Project Title: Predicting Home Prices and Related Driving Factors

Team Member Names: Binh Vu

Problem Statement:

For new home buyers, the process of purchasing a house and putting it in a bid could seem daunting due to the price tag. The determination of the appropriate worth of a home could mean the difference between thousands of savings or extra costs for the buyer. Additionally, knowing the approximate worth of a potential house could give the buyer the confidence to bid for a 'dream house' that might seem outside of range otherwise. In this paper, a dataset containing the house's physical characteristics (num bed, num bath, lot size, etc.), the external environment (street, alley, neighborhood, etc.), and the transaction information (date sold, sale type, sale condition, etc.) will be used. The goal is to predict the worth of house-based similar features and identify which characteristics are the most impactful on home price.

In summary, this paper has two main tasks to accomplish:

1. What are the features that have the most weight in determining the worth of a house?
2. For a conventional buyer, can the value of a house be predicted to be at an acceptable level of accuracy?

Data Source:

This paper uses the 'House Prices – Advanced Regression Techniques' dataset from Kaggle. This dataset has approximately 2920 data points with 80 features describing every aspect of homes in Ames, Iowa. The data types are mixed between numerical values, string classifications, binary classification, and dates. The data pre-separated 50/50 between labeled data and non-labeled data.

Methodology:

Data Filtering

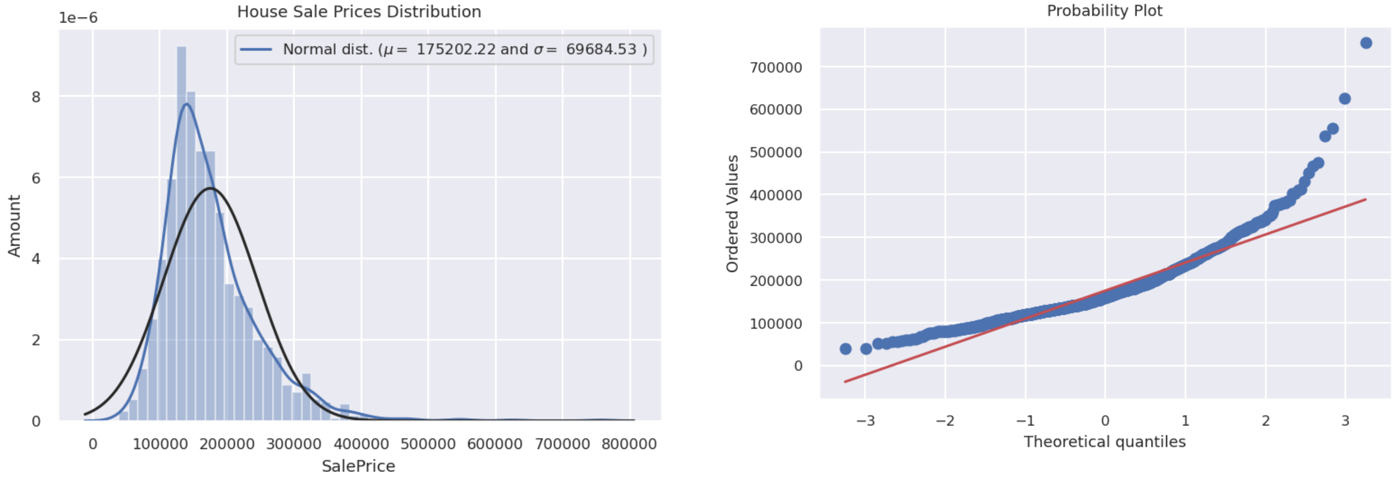
In the data, various sale conditions which include Normal Sale, Abnormal Sale, Adjoining Land Purchase, Two Lined Properties with Separate Deeds, Sale Between Family Members, and Partial home was not completed when the last assessed. Since this paper is interested in the sale between conventional buyers, only data with Normal Sales a kept for further analysis.

Initial Inspection

Using a histogram plot, an initial visual inspection of the sale price is used to determine the data distribution. In figure 1, two plots are presented. The left plot is a histogram plot of the sale price with the KDE/smoothing line (Gaussian kernel density estimate) in blue and a normal distribution line in black. The right is a QQ plot which is an intuitive way to visualize whether something is normally distributed.

