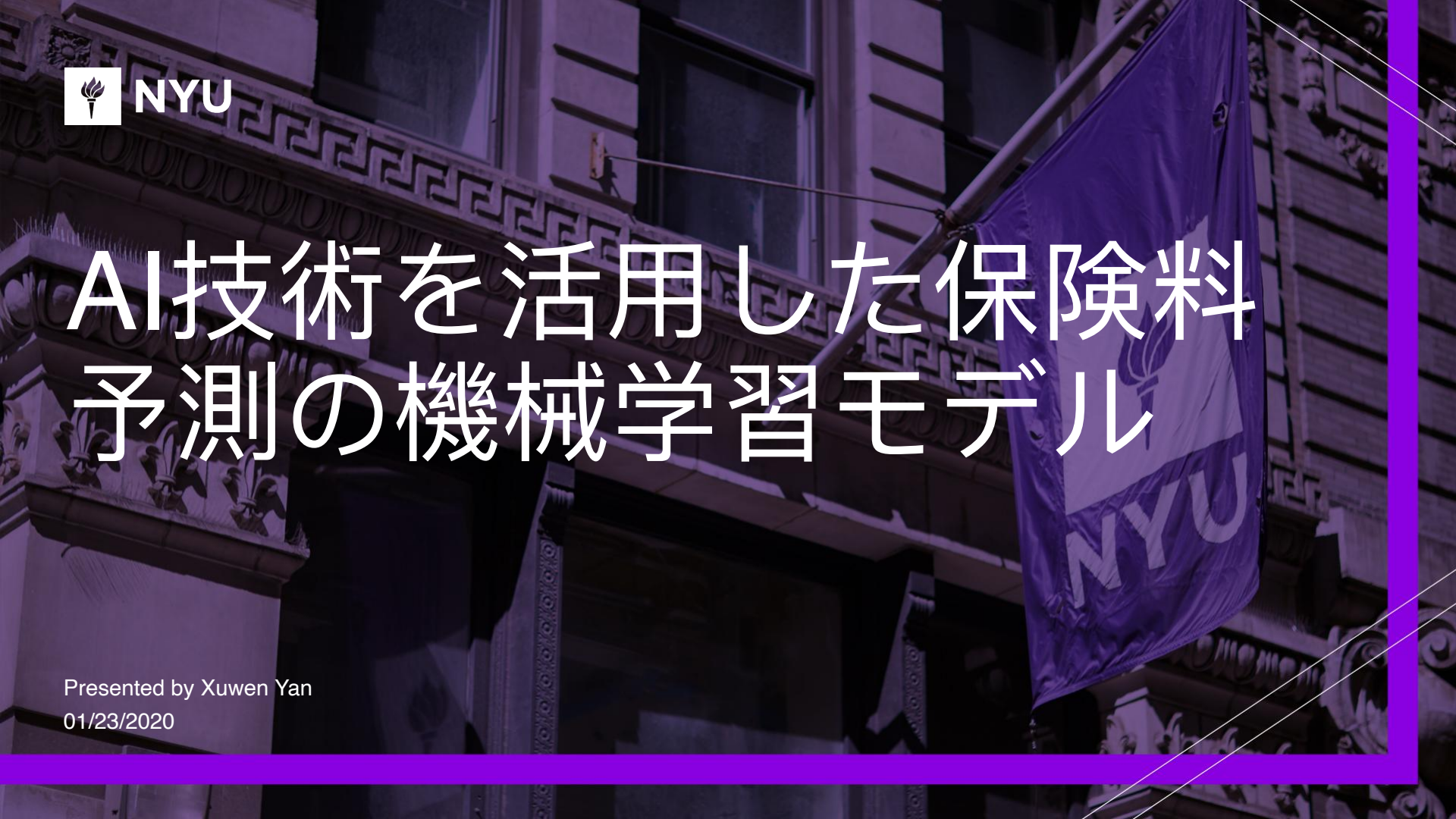


AI技術を活用した保険料予測の機械学習モデル

Presented by Xuwen Yan

01/23/2020

プロジェクトの概要

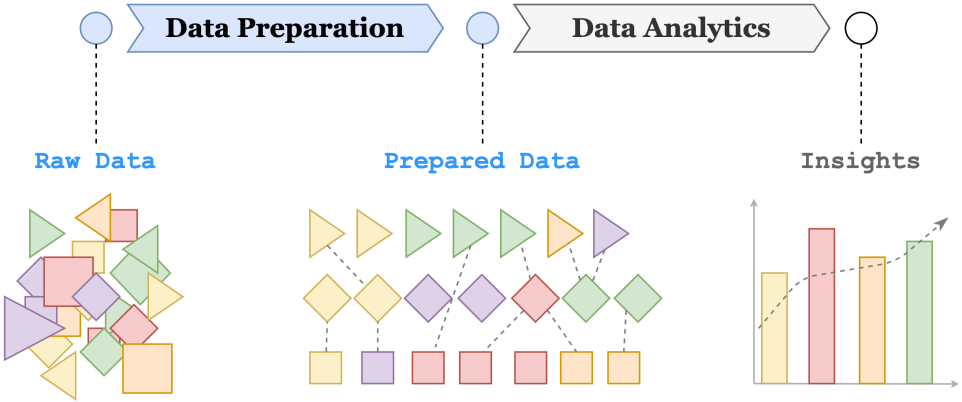
このプロジェクトの目標は、6種類の異なる機械学習回帰モデルを使用して、クライアントの保険料を予測することです。各モデルはその性能に基づいて評価・比較されました。

- データ分析と可視化(EDA)
- データ前処理
- モデルの構築と評価
- 特徴量の重要性分析

01

NYU

データ分析と可視化(EDA)



このデータセットには、年齢、性別、地域、喫煙者ステータス、保険料金などの特徴が含まれています。最初のステップでは、データの視覚化と基本的な統計量を用いてデータを理解しました。

