

SIGNATE

医療保険の費用帯予測

2025 / 6 / 27

1. 概要

コンペ名：【第 59 回_Beginner 限定コンペ】医療保険の費用帯予測

タスク：3 クラス分類（0：低, 1：中, 2：高）

使用モデル：LogisticRegressionm, SVM, RandomForest, XGBClassifier

評価指標：F1-score macro

ベースライン：0.7147（LogisticRegression）

最終スコア：0.7375（SVM）

2. 特徴量定義

特徴量名	データ型	データ範囲	概要
id	int64	0~1999	一意に識別する id
age	int64	18~64	年齢
sex	object	male, female	性別
bmi	float64	20.63~47.29	BMI
children	int64	1~5	子供の人数
smoker	object	yes, no	喫煙の有無
region	object	northeast, northwest, southeast, southwest	地区
charges	int64	0, 1, 2	目的変数

3. 前処理, 特徴量エンジニアリング内容

- ・カテゴリ変換

- sex : male / female → 1 / 0

- smoker : yes / no → 1 / 0

- ・ダミー化

- region → region_northeast, region_northwest, region_southeast, region_southwest

- ・標準化

- age : 標準化

- bmi : 標準化

4. プロセス

1. 前処理, 特徴量エンジニアリング

- カテゴリ変換、数値変数の標準化

2. ベースライン設定 (LogisticRegressionm)

- LogisticRegression (0.7147)

3. モデル構築

- SVM, RandomForest, XGBClassifier

4. モデル選択

- RandomForest に決定

5. テストデータに対し予測

5. 備考

本レポートで利用したデータセットは、コンペティション主催者から提供されたものであり、利用規約によりリポジトリに含まれない。