Report
Mini-Project 3: House Price Estimation Using Both Visual and Textual Data
CSC 215-01 Artificial Intelligence (Spring 2020)
Due Date: 23rd March 2020

Team Members
Swapnali Shrikhande (Student ID: 220252046)
Asmita Shrivastav (Student ID: 220268166)

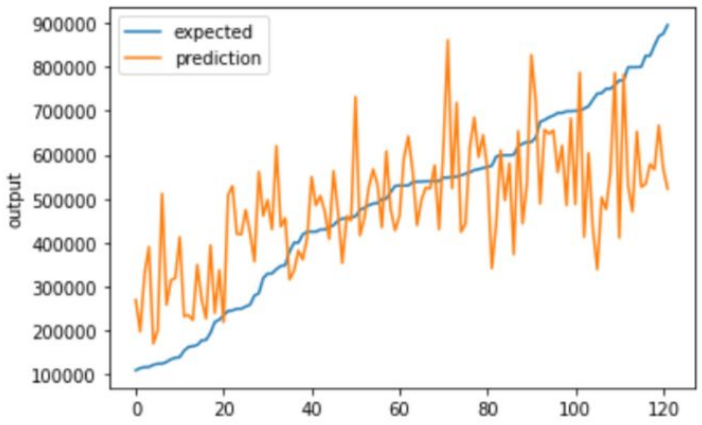
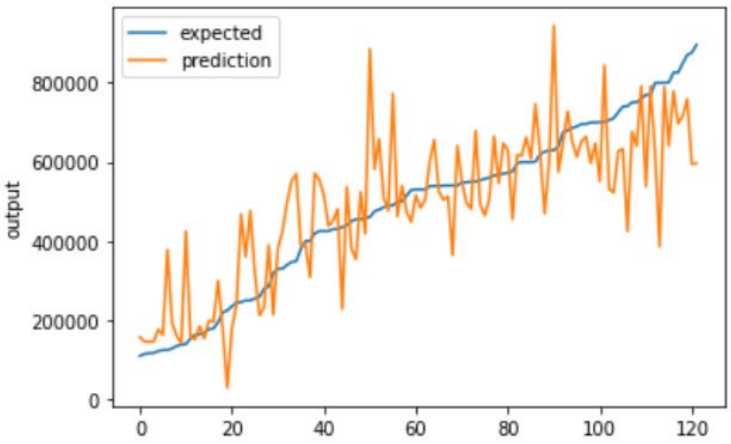
Problem Statement
With the constantly increasing house prices all around the world has become a huge hurdle for many to buy larger houses. Moreover, with limited investment opportunities in certain countries, investment in housing has become difficult. Also, it has been found that there have been discrepancies in the past between the house online images, house textual information and the actual price of the house. Here, we have combined these uncertainty factors to cumulatively make an accurate prediction for the said house. This gives us the difference between the actual price and the listed price of the house, thus helping us in making wise decisions.

Methodology
Research
We read the research paper presented in the project requirement document by Eman H. Ahmed and Mohamed N. Moustafa. The paper gave us a clear idea about the approach to be followed. Moreover, we read other online articles about the method in which prices of the houses are decided. In our research we found that the higher price range houses determined the prices of all other houses in that locality.

Data pre-processing
For textual dataset:
The text file contained data of number of bedrooms, number of bathrooms, area of the house, zipcode and its price. From this we only considered 406 rows by removing outliers of houses with prices <= 900,000 and >= 100,000. We normalized the data using z-score method and performed hot-encoding on zipcode. Further a 30-70 split was performed for train and test data. As an additional feature, we considered zipcodes with most expensive houses which was reflective of the housing price in that locality.

For image dataset:
As per the project requirement, the images were resized to 64x64 and images of rooms belonging to the same house were grouped into a 128x128 image. This grouping was performed using opencv and numpy functions. We also dropped all rows that mapped to the text outlier records. Like the textual dataset, a 30-70 split was performed for train and test data. As an additional feature we also ran the model with individual images, which improved the prediction accuracy.

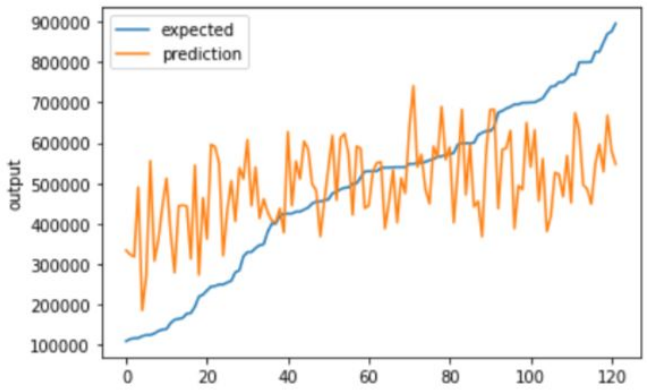
Experimental Results and Analysis of Models

Model	RMSE	R2 score	Regression lift chart
CNN: 2 conv2D & 2 maxpooling2D, relu, kernel size = (1, 5), (2, 2) + NN: 3 Dense layers of 64, 32, 20 neurons, relu + Merge layer: 1 Dense layer of 16 neurons, relu + Output layer: 1 Dense, adam	164766	0.3974	
CNN: 2 conv2D & 2 maxpooling2D, relu, kernel size = (3, 3), (1, 5) + NN: 3 Dense layers of 128, 64, 20 neurons, relu + Merge layer: 1 Dense layer of 16 neurons, relu + Output layer: 1 Dense, adam	127873	0.6370	

Best Model: RMSE score of 127873

Additional Features

1. Improving prediction accuracy by considering four images separately
   RMSE: 187364, R2-score: 0.22



2. Improving prediction accuracy by removing outliers differently
   RMSE: 211801, R2-score: 0.002

3. Comparison of best model with research paper

	Our model	Emam's model
RMSE	1.27873 * 10^5	2.79555 * 10^6

   Thus, our model achieved a slightly better accuracy than Emam's model

Task Division

Swapnali Shrikhande: Parameter tuning, Better outlier feature additional feature, Comparison with research paper additional feature, Report

Asmita Shrivastav: Data pre-processing, Model implementation, Individual image resizing, Individual image run additional feature

Project Reflection:

Through this project we learned the keras API. We learned that such multi-layer modeling of the dataset gives us a better prediction accuracy rate. Moreover, we learned how to employ using different formats of datasets (for e.g. here as image and text) and combine their prediction to make an overall prediction. We also learned how to use Google Colab GPU for modelling with heavy datasets. The use of opencv taught us more about image processing and resizing. The labs, and cheat sheet provided by the professor were of great help in our development process.