Ideally, the data distribution should resemble a normal distribution curve. However, as expected of real-world data, the original Sale Price demonstrates a right-skewed or positive skewed distribution. This characterization is determined by comparing the KDE line with the normal distribution line which does not line up and the QQ plot lower/upper quantiles curving upward. There are multiple ways of dealing with data skewing, some of which are log transform, square root transform, and box-cox transform. This paper will use log transform to deal with skewed data.

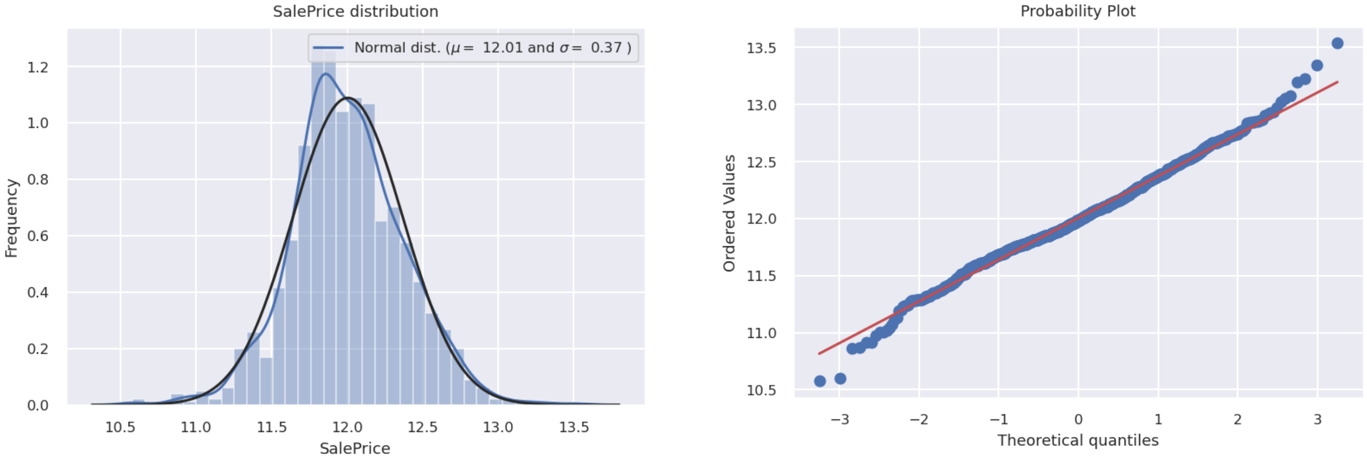
Figure 1: Original Sale Price Distribution and QQ plot



The resulting transformation of data is better represented as a normal distribution when compared to the original sale price distribution (fig 1). The KDE line closely resembles the normal distribution line and the QQ plot's head and tail is closer to the 45* line.

After adjusting for the skewness of the data, the resulting Sale Price distribution is as follows.

Figure 2: Sale Price Distribution and QQ Plot with Fixed for Skewness



Data Preprocessing

The dataset has multiple columns that have a missing blank value. In table 1, the attribute names, the amount missing, and percent missing are presented. From the data, the top 20 missing data are shown. Since the missing data types are mixed, the first task will be backfilling categorical values with 'none' and numerical/binary values with mode or mean depending on the context. This will allow us to use the full data point in machine learning (ML) model later since many algorithms do not support missing values.

Table 1: Top 20 Missing Data Counts

Attribute	Amount Missing	Percent Missing
PoolQC	1195	99.749583
MiscFeature	1148	95.826377
Alley	1127	94.073456
Fence	957	79.883139
FireplaceQu	563	46.994992
LotFrontage	237	19.782972
GarageYrBlt	61	5.091820
GarageCond	61	5.091820
GarageType	61	5.091820
GarageFinish	61	5.091820
GarageQual	61	5.091820
BsmtExposure	33	2.754591
BsmtFinType2	33	2.754591
BsmtCond	32	2.671119
BsmtQual	32	2.671119
BsmtFinType1	32	2.671119
MasVnrArea	4	0.333890
MasVnrType	4	0.333890
Electrical	1	0.083472
MSSubClass	0	0.000000

After backfilling missing data, each feature is inspected for consistency based on the provided document. That means, if 'Street' features should only have gravel/paved, then that feature should not include any other input. Otherwise, that data point should be removed. In this paper, all data points are consistence with the given data description given at the source.

