DSC520 - Statistics For Data Science

Final Project – Real Estate price Prediction

Sri Renganathan Sankaranarayanan

**Introduction :-**

In my Final project, I am planning to develop a model to compute the Log error of the Zillow real estate price predictor (“Zestimate” as it is commonly called) which are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. We will build a model to calculate the residual error of the whole process based on the real estate price sold value, date and region. Reduce error between actual home price and zillow’s estimate. We will perform EDA and correlation analysis and create a model to examine variation in estimates vs. actual home price.

**Research Questions :-**

We will use the publicly available Zillow data from Kaggle. Following are my questions that we will try to find solution.

1. Target is find the residual error of Zestimate predictor based on actual sold value (there can be other market indicators at certain time affecting selling prices in a given market, but we will only consider factual data available in public domain as Kaggle)
2. Classify the variables in the data – Categorical Vs Numerical
3. Define the correlation between various parameters in the data, obtain the main predictor variables
4. Prepare the data – Identify all missing data, use dummy data for missing fields, clean the data
5. Exploratory Data Analysis – Transform, modify and delete unwanted data, and understand relationships between different variables using Plots.
6. Compute Log Error of the Zestimate using historical transactions (Sale price, Property, Region)
7. Compute Standard Deviation of the residual Error and plot them.
8. << We will add more questions as and when we work with the Zillow Data – Professor Bushart , your feed back is welcome >>

**Approach**

Using Scatter plots, Standard deviation and regression models, we will analyze the historical Sale price, real estate property type, sold region, other dependent variable details and the variables that provide more meaningful insights, we will compute the residual error the Zillow real estate value prediction.

**Data**

We will describe three data sets that will be employed for this project. Datasets are related in the Real Estate industry but vary in context. Each dataset will be examined according to their content. A combined report will detail the findings of each dataset.

**Dataset 1**

<https://www.kaggle.com/samdeeplearning/vt-nh-real-estate>

This dataset contains features of houses in three towns in Vermont, which make up a sizable chunk of the real estate firm's business. The dataset is divided into test, train and validate data sets with test having 24 rows, train 138 and validate with 70 rows. There are 28 column describing features such as number of bedrooms, yard size, etc.

**Dataset 2**

<https://www.kaggle.com/quantbruce/real-estate-price-prediction>

Dataset columns are 'transaction\_date' , 'house\_age' , 'distance\_to\_the\_nearest\_MRT\_station' , 'number\_of\_convenience\_stores' , 'latitude' , 'longitude' , 'house\_price\_of\_unit\_area'. There are 500 records in the dataset.

**Dataset 3**

<https://www.kaggle.com/philippsp/exploratory-analysis-zillow>

There are two data sets with over 1 million records each and 58 columns. There are data records for various given years, and we will use the latest data from 2019, 2020 and 2021(if possible).

**Required Packages**

Packages needed for calculation, analysis and plotting the data sets are listed below:

library(data.table)

library(dplyr)

library(ggplot2)

library(stringr)

library(DT)

library(tidyr)

library(corrplot)

library(leaflet)

library(lubridate)

library(readxl)

library(mlogit)

**Graphs, Tables and Plots:-**

We will create histograms and density plots, scatter plots and correlation plots to examine and understand the data.

**Questions for future**

1. What are the common denominators among certain price range?
2. Does Amenities, demography, or school districts determining the price ?
3. Does location determine price?
4. Which affect the price of a home - Number of rooms and the square foot or the region?
5. How Does age affect the price of a home?

**MileStone project Step 2**

**Cleaning Data and Exploratory Data Analysis**

We will then import the dataset, perform preliminary evaluation of the dataset and prepare the dataset for the next phase which will be to test a few hypotheses, perform correlation analysis and select features to perform a regression model.

**Importing and cleaning the data sets**

There are two data sets we will import for analysis:

properties <- read.csv("data/zillow-prize-properties\_2016.csv")

This data set in indexed and contains the following attributes :-

**Graphical user interface, text

Description automatically generated**

A sample of data is shown below:

**Graphical user interface, application, table

Description automatically generated**

As shown, there are columns with numerous NA’s. The following code identifies columns with NA’s and allows us to select columns with less than 75% NA’s.

