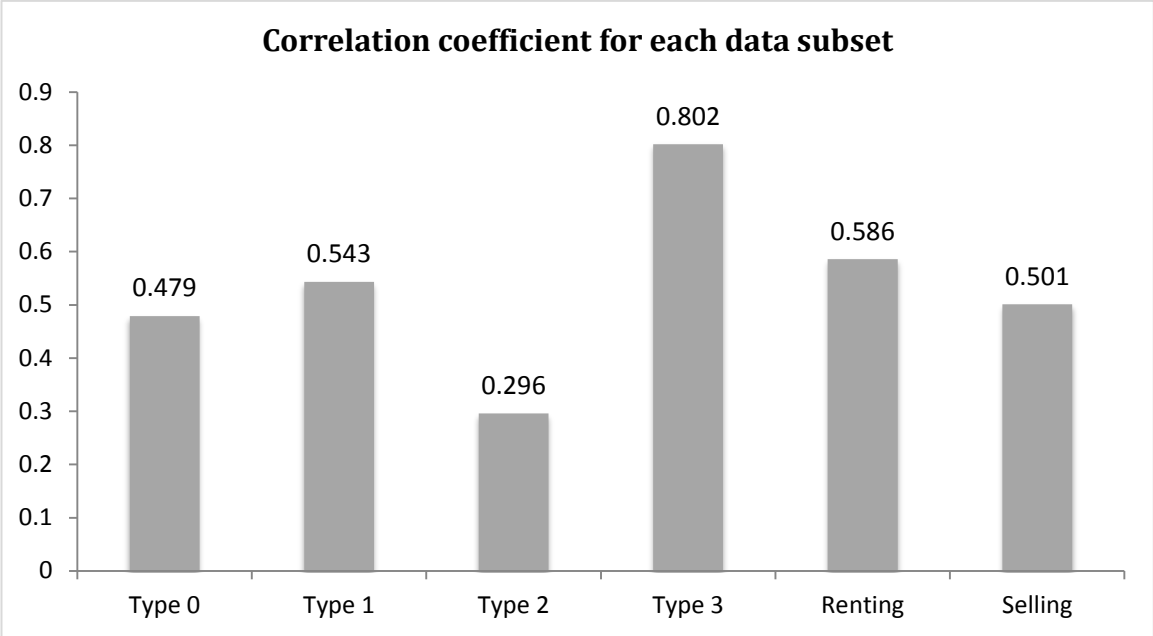

Figure 5.2 Values of correlation coefficient after applying the LR modeling experiments

Linear Regression (LR) model is built based on assigning a weight for each feature in the dataset to reflect its effect on the price. For example, when a feature has a positive weight value, it means that the feature works toward increasing the price. Similarly, having a feature with negative weight has the effect of decreasing the price.

After building the LR model for each data subset and analyzing it, it is found the Locations' features have the great effect on increasing or decreasing the price. The following are the observed notes about the effect of locations features over each data subset:

- **Type 0 subset (renting apartments)**: The top three locations in Dubai where the apartments' renting prices are the highest are: Al-Barari, Mirdiff, and Arabian Ranches. In contrary, the locations where the renting prices are the lowest are: Al-Qouz, Lotus Hotel and Sheikh Hamdan Colony locations.
- **Type 1 subset (renting villas)**: The top three locations were the villas' renting prices are the highest are: Iranian Hospital Area, Emirates Hills,

and Al-Wasel Road. Also, the locations that lead to reducing the renting prices are: Alqusais, Al-Jafiliya and Oud Al-Muteena locations.

- **Type2 subset (selling apartments):** Alnahda, Alqouz, and Albarsha are the top three locations in Dubai in terms of having highest selling prices for apartments, while International Media Production Zone, Dubai Lagoon and Falcon City have the lowest renting prices.
- **Type3 subset (selling villas):** The top three locations in terms of having the highest villas' selling prices are Palm Jebel Ali, Emirates Hills, and Dubai Lifestyle City, while Aljafiliya, Alqusais and Alrashidiya have the lowest selling prices.
- **Renting subset:** Recall that Renting dataset combines renting both apartments and villas datasets, it is observed that the Type feature has positive effect in increasing the price when it refers to renting villas (Type1), and has an effect of decreasing the price when it refers to renting apartments (Type0) and this meets the expectations as villas' renting prices are greater than the apartments' renting prices. In general, the locations that have the highest renting prices are: Iranian Hospital Area, Emirates Hills, and Al Wasel Road and the lowest prices can be found in: Aljafiliya, Oud Al Muteena, and Alrashidiya.
- **Selling subset**: Recall that Selling dataset combines both selling villas and selling apartments datasets, it is observed that the Type feature has positive effect in increasing the price when it refers to selling villas (Type3), and has an effect of decreasing the price when it refers to selling apartments (Type2) and this meets the expectations again as usually villas' selling prices are greater than the apartments' selling prices. In general, the locations that have the highest residential real estate selling prices are: Alnahda, Palm Jebel Ali, and Emirates Hills and the lowest prices can be found in: Aljafiliya, and Alrashidiya.

Table 5.2 summarizes the top locations that lead to increase or/and decrease the prices' values. When comparing the results of LR experiments on Renting

and Selling datasets, it is found that some places in Dubai are expensive in both; renting and selling residential real estate properties such as Emirates Hills. In addition, Aljafiliya and Alrashidiya are considered the cheapest places for renting or selling real estate properties regardless of the type of the property, either apartment or villa.

