ST 595
Fall 2020
10/11/20

Data Science Project 1

Advanced Regression Machine Learning Model in Python

Prepared By:

Josh Olsen

Table of Contents

INTRODUCTION

Being able to accurately predict the price of home is extremely valuable for both Sellers and Buyers.
For Sellers, homeowners and real estate agents can obtain the optimal price of a home based on features of the house such as square footage, number of bedrooms, number of bathrooms, garage size, location, etc.
Buyers can identify if a home is a favorable investment. For example, is a buyer getting a good deal by paying below market value of the property.

Both parties can benefit from knowing which variables of a house most positively and negatively affect the price from a home.
Sellers have an opportunity to address which areas of the house before putting it on the market. For Buyers looking for in investment property, there is value knowing which variables of the house negatively affect the price of the house. The buyer could than make the necessary repairs prior to reselling the house to obtain the optimal price.

Question of Interests

Explanatory Question

1. Which explanatory variables are highly correlated with the Sale Price (Response) variable?

Statistical Questions

2. Out of the Machine Learning Model, which explanatory variables most positively and negatively affect the price of the house (response variable)?

3. How accurate can a Machine Learning model predict the home price based on a series of explanatory variables?

REVIEW OF THE DATASET

The sample dataset is from residential home in Ames, Iowa. This sample is taken from the all home sales in the united states population.

Description of dataset

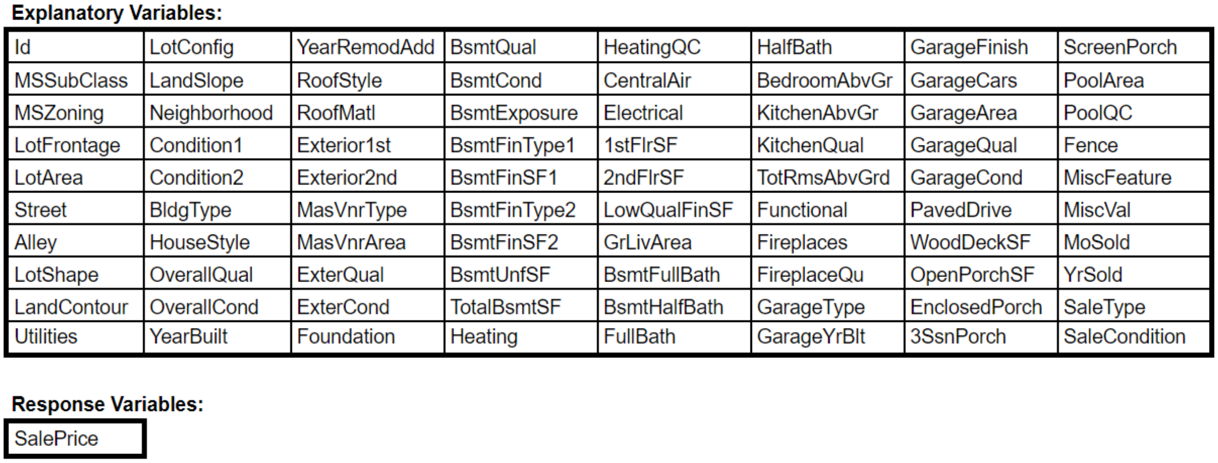
There are two datasets one to train the model and one to test the model.
There are 80 explanatory variables describing almost every aspect of residential homes. The explanatory variables and are the same in the training and test datasets.

Training Dataset

- 1,460 unique home with 80 explanatory variables that have a sale price (response variable)

Testing Dataset

- 1,458 unique homes with 80 explanatory variables that do not have a sale price (response variable)



Explanatory Variables:

Id	LotConfig	YearRemodAdd	BsmtQual	HeatingQC	HalfBath	GarageFinish	ScreenPorch
MSSubClass	LandSlope	RoofStyle	BsmtCond	CentralAir	BedroomAbvGr	GarageCars	PoolArea
MSZoning	Neighborhood	RoofMatl	BsmtExposure	Electrical	KitchenAbvGr	GarageArea	PoolQC
LotFrontage	Condition1	Exterior1st	BsmtFinType1	1stFlrSF	KitchenQual	GarageQual	Fence
LotArea	Condition2	Exterior2nd	BsmtFinSF1	2ndFlrSF	TotRmsAbvGrd	GarageCond	MiscFeature
Street	BldgType	MasVnrType	BsmtFinType2	LowQualFinSF	Functional	PavedDrive	MiscVal
Alley	HouseStyle	MasVnrArea	BsmtFinSF2	GrLivArea	Fireplaces	WoodDeckSF	MoSold
LotShape	OverallQual	ExterQual	BsmtUnfSF	BsmtFullBath	FireplaceQu	OpenPorchSF	YrSold
LandContour	OverallCond	ExterCond	TotalBsmtSF	BsmtHalfBath	GarageType	EnclosedPorch	SaleType
Utilities	YearBuilt	Foundation	Heating	FullBath	GarageYrBlt	3SsnPorch	SaleCondition

