# Encoding Categorical Variables

Analytics Vidhya

# Encoding Categorical Variables

| Id | Grade | Target |
|---|---|---|
| 1 | A | 12.0 |
| 2 | B | 11.5 |
| 3 | C | 10.5 |

Analytics Vidhya

# Encoding Categorical Variables

- Label Encoding
- One-Hot Encoding

Analytics Vidhya

Analytics Vidhya

# Label Encoding

| Id | Grade | Target |
|---|---|---|
| 1 | A | 12.0 |
| 2 | B | 11.5 |
| 3 | C | 10.5 |

| Id | Grade | Target |
|---|---|---|
| 1 | 2 | 12.0 |
| 2 | 1 | 11.5 |
| 3 | 0 | 10.5 |

Analytics Vidhya

# Label Encoding

Import StringIndexer class

Create StringIndexer object

```
from pyspark.ml.feature import StringIndexer

# String Indexer object
SI_Obj = StringIndexer(inputCol= "Grade", outputCol= "Grade_le" , handleInvalid="skip")

# Fit on train data
SI_train = SI_Obj.fit(train)

# Transform train data
train = SI_train.transform(train)
```

Fit on data

Transform data

Analytics Vidhya

# Label Encoding

| Id | Grade | Target | Grade_le |
|---|---|---|---|
| 1 | A | 12.0 | 1.0 |
| 2 | B | 11.5 | 0.0 |
| 3 | C | 10.5 | 2.0 |

Analytics Vidhya

# One-Hot Encoding

| Id | Grade | Target |
|---|---|---|
| 1 | A | 12.0 |
| 2 | B | 11.5 |
| 3 | C | 10.5 |

| Id | Grade_A | Grade_B | Grade_C | Target |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 12.0 |
| 2 | 0 | 1 | 0 | 11.5 |
| 3 | 0 | 0 | 1 | 10.5 |

Analytics Vidhya

# One-Hot Encoding

| Id | Grade | Target |
|---|---|---|
| 1 | A | 12.0 |
| 2 | B | 11.5 |
| 3 | C | 10.5 |

| Id | Grade_OHE | Target |
|---|---|---|
| 1 | (2, [1], [1]) | 12.0 |
| 2 | (2, [0], [1]) | 11.5 |
| 3 | (2, [], []) | 10.5 |

Vector

Analytics Vidhya

# One-Hot Encoding

| Id | Grade | Target | Grade_le | Grade_ohe |
|---|---|---|---|---|
| 1 | A | 12.0 | 1.0 | (2,[1],[1.0]) |
| 2 | B | 11.5 | 0.0 | (2,[0],[1.0]) |
| 3 | C | 10.5 | 2.0 | (2,[],[]) |

Analytics Vidhya

# One-Hot Encoding

| Id | Grade | Target | Grade_le | Grade_ohe |
|---|---|---|---|---|
| 1 | A | 12.0 | 1.0 | (2,[1],[1.0]) |
| 2 | B | 11.5 | 0.0 | (2,[0],[1.0]) |
| 3 | C | 10.5 | 2.0 | (2,[],[]) |

Dummy Variable Trap

Analytics Vidhya

# One-Hot Encoding

| Id | Grade_A | Grade_B | Grade_C | Target |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 12.0 |
| 2 | 0 | 1 | 0 | 11.5 |
| 3 | 0 | 0 | 1 | 10.5 |

Analytics Vidhya

# One-Hot Encoding

| Id | Grade | Target | Grade_le | Grade_ohe |
|---|---|---|---|---|
| 1 | A | 12.0 | 1.0 | (2,[1],[1.0]) |
| 2 | B | 11.5 | 0.0 | (2,[0],[1.0]) |
| 3 | C | 10.5 | 2.0 | (2,[],[]) |

Analytics Vidhya

# One-Hot Encoding

```
+--------------+
|       os_ohe|
+--------------+
|(2,[0],[1.0])|
```

Size of vector

Index

Value 1 stored at 0th index

Analytics Vidhya

# One-Hot Encoding

Import OneHotEncoderEstimator class

Create OneHotEncoderEstimator object

```
from pyspark.ml.feature import OneHotEncoderEstimator

# OHE object
OHE_Obj = OneHotEncoderEstimator(inputCols=["Grade_le"],
                                 outputCols=["Grade_OHE"])

# Fit on train data
OHE_train = OHE_Obj.fit(train)

# Transforma train data
train = OHE_train.transform(train)
```

Fit on data

Transform data

Analytics Vidhya

Thank You!!

Analytics Vidhya