IntelliML

Your Intelligent Machine Learning Companion

SAMPLE DATASET

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4	5
7.3	0.65	0.0	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5

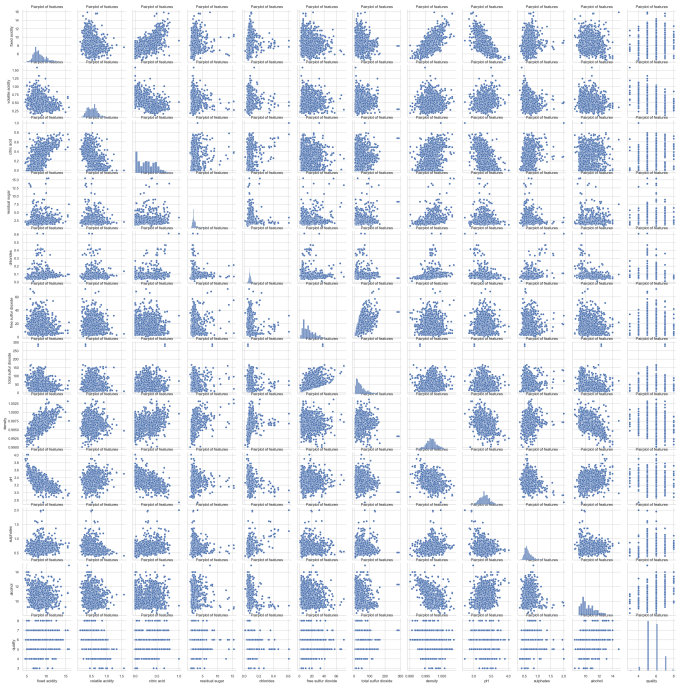
FEATURE DESCRIPTION

The dataset contains 11 features that describe different aspects of wine.

Fixed acidity is the amount of tartaric acid present in the wine. Volatile acidity is a measure of the amount of acetic acid present in the wine. Citric acid is a type of acid that is found in citrus fruits. Residual sugar is the amount of sugar that remains in the wine after fermentation. Chlorides are salts that are found in wine. Free sulfur dioxide is a type of preservative that is added to wine to prevent the growth of bacteria. Total sulfur dioxide is the amount of free sulfur dioxide plus the amount of bound sulfur dioxide. Density is a measure of the weight of a substance compared to the weight of an equal volume of water. pH is a measure of the acidity or alkalinity of a substance. Sulphates are salts that are found in wine. Alcohol is the percentage of alcohol by volume in the wine. Quality is a subjective measure of the overall quality of the wine.

INSIGHTS ON DATASET

The dataset contains 1599 data points with 11 features. The features are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. The mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile and maximum of each feature are reported.