Then, we can start removing unwanted features that will not contribute to the final price. Columns such as 'Id' could be removed because it does not have a correlated relationship with the other data point since it is a unique identifier. Fewer features mean less training time and less chance of overfitting. Then, the full dataset is checked for skewness and fixed using a similar approach given above.

Similarly, a correlation matrix would be used to determine highly correlated features. When two independent variables are highly correlated, this results in a problem known as multicollinearity, which can lead to skewed or misleading results. This paper recognized the highly correlated features but does not remove them. Removing features such as Overall Quality or Grid Living Area would be counterintuitive to predicting the final Sale Price. If the latter resulting price prediction is negatively affected by the inclusion of the highly correlated data, an attempt will be made to fix it.

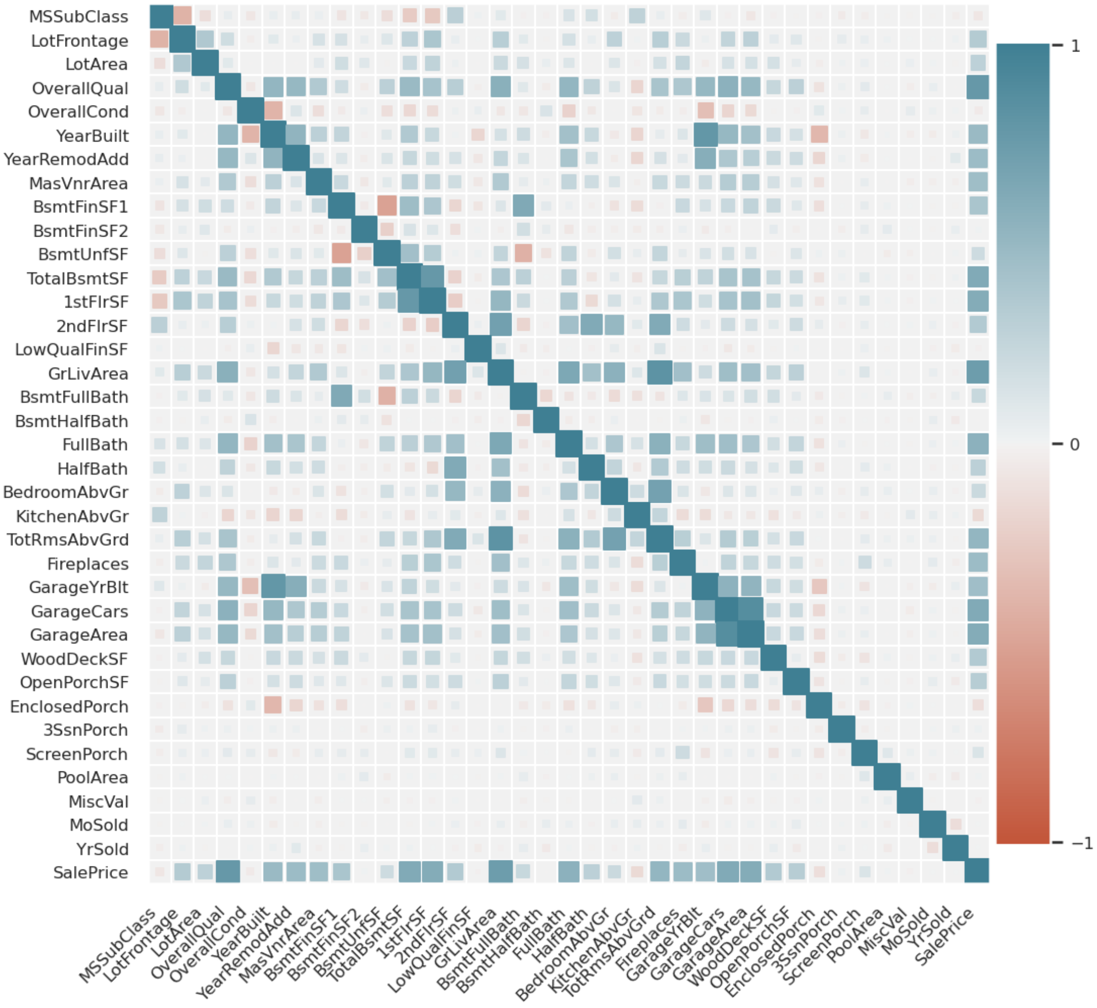


Table 2: Top 5 Highly Correlated Features to Sale Price	
OverallQual	0.786850
GrLivArea	0.744517
TotalBsmtSF	0.628819
GarageCars	0.626678
1stFlrSF	0.610918

Data Transformation

After the data is at an acceptable level, the next step is to use one-hot encoding to convert categorical to a numerical values. Which will allow ML algorithms to do a better job in prediction. Features such as 'Alley' with {Grvl:Gravel, Pave:Paved} are converted into Alley_grvl and Alley_pave with binary values for existing or not existing.

