Effects of housing characteristics on sale price

Team Members: Harikrishnan Santhosh and Prachi Jhamb

The main dataset used in our analysis is derived from the Kaggle Housing Prices competition (https://www.kaggle.com/competitions/home-data-for-ml-course/data), which is a machine learning based competition that uses 79 explanatory variables describing characteristics of residential homes in Ames, Iowa to predict the sale price of the houses. The reason why we chose this topic and dataset is because it gave us a good opportunity to understand and apply principles of machine learning in our analysis. Each of the project goals are described below:

1) Design a Web Server running a Web API service that accepts data from distributed sources (e.g. mobile phones, computers, sensors) into a persistent datastore

For designing the Web API we first load the housing data (csv file) from Kaggle on to a sqlite database. We then use FastAPI to write the python file that executes the program on a web server. This can be found under **p1/main.py** in the GitHub repository. We run various GET functions to return all readings in the database, return readings for a specific street type, for a specific year built, for average sale price for a specific street type, and the minimum and maximum sale price for a specific street type. After testing the Web API on the local server we push it on to the oracle cloud server so that it remains a persistent datastore. This can be tested at https://hariksan.ga/myapp/docs or the entire readings from the dataset can be found at https://hariksan.ga/myapp/readings.

2) Use an existing Web API to integrate relevant, publicly accessible, real-time data (e.g. a Weather API)

Since we are using housing data for our analysis, we did not have access to real-time data that was both free and accessible. Therefore, we get housing data on Iowa state from the Data USA API (https://datausa.io/about/api/). We use five different API calls to get relevant data for our project. These include data on housing property values, household income, property taxes, home ownership and average commute time for Iowa vs other states or the whole of U.S. We then convert the json file to dataframes and then extract the relevant variables and the most recent data available for each.

3) Use one or more computational notebooks and/or scripts to prepare data for analysis/sharing, including exploratory visualizations of data

We had a dataset called train data which had 79 explanatory variables describing characteristics of residential homes in Ames, Iowa. We first identify the variables with numerical data and categorical data.

To begin cleaning the data and preparing it for analysis, we first look at missing values in the data. We find that of the 79 variables, 19 of them had missing values in them. To look further, we see the percentage of missing values in each of those 19 columns.

Post that, we use the following strategy to deal with missing values:

- For the variables with categorical data: We replace the missing values with the imputed mode of the data.
- For the variables with numerical data: We replace the missing values with the imputed mean of the data.
- Variables with too many missing values: Some variables have so many missing values that we decided to drop them off completely unless the variable was of significance.

We again check if we have taken care of all NA values from the dataset. To visualize the relationship between the house characteristics and Sale Price. We use box plots for categorical variables and scatter plots for numerical variables. We find a couple of interesting relationships from these raw plots. For instance, we see that houses with access to a gravel street have a lower Sale Price as compared to houses with paved street. Similarly, houses which have Central air conditioning have a higher sale price as compared to houses that don't. Houses with a foundation of wood as well as houses with foundation of Poured Contrete have higher sale prices as compared to houses with foundation of slab, stone or brick and tile.

From the numerical data scatter plots too, we find a couple of interesting relationships. For instance, there is a positive linear relationship between the Linear feet of street connected to property (LotFrontage) and the sale price of the property. There is also a positive relationship between -Above grade (ground) living area square feet (GrLivArea) and sale price of a property.

4) Use one or more computational notebooks to analyze data to some valuable end, including exploratory and explanatory visualizations, models built from the data through machine learning algorithms, and evaluations of model performance (where applicable).

The objective here is to predict the Sale price of each house based on the house characteristics. For each Id in the test dataset, we must predict a Sale Price corresponding to it.

We have two datasets, the training data, and the testing data. The train data has one additional column than the test data – Sale Price, which is what we must predict. Using the train data, we first check the correlations of each variable with the target variable (Sale Price). We choose those house characteristics/variables which have a high correlation with the target variable and define these characteristics as "main_features". Using the train data, we create a new data frame(X) which only has these "main_features" except the sale price which we store in another data frame (y).

We then partition the X and y data frames into training and testing datasets using test_size = 0.5 and random_state =8000. Using the test dataset, we also create x_test dataset which only the features we will use for our prediction. We check for missing values in this dataset and since the columns with missing values were all numeric, we replace the NA values with the imputed mean. We run three regression models- Linear Model, Random Forest and Decision tree regressor and check the accuracy of prediction for each of the models. Since the prediction accuracy for Random Forest (model_2) was the highest, we chose it as our final model to predict Sale Price.