INSIGHTS ON NULL DATA

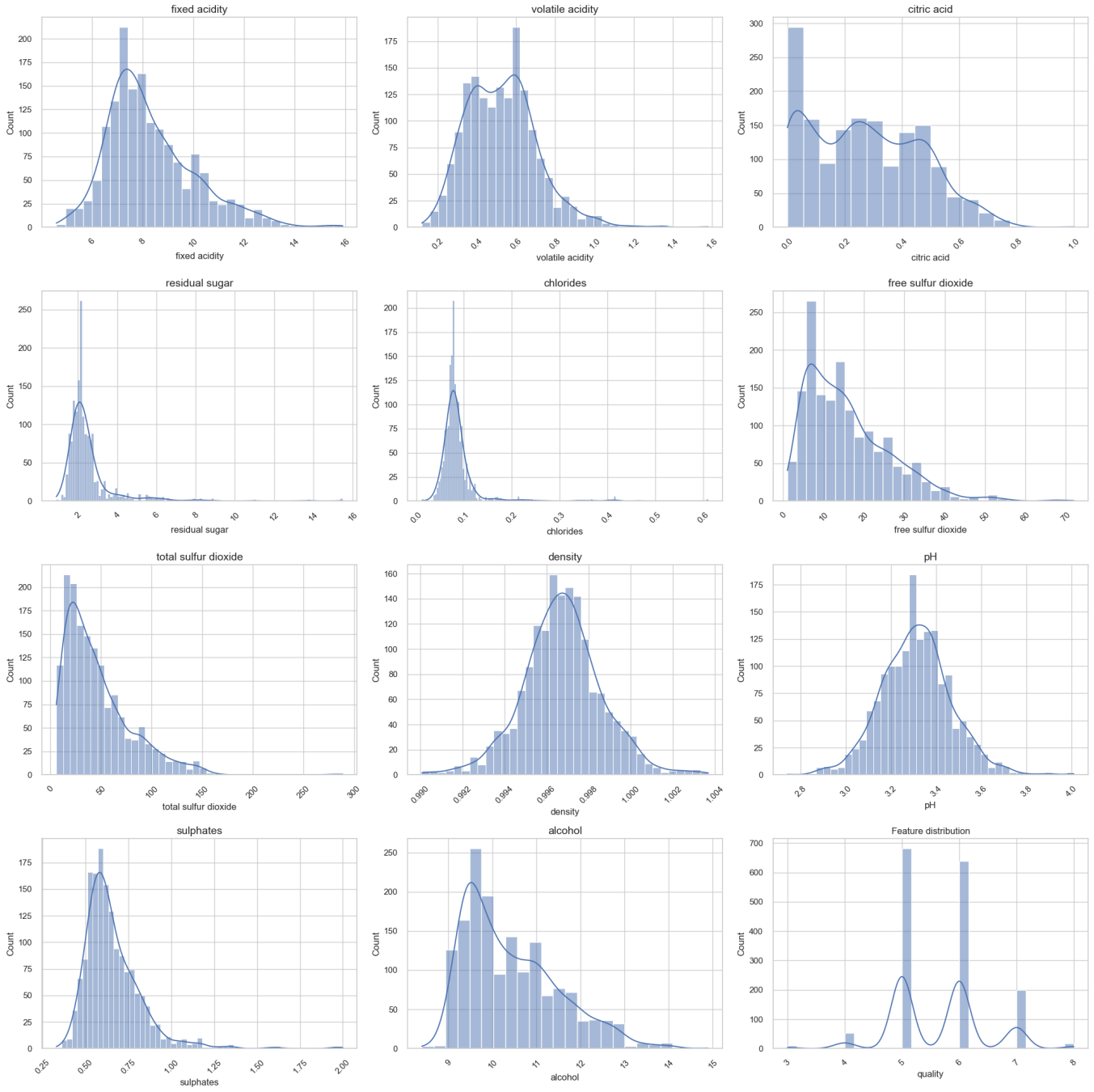
The dataset has no missing values. This is ideal, as it means that all of the data is available for analysis. However, it is important to note that this dataset is relatively small, and it is possible that a larger dataset would contain some missing values. If a dataset does contain missing values, it is important to consider how to handle them. One option is to simply remove the rows with missing values, but this can lead to a loss of data. Another option is to impute the missing values, which means to replace them with an estimated value. The best approach to handling missing values will depend on the specific dataset and the analysis that is being performed.

FEATURE DISTRIBUTION

The distribution of each feature in the dataset is as follows:

Fixed acidity: slightly left skewed, indicating that the data is more concentrated towards the lower values.
Volatile acidity: slightly right skewed, indicating that the data is more concentrated towards the higher values.
Citric acid: slightly left skewed, indicating that the data is more concentrated towards the lower values.
Residual sugar: moderately right skewed, indicating that the data is more concentrated towards the higher values.
Chlorides: moderately right skewed, indicating that the data is more concentrated towards the higher values.
Free sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the lower values.
Total sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the lower values.
Density: slightly right skewed, indicating that the data is more concentrated towards the higher values.
pH: slightly left skewed, indicating that the data is more concentrated towards the lower values.
Sulphates: moderately right skewed, indicating that the data is more concentrated towards the higher values.
Alcohol: slightly right skewed, indicating that the data is more concentrated towards the higher values.
Quality: slightly left skewed, indicating that the data is more concentrated towards the lower values.

