



# PREDICTING HOUSE PRICES WITH REGRESSION

Bhanu Bhattarai  
Michael Cook  
Jinu Daniel  
Rupinder Grewal  
Basil Marotical

# TEAM 6

## **Bhanu Bhattarai**

- Associate, Project Manager, Commercial Bank, Finance and Business Management

## **Michael Cook**

- Trade Lifecycle Associate, Corporate and Investment Bank, Securities Operations

## **Jinu Daniel**

- Vice President, Chief Quality Office, Operations Technology

## **Rupinder Grewal**

- Vice President, Software Engineer, Retail

## **Basil Marotical**








- Vice President, Data Visualization Manager, Finance Data and Insights



# The AI Canvas

## What task/decision are you examining?

The objective of this model is to predict the market value of a house for the home lending team to verify the house's appraised value.

 <b>Prediction</b>	 <b>Judgment</b>	 <b>Action</b>	 <b>Outcome</b>
The variation of housing prices across different geographical locations	True prediction will reduce risk, increase profitability of the bank. False Positive will increase risk to the bank while False Negative will have adverse impact on the bank-customer relationship.	In case of high variance between the predicted versus appraised market value of the house, the bank can initiate extra measures to minimize the risk.	The primary measure of the model's performance will be Root Mean Square Error (RMSE)
 <b>Training</b>	 <b>Input</b>	 <b>Feedback</b>	
Model needs to be trained with housing data collected from a specific geographical location, over a period of time, and Sales price adjusted for inflation	Specific features of a house such as year built, number of bedrooms/bathrooms, lot area, etc.	Compare predicted house price with the data collected, refine features, and remove the least important features from the dataset to decrease RMSE in the algorithm while operating.	

## How will this AI impact on the overall workflow?

This AI model will improve the appraisal process, which will optimize the home lending workflow from loan origination to closing. This will require modification to the business process and staff re-training.

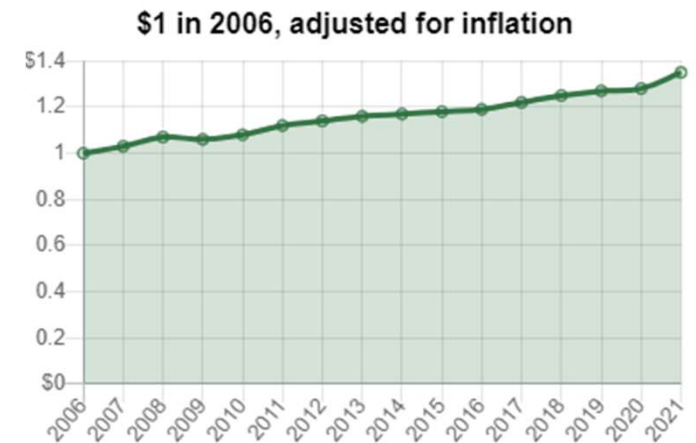
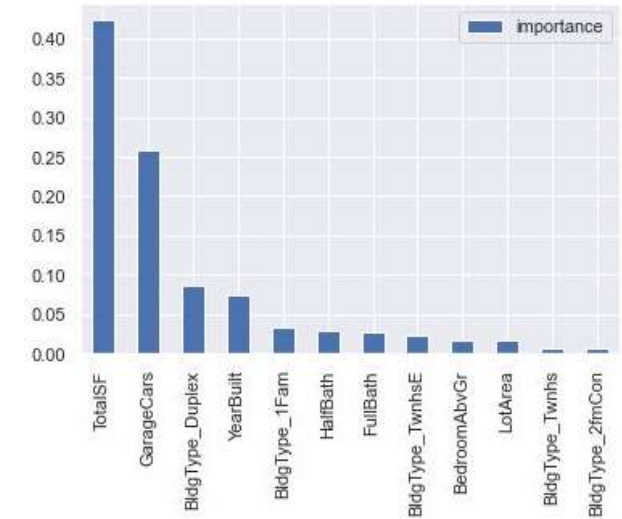
# EDA AND PREPROCESSING

## Initial Data Analysis

- Data was sourced from Ames- IA, Newark- DE, Bear-DE, Wilmington-DE
- Model was trained only on Ames- IA which were from the period of 2006-2010
- Garage car size, Year Built and Total Square Feet have high correlation with Sales Price
- Final columns selected : Lot Area, Total Square feet, Building Type, Year Built, Full Bath, Half Bath, Number of Bedroom, Number of Garage

## Inflation Adjustment

- \$1 in 2006 = \$1.4 in 2021 !



