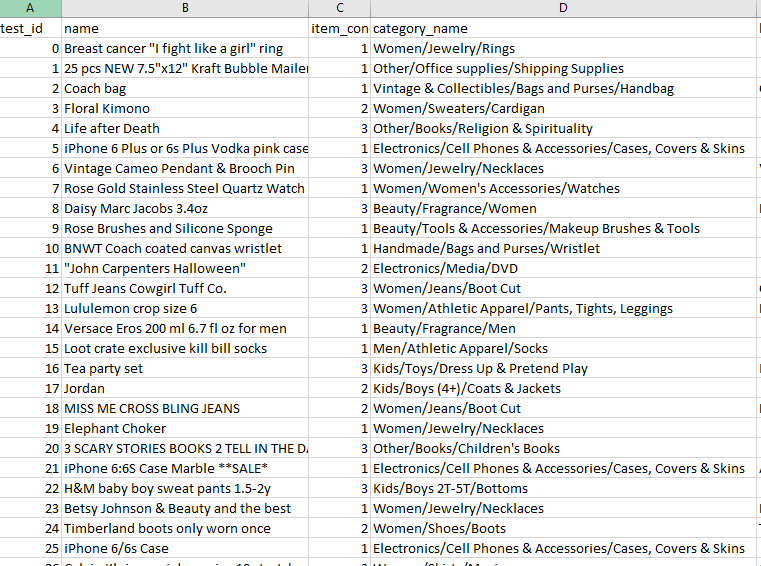
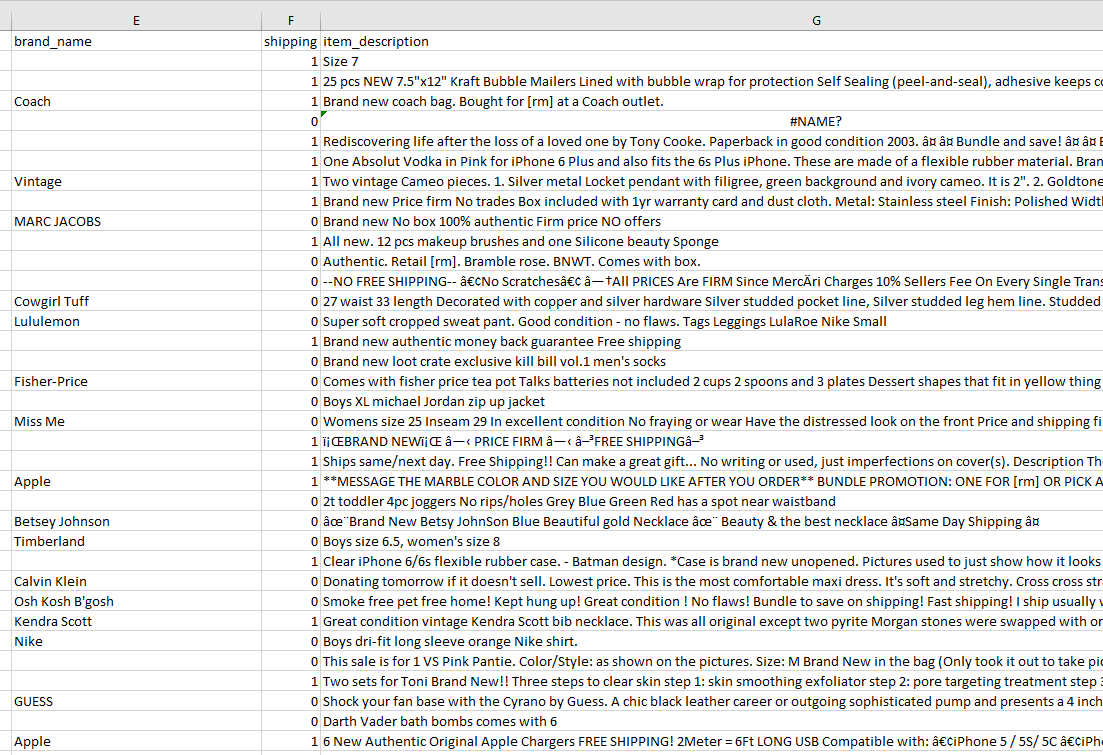
[Mercari Price Suggestion Challenge](https://www.kaggle.com/c/mercari-price-suggestion-challenge)

# Intro

Build an algorithm that automatically suggests the right product prices. You’ll be provided user-inputted text descriptions of their products, including details like product category name, brand name, and item condition.