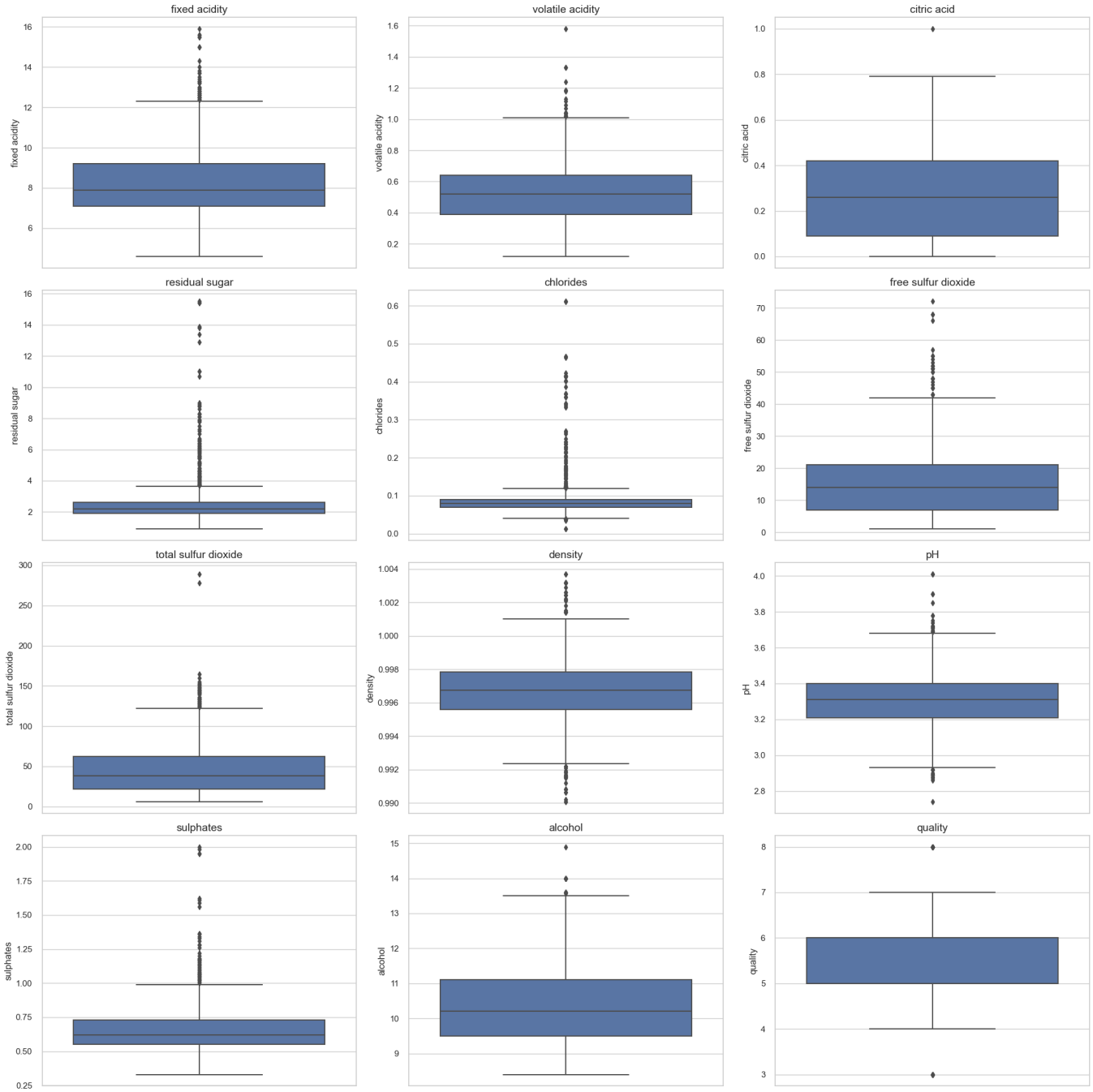
The skewness of the data can have several consequences. For example, a dataset with a high degree of skewness may be more difficult to fit with a linear model, as the data will not be evenly distributed around the mean. Additionally, a dataset with a high degree of skewness may be more susceptible to outliers, which can negatively impact the accuracy of the model.