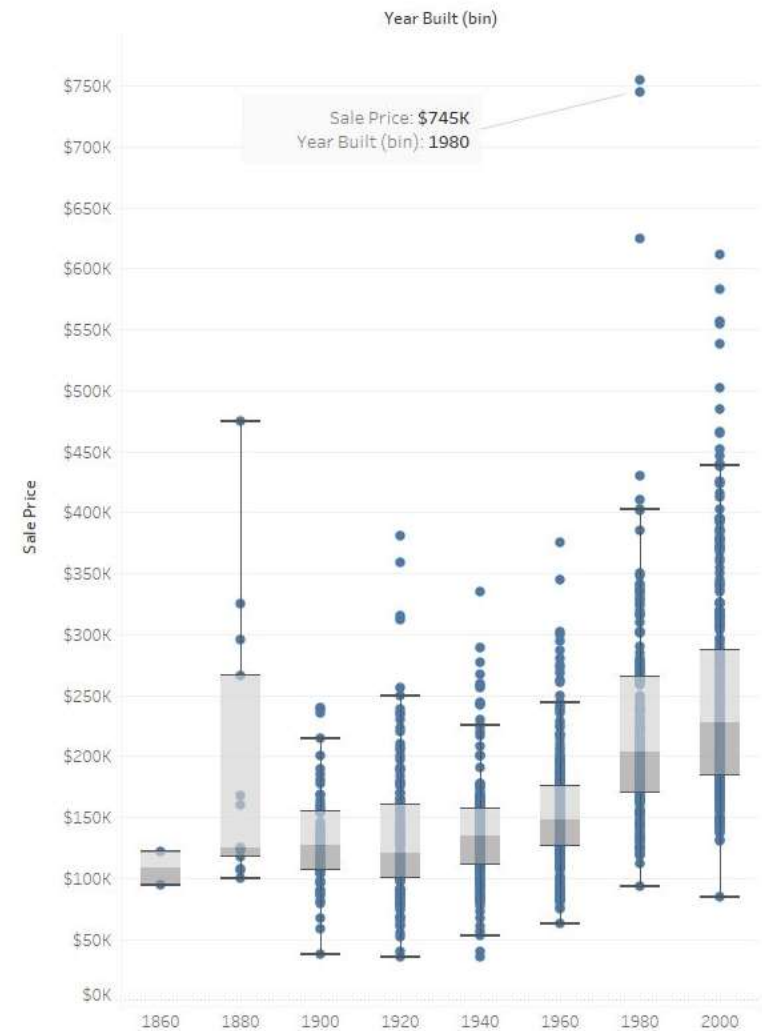
# EDA AND PREPROCESSING

## Data Cleansing

- Out of 19 attributes which had nulls, 4 attributes having more than 70 percent NULLs were dropped

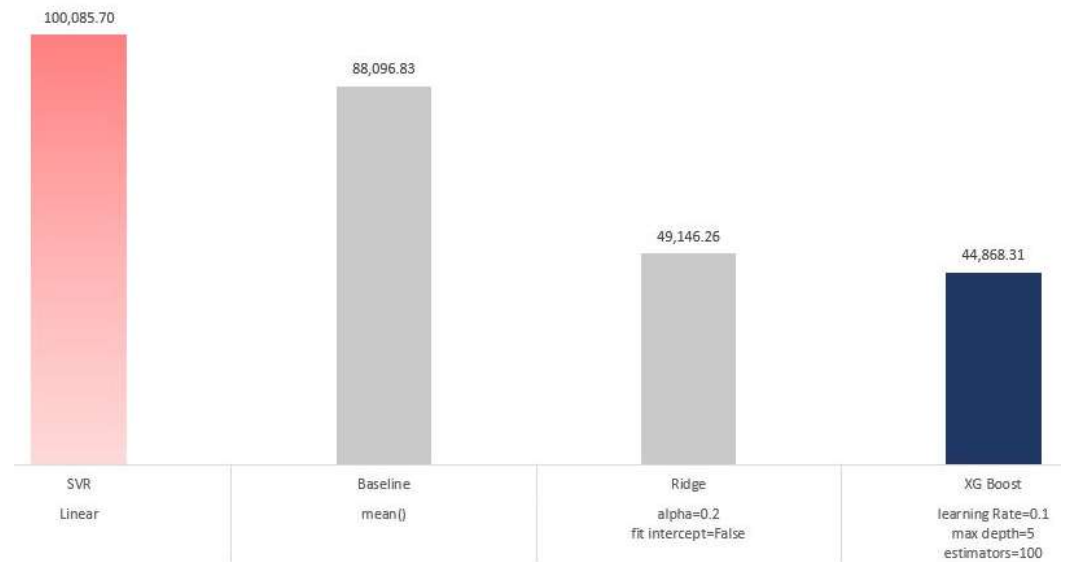
## Data Correlation

- Slight increase in Lot Area increases Sales Price by a higher margin
- Total square footage is highly correlated to Sales Price



