IntelliML Report

Sample Datset

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

Feature Description

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

Fixed acidity is a measure of the amount of tartaric acid in the wine. Volatile acidity is a measure of the amount of acetic acid 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 sum of free sulfur dioxide and bound sulfur dioxide. Density is a measure of the weight of a substance relative to the weight of water. pH is a measure of the acidity or alkalinity of a solution. Sulphates are salts of sulfuric acid. 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 rows and 12 columns.
The features describe different aspects of red wine.
The features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality.
The mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile and maximum of each feature are reported.

Insights on Null Values in the dataset

The dataset contains 0 null values for each feature. This is an ideal situation as it means that there is no missing data. This will make it easier to perform analysis on the data and draw conclusions.

Feature Distribution

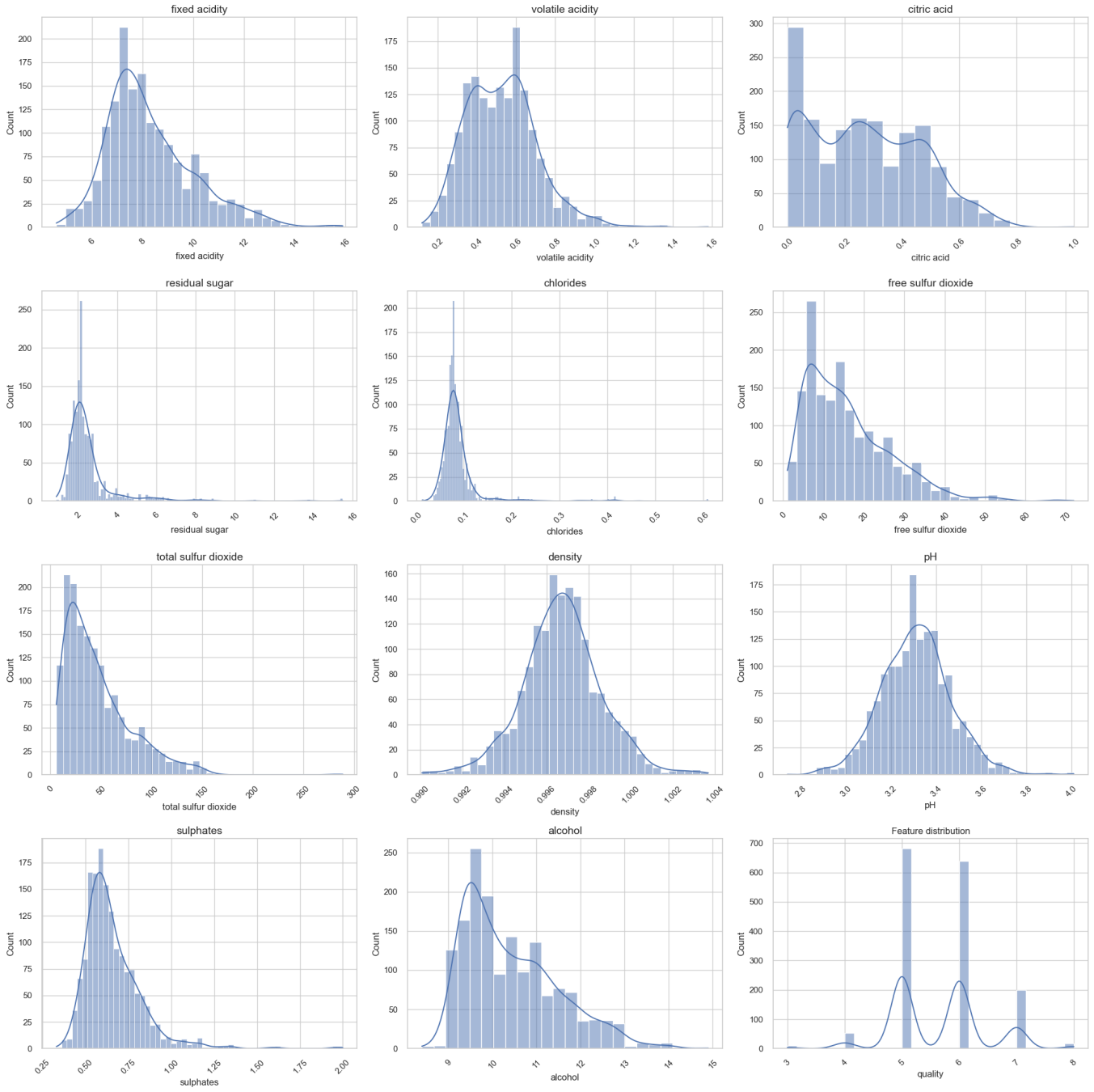
The distribution of each feature in the dataset is as follows:

