

HOUSE PRICES: ADVANCED REGRESSION TECHNIQUES PROJECT



BY SIMRAN CELLINI

PRESENTATION OUTLINE

- Project motivation
- Project aim.
- Gathering and analyzing my data.
- Heatmap
- Approach
- Data Exploration.
- Data Exploration Part 2.
- Methodology.
- Decision Tree Result
- Random Forest Results.
- Gradient Boosted Trees Results.
- Conclusion

PROJECT MOTIVATION



When researching into this topic I found out the following:

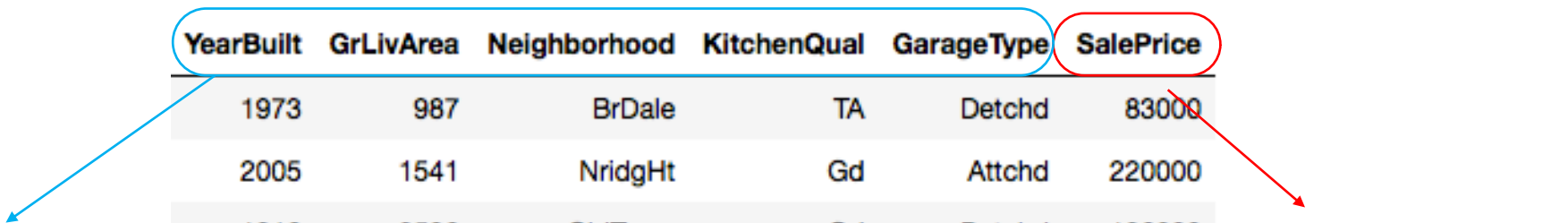
- The shortage of homes to buy has caused prices to rise sharply in many areas in the US.
- Home prices in New York City have risen 4.1 percent in the past year, still much higher than U.S. average hourly earnings.
- Eight out of 10 non-homeowners indicate that owning a home is part of their American Dream.

PROJECT AIM

BASED ON DIFFERENT HOUSE FEATURES CAN WE PREDICT WHAT THE SALE PRICE WILL BE?

GATHERING & ANALYSING MY DATA

The first step was to gather and analyze my data. The table below shows 5 different types of house features and the output of the project which is the Sale Price. My table has a total 79 different types of house feature.



YearBuilt	GrLivArea	Neighborhood	KitchenQual	GarageType	SalePrice
1973	987	BrDale	TA	Detchd	83000
2005	1541	NridgHt	Gd	Attchd	220000
1916	2526	OldTown	Gd	Detchd	136000
1920	1077	IDOTRR	TA	Detchd	67000
2005	1274	CollgCr	Gd	Attchd	203000

The different house features.

This is the output that I am trying to predict.

YEAR BUILT: The year when the house was built.

GROUND LIVING AREA: The ground floor living area in square feet.

NEIGHBORHOOD: Different neighborhood locations in the US.

KITCHEN QUALITY: The different kitchen quality ratings.

GARAGE TYPE: Whether it is a detached or attached garage.

SALE PRICE: The property's sale price in US dollars.

Heatmap visualization showing the relationship between various features (rows) and the SalePrice (columns). The color scale ranges from -0.8 (blue) to 0.8 (red), indicating the strength and direction of the correlation.

Key features and their approximate correlation with SalePrice:

- OverallQual:** Strong positive correlation (red).
- OverallCond:** Moderate positive correlation (orange).
- YearBuilt:** Moderate positive correlation (orange).
- YearRemodAdd:** Moderate positive correlation (orange).
- MasVnrArea:** Moderate positive correlation (orange).
- BmtFinSF1:** Moderate positive correlation (orange).
- BmtFinSF2:** Moderate negative correlation (blue).
- BmtUnfSF:** Moderate negative correlation (blue).
- TotalBmtSF:** Moderate negative correlation (blue).
- 1stFlrSF:** Moderate positive correlation (orange).
- 2ndFlrSF:** Moderate positive correlation (orange).
- LowQualFinSF:** Moderate positive correlation (orange).
- GLivArea:** Moderate positive correlation (orange).
- BmtFullBath:** Moderate positive correlation (orange).
- BmtHalfBath:** Moderate positive correlation (orange).
- FullBath:** Moderate positive correlation (orange).
- HalfBath:** Moderate positive correlation (orange).
- BedroomAbvGr:** Moderate positive correlation (orange).
- KitchenAbvGr:** Moderate positive correlation (orange).
- TotRmsAbvGrd:** Moderate positive correlation (orange).
- Fireplaces:** Moderate positive correlation (orange).
- GarageYrBlt:** Moderate positive correlation (orange).
- GarageCars:** Moderate positive correlation (orange).
- GarageArea:** Moderate positive correlation (orange).
- WoodDeckSF:** Moderate positive correlation (orange).
- OpenPorchSF:** Moderate positive correlation (orange).
- EnclosedPorch:** Moderate positive correlation (orange).
- 3SsnPorch:** Moderate positive correlation (orange).
- ScreenPorch:** Moderate positive correlation (orange).
- PoolArea:** Moderate positive correlation (orange).
- MiscVal:** Moderate positive correlation (orange).
- MoSold:** Moderate positive correlation (orange).
- YrSold:** Moderate positive correlation (orange).
- SalePrice:** Strong positive correlation (red).

