House Prices

Nicholas Morris

Machine Learning

Wrangling

Feature Engineering

Modeling

Missing Values

Replacing missing values in categorical columns with None. Replacing missing values in numeric columns with 0. Converting Year Sold to a date.

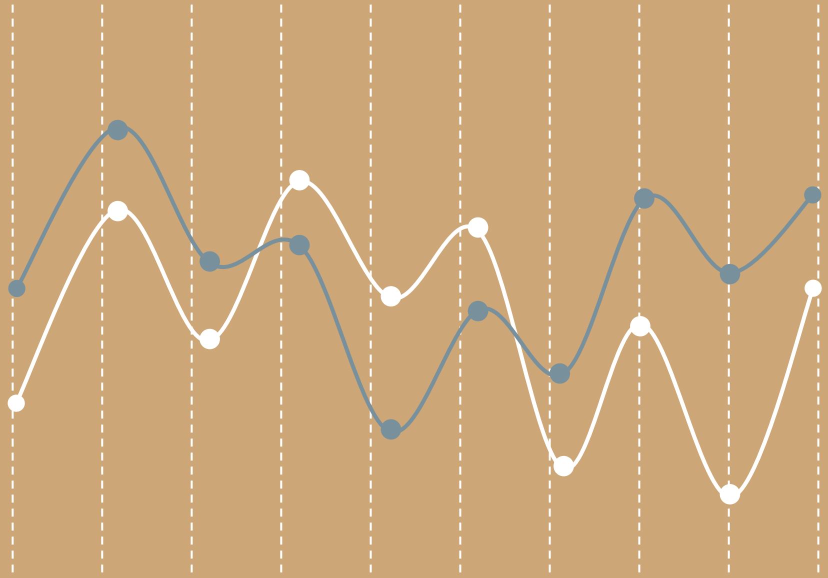
Additional Data

Adding economic trends such as unemployment and CPI. Converting categorical features to binary data points. Attempting feature transformations.

Predictions

Training linear regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

Below is the first two houses in the data. There are 1,460 total houses and 81 columns. The target we are predicting is SalePrice. [Link to the dataset and code]

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500

Removing Unnecessary Columns

Id was removed from the data.

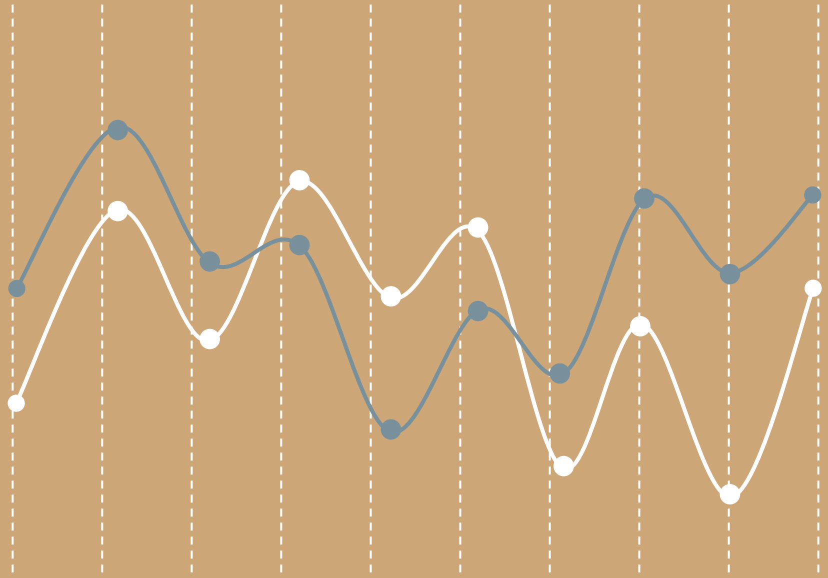
Missing Values

Replacing missing values in categorical columns with None. Replacing missing values in numeric columns with 0.

Date Sold

Converted year sold to date sold.

Feature Engineering



Binary Data

Various categorical features, some words and others are integers, were converted to binary data points.

Economic Data

The NASDAQ closing price, Unemployment rate, CPI, PPI, GDP, GDI, and Federal Funds Rate were pulled from FRED (Federal Reserve Economic Data).

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

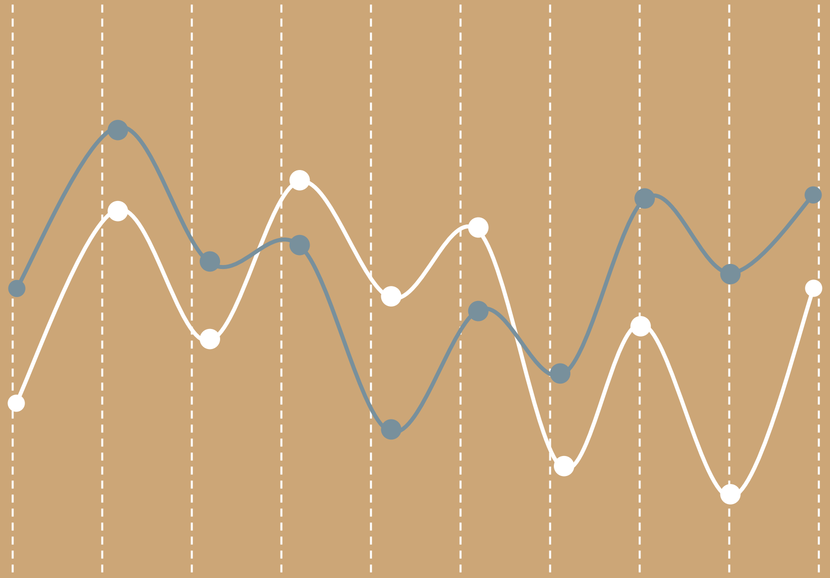
Reciprocals did not improve model performance, so, it was left out of the final model.

