Name – Varun Sinsinwar

Enrollment no. – 23112110

Branch – Chemical Engineering (B.Tech. 3rd Year)

Project Report

Satellite Imagery–Based Multimodal Property Valuation

1. Abstract

This project focuses on building a multimodal machine learning pipeline to predict residential property prices by integrating tabular housing attributes with satellite imagery. Traditional valuation models rely heavily on structured features such as size, location, and age of a property. This work extends that paradigm by incorporating visual neighborhood context extracted from satellite images.

2. Problem Statement

A Real Estate Analytics firm aims to improve its valuation framework by developing a Multimodal Regression Pipeline that predicts property market value using both tabular data and satellite imagery.

We are provided with historical housing data (including coordinates) and must programmatically acquire visual data to capture environmental context. The goal is to build a model that accurately values assets by integrating "curb appeal" and neighborhood characteristics (like green cover or road density) into traditional pricing models.

3. Dataset Description

The dataset consists of **Base Data** (tabular) with numerical attributes such as living area, latitude, longitude, grade, waterfront, and year built, and **Visual Data** which contains **Satellite images** corresponding to each property. They were collected using geographic coordinates and **Mapbox** API is used to get images and images were saved by their house id in a folder. Images have **zoom 15** and resolution of **256 x 256** pixels. A zoom of 15 is chosen so that we can get the **neighborhood** data, which includes **Greenery, House Density, water bodies and Roads**. Target Variable is “price” of the house. The tabular data has 16110 rows and 21 columns.

4. Preprocessing

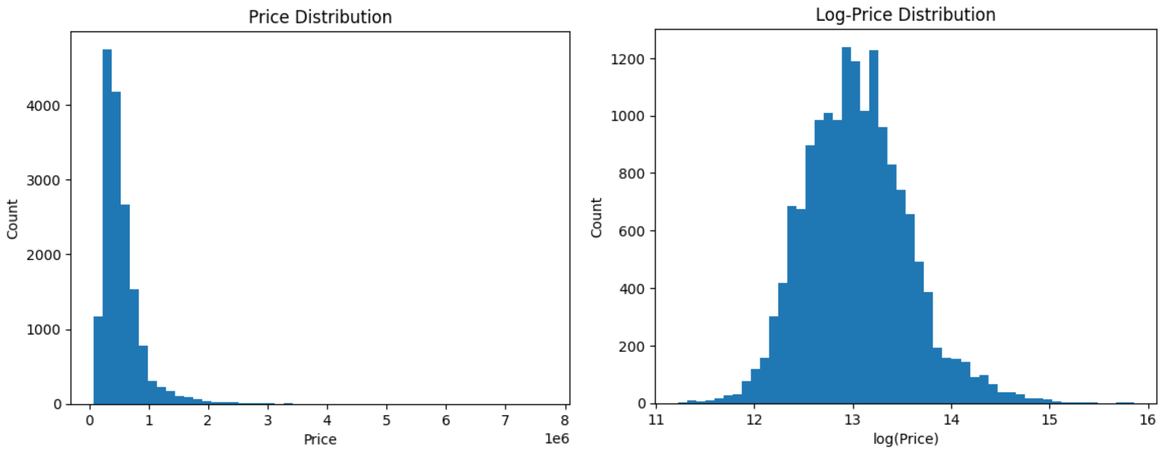
Tabular data preprocessing involved handling missing values, numerical scaling, and encoding of ordinal features. Some features (**id, date, zipcode**) were deleted as those were of no importance for training.

All the categorical variables (**waterfront, condition, grade**) were ordinal features and they don’t need any one hot encoding.

