A project report on

Virltor : AI powered real estate service

Semester Project

By

Pooja Yadav
Masters in computer science,
San Diego State University

Red ID: 826102050
Email ID: pyadav5723@sdsu.edu

INDEX

ABSTRACT

Artificial intelligence works best in industries that deal with data, and real estate has a lot of data. Historically, neighborhood comparisons and human opinion have served as the foundation for appraisals and estimations; now, these assessments are increasingly being produced by AI-based algorithms.

The manner in which real estate is advertised and sold has been changed by AI. In order to evaluate data on prospective buyers and sellers, spot trends and patterns, and develop customized marketing strategies, real estate brokers can now employ AI. Agents are able to offer more individualized services thanks to AI algorithms' ability to study consumer behavior and preferences.

Zillow, the top online marketplace for real estate, uses AI for its "Zestimates," or the unique valuation estimates it offers, as well as customized recommendations, floor plans, and images. Competitor Redfin, utilizes AI for comparable estimations and collaborated with OpenAI, to develop a chat-based search plugin that can assist home seekers in finding the ideal residence.

The real estate market can be significantly impacted by public opinion. Real estate is a special market where mood, perceptions, and feelings have a significant impact on how much a property is worth and how investors choose to invest. The perception of a neighborhood or location too can heavily influence property prices. A neighborhood with a positive reputation for safety, good schools, amenities, and a thriving community is likely to attract more buyers and command higher prices. Our online platform Virltor (virtual realtor) tried making the real estate service easy and understandable by listing all the properties in the selected neighborhood of San Diego and providing feedback and ratings based on different factors using Machine Learning and Computer Vision algorithms.

ACKNOWLEDGEMENT

I would like to convey my heartfelt gratitude to Dr. Xiaobai Liu for his tremendous support and assistance in the completion of my project**.** Your useful advice and suggestions were really very helpful to me during this project's completion. I learned that I should not use any technology just for its sheer application but understand the limitations and proper applications of each and every technology. Dr. Liu has taught me more than I could ever give him credit for here and I am eternally grateful to him.

Objective:

Renters' perceptions of the rental property, as well as their sense of security, convenience, and lifestyle preferences can all be influenced by the neighborhood. In order to draw in potential renters, property owners and managers frequently highlight the advantages of the neighborhood while marketing rental homes.

I tried to undertake analysis on the San Diego housing data extracted from the different online real estate marketplace and submit the findings. This project's purpose is to use Segmentation and Clustering analysis to rate the housing community based on the features of the community.

Process Followed:

- Collected data from multiple online real estate marketplace and created a database.
- Automated the collection of location's google street view using Selenium and Python.
- Preprocessed the collected location and housing community images using OpenCV.
- Custom trained YOLOv8 model to detect roadways and employed Segment Anything Model by Meta to segment the YOLOv8 model's output.
- Extracted meaningful features using pretrained VGG16 CNN model and reduced dimensionality using PCC. Implemented clustering analysis using K-means on the extracted features.
- Calculated silhouette score for each cluster to evaluate its quality in terms of how well samples are clustered with other samples that are similar to each other.

Dataset Information:

Sn	Columns Name	Data Type	Description
1	home_id	integer	unique string home_id
2	address	string	home address
3	city	string	city of the house
4	state	string	state of the house
5	postal_code	integer	postal or zip code of the house
6	Price	float	price
7	beds	integer	No. of beds in the apartment/house
8	baths	integer	No. of bathrooms in the apartment/house
9	square_footage	integer	size of the house in square foot
10	year_built	integer	The year in which it was built
11	latitude	string	latitude
12	longitude	string	longitude
14	redfin_url	string	redfin (online marketplace url)
15	property_type	string	indicating residential or commercial property
16	location	string	Location/City of the apartment
17	dollars_per_Sqft	float	per square feet dollar rate
18	status	string	Indicating sold or still on sale

No. of columns: 18 and number of rows: 15,000

Data Extraction and Pre-Processing:

1. **Data Extraction:**

   Redfin, a well-known home search tool and discount real estate brokerage, was mined for housing data, including house number, address, street, built year, number of rooms and bathrooms, price, status, and the type of the property to create a SQL database. Converted the SQL to Json data and utilized Python and Selenium webdriver to automate the 4 view google maps image data collection for each housing location.

2. **Data Pre-Processing:**

   To ensure consistency and optimized processing, the images have been resized and scaled to a standard resolution. This step helped reduce computational complexity and prepared images for subsequent analysis. By leveraging OpenCV's computer vision functions in Python, the raw image data from the collected housing community details was transformed into meaningful representations, enabling a more in-depth analysis of the properties and their surrounding environments. The preprocessed data can now be utilized for various purposes, including image-based machine learning models, visualizations, and further insights into the housing market.

