

データサイエンティスト育成 ベーシックステップ 機械学習

宿題
House Prices Advanced Regression Techniques

2017/09/02

Masayuki Mizuno

Agenda

- 1. 目的
- 2. アプローチ
- 3. <Step 1> 予測モデルの作成
 - a.学習データの準備
 - b.説明変数の準備
 - c.ロジスティック回帰分析による予測モデルの作成
- 4. <Step 2> ターゲットとすべきユーザー像を設定
 - a.予測モデルの解釈
 - b.期待される収益
 - c.今後のアクション
- Appendix 1.予測モデル作成プロセス

1. どのプロジェクトに取り組んだのか？












■ kaggle

□ House Prices: Advanced Regression Techniques

- <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard>

2. 最終順位

■ 順位のスクリンショット

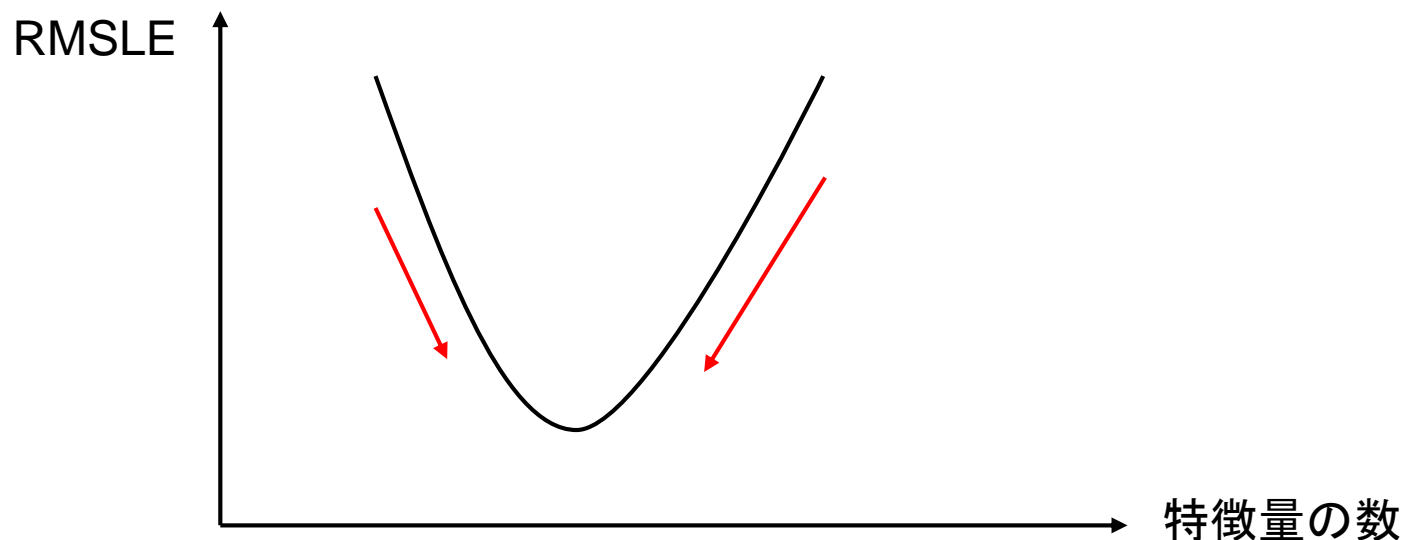
1272	▼ 22	Tang Yiming		0.15845	2	2mo
1273	▼ 22	kitivan		0.15848	1	24d
1274	▼ 22	Евгений Борисов		0.15849	2	2mo
1275	▼ 22	mocogin		0.15916	4	24d
1276	▼ 22	chen 4		0.15922	2	2mo
1277	▼ 22	MasayukiMizuno		0.15923	20	7m
Your Best Entry ↑ Your submission scored 0.16680, which is not an improvement of your best score. Keep trying!						
1278	▼ 22	keisukeuema		0.15936	6	1mo
1279	▼ 22	Giulia Rinaldi		0.15945	4	2d
1280	▼ 22	sahia		0.15953	1	1mo
1281	▼ 22	NathanaëlCarraz		0.15953	4	1mo
1282	▼ 22	MonAnjoCalma		0.15959	1	1mo

