MET CS-699: Data Mining
Term Project

Categorization of Property Type in Housing Website

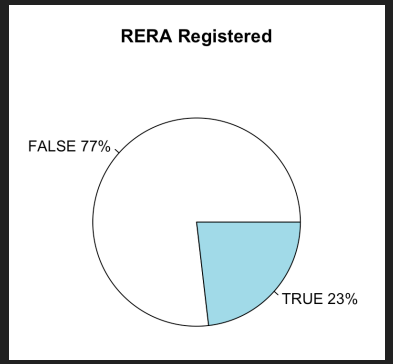
Aditya Maheshwari (U54025068)
Vaidehi Shah (U90080562)

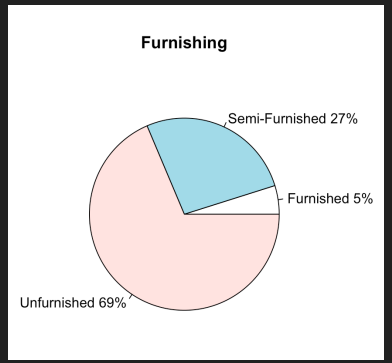
INTRODUCTION

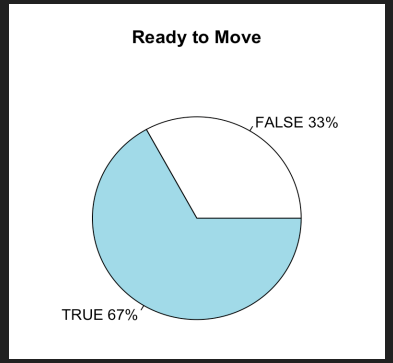
- Makaan.com is a renowned real estate portal based in India. The dataset used for this project has been prepared by scraping information off of makaan.com's website. The primary focus of this project is the listings that are stated as “sell” under the “Listing Category” attribute.

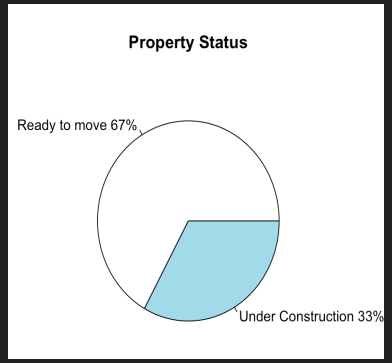
- The analysis of this project will be based on determining the type of property based on its features and characteristics. To avoid data imbalance, the categorization is done as “Apartments” and “Non-Apartments”. By addressing this, the aim is to enhance the search engine of the Makaan.com website, providing more tailored search results that are in line with customers' preferences and requirements.