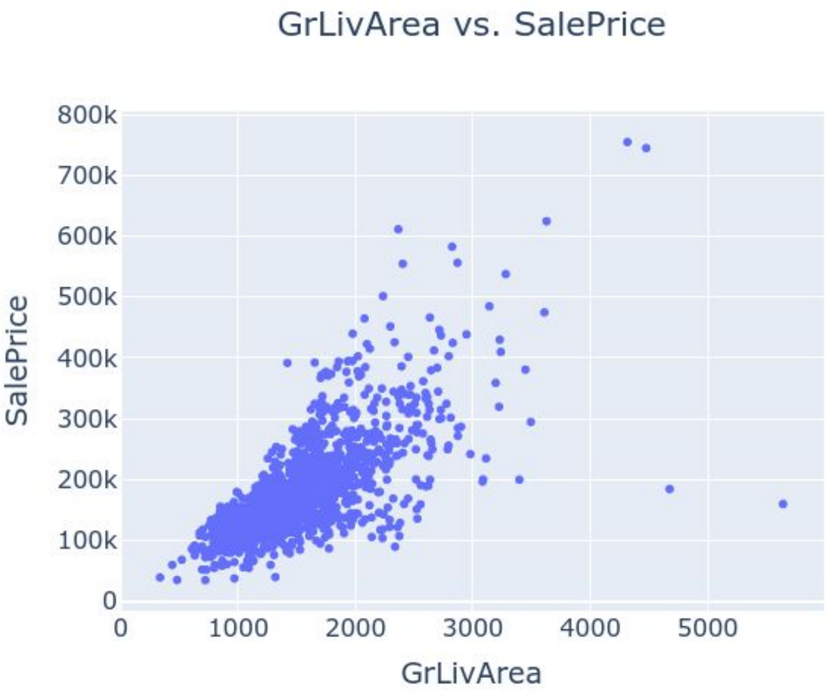
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

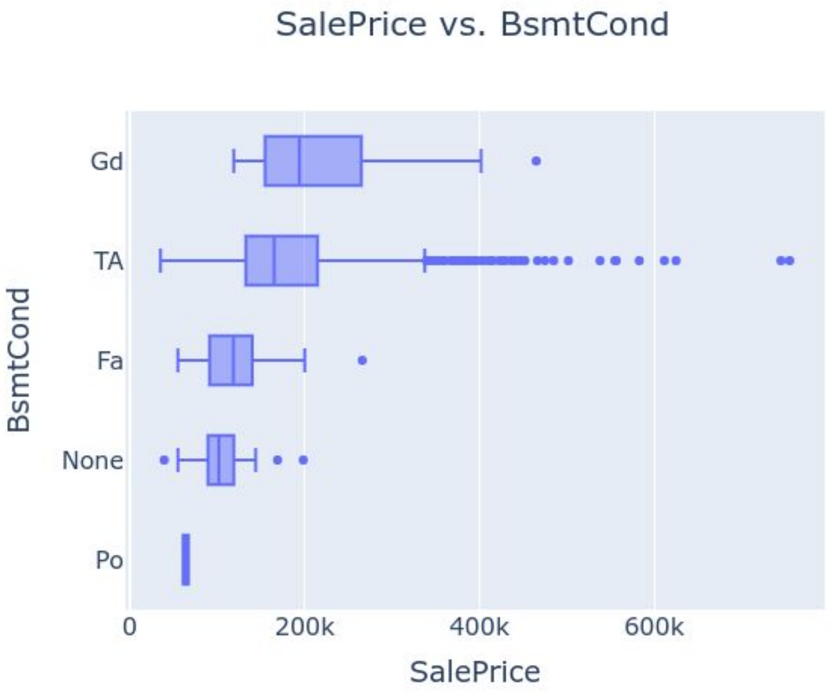
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration

