HOA Fee Classification and Prediction Model

Presented by Anny Carella Irumva

Agenda

1. Business Problem
2. Data Understanding
3. Data Preparation
4. Modeling
5. Conclusion

Business Problem

- Homeowners often struggle with the decision of whether to join a Homeowners Association (HOA) due to uncertainty about the potential fees and benefits. Many lack information on the costs involved, making it difficult to make informed choices

- Similarly, prospective buyers may want to know if a property will incur HOA fees before purchasing. Therefore, there is a need for a predictive model that can determine if a property will likely have HOA fees and provide an estimate of the potential costs

Business Problem

The primary goal of the project is to develop a machine learning model that can:

- Predict whether a property will have HOA fees based on its features, location, and other relevant data.
- Estimate the potential HOA fee amount for properties that are predicted to have HOA fees

This model could be especially useful for existing homeowners who currently do not pay HOA fees but are considering joining an HOA in their neighborhood

Data Understanding

Dataset: Real Estate Data with 5000 entries and 16 features of various property attributes

Features Used:

- Property characteristics like square footage, lot size, number of bedrooms, bathrooms, etc.
- Geographic location data: Latitude, Longitude.

Target variables:

- Category: Whether HOA fees exist or not (0 = No Fee, 1 = has Fee)
- HOA: The actual HOA fee amount for properties with HOA fees

Data Preparation

01

Missing Value Treatment:

- HOA fees imputed with 0 for missing values
- Numeric features imputed with median values
- Categorical features imputed with mode values

02

Feature Engineering:

Created a binary target for HOA fees (No Fee vs. Fee)

03

Other Steps:

- Used IQR method to filter out outliers for specific numerical features
- Applied logarithmic and square root transformations to numeric features
- Label encoding for categorical features
- Scaling numerical features
- Did an oversampling of my target variable
- Did a feature selection for every model

Modeling: Classification

Objective: Classify if a property will have HOA fees or not

Algorithm used:

- Naive Bayes
- Gaussian Bayes
- KNN (k= 1)

	Naive Bayes	Gaussian Bayes	KNN
Accuracy	0.64	0.63	0.93

Modeling: Classification

Objective: Classify if a property will have HOA fees or not

Algorithm used:

- Naive Bayes
- Gaussian Bayes
- KNN (k= 1)

	Naive Bayes	Gaussian Bayes	KNN
Accuracy	0.64	0.63	0.93

Modeling: Regression

Objective: Predict the actual HOA fee amount

Algorithm used:

- KNN regressor (k=2)
- Multivariate Linear Regression

	KNN Regressor	Linear Regression
RMSE	45.8	53.8

Modeling: Regression

Objective: Predict the actual HOA fee amount

Algorithm used:

