Processing And Transforming Features: Takeaways

Syntax

- Converting any column to the categorical data type:

```
train['Utilities'] = train['Utilities'].astype('category')
```

- Accessing the underlying numerical representation of a column:

```
train['Utilities'].cat.codes
```

- Applying dummy coding for all of the text columns:

```
dummy_cols = pd.get_dummies()
```

- Replace all missing values in a column with a specified value:

```
fill_with_zero = missing_floats.fillna(0)
```

Concepts

- Feature engineering is the process of processing and creating new features. Feature engineering is a bit of an art and having knowledge in the specific domain can help create better features.
- Categorical features are features that can take on one of a limited number of possible values.
- A drawback to converting a column to the categorical data type is that one of the assumptions of linear regression is violated. Linear regression operates under the assumption that the features are linearly correlated with the target column.
- Instead of converting to the categorical data type, it's common to use a technique called dummy coding. In dummy coding, a dummy variable is used. A dummy variable that takes the value of or to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.
- When values are missing in a column, there are two main approaches we can take:
    - Removing rows that contain missing values for specific columns:
        - Pro: Rows containing missing values are removed, leaving only clean data for modeling.
        - Con: Entire observations from the training set are removed, which can reduce overall prediction accuracy.
    - Imputing (or replacing) missing values using a descriptive statistic from the column:
        - Pro: Missing values are replaced with potentially similar estimates, preserving the rest of the observation in the model.
        - Con: Depending on the approach, we may be adding noisy data for the model to learn.

Resources

- Feature Engineering
- Dummy Coding
- pandas.Dataframe.ffillna()