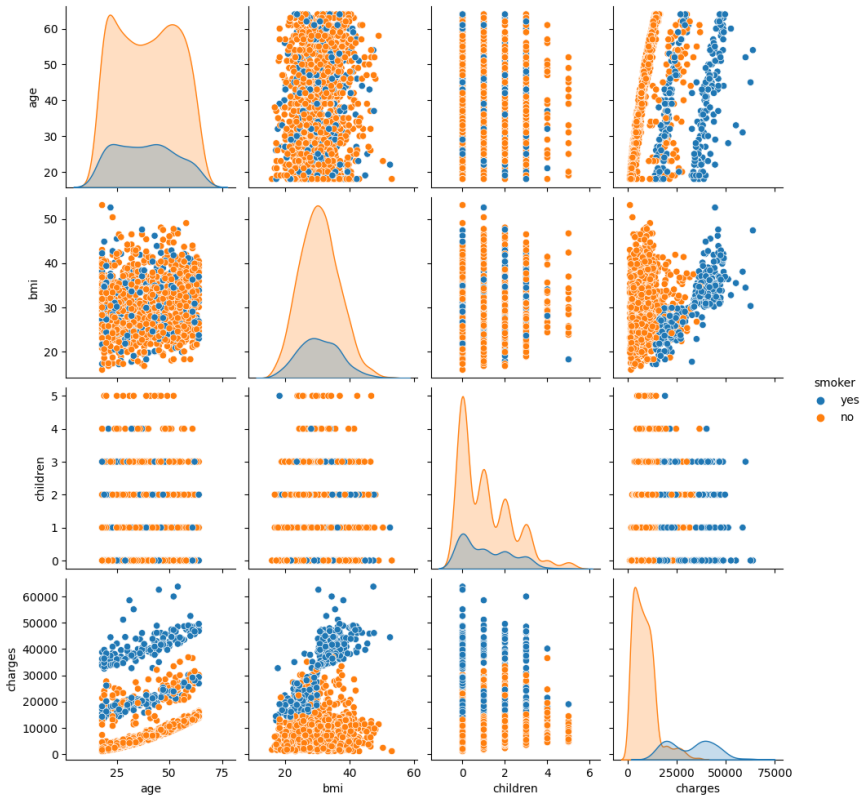
- **データの可視化**：seabornを使用して、喫煙と保険料の関係を示すボックスプロットを作成しました。喫煙者は一般的に保険料が高いことが観察されました。
- **相関行列**：相関関係を示すヒートマップを作成し、年齢、BMI、喫煙者ステータスが保険料に強く相関していることがわかりました。

NYU

喫煙ステータスによる保険料の影響

喫煙者の保険料は全般的に高く、特に年齢とBMIが高い場合、その影響が大きい

- **年齢と保険料**：喫煙者は、特に年齢が高くなると保険料が著しく高くなる。
- **BMIと保険料**：BMIが高い場合、喫煙者の保険料が顕著に上昇。
- **子供の数と保険料**：子供の数が増えても、喫煙者は全体的に保険料が高い。



