Logistic Learning with Housing Features

1st Neil Patrick Reyes *nlreyes@cpp.edu*
2nd Drake Fafard *djfafard@cpp.edu*
3rd Jason Ryan Jones *jasonjones@cpp.edu*
4th Andrew Sanford *asanford@cpp.edu*
5th Vu Nguyen *vunguyen@cpp.edu*

***Abstract*—In this project, we will employ linear regression to predict housing prices. We will leverage a data-set of housing attributes, including location, size, bedrooms, and amenities, and preprocess the data, engineer features, and train a linear regression model. Our objective will be to provide an accurate and interpretive model for housing price prediction while gaining insights into the key factors influencing prices. This study will underscore the utility of linear regression as a foundational machine learning technique in real estate applications, offering transparency and informed decision-making in the housing market.**

***Index Terms*—Regression model**

I. Introduction

In the dynamic world of real estate, accurate housing price prediction is essential for both buyers and sellers. This project explores the application of linear regression, a foundational machine learning technique, to predict housing prices. By leveraging a comprehensive dataset containing property attributes such as location, size, and amenities, we aim to provide a reliable tool for forecasting housing prices. The project encompasses data collection, preprocessing, feature engineering, model training, and evaluation, with the dual objectives of delivering accurate predictions and gaining insights into the key determinants of housing prices. Through this work, we highlight the enduring relevance of linear regression in real estate analytic, offering transparency and valuable decision-making support to stakeholders in the housing market.

II. Methodology

There are many different algorithmic variations that can be used to tackle a dataset in order to accurately predict a house's price. However, a major factor that makes (Multiple) Linear Regression a clear winner is its ability to take key features into account and present data as a unit instead of classifying it into pre-existing points, like Naive Bayes.

A. Initial Preprocessing

Initially, the data will need to be split between its dimensions and classes (features vs sale price). Following this, a reduction in data will be required in order to sift out the unnecessary data.

The first step taken, after uploading the data, is to see if there are any duplicated within the training and set data; where if any were to be found, they would be removed so they would not be included twice and affect the output.

After this step is to visualize the data with a histogram. A histogram is an illustration that depicts the range of values each feature has. With this being said, any non-numerical values can be mapped into a respective data point.

When all the features as numerical values, it is safe to create a heatmap in order to determine how the features interact with each other. From this, one can determine which is a direct correlation towards the SalePrice (Y) value that is to be predicted. Drop all the features that do not positively affect the SalePrice because it is important to keep features that increase and result in an increase in the SalePrice.

With all this being said, it is now safe to train the data points in order to get a training model that can be tested and give a proper result.

B. Training

Following the preprocessing comes the experimentation. In this portion of the project, we took the processed data and divided the data into 10 equal sections. This in turn will lead to us experimenting with the training data in order to utilize the *Leave One Out Method*. Iterating through the sections allows us to limit the possibility of overfitting the training data when it is utilized in the test model.

In succession of this, of the training sets, we will conduct a follow-up experiment using the best accurate set from the *Leave One Out Training* and use it to experiment with the test data provided; whether that be the general model or one of the iterations within the *Leave One Out* training models.

C. Testing

Following this steps have led to a prediction that the test results will be accurate upon comparison to the *Kaggle.com* test results. However if the results of the initial tests are unsatisfactory, then there will be another iteration of the initial preprocessing and training phases until a satsifactory result has been met.

III. Dataset

The data set being used in this project is provided and tested by *Kaggle.com*. Within this dataset (partial dataset shown labeled as Table I), key features like YearBuild, OverallQual, YrSold are presented for the purpose of training. Machine learning algorithms are used to predict and accurately predict a House's Sale Price based on the key features presented.

Since this is a visual snippet of the data provided by *Kaggle.com*, it cannot be emphasized how

HouseStyle	OverallQual	YearBuilt	BedroomAbvGr	FullBath	YrSold	GrLivArea	SalePrice
2Story	7	2003	3	2	2008	1710	208500
1Story	6	1976	3	2	2007	1262	181500
2Story	7	2001	3	2	2008	1786	223500
2Story	7	1915	3	1	2006	1717	140000
2Story	8	2000	4	2	2008	2198	250000
1.5Fin	5	1993	1	1	2009	1362	143000
1Story	8	2004	3	2	2007	1694	307000
2Story	7	1973	3	2	2009	2090	200000
1.5Fin	7	1931	2	2	2008	1774	129900

TABLE I
PARTIAL HOUSING DATASET

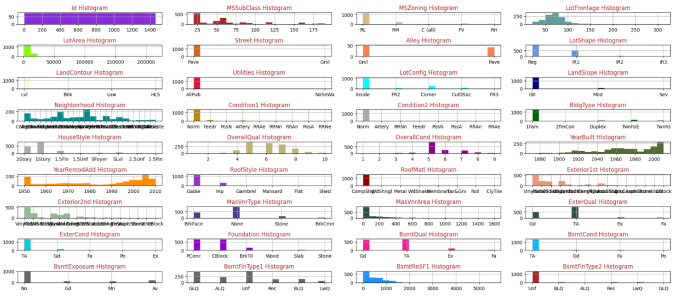

Fig. 1. Histogram

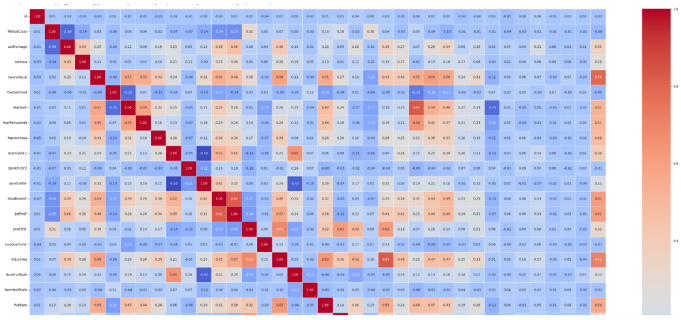

Fig. 2. Heatmap

IV. RESULTS

Results following this experiment are TBA.

V. RELATED WORK

Related work following this experiment are TBA.

VI. CONCLUSION

Conclusion following this experiment are TBA.

REFERENCES

[1] Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques

[2] J. Dunlea, "ML-driven house price prediction in 2023," Akkio, https://www.akkio.com/post/house-price-prediction-using-machine-learning (accessed Dec. 10, 2023).