Data after pre-processing:

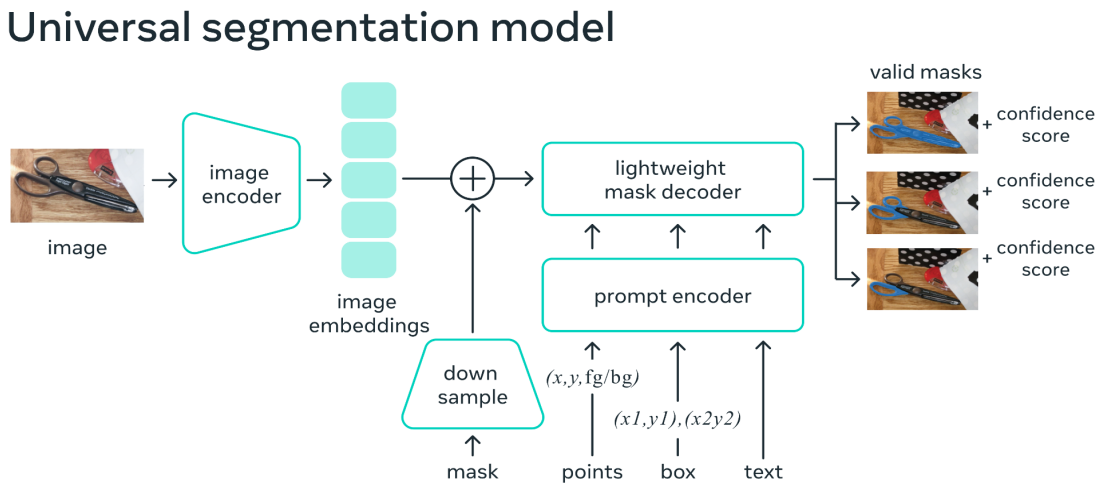
Segmentation:

Image semantic segmentation is a computer vision technique that involves dividing an image into meaningful and semantically distinct regions or segments. i.e., it aims to classify each pixel in the image into specific categories, such as objects, regions, or classes. The goal of semantic segmentation is to assign a label to every pixel, indicating which object or category it belongs to within the image.

Semantic segmentation of roadways:

To segment roadways, a custom dataset was used to train a YOLOv8 (You Only Look Once version 8) model, a popular object detection algorithm capable of detecting and localizing multiple objects in an image. The YOLOv8 model was specifically trained to identify and locate roadways within images. Once the YOLOv8 model was trained, its output was further processed using the "Segment Anything" model by Meta and classified the output images as Good or Bad segmentation based on the confidence score. The "Segment Anything" model is a semantic segmentation algorithm that assigns a semantic label to each pixel in the image, effectively segmenting the image into different classes or categories. By using the "Segment Anything" model on the YOLOv8 model's output, the roadways detected by YOLOv8 were semantically segmented into separate regions, classifying each pixel as either belonging to the roadway or to other objects in the scene.

SAM BY META:



Good Segmentation Output:

Mask 1, Score: 0.882

Mask 1, Score: 0.985

Bad Segmentation Output:

Mask 1, Score: 0.255

Cluster Analysis:

Feature Extraction:

Feature extraction using a pretrained VGG16 is a powerful technique in computer vision, enabling efficient and effective learning of meaningful visual representations for a wide range of tasks. VGG16 is a deep neural network with 16 layers, including convolutional and fully connected layers that was originally trained on a large dataset for image classification. However, its early layers are known to capture general image features that can be used for various tasks beyond classification.

Instead of training the VGG16 model from scratch on a new dataset, we can leverage the knowledge it gained from ImageNet and use it as a feature extractor for other tasks. This is known as transfer learning, where we transfer the learned features from image classification to feature extraction.

Feature Extraction Process:

To extract features, we used the early convolutional layers of VGG16 (usually up to a specified layer) because these layers captured low-level features like edges, textures, and basic patterns. These features can be generic and useful for various computer vision tasks. We discarded the fully connected layers and the final classification layer of VGG16, as they were specific to ImageNet's classification task and not relevant to feature extraction for other tasks. For each input image, we forward-pass it through the truncated VGG16 model (with removed classification layers). The output of the desired convolutional layer serves as the extracted features for that image.

PCA - Dimensionality Reduction:

Principal Component Analysis is a technique of feature extraction that maps a higher dimensional feature space to a lower-dimensional feature space. While reducing the number of dimensions, PCA ensures that maximum information of the original dataset is retained in the dataset with the reduced number of dimensions and the correlation between the newly obtained Principal Components is minimum. The new features obtained after applying PCA are called Principal Components and are denoted as PCi.

