Predicting housing prices for real estate companies

Problem Statement
Using Machine Learning
Prices of real estate properties are sophisticatedly linked with our economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available. Therefore, the goal of this project is to use machine learning to predict the selling prices of houses based on many economic factors.

A systematic method can be built to derive a layered knowledge graph and design a structured Deep Neural Network (DNN) based on it. Neurons in a structured DNN are structurally connected, which makes the network time and space efficient; and thus, it requires fewer data points for training. The structured DNN model has been designed to learn from the most recently captured data points which allows the model to adapt to the latest market trends. To demonstrate the effectiveness of the proposed approach, we can use a case study of assessing real properties in small towns.

Background
One heuristic dataset commonly used for regression analysis of housing prices is the Boston suburban housing dataset. Former analyses have found that the prices of houses in that dataset are most strongly dependent on their size and the geographical location. Until recently, basic algorithms such as linear regression can achieve 0.113 prediction errors using both intrinsic features of the real estate properties (living area, number of rooms, etc.) and additional geographical features (socio demo-graphical features such as average income, population density, etc.).

Previous work on predicting house prices has been based on regression analysis and machine learning techniques. Local linear models and random forest models, fuzzy reasoning, Backpropagation neural networks and Elman neural network can been used to forecast real estate prices. Out of these, it is found that Elman neural network can forecast more accurately and constringe faster than other approaches. Nguyen and Cripps compared the predictive performance of artificial neuralnetworks (ANNs) and multiple regression analysis for single family housing sales. They found that when enough data points were available for training, ANNs could perform better than multiple linear regressions. Additionally, a latent manifold model with two trainable components can also be used to evaluate house prices where a parametric component is used for predicting the “intrinsic” price of a house, and a non-parametric component can calculate the desirability of the neighborhood.

Unlike the existing approaches, we propose to use a deep learning approach with structured DNNs, which may outperform the conventional tools for real estate assessments.

Methodology
A framework for training a DNN in real time is shown in figure 1.

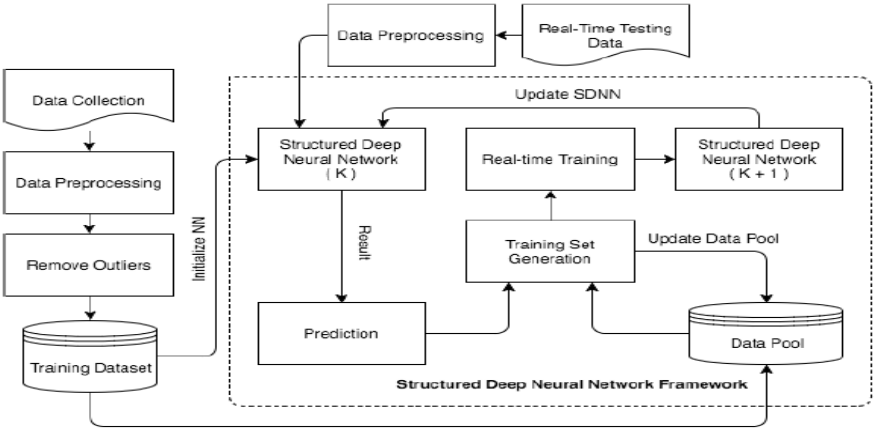

Fig 1: A framework for training a DNN in real time

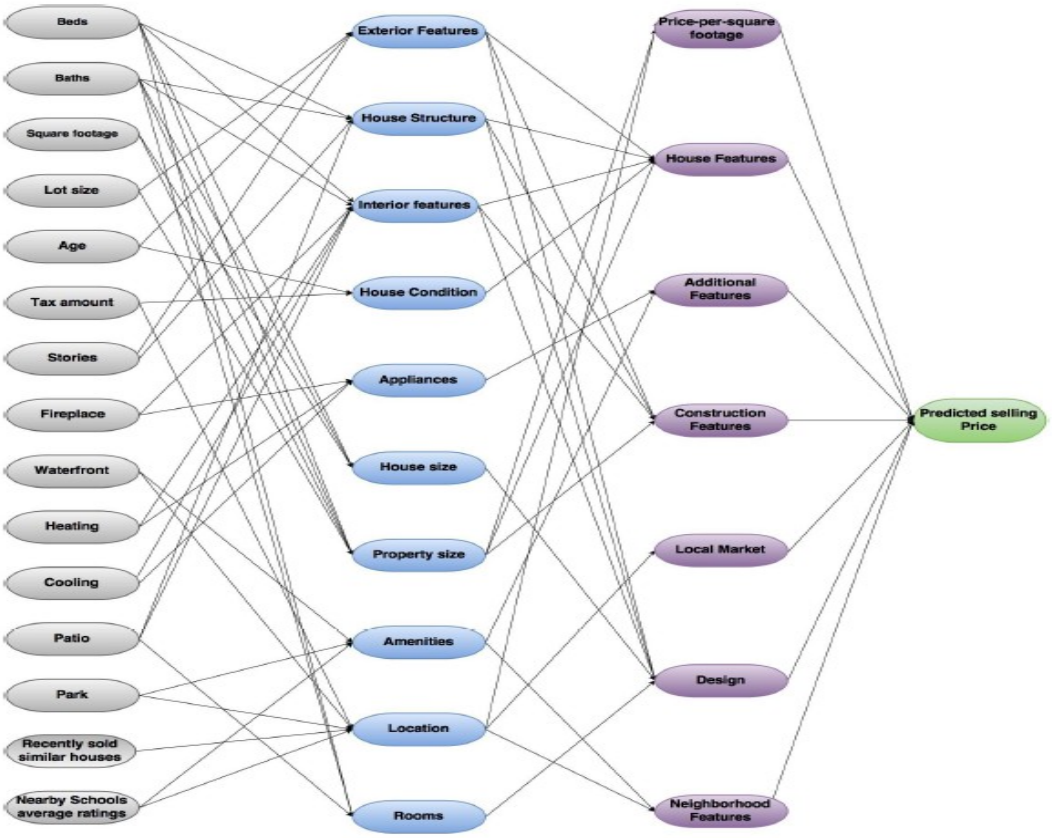

Fig 2: An example of layered knowledge graph for the real estate domain

Step 1: Data collection and dataset preparation

The data set can be taken from House sales in King County, USA and downloaded from www.kaggle.com.

Step 2: Developing a recommender system based on predictions

A structured DNN will be trained for predictive analytics. Also, feature design and selection

will be done using various machine learning approaches like Linear Regression, SVM, Random Forest, k-Nearest Neighbors (kNN) etc.

Step 3: Training and experimentation on datasets

The predictor model will be trained for the chosen dataset.

Step 4: Deployment and analysis on real life scenario

The trained and tested predictor system will be developed in real-life scenario.

Experimental Design

Dataset

The dataset House sales in King County, USA, available at https://www.kaggle.com/harlfoxem/housesalesprediction will be used for experimentation.

Evaluation measures

Measures such as accuracy will be computed by comparing the prediction and actual values for the pricing of houses.

Software and Hardware Requirements

Deep learning libraries will be exploited for the development and experimentation of the project. Training will be conducted on NVIDIA GPUs for training the DNN model.