Title
Text Mining Attributes in Real Estate: Using Textual Data and Common Features to Predict Housing Prices in the United States of America

Author
Morales, Elizabeth

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Text Mining Attributes in Real Estate: Using Textual Data and Common Features to Predict Housing Prices in the United States of America

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Elizabeth Morales

2022

ABSTRACT OF THE THESIS

Text Mining Attributes in Real Estate: Using Textual Data and Common Features to Predict Housing Prices in the United States of America

by

Elizabeth Morales
Master of Applied Statistics
University of California, Los Angeles, 2022
Professor Frederic R. Paik Schoenberg, Chair

Text mining is a powerful tool that can be used to uncover data that businesses can turn into actionable insights. The rise in technology and real estates online presence has made housing listing data readily accessible. The purpose of this paper is to turn textual data gathered from real estate listings into numerical attributes and fit regression models, alongside other common housing attributes, to predict housing prices. Text mining, machine learning, and statistical techniques were used to transform the dataset, build regression models, and select the best performing model.

The thesis of Elizabeth Morales is approved.

Akram M. Almohalwas

Mark Stephen Handcock

Hongquan Xu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2022

To my parents . . .

thank you for your endless support, love, and encouragement.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Text Mining 1

1.2 Real Estate 2

1.3 Previous Studies 2

1.4 Research Questions 4

2 Data **5**

2.1 Dataset Info 5

2.2 Attribute Info 6

2.3 Exploratory Analysis 7

2.3.1 Real Estate Listing Descriptions 7

2.3.2 Word Frequency 9

2.3.3 Word Sentiments 10

2.3.4 Distribution of Variables 10

3 Methodology **13**

3.1 Data Processing 13

3.1.1 Creating a Corpus 13

3.1.2 Corpus Transformations 13

3.2 Tokenization 15

3.3 N-grams & Bi-grams 16

3.4 TF-IDF 16

3.5 Sentiment Analysis . . . 17
3.6 Box Cox . . . 18
3.7 Regression . . . 19
3.7.1 Multiple Linear Regression . . . 19
3.7.2 Regression Trees . . . 20
3.8 Model Selection . . . 20
3.8.1 RMSE . . . 21
3.8.2 AIC . . . 21
3.8.3 K-Fold Cross Validation . . . 22

4 Analysis of Models . . . 24
4.1 Multiple Liner Regression . . . 24
4.1.1 Full Model . . . 24
4.1.2 AIC Model . . . 25
4.1.3 Cross Validation Model . . . 26
4.1.4 Reduced Model . . . 27
4.2 Random Forest . . . 29

5 Conclusion . . . 31

References . . . 33

LIST OF FIGURES

2.1 Example of 10 Attributes For One Instance In The Dataset 7
2.2 Most Frequent Words 9
2.3 Histogram of Price 11
2.4 Histogram of Price After Transformation 12

3.1 Before Transforming Corpus 14
3.2 After Transforming Corpus 15
3.3 Tokenization Example 15
3.4 Bi-gram Example 16
3.5 TF-IDF Equation 17
3.6 Box Cox Equation 18
3.7 Equation For A Multiple Regression Line 19
3.8 RMSE Equation 21
3.9 AIC Equation 22
3.10 K-Fold Example 23

4.1 Variable Importance 30

LIST OF TABLES

2.1 Sentiment Of Words 10

4.1 Full Model Output 25

4.2 AIC Model Output 26

4.3 CV Model Output 27

4.4 Reduced Model Output 28

4.5 Random Forest Output 29

CHAPTER 1

Introduction

Data is defined as a collection of facts, statistics, and items of information that can be used for reference or analysis. Data is present in everyday life from businesses, society, sciences, and other aspects of our daily life [1]. The exponential growth of data is a result of technological evolution in our society and our ability to collect and store all of this information. To turn data into valuable and organized knowledge, we need tools [1]. This has led to a field known as data mining.

Data mining is the act of discovering interesting, unexpected, or valuable structures in large datasets using machine learning and statistical techniques into useful information [2]. Data mining can be applied to various data forms such as database data, warehouse data, transactional data, data streams, ordered data, graph or networked data, spatial data, web data, multimedia data, and text data [1]. Because of these numerous data types, challenges may arise with how to mine data from these various structures and forms. One specific type of data I will focus on is text data.

1.1 Text Mining

Text data surrounds us in the form of digital libraries, repositories, and other textual information such as literature, documents, blogs, social media networks, reviews, classifieds, and e-mails. Traditional data mining tools have become inadequate in retrieving knowledgeable information from vast amounts of text data as the data stored in most text databases

are semi-structured data in that they are neither completely unstructured nor completely structured [3]. Thus, text mining has become an extension of data mining.

Text mining is the process of transforming unstructured text into structured data for easy analysis at the intersection of information retrieval, data mining, machine learning, statistics, and computational linguistics [4]. Data text mining has become a powerful tool for businesses as it allows for data-driven business decisions. One business that can benefit from text mining is real estate.

1.2 Real Estate

Real estate is an integral part of our society and economy. It is known as the act of producing, selling, and buying property. As with any product that is being sold, it is important to make sure that the price of the product is reflected by its characteristics. This is especially true in real estate where the price of homes can range from hundreds of thousands to millions of dollars. Fortunately, there is already a method that estimates the value of the characteristics of a commodity that indirectly affects its market price [5]. This is also known as the hedonic price method also known as hedonic regression.

1.3 Previous Studies

The hedonic price method has been applied to real estate as it allows for an assessment of the property value given available variables. In a paper by Herath, S. K. & Maier, G.[5], they state that regression analysis estimations are common in HPM models in real estate [5]. The hedonic regression equation can take the form of either linear, semi-log, or log-log form. When applying the hedonic price regression equation to real estate, the characteristics of the house are regressed on the price of the house. Some variables that have commonly appeared in real estate hedonic price analysis include the number of bedrooms and bathrooms, floor

area, house type, square foot, location, age, and other possible structural features [5]. Herath, S. K. & Maier, G.[5] reviewed 471 of the most cited papers related to housing and real estate that use the HPM and discovered that of the 471, only 16 papers dealt with the implicit value of structural characteristics of housing on price or value of the real estate. While previous studies have focused on regression models that contain numeric and nominal attributes, few have taken textual attributes into account.

A study done by C. Amrit et al [6] shows a method of predicting child abuse based on structured and unstructured data. This method categorizes child abuse using a model-based solely on structured data, a model based on unstructured data, and a model based on both with the purpose to determine which model might perform better in predicting child abuse. Structured data consisted of relevant features for child abuse while unstructured data was found in the form of free-text notes. All models were able to predict child abuse and were compared using a metric of AUC, accuracy, and recall with AUC being the ultimate evaluation metric. C. Amrit et al [6], found that the best performance was attained when combining the two classifiers for unstructured and structured data into an ensemble method with an AUC of 0.914, an accuracy of 0.822, and a recall of 0.870. C. Amrit et al 's [6] research shows that a combination of text mining and relative features improves models in the predictive analysis.

