PREDICTING HOUSE PRICES

USING MACHINE LEARNING TECHNIQUES

By: Sumana Chilakamarri, Ayesha Uddin, & Yash Gupta

Main topic: Use Machine Learning Techniques to analyze key features of house pricing such as location, size, amenities, etc.

Problem Statement

01 **Traditional Pricing Methods** → Are subjective and do not capture complex relationships between features

02 **Developing a model that** → identifies which features most impact pricing

Our goal for this project is to develop a scalable model that identifies which features most impact pricing and supports data-driven decision making for buyers, sellers, and investors

Methodology

Data Preprocessing

- Used an open-source dataset on Kaggle

- Removing Outliers, Filling In missing values
  - OneClassSVM(), StandardScaler()

- PCA, SMOTE
  - Limited to 75 variables

- Two data files for our two types of methods

BX
SalePrice
208500
181500
223500
140000
250000
143000
307000
200000
129900
118000
129500
144000
157000

VS.

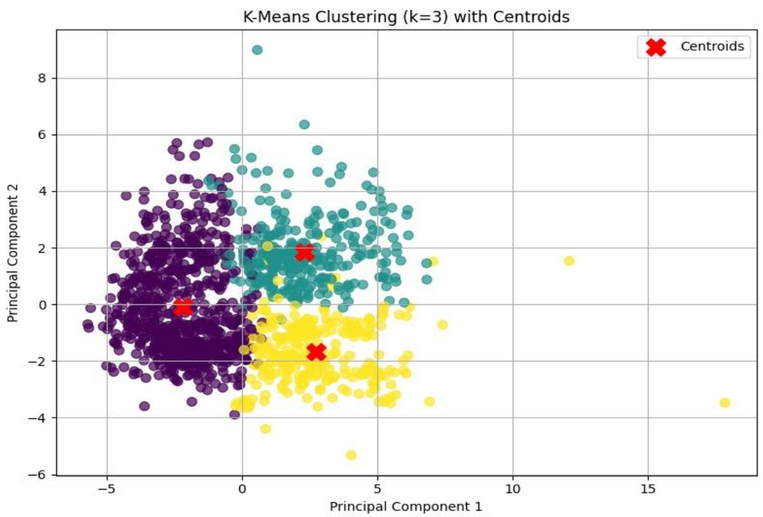
BX
PriceCategory
1
1
1
0
1
0
1
1
0
0
0
0

Models Used

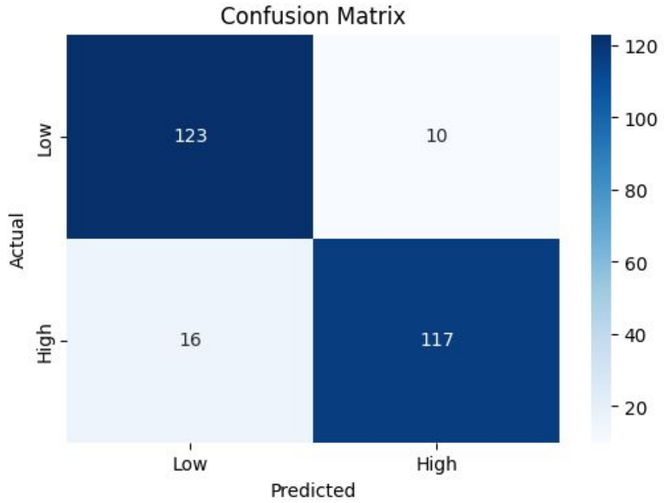
- K- Means Clustering
- Random Forest Classification

Evaluation & Results

K-Means Clustering:



- K-Means clustering on PCA-reduced data shows generally well-separated clusters
- Some overlap— between green and yellow—may reflect similarity or PCA compression effects.



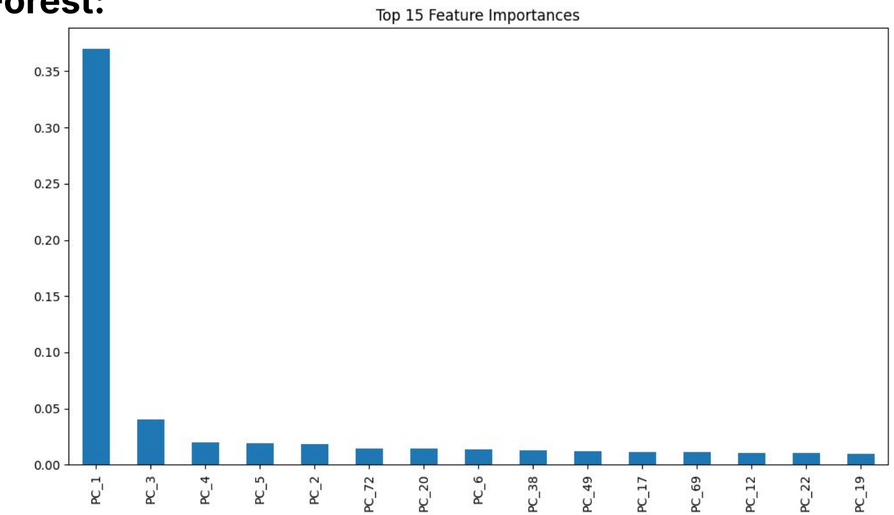
Accuracy: ~90.5%

Precision: 92.9%

Recall: 87.9%

Strong & **balanced performance** with few misclassifications

Random Forest:



- PC1 is by far the most important feature in the Random Forest model
- Most other PCs contribute minimally, with a few later components showing modest predictive value.