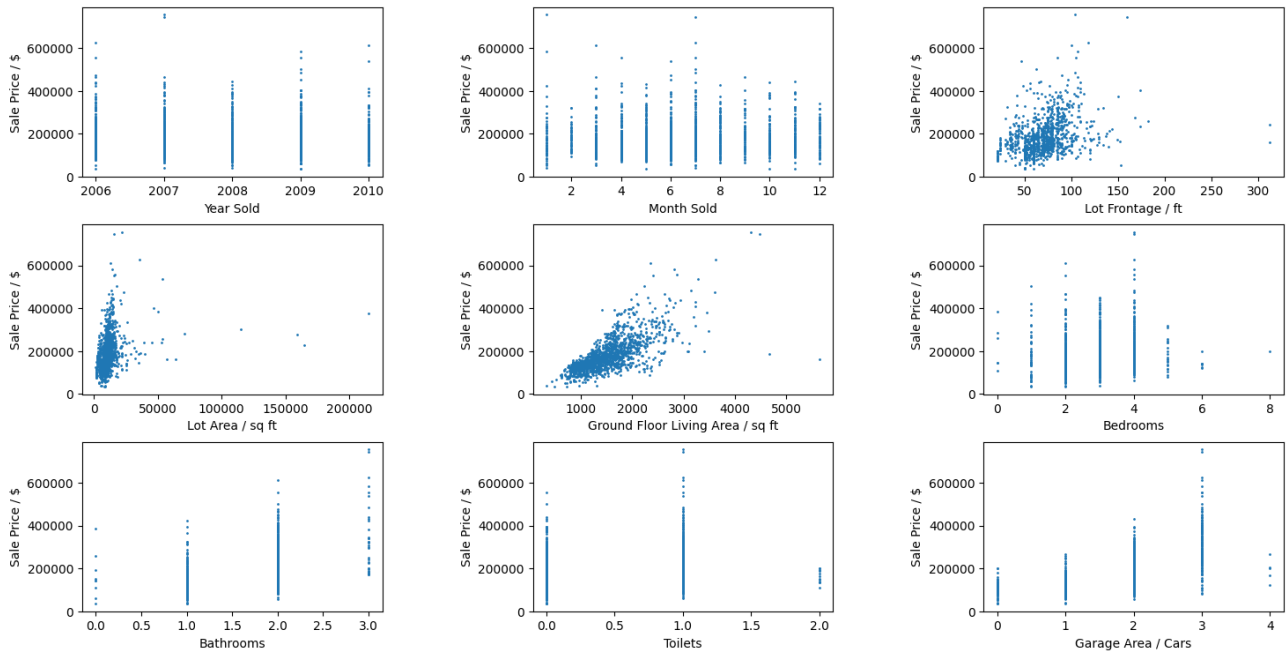
House Price Prediction

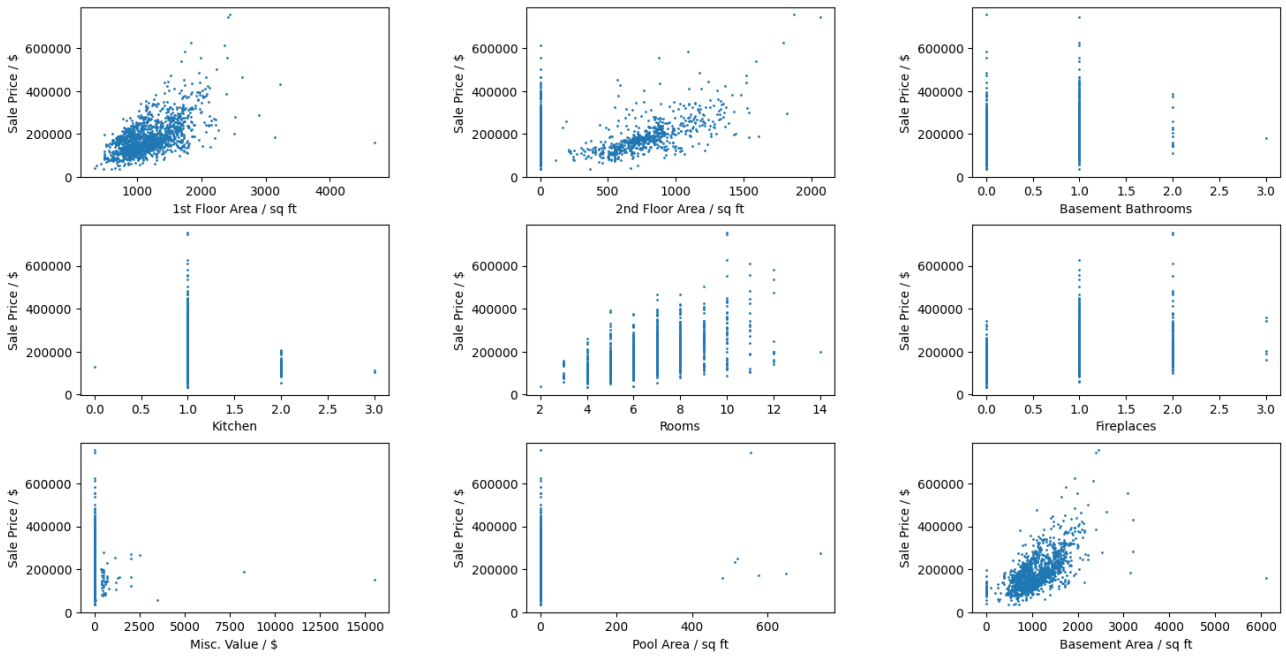
Predict the price of houses in Iowa from a large set of categorical and numerical features.