Response Variables:

SalePrice

Figure 1 - Explanatory and Response Variables

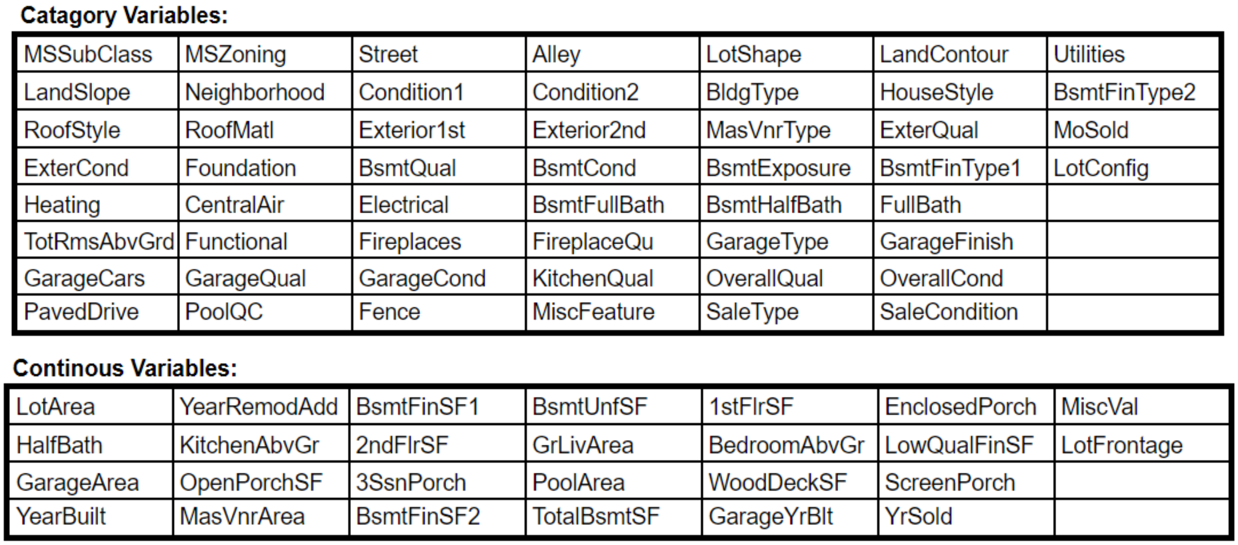
Figure 1 displays all the explanatory variables and the response variable in the dataset.

Different Types Explanatory Variables

Since all the two different types of explanatory variables, all the explanatory variables need to be segmented into two different classifications.

Variables Types:

- Category Variables
- Continuous Variables



Catagory Variables:

MSSubClass	MSZoning	Street	Alley	LotShape	LandContour	Utilities
LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	BsmtFinType2
RoofStyle	RoofMatl	Exterior1st	Exterior2nd	MasVnrType	ExterQual	MoSold
ExterCond	Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	LotConfig
Heating	CentralAir	Electrical	BsmtFullBath	BsmtHalfBath	FullBath	
TotRmsAbvGrd	Functional	Fireplaces	FireplaceQu	GarageType	GarageFinish	
GarageCars	GarageQual	GarageCond	KitchenQual	OverallQual	OverallCond	
PavedDrive	PoolQC	Fence	MiscFeature	SaleType	SaleCondition	

Continous Variables:

LotArea	YearRemodAdd	BsmtFinSF1	BsmtUnfSF	1stFlrSF	EnclosedPorch	MiscVal
HalfBath	KitchenAbvGr	2ndFlrSF	GrLivArea	BedroomAbvGr	LowQualFinSF	LotFrontage
GarageArea	OpenPorchSF	3SsnPorch	PoolArea	WoodDeckSF	ScreenPorch	
YearBuilt	MasVnrArea	BsmtFinSF2	TotalBsmtSF	GarageYrBlt	YrSold	

Figure 2 - Category and Continuous Variables

OBSERVATIONS

Missing Data

Missing data can be problematic for statistical model. Below in **Figure 5** is a visual of missing data within the dataset. **Figure 6** displays the top 6 variables with missing data. The top 6 variables accounted for 90% of the missing values.