**Text

Description automatically generated**

The ggplot graph below shows the percentage of NA’s in each column.

**Chart, bar chart

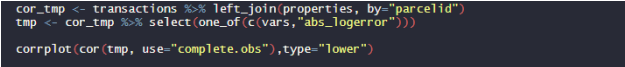
Description automatically generated**

We create a data frame that contains less than 75% NA’s. The good\_features dataframe has two columns-‘features’ and ‘missing\_pct’. The ‘features’ column contains the list of columns with less than 75% NA’s. We would still have to handle the NA’s which may appear in rows of the dataset, but this code guarantees that we remove columns with more that 75% NA’s because they would adversely affect our analysis.

We will use the features in good\_features dataframe for our analysis. For example, the code below identifies the feature with ‘\_num’ in them

****

And code below creates the correlation coefficient diagram

****

**Chart, bubble chart

Description automatically generated**

transactions <- read.csv("zillow-prize-1/train\_2016.csv") This dataset is also indexed by parceled and has additional features listed below:

**Text

Description automatically generated**

Sample of data is shown below

Table

Description automatically generated

This dataset doesn’t need to be cleaned. The parcelid of this dataset and parcelid of the ‘properties’ dataset is related and joining them would show the transaction date and the log error which is the error between the sales price of the property and the estimated price. This log error will be part of the analysis as it shows the estimate vs actual sales price. We may want to investigate the variations in the logerror and find correlations.

**Final Milestone Project Step 3.**

1. Discuss how you plan to uncover new information in the data that is not self-evident.

properties <- read.csv(“zillow-prize-1/properties\_2016.csv”) nrow(properties) # almost 3 million records sample\_20 <- properties[sample(nrow(properties), nrow(properties)\*.20),] sample\_30 <- properties[sample(nrow(properties), nrow(properties)\*.30),] nrow(sample\_20) nrow(sample\_30) write.csv(sample\_20,“zillow-prize-1/properties\_2016\_sample\_20.csv”) write.csv(sample\_30,“zillow-prize-1/properties\_2016\_sample\_30.csv”)

properties <- read.csv(“zillow-prize-1/properties\_2016\_sample\_20.csv”) nrow(properties)

Also note that the output to pdf is not supported, and the absolute log error map at the end of this file does not output to word document, so we copied it from HTML file, and saved the word document as pdf for submission.

1. What are different ways you could look at this data to answer the questions you want to answer?

# Data Preparation

## Features with percentage of missing values

Chart, bar chart

Description automatically generated

## Selected features

X, parcelid, aircon, num\_bathroom, num\_bedroom, quality, num\_bathroom\_calc, area\_total\_calc, area\_live\_finished, fips, num\_bath, num\_garage, area\_garage, flag\_tub, heating, latitude, longitude, area\_lot, zoning\_landuse\_county, zoning\_landuse, zoning\_property, rawcensustractandblock, region\_city, region\_county, region\_neighbor, region\_zip, num\_room, num\_unit, build\_year, flag\_fireplace, tax\_building, tax\_total, tax\_year, tax\_land, tax\_property, tax\_delinquency, censustractandblock

