

家賃予測

問題設定と問題分析

2018年11月1日以降の任意の日付(明記すること)時点の、以下の物件の1ヶ月あたり賃料を予想して下さい。

- ・ 8階建てのマンションの2階 (カテゴリー)
 - ・ 1K (カテゴリー)
 - ・ 専有面積 $15m^2$ (連續変数)
 - ・ 築19年 (カテゴリー)
- ・ 住所:東京都港区芝5 (カテゴリー)
- ・ 三田駅 徒歩1分、田町駅 徒歩3分、赤羽橋駅 徒歩14分 (カテゴリー)
 - ・ 礼金、敷金未定 (家賃に影響あり)
 - ・ 風呂無し、トイレ共用 (カテゴリー) (下落効果知りたい)

結論：分類問題であるXからYを予測する場合

カテゴリー変数のX多いほど、決定木型学習の手法が有効

特徴量の特定

- ・2019年2月26日、suumoにキーワード検索：
風呂なし 共同 トイレを東京都に限定して検索
 - ・該当する物件22件
- ・新宿区1件、文京区3件、台東区1件、大田区2件、中野区2件、杉並区1件、豊島区3件、足立区1件、江戸川区8件

2019年2月26日のデータ

スクレイピングデータの選択

並び替え

家賃高い順

建物種別

マンション

アパート

一戸建て・その他

間取り

ワンルーム

1K/1DK

1LDK/2K/2DK

2LDK/3K/3DK

3LDK/4K~

相場情報を更新する

予測したい特別区とそれに関連する特別区

風呂なし共同トイレに該当する特別区

- 港区の平均家賃の変わらない、中央区、渋谷区、隣の品川区を選択
- 風呂なし共同トイレに該当する物件を平均家賃の上位から選択、杉並区、足立区、江戸川区は切り捨て

市区郡	家賃相場	
港区	9.6 万円	
中央区	9.5 万円	
渋谷区	9.5 万円	
新宿区	8.9 万円	
品川区	8.9 万円	
目黒区	8.9 万円	
文京区	8.8 万円	
千代田区	8.7 万円	
台東区	8.7 万円	
江東区	8.2 万円	
中野区	8.2 万円	
墨田区	8.1 万円	
世田谷区	8.0 万円	
豊島区	8.0 万円	
大田区	7.6 万円	
杉並区	7.6 万円	
武蔵野市	7.6 万円	
荒川区	7.5 万円	
北区	7.2 万円	

川区は切り捨て

前処理

前処理 1

- ・SUUMOから港区、中央区、渋谷区、品川区、豊島区、中野区、大田区、文京区、台東区、新宿区の家賃、2019年2月5日の分をスクレイピング
 - ・標本は10コしかないため、各標本を1000コ複製して増やす
 - ・新築は築年数0年とする
 - ・文字列から数値に変換
- ・単位を揃える賃料、敷金、礼金、保証金、敷引、償却を10000にかける
 - ・予測データは、賃料+管理費にする
 - ・住所を東京都何々区、何々市町村に分割
 - ・階数を数値化して、地下の場合はマイナスにする
 - ・建物高さを数値化し、地下は算入しない。
- ・間取りを部屋数、Kの有無、Lの有無、Sの有無、DKの有無に分割する
 - ・礼金、敷金の変数を（家賃+管理費）で割る。
 - ・礼金：賃料の倍数、敷金：賃料の倍数というカラムを作る
 - ・風呂無し、トイレ共用（カテゴリー）を作る

前処理

前処理 2

- ・ターゲット：予測したいデータフレームを作
る。
- ・礼金：賃料の倍数[なし、一ヶ月、二ヶ月、
三ヶ月]、敷金：賃料の倍数[なし、一ヶ月、
二ヶ月、三ヶ月]
- ・風呂無し、トイレ共用（カテゴリー）を**1**に
する
- ・家賃の下落効果を知るために風呂無し、ト
イレ共用（カテゴリー）を**0**にする同じ予測
データフレームを作る

XGBoost(eXtreme Gradient Boosting)