APPROACH

Inputs were taken from the table.

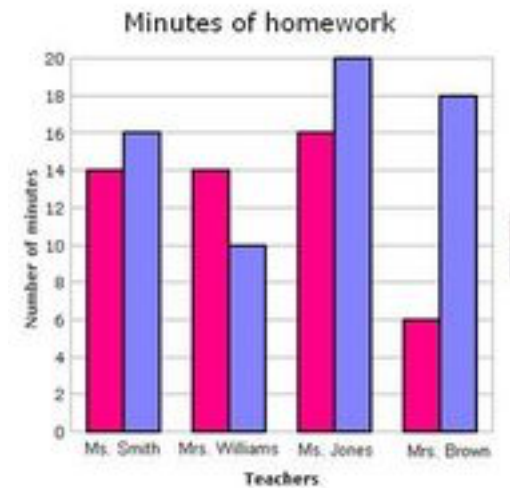
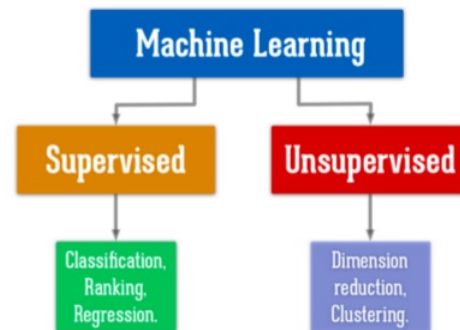
Model built on the training set to predict what the house sale price will be.

Validate / Test the model on the test set . Explore data to understand which inputs are most important.

Analyse the results.

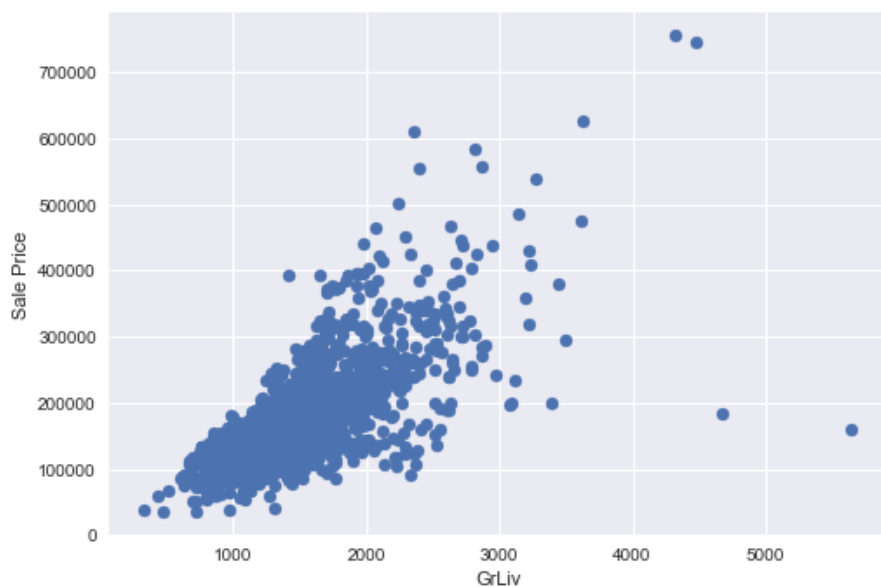
YearBuilt	GrLivArea	Neighborhood
1973	987	BrDale
2005	1541	NridgHt
1916	2526	OldTown
1920	1077	IDOTRR
2005	1274	CollgCr

