HOUSING SALES PRICE PREDICTION OF AMES, IOWA

VISHNU SIVADAS

16-05-2019

1. INTRODUCTION

1.1 Background

Ames is a city in Story County, Iowa, United States, located approximately 30 miles (48 km) north of Des Moines in central Iowa. It is best known as the home of Iowa State University (ISU), with leading agriculture, design, engineering, and veterinary medicine colleges. Housing prices of this area depends on a lot of factors. For the people who are looking for buying a house or somebody who wants to sell a house, making a wild guess is difficult and often results in bad business decisions. In this project a model is created to tackle the same.

1.2 Business Problem

When we ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. There are a lot of features to be considered before one can set the price or start negotiating. The project aims in creating a model for predicting housing sales price for Ames, Iowa considering all the important features including the neighbourhood venues

1.3 Target Audience

- House aspirants who can roughly estimate the value of a house based on its features and the average price.
- Real estate people and city planners who can decide what kind of venues to put around their products to maximize selling price.
- House sellers who can optimize their advertisements.

2. DATA DESCRIPTION

2.1 Data Sources

Data sets are prepared from the following sources:

- The Ames Housing dataset is taken from Kaggle.com which was compiled by Dean De Cock for use in data science education. It consists of 79 explanatory variables describing various aspect of residential homes in Ames, Iowa.
- **Foursquare API** is used to get the most common venues of Ames, Iowa. There is a categorical variable ‘Neighborhood’ in Ames housing dataset. Using this variable and ‘geopy’ library in python, latitude and longitude of neighbourhoods’ is found which in turn is used for finding nearby venues using Foursquare API.

2.2 Data Cleaning

In the Ames housing dataset, each neighborhood is given a code, for example 'Blmngtn' was given for 'Bloomington Rd'. Using the code directly ‘geocode’ could not translate it onto the required latitude and longitude. Further data was given in ‘kaggle.com’ describing the neighborhood code into neighborhood name. Same was extracted and made into a Data frame which was passed into ‘geocoder’ for translation after concatenating ‘Ames,Iowa’

Still there were some locations whose translation could not be run by ‘geocoder’. These were searched in the web and following actions were taken:

- If the name is different, decide which one to use after searching on the internet.
- If the neighborhood is missing from the geo data frame, add it's coordinate.
- If the neighborhood is made up, combine them into the larger neighborhood which exist in the geo data frame.

[6]:

	Neighborhood	Neigh	Latitude	Longitude
0	Blmngtn	Bloomington Rd,Ames,Iowa	42.056049	-93.625519
1	Blueste	Bluestem,Ames,Iowa	42.011170	-93.645063
2	BrDale	North Grand mall,Ames,Iowa	42.049331	-93.622661
3	BrkSide	Brookside,Ames,Iowa	42.026770	-93.617055
4	ClearCr	Clear Creek,Ames,Iowa	41.787650	-93.267011
5	CollgCr	College Creek,Ames,Iowa	42.020616	-93.693098
6	Crawfor	Crawford,Ames,Iowa	42.028029	-93.607151
7	Edwards	Edwards,Ames,Iowa	42.025819	-93.668553
8	Gilbert	Zenorsville,Ames,Iowa	42.107206	-93.717999
9	IDOTRR	Ames,Ames,Iowa	42.027910	-93.644644
10	MeadowV	Meadow Village,Ames,Iowa	42.026770	-93.617055

The Foursquare API is used to explore the neighbourhoods' and segment them. The limit was set as 100 venue and the radius 1500 meter for each neighborhood from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

[14]:
```
print(ames_venues.shape)
ames_venues.head()
```

```
(886, 7)
```

[14]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Blmngtn	42.056049	-93.625519	Ge-Angelo's Italian Restaurant	42.054871	-93.622739	Italian Restaurant
1	Blmngtn	42.056049	-93.625519	Bar	42.054446	-93.622896	Bar
2	Blmngtn	42.056049	-93.625519	Flame & Skewer	42.049287	-93.622321	American Restaurant
3	Blmngtn	42.056049	-93.625519	Anytime Fitness	42.054720	-93.622980	Gym / Fitness Center
4	Blmngtn	42.056049	-93.625519	Victoria's Secret PINK	42.049032	-93.622383	Lingerie Store

'One hot encoding' was done the 'Venue Category' and grouped by 'Neighborhood 'to make the required data set.

	Neighborhood	Accessories Store	American Restaurant	Arcade	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Automotive Shop	Bakery	Bank	Bar	Baseball Field	Bed & Breakfast
0	Blmngtn	1	1	0	0	0	0	0	0	0	1	1	0	0
1	Blueste	0	1	1	0	1	0	0	0	1	0	6	0	0
2	BrDale	1	1	0	0	0	0	0	0	0	1	1	0	0
3	BrkSide	0	2	0	0	0	0	0	0	2	2	1	0	0
4	CollgCr	0	0	1	0	0	0	0	0	0	0	0	0	0
5	Crawfor	0	2	0	0	0	0	1	1	2	2	1	2	0

2.3 Feature Selection

Pearson correlation was done on metric variables on Ames housing dataset and following variables were dropped which were found to have poor correlation:

	SalePrice
MSSubClass	-0.087910
OverallCond	-0.065785
BsmtFinSF2	-0.007437
LowQualFinSF	-0.015395
BsmtHalfBath	-0.012623
KitchenAbvGr	-0.071760
EnclosedPorch	-0.070698
3SsnPorch	0.029979
ScreenPorch	0.082308
PoolArea	0.081039
MiscVal	-0.000037
MoSold	0.067222
YrSold	-0.039064

The Ames Housing data set and venue data are merged into a single one after cleaning on which explanatory data analysis is done. .Further training is done on these datasets using Machine Learning (Regression) algorithms to make the required housing sales price prediction model.