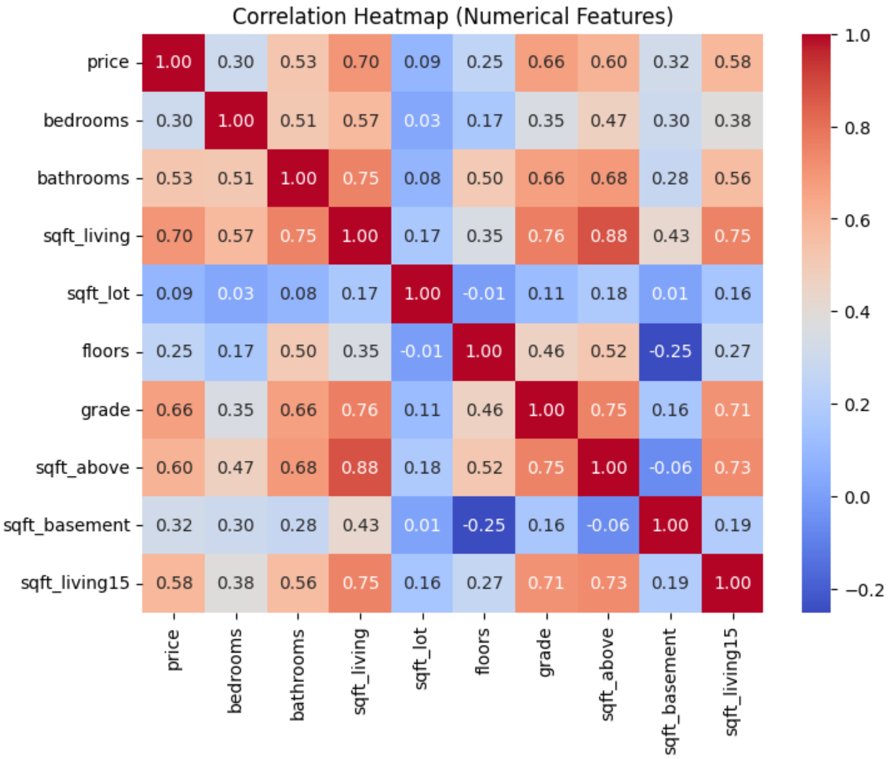
Satellite images were resized and normalized according to the demand of CNN model. A train–validation split of 0.8:0.2 was performed prior to any data training to avoid data leakage. **Target is log transformed** as log(price) represented normal distribution in better manner. This we got to know from EDA.

5. Exploratory Data Analysis

- Below Histogram plots show the price and log price distribution. Clearly the price distribution is heavily right-skewed. Most houses are in the lower-mid price range. Few luxury properties create a long tail.



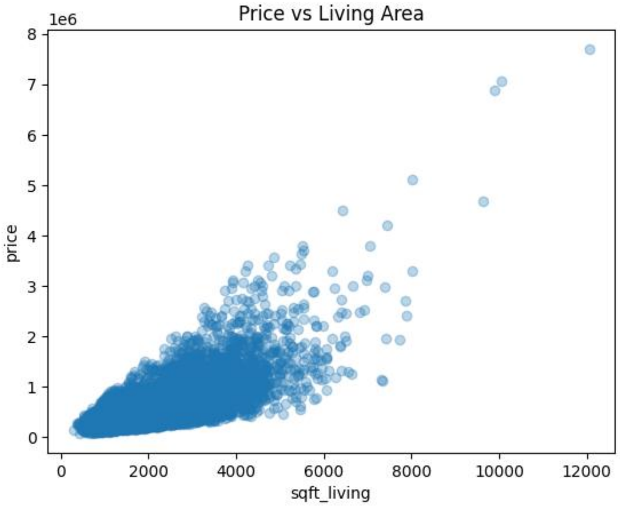
- Log(price) looks close to a normal distribution. Hence it is much better behaved for regression. So, this is used as target for modelling, and exponentiated the predictions back.
- Following is the correlation matrix heatmap.

