データサイエンスを用いた都市ガス販売量の予測精度比較

- アルゴリズム変更や部門別の予測による精度向上 -

計量分析ユニット エネルギー・経済分析グループ | 寄田 保夫

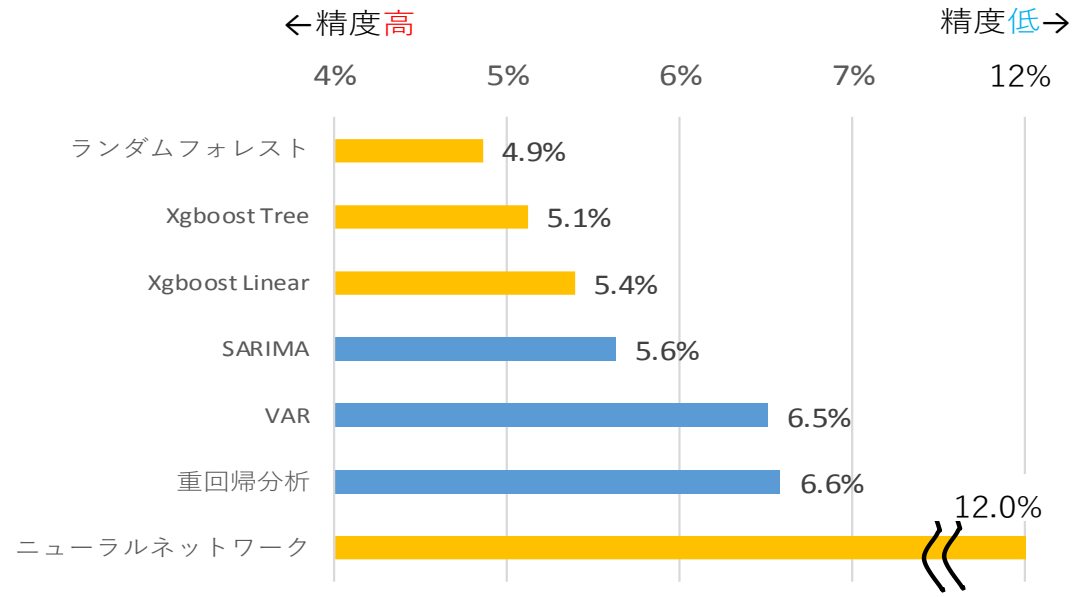
要旨

本稿では都市ガス販売量予測における高精度なアルゴリズムを検討するとともに、予測モデルの精度向上の検討を行う。1990年4月-2017年9月までの330か月間を学習期間、2017年10月-2018年9月までの12か月間を予測期間とし、全国での都市ガス販売量を予測する。

まず、第1ステップとして全体の販売量を予測し有効なアルゴリズムを検討した（図1）。ここでは、予測を導く構造の説明や解析に重きを置く統計学的手法と、データを活用した予測精度に重きを置く機械学習的手法の代表的なものを対象とした。具体的には、統計学的手法からは「重回帰分析」「ベクトル自己回帰モデル（VAR）」「季節調整済時系列モデル（SARIMA）」を、機械学習的手法からは「ランダムフォレスト」「Xgboost Tree」「Xgboost Linear」「ニューラルネットワーク」といった7つの代表的な手法を用いることとした（【参考】参照）。次に第2ステップとして部門別に予測を行うことで予測精度の向上を図った（図2）。なお、比較に用いる精度指標としては、平均絶対誤差率（各月誤差率における絶対値の平均をとったもの。以下、MAPE）を用いた。

全体の販売量の予測では、ニューラルネットワークを除き、全般に統計学的手法よりも、機械学習的手法のほうが予測精度が高くなった（図1）。特に、ランダムフォレストが最良となった。一方で、機械学習的手法のなかでもニューラルネットワークはデータ数が少なく、適切に予測できなかった。昨今、注目を集めるニューラルネットワークであるが、大量のデータを用いた分析に強く、今回のような330か月間の学習データ数では力を発揮することができなかった。