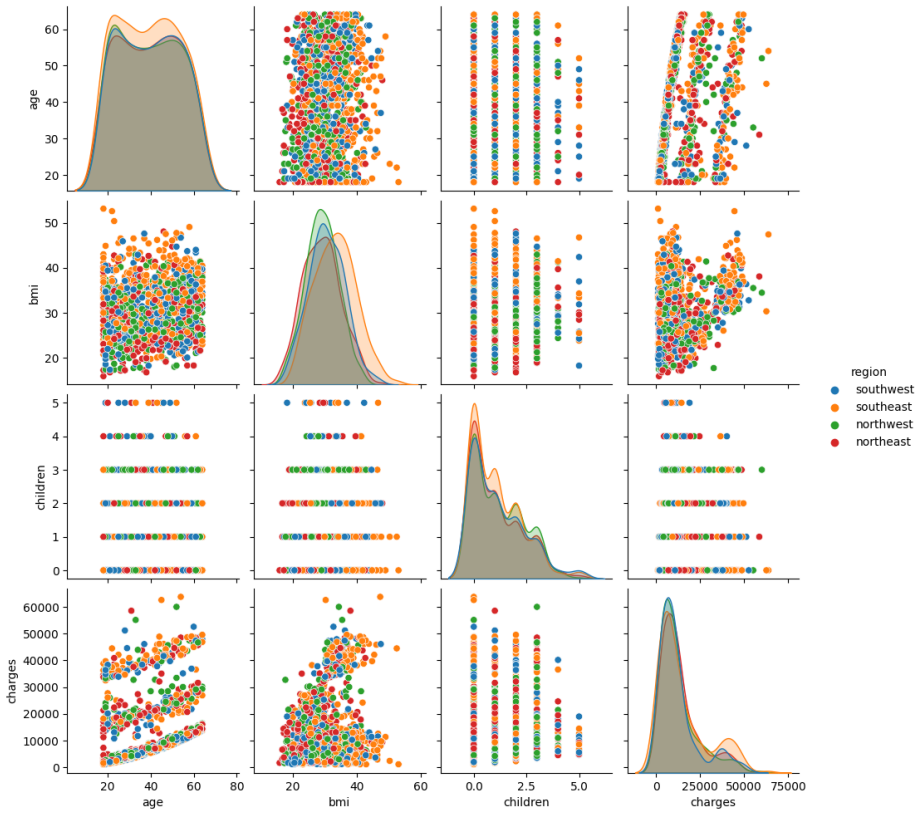

```
sns.pairplot(df, hue ='smoker')
```

NYU

地域ごとの保険料の傾向

地域は保険料に大きな影響を与えないが、年齢やBMIが保険料に影響する。

- **年齢と保険料**：地域にかかわらず、年齢が上がると保険料も上昇する。
- **BMIと保険料**：すべての地域でBMIが高いと保険料も上がる。
- **子供の数と保険料**：地域による大きな違いは見られない。




```
sns.pairplot(df, hue ='region')
```

NYU

データ前処理

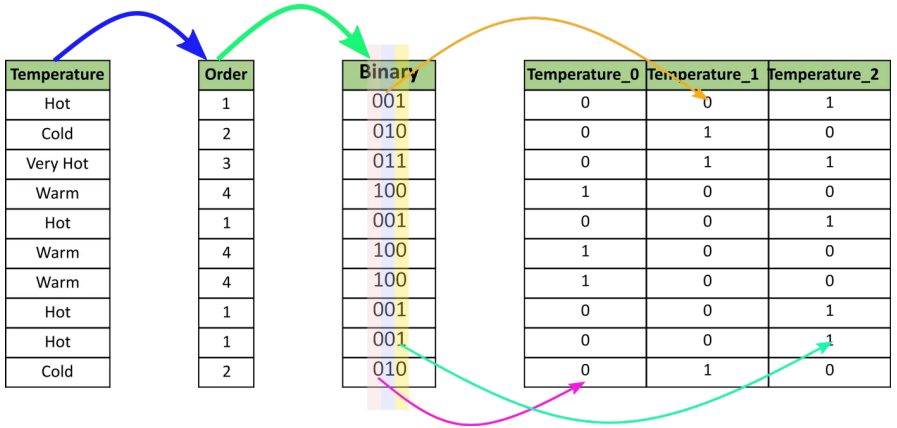
機械学習モデルにデータを入力する前に、いくつかの前処理ステップを実行しました：

- **カテゴリカル変数の処理**：get_dummiesを使用して、'smoker'、'sex'、'region'のようなカテゴリカル列にダミー変数を作成しました。
- **外れ値検出**：IQR（四分位範囲）法を使用して、'charges'列の外れ値を検出し、その影響を評価しました。データの約13%が外れ値として特定されました。
- **スケーリング**：StandardScalerを使用して、すべての変数が同じスケールであることを確認しました。これはKNNのような距離ベースのアルゴリズムに特に重要です。

$$S(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

```
scaler = StandardScaler()
X_train_scaled =
scaler.fit_transform(X_train)
X_test_scaled =
scaler.transform(X_validation)
```