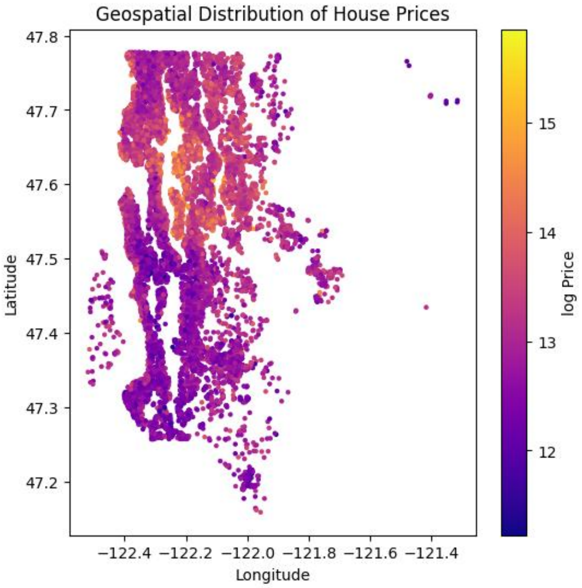


Some of the highest correlations with price:

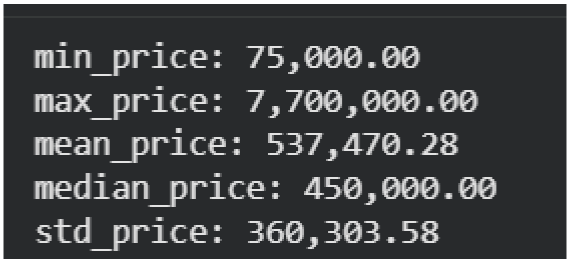
sqft_living (living area)	0.70
grade (construction quality)	0.66
sqft_above (interior space above ground level)	0.60
bathrooms	0.53
sqft_living15 (average living area of nearest 15 neighbors)	0.58

- sqft_lot (total land area) has surprisingly low correlation. This means house **interior quality & size matter more** than raw land size. Average living area is also important. If all houses in the neighbor are big, price is more.

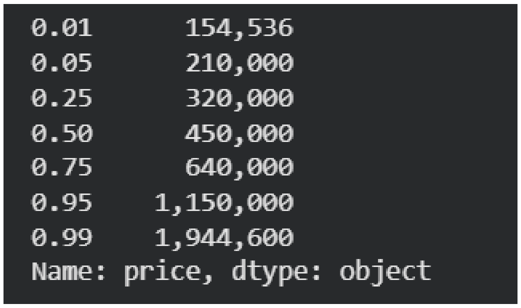




- For price vs living area, clear positive relationship, Variance increases with size.
- The geospatial distribution shows the houses are from one city. Expensive houses are spatially concentrated. Low priced houses have more concentration in the dense areas.
- Following are some more statistics. The second image shows quantiles of the price.



```
min_price: 75,000.00
max_price: 7,700,000.00
mean_price: 537,470.28
median_price: 450,000.00
std_price: 360,303.58
```



```
0.01      154,536
0.05      210,000
0.25      320,000
0.50      450,000
0.75      640,000
0.95    1,150,000
0.99    1,944,600
Name: price, dtype: object
```

- From the satellite images, we can see that houses close to a big water body (probably a lake or river) have the highest prices. The houses in a locality where density is high and less vegetation have the lowest prices. The price can also be low if the house is somewhere alone in high forest area. GradCAM also tells about these features.

Images of Top 3 highest price houses:

mapbox
© Maxar © Mapbox © OpenStreetMap
mapbox
© Maxar © Mapbox © OpenStreetMap
mapbox
© Maxar © Mapbox © OpenStreetMap