OUTLIER DETECTION

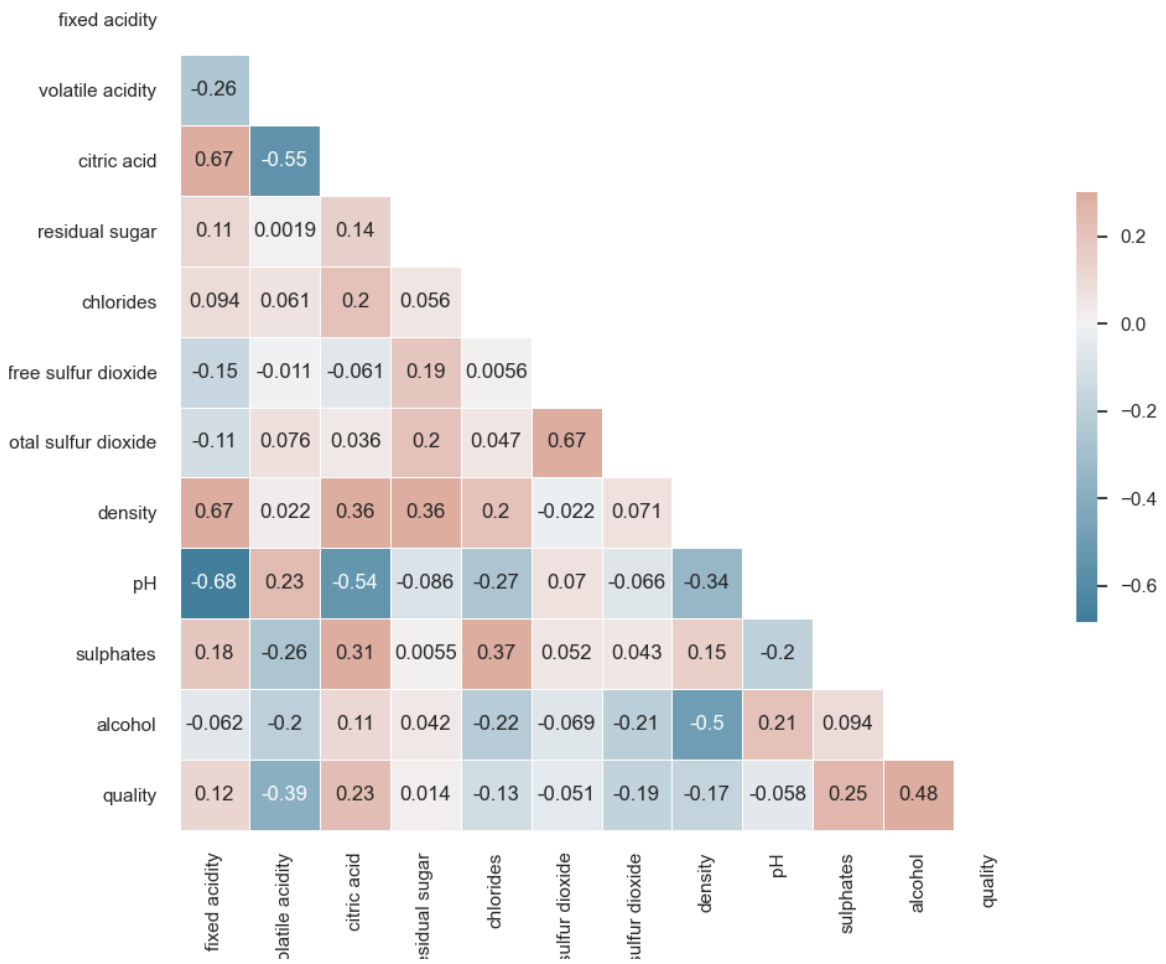
There are a few outliers in the dataset. For fixed acidity, there is an outlier at 15.9, which is more than twice the mean value. This could be due to a measurement error or a mislabeled value. For volatile acidity, there is an outlier at 1.58, which is more than three times the mean value. This could also be due to a measurement error or a mislabeled value. For citric acid, there is an outlier at 1.0, which is more than twice the mean value. This could be due to a measurement error or a mislabeled value. For residual sugar, there is an outlier at 15.5, which is more than six times the mean value. This could be due to a measurement error or a mislabeled value. For chlorides, there is an outlier at 0.611, which is more than twice the mean value. This could be due to a measurement error or a mislabeled value. For free sulfur dioxide, there is an outlier at 72, which is more than four times the mean value. This could be due to a measurement error or a mislabeled value. For total sulfur dioxide, there is an outlier at 289, which is more than six times the mean value. This could be due to a measurement error or a mislabeled value. For density, there is an outlier at 1.00369, which is more than twice the mean value. This could be due to a measurement error or a mislabeled value. For pH, there is an outlier at 4.01, which is more than one full unit away from the mean value. This could be due to a measurement error or a mislabeled value. For sulphates, there is an outlier at 2.0, which is more than twice the mean value. This could be due to a measurement error or a mislabeled value. For alcohol, there is an outlier at 14.9, which is more than four times the mean value. This could be due to a measurement error or a mislabeled value.