	Locations lead to decrease prices	**Locations lead to increase prices**
Type0 Renting Apartments	1. Al-Qouz 2. LotusHotel 3. Sheikh Hamdan Colony	1. Al-Barari 2. Mirdiff 3. Arabian Ranches
Type1 Renting Villas	1. Alqusais 2. Al-Jafiliya 3. Oud Al-Muteena	1. Iranian Hospital Area 2. Emirates Hills 3. Al-Wasel Road
Type2 Selling Apartments	1. International Media Production Zone 2. Dubai Lagoon 3. Falcon City	1. Alnahda 2. Alqouz 3. Albarsha
Type3 Selling Villas	1. Aljafiliya 2. Alqusais 3. Alrashidiya	1. Palm Jebel Ali 2. Emirates Hills 3. Dubai Lifestyle City
Renting	1. Aljafiliya 2. Oud Al-Muteena 3. Alrashidiya	1. Iranian Hospital Area 2. Emirates Hills 3. Al Wasel Road
Selling	1. Aljafiliya 2. Alrashidiya	1. Alnahda 2. Palm Jebel Ali 3. Emirates Hills

Table 5.2 List of top locations that have either increase or decrease effect on price prediction in LR experiments

Based on the above findings and analyzed results, it can be concluded that although the use of linear regression data mining technique for price prediction based on the structured numerical features has resulted in high RMSE values for some data subsets, the LR modeling reveals the key factor that controls the price of real estate property in Dubai. The factor is its location.

5.2 Linear Regression Results and Analysis after Text Mining

The linear regression experiments were repeated several times after text mining the text features, “Title” and “Description”, for each dataset. The main target of repeating the experiment for the same dataset is to add more features gradually and to record and to compare the results for each repetition. Also, having limited computing resources is an important factor that prevents the use of all generated features from text mining.

Moreover, the experiments have been repeated to study the effect of using uni-grams and bi-grams in the text mining process for generating features. This means that the experiments were repeated twice for each dataset. First experiment considered uni-grams tokens where n-grams equal to one (default) and the second experiment considered bi-grams tokens where n-grams equals to two. Table 5.3 illustrates the results of linear regression modeling and prediction before and after text mining.

Also, the results of using uni-grams and bi-grams are demonstrated for each dataset. It is important to remember that adding features to an experiment is stopped when its running time last for hours without noticeable improvement in the prediction task performance.

It is observed that the performance of linear regression has improved greatly after the inclusion of text mining in the experiments. Also, when number of n-grams in the experiments is increased from uni-grams (deafult) to bi-grams, the RMSE is declining faster. Figure 5.3, figure 5.4 and figure 5.5 show the effect of using text mining on the reduction of RMSEs and the increase of the correlation coefficients for all datasets.

Type0 – Renting apartments dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	w>=0.2	w>=0.3
No. of Features	122	321	343
RMSE	110,263.850	82,044.01	61,715.12
CC	0.479	0.757	0.871
Type1 – Renting villas dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	w>=0.7	w>=0.16
No. of Features	92	558	548
RMSE	204,650.546	99,021.76	89,710.59
CC	0.543	0.914	0.930
Type2 – Selling apartments dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	w>=0.06	w>=0.25
No. of Features	87	503	407
RMSE	8,676,287.332	5,490,825.94	3,853,490.15
CC	0.296	0.797	0.906
Type3 – Selling villas dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	w>=0.15	w>=0.2
No. of Features	76	331	508
RMSE	4,756,663.730	3,778,524.95	3,750,734.50
CC	0.802	0.88	0.882
Renting apartments and villas dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	>=0.2	>=0.2
No. of Features	143	237	361
RMSE	144,483.337	105,307.09	96,334.73
CC	0.586	0.807	0.841
Selling apartments and villas dataset			
	LR	Text Mining + LR	
		Uni-gram	Bi-gram
Selected Weight	-	>=0.1	>=0.1
No. of Features	102	239	283
RMSE	7,738,338.211	5,788,073.98	4,440,108.38
CC	0.501	0.762	0.868