- KNN regressor (k=2)
- Multivariate Linear Regression

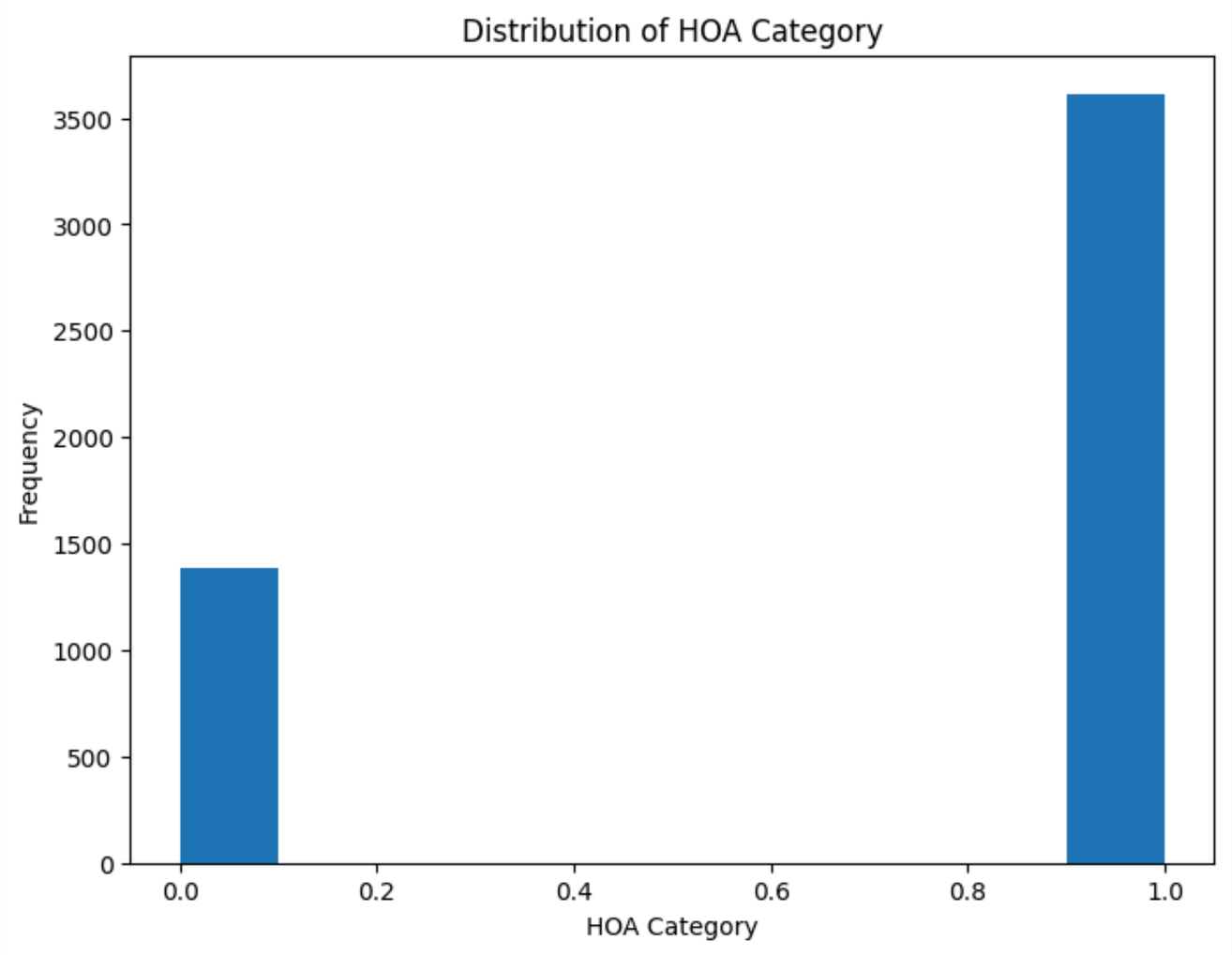
	KNN Regressor	Linear Regression
RMSE	45.8	53.8

Conclusion

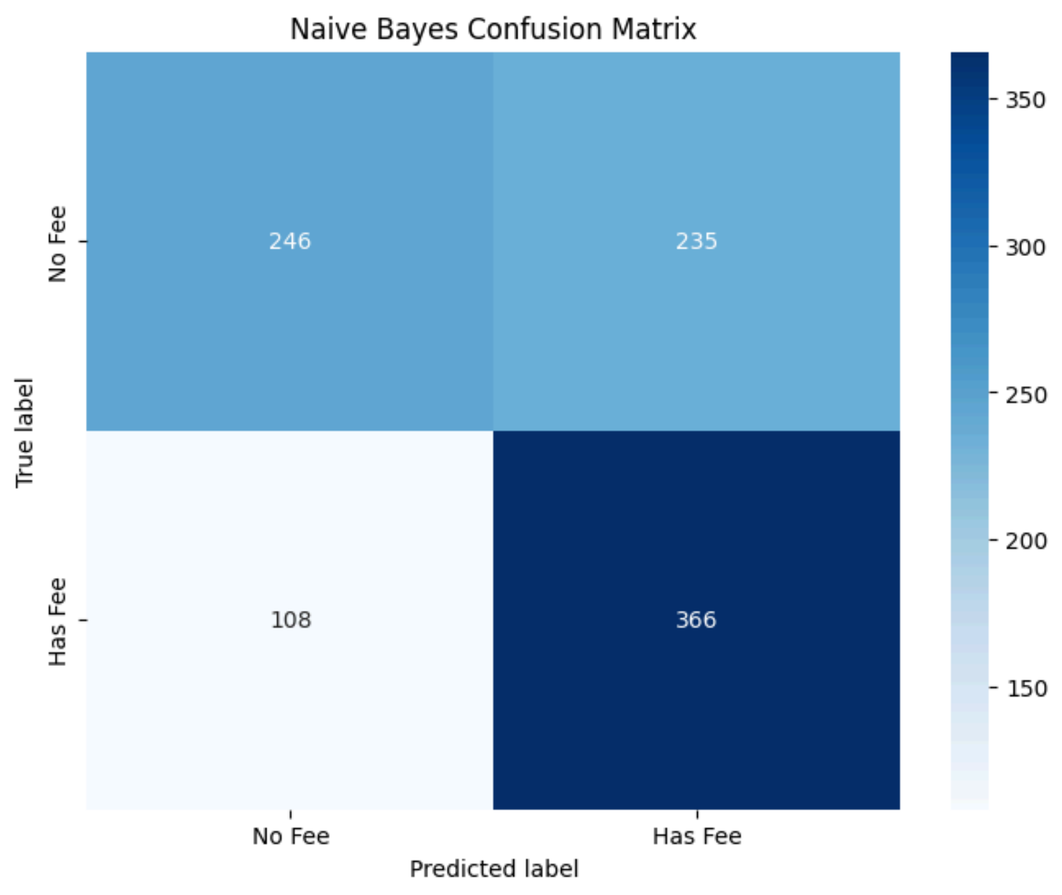
- The objective was to build a machine learning model that predict whether a property will have Homeowners Association (HOA) fees and estimate the potential fee amount
- Utilized the Raw House data with various property attributes
- Tried 3 classification models and 2 regressors
- **KNN** models achieved the highest accuracy and MAPE, making it the best choice for fees classification and prediction in this context
- This solution is useful for homeowners considering joining an HOA and can be integrated into real estate platforms to enhance the user experience by providing HOA fee insights