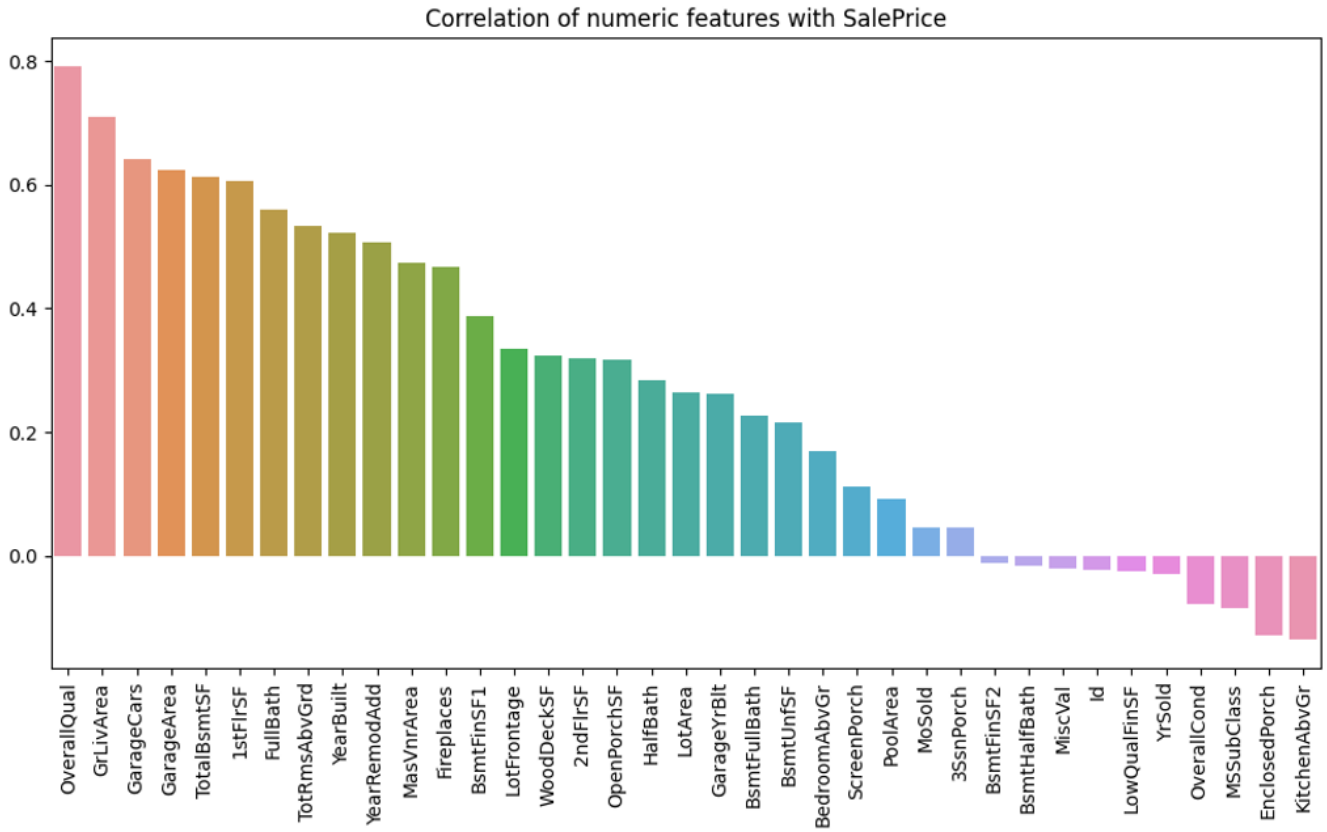
Learning Objectives

- Data correlation
- One-hot encoding
- Data bucketing
- Normalising data
- Handling outliers
- Cross-validation and hyperparameter tuning
- Gradient-boosted linear regression