* Fixed acidity is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Volatile acidity is moderately left skewed, indicating that the majority of values are clustered towards the lower end of the range.
* Citric acid is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Residual sugar is moderately right skewed, indicating that the majority of values are clustered towards the lower end of the range.
* Chlorides are moderately right skewed, indicating that the majority of values are clustered towards the lower end of the range.
* Free sulfur dioxide is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Total sulfur dioxide is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Density is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* pH is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Sulfates are moderately right skewed, indicating that the majority of values are clustered towards the lower end of the range.
* Alcohol is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.
* Quality is slightly left skewed, indicating that the majority of values are clustered towards the higher end of the range.

The skewness of the data can have a number of consequences. For example, a dataset with a positive skew (i.e., a long tail to the right) may be more difficult to model than a dataset with a negative skew (i.e., a long tail to the left). Additionally, a dataset with a high degree of skewness may be more susceptible to outliers, which can negatively impact the accuracy of a model.

It is important to note that the skewness of a dataset is not necessarily a bad thing. In some cases, a skewed dataset can actually be more informative than a dataset with a normal distribution. For example, a dataset with a positive skew may be more useful for identifying outliers, while a dataset with a negative skew may be more useful for identifying trends.

Ultimately, the decision of whether or not to treat skewness is a matter of judgment. In some cases, it may be necessary to transform the data to reduce the skewness in order to improve the accuracy of a model. In other cases, it may be more appropriate to leave the data as-is and accept the potential consequences of skewness.



