Capstone Project

Housing Price Prediction Using Advanced Regression Techniques

next slide →

By Joy Iwunor

Capstone project

Table of Contents

next slide →

03

Capstone Project

Introduction

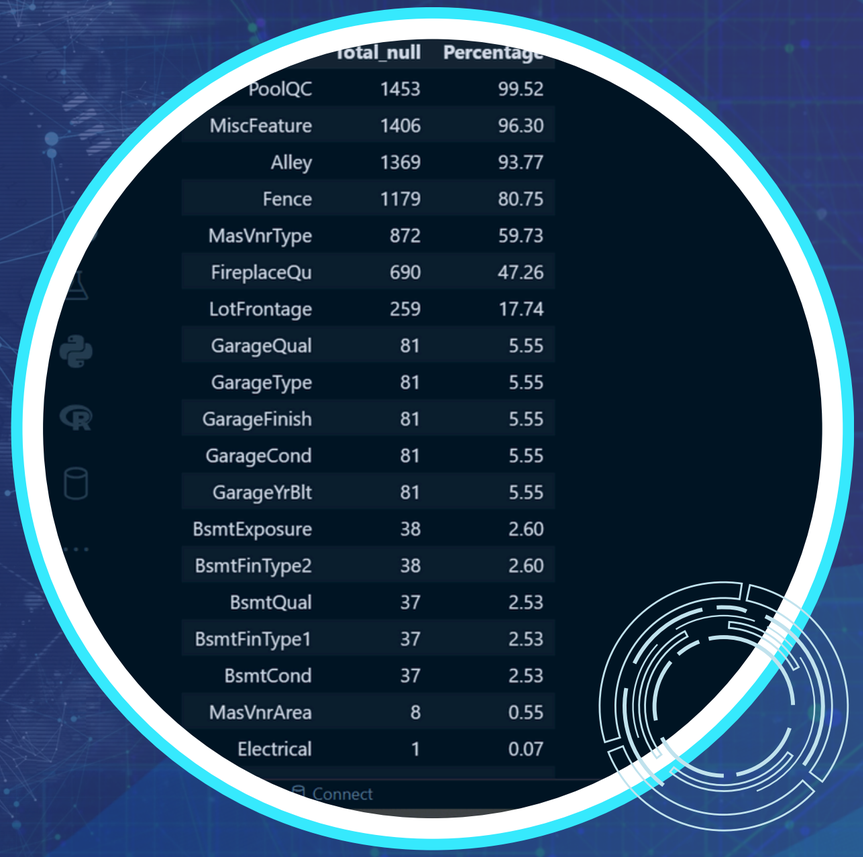
In the complex world of real estate, accurate house price prediction is a valuable tool for buyers, sellers, investors, and market analysts alike. This study focuses on the housing market in Ames, Iowa, utilizing a comprehensive dataset that captures the essence of residential properties through 80 distinct explanatory variables. These variables offer an in-depth look at almost every conceivable aspect of homes in the area, providing a root for the analysis.

next slide →

Data Cleaning

Dropped columns with high and moderate missing values from 10% above such as *PoolQC, MiscFeature, Alley, Fence, MasVnrType, FireplaceQu, LotFrontage*

Filled Low Percentage Missing Values such as. GarageQual, GarageType, GarageFinish, GarageCond, GarageYrBlt, BsmtExposure, BsmtFinType2, BsmtQual, BsmtFinType1, BsmtCond, MasVnrArea, Electrical.



	Total_null	Percentage
PoolQC	1453	99.52
MiscFeature	1406	96.30
Alley	1369	93.77
Fence	1179	80.75
MasVnrType	872	59.73
FireplaceQu	690	47.26
LotFrontage	259	17.74
GarageQual	81	5.55
GarageType	81	5.55
GarageFinish	81	5.55
GarageCond	81	5.55
GarageYrBlt	81	5.55
BsmtExposure	38	2.60
BsmtFinType2	38	2.60
BsmtQual	37	2.53
BsmtFinType1	37	2.53
BsmtCond	37	2.53
MasVnrArea	8	0.55
Electrical	1	0.07