Visualising the Data



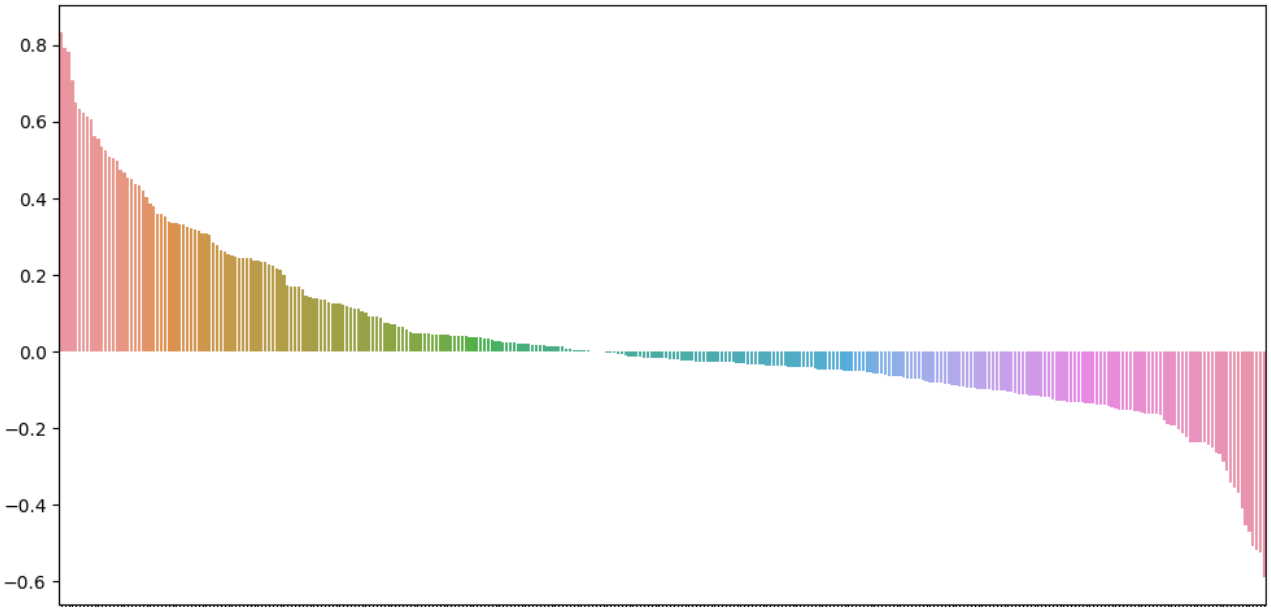

Sale Price / $
1st Floor Area / sq ft
2nd Floor Area / sq ft
Basement Bathrooms
Kitchen
Rooms
Fireplaces
Misc. Value / $
Pool Area / sq ft
Basement Area / sq ft