Types of Learning

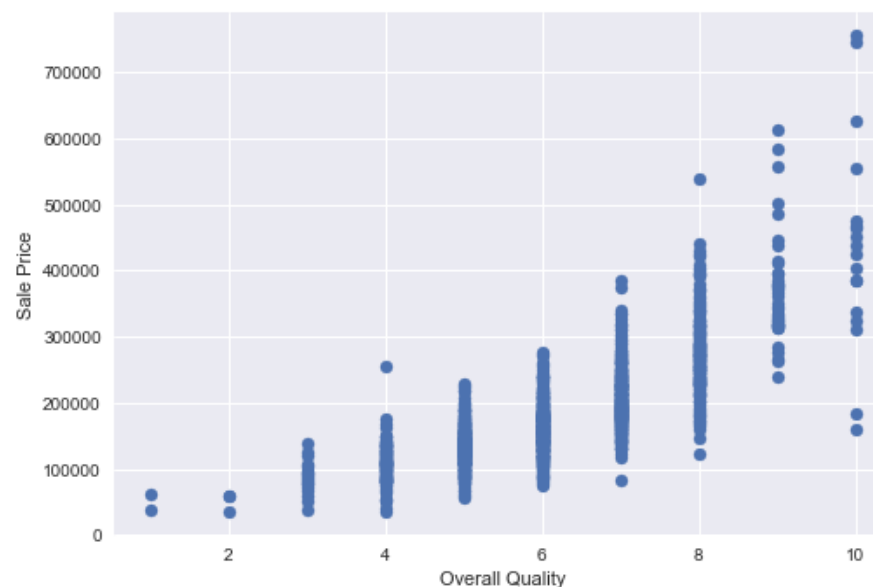


DATA EXPLORATION

In my data exploration stage I created different scatter plots to show how some important features related to the Sale Price.

