Decision Trees and Random Forests

Introduction

In the realm of business and leadership, decision-making is paramount. The likes of Steve Jobs, Sundar Pichai, and Jeff Bezos constantly navigate through a sea of choices to steer their companies toward success. These visionary leaders rely heavily on data-driven decisions, as Jeff Bezos emphasized in a recent interview, stating, “The great thing about fact-based decisions is that they overrule the hierarchy.” This philosophy underpins the essence of Decision Trees and Random Forests, where machines make decisions based on available information, mirroring the approach of successful individuals.

For instance, consider the classic Hangman game—a word-guessing challenge where players attempt to unveil a concealed word by suggesting letters. With each letter guessed, players progressively narrow down the possible options, ultimately aiming to pinpoint the correct word. This iterative process of decision-making mirrors the fundamental concept of Decision Trees.

Later in this article, we'll delve into how Decision Trees can accurately predict Hangman game words, illustrating their practical application in a familiar context.

Decoding Decision Trees

Decision Tree is machine learning algorithm which is used for regression as well as classification problems.

Let's illustrate how Decision Trees work through an example:

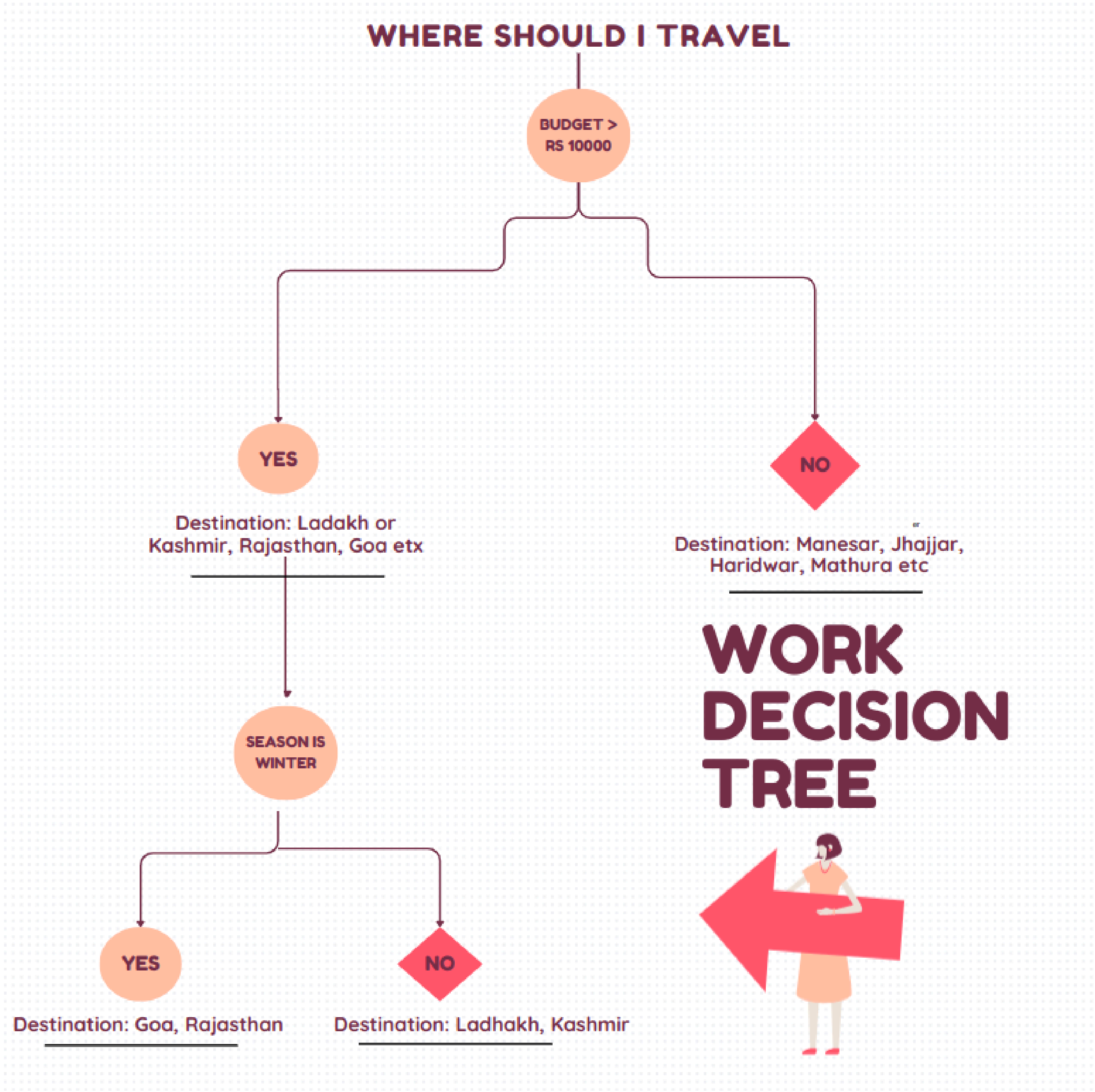
Imagine you're planning a trip for yourself. You log onto the MakeMyTrip app to find the perfect destination within your budget. After filtering based on budget constraints, you decide on Ladakh and seamlessly book your tickets. However, there's a catch—Ladakh is closed during the winter season due to harsh weather conditions, posing a safety risk for travelers. Despite your initial decision, you inadvertently chose a destination unsuitable for the current season, prioritizing basic enjoyment over safety.