By the end, there are remains 1198 data points from 1460 originally and 210 features compare to 80 originally. The clean labeled data is then split into 70% training set and test set. This paper did not split the data into validation sets because of the number of data points compared to the number of features.

Linear Regression, Lasso Regression, and Random Forest Regression

The training dataset is used to train three machine learning models: linear regression, lasso regression, and random forest regression. After the models are trained, the test dataset is used to evaluate the trained model to unknown data.

For lasso regression, multiple alphas [0.0005,0.001,0.01,0.1,1] values are tested and tested against the training dataset. The best resulting alpha value is 0.0005. This is the alpha value that is used to train the model.

Evaluation and Final Results:

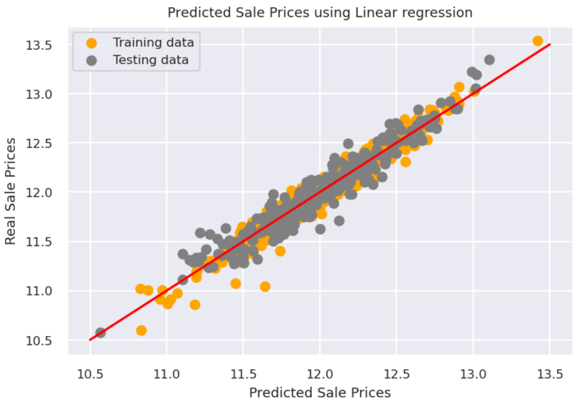
Evaluation

To evaluate the three models, there will be multiple factors to consider:

1. The Mean Absolute Error (MAE) average of the absolute differences between the actual value and the model's predicted value
2. The Root Mean Squared Error (RMSE) average of the squared differences between the actual and the predicted values.
3. R^2 score: it measures the proportion of variance of the dependent variable explained by the independent variable.

Results:

Most Impactful Feature(s): The overall quality of the house and Above grade ground living area square feet. This result is determined through a correlation map. Intuitively, this checks out. How good the house and how big the living area is has a big impact on the price of the house.



This left figure represents a comparison between a true value and a predicted value. This figure gives a visual validation of the result. With the red line representing perfect prediction, the point that is closer to the line is closer to the true value.

The resulting evaluation scores are provided below. Table 3 represents the evaluating metric for training data and Table 4 represents the evaluating metric for testing data. Using Linear Regression as a baseline model, the other two models are compared with their results.

The linear regression model's MAE, RMSE, and R-Squared all decreased from training data to testing data which is expected because the model is trained using a different set of data. When compared to linear regression, lasso regression did better in testing data for all metrics. This is also expected because it enhances regular linear regression by slightly changing its cost function, which results in less overfit models. Random forest did extremely well in training data but much worse in testing data which suggests overfitting. Overall, the lasso regression model seems to predict house sale prices best out of the three models.

Table 3: Evaluation Metric of Training Data

	MAE	RMSE	R-Squared
Linear Regression	0.054833	0.0746024	0.956707
Lasso Regression	0.063612	0.0854371	0.943219
Random Forest Regressor	0.035697	0.0517686	0.979153

Table 4: Evaluation Metric of Testing Data

	MAE	RMSE	R-Squared
Linear Regression	0.074973	0.1028687	0.9284560
Lasso Regression	0.074680	0.0992660	0.9333795
Random Forest Regressor	0.097337	0.1369508	0.8731951

The resulting MAE, RMSE, and R-Squared values show that the sale price of a house can accurately be determined.

Conclusion:
In summary, this paper attempts to identify important features that determine house sale prices given various attributes of homes in Ames, Iowa. After filtering, cleaning, and transforming the data, three regression models are used to predict the home sale prices. From the analysis above, the important features are identified, and the conclusion is that a house price can be predicted at an acceptable level of accuracy.