Project name: Predict housing prices in DFW area.

**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. This model should end up estimating housing prices better than an average person. Moreover, the result of this model can last at least a quarter to a year with small deviations since housing prices don’t fluctuate much.

**Dataset:**

1. **Data Collection:**

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:**

Datapoints that are not allowed in the dataset:

+ Any samples that have null values are either filled with 0 or removed.

+ For Categorical variables, use one hot encoding but if only 5% or below only have that specific categorical variable, drop it, or fill the value in ‘Other’ column.

+ Create a new variable called liv/lot\_ratio. If liv/lot\_ratio > 2, removed those houses since they are likely apartments or townhouses.

+ Remove houses that have price per lot Area greater than $1000 or with lot Area greater than 100000 sqft.

+ Finally, I create 3 different targets to predict: price/lotsqft, living\_price, and the actual price of the house.

+ Living price = liv/lot\_ratio \* price of the house.

+ Perform EDA on the dataset.

**Exploratory Data Analysis (EDA):**

**Chart, scatter chart

Description automatically generated**

**Scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

Based on the map, locations have significant effects on these variables, especially the target variable price/lotsqft.

**Chart, scatter chart

Description automatically generated**

Graphical user interface, application

Description automatically generated

The correlation between these 2 variables are high, which tells me that the liv/lot\_ratio will prove to be effective to predict the price/lotsqft.

Chart, scatter chart

Description automatically generated

This concept also applies when comparing with living\_price instead. As you can see, there are variables that correlate well with the price and not the living\_price. However, we need to have a separate model to predict the price of the inner part of the house itself since actual price also includes the exterior of the house.

**Modeling:**

1. **Model Selection and Training:**

Firstly, we apply cross 5-Fold validation to make sure that every single house or sample is tested. Then, we scaled all the features with MinMaxScaler() into the range 0-1 and scaled our targets with natural log. I choose ANN since it has shown to perform better when predicting price/lot\_sqft and living\_price. For more preferences, please check the folder progress\_presentation. So, I create 3 different ANN models for different purposes:

+ ANN\_location: predict price/lotsqft based on location.

+ ANN\_properties: predict price of the inner part of the house.

+ ANN\_price: predict the price of the house based on other features and the results of the first 2 models.

1. **Model Evaluation and Results:**

**Chart, line chart

Description automatically generated**

**Table

Description automatically generated**

With the combination of the other 2 models’ outputs and other features we achieved a really high R-squared for the test sets (88.15% to 91.49%).

Chart, box and whisker chart

Description automatically generated

Since our housing prices vary, the errors also vary too. The MAPE for each month is almost at 10%. However, the median APE is usually around 6.2% to 7.8%.

Chart, scatter chart

Description automatically generated

The model performs well for houses with prices ranging from $250K to $1000K. As we go above that price, the housing prices start to deviate a lot more from the actual results. Therefore, the model’s strength is that it can predict the price of ‘average’ houses but not extremely luxurious houses.

We can even improve the performance of the model if we collect older data with respect to the latest price that is posted on Zillow.