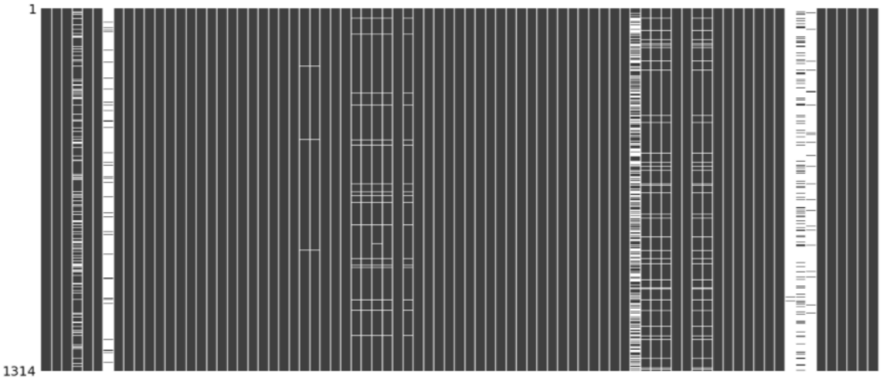
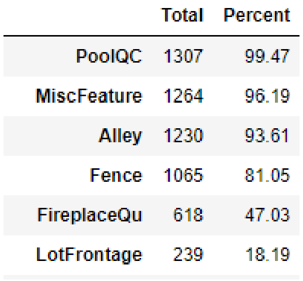


Figure 5 – Visual of Missing Data



	Total	Percent
PoolQC	1307	99.47
MiscFeature	1264	96.19
Alley	1230	93.61
Fence	1065	81.05
FireplaceQu	618	47.03
LotFrontage	239	18.19

Figure 6 – Top 6 Variables with Missing Data

There are significant amount of missing values in the dataset. This issue is addressed in the Feature Engineering #1 section.

Distribution of Sale Price (Response) Variable

Before any statically modeling, the distribution of the Sale Price (Response) Variable needs to be reviewed. The distribution of sales price against a normal distribution is displayed in **Figure 3**. The Q-Q plot of the quantile distribution is displayed in **Figure 4**.

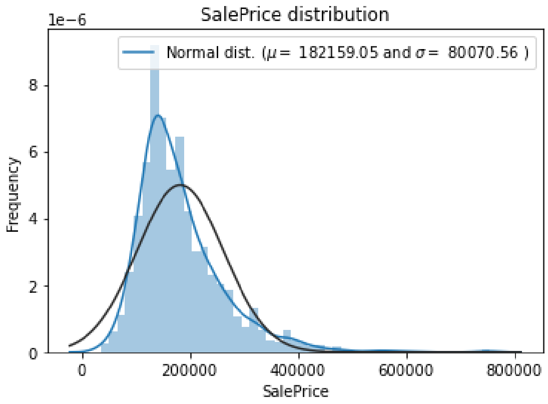


Figure 3 – Distribution of Sale Price (Response) Variable

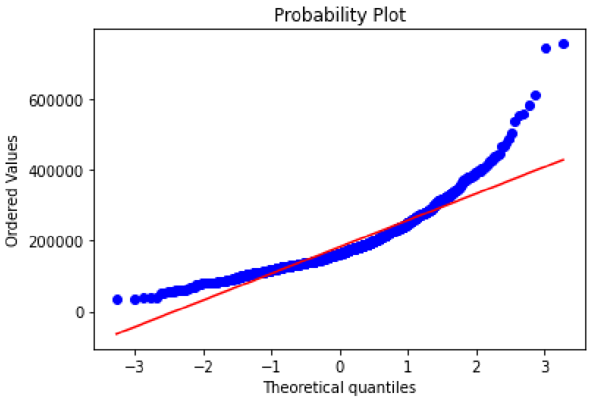


Figure 4 – Q-Q Plot of Sale Price (Response) Variable

After reviewing the two plots above, the Sale Price (Response) Variable is not normally distributed. The Sale Price (Response) Variable is left skewed. This issue is addressed in the Feature Engineering #2 section.