A study by T.P. Williams and J. Gong [7], focused on using a combined model of text data and numerical data to predict levels of cost overrun for a construction project. Text data was turned into numeric vectors by undergoing tokenization, stopping, stemming, normalization, and vector generation. Once the text was transformed into numeric variables, it was combined with other numerical data. The combined data was used to construct classification models that would predict three levels of overrun for a construction project: 1 (high overrun project), 2 (near low bid project), and 3(underrun). Of the multiple models constructed, the stacking model with combined text and numeric data had an accuracy of 43.75%, class precision of 44.32%, and class recall of 43.22% while the stacking model with

only numeric data had an accuracy of 44.15% class precision of 39.29% and class recall of 22.18% [7]. Using accuracy, recall, and precision as metrics of model performance, it was found that a stacking model containing both text and numeric data outperformed a stacking model that only contained numeric data when only looking at recall and precision. This study shows that the inclusion of text in cost overrun prediction for construction projects can improve model performance [7].

Text in real estate can appear in the form of property descriptions. Houses in general share the same attributes but that does not mean the they are not unique. A house may have a specific physical or character attribute that can be expressed through the description of a property. A description may persuade a reader to pursue one home over another. Because of this, it is worthwhile to take text attributes into account when predicting the price of a home.

1.4 Research Questions

In this paper will aim to answer the following questions:

- 1. Can real estate descriptions predict selling price?
- 2. Do certain models perform better than others?
- 3. Does the sentiment of the descriptions affect the price?

CHAPTER 2

Data

2.1 Dataset Info

Real estate properties were originally advertised in newspapers under the classified section but the rapid growth of the digital space birthed tech related real estate marketplaces. Some of the most popular online real estate websites are Zillow, Realtor.com, and Trulia. The data set was acquired from December 2021 to May 2022 by web scraping Zillow's website for homes sold in the largest cities of the United States. These cities include:

- New York, New York
- Los Angeles, California
- Chicago, Illinois
- Houston, Texas
- Phoenix, Arizona
- Philadelphia, Pennsylvania
- San Antonio, Texas
- San Diego, California
- Dallas, Texas
- San Jose, California

The data set consisted of 10 comma-separated value (CSV) files, each containing data on a number of properties listed for sale and that had been sold.

2.2 Attribute Info

The raw data contained 26 different attributes with every instance in the file representing a house. It varies per house how many of the 26 attributes contain missing values or are left blank in the data. For the purpose of this study, we decided to focus on the following numeric predictors and one text predictor. The following attributes are present in the data:

- city: the city of a house for one instance in the data.
- population: the population of a given city of a house for one instance in the data.
- state: state of a house for one instance in the data.
- address: address of a house for one instance in the data.
- zipcode: zipcode of a house for one instance in the data.
- bathrooms: number of bathrooms in a house for one instance in the data.
- bedrooms: number of bedrooms in a house for one instance in the data.
- description: verbal description of a house for one instance in the data.
- sqrft: square footage of a house for one instance in the data.
- lotSize: size of the lot a house is on for one instance in the data.
- price: price of a house for one instance in the data.
- propertyTaxRate: the property tax rate for a house in one instane in the data
- yearBuilt: the year a house was built for one instance in the data.

- age: how old a house is for one instance in the data.

city	population	state	address	zipcode	bathrooms	bedrooms	description	sqrft
Los Angeles	3967000	CA	1710 San Remo Dr	90272	6	5	Nestled in a secluded canyon in the Upper Riviera, this enchanting home offers a secluded & tranquil retreat w/stunning canyon & mountain views. Enjoy sophisticated country living in this prime Pacific Palisades neighborhood, yet just minutes away from the bustle of the city. A gated private driveway leads to a motorcourt & charming front porch. Enter into the 2-story foyer & spacious living rm w/fp & walls of windows & doors opening to the wrap-around patio that encompasses the main level. Dining rm & fabulous chefs kitchen w/prof. appliances, island & breakfast rm w/fp. The upper level is comprised of a tremendous master suite w/office/sitting rm, fp, balcony overlooking beautiful views to the Getty, walk-in closet, & spa-like bath w/fp & sauna. The lower level offers 4 addl bdrms, media lounge w/kitchenette/bar, laundry rm & storage space. The lushly landscaped exterior space offers great space for entertaining w/multiple decks/patios including a dining & BBQ area, pool & spa.	5345

Figure 2.1: Example of 10 Attributes For One Instance In The Dataset

By checking the data structure and definition of each variable, the variables were classified into two groups: numerical and categorical.

2.3 Exploratory Analysis

2.3.1 Real Estate Listing Descriptions

When looking at the description of houses, some are as few as one sentence and as long as a paragraph. Looking at the vernacular of the description, you can see that each description is unique to each house. Some include quantitative descriptions that may not be present in data while others others contain only qualitative data that cannot be quantified. Below are examples of a few descriptions found in the dataset.

"Multiple offers received - Best & Final due Feb 15th at 5pm! **New Construction** Gorgeous 4 bedroom, 2 bath brick home with stone accents. Gated driveway with room for extra parking. Walk in the custom front door and be blown away by the open concept living, kitchen and dining areas. Modern cabinetry, tray ceilings, decorative lighting, granite

counters and stone backsplash in the common areas. Luxury vinyl plank flooring and neutral paint through out. Owners suite features a walk in shower with decorative tile work. Oversized back yard and full sod for the front yard."

"27390 Smithson Valley Rd, San Antonio, TX 78261 is a single family home that contains 3,053 sq ft and was built in 2013. It contains 3 bedrooms and 3 bathrooms. The Zestimate for this house is $1,179,700. The Rent Zestimate for this home is $4,110/mo."

"Gorgeous Lakewood Heights 4 BR, 3.5 Bath Blanchard Home with amazing open family floor plan. Home features a gourmet kitchen with granite countertops, custom cabinetry, & stainless steel Kitchen-Aid appliances. Separate downstairs study, second floor gameroom with wetbar two story vaulted family room with wetbar and fireplace. Three beds up and one onsuite down, the upstairs master suite has his and her closets and a luxury bath with jetted tub, double shower, and dual vanities! Outdoor living area was recently expanded with an amazing pergola, sitting area a fireplace and a covered porch to to hang a television. Amazing home!"

"Coming soon, a modern take on classic farmhouse design in the Chappell/Amundsen school drawing area! This brand new home on quiet Berwyn Avenue is designed for comfort, shelter and luxury in today's world. Built on a nearly 38 foot wide lot, and it features almost 50 % more interior space than home built on a standard 25' wide lot! Crisp white exterior with stunning window contrasts, two great rooms and 6 bedroom/office spaces give ample space for relaxing, work, study, even entertaining friends while social distancing. Three fireplaces. Wide French doors open from the kitchen/great room onto a spacious rear deck and wide back yard with a roof deck-ready 3-car garage. All 3 levels have 10+ foot high ceilings. A grand staircase with a lofted, cathedral ceiling opens up to four second-floor bedrooms, two with suite baths and two sharing a spacious Jack-and-Jill. The owner's retreat will have a luxurious spa bath and a generously-sized dressing room, as well as a three-panel sliding door with a Juliet balcony for light and fresh air. Close to Andersonville and Lincoln Square for nightlife, shopping and restaurants, as well as train service. Minutes

away from the Peterson Target. Still under construction–slated for March/April delivery! All pictures are renderings; finished product may vary."