Correlation of numeric features with SalePrice
OverallQual
GrLivArea
GarageCars
GarageArea
TotalBsmtSF
1stFlrSF
FullBath
TotRmsAbvGrd
YearBuilt
YearRemodAdd
MasVnrArea
Fireplaces
BsmtFinSF1
LotFrontage
WoodDeckSF
2ndFlrSF
OpenPorchSF
HalfBath
LotArea
GarageYrBlt
BsmtFullBath
BsmtUnfSF
BedroomAbvGr
ScreenPorch
PoolArea
MoSold
3SsnPorch
BsmtFinSF2
BsmtHalfBath
MiscVal
Id
LowQualFinSF
YrSold
OverallCond
MSSubClass
EnclosedPorch
KitchenAbvGr

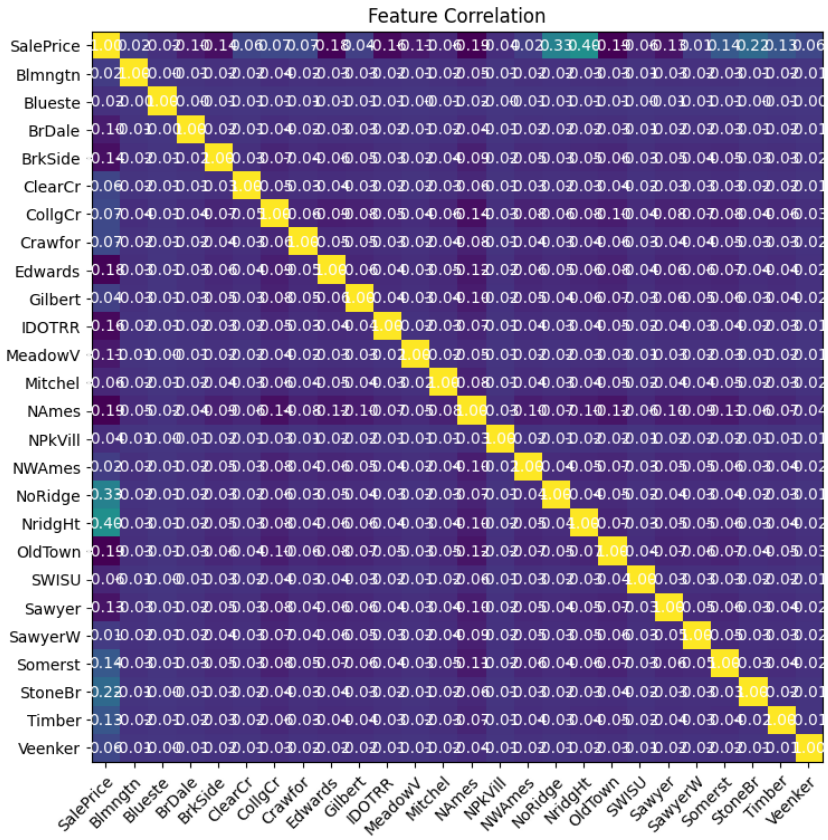
Handling Categorical Variables

All categorical variables were One-Hot Encoded. Any categoricals that were present in the training dataset but not in the testing dataset (or vice versa) were not included.

Correlation of Categorical Features with SalePrice



The 25 neighbourhoods were bucketed into groups of two or three to increase the number of occurrences of each important feature and reduce overfitting. The buckets were chosen to minimise the variance in the correlations with the sale price.