* Use Python 3
* Features
  + Name: string
  + Item\_condition\_id: Integer
  + category \_name: string
  + Brand\_name: string
  + Shipping: boolean
  + Item\_description: string





**Reference:**  Predict house prices using advanced regression techniques like random forest and gradient boosting <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

**How to get start:** <https://www.youtube.com/watch?v=ez_eVGUjIP8&feature=youtu.be>

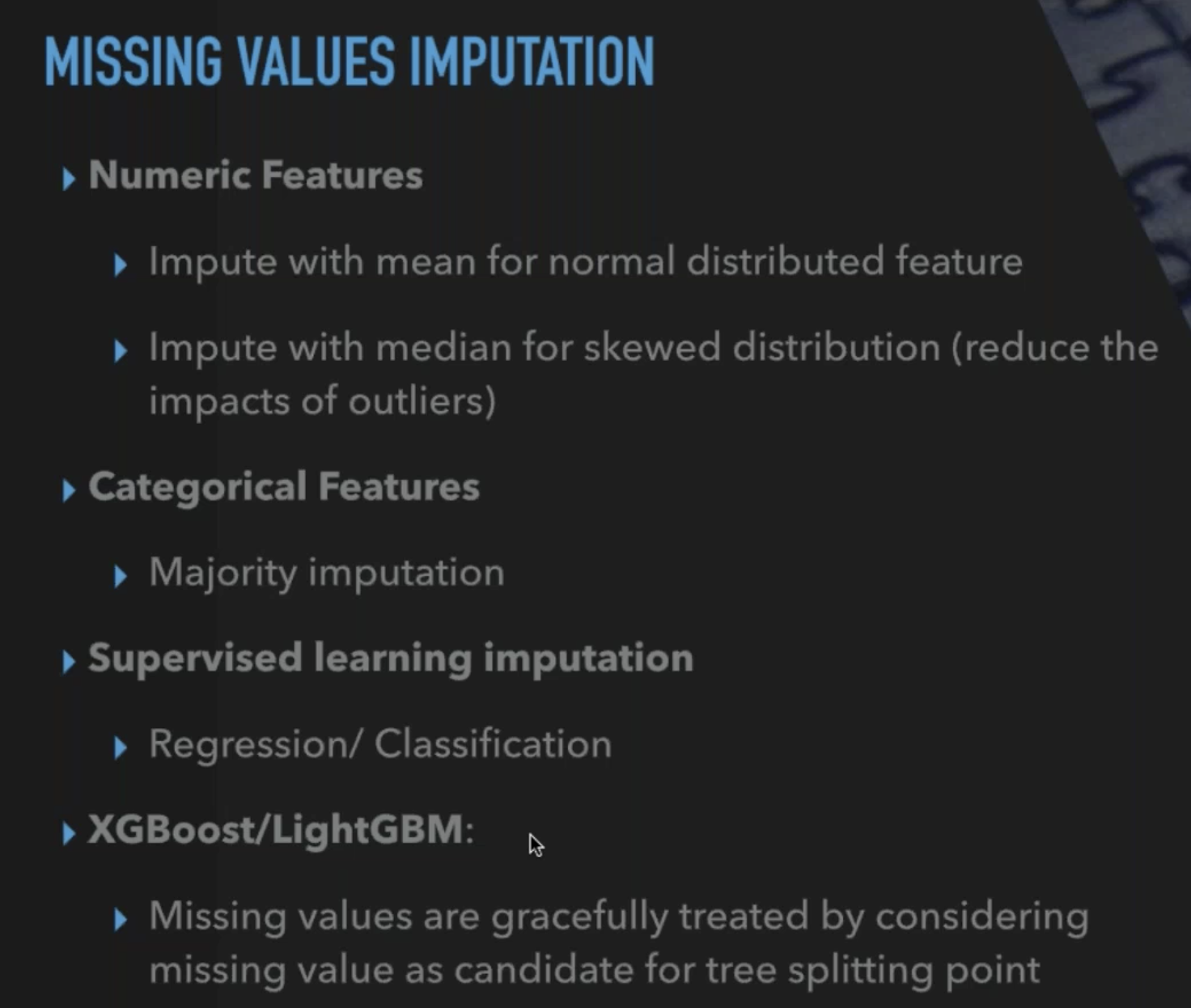
Tutorials: helpful tutorials and competition winners' interviews –[No Free Hunch](https://via.intercom-mail-200.com/e?ob=VT6FIs3MtlumEZwvBdF96qiznQtezYzCIGxsmKhy9OM%3D&h=7d16e110d3ca1a86bead991d01e11e65a5c54c86-12823998665).