Now, let's consider an alternative approach to decision-making. You revisit the MakeMyTrip app and again filter destinations by budget. This time, however, you incorporate an additional filter to screen for places available during the winter season. After applying this second filter, you're presented with a range of destinations that meet both your budget and seasonal preferences.

Although this method may take more time than the previous one, it offers a more informed decision-making process. By considering multiple factors sequentially—budget and then season—you enhance the quality of your decision, prioritizing safety and suitability over mere enjoyment.

This iterative process closely resembles the workings of a Decision Tree. Just as you sequentially filter destinations based on specific criteria, Decision Trees navigate through a series of questions or features to arrive at a conclusive decision. Each step in the decision-making process leads to a distinct path, culminating in a final outcome—similar to how branches of a tree extend from a central trunk.

That why decision tree is also known as nested if-else classification model


WHERE SHOULD I TRAVEL
BUDGET > RS 10000
YES
Destination: Ladakh or Kashmir, Rajasthan, Goa etx
NO
Destination: Manesar, Jhajjar, Haridwar, Mathura etc
SEASON IS WINTER
YES
Destination: Goa, Rajasthan
NO
Destination: Ladhakh, Kashmir
WORK DECISION TREE

Advantages to using Decision Trees:

- Simplicity: Decision Trees offer a straightforward and easy-to-understand concept.
- Accuracy: Despite their simplicity, Decision Trees often outperform complex algorithms.
- Speed: Their uncomplicated nature makes Decision Trees exceptionally fast.

Disadvantages to using Decision Trees:

- Overfitting: Decision Trees are susceptible to overfitting, especially with noisy data.
- Complexity: In certain scenarios, Decision Trees can become overly intricate, complicating interpretation.

Why did Decision Trees checks Budget first and not the Season?

This question delves into the concept of feature importance within Decision Trees. How do model decide which feature to prioritize first. Therefore these features are prioritize on the basic criteria like Gini Impurity, or Information Gain or Entropy.

These criteria help the Decision Tree algorithm assess the significance of each feature in predicting the outcome. In our scenario, budget might be prioritized over season because it's considered more influential in determining the choice of destination. For instance, a traveler's budget constraint might limit their options more significantly than seasonal preferences.