2.3.2 Word Frequency

A few of the most frequent words found in the description of homes in the dataset can be found in Figure 2.2. The five most frequently used words are home, bedroom, kitchen, floor, and bathroom. This makes sense as all homes contain each of these features and describing them in a way that's unique to a specific house can make one home stand out from other homes.

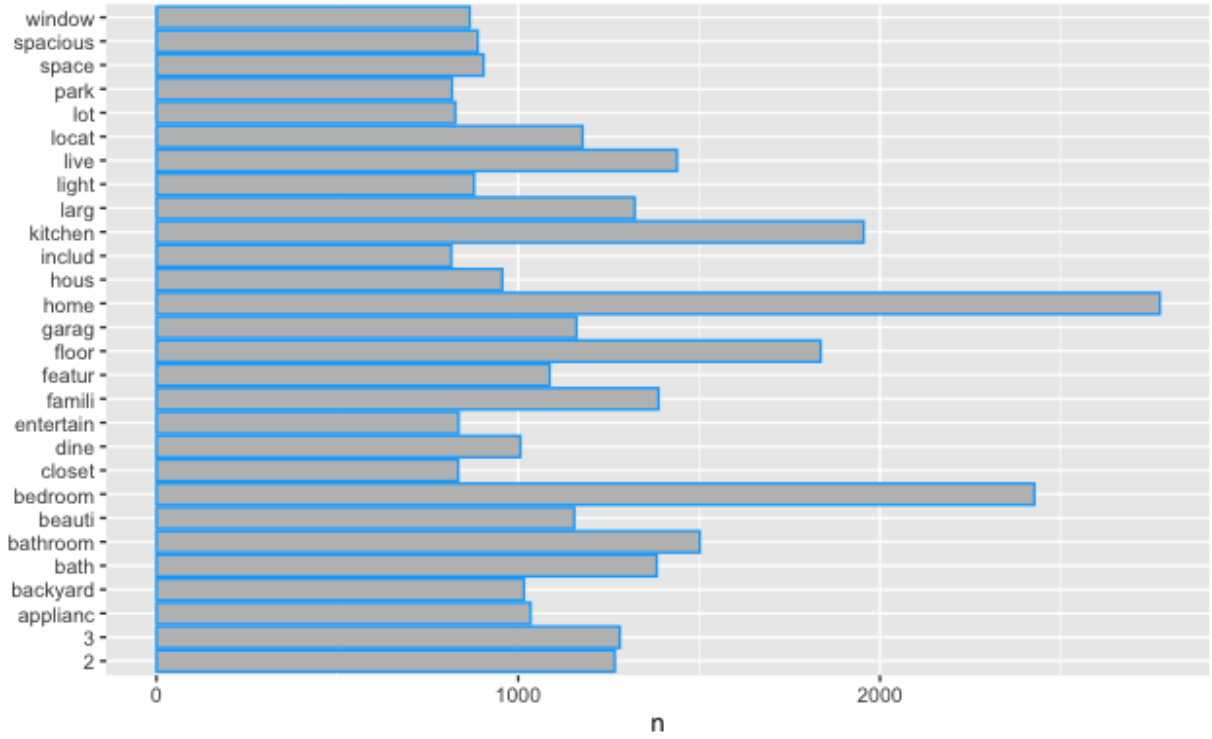

Figure 2.2: Most Frequent Words

2.3.3 Word Sentiments

Sentiment analysis on housing descriptions is a great way to look for words to include, avoid, or reframe in a listing. Looking at Table 2.1 we can see that the word with the most positive association is "superb" while the word with the most negative association is "die". Taking this into consideration, if there are words that are considered to have a negative connotation but are needed to describe less desirable features in a house, painting them in a positive light can help with that. Its always good to take sentiment into consideration as it can help with making sure descriptions are telling the story of the house and grabbing the readers attention.

Table 2.1: Sentiment Of Words

Word	**Value**
superb	5
win	4
fun	4
brilliant	4
perfect	3
drop	-1
retreat	-1
miss	-2
disappoint	-2
die	-3

2.3.4 Distribution of Variables

The data set was cleaned by removing the few missing values. When looking at the distribution of both independent and dependent variables, a few distributions looked skewed but price was heavily skewed. Looking at Figure 2.3 you can see that there is a heavy right skew in the distribution of Price. After performing a Box-Cox transformation, the histogram of price is looking normally distributed. This transformation can be seen in Figure 2.4.