Here are a few popular posts:

* [Approaching (Almost) Any Machine Learning Problem](https://via.intercom-mail-200.com/e?ob=skmtFC5JcqzymI3qU%2FycyjxjaGGupw8hx0cpt27AZXmSeKkrUtm9fIhSuSVPf9DZrZlWmPFScgY8CXs7MT2Prj0iJhgbYRQ%2BAa8C%2BeW6ooiMzSN1GlgxbdFhpQMM5v%2FtiekA05L91RUpoJsTJruqNg%3D%3D&h=0bde73f4b4b3199fef3c2c8784e90be29d17fe65-12823998665) by Abhishek Thakur
* [Scraping for Craft Beers: A Dataset Creation Tutorial](https://via.intercom-mail-200.com/e?ob=f5Hq5szPEvP9qCCP6e6WtVBffg3X0aXFB6O2c3tnAZu%2BMUC9UlWT1MApgNokRTH%2FVw2y5XX3Wmr5cmgY1ErvQj3WCs%2FWDEFhBP7UUHUGc8FFA%2F0R5fUTRzB7kvGwNr6T&h=60e9ddd8998559e745fd8b545d8835fe75daab6b-12823998665) by Jean-Nicholas Hould
* [March Machine Learning Mania, 1st Place Winner's Interview](https://via.intercom-mail-200.com/e?ob=XemabQKkEwqi%2Frp8hrqJUIEPkUWEM2e3t%2FycQiDmGGZVuU2%2BG%2BxUUAFjzdPeKU9ydNTmghyWZeW%2FR8QUUkbsw%2B4nWL71HGDQuep75JXFs7HH%2FFKO4ENkscvZrKFpg6Y80l48G59tUx7DpACF7E6Ohg%3D%3D&h=052f554fb4085f79f4500af74ea5143241c84200-12823998665) by Andrew Landgraf

# Data Exploration

* Outliers?
* Cleaning?
* Missing values?