Connect

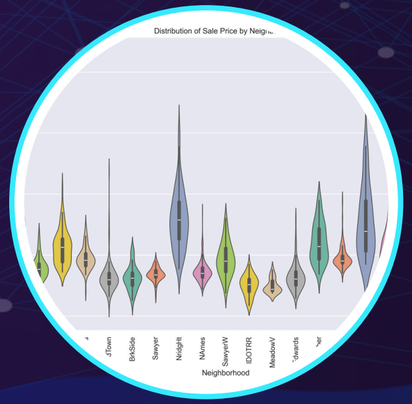
next slide →

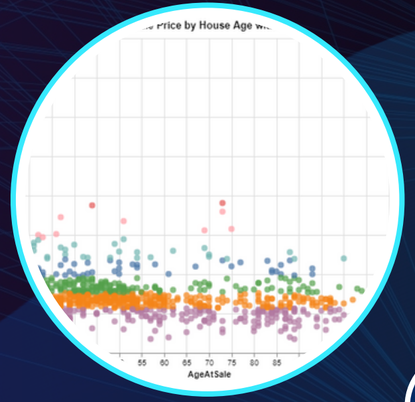
05

Capstone Project

Exploratory Data Analysis (EDA)

EDA ensures a thorough understanding of the dataset, identifies potential issues, and provides a solid foundation for further analysis or modeling. This process involves univariate analysis, bivariate analysis, multivariate analysis, future engineering and handling outliers.



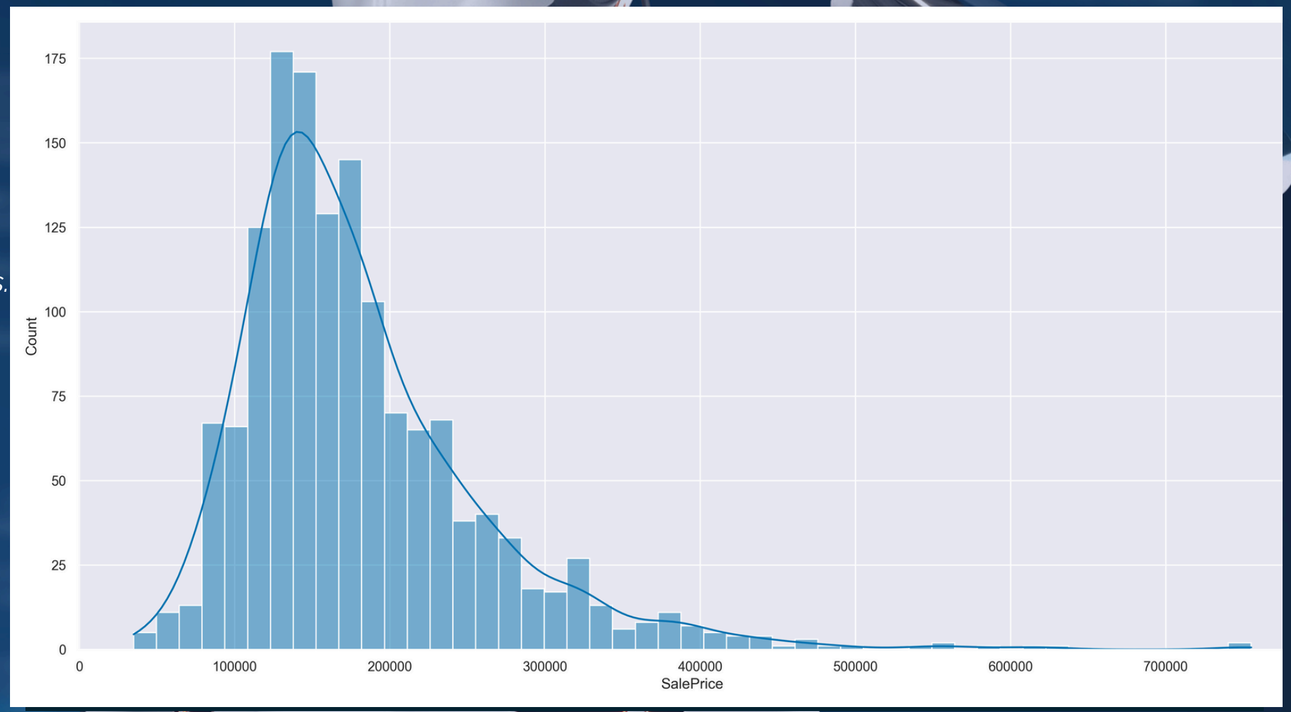


next slide →

Capstone Project

Univariate

Right-Skewed Distribution:
The distribution of SalePrice is skewed to the right, meaning that there are more houses with lower prices and fewer houses with very high prices.
Central Tendency:
The peak of the distribution occurs around the 150,000dollars - 200,000dollars range, which suggests that the majority of house prices fall within this range.
The tail on the right indicates the presence of outliers—houses that are significantly more expensive than the majority. These are the high-priced homes that are not as common but still exist in the dataset.