- ・代表的な勾配ブースティング決定木型手法
- ・XGBoostとは、GBDTとRandom Forestsを組み合わせたアンサンブル学習である
- ・精度が高くて、よくKaggleなどのコンペで使われている

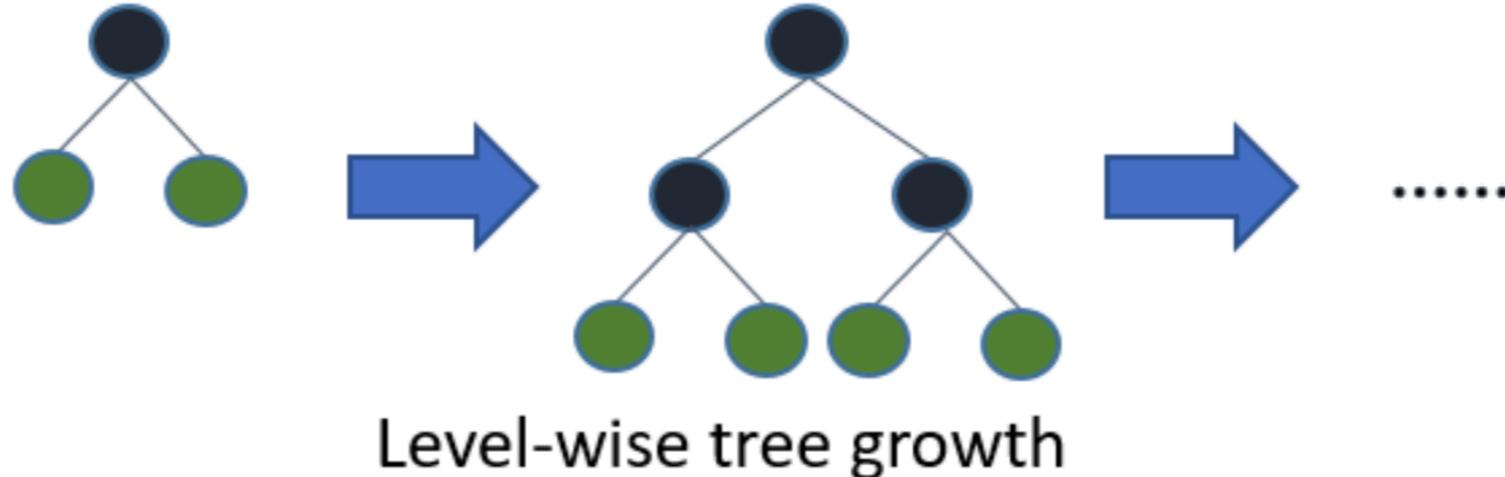
LightGBM¹

- ・2017年に出た新しいモデル
- ・XGBoostより訓練速度が速く短い時間で精度を高めることができる

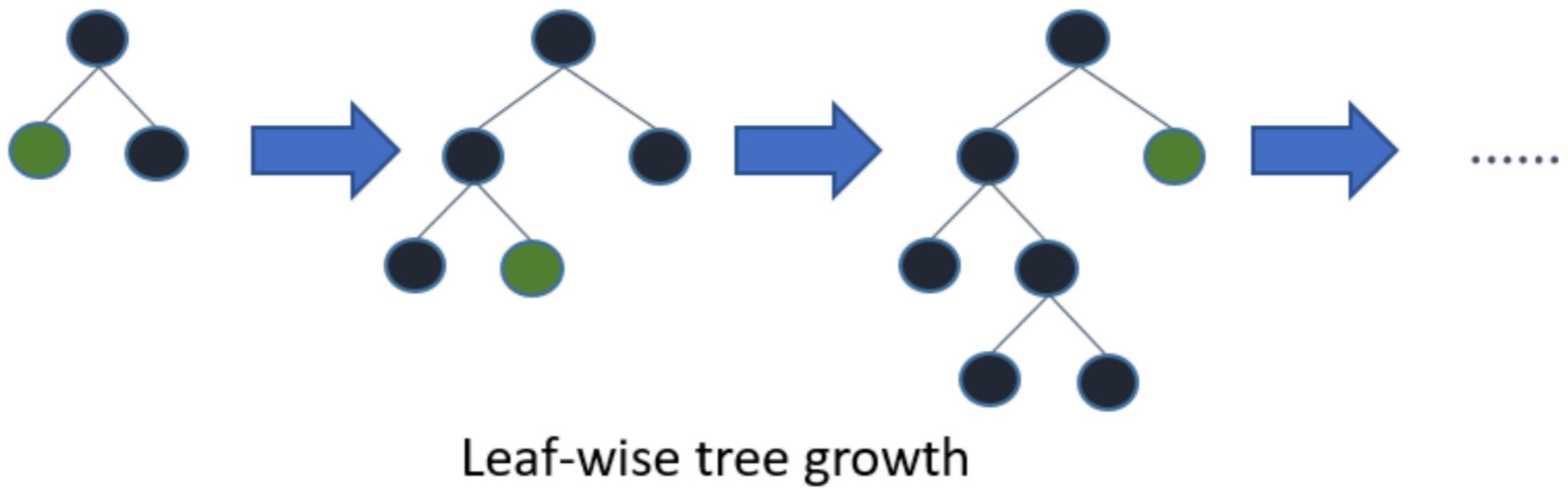
¹ Guolin Ke, Qi Meng , Thomas Finley "LightGBM: A Highly Efficient Gradient Boosting Decision Tree" at NIPS Proceedings(2017)

XGBoostとLightGBMの違い

機械学習、決定木的なアプローチ



Level-wise tree growth in XGBOOST.



Leaf wise tree growth in Light GBM.