Correlation Analysis

To get a better understanding the Explanatory Variables to the Sale Price (Response) Variable, a correlation matrix was generated in **Figure 7**.

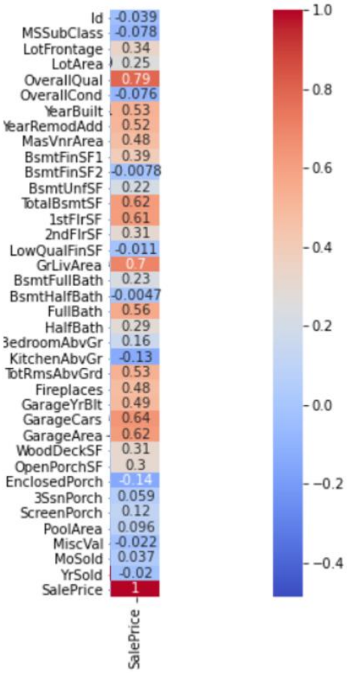


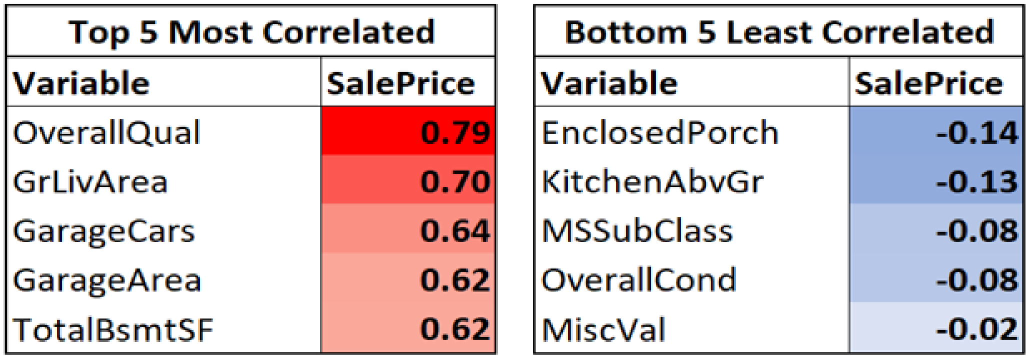
Figure 7 – Correlation Matrix of Sale Price (Response Variable) to Explanatory Variables

There are 6 Explanatory Variables that are greater than 0.60 while there are 4 Explanatory Variables that are below -0.03.

Explanatory Question of Interest

1. Which explanatory variables are highly correlated with the Sale Price (Response) variable?

Figure 8 displays the explanatory variables that are most and least correlated with the Sale Price (Response) Variable.



Top 5 Most Correlated	
Variable	**SalePrice**
OverallQual	**0.79**
GrLivArea	**0.70**
GarageCars	**0.64**
GarageArea	**0.62**
TotalBsmtSF	**0.62**

Bottom 5 Least Correlated	
Variable	**SalePrice**
EnclosedPorch	**-0.14**
KitchenAbvGr	**-0.13**
MSSubClass	**-0.08**
OverallCond	**-0.08**
MiscVal	**-0.02**

Figure 8 – Summary of Top 5 Explanatory Variables that are Most and Least Correlated to Sale Price (Response) Variable

FEATURE ENGINEERING

1. Addressing Missing Values

There are several category variables that has “Missing Values” that needed to be addressed. This was found in the **Alley, PoolQC, MiscFeature, Fence,** an**d FireplaceQu,** variable which would indicate that there is no Alleyway on the house or property. “NA’s” or “missing values” where replaced with “No Alley” for this variable.

There are six continuous variables with “Missing Values” that needed to be addressed. This was found in the **LotFrontage, MasVnrArea**, **BsmtFinSF1**, **BsmtFinSF2**, **BsmtUnfSF**, **TotalBsmtSF**, and **GarageArea**. This indicates that the home does not have any of these features. A zero value “0” was replaced for these variables with “Missing Values”.

After accounting for the “NA’s” and “Missing Values” Displayed below in **Figure 9** is a visual of missing data within the dataset.