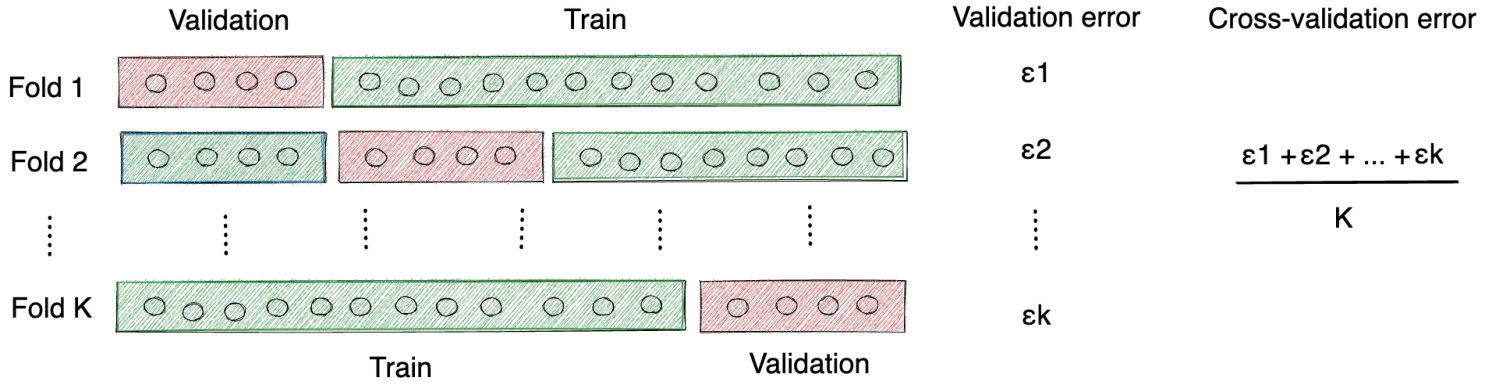
$$\mathrm{IQR}(x) := x_{0.75} - x_{0.25}$$



モデルの構築

1. **線形回帰**：特徴と保険料の間の線形関係を捉える基本的な線形モデル。
2. **Lasso回帰**：過学習を防ぐための正則化を導入したモデル。
3. **ElasticNet回帰**：Lassoとリッジ回帰のペナルティを組み合わせたもの。
4. **K-Nearest Neighbors (KNN)**：最も近い隣人に基づいて値を予測する非パラメトリックな手法。
5. **決定木**：非線形の関係を捉える木構造のモデル。
6. **Gradient Boosting**：弱い学習者を多数構築して性能を向上させる強力なアンサンブル手法。

各モデルは、**10分割交差検証(k-fold cross validation)**を使用して評価され、**R^2**スコアを評価指標として使用しました。



モデルの結果比較

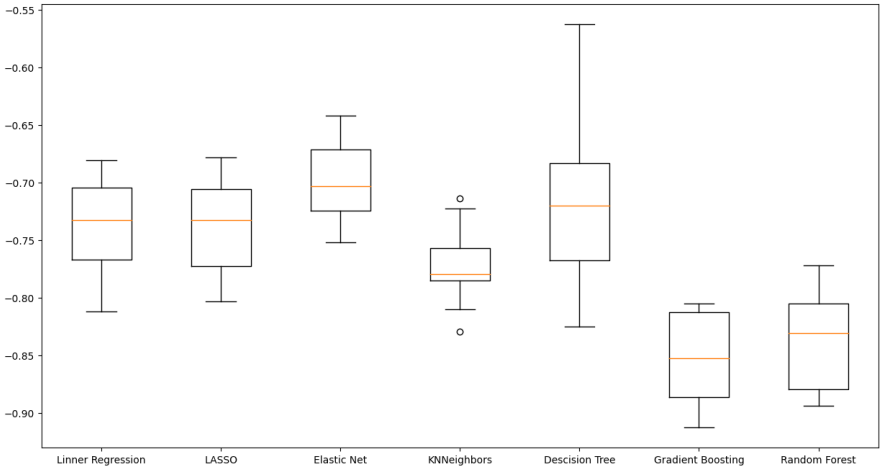
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