Table 5.3 Results of linear regression experiments before and after text mining

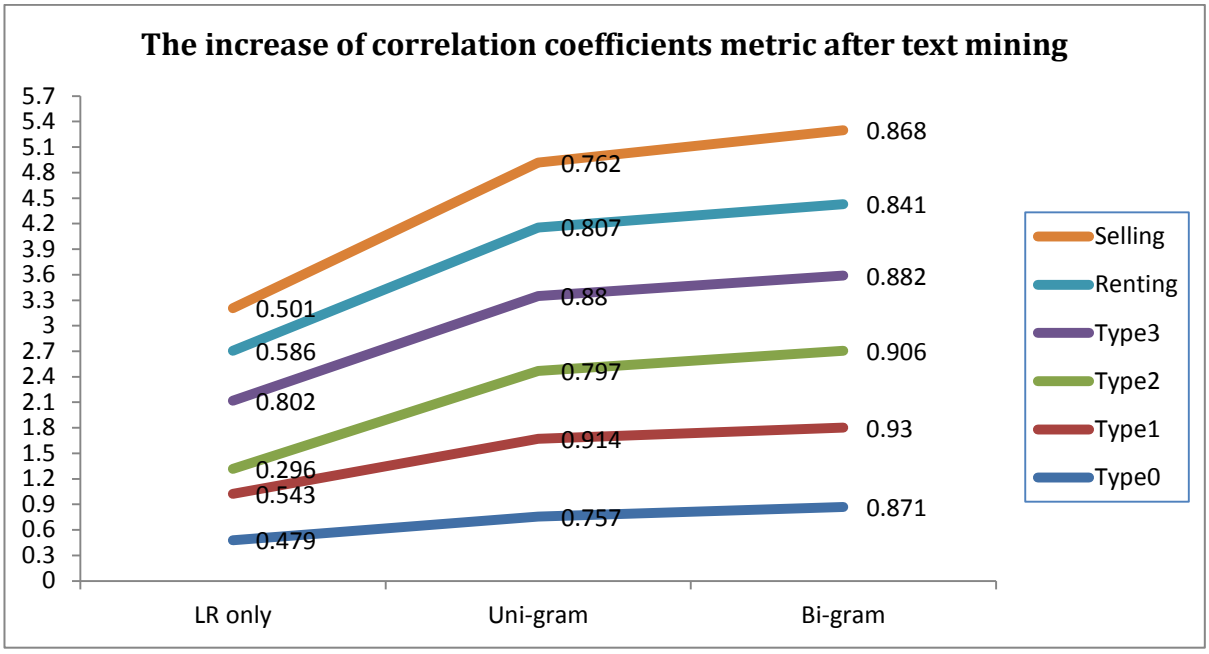

Figure 5.3 Comparison between resulted correlation coefficients before and after text mining

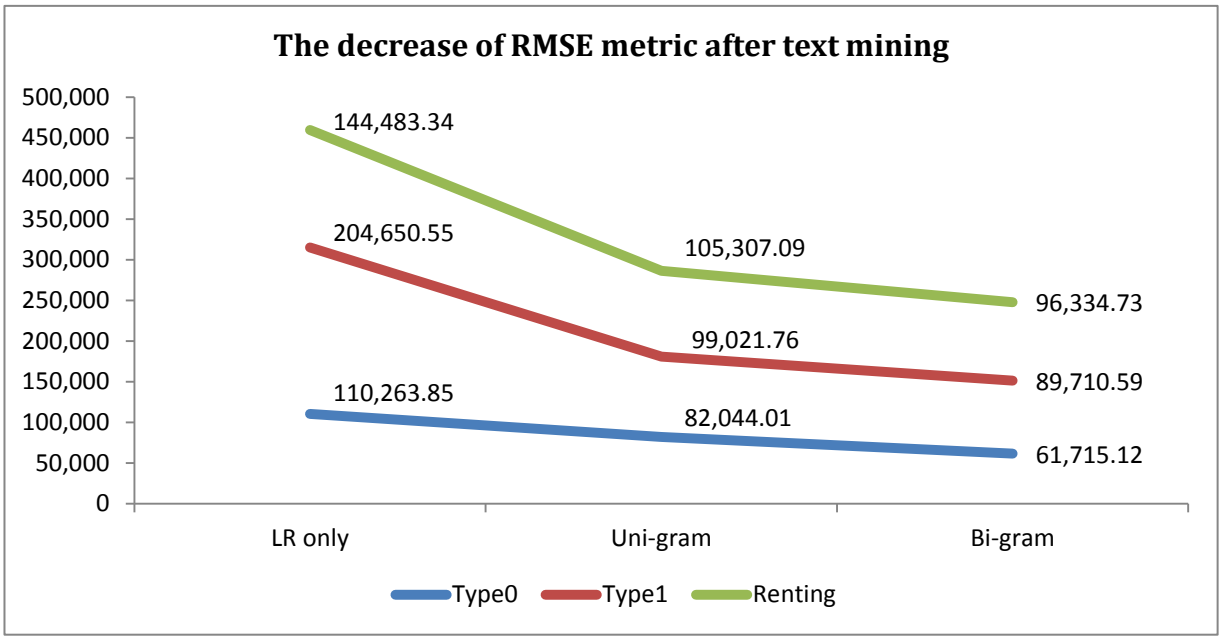

Figure 5.4 Comparison between resulted RMSEs before and after text mining for Renting, Type0 (renting apartments), and Type1 (renting villas) datasets.