評価

$$RMSLE = \sqrt{\frac{1}{N} \sum_{i=0}^n (\log(y_i + 1) - \log(y'_i + 1))^2}$$

- RMSLEは対数を取っているので一つの大きな間違いでの差が出にくくなる

交差検証法

テストデータ

トレーニングデータ

トレーニングデータ

テストデータ

トレーニングデータ

トレーニングデータ

テストデータ

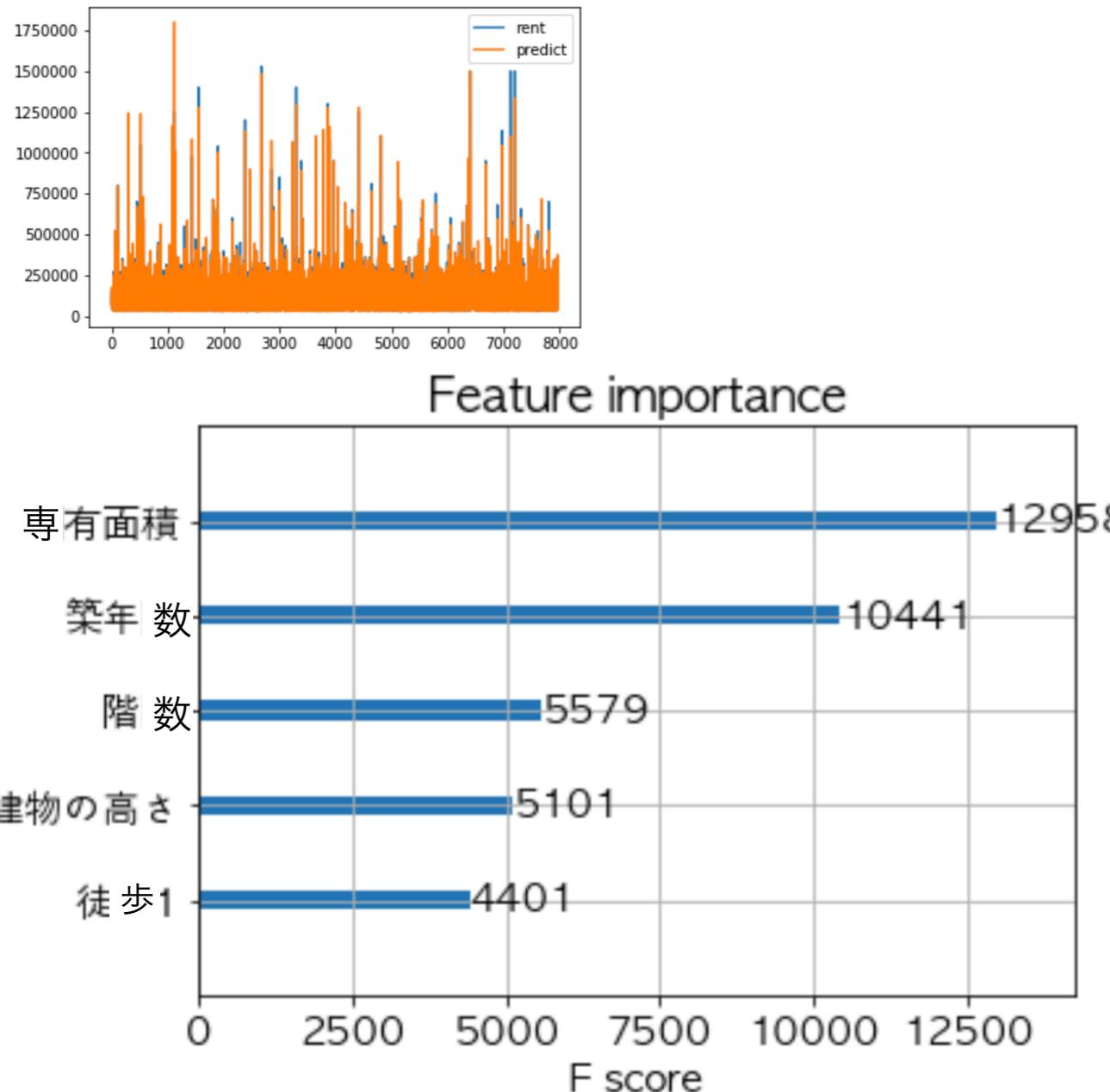
XGBoostモデル

パラメーターの設定

元データ	トレーニングデータ (80%)	+	検証データ (10%)	+	テストデータ (10%)
各木に抽出される標本の割合			0.8		
各木に抽出される列の割合			0.8		
学習率			0.1		
木の深さの最大値			15		
学習回数			100		

XGBoostモデル

XGBoostモデル



風呂なし トイレ共同	礼金	なし	1ヶ月	2ヶ月	3ヶ月
敷金					
なし	82401.734375	82770.210938	90471.726562	87936.039062	
1ヶ月	76094.421875	76569.257812	85109.414062	82573.718750	
2ヶ月	77051.281250	77078.828125	85771.007812	83235.312500	
3ヶ月	83947.171875	83862.031250	88712.468750	86176.781250	
なしの場合					
礼金					
なし	86798.671875	87167.148438	95335.039062	92799.351562	
1ヶ月	80491.382812	80966.218750	89972.726562	87437.031250	
2ヶ月	79050.398438	79077.945312	88236.492188	85700.796875	
3ヶ月	86110.984375	86025.843750	91342.640625	88806.953125	