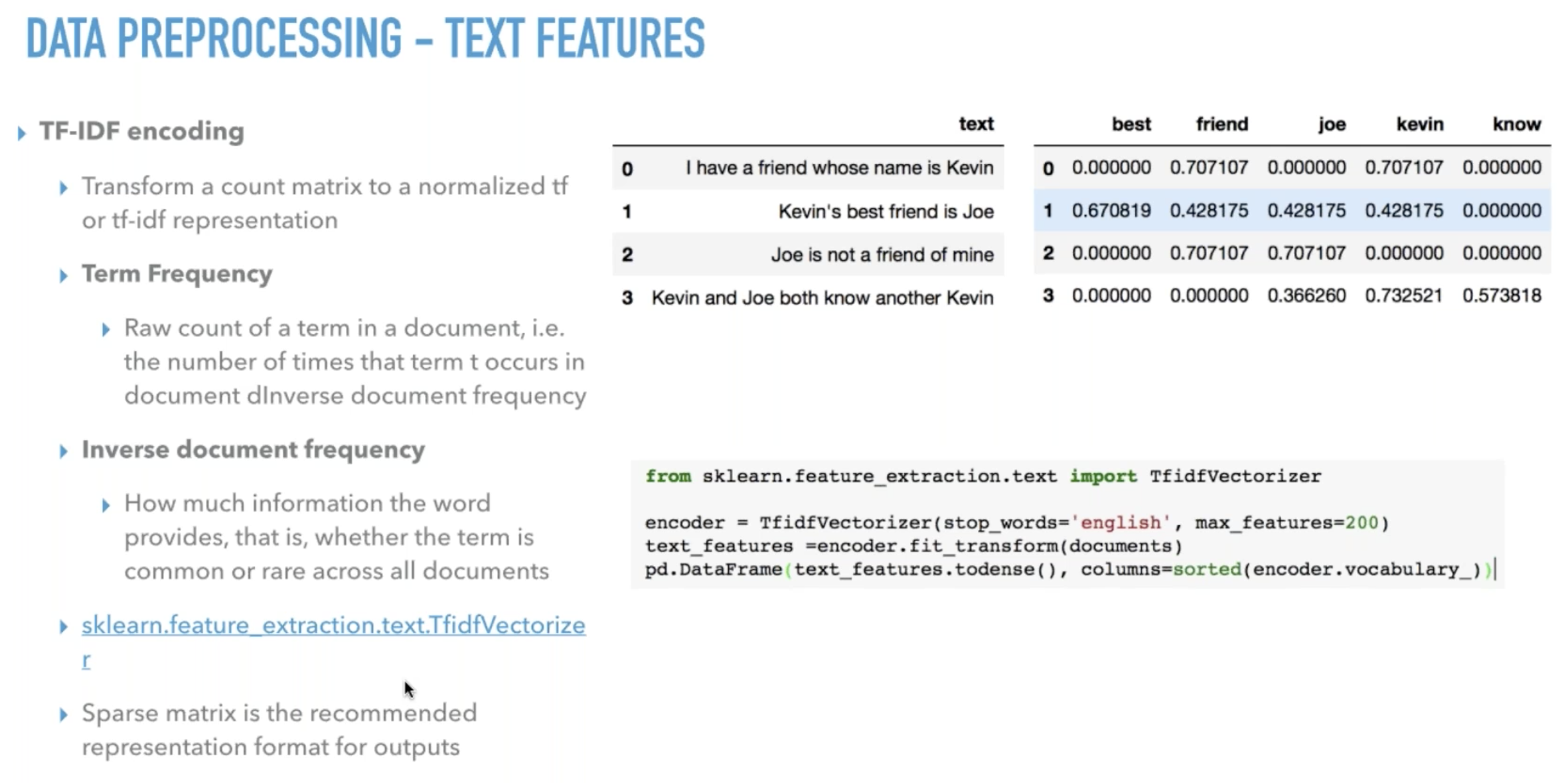
# Feature Engineering

* Transform: PCA
* Selection: KPERCENTILE