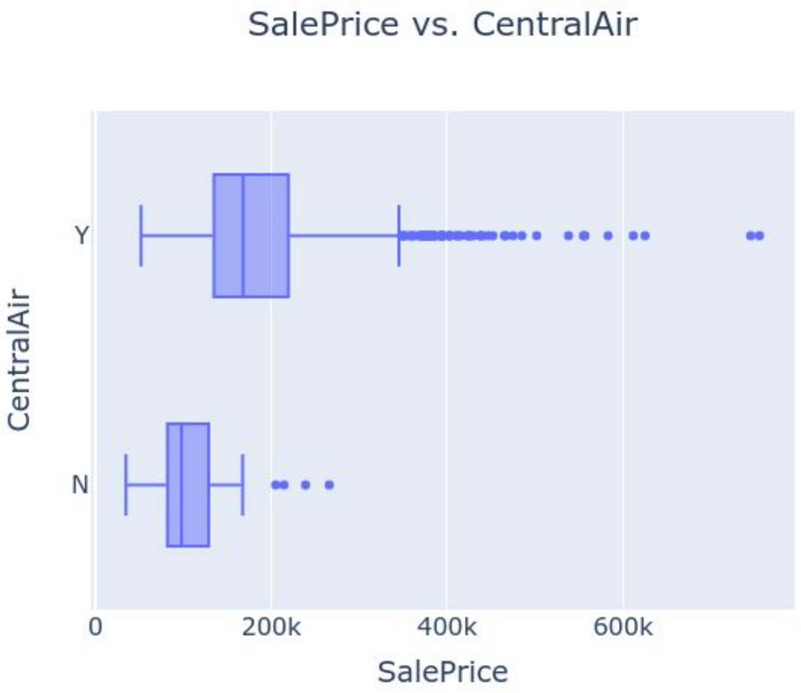




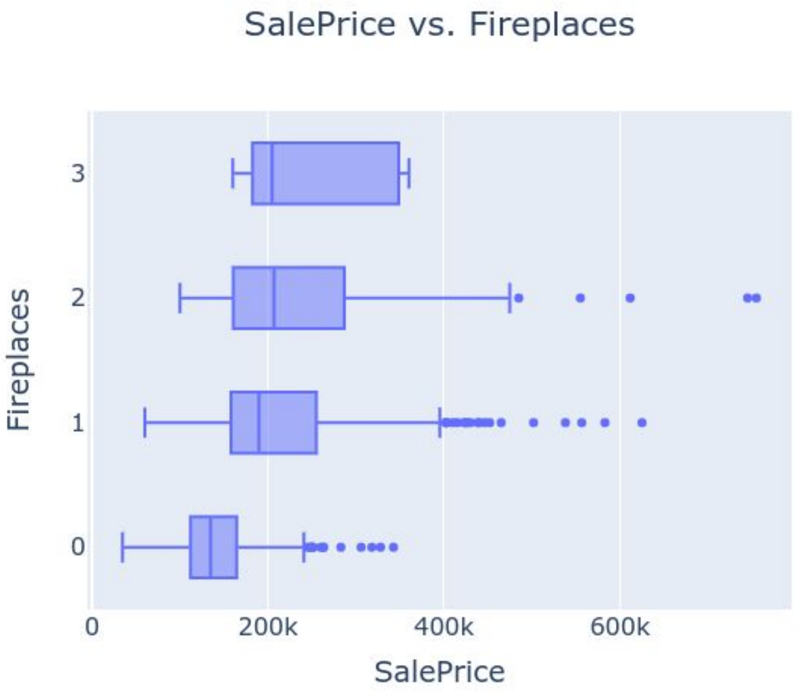
We can see that as the ground living area increases the price of the house is higher.



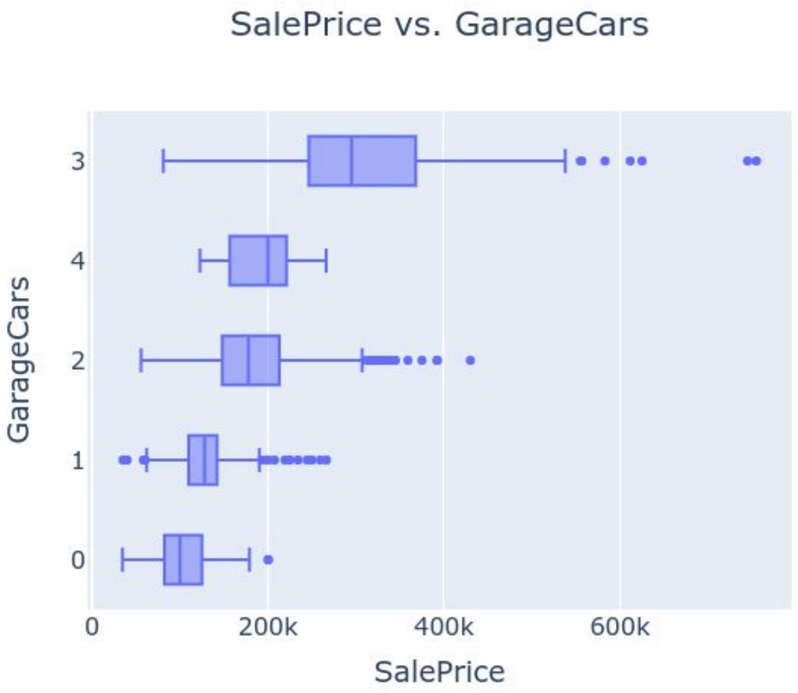
We can see that the price varies according to the basement condition.



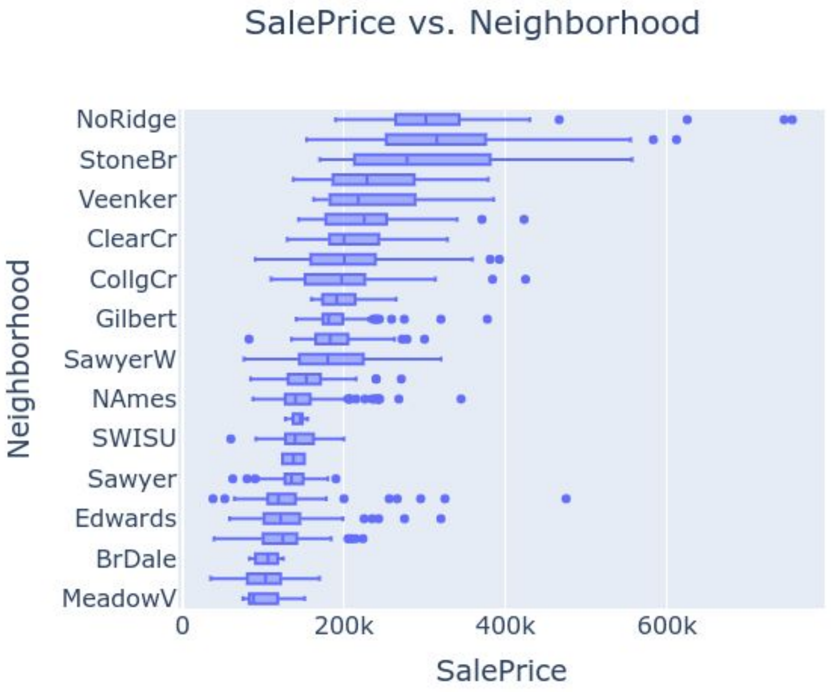
We can see that if the house has central air, then the price tends to be higher.