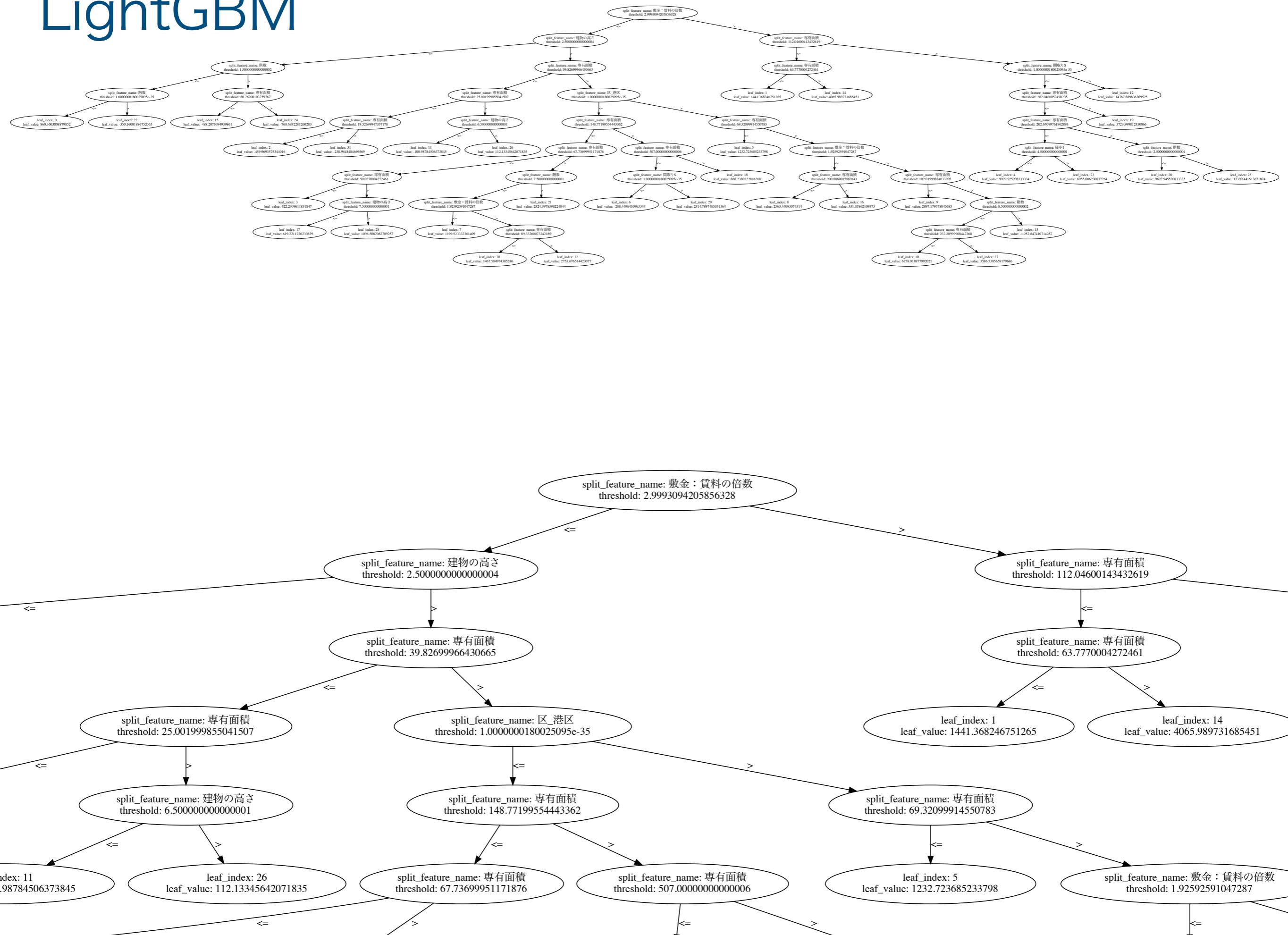
- 精度RMSLE : 0.08534621217742563
- 風呂なしトイレ共同による家賃の下落効果が確認
- しかし、敷金3ヶ月の方が過小評価（高級マンションの場合は10万を超えるはず）
- 木の本数が多いため、敷金3ヶ月のサンプル数が少ないケースに対する検証力が弱い