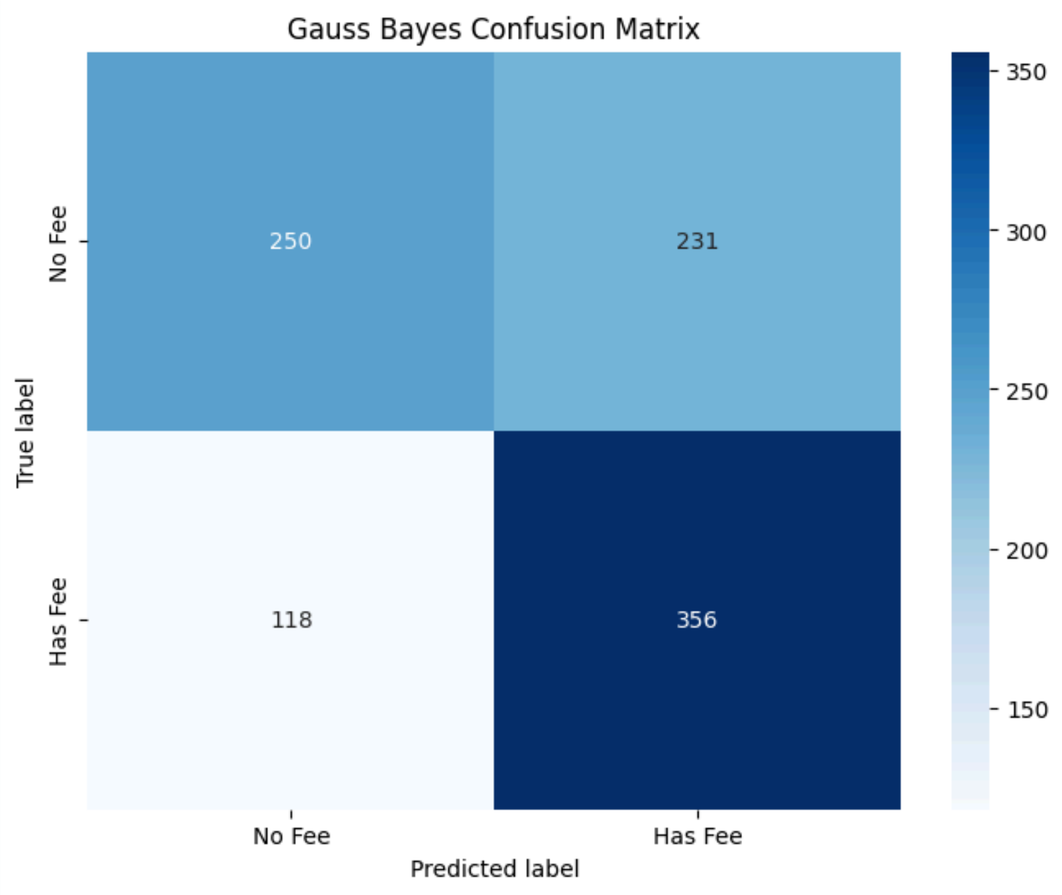
Thank You

Appendix

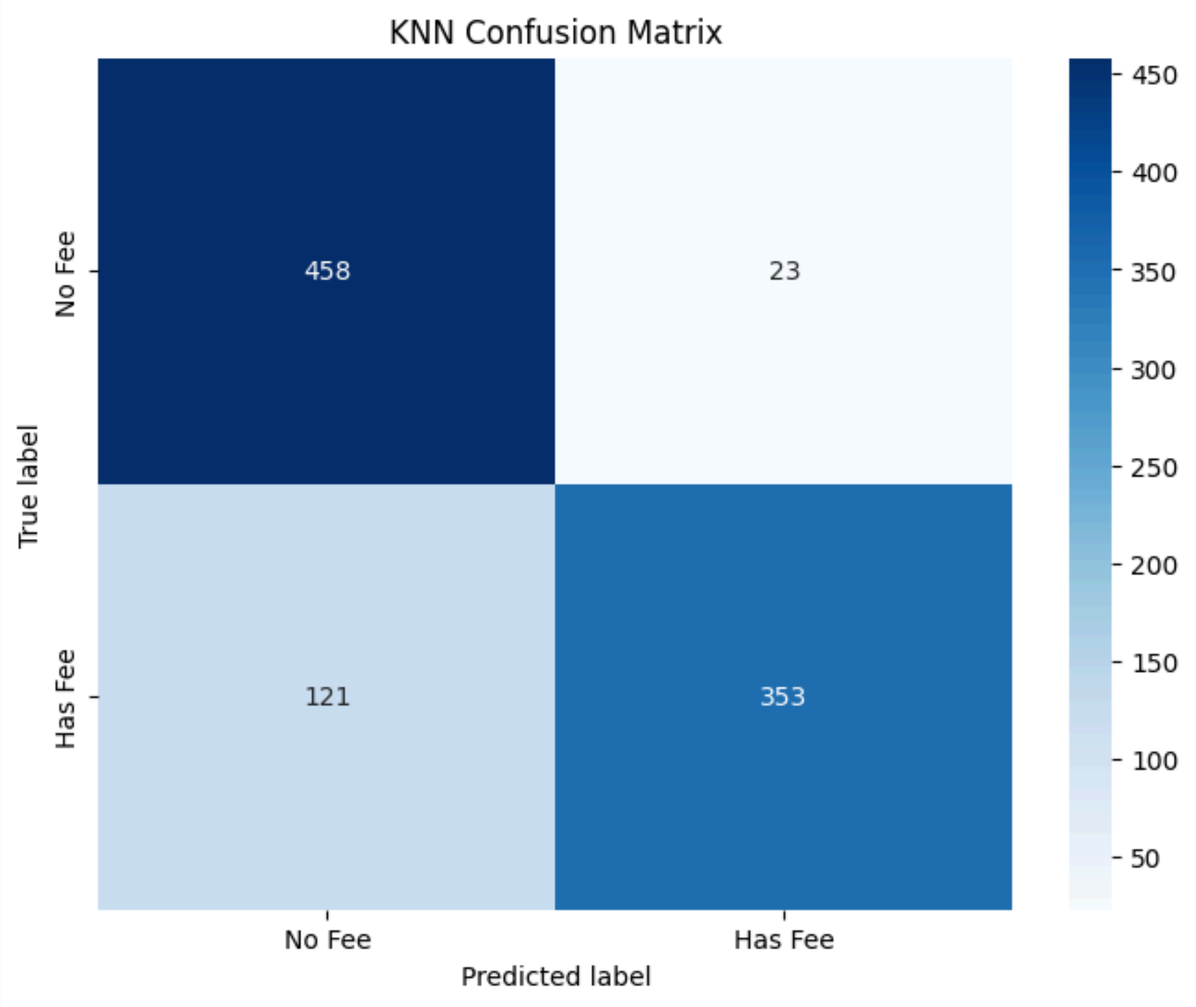

Distribution of HOA Category
Frequency
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
HOA Category

Appendix


Naive Bayes Confusion Matrix
246
235
108
366
No Fee
Has Fee
True label
No Fee
Has Fee
Predicted label
350
300
250
200
150


Gauss Bayes Confusion Matrix
250
231
118
356
No Fee
Has Fee
True label
No Fee
Has Fee
Predicted label
350
300
250
200
150

Appendix


KNN Confusion Matrix
458
23
121
353
No Fee
Has Fee
True label
No Fee
Has Fee
Predicted label
450
400
350
300
250
200
150
100
50