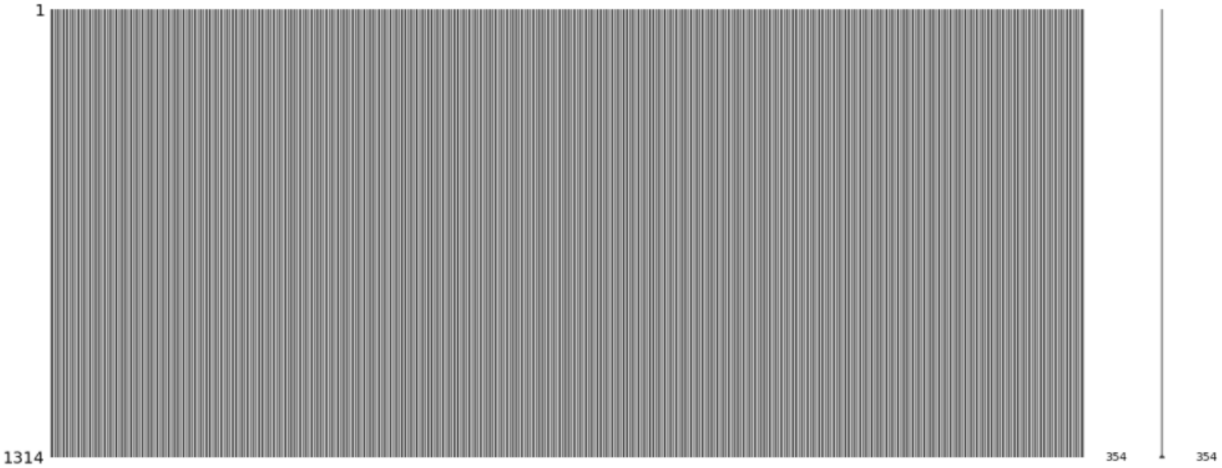
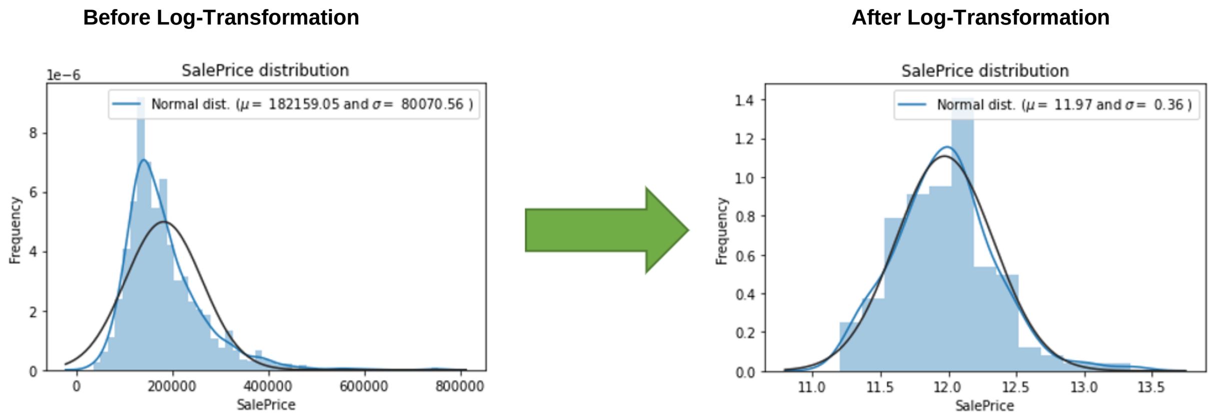


Figure 9 – Updated Visual of Missing Data

Accounting to **Figure 9**, there are no more missing values.