* Scaling: Standard scaler

## New features

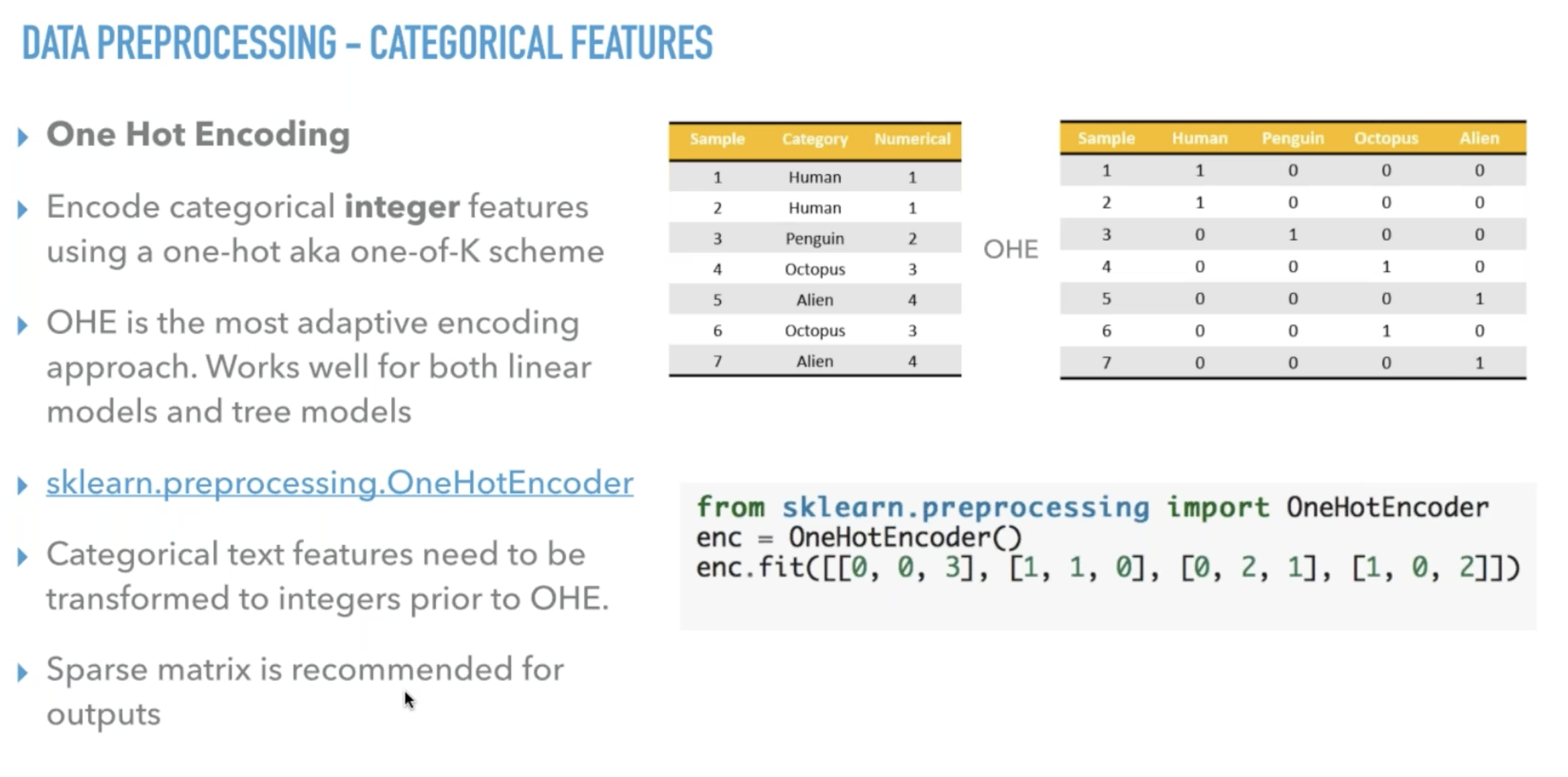
## Text Features

Tfidf?