# Data set info

## Summary of transactions data set

## parcelid logerror date

## Min. : 10711738 Min. :-4.60500 2016-07-29: 910

## 1st Qu.: 11559500 1st Qu.:-0.02530 2016-04-29: 902

## Median : 12547337 Median : 0.00600 2016-09-30: 894

## Mean : 12984656 Mean : 0.01146 2016-06-30: 874

## 3rd Qu.: 14227552 3rd Qu.: 0.03920 2016-05-27: 863

## Max. :162960842 Max. : 4.73700 2016-08-31: 737

## (Other) :85095

## head of transactions data set

## parcelid logerror date

## 1 11016594 0.0276 2016-01-01

## 2 14366692 -0.1684 2016-01-01

## 3 12098116 -0.0040 2016-01-01

## 4 12643413 0.0218 2016-01-02

## 5 14432541 -0.0050 2016-01-02

## 6 11509835 -0.2705 2016-01-02

## head of properties data set

## X parcelid aircon architectural\_style area\_basement num\_bathroom

## 1 2342182 12484350 1 NA NA 4

## 2 541880 12181530 NA NA NA 1

## 3 1617969 14222705 NA NA NA 0

## 4 1729095 11467838 NA NA NA 2

## 5 84189 13015665 NA NA NA 1

## 6 1570086 12633721 NA NA NA 4

## num\_bedroom framing quality num\_bathroom\_calc deck area\_firstfloor\_finished

## 1 6 NA 4 4 NA NA

## 2 2 NA 7 1 NA NA

## 3 0 NA NA NA NA NA

## 4 3 NA 7 2 NA NA

## 5 2 NA 7 1 NA NA

## 6 6 NA 7 4 NA NA

## area\_total\_calc area\_live\_finished area\_liveperi\_finished area\_total\_finished

## 1 3315 3315 NA NA

## 2 1088 1088 NA NA

## 3 4231 NA NA NA

## 4 1327 1327 NA NA

## 5 976 976 NA NA

## 6 3030 NA NA 3030

## area\_unknown area\_base fips num\_fireplace num\_bath num\_garage area\_garage

## 1 NA NA 6037 NA 4 NA NA

## 2 NA NA 6037 NA 1 NA NA

## 3 NA 4231 6059 NA NA 0 0

## 4 NA NA 6037 NA 2 NA NA

## 5 NA NA 6037 NA 1 NA NA

## 6 NA NA 6037 NA 4 NA NA

## flag\_tub heating latitude longitude area\_lot num\_pool area\_pool pooltypeid10

## 1 2 33852329 -118125098 6548 NA NA NA

## 2 NA 33986932 -118297793 6008 NA NA NA

## 3 NA 33819470 -117832954 9118 NA NA NA

## 4 7 33958756 -118398254 5145 NA NA NA

## 5 7 34138321 -117908275 7000 NA NA NA

## 6 NA 33782587 -118251435 7000 NA NA NA

## pooltypeid2 pooltypeid7 zoning\_landuse\_county zoning\_landuse zoning\_property

## 1 NA NA 0100 261 LKR1YY

## 2 NA NA 0100 261 LAR2

## 3 NA NA 96 248

## 4 NA NA 0100 261 LAR1

## 5 NA NA 0100 261 AZR1C\*

## 6 NA NA 0200 246 LAR2

## rawcensustractandblock region\_city region\_county region\_neighbor region\_zip

## 1 60375708 12292 3101 NA 96212

## 2 60372372 12447 3101 118208 96025

## 3 60590758 33252 1286 NA 97063

## 4 60372780 12447 3101 7877 96026

## 5 60374006 37015 3101 NA 96464

## 6 60372947 12447 3101 48516 96228

## num\_room story num\_75\_bath material num\_unit area\_patio area\_shed build\_year

## 1 0 NA NA NA 1 NA NA 1950

## 2 0 NA NA NA 1 NA NA 1940

## 3 0 NA NA NA 4 NA NA 1973

## 4 0 NA NA NA 1 NA NA 1944

## 5 0 NA NA NA 1 NA NA 1921

## 6 0 NA NA NA 2 NA NA 1921

## num\_story flag\_fireplace tax\_building tax\_total tax\_year tax\_land

## 1 NA 255065 277073 2015 22008

## 2 NA 22500 172568 2015 150068

## 3 2 212125 353550 2015 141425

## 4 NA 97880 370695 2015 272815

## 5 NA 94231 262698 2015 168467

## 6 NA 225985 242687 2015 16702

## tax\_property tax\_delinquency tax\_delinquency\_year censustractandblock

## 1 3766.44 NA 6.037571e+13

## 2 2186.02 NA 6.037237e+13

## 3 4655.98 NA 6.059076e+13

## 4 4610.51 NA 6.037278e+13

## 5 3303.92 NA 6.037401e+13

## 6 3144.99 NA 6.037295e+13

1. Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information?

