DataScientist Society

リークとビジネス応用での注意点

～コンペの振り返り，リーク(Leakage)，データ分析で大切なこと～

今回の目標

- リーク(Leakage)の理解
- 検証の枠組みが大切なことの理解
- インプットとアウトプットが大切なことの理解

コンペの振り返り

➢ 問題設定

コンペサイト(https://signate.jp/competitions/1)には「コールセンター運用に活用」とありました

➢ AUCが高すぎる

皆さんは今回の銀行の顧客ターゲティング問題でAUC:0.9を超えるスコアを出すことができました. AUC:0.9とはどういうことでしょうか?

学習データはy=0(非申込者)が23,954件で, y=1(申込者)が3,174件だったので, 右図の例では, モデルの出力確率の上位から電話すると7965件で申込者を全て捕捉できます.

$$23,954 \times 0.2 + 3,174 = 7964.8$$

今までは27,128件だったので, 従来の約3割で同じ効果を期待できます.
→ 残念ながらリーク(Leakage)です

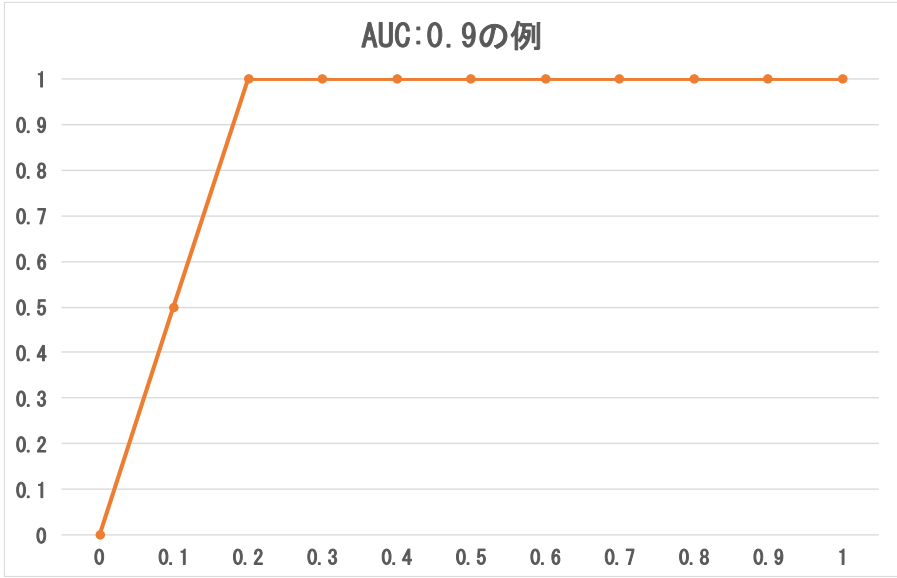

図: ROC曲線

リーク(Leakage)とは

- **リークとは**

本来想定されていない方法で，目的変数あるいはそれに近いものが漏れていて，機械学習できてしまうこと

参考 → https://www.kaggle.com/docs/competitions#leakage

- **リークの例**
  - ✓ 答えと同等の変数が元々入っている
  - ✓ 「簡単な計算」で答えと同等の変数が元々入っている
  - ✓ 未来の情報が入っている
  - ✓ 処理の差異が逆算できる
  - ✓ 無意味と思ったIDに意味がある
  - ✓ 検証データとしてホールドアウト分割したつもりが，正しく分割できていない(実質同じデータが構築と検証に混在しているなど)
  - ✓ 時系列データをクロスバリデーションでごちゃまぜにしてしまっている

  - ✓ KDDCUP2008：病院IDで患者の重篤(じゅうとく)度が分かる
  - ✓ KDDCUP2013：紐づけキーの同じデータの個数が強い説明変数
  - ✓ KDDCUP2015：同じ人の未来の行動ログが利用できてしまう

リーク(Leakage)とは

➢ コンペでのリーク

機械学習コンペでも稀に起こりますが，コンペでは不正ではなく問題作成者が悪いので，不正ではありません.