CATEGORICAL VARIABLE DISTRIBUTION





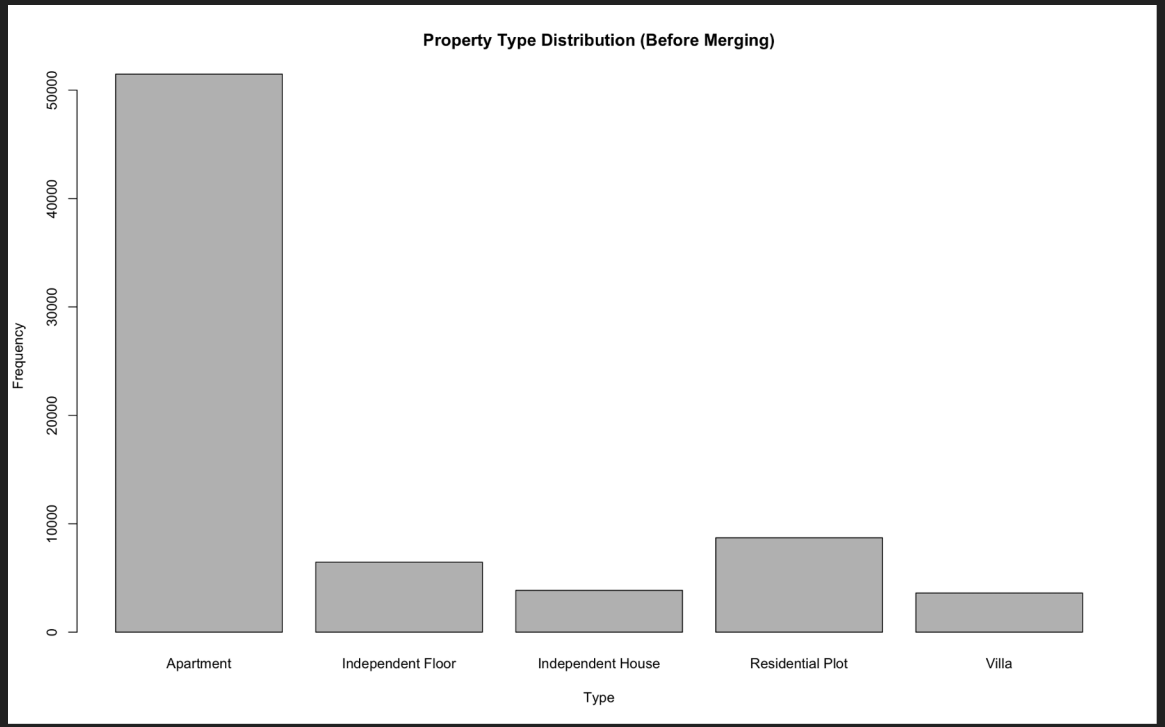




PROBLEM STATEMENT

Predicting if a property is an apartment or not is crucial for real estate. It helps investors, managers, lenders, and buyers analyze the market, manage properties, and make investment decisions. Analyzing the property's features, location, size, amenities, and demographics using machine learning algorithms can lead to informed decisions and better outcomes.

Transforming data according to the Problem Statement:



Before

After

DATA CLEANING & TRANSFORMATION

CLASSIFICATION ALGORITHMS

Classification algorithms are used to predict the class or category of a data point based on its features or attributes. We used the following five supervised machine learning classification algorithms

ATTRIBUTE SELECTION METHODS

Attribute Selection Methods involve selecting the most relevant and useful features (or attributes) from a dataset to use as input for a model. We used the following five classification algorithms:

FEATURES SELECTION METHOD	NUMBER OF SELECTED FEATURES
Random Forest Feature Selection	4
Boruta Model	12
Chi-Square	5
R-Part	10
Recursive Feature Elimination	5

PERFORMANCE METRICS - ACCURACY

Model / Method	Random Forest Feature Selection	Boruta Model	Chi Square	R-Part	Recursive Feature Elimination	All Features
Logistic Regression	83.63%	83.49%	83.49%	87.66%	84.26%	89.7%
Naive Bayes	79.25%	73.64%	73.64%	86.00%	72.18%	79.3%
K-Nearest Neighbors	85.72%	85.52%	85.52%	94.27%	88.46%	93.94%
Support Vector Machine	84.73%	84.87%	84.87%	90.46%	87.63%	90.85%
Random Forest	87.05%	95.51%	87.11%	95.21%	88.94%	95.69%

PERFORMANCE METRICS - PRECISION

Model / Method	Random Forest Feature Selection	Boruta Model	Chi Square	R-Part	Recursive Feature Elimination
Logistic Regression	0.8639	0.8871	0.8607	0.8726	0.8371
Naive Bayes	0.7569	0.8291	0.8170	0.8326	0.7578
K-Nearest Neighbors	0.8489	0.9324	0.8458	0.9371	0.8786
Support Vector Machine	0.8699	0.8973	0.8639	0.9046	0.8837
Random Forest	0.8756	0.9498	0.8765	0.9453	0.8816

PERFORMANCE METRICS - RECALL

Model / Method	Random Forest Feature Selection	Boruta Model	Chi Square	R-Part	Recursive Feature Elimination
Logistic Regression	0.7495	0.8619	0.7473	0.8305	0.7781
Naive Bayes	0.7534	0.8493	0.5753	0.8648	0.5533
K-Nearest Neighbors	0.8061	0.9235	0.8032	0.9274	0.8429
Support Vector Machine	0.7687	0.8795	0.7738	0.8677	0.8192
Random Forest	0.8131	0.9445	0.8138	0.9420	0.8522

CONCLUSION: BEST MODEL

- Looking at the average accuracy, precision, and recall values for each model, we can see that the Random Forest model consistently performs well across all attribute selection methods, followed by the K-Nearest Neighbors and Support Vector Machine models. The Logistic Regression and Naive Bayes models generally have lower average performance across all attribute selection methods. Therefore, we can conclude that the Random Forest model is the best performing model across all classifiers.

- Among the attribute selection methods, Boruta Attribute Selection takes the win with Random Forest Classification Model. It is closely competing with r-part attribute selection method which has overall higher statistics for all models except for Random Forest.

THANK YOU