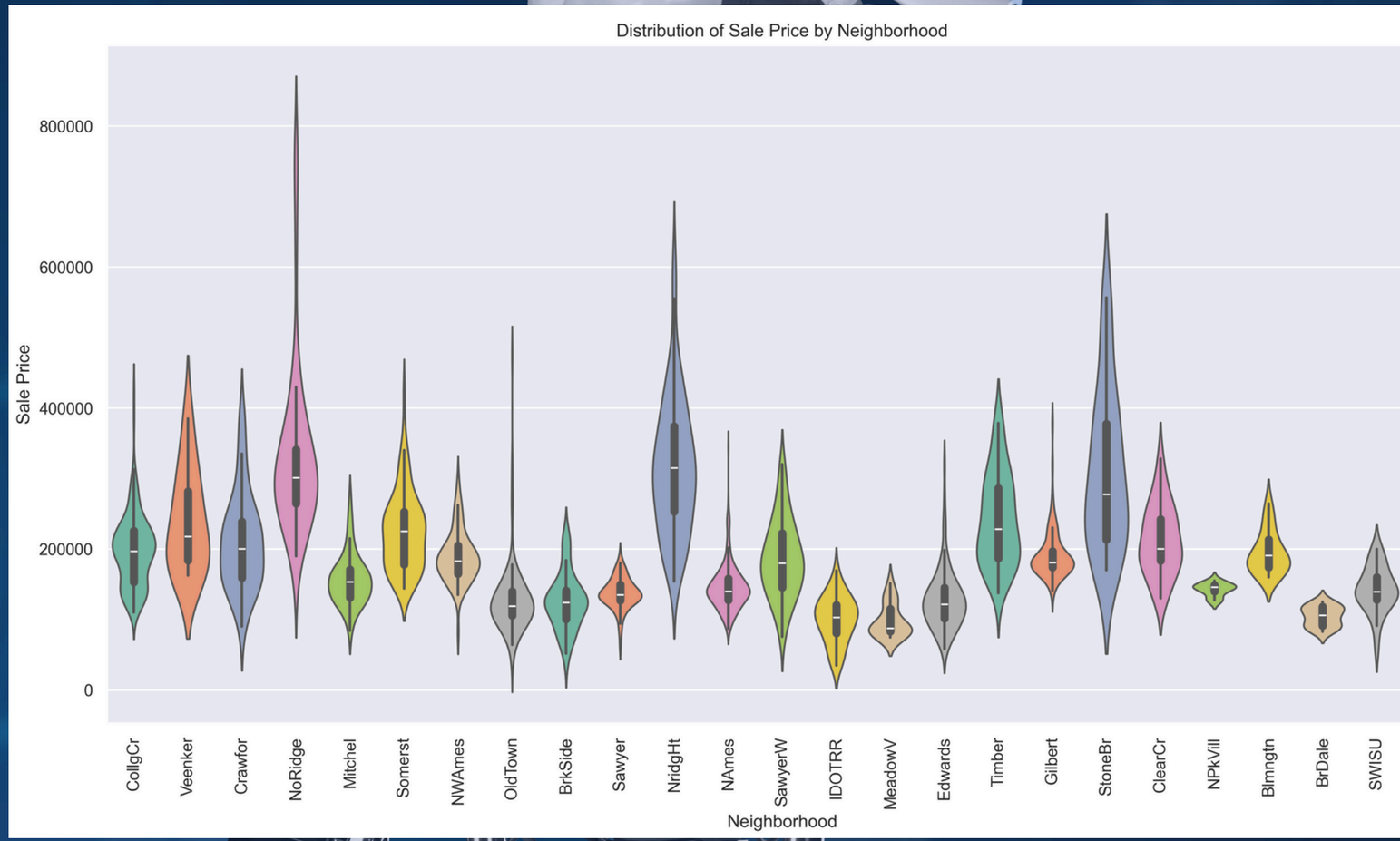
next slide →

Capstone Project

Bivariate

The distribution of sale prices varies significantly across different neighborhoods. NoRidge, StoneBr, and NridgHt have the highest sale prices, with median values well above 400,000. On the other hand, neighborhoods like BrDale, SWISU, and IDOTRR have much lower sale prices, with medians under 200,000

Affluent Areas: Neighborhoods like NoRidge, StoneBr, and NridgHt are more affluent with higher property values, as seen from their higher median sale prices and wider distributions.
More Affordable Areas: Neighborhoods like BrDale, SWISU, and MeadowV are more affordable, with lower median prices and tighter distributions, indicating less variance in property values.