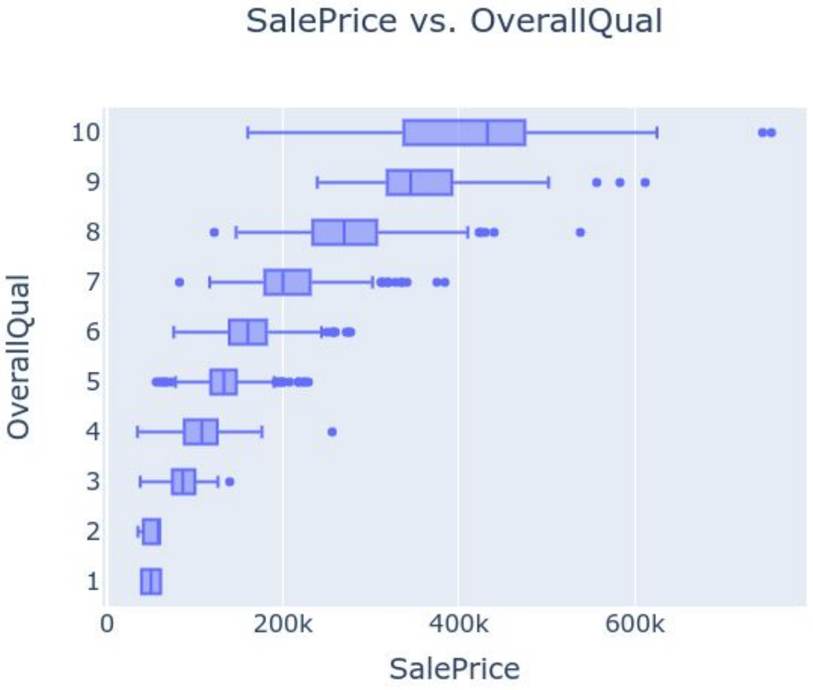
The more fireplaces a house has the higher the price.



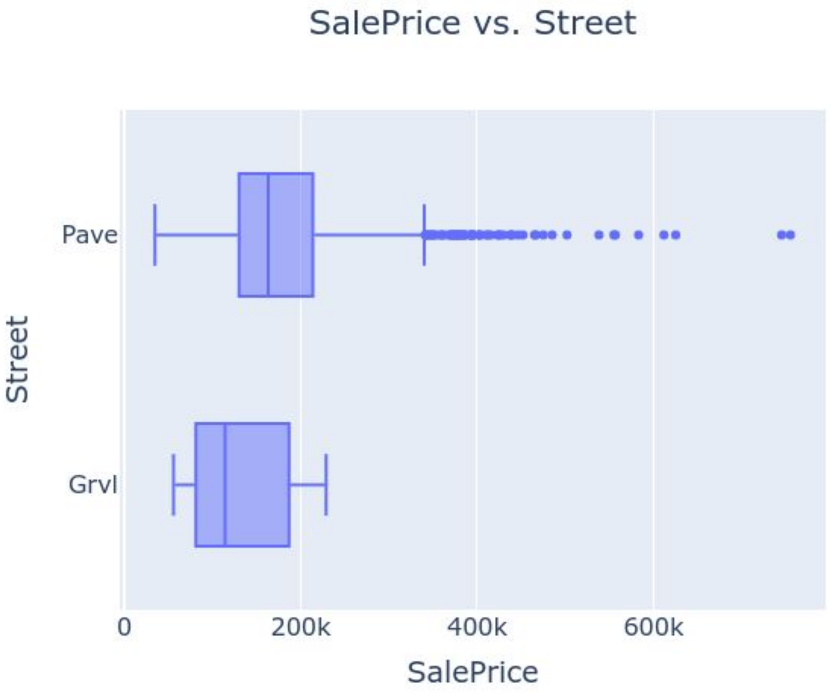
The more car spaces in the garage the higher the price of the house; but there’s a tipping point where 3 car spaces tends to be a more expensive house than 4 car spaces.



We can see that the price varies according to the type of neighborhood the house is in.

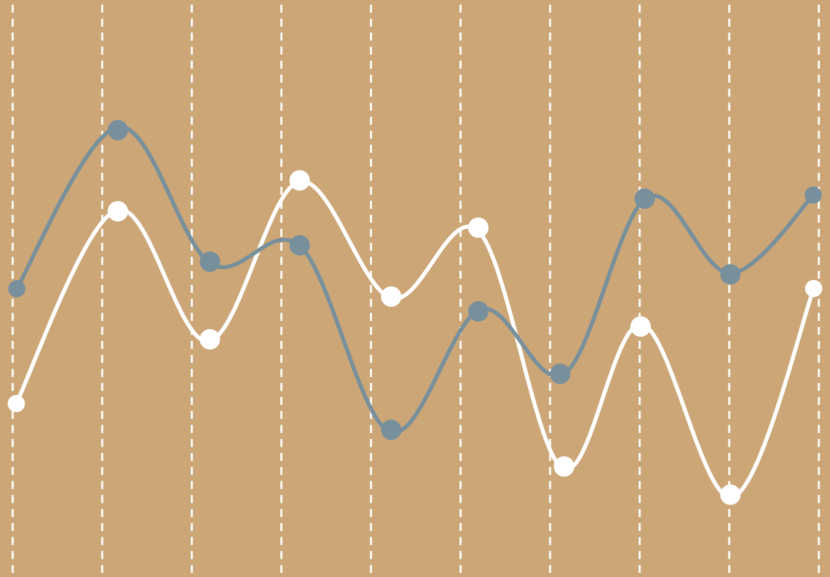


We can see that the higher the quality of the house the higher the the price.



We can see that a house with a paved street tends to be more expensive than a house with a gravel street.

Modeling



Model Parameters

Linear Regression

Library: scikit-learn
Length Of Path: 1e-9
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 500

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Linear Regression

R2: 0.89
RMSE: 23,499
In Control: 97.59%

Model Indicators:
1. RoofMatl_ClyTile
2. GrLivArea
3. Condition2_PosN
4. OverallQual_10
5. OverallQual_9

