

機械学習を用いた回転寿司店の販売数量予測

苫小牧工業高等専門学校 ○伊藤将大, 三上 剛, 山本椋太, 三河佳紀 (株)久恵比寿 畑中 稔

要 旨

外食産業における食品ロスの削減のためには、高精度な需要量予測が必要とされている。本研究では、苫小牧市周辺に店舗を持つ(株)久恵比寿の回転寿司店における寿司ネタの販売数量に対し、気象データ(天候、最高気温、降水量、湿度)、地域特性(小中学校の行事、各種イベント等)、その他の要因(緊急事態宣言の発令、新型コロナウイルス感染者数など)に関する各種データを加え、機械学習を用いて高精度な販売数量の予測を目指す。

1. はじめに

日本では2030年度までに事業系の食品ロス量を273万tまで削減し、2000年度比で半減する目標があり[1]、そのための有効なアプローチの1つとして、販売数量およびその元となる来客数の予測を正確に行うことが重要である。

本研究では、苫小牧市内の回転寿司店における日毎の来客数の予測を行う。このような時系列データの予測は古くから行われており、たとえばSARIMA[2]や近年ではLSTM[3]などがその手法として用いられている。しかし、本研究では説明変数として、量的変数のみならず質的変数も含むため、これらの手法ではなく、決定木のアンサンブル学習であるランダムフォレストを用いることにし、それにトレンド予測を組み合わせることにした。

2. 手法

2.1. データ

使用するデータは、(株)久恵比寿が展開している苫小牧市内の回転寿司店における2021年4月1日から2022年1月31日までの10ヶ月の来客数であり、それを図1に示す。10ヶ月に計12日ある定休日は除外している。このデータを機械学習で用いるために、4月1日から11月29日までの前半の80%を訓練データ、11月30日以降の残り20%をテストデータとして用いた。なお、図1における縦軸の来客者数の値は非表示としている。

学習する際には、図1のデータに販売数量に影響を与えると思われる様々な要因をデータとして追加した。まず、土日祝日に来客者数が多いことから、曜日と祝日か否かを識別する値を加えた。そして、気象庁から取得した天候、最高気温、最低気温、平均気温、18時の気温、18時の湿度、日積算降水量の7つの気象データを加えた。さらに、直近1日前の来客者数、2日前の来客者数、3日前の来客者数、および1週間前の来客者数のデータを加えた。なお、1週間前の来客者数のデータを追加する際に、1週間前が定休日であった場合、その前日と翌日の来客者数の平均を使用した。

また、北海道のホームページより期間中の新型コロナウイルス感染者数、緊急事態宣言の発令の有無などのデータを加え、さらに、苫小牧市のホームページから、市内の小中学校の入学式、卒業式、運動会などのイベントの有無も加えた。

2.2 評価方法

モデルの評価は、RMSE(Root Mean Squared Error)と R^2 (R Squared)を評価の指標として使用する。RMSEの式を式(1)

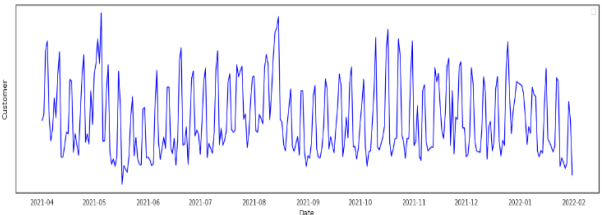


図1 来客者数の時系列データ

に記す。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

式(1)において y_i は観測値、 \hat{y}_i は予測値で、これらの差の2乗の平均を取る事でモデルの評価を行うため、モデルの精度が高くなるほどRMSEの値は小さくなる。 R^2 を式(2)として以下に記す。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

式(2)において y_i は観測値、 \hat{y}_i は予測値、 \bar{y} は平均値である。式(2)の右辺において分数の分子は残差平方和、分母は全平方和である。 R^2 は1から残差平方和を全平方和で除算した結果(0以上)を引く計算式で算出される。 R^2 はモデルの精度が高いほど1に近づく。

2.3.トレンド抽出

今回、時系列データを周期成分とトレンド成分に分解する事で、予測精度の向上を狙った。以下にその手順を記す。

式(3)の y を来客者数、 x を日付として15日間のデータに線形近似を行うことで a_0 と a_1 を決定し、 x に16を代入することで16日目のトレンドを予測する。

$$y = a_0 + a_1 x \quad (3)$$

これを繰り返して16日目以降の来客者数のトレンドを作成する。なお、線形近似には、Pythonの数値解析ソフトウェアであるScipyの`cure_fit`関数を用いた。(3)式で予測したトレンドの値を来客者数から減じた差分のデータに対して機械学習が学習する。

なお、トレンドを作成するのに15日間のデータを使用する都合上、トレンド抽出を用いた予測を行う際には、訓練データの最初の15日間のデータを除外している。

2.3. ランダムフォレスト

ランダムフォレストは、決定木アルゴリズムのアンサンブル学習により予測する標準的なものを用いた。ランダムフォレストの深度と、葉ノードの数がパラメータとなるが、これらの値を変えてテストデータに対する RMSE と R^2 による評価を行い、その評価の平均を比較する事で、使用した特徴量毎に最適な深度と葉ノードの数を見積もった。

3. 結果と考察

3.1. 来客者数の予測結果

データの属性の中で、土日および祝日か否かという情報が最も予測の精度に影響を与えることが分かった。また、8月と5月、新型コロナの日陽性者数のデータによる影響も大きかった。ランダムフォレストのパラメータは、深度を6、葉ノード数を9に設定したものが、最も予測精度が高くなった。この設定でランダムフォレストによる予測結果を図2として以下に記す。なお、実際の来客者数を青線、学習データの予測結果を赤線、テストデータの予測結果を緑線として記す。尚、図2のテストデータのRMSEは78.92、 R^2 は0.71だった。図2より前日との来店者数の差が大きい時の予測精度が低くなっている事がわかる。

