**Supervised Learning – Regression**

**Business Understanding**

* **Problem Statement**

The problem we are addressing is predicting target variable based on various attributes in the California dataset. This is a supervised learning problem where the goal is to accurately predict the target variable.

* **Importance of the Problem**

Predicting this target variable is important because it helps in understanding the factors that influence the outcome. Accurate predictions can lead to better decision-making and strategic planning.

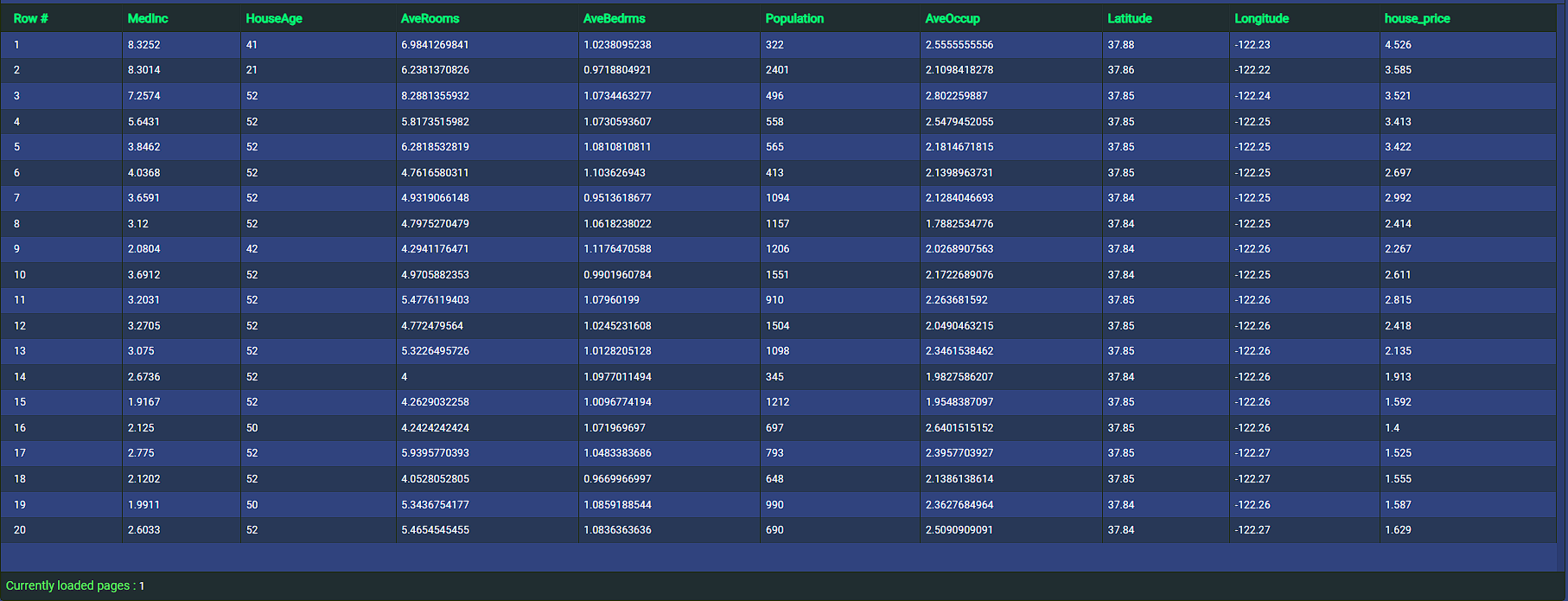
* **Data Source**

The dataset was downloaded from Kaggle. The specific dataset used is the "California dataset". The dataset is named california.csv and contains the following columns:

**Data Collection**

This dataset offers details on a range of Californian housing units.

Variables:

* MedInc: median income in block group
* HouseAge: median house age in block group
* AveRooms: average number of rooms per household
* AveBedrms: average number of bedrooms per household
* Population: block group population
* AveOccup: average number of household members
* Latitude: block group latitude
* Longitude: block group longitude

**Data Understanding**

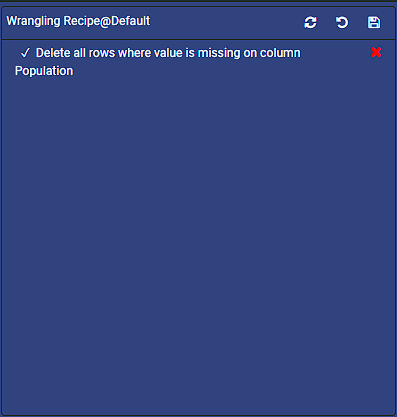
* **Exploratory Data Analysis (EDA)**

Initial data exploration reveals the following characteristics about the dataset: The dataset comprises 20,640 entries, each with 9 features.

The target variable in this context can be considered as house\_price, which we aim to understand and predict based on other housing and demographic features.

**Data Preparation**

Handling Missing Values

Missing values in the dataset were handled by deleting.

**Data Splitting**

The dataset was split into training and testing sets: 80% of the data was used for training and 20% was reserved for testing.

* 80% of the data was used for training.
* 20% of the data was reserved for testing.

**Methodology**

**Model Selection**

For this task, we used the following algorithms:

* **RandomForestRegressor**

A random forest is a meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

* **ExtraTreeRegressor**

Extra-trees differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max\_features randomly selected features and the best split among those is chosen.

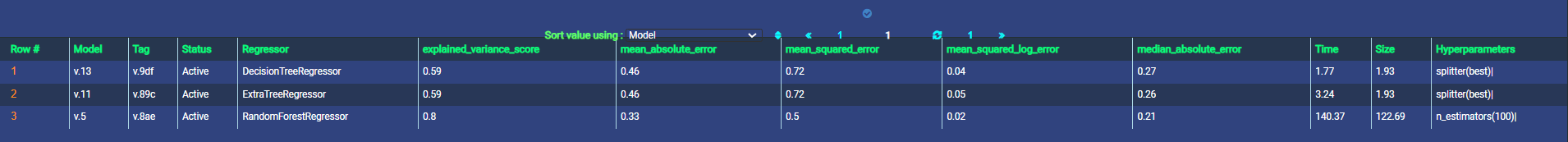
* **DecisionTreeRegressor**

Decision Tree Regressor tries to predict a continuous target variable by cutting the feature variables into small zones, and each zone will have one prediction.

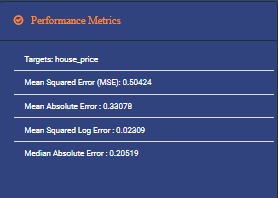
**Model Evaluation**

**Evaluation Metrics**

The models were evaluated on the test dataset using the following metrics:



**Best Model**

The best-performing model is RandomForestRegressor with 0.33 error rate.



**Predicted vs. Target**

**Performance Metrics**

**Model Accuracy and Sample Accuracy**



**Conclusions**

**Improvements**

Future improvements could include adding more relevant features and applying advanced techniques like ensemble learning to boost model performance.

**Key Learnings**

This project highlights the importance of data preprocessing, feature engineering, and the value of model evaluation metrics in selecting the best model for deployment.