XGBoost

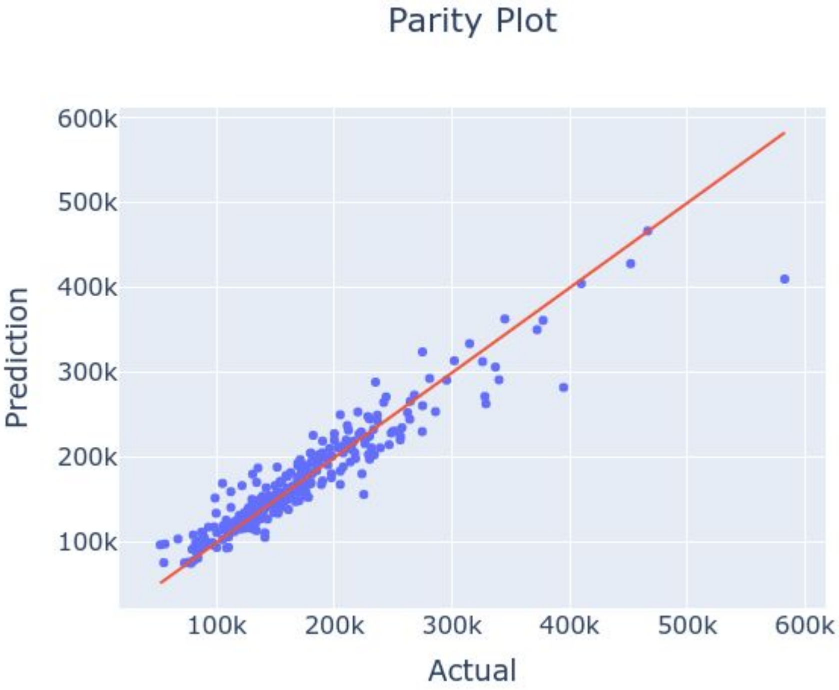
R2: 0.90
RMSE: 22,701
In Control: 98.28%

Model Indicators:
1. ExterQual_TA
2. GarageCars_3
3. BsmtQual_Ex
4. KitchenQual_TA
5. KitchenQual_Ex

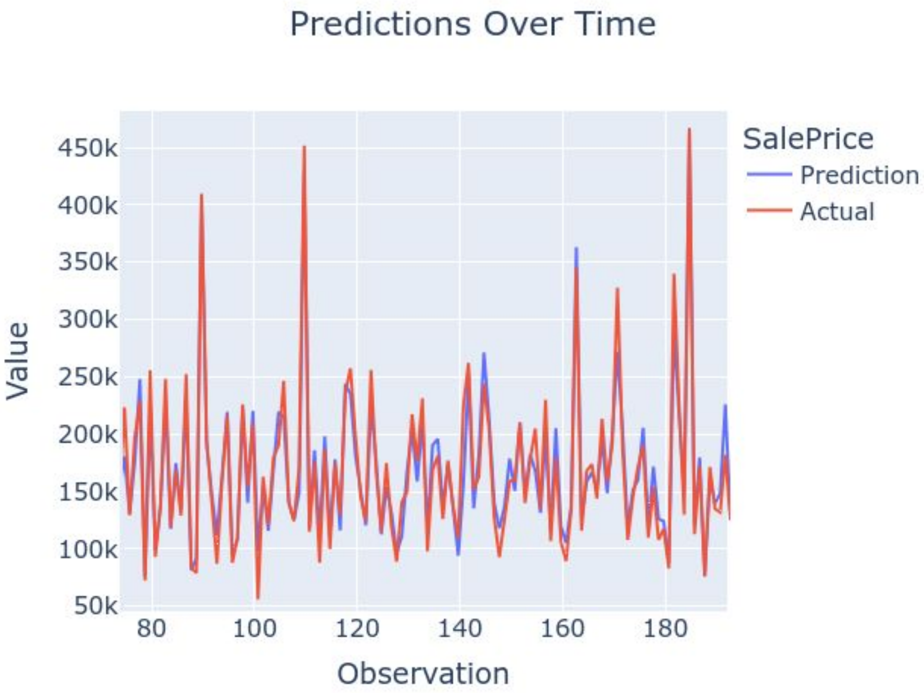
Neural Network

R2: 0.85
RMSE: 26,841
In Control: 97.59%

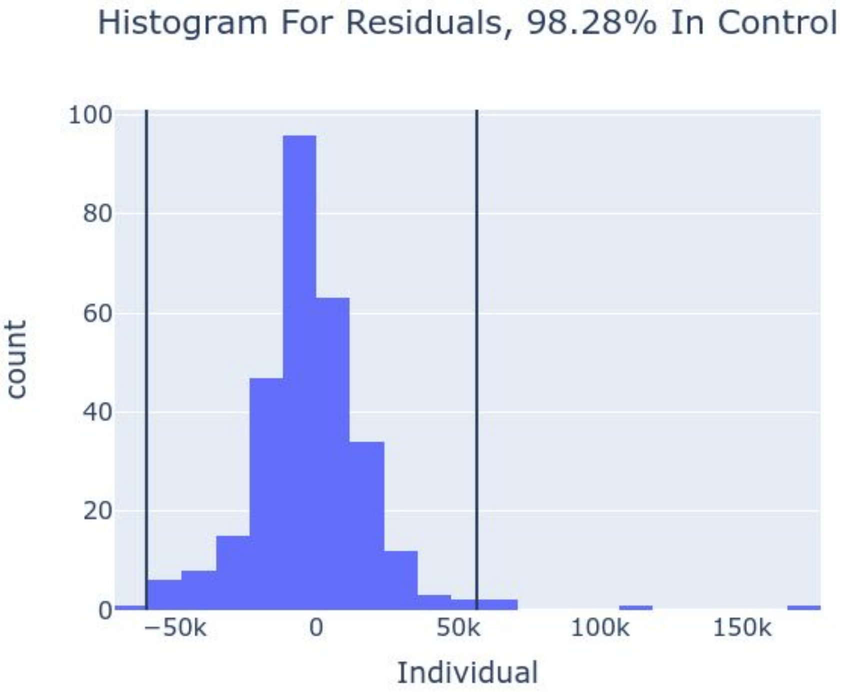
Model Indicators:
1. LowQualFinSF_397
2. LowQualFinSF_420
3. LowQualFinSF_481
4. LowQualFinSF_392
5. LowQualFinSF_473



These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. The predictions are centered on the red line (perfect predictions). There is a tendency to under-predict the price of more expensive houses.