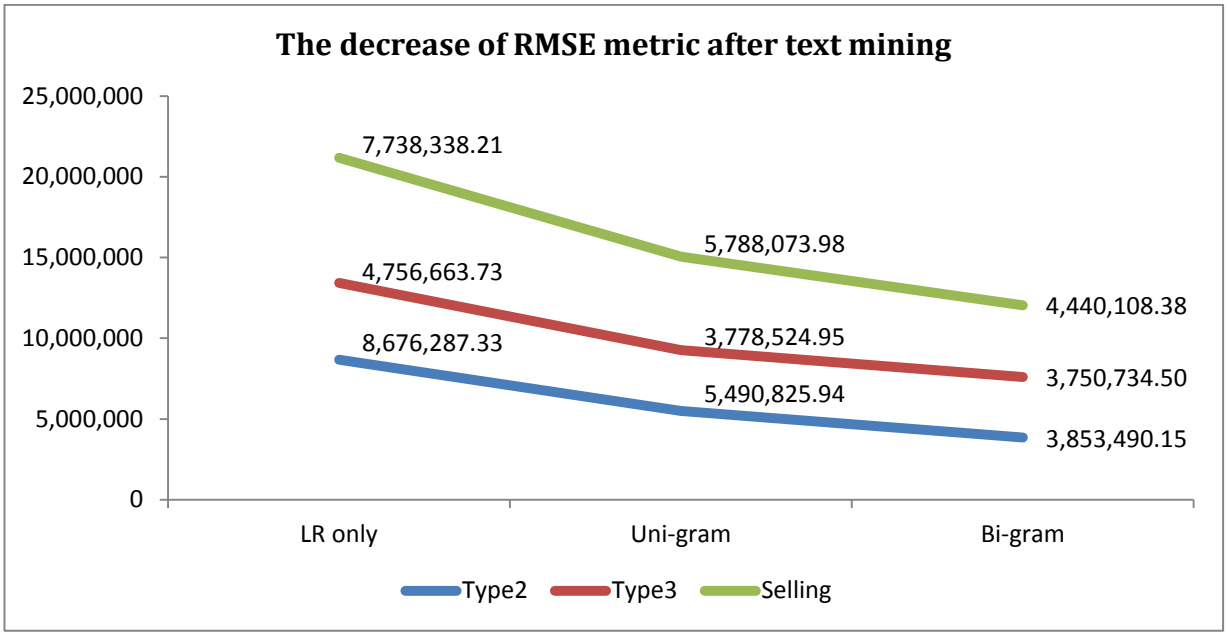

Figure 5.5 Comparison between resulted RMSEs before and after text mining for Selling, Type2 (selling apartments), and Type3 (selling villas) datasets.

Based on listed findings in tables 5.4 and 5.5, the following points are noted about the results of linear regression prediction task before and after using text mining:

Dataset	Root Mean Squared Error – RMSE			
	Before	After	Difference	Percent of Decrease
Type 0- Renting apartment	110,263.85	61,715.12	48,548.73	44.03%
Type 1- Renting villas	204,650.55	89,710.59	114,939.96	56.16%
Type 2- Selling apartments	8,676,287.33	3,853,490.15	4,822,797.18	55.59%
Type 3- Selling villas	4,756,663.73	3,750,734.50	1,005,929.23	21.15%
Renting	144,483.34	96,334.73	48,148.61	33.32%
Selling	7,738,338.21	4,440,108.38	3,298,229.83	42.62%

Table 5.4 Comparison between values of RMSE of linear regression before and after adding text features

Dataset	Correlation Coefficient-CC		
	Before	After	Percent of Increase
Type 0	0.479	0.871	81.84%
Type 1	0.543	0.930	71.27%
Type 2	0.296	0.906	206.08%
Type 3	0.802	0.882	9.97%
Renting	0.586	0.841	43.52%
Selling	0.501	0.868	73.25%

Table 5.5 Comparison between values of CC of linear regression before and after adding text features

- The root mean squared errors (RMSEs) for all datasets have decreased noticeably after adding the text features to LR experiments.

  For example, the RMSEs of the datasets Type1 (renting villas) and Type2 (selling apartments) have almost the same decrease of 56%. While the RMSE percent of decrease for Type3, selling villas, was the smallest, 21.15%, from 4,756,663.730 to 3,750,734.50.

- The correlation coefficients between the input features that represent the characteristics of apartments or villas and the "Price" feature have increased in all experiments for all datasets.

  For instance, the highest increase in the linearity between input features and "Price" feature has observed for dataset Type2, selling apartments, where it was 0.296 and it became 0.906 after adding text features with increase of 206.08%. On the other hand, the correlation coefficient of Type3 dataset, selling villas, has increased by 9.97% only, from 0.802 to 0.882.

- Whenever more text features are added to an experiment, the RMSE is falling down and the correlation coefficient is increasing. This means that adding more features that were generated by text mining to linear regression experiments leads to improving the price prediction.
- The top text features that affect the increase or decrease of the predicted price are mostly the ones that refer to locations in Dubai.

Figures 5.6 and 5.7 show clearly the improvements in linear regression prediction results after using text mining, the increase in correlation between the input features and the price and the reduction of root mean squared error, RMSE for all data subsets.