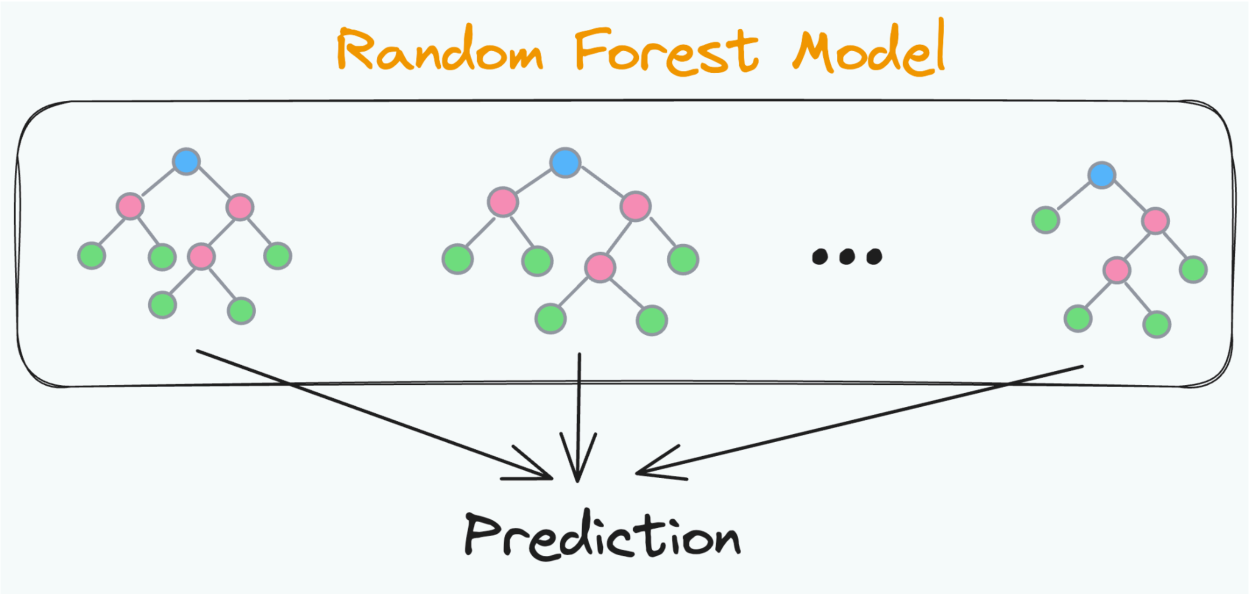
By employing these evaluation criteria, Decision Trees can effectively judge the importance of features, ensuring that the decision-making process is guided by the most relevant factors.

Random Forests - The Ensemble Advantage -

One drawback of Decision Trees is their tendency to become dense and overly complex, often leading to overfitting. To mitigate this issue and reduce variance, Data Scientists commonly turn to Random Forests.

Random Forests offer a solution by aggregating predictions from multiple decision trees. Essentially, a Random Forest is a collection of numerous decision trees, where the outcomes of individual trees are combined into a final result. This ensemble approach creates a robust and generalized model, akin to a forest comprising various trees.

The term "Random" in "Random Forest" stems from the fact that decision trees within the ensemble are randomly created. Each tree is trained on a random subset of the data, introducing diversity and preventing individual trees from overly fitting the training data. This randomness enhances the overall performance and reliability of the Random Forest model, making it a preferred choice for addressing the limitations of Decision Trees.



Advantages to using Random Forest:

- Overfitting Reduction: Random Forests effectively combat overfitting, resulting in more generalized models.
- Scalability: They demonstrate efficacy with large datasets due to their parallelizable nature.

Disadvantages to using Random Forest:

- Training Time: Random Forests may require more time for training compared to Decision Trees.

How do Entropy Works?

Understanding how entropy works is crucial in grasping the concept of feature importance within Decision Trees. Entropy serves as a fundamental metric used by the model to determine the significance of features during the decision-making process.

Before delving into the problem-solving task, Decision Tree models begin by calculating entropy. But what exactly is entropy? In simple terms, entropy represents the level of disorder or impurity within the dataset. By quantifying this disorder, the model gains insights into the distribution of data across different features.

Entropy-based calculations enable the model to evaluate the information gain associated with selecting a particular feature for splitting. The goal is to minimize entropy, indicating a more homogeneous subset of data after the split. Features that result in significant reductions in entropy are considered more informative and are prioritized during the decision-making process.

In summary, entropy serves as a cornerstone in the decision-making process of Decision Trees, guiding the model in selecting the most relevant features to optimize predictive performance. By understanding and leveraging entropy effectively, Decision Trees can navigate through complex datasets and make informed decisions to achieve accurate predictions.

Formula of Entropy

$$Entropy\ (p) \ = \ - \sum_{i=1}^{N} p_i \ log_2 \, p_i$$

$h(X) = – [(P_i * \log_2 P_i) + (Q_i * \log_2 Q_i)]$

Here P_i = Probability of Y = 1 i.e. probability of success of the event
Q_i = Probability of Y = 0 i.e. probability of failure of the event

Conclusion:

In conclusion, the concepts of Decision Trees and Random Forests offer powerful tools for making data-driven decisions in various domains. Decision Trees provide a structured approach to decision-making, resembling the sequential process we use in everyday life, while Random Forests offer an ensemble method to improve accuracy and reduce overfitting. By understanding these concepts and their applications, individuals and organizations can harness the potential of machine learning to make informed choices and achieve better outcomes. Whether it's planning a trip, predicting outcomes in business, or solving complex problems, Decision Trees and Random Forests serve as invaluable resources for enhancing predictive performance and driving success.