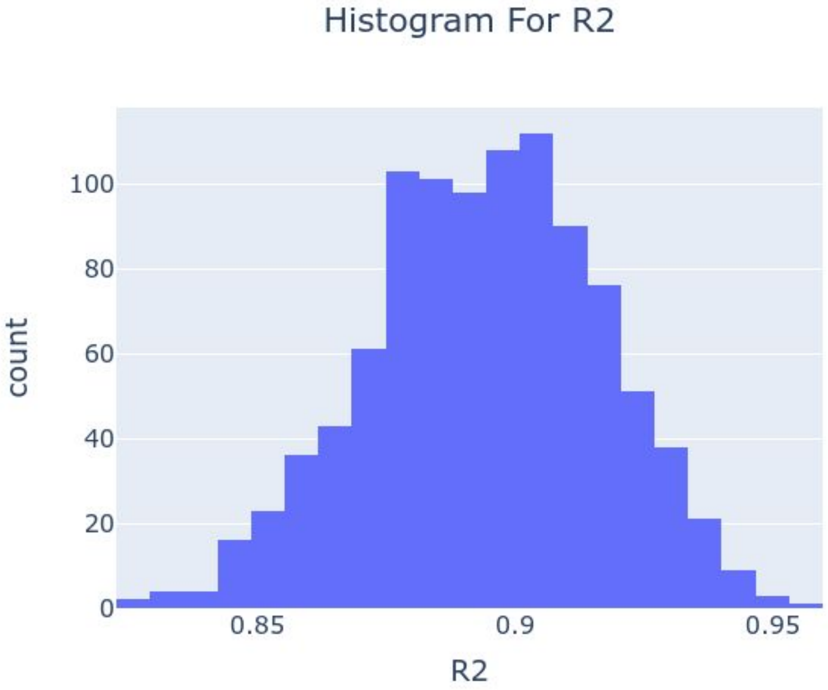


Here’s a snapshot of the predictions over time. We can see the the blue predictions follow the actual values well. There are some instances where the predictions go above the actual values. We can see there is no seasonality in the data.



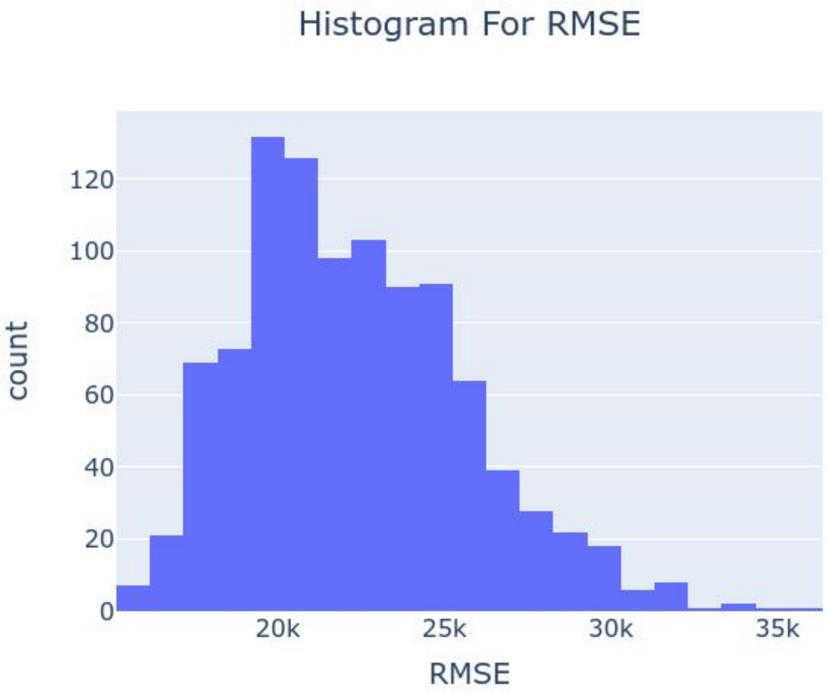
The residuals are prediction error = actual - predicted.

The residuals have a tight bell shape, which is good, and they are centered on zero. Control limits were computed on the residuals and we can see that the prediction error is mostly under control. We can see a skew to the right, which shows the model has a tendency to under-predict the price.



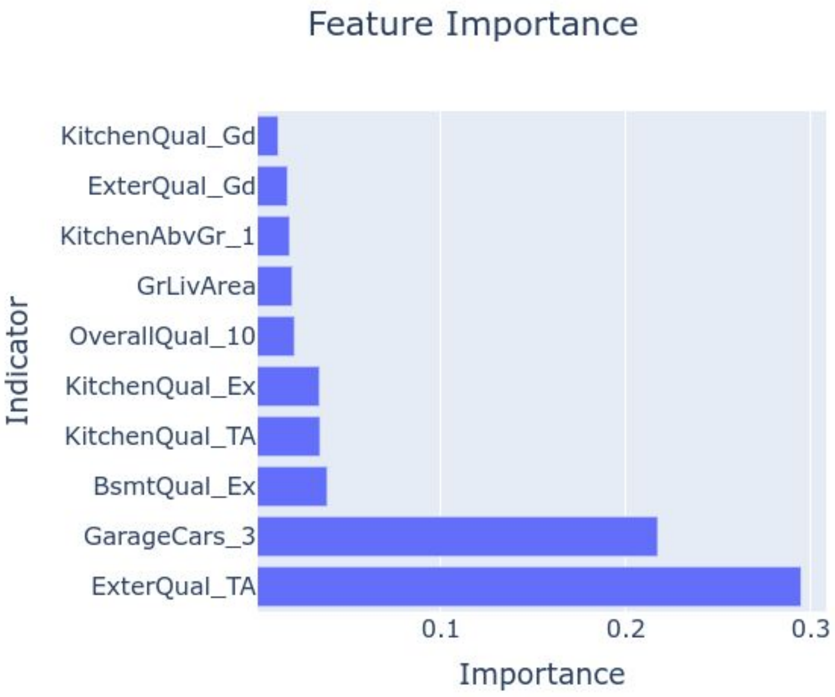
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then R2 was computed on each sample to get a distribution.

R2 has a reasonable range between 0.85 and 0.95, which is good. R2 has a bell shape, which is good.