各モデルは、**10分割交差検証(k-fold corss validation)**を使用して評価され、**R^2**スコアを評価指標として使用しました。

結果は次の通りです:

- 線形回帰: R^2 = -0.737
- **Lasso回帰**: R^2 = -0.737
- **ElasticNet**: R^2 = -0.699
- **KNN**: R^2 = -0.773
- 決定木: R^2 = -0.713
- **Gradient Boosting**: R^2 = -0.853

これらの結果から、Gradient Boostingが最も高い性能を示し、次いで決定木とElasticNetが続きました。



Total sum of squares	Explained sum of squares	Residual sum of squares
$SS_{tot} = \sum_{i=1}^{m}(y_i - \bar{y})^2$	$SS_{reg} = \sum_{i=1}^{m}(f(x_i) - \bar{y})^2$	$SS_{res} = \sum_{i=1}^{m}(y_i - f(x_i))^2$
Real mean-centered	Prediction mean-centered	Real Prediction Errors

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

NYU

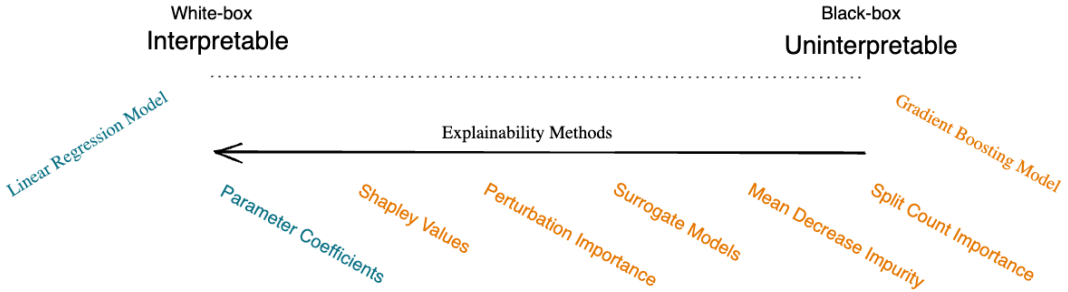
Model Tuning and Grid research

Gradient Boostingモデルのハイパーパラメータを調整するために、GridSearchCVを使用しました。具体的には、n_estimators（木の数）とlearning_rateのチューニングを行いました。

最適な組み合わせは以下の通りです:

- n_estimators = 100
- learning_rate = 0.05

チューニング後、モデルはトレーニングセットで**R^2 = 0.87**、テストセットで**R^2 = 0.86**という高いスコアを達成しました。




```
param_grid = {
    'n_estimators': [50, 100, 150, 200, 250, 300, 350,
400],
    'learning_rate': [0.01, 0.05, 0.1, 0.2, 0.3, 0.4],
}
model = GradientBoostingRegressor()
kfold = KFold(n_splits=num_folds, shuffle=True)
grid = GridSearchCV(estimator=model,
param_grid=param_grid, scoring=scoring, cv=kfold)
grid_result = grid.fit(X_train_scaled, Y_train)
```

