Decision Tree Challenge

Feature Importance and Categorical Variable Encoding

Decision Tree Challenge - Feature Importance and Variable Encoding

Challenge Overview

Your Mission: Create a simple GitHub Pages site that demonstrates how decision trees measure feature importance and analyzes the critical differences between categorical and numerical variable encoding. You'll answer two key discussion questions by adding narrative to a pre-built analysis and posting those answers to your GitHub Pages site as a rendered HTML document.

Discussion Questions for Challenge

Your Task: Add thoughtful narrative answers to these two questions in the Discussion Questions section of your rendered HTML site.

1. **Numerical vs Categorical Encoding:** There are two modelsin Python written above. For each language, the models differ by how zip code is modelled, either as a numerical variable or as a categorical variable. Given what you know about zip codes and real estate prices, how should zip code be modelled, numerically or categorically? Is zipcode and ordinal or non-ordinal variable?
2. **R vs Python Implementation Differences:** When modelling zip code as a categorical variable, the output tree and feature importance would differ quite significantly had you used R as opposed to Python. Investigate why this is the case. What does R offer that Python does not? Which language would you say does a better job of modelling zip code as a categorical variable? Can you quote the documentation at https://scikit-learn.org/stable/modules/tree.html suggesting a weakness in the Python implementation? If so, please provide a quote from the documentation.

3. **Are There Any Suggestions for Implementing Decision Trees in Python With Prioper Categorical Handling?** Please poke around the Internet (AI is not as helpful with new libraries) for suggestions on how to implement decision trees in Python with better (i.e. not one-hot encoding) categorical handling. Please provide a link to the source and a quote from the source. There is not right answer here, but please provide a thoughtful answer, I am curious to see what you find.

Solution

i Numerical vs Categorical Encoding

The two models which were shared considered **ZipCode** as `numerical` variable and other as `categorical` variable. Based on my understanding of how zipcodes have a very significant effect on pricing of the real estate (Fun fact: Wife is a Real Estate Agent). For decision tree models, which are **non-parametric** and **insensitive** to monotonic transformations, the best way to encode zipcode depends on how you want the model to **interpret geographic information**.

Is Zip Code Ordinal or Non-Ordinal?

- **Zip code is a non-ordinal categorical variable.**
- Although zip codes are numeric, their values **do not imply order, magnitude, or proximity.**
  - For example, zip code 10001 (NYC) and 90210 (Beverly Hills) are far apart geographically, yet numerically close.
- They are **nominal identifiers** for geographic regions — not ranked or scaled

How Should Zip Code Be Modeled?

Categorically, but with nuance depending on model type and goals
For **Decision Tree**:

- Label Encoding
- Target Encoding

Trees handle arbitrary labels well; target encoding can capture price patterns effectively