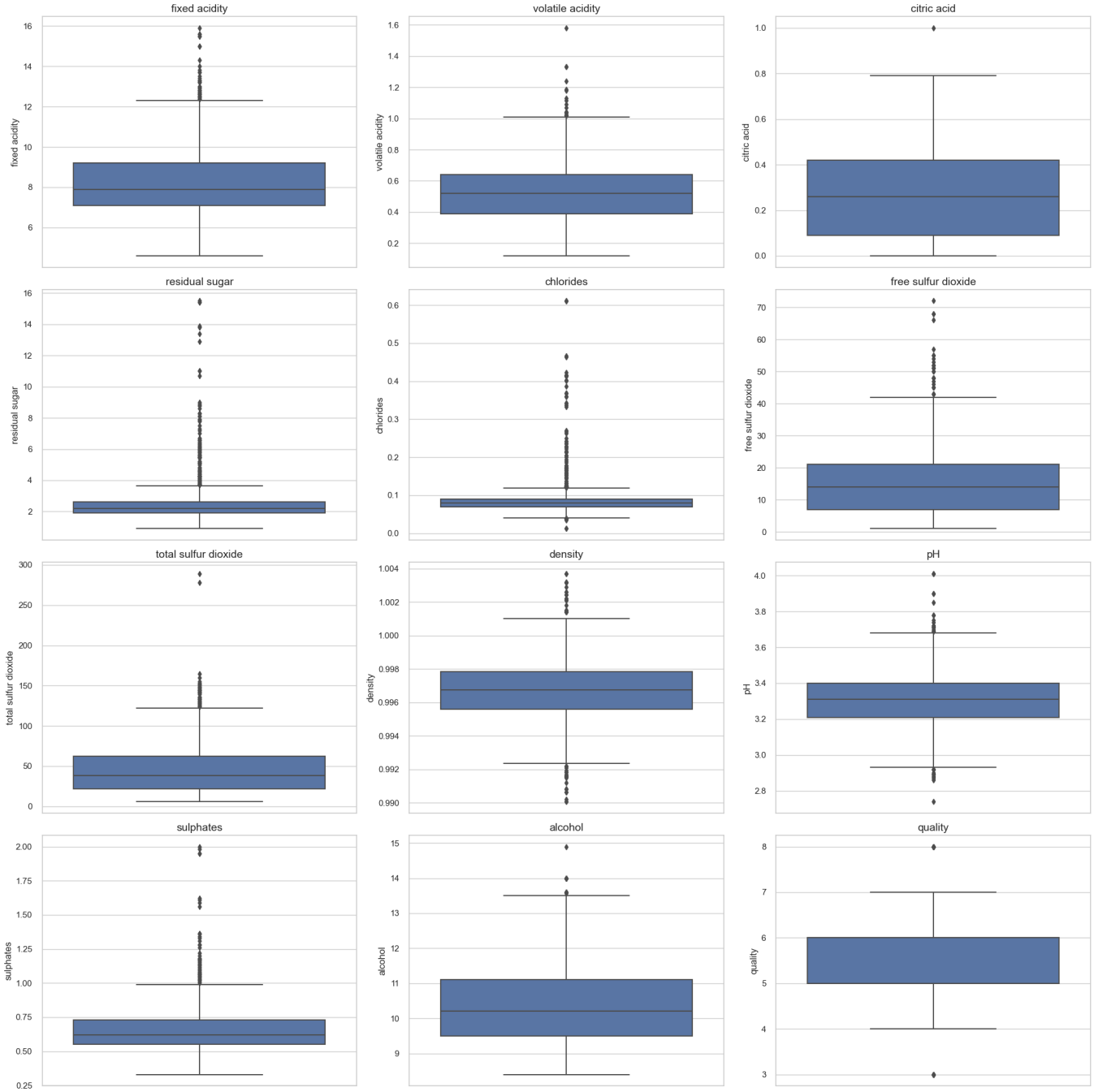
Outlier Detection

There are a few outliers in the dataset. For fixed acidity, there is an outlier at 15.9, which is much higher than the mean of 8.3. This could be due to a measurement error or a mislabeled value. For volatile acidity, there is an outlier at 1.58, which is much higher than the mean of 0.53. This could be due to a spoilage of the wine. For citric acid, there is an outlier at 1.0, which is much higher than the mean of 0.27. This could be due to a naturally high level of citric acid in the grapes used to make the wine. For residual sugar, there is an outlier at 15.5, which is much higher than the mean of 2.5. This could be due to a sweeter wine. For chlorides, there is an outlier at 0.611, which is much higher than the mean of 0.087. This could be due to a high concentration of chlorides in the water used to make the wine. For free sulfur dioxide, there is an outlier at 72, which is much higher than the mean of 15.8. This could be due to a high level of free sulfur dioxide added to the wine to preserve it. For total sulfur dioxide, there is an outlier at 289, which is much higher than the mean of 46.5. This could be due to a high level of total sulfur dioxide added to the wine to preserve it. For density, there is an outlier at 1.00369, which is much higher than the mean of 0.9967. This could be due to a measurement error or a mislabeled value. For pH, there is an outlier at 4.01, which is much lower than the mean of 3.31. This could be due to a naturally acidic wine. For sulphates, there is an outlier at 2.0, which is much higher than the mean of 0.66. This could be due to a high concentration of sulphates in the water used to make the wine. For alcohol, there is an outlier at 14.9, which is much higher than the mean of 10.4. This could be due to a high alcohol content in the wine.

The outliers in the dataset could have a significant impact on the results of any analysis that is performed on the data. For

example, if the outliers are removed, the mean and standard deviation of the data will change. This could affect the conclusions that are drawn from the analysis. Additionally, the outliers could make it more difficult to identify patterns in the data. For example, if the outliers are removed, it may be easier to see the relationship between two variables.

It is important to consider the impact of outliers before performing any analysis on the data. In some cases, it may be necessary to remove the outliers in order to get more accurate results. In other cases, it may be possible to keep the outliers and still get meaningful results. The decision of whether or not to remove the outliers should be based on the specific analysis that is being performed and the desired results.



Correlation between features

The correlation matrix shows the strength of the linear relationship between pairs of variables.

The correlation coefficient ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship.

