

予測システムの 開発と評価

総合情報学科

知能情報システム研究室

永井ゼミ 4年 門間大知

関連研究 参考ページ 使用データ

- ◇ <https://urtrip.jp/> 2022年～今日に至るまで毎時のデータ(2021以前は欠損地が多く扱えないため)
- ◇ https://nttdocomo-developers.jp/entry/202212241200_2 CSV参考
- ◇ <https://jssst.or.jp/files/user/taikai/2023/papers/54-R-S.pdf> 機械学習を用いたテーマパークのアトラクション待ち 時間予測手法の確立(重回帰分析、曜日ごと)

◆作成するシステム

- ◆ アトラクションごとの混雑予測、待ち時間を予測する。
- ◆ その日の天気、祝日などの要因を組み合わせて予測する。
過去の期待度などから新しいアトラクションなどが出た場合の予測もしていく。
- ◆ チケット高騰による待ち時間への影響なども予測していく
売り上げ向上予測からその日の入園数などを予測

開発詳細

収集方法

過去待ち時間など、から研究材料を収集。

PythonによるWebスクレイピングを用いてデータを収集し、csvファイル
化データの特徴や傾向を整理

データのクリーニングで欠損地や異常値を処

環境

- ◆ Jupyter Notebook(ジュピター・ノートブック)を使う
- ◆ pandas(パンドラス)を使いデータフレーム作成
- ◆ scikit-learn(サイキット・ラーン) データの前処理や、機械学習のアルゴリズムを使った学習・予測、そしてモデルの評価など、データ分析や機械学習に必要な機能をひとまとめたパッケージ
- ◆ Matplotlib(マットプロットリブ)を使用してデータを可視化

モデル

◆重回帰分析

- ◆ 複数の要因が特定の結果にどのように影響しているかを理解するための統計的手法

◆LightGBM(ライトGBM)

- ◆ 機械学習における分析アルゴリズムで、与えられたデータから、目的となる変数を表現する「教師あり学習」と呼ばれる分野のデータ分析方法の1つ。目的変数に応じて、説明変数を「分類」するための手法で、高精度で信頼性が高く、また汎用性も高いことが特徴。

◆GBDT(勾配ブースティング木)

- ◆ アンサンブル学習によって、一つ前の決定木の誤りを修正するように決定木を作っていく、バイアスが低く予測精度が高いモデルを作る学習手法

評価

◇ 平均絶対誤差(MAE)

予測値と実際の値との差(誤差)の絶対値の平均を取ったもの

◇ 平均二乗誤差(MSE)

予測値と実際の値との差(誤差)を二乗したもの(二乗誤差)をデータ点数で割ったもの

◇ 二乗平均平方根誤差(RMSE)

実際の値と何らかのモデルに基づく予測値があるとき、両者の差を二乗して平均し、平方根(ルート)を取った値

◇ 、決定係数(R²スコア)

統計学において独立変数(説明変数)が従属変数(目的変数)のどれくらいを説明できるかを表す値

今後の目標

- ◇ 天気などの要因を加味して予測を行う。
 - ◇ チケット高騰などが起こった場合の混雑予測を行う
 - ◇ 様々なモデルを用いて予測を行い結果の変化を考察する
-
- ◇ 天気などの要因をどのように予測と組み込んでいくか。
 - ◇ チケット高騰の予測の正確さの確認方法
 - ◇ 関連研究など

メモ（スライドとは関係ありません）

線形回帰（重回帰）

1. Ridge回帰
 2. Lasso回帰
 3. ElasticNet回帰
 4. RandomForest回帰
 5. SVR（サポートベクターマシーン）
-
6. やること
 - ◇ モデル、評価まとめを詳しく なぜそのモデル、評価を使うのか。
 - ◇ 図を載せる（CSVファイルした結果など）
 - ◇ スケジュール

- ◇ 重回帰分析メリット

- ◇ より現実に対応した分析ができる

- ◇ 重回帰分析は、複数の変数を同時に考慮できるため、現実の複雑な状況をより正確に反映した分析が可能だ。

- ◇ デメリット

- ◇ 費用対効果の高いマーケティングにつなげられる

- ◇ 重回帰分析を用いることで、マーケティング活動の各要素が売上にどの程度影響しているかを定量的に評価できるようになる

- ◇ 多重共線性により分析が困難になるケースも

- ◇ 重回帰分析では、分析に用いる変数間で高い相関がある場合、正確な結果を得ることが困難になることがある。

- ◇ LightGBM（ライトGBM） メリット

- ◇ ビッグデータ高性能

- ◇ デメリット パラメータチューニング必要

- ◇ GBDT

- ◇ 精度が高い 汎用性が高い

- ◇ 学習に時間がかかる 過学習を起こしやすい その分ノイズに対して適合してしまうリスクも大きいです。

◇ 平均絶対誤差 (MAE)

外れ値の影響を受けにくいという利点があります。また、元の数値から単位が変化しないため、シンプルで理解しやすい指標です

◇ 平均二乗誤差 (MSE)

◇ 予測誤差が大きいほどMSEは大きくなるため、MSEが小さいほどモデルの予測性能が高いと言えます。

◇ MSEは誤差を二乗するため、大きな誤差を重視する特性があります。一方で、二乗されているので指標の単位がわかりにくいという欠点があります。

◇ 二乗平均平方根誤差 (RMSE)

◇ 値が小さいほどモデルの予測精度が高いことを示します。MSEの平方根を取るため元の数値と単位が同じになり、解釈がしやすくなっています。

、決定係数 (R²スコア)

◇ 決定係数は全体の変動に対する誤差の割合を示しており、0から1の範囲の値を取ります。1に近いほど良いモデルであることを示し、逆に0に近い場合はモデルの説明力が低いことを示します。

◇ 説明変数を増やすほど値が1に近づく傾向がある点には注意が必要です

評価指標	説明	メリット	デメリット
平均絶対誤差(MAE)	予測値と実際の値との絶対誤差の平均	予測エラーの大きさを直感的に理解できる	大きな誤差を重視しない
平均絶対パーセント誤差(MAPE)	予測値と実際の値との相対的な誤差の平均	予測エラーが実際の値に対してどれだけ大きいかをパーセンテージで表せる	実際の値が0に近い場合、誤差が無限大になる可能性がある
平均二乗誤差(MSE)	予測値と実際の値との誤差を二乗したものの平均	大きな誤差を重視する	単位が二乗されるため、直感的に理解しにくい
二乗平均平方誤差(RMSE)	MSEの平方根	予測エラーの大きさを直感的に理解でき、大きな誤差を重視する	単位が元の尺度と異なるため、解釈が難しい場合がある
対数平均二乗誤差(RMSLE)	予測値と実際の値の対数を取ったものの二乗誤差の平均の平方根	比率を重視し、絶対的な誤差よりも相対的な誤差を重視する	対数を取るため、負の予測値を扱うことができない
決定係数	モデルがデータにどれだけフィットしているかを示す指標	モデル全体の性能を一つの数値で評価できる	モデルが複雑になるほど1に近づくため、過学習を引き起こす可能性がある