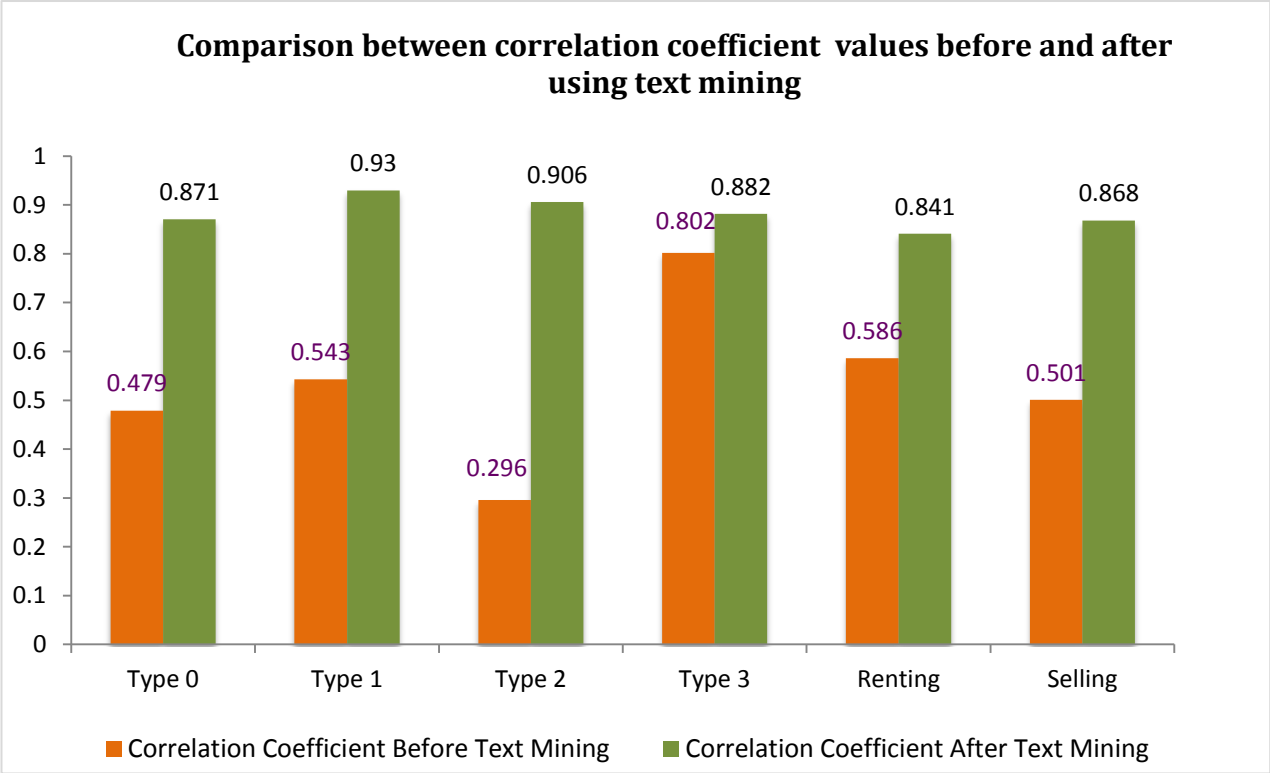

Figure 5.6 Comparison between correlation coefficient values before and after text mining in LR experiments

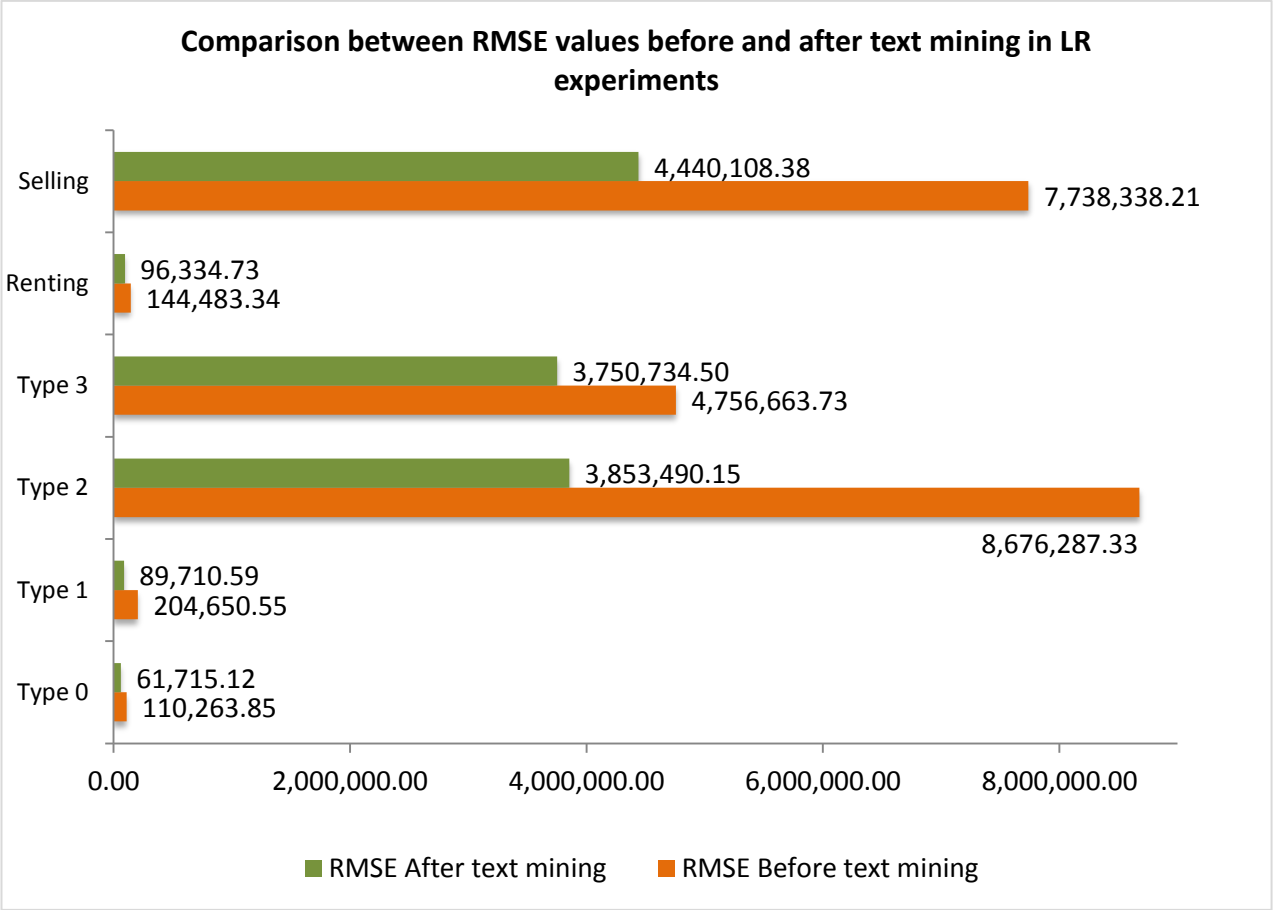

Figure 5.7 Comparison between RMSE values before and after text mining in LR experiments

The following table, 5.6, has some of the top text features that lead to the increase of the predicted price in linear regression experiments. It is noticed that the top text features are either refer to locations in Dubai or refer to extra description for a property.