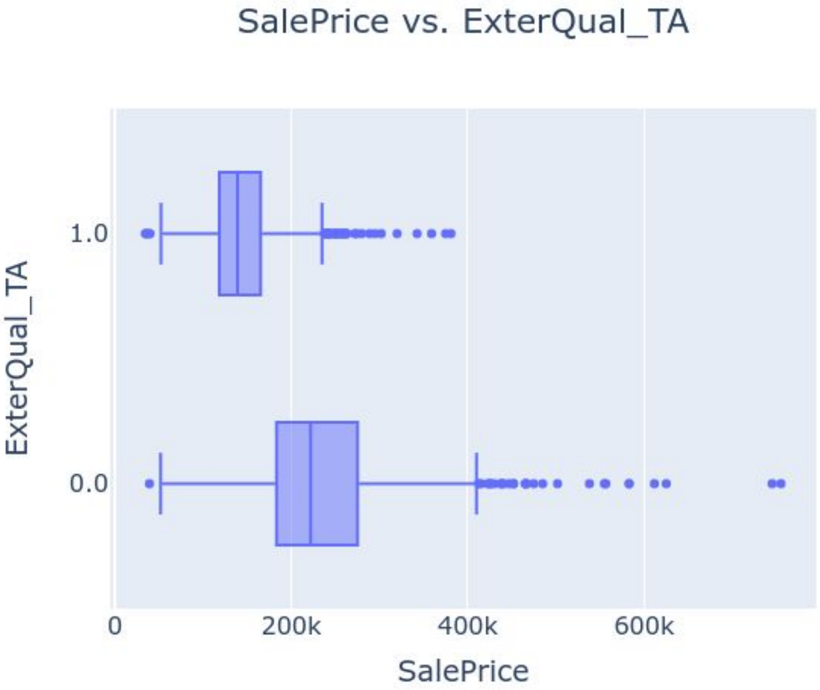


The prediction error was resampled as previously mentioned to get a distribution for RMSE.

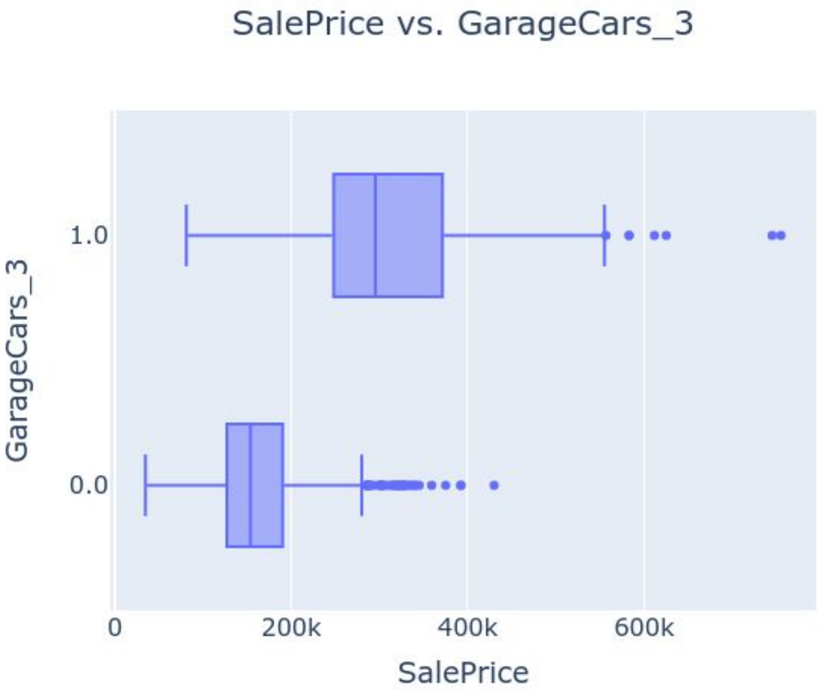
On average, the predictions are off by 20 thousand to 30 thousand, which is a reasonable range. There is a skew to the right, which isn't good.



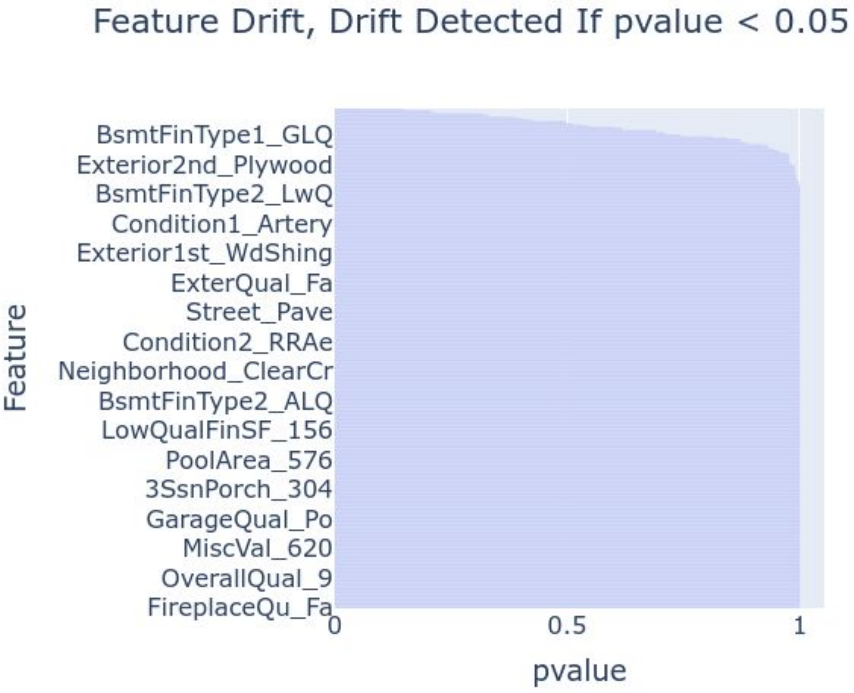
These are the top ten indicators of house price. Typical exterior quality and 3 car garages are the most import indicators with the remaining indicators tapering off.



When the exterior quality is typical the price of the house is less with a median of 139 thousand, and the other houses have a median price of 222 thousand.