NYU

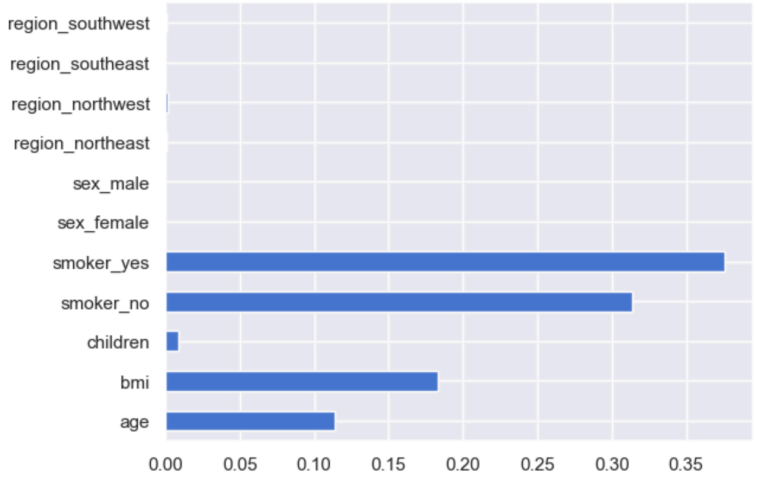
特徴量の重要性

Gradient Boostingの組み込み機能を使用して、保険料の予測に重要な特徴量を分析しました。

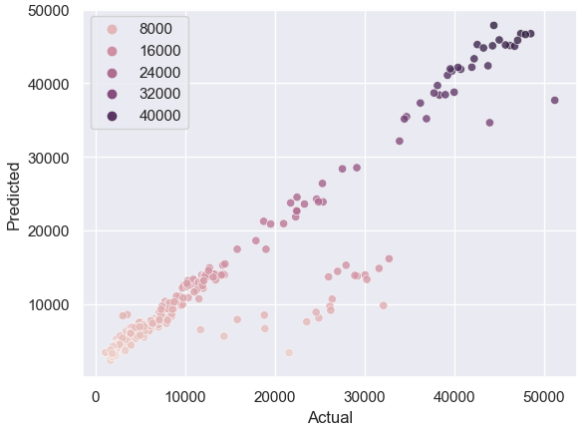
- **最も重要な特徴**は、年齢、BMI、喫煙者ステータスでした。
- また、**SHAP**（SHapley Additive exPlanations）解析を行い、モデルの予測を視覚化し、各特徴がどのようにモデルの出力に影響を与えるかを説明しました。

```
print("Feature Importance: ",model.feature_importances_)
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.plot(kind='barh')
plt.show()
```

```
Feature Importance:  [1.13792387e-01 1.83519638e-01 8.61243454e-03 3.13440660e-01
 3.76062068e-01 3.56846747e-04 2.88315807e-04 9.39321154e-04
 1.98731958e-03 2.51583253e-04 7.49426098e-04]
```

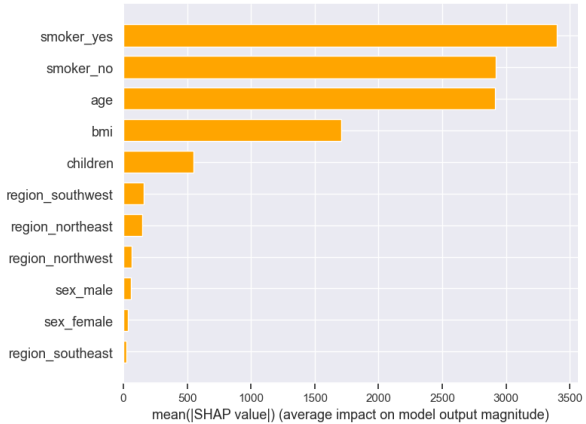


結果



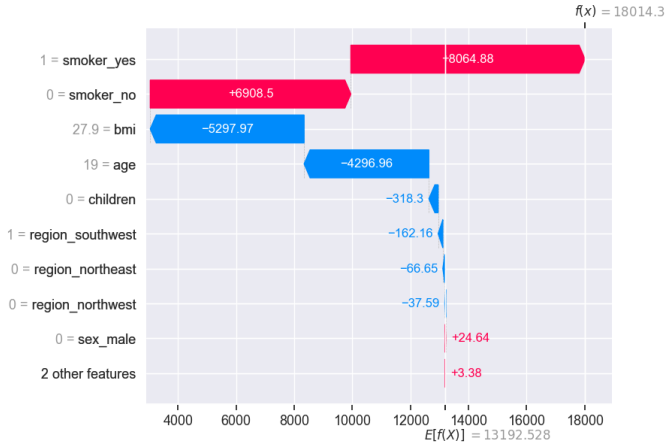
実際値 vs 予測値

モデルの予測値と実際の保険料の値を比較しています。点が対角線に近いほど、予測が正確であることを示します。全体的にモデルは良好な精度を示していますが、高額の保険料では若干の偏差が見られます。



特徴量の重要度

SHAP値を使用して、保険料予測に最も影響を与える特徴量を示しています。`smoker_yes`（喫煙者）が最も大きな影響を与え、次いで年齢とBMIが保険料に影響を及ぼしています。



SHAP Waterfallプロット

特定の予測に対する各特徴量の影響を示しています。喫煙者ステータスが保険料を大きく押し上げており、BMIや年齢も大きく影響しています。

NYU

結論

- Gradient Boostingは最も優れたモデルで、最も正確な予測を提供しました。
- 喫煙者ステータスは保険料を予測する上で最も重要な要因でした。
- Hyper parameterの調整により、モデルの性能が大幅に向上し、保険料予測に対する堅牢な解決策となりました。

NYU


“
THANKS FOR WATCHING
NYU