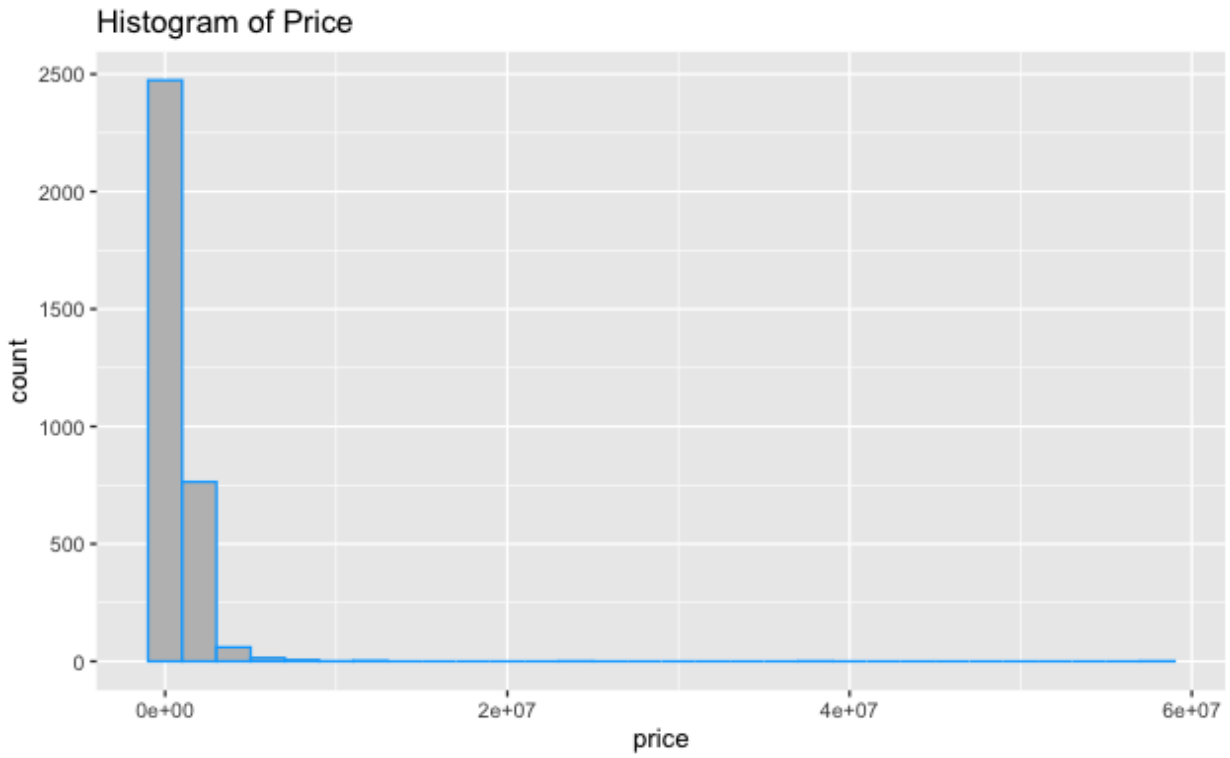

Figure 2.3: Histogram of Price

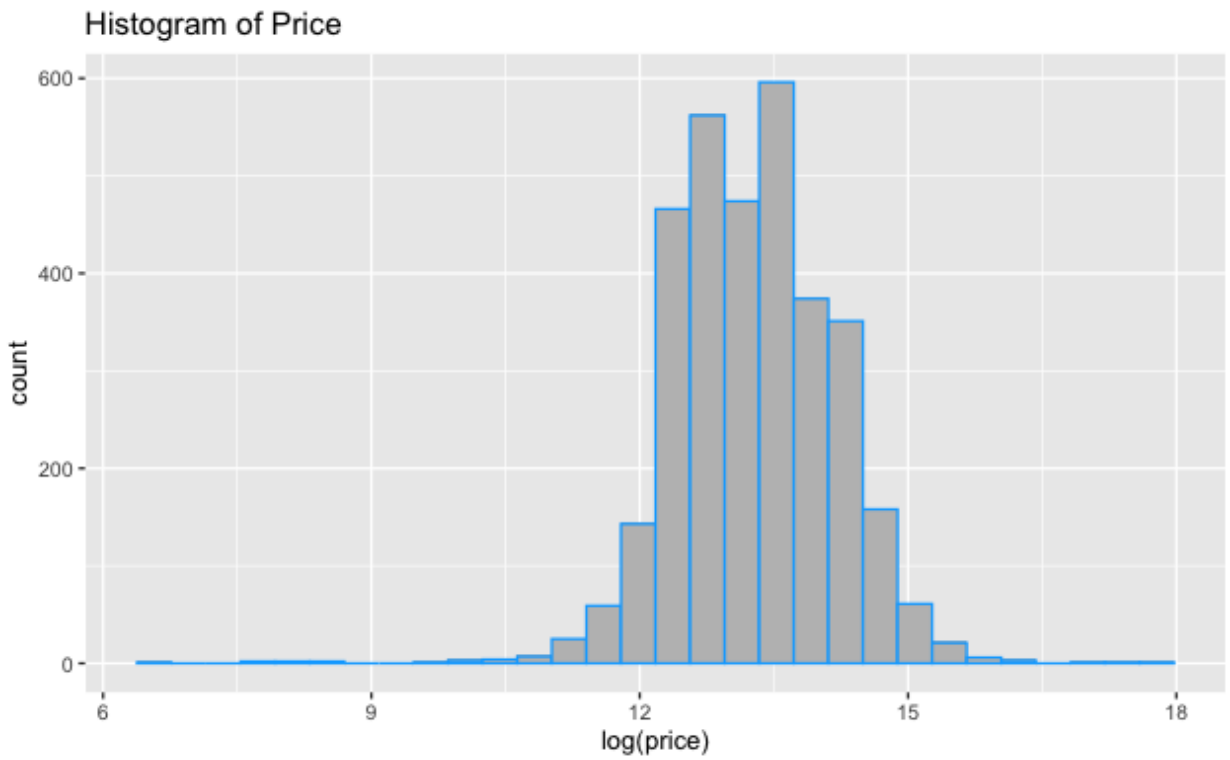

Figure 2.4: Histogram of Price After Transformation

CHAPTER 3

Methodology

3.1 Data Processing

As previously mentioned, numeric data and text data can be used together in a model to help improve pricing predictions. Each row in the dataset contains a description specific to that home and in order to process the text for analysis, it has to be separated from all the other numeric variables. Before any analysis of text can be done, we need to perform some pre-processing of the text. All of the data was processed in R using the following data mining and natural language processing packages: readr, tm, tidyverse, tidytext, dplyr, and ggplot2.

3.1.1 Creating a Corpus

The description of homes were pulled from the data and turned into a corpus. A corpus is a collection of text documents, and in our case it's a collection of the descriptions for all homes in the dataset [8]. Each observation in the dataset contains a description each description from all observations where pulled from the dataset to form a corpus.

3.1.2 Corpus Transformations

With all of the text in one corpus, we can continue to clean textual data by performing several transformations on the corpus. We begin by converting all text characters to lowercase as it removes unwanted noise in our data. We then remove punctuation as its common to do

so in preprocessing of text and can provide grammatical context which supports understanding. We continue with removing stop words as they may take away from important words found in the descriptions. Stop words are common words found in a given language. Words like for, very, and, of, are, etc, are common stop words. In addition to removing English words, we may choose to remove our own words. Through exploratory analysis, it was found that the word "zestimate" appeared very frequently. Zestimate refers to the estimate given by Zillow on price prediction for a home. Since this is not our focus, we proceeded to remove it. We finished cleaning the corpus by removing white spaces and performing stemming on the corpus which allowed us to return verbs and words to their infinitive forms. Many times, large texts can be reduced by a few words and possibly down to a few words.

"Nestled in a secluded canyon in the Upper Riviera, this enchanting home offers a secluded & tranquil retreat w/stunning canyon & mountain views. Enjoy sophisticated country living in this prime Pacific Palisades neighborhood, yet just minutes away from the bustle of the city. A gated private driveway leads to a motorcourt & charming front porch. Enter into the 2-story foyer & spacious living rm w/fp & walls of windows & doors opening to the wrap-around patio that encompasses the main level. Dining rm & fabulous chefs kitchen w/prof. appliances, island & breakfast rm w/fp. The upper level is comprised of a tremendous master suite w/office/sitting rm, fp, balcony overlooking beautiful views to the Getty, walk-in closet, & spa-like bath w/fp & sauna. The lower level offers 4 addl bdrms, media lounge w/kitchenette/bar, laundry rm & storage space. The lushly landscaped exterior space offers great space for entertaining w/multiple decks/patios including a dining & BBQ area, pool & spa."

