Decision Tree Challenge

Feature Importance and Categorical Variable Encoding

Decision Tree Challenge - Feature Importance and Variable Encoding

How does encoding categorical variables as numbers affect our understanding of feature importance?

For detailed analysis and conclusions, see the [Discussion Section] {#sec-Discussion} .

The Ames Housing Dataset

We are analyzing the Ames Housing dataset which contains detailed information about residential properties sold in Ames, Iowa from 2006 to 2010. This dataset is perfect for our analysis because it contains a categorical variable (like zip code) and numerical variables (like square footage, year built, number of bedrooms).

The Problem: ZipCode as Numerical vs Categorical

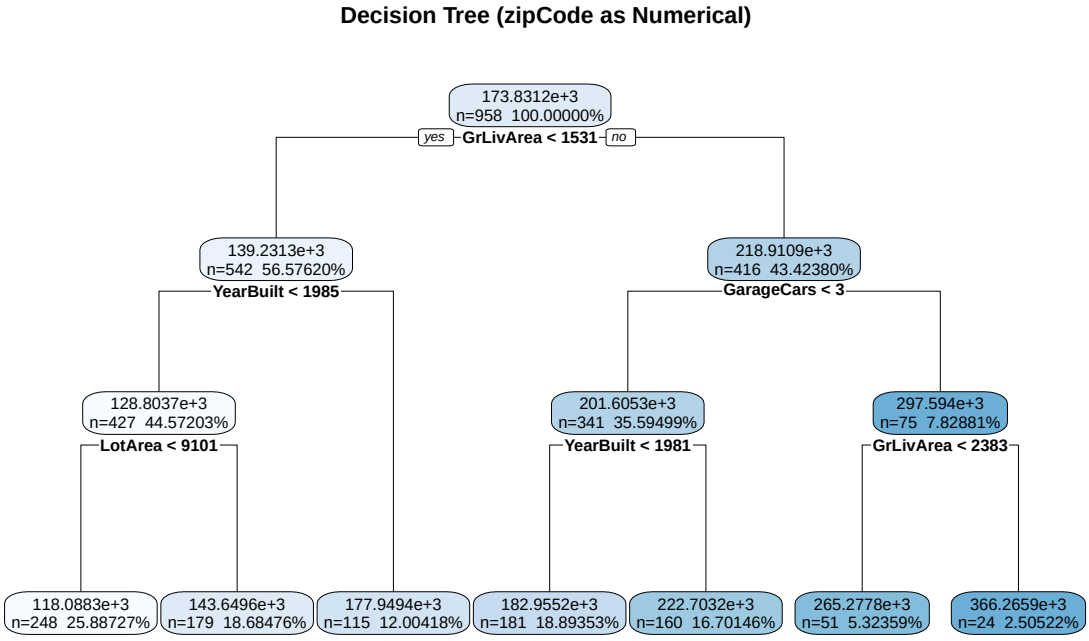
Key Question: What happens when we treat zipCode as a numerical variable in a decision tree? How does this affect feature importance interpretation?

The Issue: Zip codes (50010, 50011, 50012, 50013) are categorical variables representing discrete geographic areas, i.e. neighborhoods. When treated as numerical, the tree might split on "zipCode $>$ 50012.5" - which has no meaningful interpretation for house prices. Zip codes are non-ordinal categorical variables meaning they have no inherent order that aids house price prediction (i.e. zip code 99999 is not the priceiest zip code).

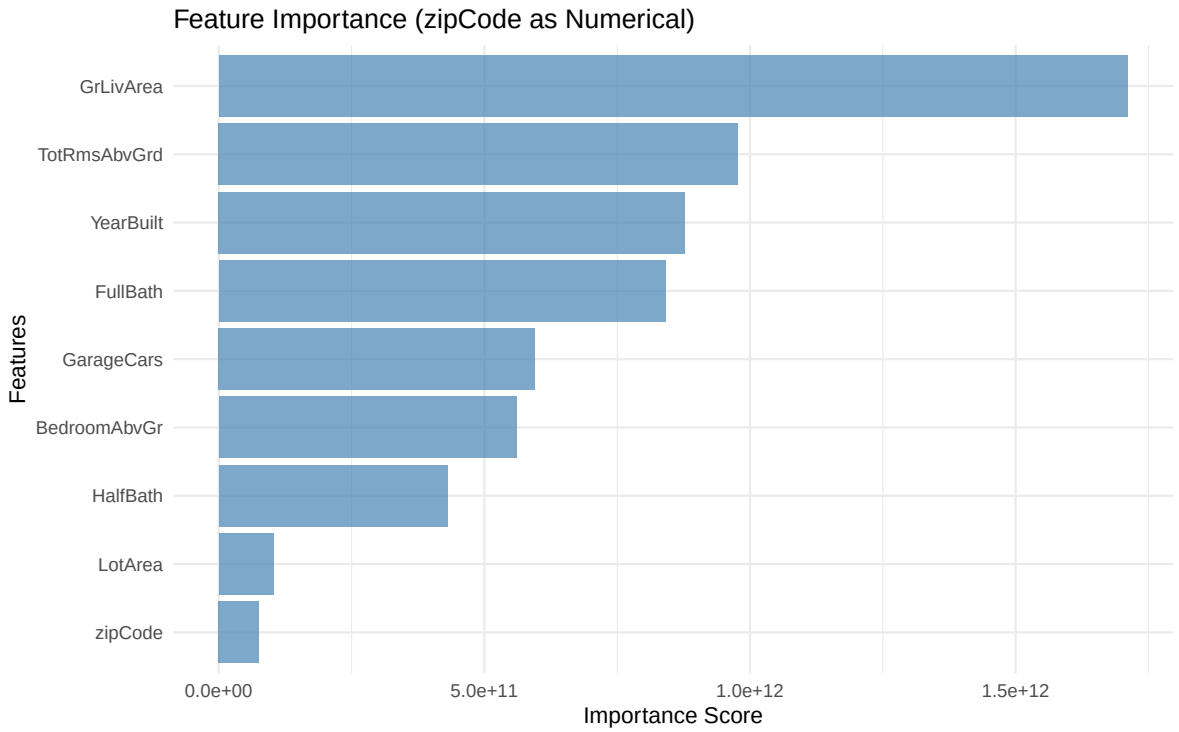
Data Loading and Model Building

```
Model built with 7 terminal nodes
```