図1 | 都市ガス月別販売量のアルゴリズム別予測精度（MAPE）

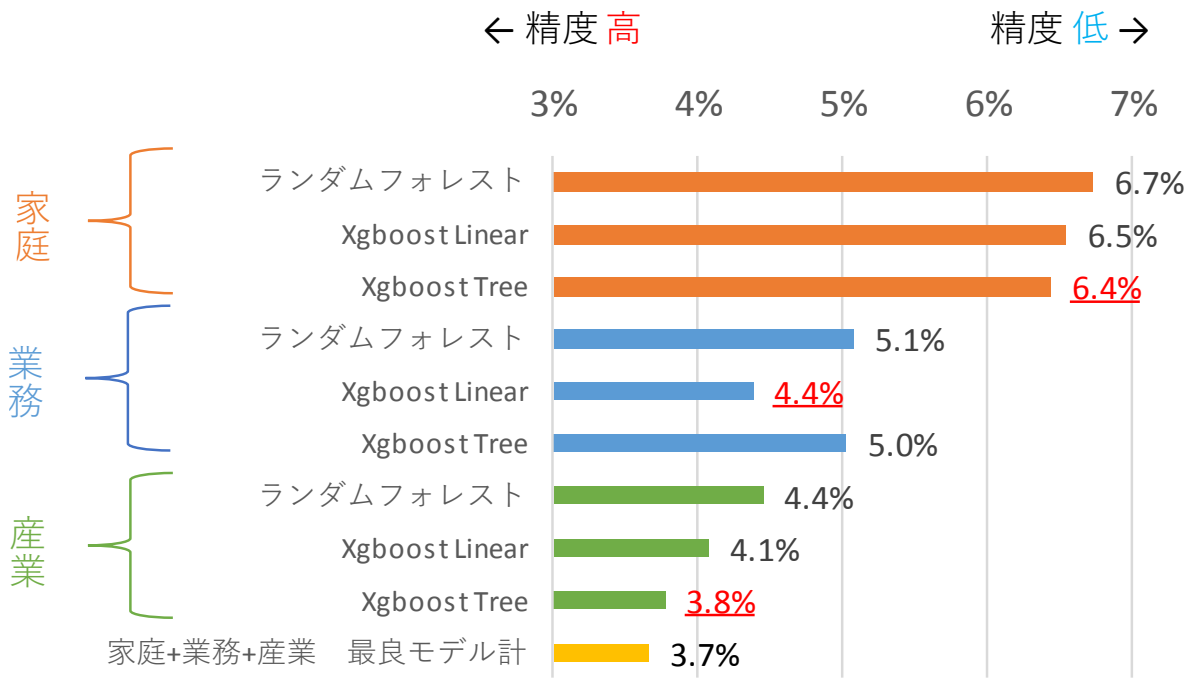


出所:筆者算出。
(注) 黄色グラフは機械学習的手法、青色グラフは統計学的手法。

部門別の販売量の予測では、全体の販売量の予測で精度の高かった3つのアルゴリズム（ランダムフォレスト、Xgboost Tree、Xgboost Linear）に絞り検討した結果、Xgboost LinearやXgboost Treeのほうがランダムフォレストよりも予測精度が高い場合が多数を占めた（図2）。最も予測精度の高いアルゴリズムは、家庭部門ではXgboost Tree、業務部門ではXgboost Linear、産業部門（発電用除く）ではXgboost Treeとなっている。このように、アルゴリズムは適用課題次第で結果の精度が変わるため、工学的に試す必要がある。

最終的な結果となる家庭+業務+産業のそれぞれの最良モデルを合計した場合のMAPEは3.7%まで改善した。部門別よりも合計したモデルのほうが精度が良いのは、誤差が正と負で打ち消しあうためである。なお、SARIMAモデルも部門別に積み上げる形で予測をしたが、こちらのMAPEは5.7%となり、全体販売量予測の5.6%から悪化した。部門別の積み上げでは、説明変数がより直接的に予測に寄与できることから、説明変数を用いたモデルのほうが適しているということだろう。

図2｜都市ガス月別販売量の部門毎のアルゴリズム別予測精度（MAPE）



出所：筆者算出。
(注) 産業は一般工業用のみ（発電用を除く）。「最良モデル計」は、家庭・業務・産業で最も精度の高かったモデル（赤字で表示）を合計したもの。

今回の予測精度比較では機械学習的手法が上回ったが、統計学的手法は構造の説明や解析に優れている。それぞれのアルゴリズムの長所と短所を把握し、状況に応じて統計学的手法と機械学習的手法を使い分けることが重要である。

<キーワード> 都市ガス、予測、データサイエンス、統計学、機械学習、AI

お問い合わせ: report@tky.ieej.or.jp