LightGBM

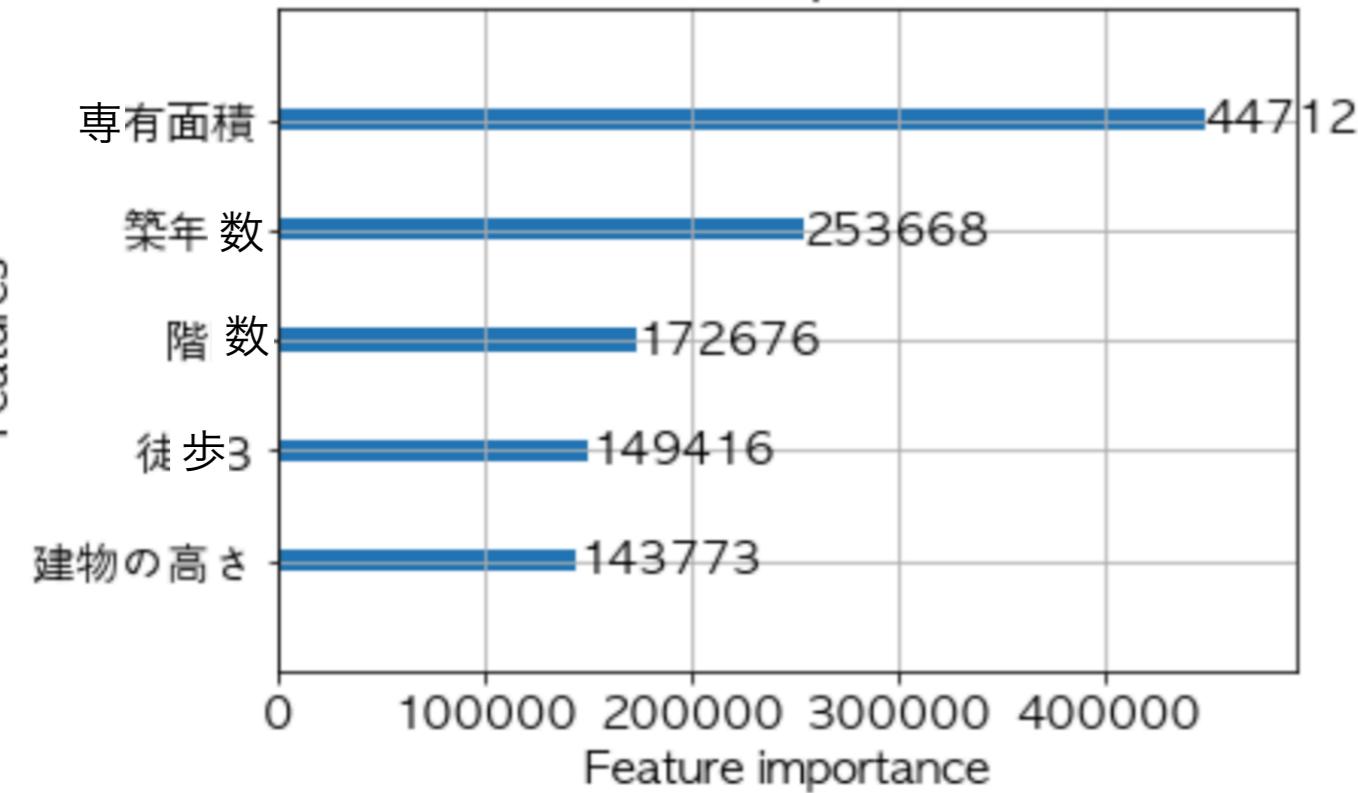
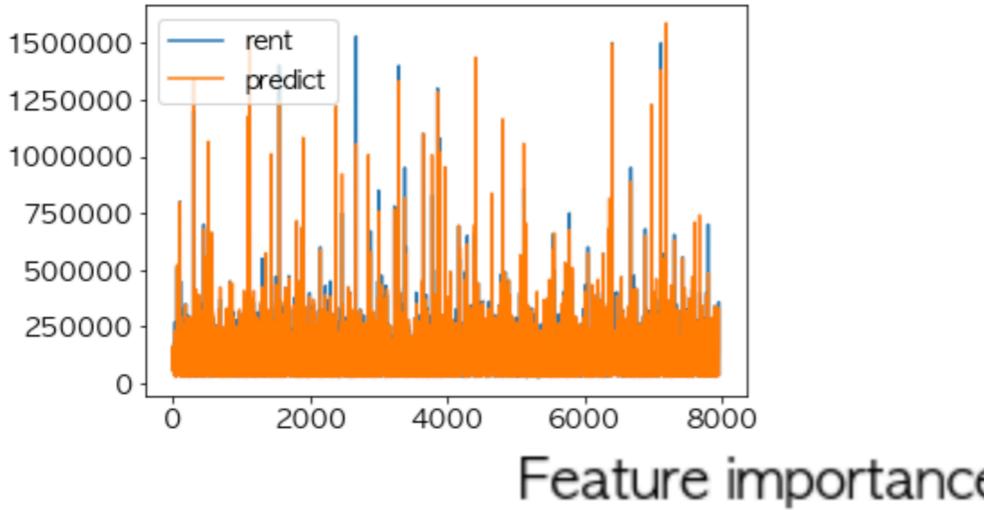
パラメーターの設定