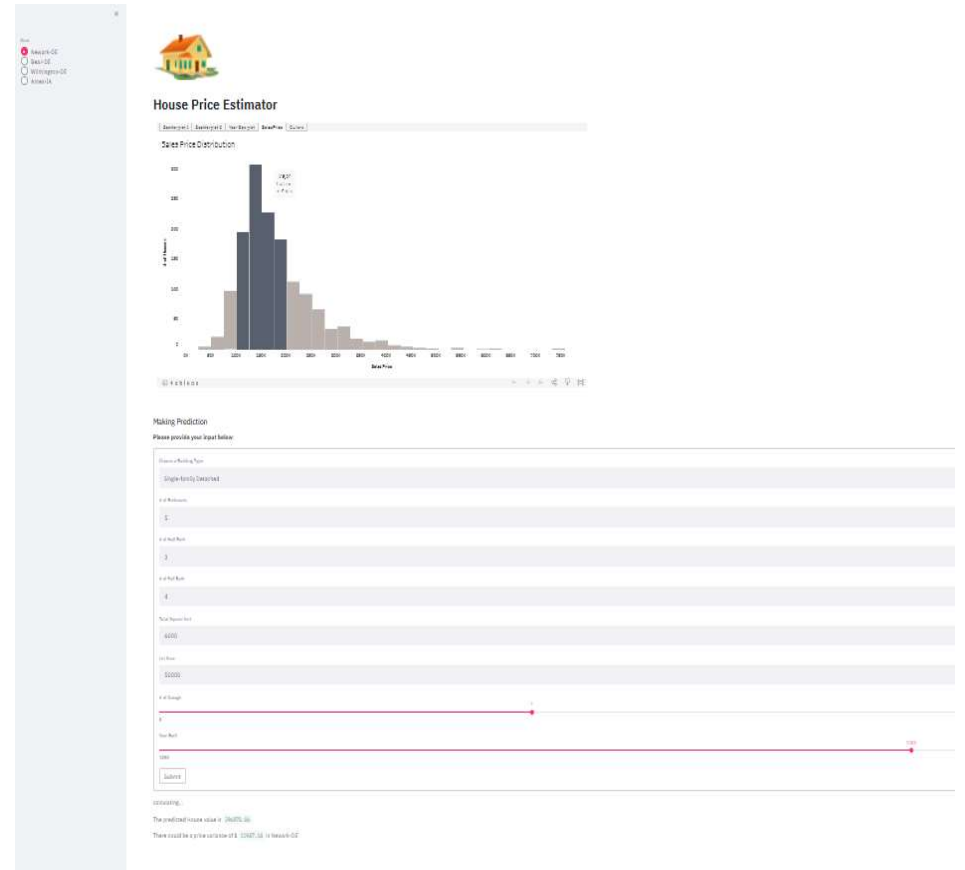
# MODEL TRAINING, TUNING, AND SELECTION

- Selected XG Boost as the best performing model based on RMSE
- Checked the important features during tuning and removed one of the least important features (House Style) from the dataset which decreased RMSE



# DEMO OF ML WEB APP

<https://housing-demo-team6.herokuapp.com/>



# MODEL LIMITATIONS/GENERALIZATION ISSUES

- The current data source was more specific to one region and was lacking key features such as School District/School Rating, Future Development (Commercial Complex, Highways etc.), Crime rate
- Our project shows that it is not possible to train a housing model on only one city and generalize it to others
  - Real Estate values are highly dependent upon location and geographical region
- To build a highly predictive model, multiple geographical regions are needed to train model.
- Next Steps:
  - Interview Real Estate Agents to value the area based on
    - School District/School Rating
    - Future Development (Commercial Complex, Highways etc.)
  - Find proximity to amenities



# APPENDIX

## Github:

- <https://github.com/bmarotical/capstoneTeam6>

## Dataset:

- <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

## CPI:

- <https://github.com/datadesk/cpi>

## RapidApi (datasets):

- <https://rapidapi.com/datascraper/api/us-real-estate/>