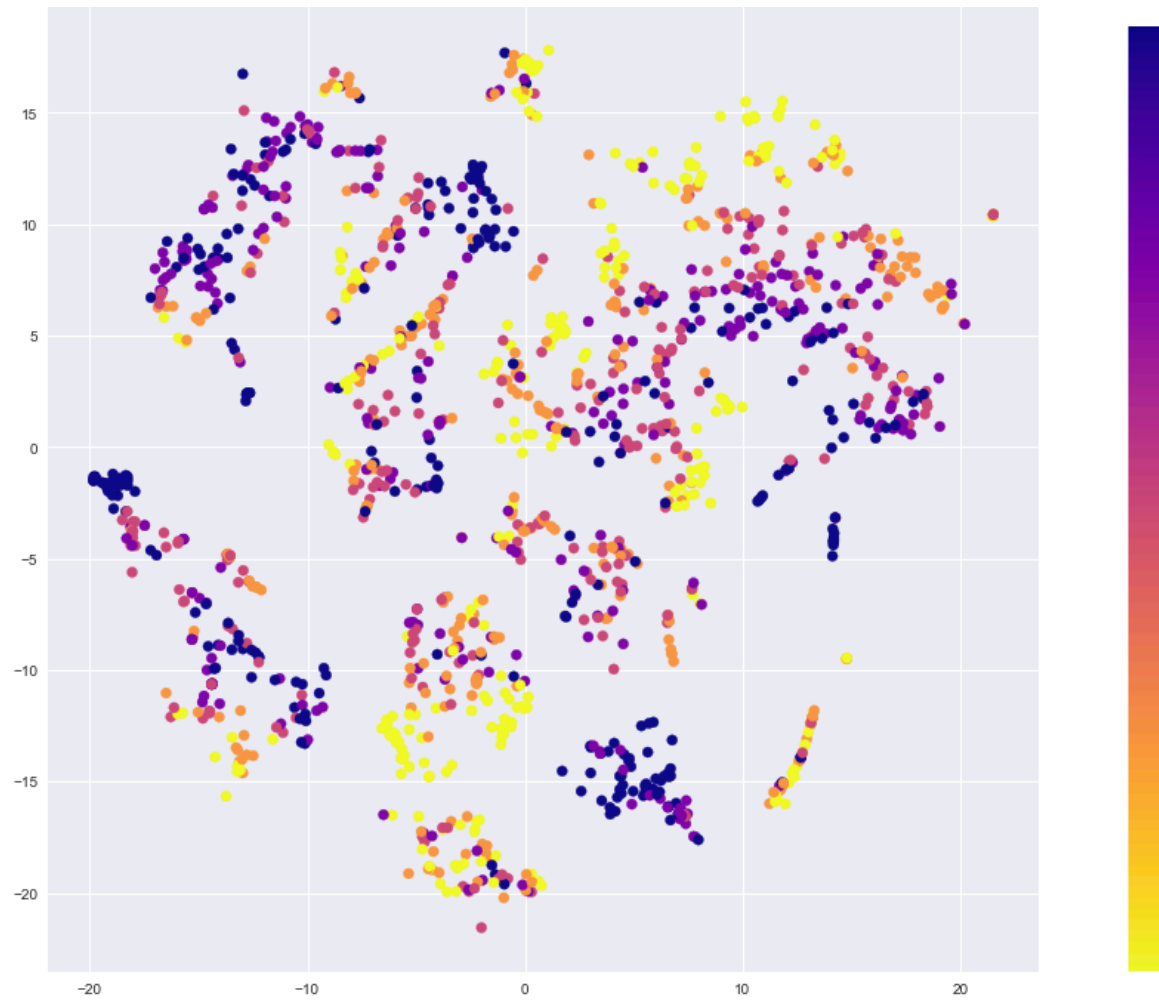


Ground Living Area



Overall Quality

DATA EXPLORATION PART 2



METHODOLOGY

Below are the three machine learning models that I used.

Models

Decision Trees (Makes predictions by choosing one input at a time splitting the data)

Random Forest (By using many decision trees, it takes a random selection of inputs, combines the outputs then takes average of outputs.)

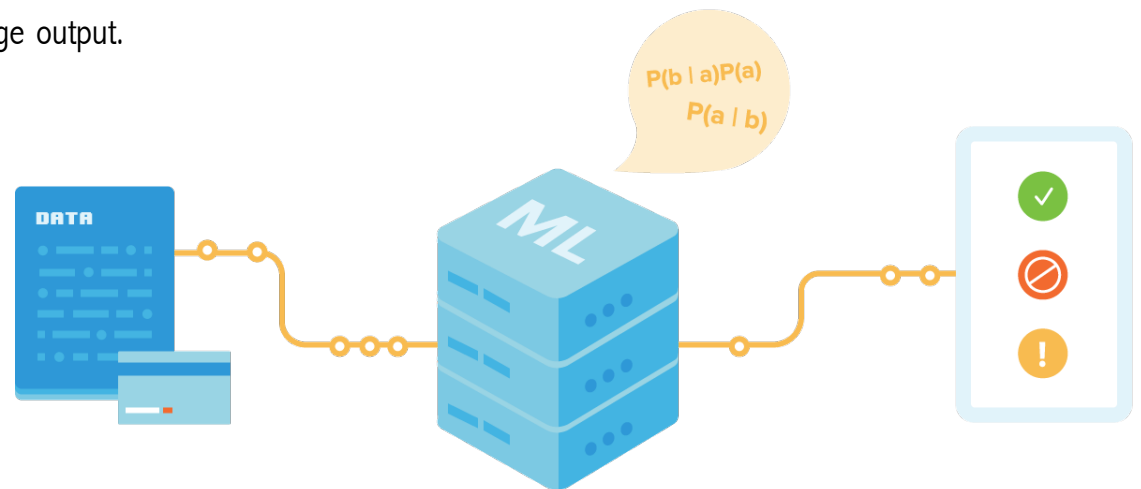
Gradient Boosted Trees (Made up of decision trees, each tree is a weak classifier, adds the outputs and then takes the average output.)

Evaluation

Regression Output

Root Mean Squared Error over test set.