Figure 3.1: Before Transforming Corpus

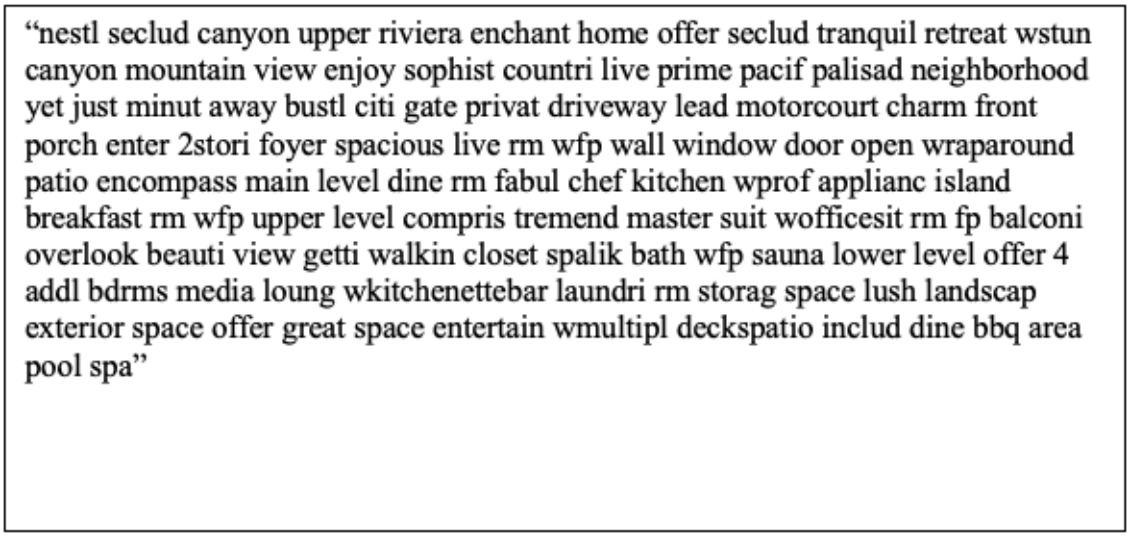

Figure 3.2: After Transforming Corpus

3.2 Tokenization

Proceeding with pre-processing of the text, we began by tokenizing our corpus. Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph. This allows us to gather textual information from individual words.

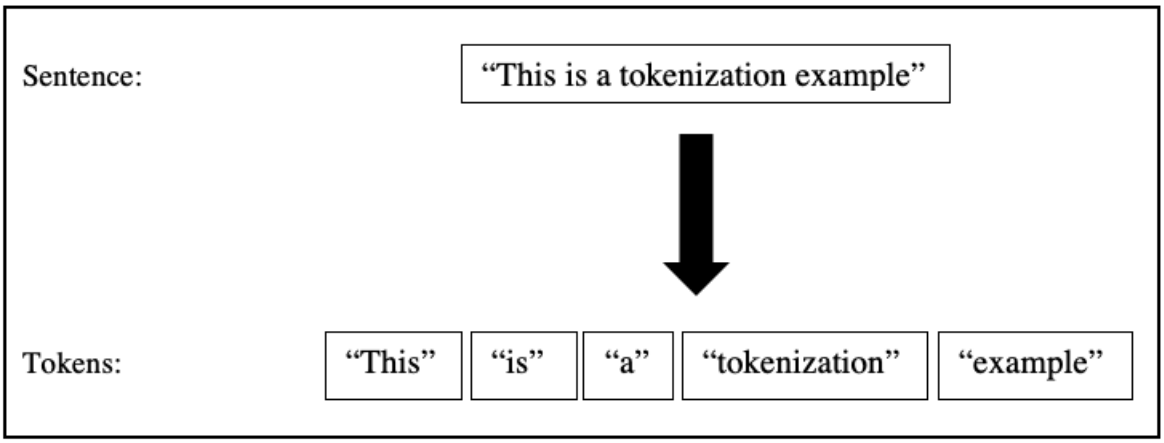

Figure 3.3: Tokenization Example

3.3 N-grams & Bi-grams

Many interesting text analyses are based on the relationships between words, from examining which words tend to follow others immediately to words that tend to co-occur within the same documents. N-grams are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward [9]. For the purpose of our textual analysis, we will look at bi-grams. A bi-gram is a two-word sequence of words and makes a prediction for a word based on the one before. The main purpose is to calculate and visualize relationships between words in our text dataset. Using the bi-grams, we were able to create a vocabulary for analysis.

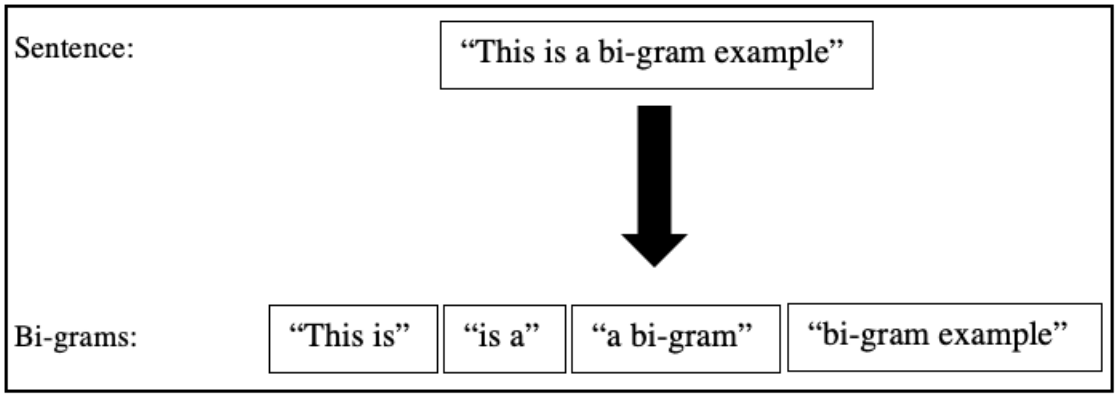

Figure 3.4: Bi-gram Example

3.4 TF-IDF

Quantifying what a document is about is one of the main goals of text mining. One measure of how important a word may be is its Term Frequency(TF), how frequently a word occurs in a document. If we are to look at a term's Inverse Document Frequency (IDF), it decreases the weight for commonly used words and increases the weight for words that are not used very much in a corpus. This can be combined with Term Frequency to calculate a

term's TF-IDF, the frequency of a term adjusted for how rarely it is used.

TF-IDF is intended to measure how important a word is to a document in a corpus, or in our case a corpus of housing descriptions from all cities. Once the TF-IDF's for all bi-grams are calculated, we can use the sum of TF-IDF scores for one given observation to quantify the description of a house for one instance in our data.

$$\text{TF-IDF score for a term } i \text{ in document } j = \text{TF}(i,j) * \text{IDF}(i)$$

where

IDF = Inverse Document Frequency

TF = Term Frequency

$$\text{TF}(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ words\ in\ document\ j}$$

$$\text{IDF}(i) = \log_2 \left(\frac{Total\ documents}{Documents\ with\ term\ i}\right)$$

and

t = Term

j = Document

Figure 3.5: TF-IDF Equation

3.5 Sentiment Analysis

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. The AFINN package in R allows us to get specific sentiment lexicons with the appropriate measures for each one [10].

The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. We used this to perform sentiment analysis on our textual data. Similar to the procedure for TF-IDF, once the sentiment scores for all bi-grams are calculated, we can use the sum of sentiment scores for one given observation to quantify the sentiment of a description for one instance in our data.