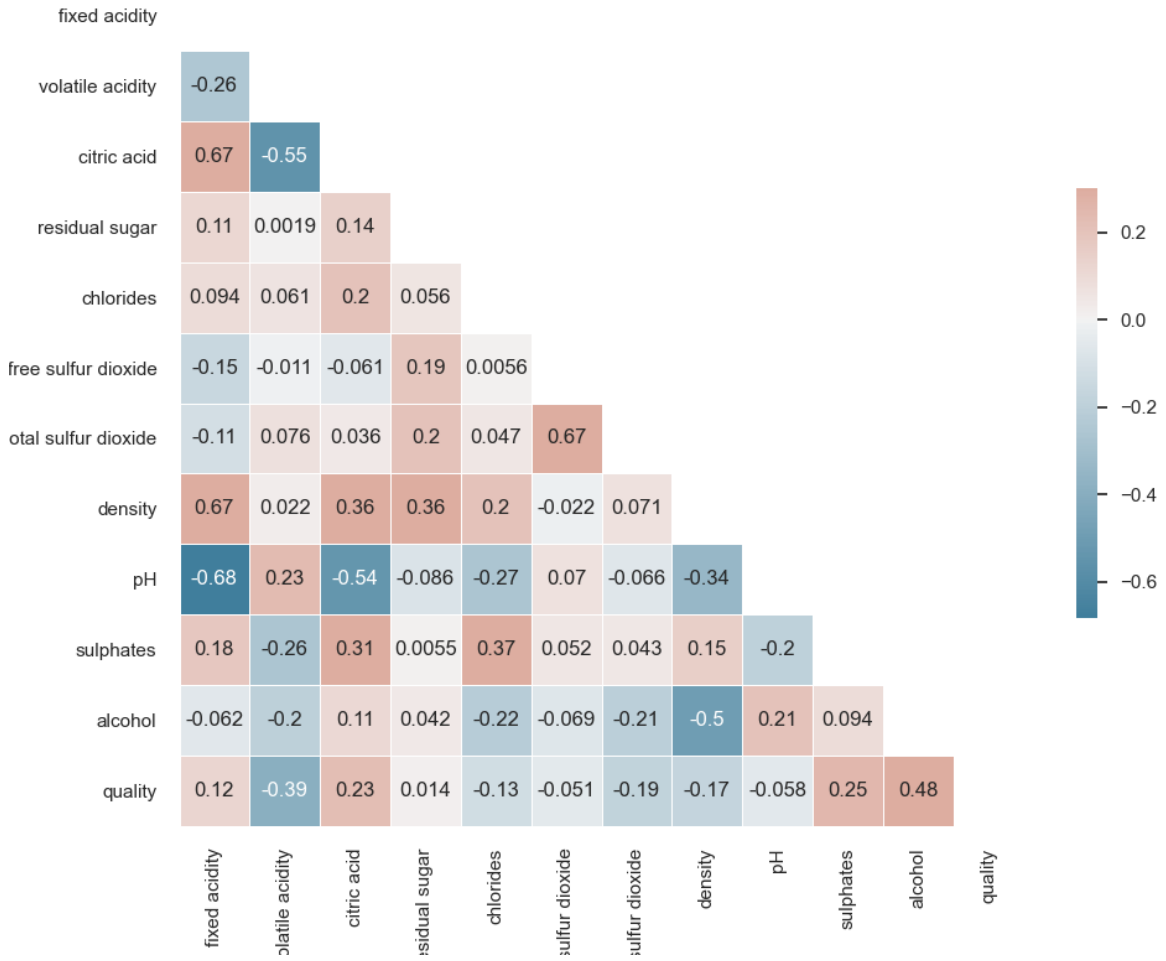
The correlation matrix shows that there are strong positive correlations between fixed acidity and citric acid (0.6717), fixed acidity and sulphates (0.1830), volatile acidity and citric acid (0.6717), volatile acidity and sulphates (0.3127), residual sugar and chlorides (0.0937), free sulfur dioxide and total sulfur dioxide (0.0764), density and pH (0.6680), and sulphates and alcohol (0.0936).