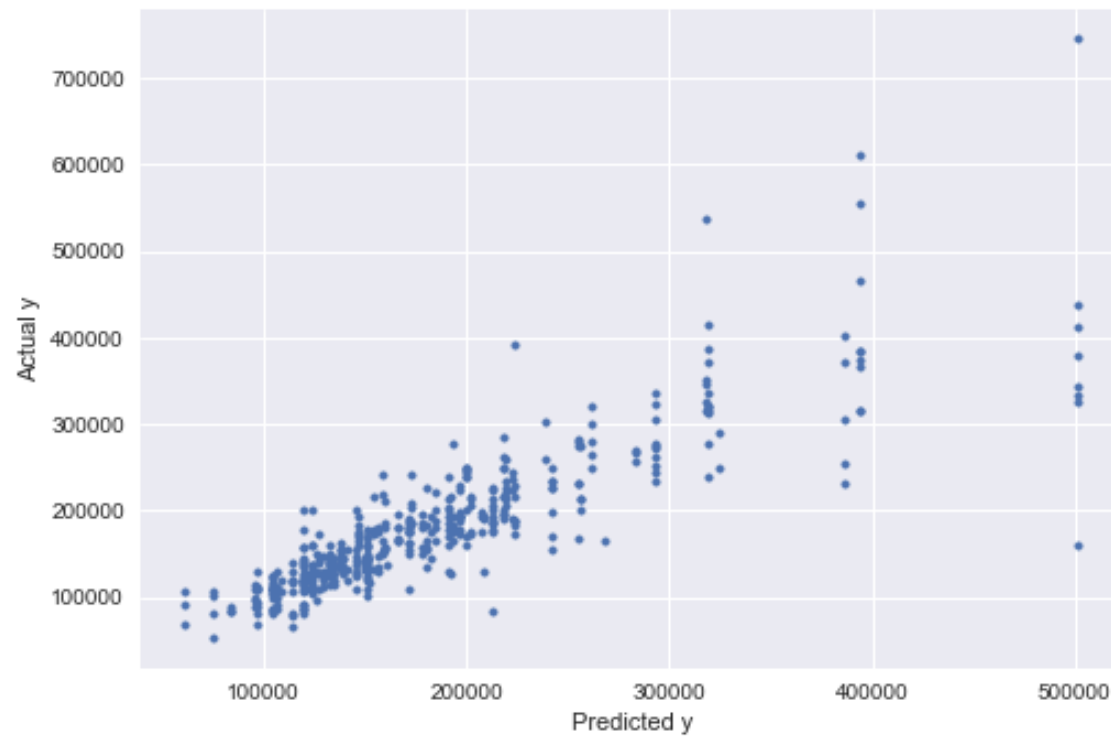
R2 over the test set (how close you are to perfect prediction)



