Train In Data

Rare Labels

Rare Labels

- Rare labels are those that appear only in a tiny proportion of the observations in a dataset

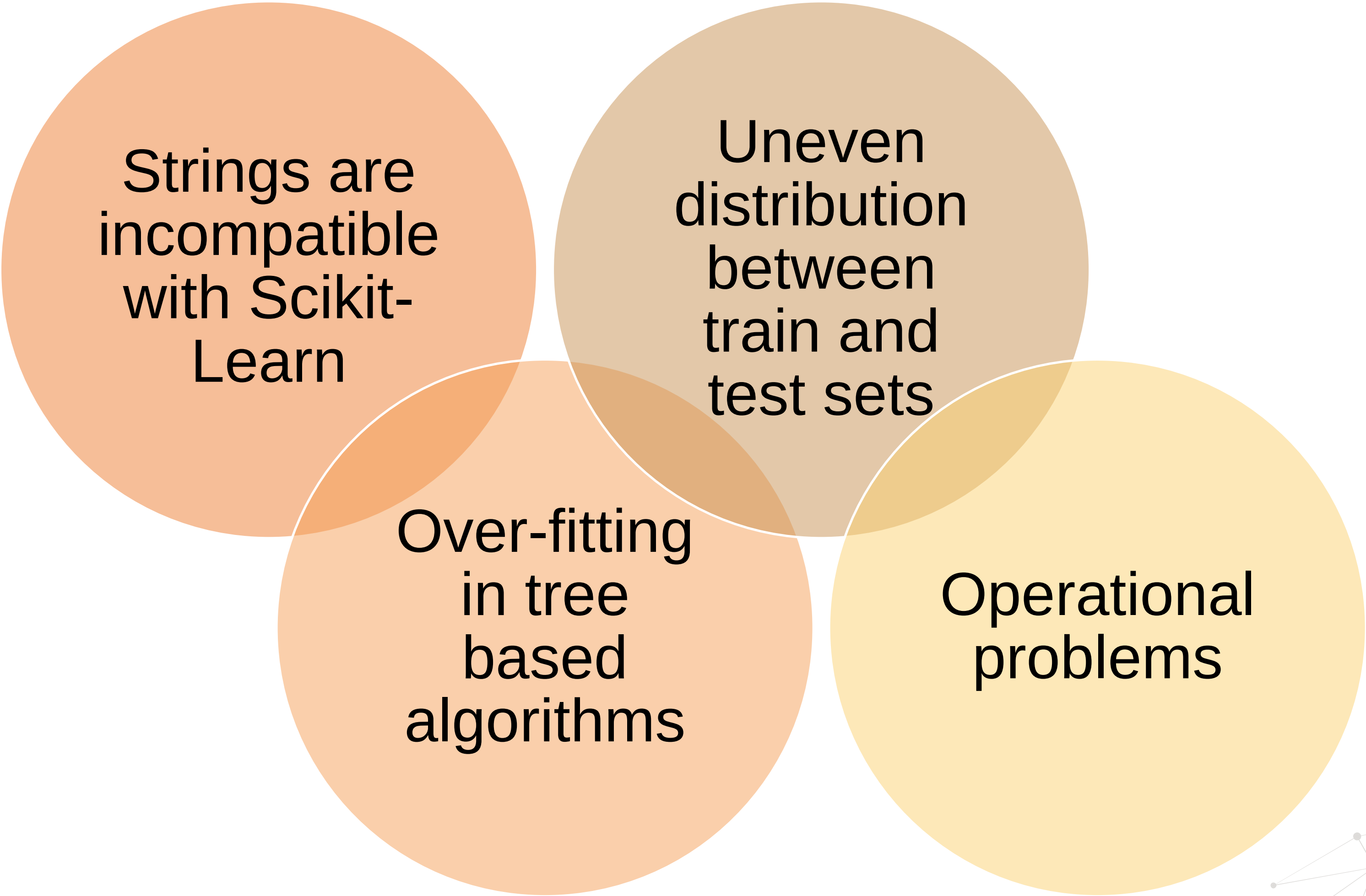
- For example, for the variable “city” where a US citizen lives:
  - New York is a frequent category (many people live in New York)
  - Leavenworth is a rare category (few people live there)

Rare Labels effects

How do Rare labels affect the performance and operationalization of Machine Learning Models?