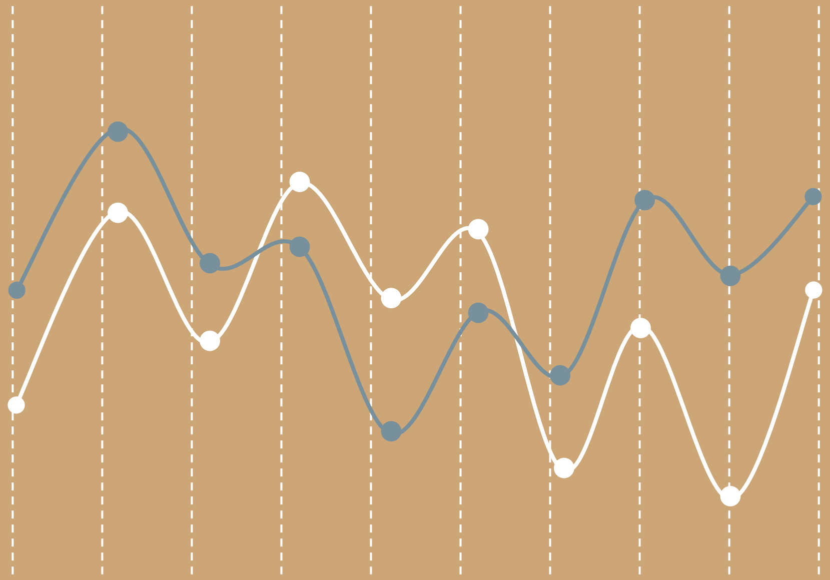


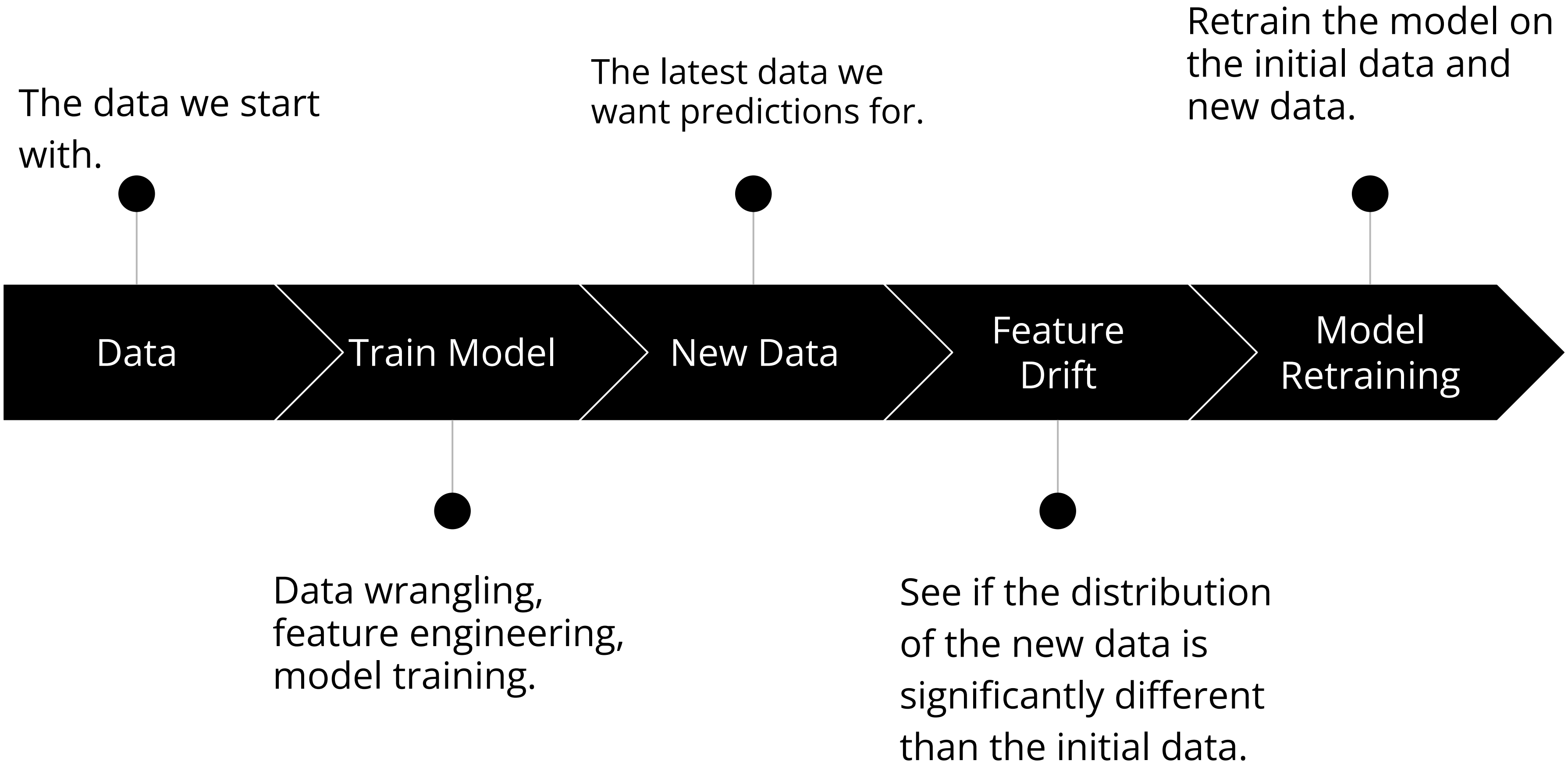
3 car garages are much more expensive than other houses. A 3 car garage house has a median price of 295.5 thousand, and the other houses have a median price of 153.9 thousand.



A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. All of the columns do not experience a drift, which is good.

Deployment





Data

The data we start with.

Train Model

Data wrangling, feature engineering, model training.

New Data

The latest data we want predictions for.

Feature Drift

See if the distribution of the new data is significantly different than the initial data.

Model Retraining

Retrain the model on the initial data and new data.

Thank You