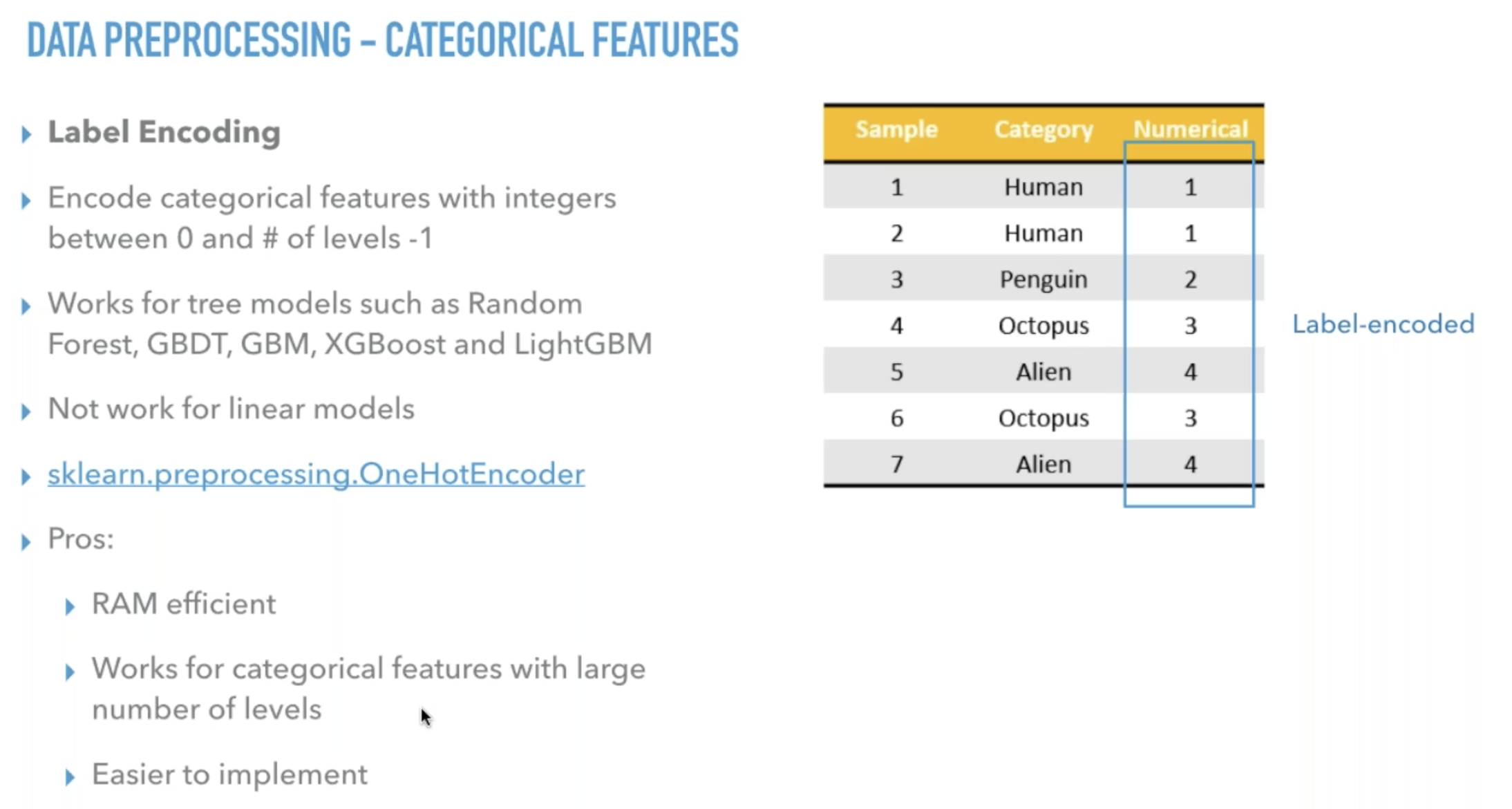


## Categorical features

One hot encoding?



Label encoding?

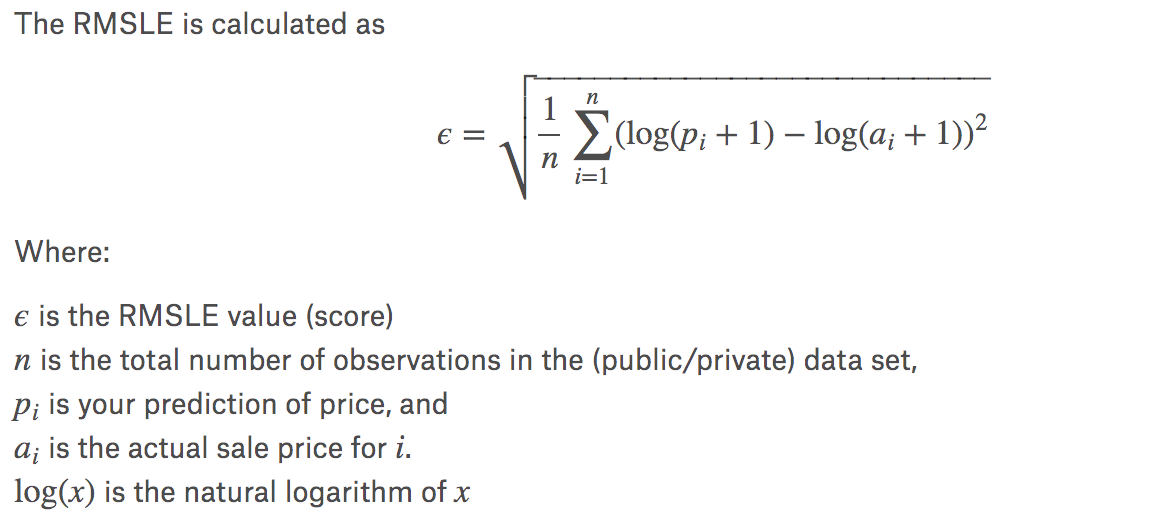


# Algorithm

* Use gridsearchcv
* Supervised regression
  + Linear regression
  + Lasso regression
  + Decision Tree regression
  + SV regression

# Evaluation

* Validate
  + train/test split
  + K-fold
* The evaluation metric for this competition is [Root Mean Squared Logarithmic Error](https://www.kaggle.com/wiki/RootMeanSquaredLogarithmicError).



* Submission File
  + For every row in the dataset, submission files should contain two columns: test\_id and price. The id corresponds to the column of that id in the test.tsv. The file should contain a header and have the following format:

|  |
| --- |
| * test\_id,price * 0,1.50 * 1,50 * 2,500 * 3,100 * etc. |

Ideas:

How to process categories:》Section : category features

Method 1: convert categories to numeric like shipping and brand name

Method 2: add it to text features for a little bit more info

Addition features: -》Section : new features

1. Length of description
2. Sum of tfidf score