## statistics of logerror in transactions

* Mean : 0.0114572
* median : 0.006
* std : ’ 0.1610788
* Max: 4.737
* Min: -4.605

# Data exploration

## Scatter plot of logerror

Chart

Description automatically generated

## histogram of logerror

Chart, histogram

Description automatically generated

## histogram of logerror binned

Chart, histogram

Description automatically generated

## parcelid logerror date year\_month abs\_logerror

## 1 11016594 0.0276 2016-01-01 2016-01-01 0.0276

## 2 14366692 -0.1684 2016-01-01 2016-01-01 0.1684

## 3 12098116 -0.0040 2016-01-01 2016-01-01 0.0040

## 4 12643413 0.0218 2016-01-02 2016-01-01 0.0218

## 5 14432541 -0.0050 2016-01-02 2016-01-01 0.0050

## 6 11509835 -0.2705 2016-01-02 2016-01-01 0.2705

## Scatter plot of abs\_loggerror

Background pattern

Description automatically generated

## parcelid logerror date year\_month abs\_logerror

## 1 11016594 0.0276 2016-01-01 2016-01-01 0.0276

## 2 14366692 -0.1684 2016-01-01 2016-01-01 0.1684

## 3 12098116 -0.0040 2016-01-01 2016-01-01 0.0040

## 4 12643413 0.0218 2016-01-02 2016-01-01 0.0218

## 5 14432541 -0.0050 2016-01-02 2016-01-01 0.0050

## 6 11509835 -0.2705 2016-01-02 2016-01-01 0.2705

1. What types of plots and tables will help you to illustrate the findings to your questions?

## Histogram of abs\_loggerror

Chart, histogram

Description automatically generated

## Histogram of abs\_loggerror binned

Chart, histogram

Description automatically generated

## geom\_bar: na.rm = FALSE

## stat\_bin: binwidth = NULL, bins = NULL, na.rm = FALSE, pad = FALSE

## position\_stack

## graph of abs\_logerror groupped by month of year

Chart, line chart

Description automatically generated

## graph of abs\_logerror within strandard deviation

Chart, line chart

Description automatically generated

## Correlations

## num\_bathroom num\_bedroom num\_bathroom\_calc num\_bath num\_garage num\_room

## 1 NA NA NA NA NA NA

## 2 NA NA NA NA NA NA

## 3 3 2 3 3 NA 0

## 4 NA NA NA NA NA NA

## 5 NA NA NA NA NA NA

## 6 4 4 4 4 NA 0

## num\_unit logerror abs\_logerror

## 1 NA 0.0276 0.0276

## 2 NA -0.1684 0.1684

## 3 1 -0.0040 0.0040

## 4 NA 0.0218 0.0218

## 5 NA -0.0050 0.0050

## 6 1 -0.2705 0.2705

### correlation using ggpairs using a subset of data(~10%)

Table

Description automatically generated with medium confidence

### correlation using corrplot

Chart, bubble chart

Description automatically generated

## parcelid logerror date year\_month abs\_logerror percentile

## 1 11016594 0.0276 2016-01-01 2016-01-01 0.0276 3

## 2 14366692 -0.1684 2016-01-01 2016-01-01 0.1684 5

## 3 12098116 -0.0040 2016-01-01 2016-01-01 0.0040 1

## 4 12643413 0.0218 2016-01-02 2016-01-01 0.0218 3

## 5 14432541 -0.0050 2016-01-02 2016-01-01 0.0050 1

## 6 11509835 -0.2705 2016-01-02 2016-01-01 0.2705 5

### Density plot shows density of worst predicitons is lower in lower latitudes , but higher in median lattitude and lowwer around and above 34400000.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

### denisity of longitude. Worst predictions are in the -118500000 and -118000000

Chart, line chart, histogram

Description automatically generatedChart, line chart, histogram

Description automatically generated

### area\_total\_finished

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

### num\_room

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

### maps based on a sampling of 10000 properties

#### Map of Los Angels county

Map

Description automatically generated

#### Map of Orange County

Map

Description automatically generated

#### Map of Ventura County

Map

Description automatically generated