next slide →

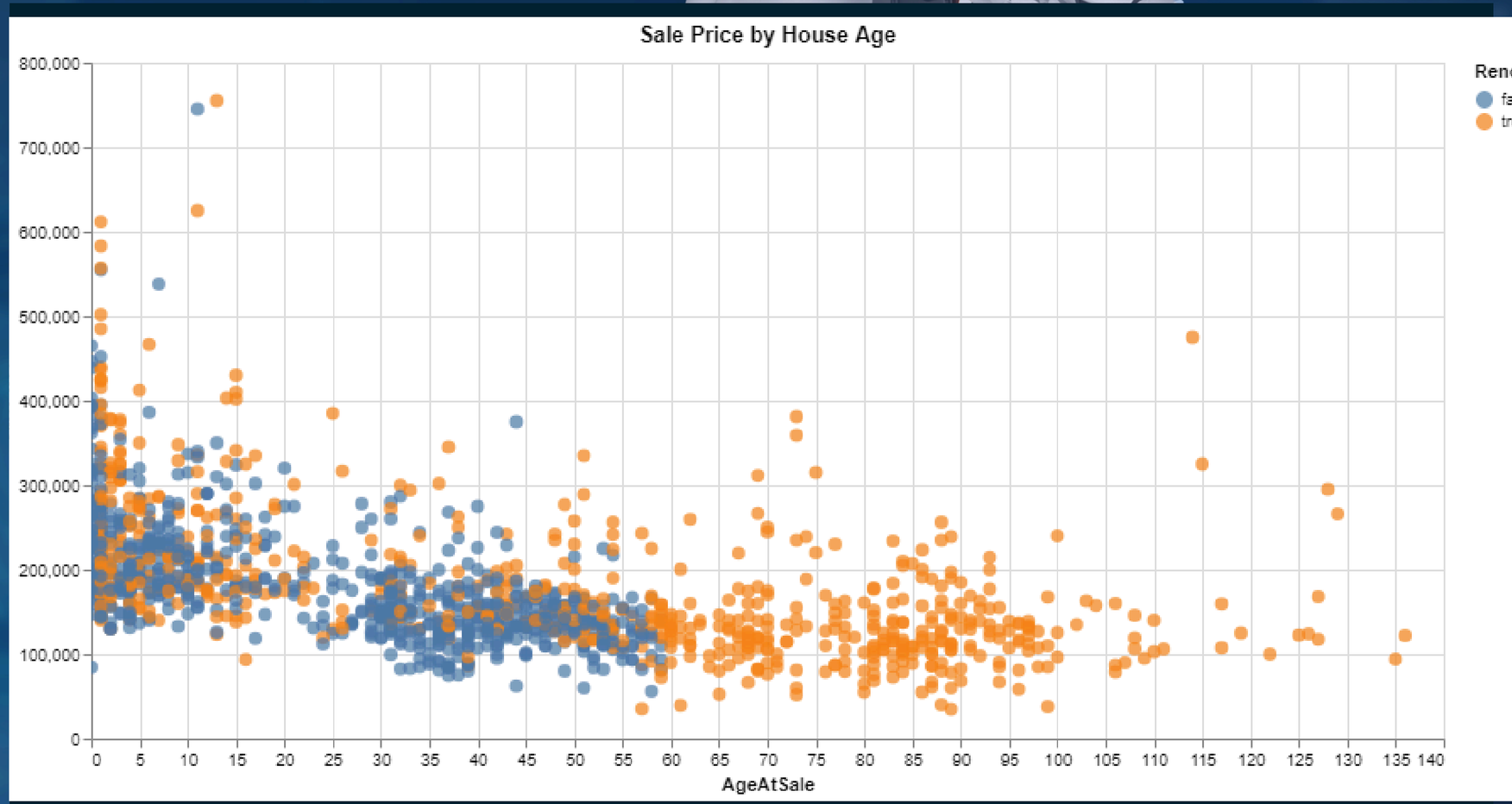
Capstone Project

Multiivariate

There is a downward trend in sale price as the age of the house increases, suggesting that newer homes tend to sell for higher prices compared to older ones.

Market Value Trends:
Newer homes generally command higher prices, but strategic renovations can help maintain property values over time..

Renovation Significance:
Renovation appears to significantly mitigate the effect of aging on property value, helping older homes retain or even increase their market value.



next slide →

09

Capstone project

Model Performance

Models	RMSE	R^2	Adj-R^2
LR	35220.343457	0.838276	0.822409
RF	25208.334565	0.917153	0.909025
XGB	20384.374629	0.945827	0.940512
GB	20326.376625	0.946135	0.940850

next slide →

10

Conclusion

This study aimed to predict housing prices using advanced regression techniques, incorporating careful feature selection through exploratory data analysis (EDA). I implemented and compared several models including Linear Regression (LR), Random Forest (RF), XGBoost (XGB), and Gradient Boosting (GB).

Among the techniques employed, Gradient Boosting (GB) emerged as the most accurate model for predicting housing prices in our dataset. This aligns with GB's known strengths in handling complex, non-linear relationships and its ability to capture subtle interactions between features.