DECISION TREE RESULTS

THE ROOT-MEAN SQUARED ERROR = \$43596

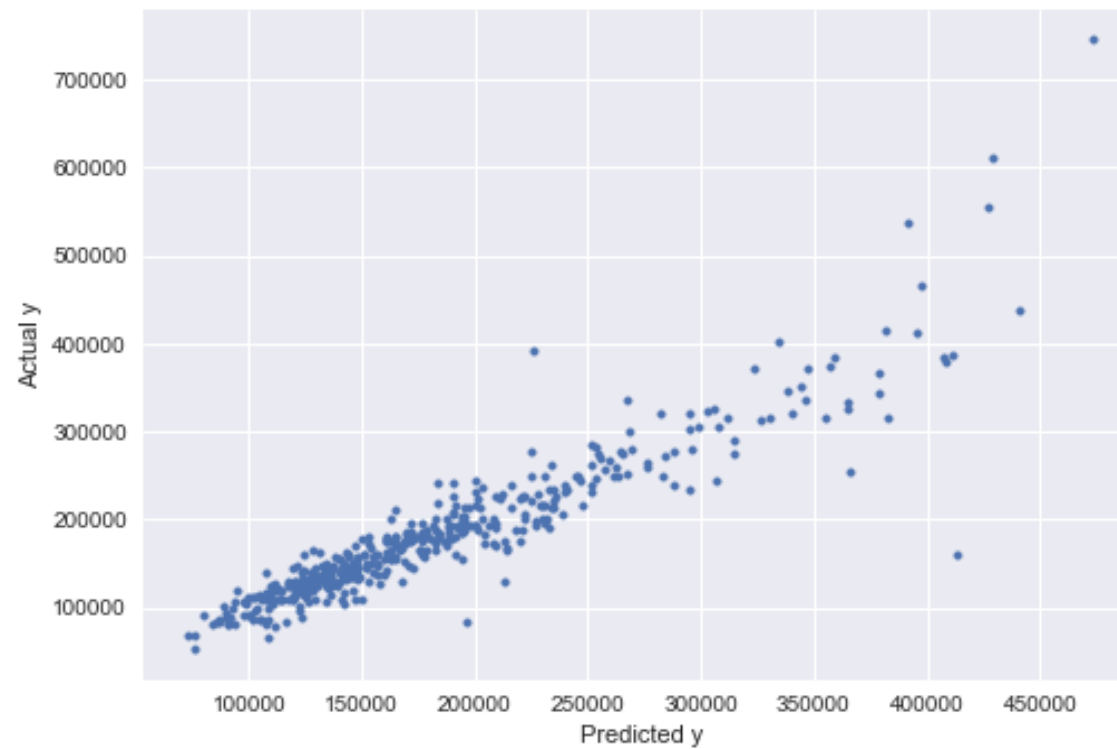
R2 MODEL SCORE = 0.720



RANDOM FOREST RESULTS

THE ROOT-MEAN SQUARED ERROR = \$31395.

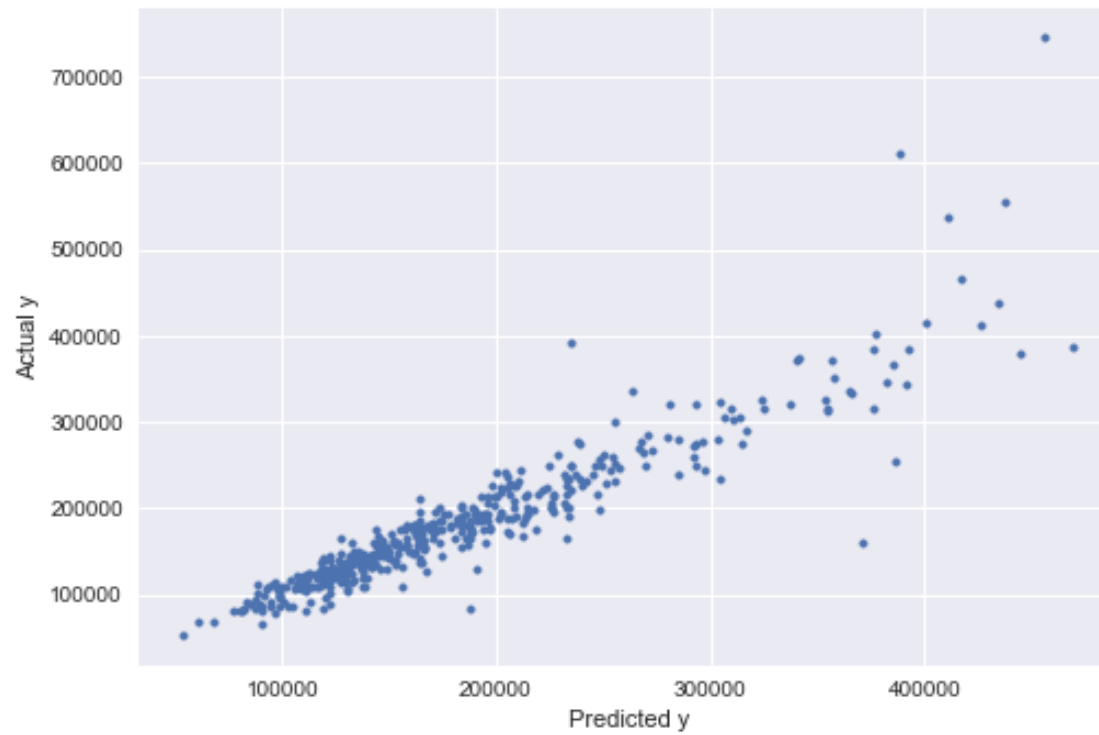
R2 MODEL SCORE = 0.854



GRADIENT BOOSTED TREES

THE ROOT-MEAN SQUARED ERROR = \$31033.

R2 MODEL SCORE = 0.858



CONCLUSION

- To sum up the best models to use were Random Forests and Gradient Boosted Trees. The Root Mean Squared Error for both was around \$30,000. It shows that the predictions are more accurate for the lowest house prices.
- The machine learning models also showed me what the top ten important features are relating to the Sales Price. Number 1 represents the best model feature.
 - 1. Ground Living Area
 - 2. Overall Quality
 - 3. Lot Area
 - 4. Garage Area
 - 5. Total SF Basement.
 - 6. Overall Condition
 - 7. Year Built
 - 8. Unfinished SF of Basement
 - 9. Type 1 Finished SF
 - 10. Open Porch SF

**ANY
QUESTIONS
?**