2. Addressing Non-Normally Distributed Sale Price (Response) Variable

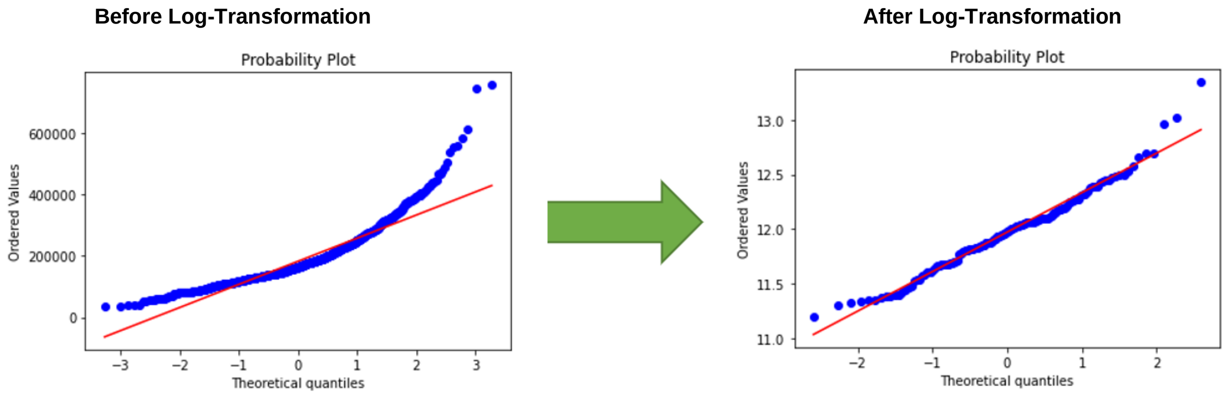
In order to account for non-normally distributed of the Sale Price (Response) Variable, a Log-Transformation of the Sales Price (Response) Variable was performed.
A before and after Log-Transformation of the Sale Price (Response) Variable is displayed below in **Figure 10 and 11**.



Before Log-Transformation

After Log-Transformation

Figure 10 – Distribution of Log-Transformation Sale Price (Response) Variable



Before Log-Transformation

After Log-Transformation

Figure 11 – Q-Q Plot of Log-Transformation Sale Price (Response) Variable

The Log-Transformation of the Sale Price fixes the non-normally distributed issue.
The Sales Price (Response) Variable is now normally distributed in the after-plot **Figure 10**.
Figure 11 displays before and after Log-Transformation Q-Q plot of the quantile distribution
The after plot in **Figure 11** displays that the normal distribution is symmetric. This means that Sale Price (Response) Variable displays significantly less skewness. There are a few dots that do not follow the straight line. This is probably due to a few outliers in the dataset.

STATISTICAL MODELING

Experimental Design

There are two main components of design a machine learning model. First is training the model and the second is testing the model.

Below in Figure 3 is a basic high level visual of the experimental design for training and testing a machine learning model.

