

DBC Final Presentation

Charlie, Ryan, and Zach

The Objective

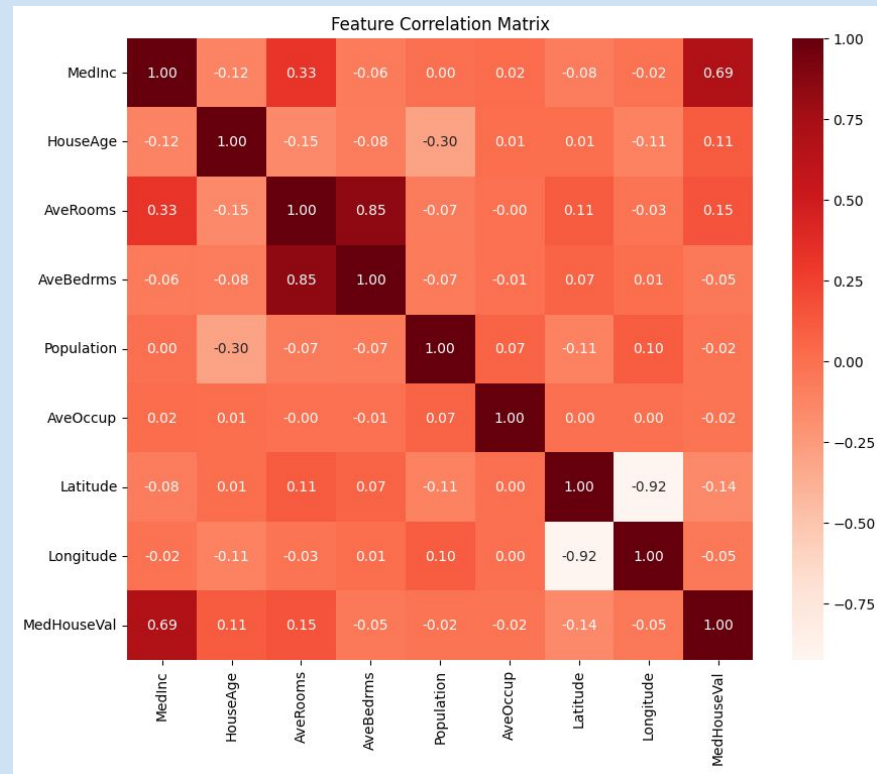
- Goal: To build a predictive model that learns parameters to map input features to a specific variable
 - Target Variable: MedHouseVal (Median House Value)
- Approach: A regression task utilizing historical data to predict continuous values
- Why it Matters: Understanding the drivers of housing value bolsters investment strategy, urban planning, and economic forecasting

Data Composition

- Dataset Used: California Housing Dataset
- Volume: 20,640 records with 9 attributes
- Key Features (Inputs):
 - Demographic: Median Income (MedInc), Population, Occupancy
 - Structural: House Age, Average Rooms, Average Bedrooms
 - Geographic: Latitude, Longitude
- Implication: The dataset mixes physical property attributes with socio-economic context, allowing us to test which category drives value more efficiently