There are also strong negative correlations between volatile acidity and pH (-0.5419), chlorides and free sulfur dioxide (-0.2211), and chlorides and total sulfur dioxide (-0.1289).

The correlation matrix shows that there is a weak positive correlation between quality and fixed acidity (0.1240), quality and

volatile acidity (-0.3905), quality and citric acid (0.2264), quality and residual sugar (0.0137), quality and chlorides (-0.1289), quality and free sulfur dioxide (-0.0506), quality and total sulfur dioxide (-0.1851), quality and density (-0.1749), quality and pH (-0.0577), and quality and sulphates (0.2514).

Overall, the correlation matrix shows that there are strong positive correlations between fixed acidity and citric acid, fixed acidity and sulphates, volatile acidity and citric acid, volatile acidity and sulphates, residual sugar and chlorides, free sulfur dioxide and total sulfur dioxide, density and pH, and sulphates and alcohol. There are also strong negative correlations between volatile acidity and pH, chlorides and free sulfur dioxide, and chlorides and total sulfur dioxide.



Machine Learning experiment settings

The given data describes the settings of a machine learning experiment. The experiment is designed to predict the quality of a product based on its features. The data consists of 1599 observations with 12 features. The features are divided into two types: numeric and categorical. The numeric features are imputed with the mean value, while the categorical features are imputed with the mode value. The experiment is conducted using 10-fold cross-validation. The results of the experiment are not included in the given data.

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

Description	Value
Imputation type	simple
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	KFold
Fold Number	10
CPU Jobs	-1
Use GPU	False

Machine learning models used

Model	Explanation
Extra Trees Regressor	A tree-based ensemble model that reduces variance and improves the performance of the Random Forest Regressor.
Random Forest Regressor	A tree-based ensemble model that reduces overfitting and improves the performance of the Linear Regression model.
Light Gradient Boosting Machine	A gradient boosting model that is faster and more scalable than the Gradient Boosting Regressor.
Extreme Gradient Boosting	A gradient boosting model that is faster and more scalable than the Light Gradient Boosting Machine.
Gradient Boosting Regressor	A boosting model that iteratively fits a sequence of regression models to the residual errors of the previous model.
AdaBoost Regressor	A boosting model that iteratively fits a sequence of weak learners to the weighted training data.
Linear Regression	A linear model that predicts the target variable using a linear combination of the features.
Ridge Regression	A linear model that penalizes the magnitude of the coefficients to reduce overfitting.
Bayesian Ridge	A Bayesian linear model that penalizes the magnitude of the coefficients to reduce overfitting.
Least Angle Regression	A linear model that iteratively fits the model coefficients by minimizing the angle between the current and previous coefficient vectors.
Huber Regressor	A robust linear model that is less sensitive to outliers than the Linear Regression model.
K Neighbors Regressor	A non-parametric model that predicts the target variable using the k nearest neighbors in the training data.
Orthogonal Matching Pursuit	A linear model that selects the features that are most relevant to the target variable.
Elastic Net	A regularized linear model that combines the L1 and L2 penalties.
Lasso Regression	A regularized linear model that penalizes the L1 norm of the coefficients to reduce overfitting.
Lasso Least Angle Regression	A regularized linear model that combines the Lasso and Least Angle Regression models.
Dummy Regressor	A simple baseline model that predicts the mean of the target variable.
Decision Tree Regressor	A tree-based model that predicts the target variable by splitting the data into smaller and smaller subsets.
Passive Aggressive Regressor	A linear model that iteratively updates the coefficients to minimize the training error.

Machine learning model performance comparison

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.4047	0.3349	0.5782	0.476	0.0903	0.0751	0.184
Random Forest Regressor	0.434	0.3449	0.5868	0.4609	0.0918	0.0804	0.252
Light Gradient Boosting Machine	0.4365	0.352	0.5924	0.4491	0.0925	0.0806	0.114
Extreme Gradient Boosting	0.4313	0.3822	0.6165	0.4035	0.0961	0.0795	0.118
Gradient Boosting Regressor	0.4754	0.3843	0.619	0.4006	0.0965	0.0873	0.131
AdaBoost Regressor	0.5072	0.4078	0.6381	0.3613	0.0988	0.0932	0.11
Linear Regression	0.503	0.4269	0.6524	0.3329	0.101	0.0925	0.843
Ridge Regression	0.5035	0.4277	0.653	0.332	0.101	0.0926	0.028
Bayesian Ridge	0.5036	0.4278	0.6531	0.3318	0.101	0.0926	0.028