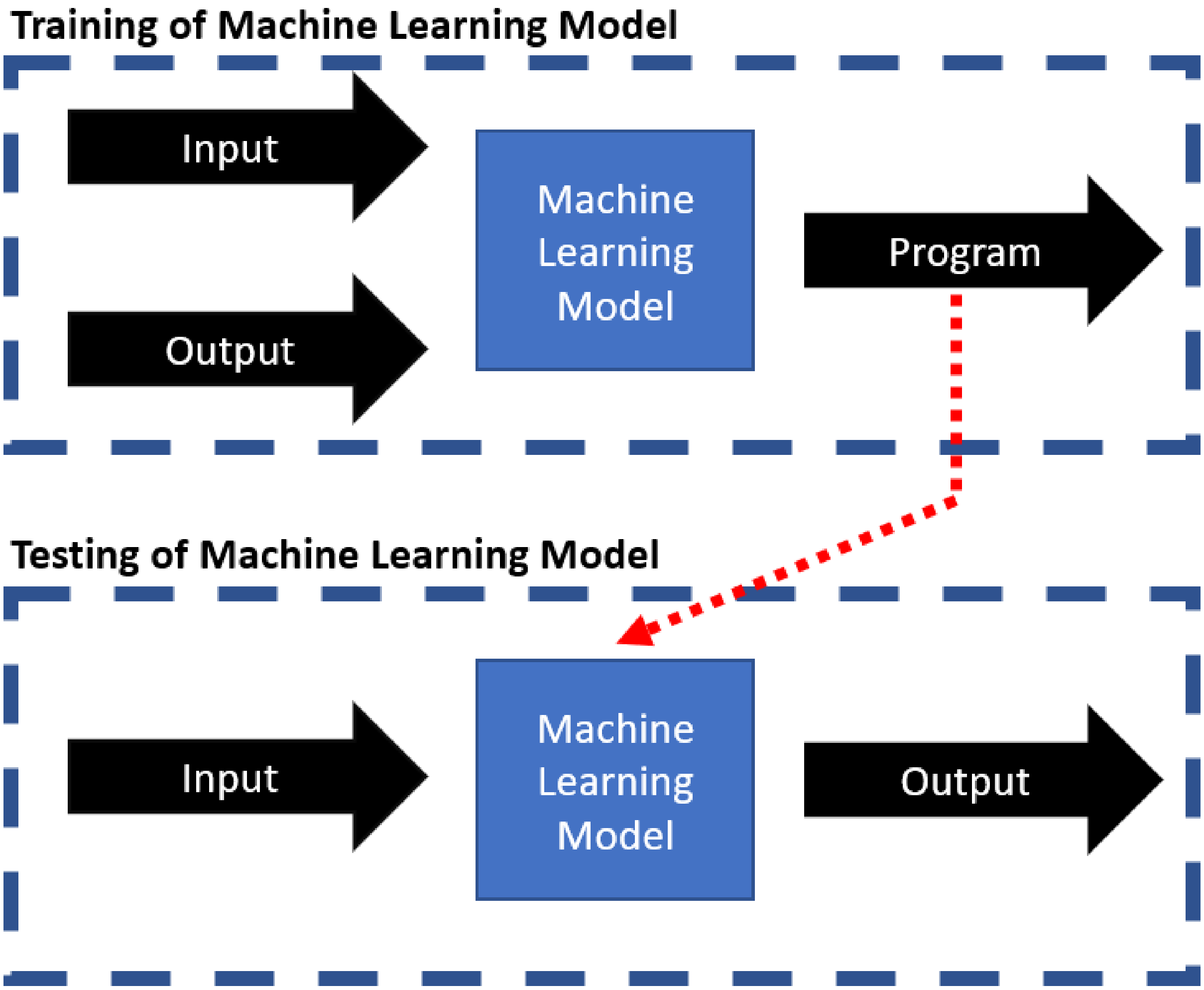


Figure 12 - Training and Testing Design

Machine Learning Model

There are a variety of different Machine Learning Models.
A Linear Polynomial Regression Machine Learning Model was used to fit the training dataset.
Figure 4 displays the model notation of the Linear Polynomial Regression Model

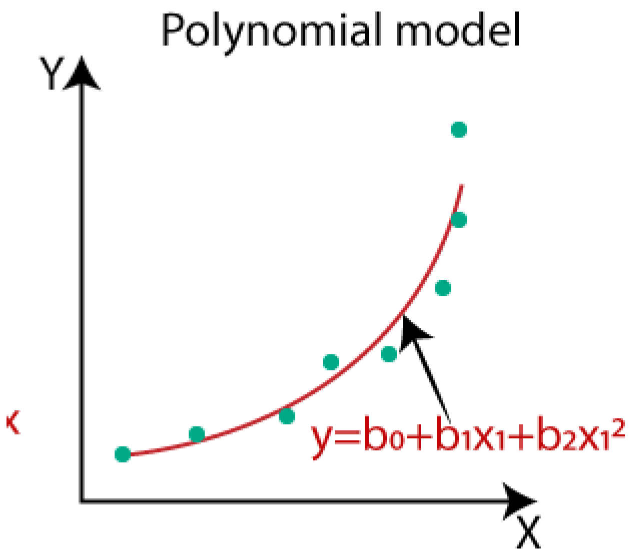


Figure 13 - Linear Polynomial Regression Model
https://www.javatpoint.com/machine-learning-polynomial-regression

Evaluation Metric

The Root Mean Squared Log Error (RMSLE) was performed to measure the error rate of the model. RMSLE measures the ratio between actual and predicted values.
RMSLE puts more weight on the small amount of large numbers. Being off 10k in the home price would be proportional instead of being treated the same.

For example, being 10k off from a 200k home price is not the same as being off 10k for a 500k home price.

Results of Analytical Questions of Interest:

2. Out of the Machine Learning Model, which explanatory variables most positively and negatively affect the price of the house (response variable)?

Below are the top 10 positive and negative feature weights for the explanatory variables.
The green displays the positive while the red displays negative features.

