Boston University

CS 699 A2 – Data Mining

Project Assignment

Classification on Massachusetts housing data

Authors: Jan Allemann

Mike Zhong

Due: April 06th, 2020

Content

[1. Data mining goal 1](#_Toc36990605)

[2. Description of dataset 1](#_Toc36990606)

[3. Tools 2](#_Toc36990607)

[4. Attribute Selection 2](#_Toc36990608)

[5. Mining algorithms 2](#_Toc36990609)

[6. Conclusion 2](#_Toc36990610)

# Data mining goal

The class attribute is the PRICE, which is a continuous dollar value variable. To convert this into categorical ordinal data, we will bin the list prices using an equal width binning process. The widths of the bins is yet to be determined. The application of this classification will be suggest list prices for sellers.

The goal of the project is, to predict the PRICE category

# Description of dataset

The dataset is obtained from <https://www.redfin.com>. Redfin is a commercial real estate listing company that aggregates real estate data for consumer and business use. This data set was obtained by performing a search with no filters on the Boston, Massachusetts area and expanding the scope until a sufficient number of listings were included. The data set contains 350 tuples with 27 attributes each. The attributes are described in the following table.

|  |  |  |
| --- | --- | --- |
| Attribute | Format | Description |
| SALE.TYPE | Categorical | They are all MLS listing |
| SOLD.DATE | Datetime | N/A |
| PROPERTY.TYPE | Categorical | Categorical nominal |
| ADDRESS | String | House number and street name |
| CITY | String | Categorical nominal |
| STATE.OR.PROVINCE | String | Categorical nominal |
| ZIP.OR.POSTAL.CODE | String | Categorical nominal |
| PRICE | Integer | Continuous float |
| BEDS | Integer | Discrete integers |
| BATHS | Integer | Discrete ordinal |
| LOCATION | String | Categorical nominal |
| SQUARE.FEET | Integer | Discrete integers |
| LOT.SIZE | Integer | Discrete integers |
| YEAR.BUILT | Integer | Discrete integers |
| DAYS.ON.MARKET | Integer | Discrete integers |
| X..SQUARE.FEET | Integer | Continuous float |
| HOA.MONTH | Integer | Discrete integers |
| STATUS | Categorical | Categorical nominal |
| NEXT.OPEN.HOUSE.START.TIME | Datetime | Mostly empty |
| NEXT.OPEN.HOUSE.END.TIME | Datetime | Mostly empty |
| URL | String | URL to home on redfin |
| SOURCE | Categorical | Categorical nominal |
| MLS | Integer | Unique ID |
| FAVORITE | Binary | User specific variable |
| INTERESTED | Binary | User specific variable |
| LATITUTDE | Numerical | Continuous float |
| LONGITUDE | Numerical | Continuous float |

# Tools

The different steps of the

# Preprocessing

## Dropping columns

The distribution for each attribute is plotted. There are multiple columns (i.e. ‘Sale Type’ Figure 1) containing the same value in each tuple.

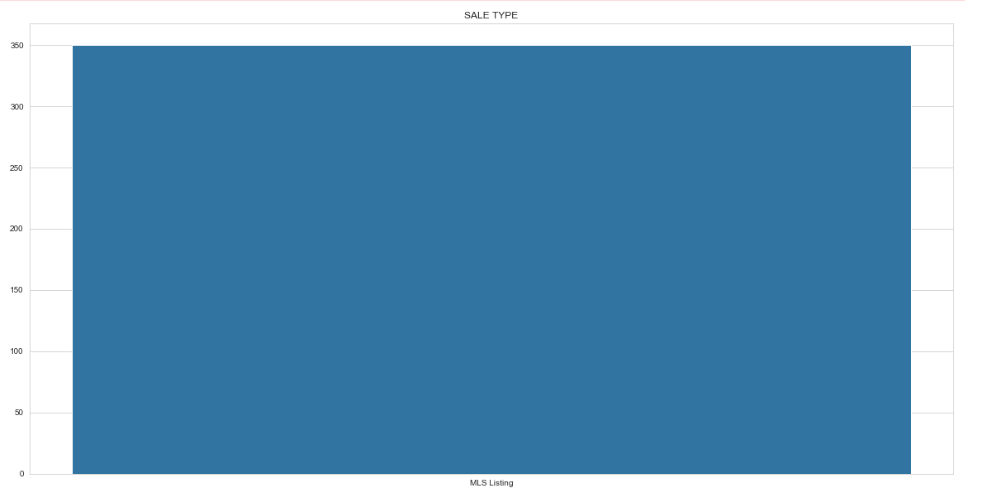


Figure 1: Distribution of the sale type column

Such columns don’t add any information for the classification and therefore the following columns are dropped:

* INTERESTED
* FAVORITE
* STATUS
* SALE TYPE

One column has a unique value in each row and would therefore lead to overfitting on the training set. The column is dropped.

* MLS#

Multiple columns are either mostly empty or are not related to the price and are therefore dropped.

* URL
* SOURCE
* SOLD DATE
* NEXT OPEN HOUSE START TIME
* NEXT OPEN HOUSE END TIME

## Missing Values and Outliers

The column HOA.MONTH describes the monthly HOA-fee for each listing. For listings without any HOA-fee the value should be set to 0 instead of NaN.

By plotting the LOT.SIZE against SQUARE.FOOT we can detect an outlier, that has to be removed.

|  |  |
| --- | --- |
| LOT.SIZE |  |
|  | SQUARE.FOOT |

On the remaining LOT.SIZE, SQUARE.FOOT pairs, we can do a linear regression to interpolate missing values in both columns.

|  |  |
| --- | --- |
| LOT.SIZE |  |
|  | SQUARE.FOOT |

If the attributes BED or BATHS are not containing any value, the tuple can be dropped. These tuples most likely belong to empty lots and are therefore not of interest.

## Binning

The continuous, numeric target attribute PRICE has to be converted into a nominal attribute to create a classification problem.

# Attribute Selection

4 different attribute selection methods are chosen from the WEKA library. The following screenshots show the results of each of the algorithms as well as the self-chosen subset.

|  |  |
| --- | --- |
| CfsSubset | Top five attributes:   * Property Type * Address * City * State or Province * ZIP or Postal Code |
| OneRAttributeEval  A screenshot of a cell phone  Description automatically generated | Top five attributes:   * Square feet * Location * City * Baths * ZIP or postal code |
|  |  |
|  |  |

# Mining algorithms

# Conclusion

Appendix