Tree Visualization



Feature Importance Analysis



Critical Analysis: The Encoding Problem

The Problem Revealed

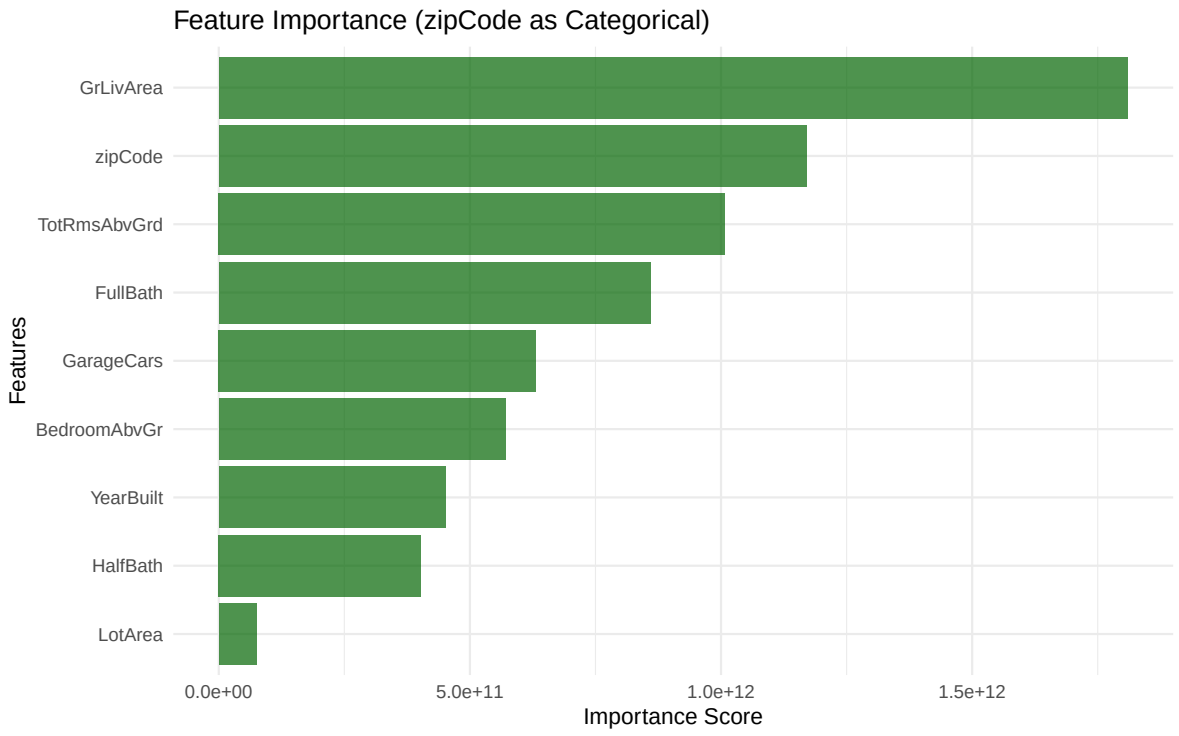
What to note: Our decision tree treated `zipCode` as a numerical variable. This leads to zip code being unimportant. Not surprisingly, because there is no reason to believe allowing splits like "zipCode $<$ 50012.5" should be beneficial for house price prediction. This false coding of a variable creates several problems:

1. **Potentially Meaningless Splits:** A zip code of 50013 is not "greater than" 50012 in any meaningful way for house prices
2. **False Importance:** The algorithm assigns importance to zipCode based on numerical splits rather than categorical distinctions OR the importance of zip code is completely missed as numerical ordering has no inherent relationship to house prices.
3. **Misleading Interpretations:** We might conclude zipCode is not important when our intuition tells us it should be important (listen to your intuition).

The Real Issue: Zip codes are categorical variables representing discrete geographic areas. The numerical values have no inherent order or magnitude relationship to house prices. These must be modelled as categorical variables.

Categorical Encoding Analysis

Feature Importance: Categorical zipCode



Discussion {sec-Discussion}

1. **Numerical vs Categorical Encoding:**

I know that there is no difference higher zip codes vs lower zip codes which means that modeling numerically does not make much sense because the value of the numbers within a zip codes does not tell us anything. But, modeling categorically allows us to treat each zip code as its own category which helps with analysis.

2. **R vs Python Implementation Differences:**

R does a better job when modeling a categorical variable because the Python importance chart attempts to show the importance of each zip code in the dataset which does not help us since there are so many. This is because Python creates a dummy variable for each zip code while R uses a factor to make zip codes categorical. This is why each individual zip code shows up on the Python feature importance graph. There is some documentation saying that the langugage Julia can handle categorical variables effectively. Julia documentation states, "CategoricalVector … is designed to additionally provide full support for working with categorical variables, both with unordered (nominal variables) and ordered categories (ordinal variables)" which shows Julia's versatility when it comes to these kinds of variables.