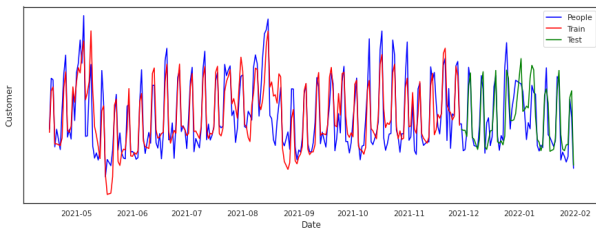


図2 トレンド抽出とランダムフォレストによる予測結果

3.2. 特徴量の考察

図3に今回使用したデータでランダムフォレストが予測するのに用いた各特徴量の重要度を示す。その結果、日曜、土曜、休日、コロナ新規日陽性者数、8月、5月、降水量では十分であることが分かった。特に土日祝日は来客者数予測に最も重要な因子であることが分かった。また、5月と8月が高い値だったのは、ゴールデンウィークおよびお盆の時期によるものと考えられる。

一方、小中学校の行事は、予想に反してほとんど影響がないことが分かった。気象データは降水量の影響がやや高かった。この2つは地域の特性によると思われるが、その理由は今のところはっきりしない。今後検討が必要と考えられる。

また、新型コロナの日毎の新規感染者数が比較的大きく寄与しており、感染者数が増えると来客者数が増えるという結果が見られた。これは一見すると矛盾するが、感染者数が増加した時期がゴールデンウィークやお盆の時期と重なっており、これら連休の時期に人の流れが多くなり来客者数が増えたことを間接的に表したものと考えられる。

3.3. トレンドの考察

図2との比較のために、トレンド抽出を行わずランダムフォレストのみ用いた場合の予測結果を図4に示す。またトレンドのみ抽出した結果を図5に記す。

図4のテストデータのRMSEは91.43、 R^2 は0.47であるため、トレンド抽出と併用した場合の方が、RMSEで12.51、 R^2 で0.23だけ向上している。そのため、トレンド抽出が有効に機能しているものと考えられる。

特に図2と図4を比較すると、図4のランダムフォレスト

のみで予測した場合は、来客者数の高いピークの予測は優れているが、低いところの予測があまり優れていない。トレンド抽出を行うことで、ピークの予測のみならず来客者数が低いところの予測精度も向上している。しかし、図2を見ると、過剰に低く予測しているところや、高く予測しているところも散見されるが、全体的には、トレンド抽出が予測に良い結果をもたらしている事がわかる。図5のトレンドの値のみをみると、土日祝日の後はピークの影響を受けてトレンドの値も高めにでており、より良い予測精度を出すためにはトレンド抽出の方法を改めなければならない事もわかる。

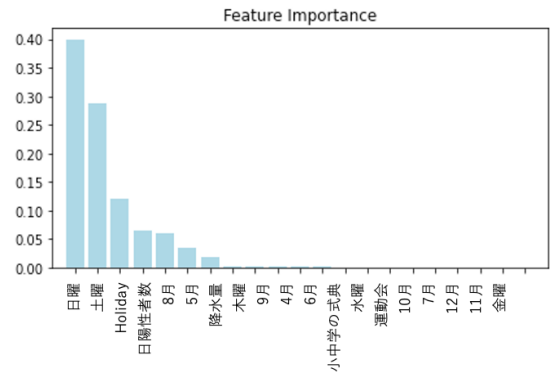


図3 データの各特徴量の重要度

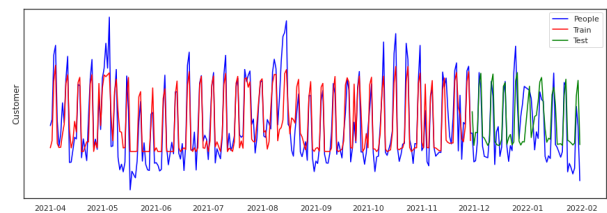


図4 トレンド抽出を行わなかった場合の結果

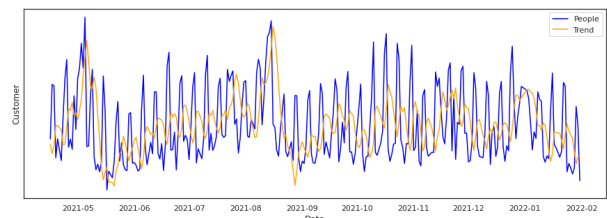


図5 作成したトレンド

5. おわりに

本研究は(株)久恵比寿が苫小牧市内に展開する回転寿司店の売上予測をするために、ランダムフォレストおよびトレンド抽出を組み合わせ、同店舗の来客者数の予測を行った。今後の課題として、予測精度の向上が見込める特徴量の追加についてさらに検証を行う。また、今回使用しているトレンド抽出が前日の影響を受け過ぎている事がわかったため、トレンドを生成する方法を改善する必要がある。

参考文献

- [1] 農林水産省 食品ロス量の公表
<https://www.maff.go.jp/j/press/shokusan/kankyoi/210427.html>
- [2] Naim, I., et al, Effective short-term forecasting for daily time series with complex seasonal patterns. *Procedia Comput. Sci.* **132**, 1832–1841 (2018)
- [3] Greff, K., et al, LSTM: A Search Space Odyssey, *IEEE Trans. Neural Netw. Learn. Syst.* **28** (10), 2222-2232, (2016)