5. Model risk and governance

- machine learningモデルはいくつもあって，それぞれ長所短所がある．

Model Chioce

いわゆる機械学習でよく使われるモデルは大きく分けて4つ

1. Trees
2. clustering (unsupervised approaches)
3. neural network
4. regression

Trees

最もよく用いられるモデルの一つ．

decision tree(決定木)：アルゴリズム簡単だか，オーバーフィットしてしまう．ただ，サンプルとしては多く用いられている．（データを過剰にフィットしてしまうので，判別のthresholdが高いモデルになる？？）

random forest:最も多く用いられているアルゴリズム

Gradient boosted trees(GBTs):木が前の木(gradient descentと呼ばれる)から学ぶ．平均絶対誤差(mean absolute error)を最小にするアルゴリズム.(誤差の絶対値の和をサンプル数で割った値:回帰モデルの誤差を評価するのに用いられる)

treeのアルゴリズムは最もよく用いられている．

ある銀行:tree basedのアルゴリズムは，バケットわけとかランクわけをしないで機械学習できるので，情報ロスがない．

ある２つの銀行：あまり整っていないデータを元にして協力なpredictiveなモデルを作っているらしい．(どんな題材だろう…)

Clustering

ポートフォリオ細分化(segmentation)に用いられている
データを探索したり，変数の関係とか，あるいはデータの変数の自由度を減らしたりとか．

Neural network

多くの銀行は難しいモデルは不明瞭であって，あまり望ましいものではない．扱うデータ量もおおいし，最先端のITインフラが専門的すぎるから…

ただ，将来予測の威力は絶大で，複雑なデータパターンもモデル化する．これらのアルゴリズムは主により人工的な知能が絡む画像解析やアルファ碁と馴染み深い．

いわゆるシナプスっぽいものを介して繋がったいくつものノードから形成されるのがneural network.

> 例えば，赤いものを見たら，りんごと認識するような信号経路(シナプス)をつける．その後，同じことを繰り返し行うと，生物みたいに覚えてきて，りんごをみたら赤を思い浮かべるようになる(ネットワークが形成されたとみなす)
> 関連のつよりニューロンと関連の弱いニューロンが形成されていく．
> 一通り学習させたあとは，ニューロンの状態をどこかプラス１にして，他のニューロンのプラスマイナスが変わらない安定状態になるまで繰り返す．この状態が記憶と解釈される．

12%の銀行がディープラーニングを使っていて，一銀行のみcomvolutional recurrent network()を応用している．複雑なヒストリカルのプロセスを理解するために．あるいは他の銀行は，ディープラーニングをテキストマイニングや欠損データ
の代入に用いている．ある銀行はリソース(PCの？)がかかるのでもうすでにディープラーニングを用いるのをやめている

結局regressionのアルゴリズムを用いている銀行がスタンダードで，多くの銀行で用いられている．LASSO(least absolute shrinkage and selection operator)