3.6 Box Cox

Taking a graphical look at our data, we saw that the histogram of price (Figure 2.3) was heavily skewed. One way of normalizing skewed data is to perform a transformation on the data. The transformation performed on the price of homes is called a Box Cox transformation. The Box Cox transformation on a response variable can be mathematically expressed in Figure 3.6.

$$t_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ log(y), & \text{if } \lambda = 0. \end{cases}$$

Figure 3.6: Box Cox Equation

When finding λ for our transformation, it was very close to zero and therefore rounded to the nearest whole number which was zero. Thus when looking at the Box Cox equation (Figure 3.6) we see that if λ=0 then our transformation would be to take the log of our response variable. The log of prices was taken to perform a transformation on the price of homes.

3.7 Regression

The objective of regression analysis is to predict a numerical outcome from one or more predictors. In regression we divide the variance in the outcome variables to two parts. The part due to regression or the part we can explain as a result of correlation between predictor and outcome.

3.7.1 Multiple Linear Regression

Multiple linear regression attempts to model the relationship between two or more predictor variables and an outcome variable by fitting a linear equation to observed data.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

Figure 3.7: Equation For A Multiple Regression Line

Figure 3.7 shows the linear equation for a multiple regression model. The equation is further defined below:

- Y = the predicted or expected value of the dependent variable
- X1...Xp = p distinct independent or predictor variables,
- b0 = the value of Y when all of the independent variables (X1...Xp) are equal to zero
- b1...bp = the estimated regression coefficients.

Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. In the multiple regression situation, b1, for example, is the change in Y relative to a one unit change in X1, holding all other independent variables constant. Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

3.7.2 Regression Trees

A regression tree is a type of decision tree used to predict or model a numeric response variable. When choosing where to create a partition, the tree will consider all possible partitions of the data. Decision tree algorithms are greedy, and will choose the split that minimizes the Residual Sum of Squares (RSS) of the data after making the split.

A Random forest is a way to improve decision trees. A Random forest randomizes the variables used in creating a tree. The individual trees are thus impaired in that they cannot use all of the available variables to create a split. The hope is that by having many sub-optimal, yet highly diverse trees, the resulting prediction made by the aggregate of all trees will be better.

Random Forest works like this:

1. create a bootstrap resample of the training data.
2. when fitting a tree, choose only a random subset of the predictor variables in the data set.
3. fit many trees - each tree will use a different random subset of predictor variables and different random subset of training data.
4. make a prediction by averaging the prediction of all the different trees (for a regression tree), or by majority voting for a classification tree

3.8 Model Selection

Across the following models, we used an 80/20 split among our training and testing data. This gives us the opportunity to test each model equally and compare some of the model statistics as well as RMSE values.

3.8.1 RMSE

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Figure 3.8: RMSE Equation

RMSE can be represented by the following:

- Σ is the summation of all values
- $\hat{y}_i$ is the predicted value
- y_i is observed or actual value
- $(y_i - \hat{y}_i)^2$ are the squared differences between predicted and observed values
- n is the total sample size

It is also a helpful criterion in determining the best performing model among different models that we may have trained on one particular dataset. To do so, we simply compare the RMSE values across all models and select the one with the lowest value on RMSE.

3.8.2 AIC

Akaike Information Criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. AIC will take each model and rank them from

best to worst with the best model being the one that neither under-fits nor over-fits. Figure 3.9 shows the mathematical expression of AIC.

$$\text{AIC} = -2(\text{log-likelihood}) + 2K$$

Figure 3.9: AIC Equation

AIC can be represented by the following:

- K is the number of model parameters (the number of variables in the model plus the intercept)
- Log-Likelihood is a measure of model fit

3.8.3 K-Fold Cross Validation

Cross Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. K-Fold Cross Validation is a form of CV where the data is divided into k random subsets. A total of k models are fit, and k validation statistics are obtained. Figure 3.10 shows an example of a K-Fold Cross Validation [11].

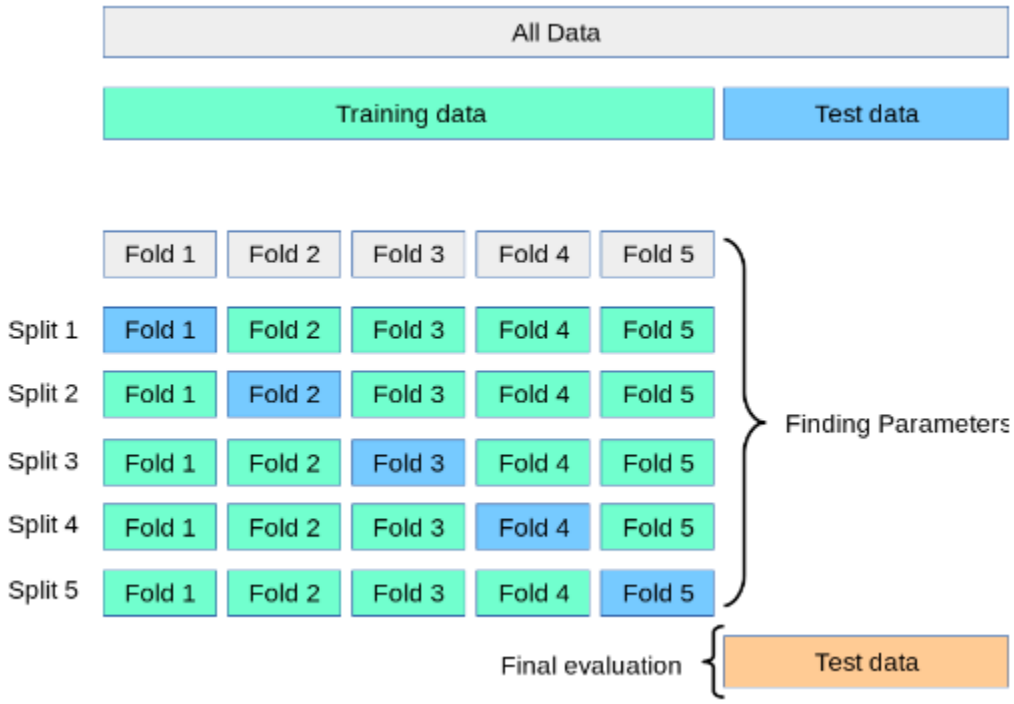

Figure 3.10: K-Fold Example

CHAPTER 4

Analysis of Models

In order to obtain the test validation of our models, the dataset was divided into two parts with an 80-20 split: 80% of data instances are the training data while the rest of 20% are test data. We already know that there are total 3,308 different instances in this case. Therefore, the training set contains 2,573 data points and the test set contains 735 data points. Our goal is to find a model that best fits the regression line for predicting housing prices given the predictor variables: population, bathrooms, bedrooms, sqrft, lotsize, propertytaxrate, age, tfidf_sum, and sentiment_sum.

