

# Feature Neutralization のちょっとディープな話

2021-12-18 / Numerai Meetup / [@yoshiso44](#)

# Self Introduction

Name:

よしそ(@yoshiso44)

Job:

botter / marker taker

Background:

ソフトウェアエンジニア/データサイエンティスト

Interests:

Financial Time Series Deep Learning /

Deep Learning Generalization on Noisy Data

# Agenda

- Feature Neutralizationとは?
- 古典的な投資理論との関係
- Feature Neutralizationの限界
- 限界を越えて

# Feature Neutralization(FN)とは

- 予測モデルのリスクを抑えるために、予測値と特徴量との間の順位相関(=Feature Exposure)を下げるための手法。
- 予測値をtargetに線形回帰し、その残差を新しい予測値とする。
- Numeraiでは、この処置を行うと安定したパフォーマンスが得られる傾向にあることが共有されている([@jrb](#))
- 生の特徴量を予測値として評価した場合スコアの分散が高く、相関が高いほど予測モデルのスコアも高い分散に。

# 古典的理論との関係

- FNで係数を推定する線形回帰の式は、マルチファクターリスクモデルを表現する式とみなせる。
- FN後の予測値は、ファクターと線形独立。古典的なリスクモデルでの評価上、ファクターリスクへのエクスポージャーが大きく減少する。

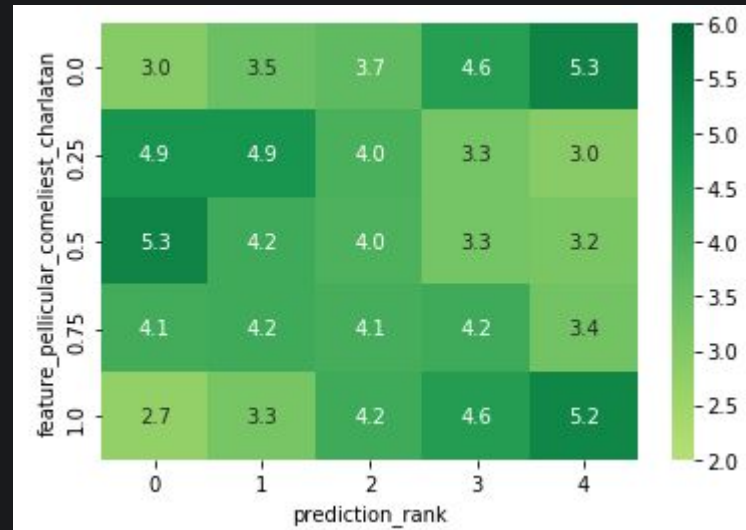
マルチファクターリスクモデル

$$r_t = \sum_{k=1}^k X_{tk} b_k + u_t$$

- $x_{tk}$  = ファクターkのエクスポージャー
- $b_k$  = ファクターkのファクターリターン
- $u_t$  = 銘柄nのスペシフィックリターン

# Feature Neutralizationの限界

- 線形なリスク中立化手法であり、非線形なリスクに対して脆弱
- Feature Exposureも同様に非線形なリスクを計測困難。
- 左記ヒートマップにて典型的な例を可視化



5分位化した予測スコアと特徴量のグリッドに何%のサンプルが存在しているか(疑似scatter plot)

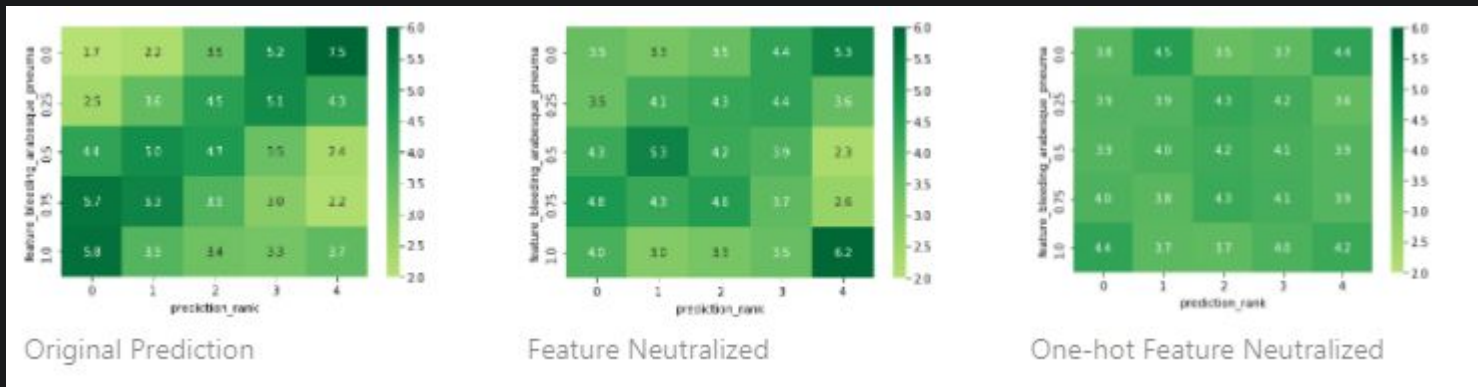
# 限界を越えるために - 新しいリスク指標

- 非線形なリスクも捉えるGross Feature Exposure
- 予測スコアの分位ごとにFeature Exposureの偏りを評価する

```
def gross_feature_exposure(  
    feature: pd.Series,  
    prediction: pd.Series,  
) -> float:  
    group = pd.qcut(prediction, 5, labels=False)  
    return (  
        feature.sub(feature.mean()).div(feature.std())  
        .groupby(group).mean().abs().sum()  
    )
```

# 限界を超えるために - Onehot FN

- Onehot Feature Neutralizationの導入
- Onehot Encodingした特徴量に対してFNを行うことで、非線形な依存関係についてもリスク中立化を行う





## 限界を越えるために - アイデア集

- Risk Exposure Budgeting([@mdo](#))
- Riskiest Feature Neutralization([@nyuton](#))
- Solve as Constrained Optimization Problem

**Fin.**