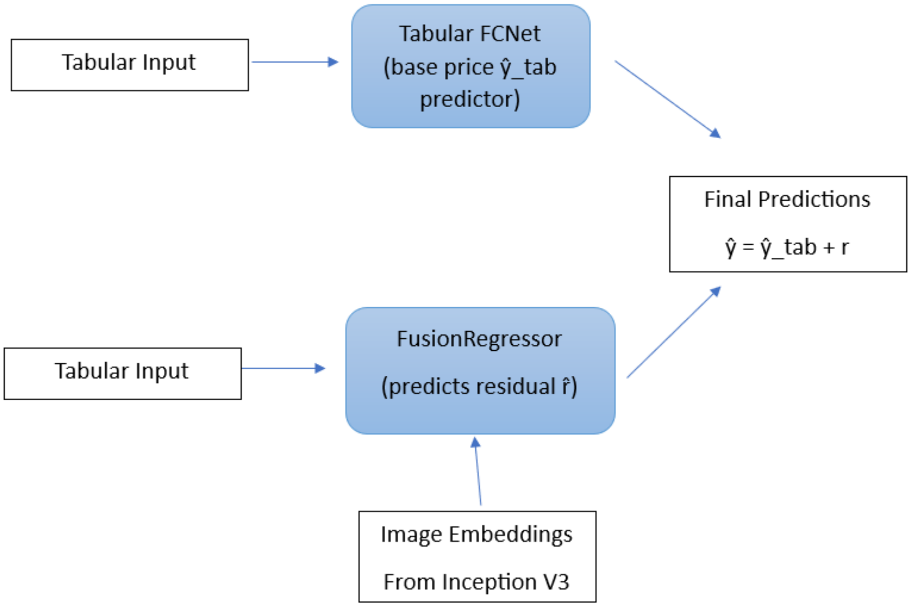
Images of Bottom 3 lowest price houses:

mapbox
© Maxar © Mapbox © OpenStreetMap
mapbox
© Maxar © Mapbox © OpenStreetMap
mapbox
© Maxar © Mapbox © OpenStreetMap

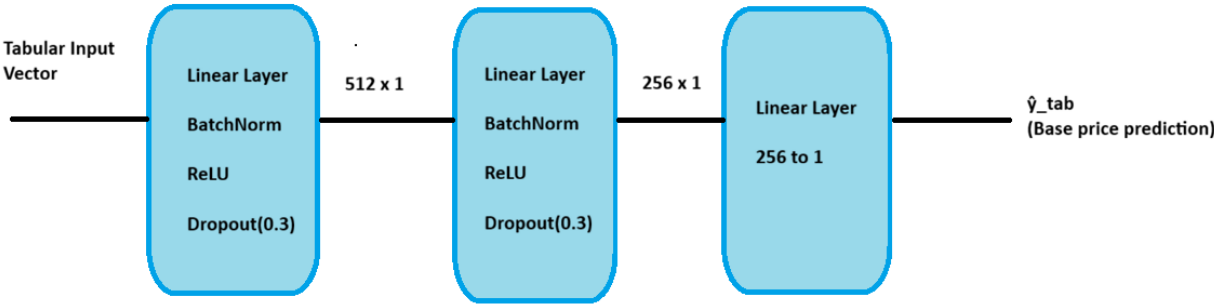
6. Model Architecture

The final model used has a multimodal architecture which can be understood from the figures below. It is a residual network-based model. It can be divided into two parts: (a.) the base price predictor which uses only tabular dataset, and (b.) a bimodal fusion network which uses both Tabular and Image embedding data to predict the residuals. Finally, they are added to get the final predictions. The image embeddings are extracted using pretrained **Inception V3** model which is trained ImageNet Dataset. It makes **2048**-dimensional embeddings which are then fed into a fully-connected network to get 64-dimensional vector. This vector is concatenated with the 40-dimensional coming from the FC layer of the tabular data. After passing through a classifier layer, it predicts the residual prices.