A full model can be represented by the following equation:

$$Y_{price} = Beta_0 + Beta_1 X_{population} + Beta_2 X_{bathrooms} + Beta_3 X_{bedrooms} + Beta_4 X_{sqrft} +$$
$$Beta_5 X_{lotsize} + Beta_6 X_{propertytaxrate} + Beta_7 X_{age} + Beta_8 X_{tfidf\_sum} + Beta_9 X_{sentiment\_sum}$$

4.1 Multiple Liner Regression

4.1.1 Full Model

A multiple linear regression model was fitted on our training data set. The output of this model can be seen in Table 4.1 along with other statistics in the figure below. From looking at the given output in Table 4.1 we can see that some predictors are not significant in predicting the price of a home at the α =0.05 level. These predictors include lotsize and age. When looking at the RMSE for this model, we get an RMSE of 1841651.

Table 4.1: Full Model Output

Coefficients	**Estimate**	**Std. Error**	**P-value**
(Intercept)	1.30e+01	1.13e-01	$< 2e-16 * **$
population	-5.28e-08	6.38e-09	$< 2e-16 * **$
bathrooms	3.80e-01	1.67e-02	$< 2e-16 * **$
bedrooms	-4.19e-02	1.81e-02	0.0208 *
sqrft	1.40e-05	4.34e-06	0.0013 **
lotsize	7.00e-10	3.78e-09	0.8529
propertytaxrate	-6.81e-01	1.83e-02	$< 2e-16 * **$
age	8.35e-04	4.45e-04	0.0605 .
tfidf_sum	1.02e-01	1.71e-02	2.8e-09 ***
sentiment_sum	6.64e-03	2.17e-03	0.0023 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6276 on 2563 degrees of freedom

Multiple R-squared: 0.509

Adjusted R-squared: 0.508

F-statistic: 296 on 9 and 2563 DF

p-value:$< 2e-16$

4.1.2 AIC Model

A multiple linear regression model was fitted on our training data set and a model selection criterion of AIC was used on the model. The output of this model using AIC can be seen in Table 4.2 along with other statistics in the figure below. When looking at the RMSE for this model, we get an RMSE of 1841651.

Table 4.2: AIC Model Output

Coefficients	**Estimate**	**Std. Error**	**P-value**
(Intercept)	1.30e+01	1.13e-01	$< 2e-16 * **$
population	-5.28e-08	6.37e-09	$< 2e-16 * **$
bathrooms	3.80e-01	1.67e-02	$< 2e-16 * **$
bedrooms	-4.18e-02	1.81e-02	0.0210 *
sqrft	1.40e-05	4.34e-06	0.0013 **
propertytaxrate	-6.81e-01	1.83e-02	$< 2e-16 * **$
age	8.37e-04	4.44e-04	0.0597 .
tfidf_sum	1.02e-01	1.71e-02	2.7e-09 ***
sentiment_sum	6.64e-03	2.17e-03	0.0023 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.628 on 2564 degrees of freedom

Multiple R-squared: 0.509

Adjusted R-squared: 0.508

F-statistic: 333 on 8 and 2564 DF

p-value:$< 2e-16$

4.1.3 Cross Validation Model

A multiple linear regression model was fitted on our training data set and a model selection criterion of K-Fold Cross Validation was used on the model. The output of this model using a 5-Fold CV can be seen in Table 4.3 along with other statistics in the figure below. When looking at the RMSE for this model, we get an RMSE of 1841651.

Table 4.3: CV Model Output

Coefficients	**Estimate**	**Std. Error**	**P-value**
(Intercept)	1.35e+01	7.42e-02	$< 2e-16 * **$
population	-5.20e-08	6.39e-09	6.1e-16 ***
bathrooms	4.01e-01	1.66e-02	$< 2e-16 * **$
bedrooms	-4.38e-025	1.83e-02	0.0167 *
sqrft	1.42e-05	4.39e-06	0.0012 **
lotsize	1.09e-09	3.82e-09	0.7750
propertytaxrate	-7.04e-01	1.82e-02	$< 2e-16 * **$
age	1.29e-03	4.45e-04	0.0037 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.635 on 2565 degrees of freedom

Multiple R-squared: 0.498

Adjusted R-squared: 0.497

F-statistic: 364 on 7 and 2565 DF

p-value:$< 2e-16$

4.1.4 Reduced Model

A criterion that was used in determining the best Liner Regression Model was RMSE. In our case, all of our models happened to have the same RMSE (1841651) value. Because p-value is also used as another criterion for variable removal in a model, the full model was chosen. Both lotsize and age were removed from the full model as they had p-values greater than α=0.05. Removing these variables didn't change the RMSE, but it did decrease the individual p-values of the variables which only made them more significant in predicting the price of a house. Additionally, when fitting the model on the trained data to our testing dataset, we get and RMSE of 1443382. The regression models used y=log(price) as the

response, but the computed RMSE used non transformed price and predicted price. A lower RMSE is desired and a good indicator that we did not overfit our training model. Now that we have finalized the best MLR model, it can be expressed as the following:

$$Y_{price} = 1.30e^{+01} - 5.13e^{-08}X_{population} + 3.69e^{-01}X_{bathrooms} + 3.92e^{-02}X_{bedrooms}$$

$$+1.44e^{-05}X_{sqrft} - 6.82e^{-01}X_{propertytaxrate} + 1.05e^{-01}X_{tfidf\_sum} + 6.88e^{-03}X_{sentiment\_sum}$$

Table 4.4: Reduced Model Output

Coefficients	**Estimate**	**Std. Error**	**P-value**
(Intercept)	1.30e+01	1.10e-01	$< 2e-16$ **
population	-5.13e-08	6.33e-09	7.6e-16 ***
bathrooms	3.69e-01	1.55e-02	$< 2e-16$ **
bedrooms	3.92e-02	1.80e-02	0.03000 *
sqrft	1.44e-05	4.34e-06	0.00093 ***
propertytaxrate	-6.82e-01	1.83e-02	$< 2e-16$ **
tfidf_sum	1.05e-01	1.70e-02	7.2e-10 ***
sentiment_sum	6.88e-03	2.17e-03	0.00155 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.628 on 2565 degrees of freedom

Multiple R-squared: 0.509

Adjusted R-squared: 0.507

F-statistic: 379 on 7 and 2565 DF

p-value:$< 2e-16$

As you can see in the model expressed above, each variable has a coefficient. Some are positive while some are negative and will affect price accordingly for every unit increase in a single variable holding all other variables constant. If we take a look at the tfidf_sum variable, we can say that for every unit increase in tfidf_sum, price will increase by 11%

holding all other variables constant. Similarly, if we look at the propertytaxrate variable, we can say that for every unit increase in propertytaxrate, price will decrease by -49.4% holding all other variables constant. Additionally, given an R^2 value of 0.509, our model explains 50.9% of the variance present in our data.