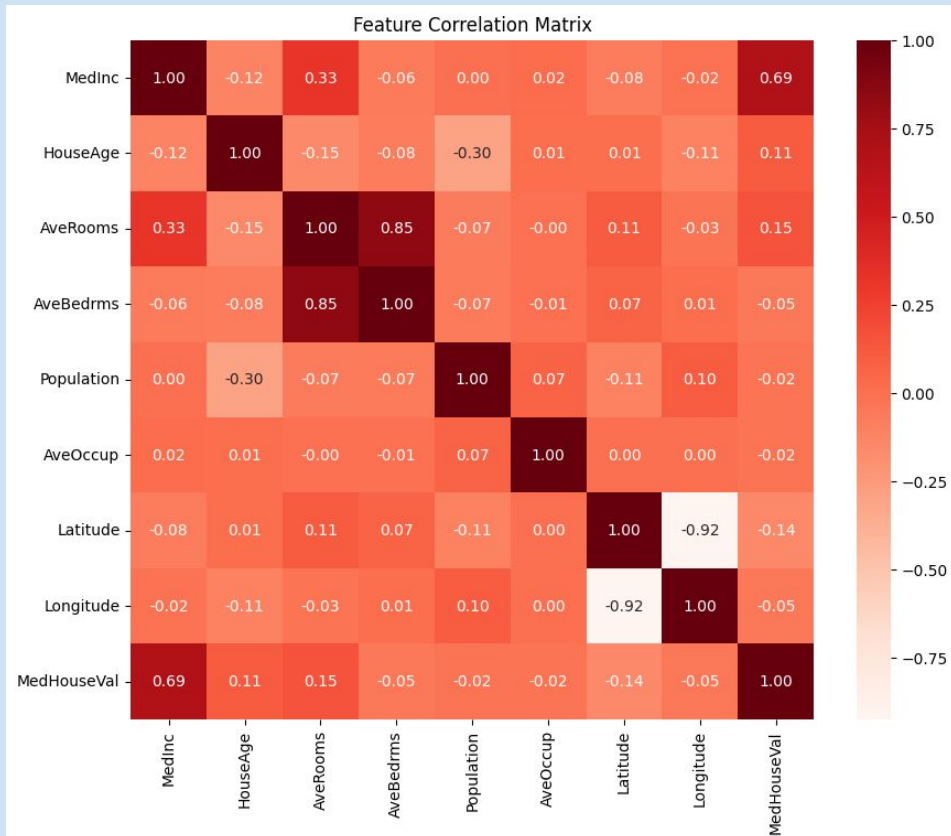
Exploratory Analysis - Geographic Value

- Observation: High-value homes are not randomly distributed; they are tightly clustered in specific latitude/longitude regions (like coastal/urban centers)
- Implication: Location holds true - the heatmap proves that locational features are critical predictors
 - Models that ignore geospatial context will likely fail to understand the depth of this market



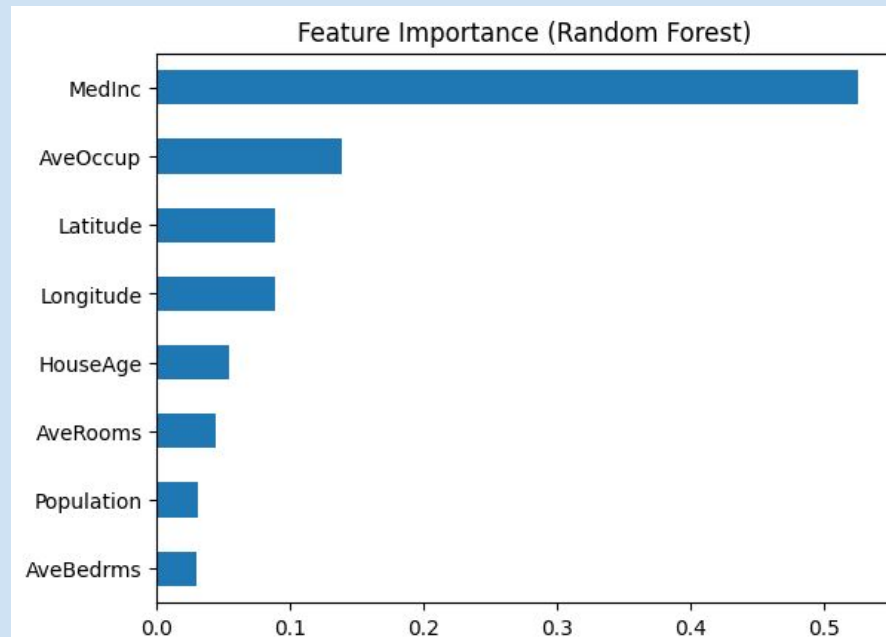
Exploratory Analysis - Income Correlation

- Key Finding: MedInc (Median Income) has the strongest positive correlation with the House Value
- Secondary Finding: AveRooms (Average Rooms) has a very weak correlation
- Implication: Purchasing power is a stronger predictor of home value than the physical size of the home
 - A market driven by affordability and exclusivity rather than pure utility or square footage



Model Strategy - Linear VS. Non-Linear

- Model 1: Linear Regression - A simple model establishing a direct relationship between features and value
- Model 2: Random Forest Regressor - An “overpowered” estimator that fits multiple decision trees to improve accuracy and control fitting the mode



Performance Results

- Metric Uses: R2 Score (Percentage of variation) and MSE (Mean Squared Error)
- Linear Regression:
 - MSE: 0.5559
 - R2: 0.5758
- Random Forest:
 - MSE: 0.2552
 - R2: 0.8053
- Implication: The housing market is non-linear
 - The interaction between location, age, and income cannot be captured by a single line
 - Random Forest is the superior choice

Drivers of Prediction

- Dominant Feature: MedInc is overwhelmingly the most important predictor
- Secondary Features: Location (Latitude/Longitude) and AveOccup
- Minor Features: HouseAge, AveRooms, AveBedrooms have low predictive power
- Implication: The condition of the asset (age, room count) is less relevant to valuation than the socio-economic status of the neighborhood

Strategic Implications & Risks

- The model relies heavily on income to predict value
 - Using this for future planning may reinforce economic segregation
 - As the model shows that high income equals high value overlooking up-and-coming- areas with lower current income
- While the reliability with the Random Forest is high (0.80), it is not perfect
 - Still 20% unexplained variance - suggests external factors like interest rates, schools districts, etc. are missing from the data