Feature Engineering

Several new features were created from the existing features, aiming to maximise their correlation with the sale price. These are summarised below

Feature	Construction	Correlation
IsNew	Is this a newly built house?	0.3575
Is2Stories	Is this a two-storey house?	0.2429
Is2Fam	Is this a two-family house?	–0.0973
TotalSF	TotalBsmtSF + 1stFlrSF + 2ndFlrSF	0.7823
TotalBathrooms	FullBath + BsmtFullBath + 0.5x(HalfBath + BsmtHalfBath)	0.6317
HouseAge	YrSold – YearBuilt	–0.5234
RemodAge	YrSold – YearRemodAdd	–0.5091
QualityArea	OverallQual x GrLivArea	0.8321

Normalising Data

The house prices in the training dataset were not normally distributed. This made fitting the linear regression model more difficult and produced less accurate predictions. The house prices were transformed into a normal distribution by taking the natural log of each house price. This reduced the skew in the data, lessening the effect of outliers with very high or very low house prices.

Handling Outliers

As can be seen in the data visualisation, many features that were strongly correlated with sale price also had several significant outliers.

For example, the house price increases with the number of bedrooms until there are 4, after which it stagnates. Similarly, the house price increases with the size of the garage until it fits 3 cars, after which it stagnates. To handle these outliers, values outside the linear region were snapped back to this edge value (e.g. 4 bedrooms or 3 cars).

Some houses were extreme outliers in their Lot Frontage, Lot Size, Ground Floor Living Area, 1st Floor Area, etc, relative to their sale price. In the training data, these houses were removed manually to provide a better fit.

Hyperparameter Tuning

A gradient-boosted linear regression model, XGBoost, was chosen for this prediction task. Its hyperparameters were chosen by random sampling of the hyperparameter-space with 5-fold cross-validation. 50 different sets of hyperparameters were tested from the following options:

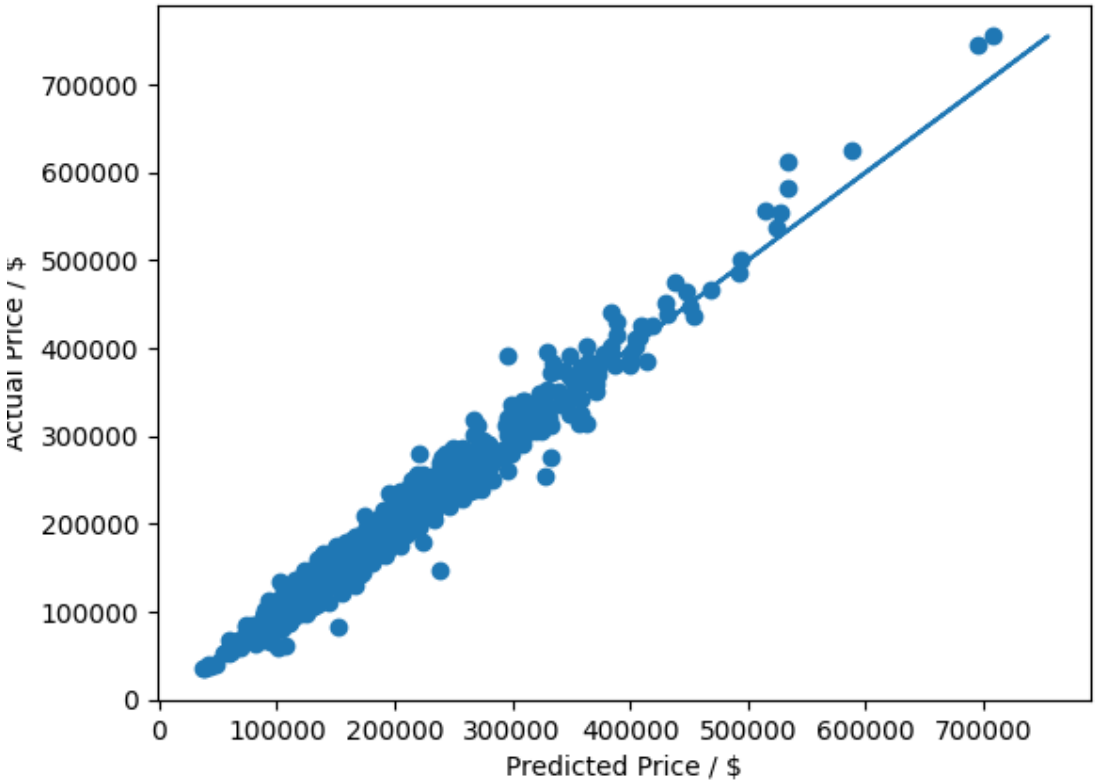
n_estimators: 100, 200, 300, 500, 800, 1000
max_depth: 2, 3, 4, 5
subsample: 0.4, 0.6, 0.8, 1.0
learning_rate: 0.05, 0.01, 0.1, 0.2
colsample_bytree: 0.4, 0.6, 0.8, 1.0

