House Recommendation System: Should I buy this or not!

Team: Excellors (Aditya Gollapalli and Vaishnavi Vijayashankar)

A. Problem Statement:

As humans, we all have a sense of belonging, a place we all call "home". It is an essential part of our daily lives, and something we cannot do without. Every individual desire to own their home(s). But are they making the right decisions with their purchases? What should they consider before buying? What are the criteria we should look for? This model helps users decide on whether to buy this particular property.

B. Data Sources:

The data has been outsourced from Open Intro .org (Kaggle House price Competition):

The links are as follows:

1. The link is the what the project will be based on:

https://www.openintro.org/data/index.php?data=ames

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Iowa Crime Rates and Statistics - NeighborhoodScout

The dataset consists of 82 exploratory data attributes, that pertain to the specifics of a house respectively.

The dataset provided here is mostly uncleaned and the observation made here is that is raw data collected from numerous surveys of the Iowa state.

The data is pre-processed as follows and named respectively:

- ***Raw_data.csv:*** The raw data previously found.
- ***Processed_data.csv***: The data is currently being stored in this name, which is also currently being processed from the EDA process.
- ***New_data.csv:*** The data after processing, cleaning will be stored here.

C. Methodology:

The proposal on this project is as follows:

1. **Exploratory Data Analysis on the Existing data available:** The main goal of this exploration is to find the existing trends in terms of people's purchases, their liking, interest in the type of house they desire. This is to quite understand the market fluctuations. This is also to understand the way people choose to liven up their home. The exact specifications that were highly sold, or the price point at which it was sold.
2. **Select the appropriate attributes:** Attribute Selection would be based of some of the chosen criterion that satisfies the home-owner or renter such as Nearest School Availability, Crime rate of thar neighbourhood, Sale Price, Nearby facilities available (community park, gym, pool, libraries etc). Some of the EDA we have done is shown below with respect to the available prospects derived from both external sources and the dataset itself.

3. **Classification Application:** As the user fills the response, we need to create a **Classifier Model** for analysing whether this house is worthy or not.

4. **Algorithms/ Solution Technologies:**

   The algorithm that we have thought of is to use Naïve Bayes Classification and Decision Trees. We also want to compare and contrast with other algorithms that make use of categorical variables and textual data.

   Sometimes, we have noticed that data is not completely utilized, to compensate we might add on a boosting algorithm.

5. **Risks:**
   The risks that we could face here are namely the misuse or misinterpretation of attributes that have similar meanings but have different interpretation such as:
   We have "Condition1" and "Condition2" , they both refer to the condition of the house, but they differ in the locality of their mention, as one is internal and the other is external condition respectively.

6. **Challenges:**

   The dataset is diverse and lots of choices may not be accurate, or the pattern generated from the analysis could be vague in terms of general understanding.

   With the existing attributes, questionaries can be made for the customers and build a classifier model depending upon their response in the questionary. However, there are more attributes, so expectations might differ from customer choice. We need to analyse what might be the customer's attribute choice.

7. **Citations:**

   Supervised learning — scikit-learn 1.1.3 documentation

   1. Supervised learning — scikit-learn 1.1.3 documentation

   1.9. Naive Bayes — scikit-learn 1.1.3 documentation

   1.9. Naive Bayes — scikit-learn 1.1.3 documentation

   The idea for usage of *crime rate* and also *neighbourhood scanning* was given by: ***Professor Wan Bae***.

8. **Project Link:**

   Github: Aditya-Gollapalli/ITDS-Project-Team-Excellors (github.com)

Group Dynamics:

1. General communication: We are synchronizing via WhatsApp chat and Microsoft teams to share ideas and documentations, Outlook to communicate with the professor and Canvas portal to share important links and documents for collaboration.

2. Sharing data and code: GitHub repository is used for sharing and uploading the code and datasets

3. Periodic Meeting: We are meeting every alternate day to discuss what has been done, what we'll be doing next and whether are there any roadblocks.

Follow up:

Our next step is to further the EDA process and also to rediscover new things from the dataset, and also building the model such that we can proceed to the classification step, to yield results.