

DATA 3464: Fundamentals of Data Processing

<< Long form title >>

Charlotte Curtis

January 20, 2026

Topic overview

- Exploring categorical data
- Dealing with missing values
- Categorical data encoding strategies

Resources used:

- [Feature Engineering Chapter 5](#)

What is categorical data?

- Samples can take on one of several discrete values or groups
 - **Nominal**: no particular order to the groups
 - **Ordinal**: groups relate to each other in a specific order
- Categories can be represented as strings *or* numeric types
 - Domain knowledge is necessary!

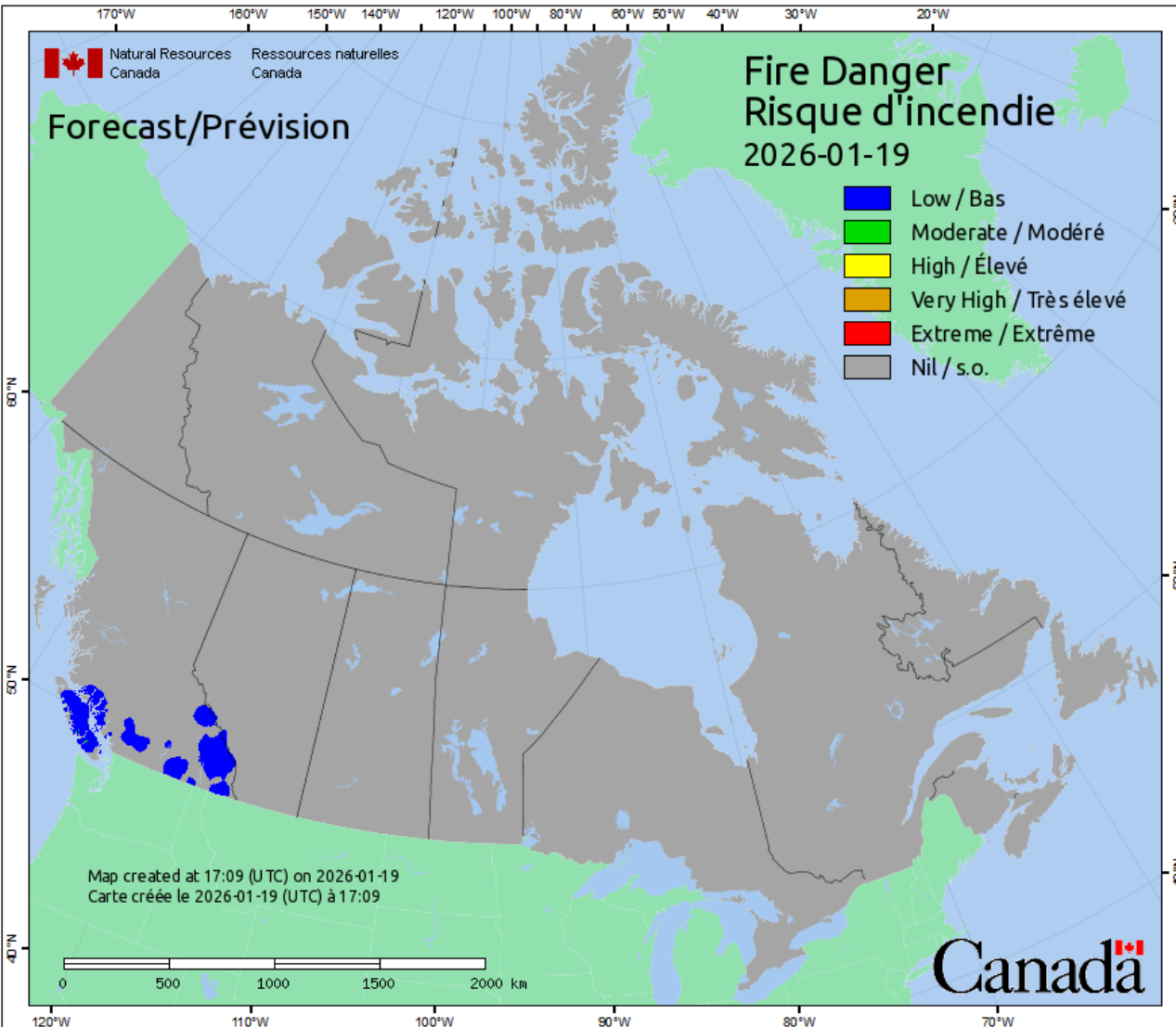
Let's take a few minutes to brainstorm some examples

Representing categorical data

- Tree-based models can handle string-based categories as-is
- Most other models need numbers
- Consider:
 - Ordinal or nominal?
 - How many possible categories?
 - Any chance new ones might show up?

How could we encode the examples?

Ordinal encoding



Category	Feature
Nil	0
Low	1
Moderate	2
High	3
Very High	4
Extreme	5

Nominal categories: one-hot encoding

- Categories have no natural relationship
- Create k new features from k categories, very sparse matrix

Animal		cat	dog	rabbit
cat	→	1	0	0
dog	→	0	1	0
rabbit	→	0	0	1

What kinds of problems could occur with this encoding scheme?

Another approach: target encoding

Getting fancy

- Feature hashing or the "hash trick"