The highest scoring set of hyperparameters was:

n_estimators: 300, max_depth: 4, subsample: 0.8, learning_rate: 0.05, colsample_bytree: 0.8

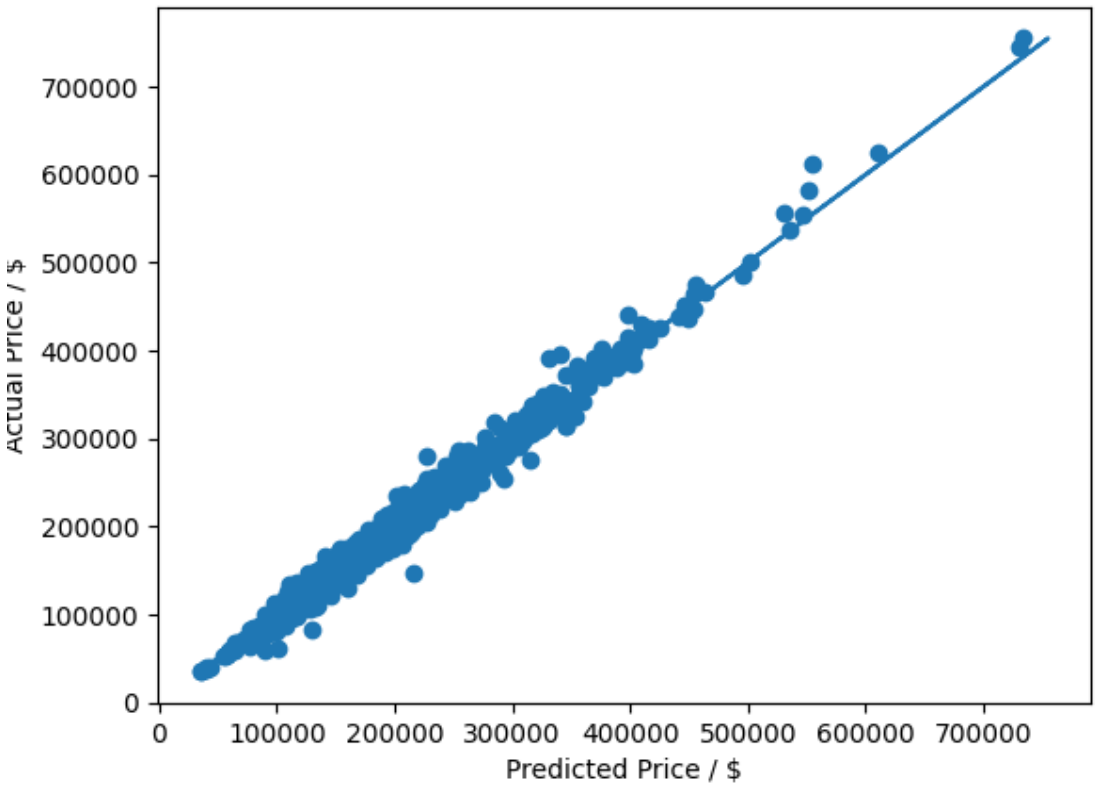
This model achieved a loss of 0.1261 on the test dataset.

