Decision Tree vs Logistic Regression

TL;TR: Decision tree is superior over a logistic regression when the data set is large and when the relationships between the different features and the target variable are complex and non-linear.

Decision trees and logistic regression are both popular machine learning algorithms used for classification problems. Both algorithms have their own strengths and weaknesses, and the choice between them depends on the specific characteristics of the dataset and the problem at hand.

When to use a decision tree over a logistic regression: Decision trees are often better suited than logistic regression in certain use cases such as:

1. **Handling Complex and Non-Linear Relationship:** Decision trees are able to handle both categorical and numerical features, as well as non-linear relationships between features and the target variable. This makes decision trees well suited for datasets where the relationship between the predictors and the response is not easily modeled by a linear equation. This makes decision tree a good choice when the relationships between different features and the target variable are complex and non-linear.
2. **Handling outliers:** Decision trees are able to handle missing values and outliers in the data much better then a logistic regression. A decision tree is not affected by outliers because it splits the data based on the feature values. Outliers are just considered as another data point with a specific feature value, and the tree will split the data accordingly. Even if an outlier falls into a leaf, the leaf will still have a mixture of examples from different classes. Therefore, the leaf will be impure, and the label will be determined by the majority class.
   On the other hand, logistic regression is affected by outliers because it is a linear model, which assumes a linear relationship between the predictors and the log-odds of the response variable. The algorithm finds the best-fitting line by minimizing the sum of squared errors. Outliers can have a significant impact on the line and can result in overfitting. The presence of outliers can pull the line towards them, and cause the model to be less generalizable to new data. Additionally, the logistic regression algorithm uses Maximum Likelihood Estimation (MLE) to estimate the parameters of the model, which is sensitive to outliers. MLE assumes that the data is normally distributed, but outliers can cause the distribution to deviate from normality, leading to biased estimates of the parameters.
3. **Handling missing values:** Decision trees are able to handle sparse data, which means that they can handle data where many of the features have missing or zero values. Logistic regression, on the other hand, assumes that all features have non-zero values, so it can't handle sparse data as well as decision trees. It assumes that the data is Missing Completely at Random (MCAR), which means that the probability of missing data is unrelated to both observed and unobserved variables. But in practice, it's often not the case, and the missing data may be related to the outcome variable, this can lead to bias in the model's estimates and predictions. Another reason is that logistic regression is a parametric model, it makes assumptions about the distribution of the data, and requires a large sample size to estimate the parameters accurately. When data is missing, it can lead to a decrease in sample size and make it harder to estimate the parameters accurately. Another reason is that logistic regression is a parametric model, it makes assumptions about the distribution of the data, and requires a large sample size to estimate the parameters accurately. When data is missing, it can lead t

4. **Handling Large and High-Dimensional Data:** Decision trees partition the feature space into smaller regions, which makes them more robust to high-dimensional data. This is because the partitioning reduces the effective dimensionality of the data, making it easier to find patterns and make predictions. For example, a decision tree can split the data into smaller subsets based on different combinations of features, whereas logistic regression uses all features at once and the model becomes complex and harder to interpret as the number of features increases. Additionally decision trees are able to handle high-dimensional data by selecting the most important features to split the data. The algorithm starts with all features and at each split, it selects the feature that best separates the data into different classes. This means that decision trees can automatically select the most important features, reducing the dimensionality of the data. Logistic regression, on the other hand, uses all features at once, which can make it harder to interpret and understand the model when the number of features is high.

In summary, decision trees are a good choice when the relationship between the predictors and the response is complex and non-linear, when interpretability is important, when dealing with missing values or outliers, when handling large and high-dimensional data, and when the target variable has more than two classes.

When to use a logistic regression over a decision tree

1. **Predicting a binary outcome:** Logistic regression is specifically designed for binary classification problems, where the goal is to predict one of two possible outcomes. It estimates the probability of a binary outcome based on one or more predictor variables. Logistic regression is simple to interpret and it doesn't require much computational resources. This makes it a good choice when the target variable is binary and the relationship between the predictors and the response is linear.
2. **Linear relationship between predictors and response:** Logistic regression assumes that there is a linear relationship between the predictors and the log-odds of the response variable. This makes it well-suited for datasets where the relationship between the predictors and the response can be described by a linear equation.
3. **Small Sample Size:** Logistic regression tends to perform better with small sample sizes than decision trees. Decision trees require a large number of observations to create a stable and accurate model, and are more prone to overfitting with small sample sizes.
4. **Dealing with Categorical Predictors:** Logistic regression can handle categorical predictors by creating binary or dummy variables. This makes it well-suited for datasets where some of the predictors are categorical.

In summary, Logistic regression is better than a decision tree when the relationship between the predictors and the response can be modeled by a linear equation, when interpretability and transparency are important, when dealing with continuous predictors, when the sample size is small and when it's needed to predict class probabilities directly.