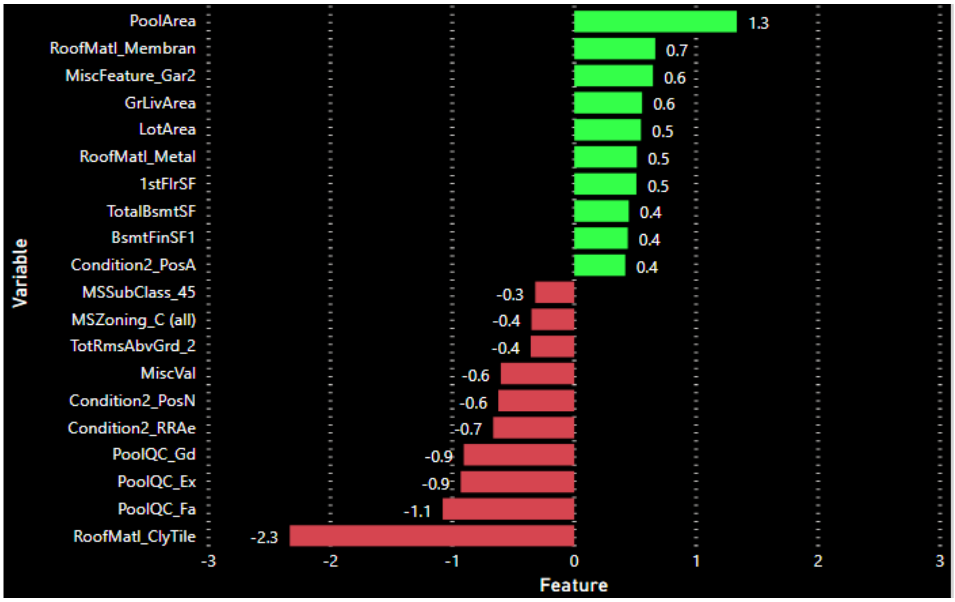


Figure 14 - Top 10 Positive and Negative Variable Features

There is significant value knowing which variables positively and negatively affect the house price.

If someone is selling a house with Roof with Clay Tile, they might consider updating the roof to a different material prior (like Membran which has high postive feature) to selling the house.

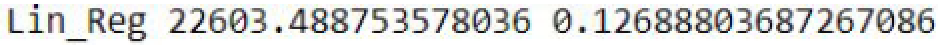
If someone is buying a house, it is value to know that there is a significant premium if there is there is a pool with the house.

Results of Analytical Questions of Interest:

3. How accurate can a Machine Learning model predict the home price based on a series of explanatory variables?

To determine the results of the machine learning model, the trained machine learning model was tested against the test dataset.

Below are the results from trained machine learning model against on the testing data.



```
Lin_Reg 22603.488753578036 0.12688803687267086
```

The Root Mean Squared Log Error (RMSLE) is 0.1268

The RMSLE is a fraction that measures how far the machine learning model is off from predicting the correct price.
Since the RMSLE is in log form, it needs to be transformed.
abs(1 - exp(.1269)) is the formula to obtain the error rate from the RMSLE.

The machine learning model is on average **13.5%** off from predicting the actual home price.

There is significant value knowing how well this machine learning model performs.
A real-estate investor could find value in using this machine learning model to compare to the asking price to determine if the investment is a favorable deal.

SUMMARY OF QUESTIONS OF INTEREST

Explanatory Question

1. Which explanatory variables are highly correlated with the Sale Price (Response) variable?

The top 3 variables that are most correlated with the Sale Price (Response) variable:

1. **OverallQual**
2. **GrLivArea**
3. **GarageCars**

The top 3 variables that are Lease correlated with the Sale Price (Response) variable:

1. **EnclosedPorch**
2. **KitchenAbvGr**
3. **MSSubClass**

Statistical Questions

2. Which explanatory variables most positively and negatively affect the price of the house (response variable)?

The top 3 variables that are positively affect the house price are:

1. **PoolArea**
2. **RoofMatl_Membran**
3. **MiscFeature_Gar2**

The top 3 variables that are negatively affect the house price are:

1. **RoofMatl_ClyTile**
2. **PoolQC_Fa**
3. **PoolQC_Ex**

3. How accurate can a Machine Learning model predict the home price based on a series of explanatory variables?

The machine learning model is on average **13.5%** off from predicting the actual home price.

FUTURE WORK

Upon putting together this technical report, there are two areas where I plan to explore in the future.

- Further examine PoolQC_Ex variable. It seems odd that a pool in excellent condition would have a negative effect on the price of houses within the machine learning model.

- Fine tuning the machine learning model to decrease the error rate.

  Two areas to explore are:

  - Review of outliers in the dataset to see if adjusting for outliers increases the accuracy of the machine learning model.

  - Remove variables that are not statically significant