大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

時系列データを用いた機械学習の実践（Kaggleコンペへの参戦）

NTTデータアイ　荒山 泰佑
日本総合研究所　岡野 文香
鹿島建設　金子 晴紀
日本ユニシス　横井 康司

演習テーマ

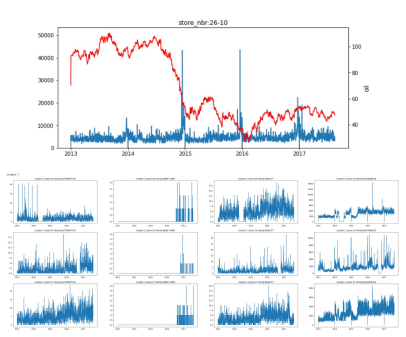
- KaggleのStoreSales※を題材として、可視化分析やアルゴリズム選定、モデル構築など**機械学習における一連の分析工程を体得**する。
  ※エクアドルの総合スーパーの店舗×商品の時系列売上を予測するコンペ
- 機械学習特有の難しさや注意点について考察を得る

適用した手法/ツール

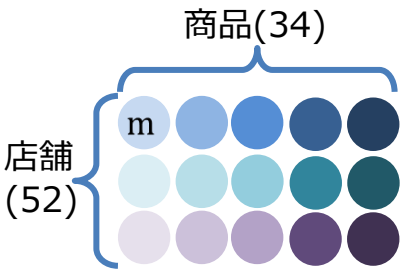
- Pythonにてデータを可視化し、売上に関係するデータを検討し仮説検証
- 学習・予測には時系列データ予測アルゴリズムであるProphetを使用
- コーディング部分はモブプログラミングによりチーム全員で課題に取り組み

機械学習分析の実践

可視化分析・仮説抽出 → アルゴリズム選定・初期モデル構築 → モデル改善



Prophet $y(t) = g(t) + s(t) + h(t) + Bx(t)$



- 仮説1：店舗別・商品別に予測
- 仮説2：トレンド・周期性
- 仮説3：祝日効果＝国民休日 → 改善モデル1
- 仮説4：イベント効果＝給与支給日 → 改善モデル2
- 仮説5：外部説明変数＝販売促進データ → 改善モデル3

✓ 売上を様々な角度から分析
✓ 学習モデルに組み込む**仮説を設定**

✓ Facebook社が開発した時系列解析用のライブラリ
✓ **初期モデル**としてトレンド、周期性を考慮したモデルを構築

✓ 初期モデルに対して休日効果、給与支給日、販促データをそれぞれ考慮し**改善モデル**を構築

評価

モデル	定量評価（RMSLE） 検証データ	Kaggle
初期モデル	0.5669	0.5428
改善モデル1(休日)	0.5805	△ 0.5408
改善モデル2(給与支給)	0.5694	0.5450
改善モデル3(販売促進)	○ **0.3939**	× 1.4184

✓ 改善モデル1,2では大きな改善はせず、また改善しない原因も推定できなかった
✓ 改善モデル3ではテストデータとして与えられた販売促進情報を修正することで改善が図れる見込み

課題

■予測改善
✓ 改善へ向けた原因分析、仮説設定等を含め**要因分析**を進める必要がある
✓ **ドメイン知識**、**他のアルゴリズム**を用いる等を分析・改善に組込み、多方面から検討する必要がある

■演習課題（分析工程の学習）
✓ 分析工程の全体像や改善へ向けた課題・注意点などに関して学習することができた
✓ データ収集、時系列性の確認等省略した工程を習得する必要がある

データ収集 → 可視化 → 時系列性確認 → 仮説抽出 → アルゴリズム → モデル構築 → モデル改善