

4.4 单变量非線形变换

15t4034s 莊司響之介

ほとんどのモデルは、個々の特徴量が、おおよそガウス分布に従っているときに最もうまく機能する。

→logやexpを使って実現する。

このような変換は、カウントデータで特に有効

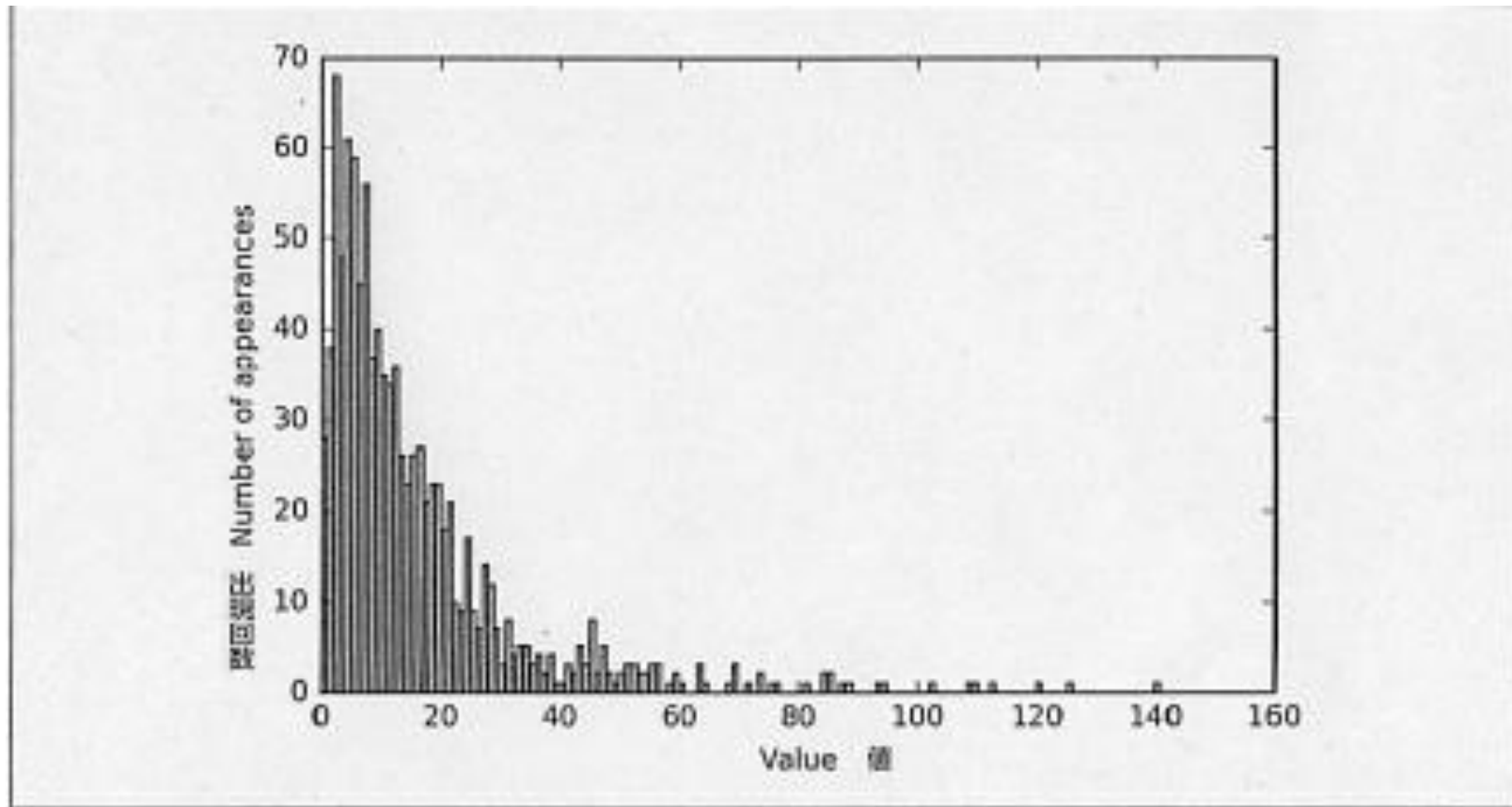


図4-7 特徴量X[0]の値のヒストグラム

上の図は、データXの特徴量X[0]の値のヒストグラムである。このデータを用いてリッジ回帰でモデリングを行った結果、テストスコアは0.622であった。

データ X を対数変換

```
X_train_log=np.log(X_train+1)
```

```
X_test_log=np.log(X_test+1)
```

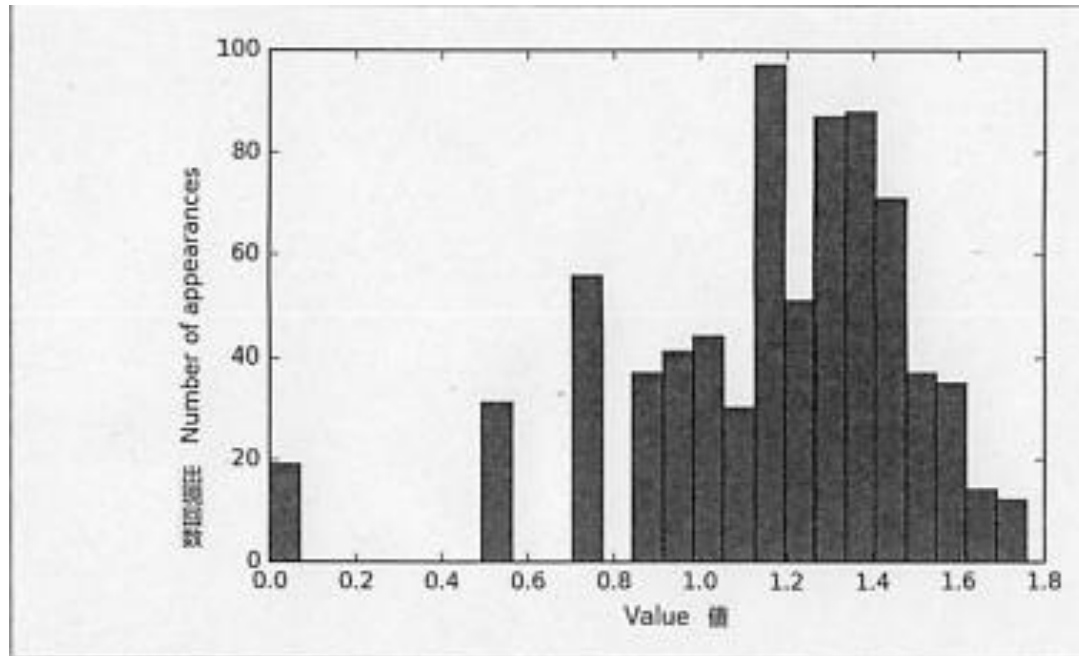


図4-8 対数変換後の特徴量X[0]の値のヒストグラム

テストスコア：0.622→0.875

例のような変換は、決定木ベースのモデルには関係ないが、線形モデルにとっては本質的である。回帰モデルの場合は、ターゲット変数 y も変換した方がよいこともある。