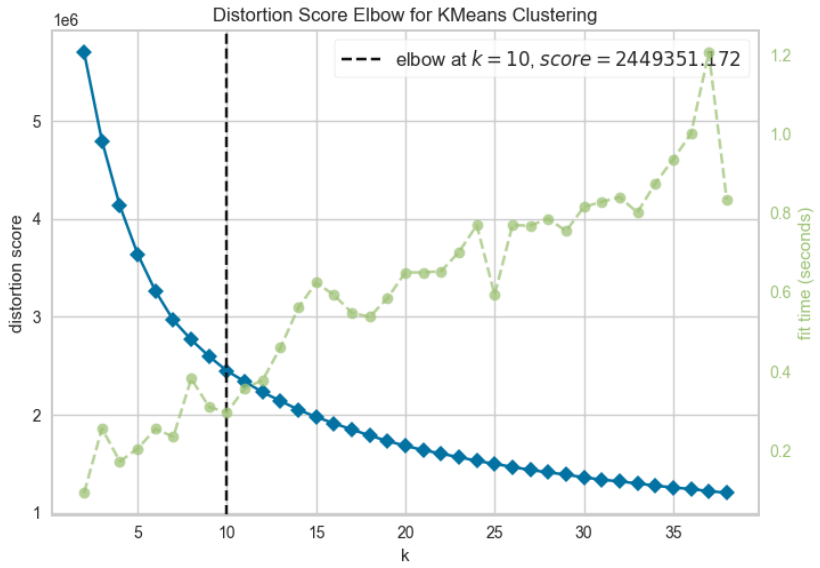
After feature extraction we normalized the range of our features so that our algorithm can interpret them on the same scale. We transformed our normalized features using PCA to get a dimension of (38318 , 5) and applied K-means clustering on the transformed features.

K-means:

The k-means algorithm is a method for discovering the centers of clusters. It is called an unsupervised learning method because the algorithm is not told what the correct clusters are; it must infer clusters from the data. The k-means algorithm finds k centroids within a dataset that each correspond to a cluster of inputs.
Once features were transformed using PCA, applied K-means clustering with default k-means++ initialization and clustered our data into 10 different clusters. Elbow method implementation was done for k values ranging from 3 to 50, and it was found that 10 was the best k value.

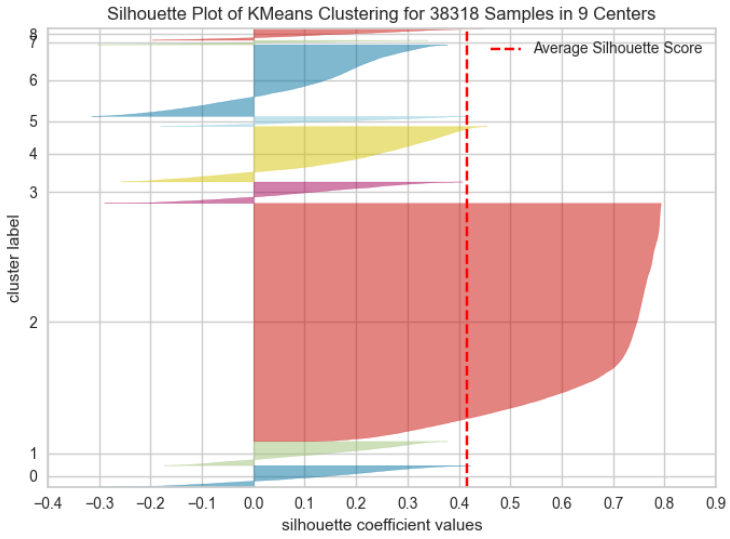
Elbow Method:

The elbow method is a popular unsupervised learning algorithm used in K-Means clustering. Unlike supervised learning, K-Means doesn't require labeled data. It involves randomly initializing K cluster centroids and iteratively adjusting them until they stop moving. From the below graph we can observe that we have an elbow point at k = 10. i.e 10 is an optimal number for the clusters.



Silhouette Score:

The Silhouette score in the K-Means clustering algorithm is between -1 and 1. This score represents how well the data point has been clustered, and scores above 0 are seen as good.