“Same impacts and considerations as with high cardinality ”

Rare labels impacts



Strings are incompatible with Scikit-Learn

Uneven distribution between train and test sets

Over-fitting in tree based algorithms

Operational problems

Rare label example

Obs	Gender	Vehicle Make
1	Male	Mercedes
2	Male	Ford
3	Male	Ford
4	Male	Renault
5	Male	Seat
6	Male	Renault
7	Female	Citroen
8	Female	Toyota
9	Female	Kia
10	Female	Kia
11	Female	Nissan
12	Female	BMW

Train Set

Obs	Gender	Vehicle Make
1	Male	Mercedes
3	Male	Ford
6	Male	Renault
7	Female	Citroen
9	Female	Kia
11	Female	Nissan

Test Set

Obs	Gender	Vehicle Make
2	Male	Ford
5	Male	Seat
4	Male	Renault
8	Female	Toyota
10	Female	Kia
12	Female	BMW

Rare label example

Obs	Gender	Vehicle Make
1	Male	Mercedes
2	Male	Ford
3	Male	Ford
4	Male	Renault
5	Male	Seat
6	Male	Renault
7	Female	Citroen
8	Female	Toyota
9	Female	Kia
10	Female	Kia
11	Female	Nissan
12	Female	BMW

Train Set

Obs	Gender	Vehicle Make
1	Male	Mercedes
3	Male	**Ford**
6	Male	**Renault**
7	Female	Citroen
9	Female	**Kia**
11	Female	Nissan

Test Set

Obs	Gender	Vehicle Make
2	Male	**Ford**
5	Male	Seat
4	Male	**Renault**
8	Female	Toyota
10	Female	**Kia**
12	Female	BMW

Rare label example

Obs	Gender	Vehicle Make
1	Male	Mercedes
2	Male	Ford
3	Male	Ford
4	Male	Renault
5	Male	Seat
6	Male	Renault
7	Female	Citroen
8	Female	Toyota
9	Female	Kia
10	Female	Kia
11	Female	Nissan
12	Female	BMW

Train Set

Obs	Gender	Vehicle Make
1	Male	Mercedes
3	Male	**Ford**
6	Male	**Renault**
7	Female	Citroen
9	Female	**Kia**
11	Female	Nissan

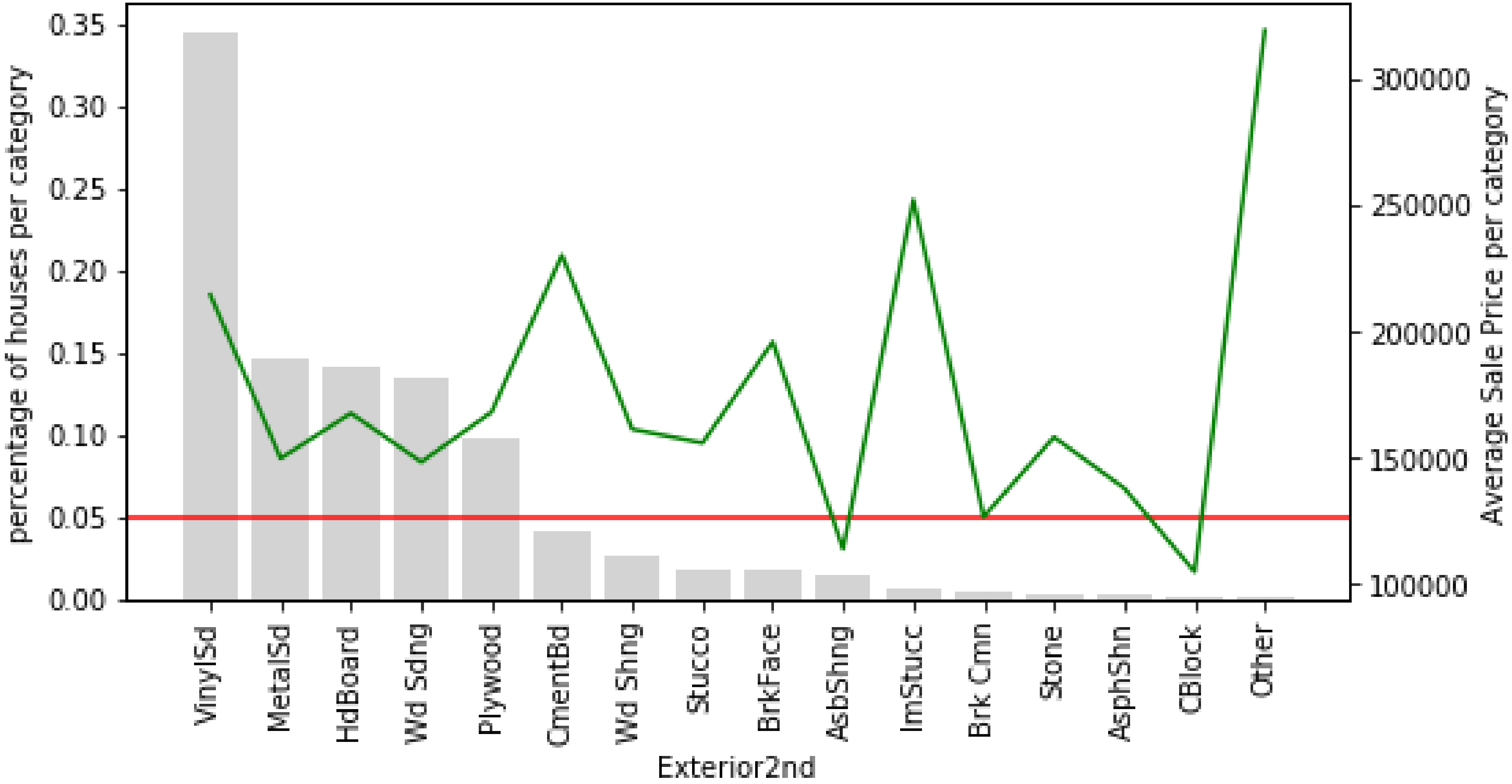
Potential Overfit

Test Set

Obs	Gender	Vehicle Make
2	Male	**Ford**
5	Male	Seat
4	Male	**Renault**
8	Female	Toyota
10	Female	**Kia**
12	Female	BMW