Dataset	Words refer to locations	Words refers to services	Other Words
Renting Apartments Type 0	diamond_atlantis, diamond_palm, uptown, barsha_hotel, plan_jumeirah.	Service_ valet, facilities_swimming pool, balcony_restaurant, kitchen_commercialheavy, facilities_decorated, investment_maids, upgraded_maids, (were maids means maid room)	Sale, good_investment, sale_binayah, binayah_info, call_binayah, (where binayah is a name of a real estate agency), lavish_lobby, room_lavish, (were Lavish is an adjective).
Renting Villas Type 1	type_novella, novella_village, residences_shores, shores_palm, embody_arabian dubai_embody , arabian_european, village_plot, meadows_favored, layouts_meadows, kempinski_residence, sequim_manara, golfer_paradise, june_signature, paradise_emirates	Wardrobes, furniture_total, beach_car, min_driving , driving_mall (refers to the time to reach a mall)	million_engel million_study exclusive_place themed_residences, collection_themed, paramount_wealth, beauty_unrivalled,
Selling Apartments Type 2	mixed_jumeirah, hotel_barsha, khalifa_plot, dubai_plot, kempinski_plam, plam_royal, nahda_hotel, barsha_sale, interchange_dubai, sale_zayed, armani_furnishe,	penthouse_car,	millions_full, buyer_client, plans_affection, road_units,
Selling Villas Type 3	referred_beverly, paradise_emirates, hills_prestigious, clubs_emirates, hills_interchange, fairways_montgomerie, paradise_development, hills_bedrooms,emirates_emirates, undisputed_beverly, address_emirates, sector_emirates, montgomerie_hole, views_beachfront,	pool_basement, billiard, house_spa, guard_room, master_architect, entrance_halls club_communities, communities_recreation, stairwell_servicing, servicing_floors,	general_market, areas_development , rich_famous dream_bespoke project_sold class_montgomerie wealth_beauty development_place
Renting	residences_diamond novella_village type_novella diamond_atlantis diamond_palm dubai_embody embody_arabian shores_palm residences_shores arabian_european village_plot meadows_favored layouts_meadows barsha_hotel Iranian Hospital	luxurious_beach manager_room service_valet beach_car parks_avalibale rest_floor floor_affordable economical_bedroom brand_hotel beach_creates resorts_residences sitting_kitchen	themed_residences sale_binayah beach_life signature_villas collection_themed villas_collection
Selling	mixed_jumeirah ,dubai_plot , nahda_hotel, sale_zayed, hotel_sale, hotel_newly, level_plot, building_nahda, nahda_dubai, golfer_paradise,	brand_hotel, height_ground, penthouse_car, full_building, room_meeting, fully_operating, roof_upper, tower_entirely, upper_roofs, bayproject_basements, bedroom_unitsbrand,	millions_full hearts_content splash_hearts use_building course_created european_order client_details plans_affection

Table 5.6 Some top text features that lead to the increase of the predicted price

Also, table 5.6 has proven that the most important factor that affects the price of villas and apartments that are either offered for renting or for selling in Dubai is the location. When more information about the location of a property is given in the descriptive features such as the surrounding area, streets, facilities, malls, hotels, etc., then the accuracy of predicting the price is enhanced.

On the other hand, table 5.7 has some of the top regular features and text features that lead to the decrease of the predicted price in the linear regression experiments.

Dataset	Words refer to locations	Other Words
Renting Apartments Type 0	burj_khalifa, Loc = 78 (Al Quoz), Loc = 255 (Nad Al Sheba), Loc = 231(Lotus Hotel), Loc = 293 (Shaikh Hamdan Colony).	host_recreational, unfurnished_rented, bus, gardens
Renting Villas Type 1	montgomerie_hole, class_montgomerie, hills_world, fairways_montgomerie, montgomerie_makes, fronds_palm, hills, shores,	unrivalled_master, truly_luxurious, wealth_beauty
Selling Apartments Type 2	hotel_sheikh, views_kempinski, road_plot	ground_floors, plot_size, facilities_restaurant, additional_facilities, dubai_skyscrapers
Selling Villas Type 3	hillls_golfer, hillls, emirates_hillls, vicinity_dubai, cove_townhomes, montgomerie_makes, Beverly, marinas, colin_montgomerie, montgomerie_world, live_emirates, palm_offer, tree_palm, hills_unquestionably, school_springs, jumeirah_exciting,	prestigious_gated,fairways_world, crescent_shaped, desmond,muirhead (names of real estate agents).
Renting	european_mediterranean, Loc = 44 (Al Jafliya), Loc = 269 (Oud Al Muteena), Loc = 310(The Villas), Loc = 88(Al Rashidiya), Loc = 34(Al Hamriya), Loc = 294(Shangri-La Hotel), Loc = 317(Umm Al Sheif), Loc = 68(Al Mizhar), Loc = 231(Lotus Hotel), Loc = 305(The Gardens), palm_jumeirah	luxurious_resorts building balcony amenities_daylight floor_rest
Selling	Beverly, Loc = 43(Al Jafiliya), Loc = 88(Al Rashidiya), Loc = 306(The Green Community), Loc = 245 (Mirador), marina_residential, Loc = 188(Falcon City)	time_european, merit, order_merit, properties_look, architect_desmond, desmond, desmond_muirhead, green_fairways, hills, apartment, bldg_sale, villa

Table 5.7 Some regular features and text features that lead to decrease the predicted price

Chapter 6

Conclusion and Future Work

The conclusion of this thesis is discussed in this chapter along with highlighting its main contributions. In addition, the answers of the research questions and the suggested future work are included as well.