3. どのような仮説でアプローチしたのか？

- 前提：ランダムフォレストを利用

- ランダムフォレストの学習のため

- 仮説：特徴量の数を増減すると RMSLE の最小値が見つかる。以下のイメージ。

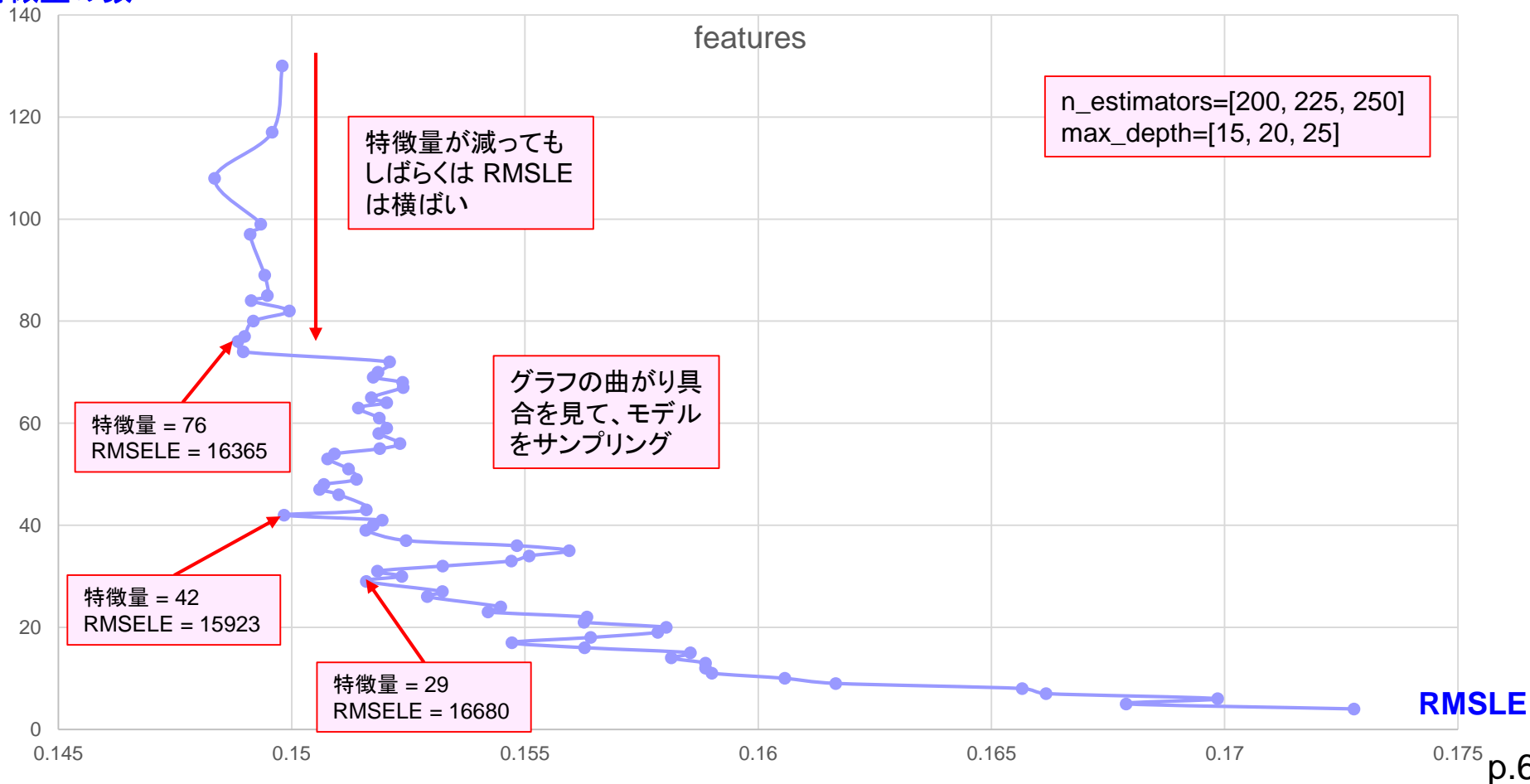


4. うまく行ったこと

■ 特徴量を減らして RMSLE の変化を確認

□ モデルベース特徴量選択 (SelectFromModel)

特徴量の数



5. うまく行かなかったこと

■ 変数増加法

- ブートキャンプ 5 回目のテキストの方法
- 11 個選択され、スコア結果は 0.16370

6. 今後の改善の可能性について

■ ビジネスドメインの考察

- 例) 住宅ローン金利の追加など

■ 各種テクニックの利用

- 特徴量選択

- 単変量統計 (SelectKBest / SelectPercentile)
- 反復選択 (RFE)

- 評価

- 交差検証
- `n_estimators` / `max_depth` 以外のグリッドサーチ

- (以下は、今回活用済み)

- カテゴリ変数の整数化
- 割り算で特徴量の作成

■ 複数のモデルでの試行

- 重回帰分析など