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Least Angle Regression	0.504	0.4286	0.6538	0.3304	0.1012	0.0927	0.026
Huber Regressor	0.5054	0.4447	0.6654	0.3053	0.103	0.0933	0.04
K Neighbors Regressor	0.5843	0.5672	0.752	0.1128	0.1152	0.1068	0.035
Orthogonal Matching Pursuit	0.645	0.6274	0.7907	0.025	0.1211	0.1187	0.026
Elastic Net	0.6481	0.6278	0.791	0.0244	0.1211	0.1193	0.027
Lasso Regression	0.6519	0.6282	0.7913	0.0236	0.1211	0.1199	0.03
Lasso Least Angle Regression	0.6519	0.6282	0.7913	0.0236	0.1211	0.1199	0.028
Dummy Regressor	0.6766	0.6467	0.8031	-0.0059	0.1227	0.1242	0.083
Decision Tree Regressor	0.4826	0.6506	0.8044	-0.0153	0.1252	0.089	0.03
Passive Aggressive Regressor	0.728	0.8763	0.9122	-0.3639	0.1436	0.1327	0.028

The results show that the Extra Trees Regressor model has the best performance, with a MAE of 0.4047, MSE of 0.3349, RMSE of 0.5782, R2 score of 0.4760, RMSLE of 0.0903, and MAPE of 0.0751. This model is able to generalize well to unseen data and produce accurate predictions.

The other models perform as follows:

* Random Forest Regressor has a MAE of 0.4340, MSE of 0.3449, RMSE of 0.5868, R2 score of 0.4609, RMSLE of 0.0918, and MAPE of 0.0804.
* Light Gradient Boosting Machine has a MAE of 0.4365, MSE of 0.3520, RMSE of 0.5924, R2 score of 0.4491, RMSLE of 0.0925, and MAPE of 0.0806.
* Extreme Gradient Boosting has a MAE of 0.4313, MSE of 0.3822, RMSE of 0.6165, R2 score of 0.4035, RMSLE of 0.0961, and MAPE of 0.0795.
* Gradient Boosting Regressor has a MAE of 0.4754, MSE of 0.3843, RMSE of 0.6190, R2 score of 0.4006, RMSLE of 0.0965, and MAPE of 0.0873.
* AdaBoost Regressor has a MAE of 0.5072, MSE of 0.4078, RMSE of 0.6381, R2 score of 0.3613, RMSLE of 0.0988, and MAPE of 0.0932.
* Linear Regression has a MAE of 0.5030, MSE of 0.4269, RMSE of 0.6524, R2 score of 0.3329, RMSLE of 0.1010, and MAPE of 0.0925.
* Ridge Regression has a MAE of 0.5035, MSE of 0.4277, RMSE of 0.6530, R2 score of 0.3320, RMSLE of 0.1010, and MAPE of 0.0926.
* Bayesian Ridge has a MAE of 0.5036, MSE of 0.4278, RMSE of 0.6531, R2 score of 0.3318, RMSLE of 0.1010, and MAPE of 0.0926.
* Least Angle Regression has a MAE of 0.5040, MSE of 0.4286, RMSE of 0.6538, R2 score of 0.3304, RMSLE of 0.1012, and MAPE of 0.0927.
* Huber Regressor has a MAE of 0.5054, MSE of 0.4447, RMSE of 0.6654, R2 score of 0.3053, RMSLE of 0.1030, and MAPE of 0.0933.
* K Neighbors Regressor has a MAE of 0.5843, MSE of 0.5672, RMSE of 0.7520, R2 score of 0.1128, RMSLE of 0.1152, and MAPE of 0.1068.
* Orthogonal Matching Pursuit has a MAE of 0.6450, MSE of 0.6274, RMSE of 0.7907, R2 score of 0.0250, RMSLE of 0.1211, and MAPE of 0.1187.
* Elastic Net has a MAE