4.2 Random Forest

Random Forest creates the best model by using fitting many trees with different random subset of predictor variables and training data. Our Random Forest Regression model with Number of trees:500 and No. of variables tried at each split: 3, predicts price by averaging the prediction of all the different trees. Looking at Table 4.5, we see that the mean of squared residuals is $9.05e^{+11}$ and 65.2% of variance explained indicate how well the model fits the data. Since all of our other models are being measured using RMSE, we will also need to calculate RMSE for this model. This was simply done by taking the square root of the mean squared residuals which was provided in our model output. The RMSE value for this Random Forest model is 951315. When fitting our Random Forest model on the testing data, we get an RMSE value of 401911.

Table 4.5: Random Forest Output

Mean of squared residuals	**% Var explained**
$9.05e^{+11}$	65.2

Additionally, when looking at figure 4.1 we can visualize variable importance in our Random Forest Model. The x-axis labeled, %IncMSE, denotes the importance of the variable. The higher the number is, the more important that variable is. Looking at figure 4.1, there is a legend for IncNodePurity. IncNodePurity is a measure of variable importance based on the Gini impurity index used for the calculating the splits in trees. The higher the IncNodePurity, the higher that variable importance is. Again, looking at figure 4.1, we can see

that the three most important variables are propertytaxrate, population, and sqrft. We can also see that the three least important variables are tfidf_sum, sentiment_sum, and age.

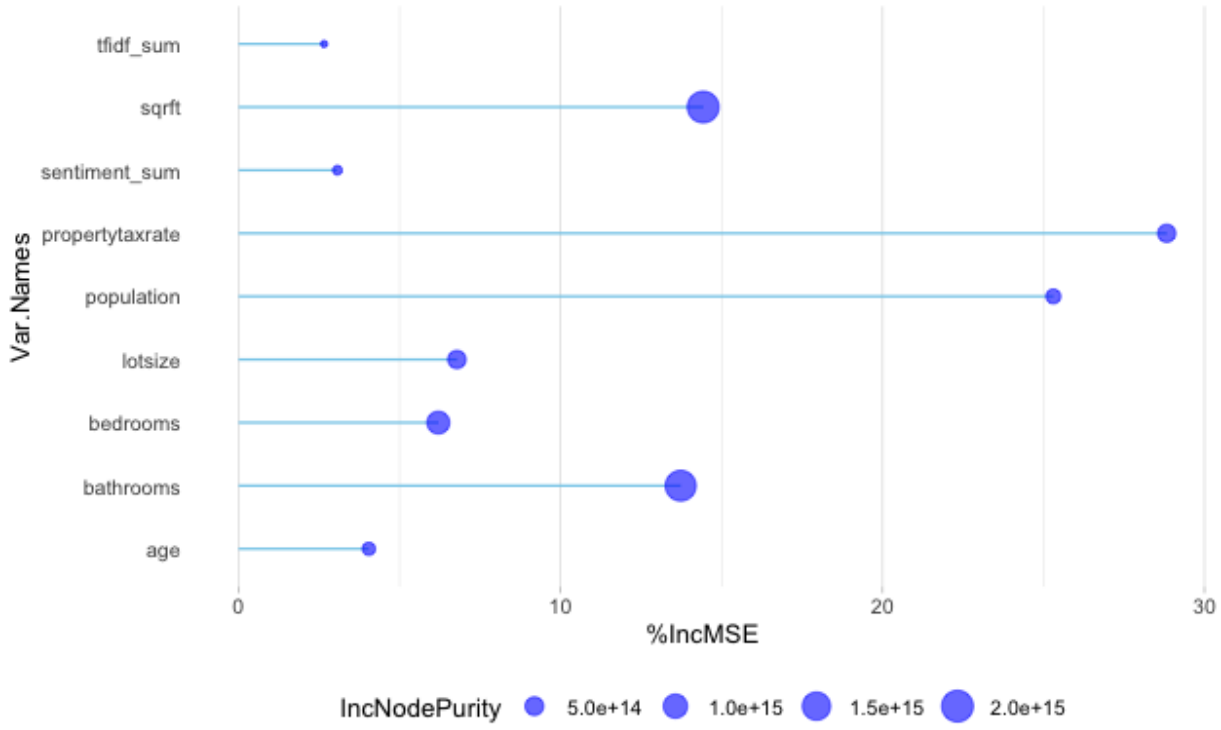

Figure 4.1: Variable Importance

CHAPTER 5

Conclusion

The main objective of this paper is to determine whether the inclusion of real estate descriptions in a regression model can predict selling price. We first explored our dataset by using exploratory data analysis techniques including cleaning missing values, exploring the distribution of the data, creating a corpus with only the textual data of real estate listings, finding the Term Frequency Inverse Document Frequency (TF-IDF) of words, and calculating sentiment scores for words.

We began our analysis by fitting a multiple linear regression model. We used two criterion based methods (AIC and K-Fold CV) to select the best performing multiple linear regression model. We found that the model that performed the best is expressed by

$Y_{price} = 1.30e^{+01} - 5.13e^{-08}X_{population} + 3.69e^{-01}X_{bathrooms} + 3.92e^{-02}X_{bedrooms}$

$+1.44e^{-05}X_{sqrft} - 6.82e^{-01}X_{propertytaxrate} + 1.05e^{-01}X_{tfidf\_sum} + 6.88e^{-03}X_{sentiment\_sum}$ with an RMSE value of 1841651. We then proceeded with fitting a Random Forest model on our data and while it can't be easily represented by a function, it can still be measured with RMSE which happened to be 951315. Both our regression and Random Forest models showed that real estate listing textual data is a significant predictor of real estate housing prices. Additionally our models showed that sentiment is also a significant predictor of real estate housing prices and will affect price positively. In selecting the best model to predict housing prices, a random forest seems to perform the best as it greatly decreases the RMSE of the testing dataset.

These observations also mean that further improvements can be made to improve our

current models. Aside from obtaining a larger dataset with more attributes, future studies can also perform additional hyperparameter tuning on models, which will most improve it's performance. Additionally, future research can focus on disaggregating the model by city as markets can be regional. Nevertheless, the final model does a decent job of including textual data as a predictor of the price of homes.

REFERENCES

[1] Micheline Kamber Jiawei Han, Jian Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2012.

[2] David J Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

[3] S. Roshni R. Sagayam, S.Srinivasan. A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal Of Computational Engineering Research*, Vol. 2(5):1443–1446, 2012.

[4] Shaeela Ayeshaz Ramzan Talib, Muhammad Kashif Hanify and Fakeeha Fatimax. Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11):414–418, 2016.

[5] Gunther Herath, Shanaka & Maier. The hedonic price method in real estate and housing market research. a review of the literature. *Institute for Regional Development and Environment*, pages 1–21, 2010.

[6] Chintan Amrit, Tim Paauw, Robin Aly, and Miha Lavric. Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, 88:402–418, 2017.

[7] Trefor P. Williams and Jie Gong. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, 43:23–29, 2014.

[8] Lorna Maria A. Understanding and writing your first text mining script with r.

[9] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41:853–860, 2014.

[10] Julia Silge and David Robinson. Sentiment analysis with tidy data.

[11] Scikit learn Developers. Cross-validation: evaluating estimator performance.