Name: Khoa Truong

Project name: Predict housing prices in DFW area.

UTA ID: 1001626465

1. **Introduction:**

The housing market has always been an interesting topic and a lot of financial professionals want to predict the price so they can have an advantage in the market. Moreover, it’s not an easy task since there are so many factors involved in these predictions. However, a lot of predictions can be inaccurate because of human biases. Therefore, I wonder if I can create an algorithm that can sufficiently predict housing prices as unbiased as possible. In this white paper, I will mention on these topics: How I will collect the data and clean it, EDA, modeling, and evaluation.

1. **Dataset:**

To have an accurate algorithm, I want the real and up-to-date house data. So, I get information from the current single-family house sales on Zillow. To be able to do so, I use an API called ‘Zillow.com’ on the website rapidapi.com. This dataset contains a lot of information of a house. For instances, the number of bedrooms and bathrooms of a house, the living area in square feet, the latitude and longitude of the house, the school district the house located in, etc. Overall, there are more than 20 variables in this dataset and more than 10000 houses (samples) that we can analyze.

1. **Data Preparation:**

I will remove any samples that have null values in the living Area Value variable and housing prices. The reason that I decide not to fill these variables is that it can make my model poorly generalize. Not only that, but I will also look into abnormal datapoint. For instance, I notice that there is a house that is only around 700 square feet with 2 bedrooms and 1 bathroom, but the price is $900K. Most likely, this could be a data entry error and it should be 7000 square feet would be instead. However, this could be that some of these samples are not actually houses and they are farms/barns. We have the same approach towards other variables in the dataset too. Overall, I would rather have a good understanding of the variables that I am using than just filling the null values out.

1. **Exploratory Data Analysis (EDA):**

After the data is cleaned, I would perform some analysis on the variables. I would look at the distribution of continuous variables and look for the best normalization approach for these variables before feeding them into ML models. I will also analyze categorical variables such as the frequency of each category and if the categorical variables are correlated with one another. It’s important to check for multicollinearity such as the number of spaces in a garage is highly correlated with the number of spaces in a variable called covered. Graphs are actually very useful for visualization. For instance, a comparison of price distribution between different categories of categorical variable can help us see if that category has a heavy effect on the housing price distribution.

1. **Modeling:**

After picking the variables, it’s time to create a model that predict housing prices. I would start with traditional ML models such as KNN and Random Forest. Then, I would create a Deep Learning model to see which would perform better. These models will use the characteristics of these properties to predict. Moreover, while building the best model, we will learn which variables get the models to perform great and which hold the model back. Various modeling techniques will be used to achieve the best result.

1. **Evaluation:**

Finally, we will evaluate these models with R2 and adjusted R2. Of course, a higher R2 means a better model.