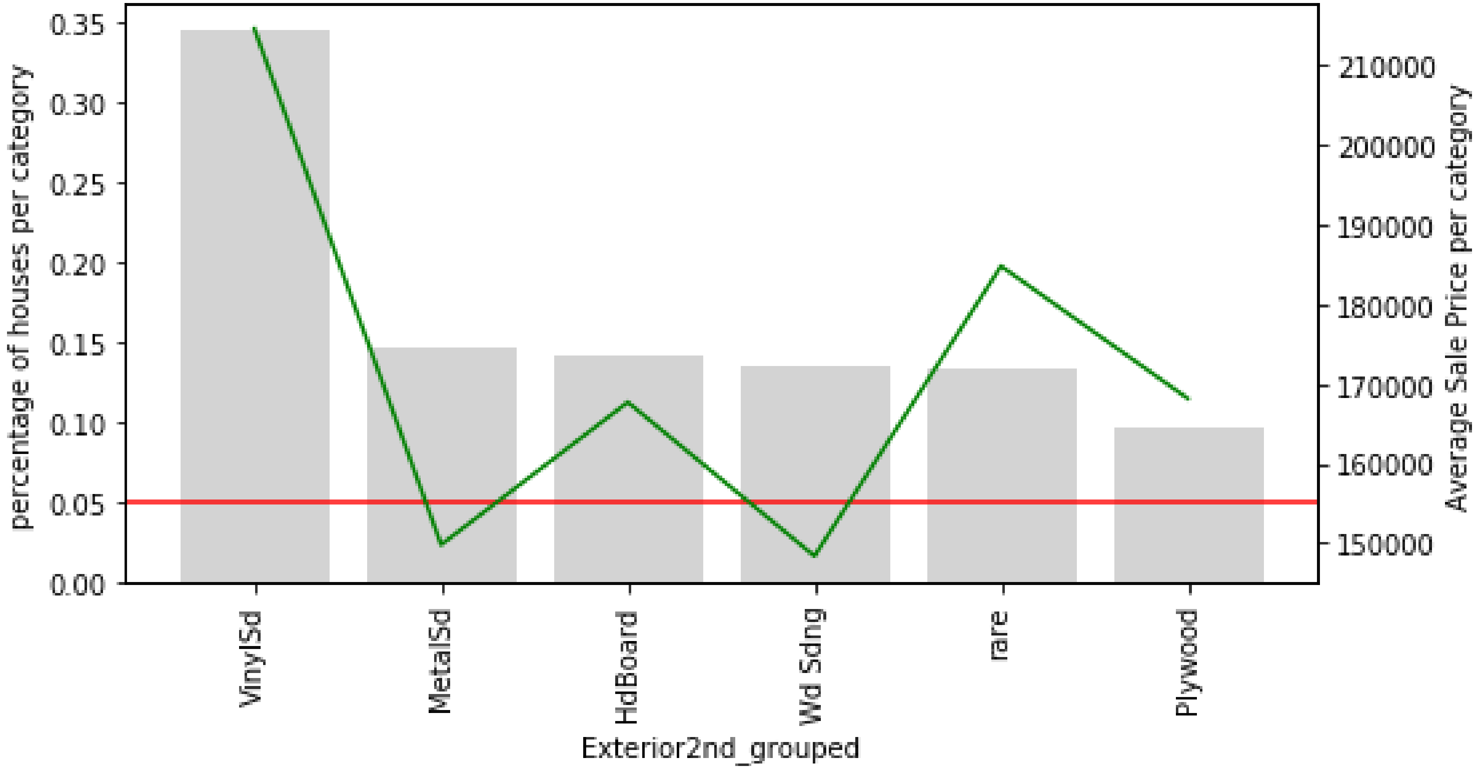
Potential Operationalisation problem

Rare labels and deriving information



Hard to understand the true effect of the rare label on the outcome

Rare labels and deriving information



Hard to understand the true effect of the rare label on the outcome

Summary

- Strings need to be encoded to numbers for use with Scikit-Learn
- Rare labels may cause over-fitting and operationalisation problems
- Hard to understand the role of the rare label on the outcome prediction
- Removing rare labels may improve model performance

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook

- How to quantify category frequency
- Examples of frequent and rare labels
- Example of difficulty to derive reliable information from rare labels
- Example of uneven distribution of rare labels between train and test sets

Train In Data

THANK YOU

www.trainindata.com