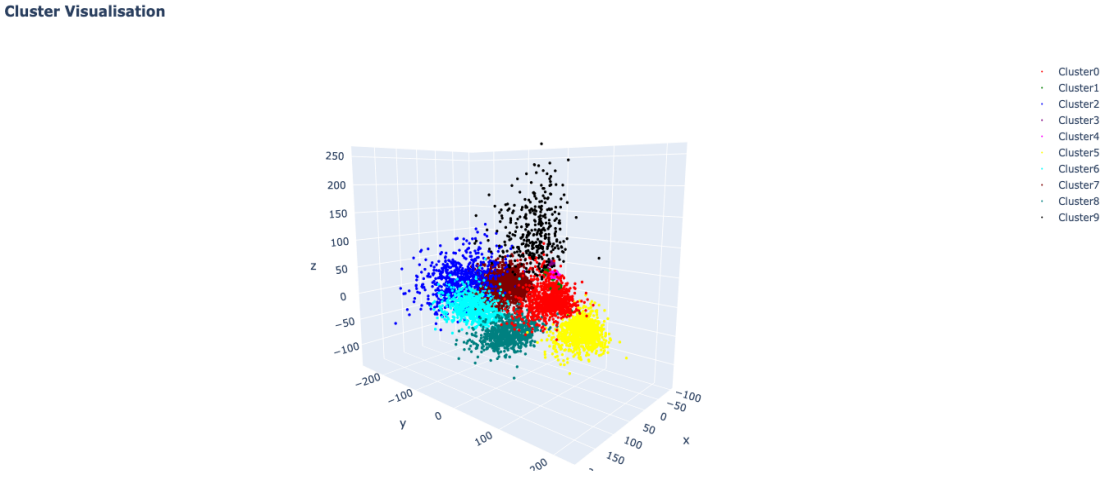


From the above graph it can be inferred that our average silhouette score is greater than 0, i.e our k-means algorithm was able to properly cluster most of our image data.

Silhouette Score for K = 10 Clusters:

[0.42318946, 0.06263669, 0.15474442, 0.048972648, 0.12355146, 0.14633566, 0.15660276, 0.59133714, 0.10567224, 0.19421431]

Cluster Visualization:



Conclusion:

The housing data analysis has provided understanding and valuable insights into the real estate sector. By implementing advanced Machine Learning and Data Analysis we were able to collect and process millions of housing communities of San Diego enabling us to rate and evaluate some meaningful insights.

The implementation of the Segment Anything Model allowed us to achieve accurate semantic segmentation with a very high confidence rate. This approach enhanced our real estate service and empowered homebuyers to make informed decisions.

Clustering analysis allowed us to separate our data into different groups based on their visual features and attributes. It not only allowed us to separate our data into different groups but also revealed hidden relationships and trends within the data. These findings can inform the potential stakeholders about the housing trends, communities, investment options, etc.

The combination of data analysis and machine learning with the traditional housing market data has enriched our understanding of how visual features impact housing prices, amenities, and overall demand.

This research of home image data has shown the strength and potential of data-driven approaches in altering how we perceive and engage with the real estate market. We can continue to promote innovation and enhance decision-making processes within the real estate sector by utilizing artificial intelligence and data analytics, ultimately leading to the creation of a more open and effective housing market for all parties.

Future Scope:

The results and findings derived from this data analysis have far-reaching implications for the real estate industry. Beyond the immediate applications of our real estate service, the methodologies and techniques developed can be adapted to analyze housing data in other regions, cities, or states. further empowering potential buyers with valuable insights. As with any data analysis project, there are areas for further improvement Fine-tuning clustering algorithms, incorporating additional features, and validating results through external sources can enhance the accuracy and reliability of our findings. Additionally, ongoing updates to the dataset will ensure that our analysis remains relevant and reflective of the dynamic real estate market.

As image analysis and computer vision technologies continue to grow, we can develop more advanced algorithms that can extract and process intricate visual details of the housing data. Clustering-based analysis can contribute to improving transparency in the real estate market. By providing comprehensive and granular insights into different housing clusters, empowering potential buyers and sellers with accurate information, reducing information asymmetry, and promoting a more equitable market.

References:

1. https://stephenallwright.com/good-clustering-metrics/#:~:text=The%20most%20common%20ways%20of,Calinski%2DHarabaz%20Index
2. https://www.sciencedirect.com/science/article/pii/S0968090X1830891X#:~:text=Feature%20extraction%20is%20one%20of,patterns%20distinguishable%20from%20each%20other.
3. https://medium.com/@joel_34096/k-means-clustering-for-image-classification-a648f28bdc47#:~:text=Yes!,Image%20Classification%20of%20MNIST%20dataset.
4. https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918?source=-----7315defb5918----2----------------------------
5. https://towardsdatascience.com/k-means-and-pca-for-image-clustering-a-visual-analysis-8e10d4abba40?source=-----8e10d4abba40----4----------------------------
6. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c?source=-----e976bb81d10c----8----------------------------
7. https://paperswithcode.com/task/image-clustering
8. https://segment-anything.com/
9. https://docs.ultralytics.com/
10. https://universe.roboflow.com/
11. https://ultralytics.com/yolov8
12. https://www.redfin.com/
13. https://www.zillow.com/
14. https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918
15. https://en.wikipedia.org/wiki/Principal_component_analysis