→ コンペなどでは，上位陣だけスコアが高い場合，強い説明変数かリークを見つけたことが推測されます

➢ ビジネス上でのリーク

もし，現実のビジネス問題設定においてリークが見つかった場合は，すぐに対処する必要があります.

→ 実際の運用では想定と大きく乖離した結果となります

今回のリーク例

➢ 説明変数の妥当性

今回，決定木モデルを可視化した際に強力に効いた変数であったdurationにリークの疑いがあります．コンペサイトのデータ説明にはdurationは「最終接触時間（秒）」と記載されていました．

✓ 仮説

最終的に接触した時間が短い → 口座開設の勧誘を断っている → 申込みしない
最終的に接触した時間が長い → 口座開設の詳細を聞いている → 申込する

仮に，上記の仮説が正しいとすると，それはコールセンターの人間が実際に電話するまで分からないこと．

つまり，「申込するから最終接触時間が長い」ので，因果が逆になっています．したがって，これは「電話する対象を探す分析」という意味では，durationは未来データであり，リークです．

リークへの対処法

➢ 検証の枠組みを「人間」がしっかり考える

検証の枠組みは，統計や機械学習は役に立ちません．元となるデータの生成過程やデータ周りの知識が重要になります．

→ 時系列で学習データと検証データをホールドアウトするのが大切です

➢ 確認項目例

✓ 説明変数は妥当なものか
→ 同じ説明変数が運用時でも取得できるか
→ データ処理にミスはないか

✓ 簡単なモデルと比較して，異常に精度が高くないか

✓ 過去データと未来データはどれくらい同質か
→ 未来と過去でデータは同様と考えて、本当にいいのか
→ 地域や年代が偏ってないか

データ分析で大切なこと

➢ 業務フローと大切なポイント

業務フローの例（あくまで一例であることに，ご留意下さい）

1. 何のために，何を予測するのかを決める
2. 過去のデータ収集
3. 利用する変数の選定
4. 使用するモデル（手法）の選定
5. 特徴量の作成，変数の加工
6. 学習
7. 精度評価
8. 精度評価を踏まえて運用方法の決定

データ分析で大切なこと

➢ 業務フローと大切なポイント

業務フローの例（あくまで一例であることに，ご留意下さい）

1. 何のために，何を予測するのかを決める
2. 過去のデータ収集
3. 利用する変数の選定
4. 使用するモデル（手法）の選定
5. 特徴量の作成，変数の加工
6. 学習
7. 精度評価
8. 精度評価を踏まえて運用方法の決定

ここが重要なポイント

これは誰でもできる

データ分析で大切なこと

- 大切なポイントの詳細

3. 利用する変数の選定

- ✓ 収集したデータをじっくりと分析する
- ✓ 異常データや変数の傾向などを，データ収集した人と認識合わせすることでデータの質を確認する
  - → データの途中から単位が変わっていることもある
- ✓ 問題設定に対して適切な変数選定（汎用性のない変数を除くなど）

4. 使用するモデル（手法）の選定

- ✓ 目的，データ，運用を踏まえて使用するモデルを選定
- ✓ 使用するモデルの中身をきちんと把握していることが大事

データ分析で大切なこと

- 大切なポイントの詳細

5. 特徴量の作成，変数の加工

- ✓既存の変数を複数組み合わせることで新たな特徴を作成する
- ✓線形回帰であれば，各変数について線形化，多重共線性の確認，外れ値の丸めを行う

→他のモデルでも前処理はとても大切な作業となります

7. 精度評価

- ✓目的に合わせた評価指標の選定
- ✓リークに気を付けた検証の枠組み設計

8. 精度評価を踏まえて運用方法の決定

- ✓精度評価の結果から運用方法を決める
- ✓例：モデルの予測違反確率の上位10%で，違反者の90%を捉えられるならば，残りの90%に含まれる違反者(違反者の10%)に目をつぶれば，業務の90%を削減できる

→使用する精度評価指標の中身の理解が大切