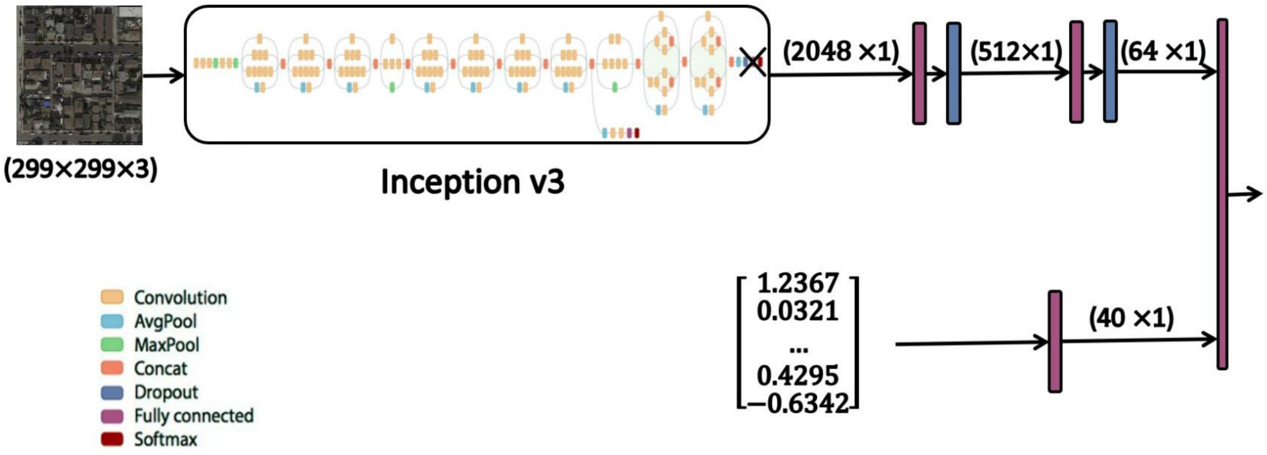
This particular model was chosen after a lot of experiments with baseline models and different experiments. The problem was that images were not able to explain the variance on their own (their R2 score was close to zero). This was harming our final predictions. So, the fusion network was used only for residual predictions, and Tabular data provided most of the information. In below sections, we can also see the results with this and other models.



Tabular data network (TabularFCNet)



Fusion Network



Training Strategy

Models were trained using **mean squared error loss** with **Adam** optimization. Weight decay (L2 normalization) and dropout were applied to mitigate overfitting. Model performance was monitored using RMSE and R^2 metrics at each epoch.

Experiments and Results

Baseline experiments using only tabular data achieved strong performance with R^2 close to 0.87–0.89. Then using only images for prediction gave close to zero R^2, which tells that the images are not good for this experiment. Tabular data is on its own capable of explaining almost all the patterns. There is large correlation of price with many tabular attributes. The reasons for images not being useful can be multiple:

- There are multiple variables which decide the price of a house. We got info about interior and size of a house from the tabular data, and about the neighborhood from image data, but the location of the house cannot be known. Like if the house is at the center of the city, it can have high price compared to those on outskirts. The images can only tell about 15 – 20 houses around the target house, but not about the location.
- There is no model trained specifically for satellite image data. Training our own model is not feasible for 16110 images only. It can also take a lot of time and resources. So, Inception V3 is used for image embeddings, but being trained on ImageNet dataset, does not provide embeddings especially designed for satellite images.

Naive fusion (using only FusionRegressor) initially caused performance degradation due to overfitting. Residual fusion and regularization strategies led to stable training and gradual improvements, narrowing the gap with the tabular-only baseline. The final results for the experiments are given in the following table:

Model	RMSE (validation)	R^2 score (validation)
TabularFCNet (only tabular data)	0.1900	0.8692
FusionRegressor (Fusion network)	0.2399	0.7914
Residual Network (Final Model used)	**0.1787**	**0.8865**

Challenges

Key challenges included overfitting in multimodal fusion, training instability, and increased computational cost. Getting the perfect architecture which could combine the features from table and images both was also a challenge. Grad-CAM interpretability on regression models also required careful architectural handling.

Conclusion

This project demonstrates that satellite imagery can complement tabular real estate data when fused carefully. Residual learning proved more effective than direct feature concatenation. Future work may explore transformers, attention-based fusion, and stronger data augmentation strategies.

Future Work

Future extensions include hyperparameter optimization, self-supervised pretraining on satellite imagery, temporal market modeling, and improved explainability using visual attribute on methods. We need a model which is particularly trained on satellite images data. If proper computational resources are provided, this can be achieved by experimentation with model architectures. This can greatly improve the performance as in that case the images will really contribute to the feature space.