The presence of these outliers could have a significant impact on the results of any analysis that is performed on the data. For example, if a linear regression is performed on the data, the outliers could skew the results and lead to inaccurate conclusions. It is important to take into account the presence of outliers when performing any analysis on the data.



CORRELATION BETWEEN FEATURES

The correlation matrix shows that fixed acidity is positively correlated with citric acid and sulphates, and negatively correlated with volatile acidity. Residual sugar is positively correlated with alcohol and negatively correlated with pH. Chlorides is positively correlated with free sulfur dioxide and total sulfur dioxide, and negatively correlated with density. Density is negatively correlated with pH and sulphates. Alcohol is positively correlated with quality.



MACHINE LEARNING EXPERIMENT SETTINGS

This experiment is a regression task with the target variable `quality`. The original data shape is `(1599, 12)`. After transformation, the data shape becomes `(1599, 12)`. The train set has 1119 samples and the test set has 480 samples. The experiment uses 10-fold cross-validation. The preprocess method is `simple` with mean imputation for numeric features and mode imputation for categorical features.

Description	Value
Target	quality
Target type	Regression
Original data shape	(1599, 12)
Transformed data shape	(1599, 12)
Transformed train set shape	(1119, 12)
Transformed test set shape	(480, 12)
Numeric features	11
Preprocess	True
Imputation type	simple
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	KFold
Fold Number	10

Description	Value
CPU Jobs	-1
Use GPU	False

MACHINE LEARNING MODELS USED

Model	Explanation
Extra Trees Regressor	A tree-based ensemble model that reduces variance by using multiple decision trees
Random Forest Regressor	A tree-based ensemble model that reduces overfitting by using multiple decision trees
Light Gradient Boosting Machine	A gradient boosting model that uses a single decision tree
Gradient Boosting Regressor	A gradient boosting model that uses multiple decision trees
Extreme Gradient Boosting	A gradient boosting model that uses multiple decision trees and is optimized for speed
AdaBoost Regressor	An ensemble model that combines multiple weak learners into a strong learner
Linear Regression	A linear model that predicts the target variable using a linear function of the features
Ridge Regression	A linear model that reduces overfitting by adding a penalty to the model coefficients
Least Angle Regression	A linear model that uses an iterative algorithm to find the best model coefficients
Bayesian Ridge	A linear model that uses Bayesian inference to find the best model coefficients
Huber Regressor	A robust linear model that is less sensitive to outliers
K Neighbors Regressor	A non-parametric model that predicts the target variable using the k nearest neighbors of the input data
Elastic Net	A linear model that combines the features of ridge regression and lasso regression
Orthogonal Matching Pursuit	A linear model that uses an iterative algorithm to find the best subset of features
Lasso Regression	A linear model that reduces overfitting by shrinking the model coefficients towards zero
Lasso Least Angle Regression	A linear model that combines the features of lasso regression and least angle regression
Dummy Regressor	A simple model that predicts the mean of the target variable
Decision Tree Regressor	A tree-based model that predicts the target variable using a decision tree
Passive Aggressive Regressor	A linear model that iteratively updates the model coefficients to reduce the training error

PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.3988	0.3379	0.5786	0.4549	0.09	0.074	0.354
Random Forest Regressor	0.4382	0.3542	0.5928	0.4274	0.0922	0.081	0.41
Light Gradient Boosting Machine	0.4545	0.3819	0.6149	0.3846	0.0954	0.0839	0.314
Gradient Boosting Regressor	0.4853	0.4	0.6298	0.3549	0.0976	0.0892	0.298
Extreme Gradient Boosting	0.439	0.402	0.6308	0.3494	0.0981	0.0814	0.348
AdaBoost Regressor	0.5047	0.4055	0.6339	0.3476	0.0977	0.0924	0.262
Linear Regression	0.4962	0.4118	0.6397	0.3362	0.0989	0.0913	3.352
Ridge Regression	0.4967	0.4119	0.6397	0.3362	0.0989	0.0914	0.164
Least Angle Regression	0.4962	0.4118	0.6397	0.3362	0.0989	0.0913	0.168
Bayesian Ridge	0.4969	0.4122	0.6399	0.3358	0.0989	0.0914	0.161
Huber Regressor	0.4958	0.42	0.6454	0.3245	0.0999	0.0915	0.194
K Neighbors Regressor	0.5855	0.5672	0.7513	0.0822	0.115	0.1069	0.188
Elastic Net	0.6438	0.6095	0.7781	0.0217	0.119	0.1184	0.162
Orthogonal Matching Pursuit	0.6412	0.6096	0.7782	0.0213	0.119	0.118	0.164
Lasso Regression	0.6481	0.6105	0.7788	0.0198	0.119	0.1192	0.153

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Lasso Least Angle Regression	0.6481	0.6105	0.7788	0.0198	0.119	0.1192	0.163
Dummy Regressor	0.6714	0.6272	0.7896	-0.0083	0.1205	0.1233	0.243
Decision Tree Regressor	0.4844	0.631	0.789	-0.0238	0.1233	0.0891	0.176
Passive Aggressive Regressor	0.6284	0.6735	0.7969	-0.0746	0.1228	0.1167	0.189

The results show that the Extra Trees Regressor model achieved the best performance with a MAE of 0.3988, MSE of 0.3379, RMSE of 0.5786, R2 score of 0.4549, RMSLE of 0.0900, and MAPE of 0.0740. This is likely due to the fact that Extra Trees Regressor is a very powerful and versatile algorithm that can be used to fit a wide variety of data. Additionally, the model was trained with a large number of trees (n_jobs=-1) and a random state of 5976, which likely helped to improve its performance.

Other models that performed well include the Random Forest Regressor (MAE of 0.4382, MSE of 0.3542, RMSE of 0.5928, R2 score of 0.4274, RMSLE of 0.0922, and MAPE of 0.0810), the Light Gradient Boosting Machine (MAE of 0.4545, MSE of 0.3819, RMSE of 0.6149, R2 score of 0.3846, RMSLE of 0.0954, and MAPE of 0.0839), and the Gradient Boosting Regressor (MAE of 0.4853, MSE of 0.4000, RMSE of 0.6298, R2 score of 0.3549, RMSLE of 0.0976, and MAPE of 0.0892). These models all achieved relatively good performance, suggesting that they are all viable options for regression tasks.