House Price Prediction Performance of the Final XGBoost Model on the Training Dataset

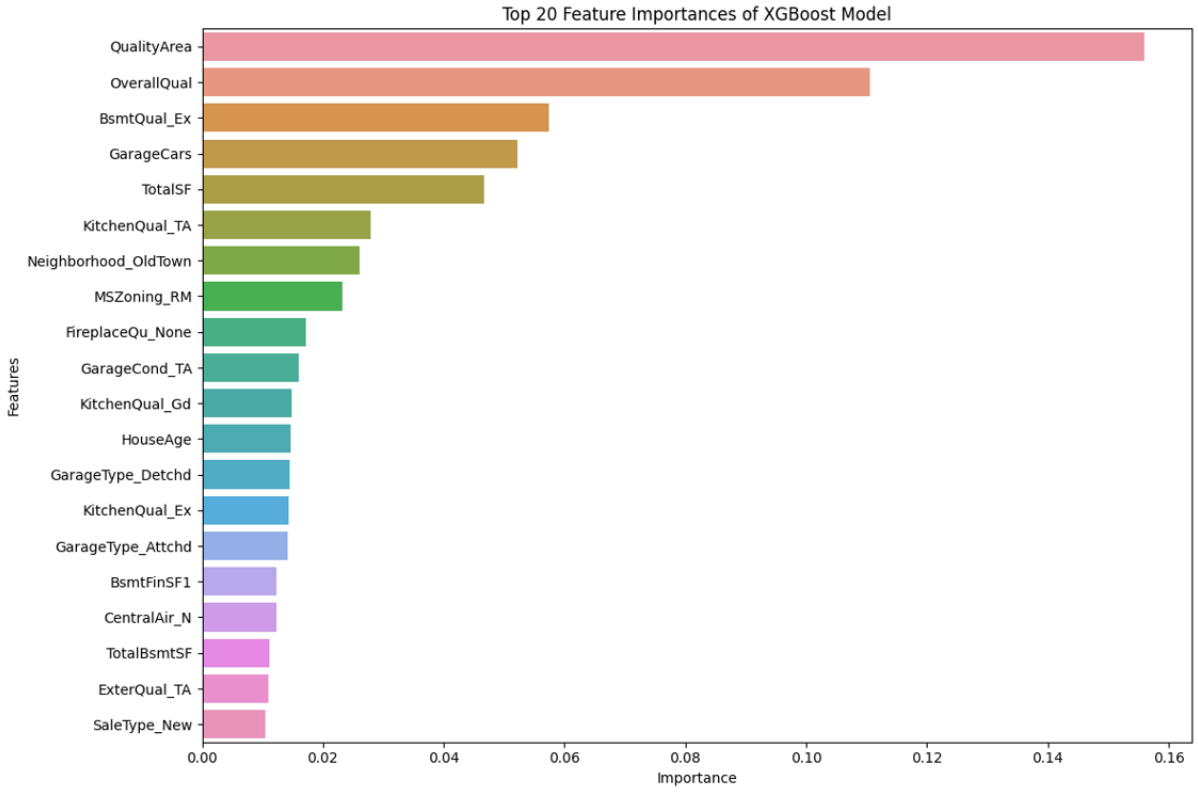


The 5-fold cross-validation step was important to prevent overfitting on the training dataset. To demonstrate this, hyperparameter tuning was performed without cross-validation, producing a model that gave a closer fit to the training dataset but a much greater loss of 0.3405 on the test dataset.

House Price Prediction Performance of the Overfitted Model on the Training Dataset



Feature Importance



Conclusion

This project successfully predicted house prices in Iowa using a pipeline of data preprocessing, feature engineering, and hyperparameter tuning. Normalising skewed data, handling outliers, and encoding categorical variables made the dataset more suitable for linear regression. Feature engineering significantly improved the model's performance by introducing new features that were highly correlated with sale price, such as TotalSF and QualityArea. A gradient-boosted XGBoost model was selected for its robustness and accuracy, and its performance was optimised through extensive hyperparameter tuning with 5-fold cross-validation. The final model achieved a strong test loss of 0.1261, demonstrating its strong ability to generalise trends and make accurate predictions.