6.1 Conclusion

The use of data mining teachniques for discovering significant hidden information in structured stored data has proven the effectivness of these techniques in several domains. Also, text mining allows getting high quality information from text. In this thesis, an integration between the use of data mining technique, linear regression, and text mining task is proposed for the sake of enhancing the performance of the data mining prediction task. The improvment is achieved by using text mining to convert text features into structured numerical format so that linear regression prediction task takes into account the hidden information in text data in addition to the original stored structured features.

The proposed system architecture is mainly based on using linear regression data mining technique in predicting the price of a real estate property using a dataset that were collected and extracted from online real estate classifieds. The dataset is divided into six subsets as shown in table 6.1 based on the type of the classifieds: renting apartments, renting villas, selling apartments, selling villas, renting, and selling.

Subset Name	Type	Number of Records
Type 0	Renting apartment	24,424
Type 1	Renting villa	8,723
Type 2	Selling apartment	22,672
Type 3	Selling villa	10,579
Renting	Renting apartment/villa	33,137
Selling	Selling apartment/villa	33,251

Table 6.1 The six subsets used in experiments

Within each experiment, a data subset is divided into two portions based on the data type of the features. The structured numerical features: number of bedrooms, location of property and price, were fed directly into the linear regression model. In the same time, the text descriptive features, title of the classified and the property's description, have gone under the process of text mining in order to convert the text content into suitable numerical format for linear regression task. As a result of text mining, thousands of features are generated and each feature reflects the importance of a word using its term frequency-inverse document frequency TF-IDF value.

The results of carrying out the linear regression experiments on the original structured data are considered the baseline of the research. Hence, the target was to compare the baseline results with the results of linear regression prediction after text mining the text features and adding them to the experiments. Based on the results and the carried out analysis, the answers of research questions are as follows:

- *How will traditional data mining techniques, relying only on structured data, perform when used on Dubai real estate classifieds?*

  The baseline results of using linear regression modeling as a data mining technique in predicting the price of real estate properties are encouraging to continue investigating and researching in the field. The accuracy of price predictions was below average and the root mean squared error was high for most data subsets. However, the linear regression analysis allows highlighting the key player factor in controlling the prices of real estate properties in Dubai. The factor is the location of the property. It is found that some places in Dubai are

expensive in both; renting and selling residential real estate properties such as Emirates Hills. In addition, Aljafiliya and Alrashidiya are considered the cheapest places for renting or selling real estate properties regardless of the type of the property, either apartment or villa.

As a conclusion, the use of traditional data mining techniques in prediction is encouraged and requires working on improving the results. The answer for the second research question is the trial for improvement.

- *Will the use of text mining improve the accuracy of price predictions?*

  Definitely the results show that the use of text mining to prepare the text features to be added to the task of linear regression prediction has highly improved the accuracy of prediction. Table 6.2 shows how greatly the root mean squared error, RMSE, values decreased after integrating text-mining task in linear regression model. For instance, the RMSE's for datasets Type1 and Type2 have decreased by 56%.

  Also, the linear correlation between regular features and the price feature have increased for all data subsets. For example, for the dataset Type2, the correlation coefficient has increase by 206% from 0.296 to 0.906.

Dataset	Root Mean Squared Error – RMSE			Correlation Coefficient – CC		
	Before Text Mining	After Text Mining	Percent of Decrease	Before Text Mining	After Text Mining	Percent of Increase
Type 0- Renting Apartments	110,263.85	61,715.12	44.03%	0.479	0.871	81.84%
Type 1- Renting Villas	204,650.55	89,710.59	56.16%	0.543	0.930	71.27%
Type 2- Selling Apartments	8,676,287.33	3,853,490.15	55.59%	0.296	0.906	206.08%
Type 3- Selling Villas	4,756,663.73	3,750,734.50	21.15%	0.802	0.882	9.97%
Renting	144,483.34	96,334.73	33.32%	0.586	0.841	43.52%
Selling	7,738,338.21	4,440,108.38	42.62%	0.501	0.868	73.25%

Table 6.2 Results of linear regression experiments before and after using text-mining

To the best of our knowledge, this thesis has achieved the following contributions:

1. It is the first scientific analysis of Dubai's real estate classifieds. The analysis reveals important facts about significant factors that play an important role in determining the renting or selling price of a real estate property in Dubai. One of the important factors is the property's location.

   It is found that giving detailed information about where does a property located and what are the surrounding places, hotels, roads, malls, etc. can highly improve the price estimation value. Moreover, another important factor is revealed by text mining which is giving more descriptive information about the available facilities and features of an apartment or a villa has direct effect in determining the property's price. Examples of such features: swimming pool, balcony, maid's room, furnished, spa, etc.