元データ	トレーニングデータ (80%)	+	検証データ (10%)	+	テストデータ (10%)
各木の特徴量利用の割合			0.8		
バギング			0.8		
学習率			0.01		
葉の数			33		
学習回数			100000		

LightGBM



LightGBM



風呂なしトイレ共同

	礼金	なし	1ヶ月	2ヶ月	3ヶ月
敷金					
なし	82405.295089	89135.452032	90455.482608	95636.432316	
1ヶ月	76902.972926	75855.582087	76647.184439	81864.091526	
2ヶ月	74481.616603	74257.094451	74643.605353	79855.120760	
3ヶ月	123808.166902	100369.438926	96800.587914	101216.565663	

なしの場合

	礼金	なし	1ヶ月	2ヶ月	3ヶ月
敷金					
なし	86525.248354	93255.405297	94575.435874	99756.385582	
1ヶ月	81022.926192	79975.535353	80767.137705	85984.044792	
2ヶ月	78669.475782	78444.953630	78831.464533	84042.979940	
3ヶ月	127676.130782	104237.402806	100668.551794	105084.529543	

予測結果

- 風呂なしトイレ共同による家賃の下落効果が確認
- しかし、精度はRMSLE : 0.09011026284563534、XGBoostに負けた。
- 特徴量の重要度として最初に来るのは駅一番近いの、徒歩 1 のはずだが、徒歩 3 になっている。
- 木の本数が少ないため、今回の平均家賃のばらつきの多いデータ、(区、市町村、分類データ、間取りなどによって家賃の違い)に対応しきれない可能性。

実際の比較 風呂なし共同トイレに該当しない場合

XGBoost

礼金	なし	1ヶ月	2ヶ月	3ヶ月
敷金				
なし	86798.671875	87167.148438	95335.039062	92799.351562
1ヶ月	80491.382812	80966.218750	89972.726562	87437.031250
2ヶ月	79050.398438	79077.945312	88236.492188	85700.796875
3ヶ月	86110.984375	86025.843750	91342.640625	88806.953125

予測した家賃の条件

8階建てのマンションの2階

1K

専有面積15m²

築19年

住所: 東京都港区芝5

三田駅 徒歩1分、田町駅 徒歩3分、赤羽橋駅 徒歩14分

結論

- 全体精度の方はXGBoostの方が良いが、稀な物件や、サンプル数が少ないデータを予測する場合は、LightGBMの方が良いかもしれない

反省

- 今回予測に使うデータが東京全ての特別区ではない、学習しやすいために、1kの平均家賃が近い特別区と、風呂なし共同トイレに該当する特別区のデータを選んだ。人為的な操作による学習のバイアスが存在しないことは否定できない。

LightGBM

礼金	なし	1ヶ月	2ヶ月	3ヶ月
敷金				
なし	86525.248354	93255.405297	94575.435874	99756.385582
1ヶ月	81022.926192	79975.535353	80767.137705	85984.044792
2ヶ月	78669.475782	78444.953630	78831.464533	84042.979940
3ヶ月	127676.130782	104237.402806	100668.551794	105084.529543

それに近い実際の家賃



賃貸マンション
ラグーンシティ芝公園

東京都港区芝3
J R山手線/田町駅 歩7分
都営三田線/三田駅 歩4分
都営三田線/芝公園駅 歩5分

築16年
8階建

階	賃料/管理費	敷金/礼金	間取り/専有面積	お気に入り
4階	8.5万円 3000円	敷 - 礼 8.5万円	1K 19.39m ²	

[詳細を見る](#)