2. It is the first trail in using text mining to improve the accuracy of real estate property price prediction. All previous researches have been carried out on structured numerical data that describes the basic features of a property to run the prediction task. Therefore, this thesis has achieved the goal of making use of descriptive text features in enhancing the prediction task.

3. The results of the experiments were verified based on a real world dataset. The used dataset is extracted from online real estate classifieds. Therefore, the information and the characteristics of the properties are considered up to date and reflect the current trends in the real estate market in Dubai.

6.2 Future Work

As a future work, it is planned to examine the performance of using another data mining technique to replace the linear regression model in price prediction task. For example, neural network technique may replace the linear regression model.

Also, the proposed feature selection scheme in the text mining process depends on measuring the weights of the generated text features in correlation with the targeted feature, price. Then, the features with the highest weights (highly correlated with the price) are given the priority to be involved in the experiments with comparison with other features. Therefore, in order to increase the accuracy of the proposed model with less number of text features as possible and as fast as possible, it is planned to find out another technique for feature selection that can help in achieving the target.

Finally, while analyzing the results of text mining and by analyzing the content of the "Description" feature, it was found that the performance of the system can be further enhanced if more cleaning preprocessing steps to be done. Many records have incorrect information. For example, some records have information about monthly renting price in the description feature, while the value of the price feature reflects a selling price. Also, a renting dataset might have records for properties that are offered for selling. In addition, the apartments' classifieds are mixed with studio classifieds. Therefore, extra focused preprocessing could highly enhance the performance.

References

Acciani, C., Fucilli, V., and Sardaro, R. (2011). Data Mining in Real Estate Appraisal: A Model Tree and Multivariate Adaptive Regression Spline Approach. *Aestimum, 2011*, Vol. 58, pp.27-45

Asilkan, Ö., Ismaili, A., and Nuredini, K. (2011). An Exemplary Survey Implementation on Text Mining with Rapid Miner. *1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)*, pp. 221-234

Azevedo, A., and Santos, M. (2008). KDD, SEMMA AND CRISP-DM: A Parallel Overview in Ajith Abraham, ed., *IADIS European Conference on Data Mining, IADIS*, pp. 182-185

Berson, A., Smith, S., and Thearling, K. (2011). An Overview of Data Mining Techniques [online]. [Accessed 28 November 2011]. Available at: http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Buana P., Jannet S. and Putra I. (2012). Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News. *International Journal of Computer Applications*, Volume 50 – No.11, pp. 37-42

Cacho C. (2010). A Comparison of Data Mining Methods for Mass Real Estate Appraisal. University Library of Munich, Germany. [online]. [Accessed 1 July 2013]. Available at: http://EconPapers.repec.org/RePEc:pra:mprapa:27378

Chibelushi, C., and Thelwall, M. (2009). Text Mining for Meeting Transcript Analysis to Extract Key Decision Elements. *Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2009,* Vol I, pp.710-715

Dukova, R., Gacovski, Z., Kolic, J. and Markovski, M. (2012). Data Mining Application for Real Estate Valuation in the city of Skopje. *CT Innovations 2012 Web Proceedings.* [online]. [Accessed 1 June 2013]. Available at: http://ictinnovations.org/2012/htmls/papers/WebProceedings2012.pdf

Guan J., Levitan A., and Zurada, J. (2008). An Adaptive Neuro-Fuzzy Inference System-Based Approach to Real Estate Property Assessment. *Journal of Real Estate Research (JRER)*, Vol. 30, No. 4, pp. 395-421

Han, J and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers

Hotho, A., Nürnberger, A. and Paaß, G. (2005). A Brief Survey of Text Mining. *Journal for Computational Linguistics and Language Technology* 20, pp. 19-62

Jaen, R. (2002). Data Mining: An Empirical Application in Real Estate Valuation. *FLAIRS-02 Conference Proceedings*, pp. 314-317

Janicic, P., Keselj, V. and Tomovic, A. (2006). n-gram-based Classification and Unsupervised Hierarchical Clustering of Genome Sequences. *Computer Methods and Programs in Biomedicine*, vol. 81, pp. 137-153

Jungermann , F. (2011). Documentation of the Information Extraction Plugin for RapidMiner [online]. [Accessed 30 March 2014]. Available at: http://www-ai.cs.uni-dortmund.de/PublicPublicationFiles/jungermann_2011c.pdf

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice*, pp. 137–153

Massey University.(2013). Real Estate & Mortgage Insights [online]. [Accessed 30 July 2013]. Available at: http://www.realestateabc.com/insights/mls.htm

RapidMiner (2010). RapidMiner 5.0 User Manual [online]. [Accessed 15 July 2013]. Available at: http://rapidminer.com/documentation/

Rexer, K. (2013). 2013 Data Miner Survey. [online]. [Accessed 15 December 2013]. Available at: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2013.html

Roshni, S., Sagayam R., and Srinivasan, S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. *International Journal of Computational Engineering Research*, Vol. 2 Issue. 5 pp. 1443-1446

Saravanan, D., and Chonkanathan, K. (2010). Text Data Mining: Clustering Approach. *International Journal of Power Control Signal and Computation (IJPCSC)*, Vol. 1 No. 4, pp. 13-16

Wedyawati, W. and Lu, M. (2004). Mining Real Estate Listings Using ORACLE Data Warehousing and Predictive Regression. *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration-IRI*, pp. 296-301

Witten, I., and Frank, E. (2005). Data Mining Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers.