中古マンション価格予測

氏名：山内 雄登
提出日：2025年1月9日(木)

やまのうち　ゆうと

山内 雄登　31歳

2017年3月東京電機大学中退
2022年前後に遺跡発掘作業員として3現場分就労
2023年8月Neuro Dive渋谷 通所開始
機械学習コンペでは入賞経験有

診断名	うつ病,発達障害
得意なこと	論理的に考えること。粘り強く取り組むこと。
苦手なこと	音や視界の動きに敏感(ノイズキャンセリングイヤホンで対処)
配慮事項	プログラミング中はイヤホンを使えると助かります。 (通所開始～今日まで「うつ症状」は軽く、自己対処もできています)

目次

- 不動産業界の課題と現状
- 成果物の目的
- 作成環境
- 中古マンションの価格変動要因
- 使用データの具体的な内容と工夫

  データ前処理（第１段階）,追加データとデータの前処理（第２段階）,特徴量の追加
- EDA

  時系列ごとの中古マンション価格、学習データから分かる傾向
- AIで中古マンション価格予測をする

  使用したモデルの説明, 複数モデルを使用した予測フロー,
  モデルのアンサンブル方法と効果, 予測精度, 予測結果,
  目的変数と平均絶対誤差（MAE）の関係, 平均絶対誤差（MAE）のヒートマップ分析,
  特に効果があった特徴量，Geohash-レベル別範囲
- 考察
- 課題点・今後の施策

不動産業界の現状と課題

背景

近年の日本の不動産業界は、人口減少や少子高齢化に伴う住宅需要の低下や、都市と地方の地価格差の拡大といった課題に直面しています。特に地方では空き家が増加し、地価下落リスクが高まっている一方、都市部ではマンション価格が高騰し、中古マンション市場は活況を呈しています。

具体例

東京都の中古マンションの平均成約価格は2023年12月時点で5,892万円と、10年前と比べて倍近い水準になっています。

正確な予測の重要性

中古マンションの売買を主な業務とする不動産会社にとって、将来の価格予測ができないと経営リスクが高まり、適切な戦略を立てることが困難になります。

成果物の目的

不動産の売買や賃貸、投資や開発などの各種事業において、取引の効率化や収益の最大化を図る事です。

作成環境

使用環境	OS:Windows11,CPU:11th Gen Intel Core i7-11370H (4コア/8スレッド),GPU: NVIDIA GeForce RTX 3050 Ti (4GB GDDR6),メモリ: 32GB LPDDR4x,ディスクサイズ: 2TB SSD Google Colaboratory Pro+
使用言語	Python
使用した 機械学習モデル	LightGBM,Catboost,SARIMAX
使用したデータ	Nishika「中古マンション価格予測2024夏の部」 学習データ：2005年第3四半期～2023年第2四半期 予測期間：2023年第3四半期～2023第4四半期

中古マンションの価格変動要因

・中古マンションの価格上昇と在庫増加による売れ行き鈍化リスク

中古マンションの価格→高騰,在庫→増加

・人口減少と少子高齢化による住宅需要の低下

特に地方でこの影響が顕著

・地価の二極化

都市部→高騰,地方→下落

・環境への対応、省エネ型住宅やスマートハウスの普及

省エネ性能の高い住宅→光熱費の削減,スマートハウス→快適性や利便性が向上

・DXの進展

・政府の住宅政策や金利政策の影響

使用データの具体的な内容と工夫

学習データの概要

1 データ期間

2005年第3四半期から2023年第2四半期までの
中古マンション取引データを使用しています。

2 主要な特徴量

都道府県名、市区町村名、地区名、最寄駅：名称、最寄駅：距離（分）、建築年、
取引時点、面積（㎡）などが含まれています。(計27個)

3 テストデータ

2023年第3四半期から第4四半期の目的変数(取引価格（総額）_log)が欠けた
データが用意されています。

実際の価格[円]の$\log_{10}$を取った値

データの前処理（第１段階）

1

表記修正

数値で表せる列に文字列で記載されている値等を修正しました。

2

不要なデータの削除

値が1種類しかない列や、全て欠損している列を削除しました。

追加データとデータの前処理（第２段階）

取得した追加データ

駅データ

駅データ.jpファイルから全国の駅の住所と経度緯度を取得しました。

地理情報

国土地理院APIから住所の経度緯度を取得しました。

1. **表記ゆれの修正**
   学習データ、駅データ、住所と駅名の表記ゆれを修正
2. **データの統合**
   学習データに駅データ、地理情報を統合し１つのテーブルデータにまとめました。
3. **欠損値の補間**
   地区名や駅名の欠損値を、テーブル内の情報を元に補間しました。

特徴量の追加

追加した特徴量の中でも精度に大きく貢献したもの

購入までの築年数

建築年から取引時点までの経過年数を計算し、物件の経年状態を反映させました。

旧耐震フラグ

1981年以前に建築された物件を旧耐震基準として識別し、安全性の指標としました。

geohash(2~7)

位置情報を高精度で表現するgeohashを導入し、地理的特性をより詳細に反映させました。

geohash(2~7)の集約値特徴量

geohashに基づいた統計値を算出し、地域ごとの特性を数値化しました。

中古マンション価格を可視化

全国の中古マンション価格は…

・価格は2005年第3四半期から2012年第4四半期までは低下傾向にあります。

・2013年第1四半期以降は上昇傾向にあります。

学習データから判明したデータの傾向・パターン

発表の際に、Pythonの動的な画面をお見せしながらご説明します

AIで中古マンション価格予測をする

もし、価格予測ができれば…

・戦略的な在庫管理が可能になります

・手動でのデータ分析や価格設定作業が削減されます

・データに基づいた価格予測を提供することで、顧客に対する信頼性が高まります

過去データ	価格
説明変数	判明している

学習

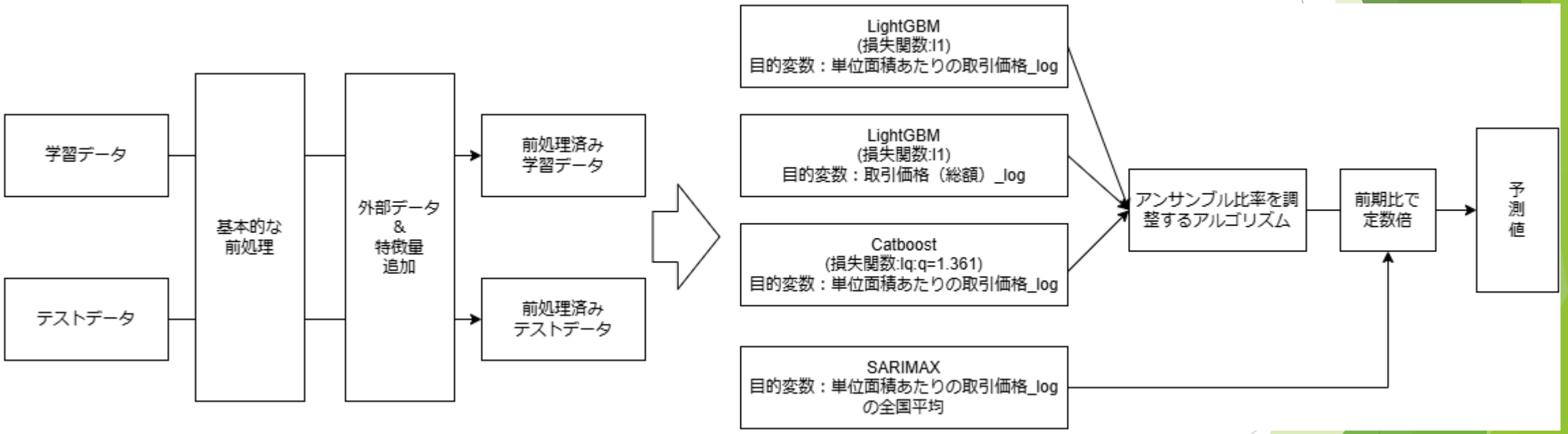
予測

未来データ	価格
説明変数	？

使用した予測モデルの説明

	強み	弱み
SARIMAX	**周期的なパターンを持つデータに適している。** 過去のデータから季節性やトレンドを捉えて比較的高い精度で予測を行うことができます。	**予測するステップ数が長いほど予測の不確実性が増加します。** ・長期的な予測では外部要因(市場の変動、経済状況等)の影響が大きくなるためです。
LightGBM	**複雑な関係性や相互作用を持つデータに適している** ・分類や回帰などの幅広い機械学習タスクにおいて、複雑なデータ構造を捉える能力を持っています。	**学習データに存在する範囲の値しか予測できません** ・学習データの範囲外の値（外挿）に対しては予測精度が低下する傾向があります。 LightGBMは決定木ベースのモデルであり、学習データに基づいて分岐を作成するのが原因です。
Catboost	**複雑な関係性や相互作用を持つデータに適している** ・分類や回帰などの幅広い機械学習タスクにおいて、複雑なデータ構造を捉える能力を持っています。 特にカテゴリカルデータの扱いに優れています。	**学習データに存在する範囲の値しか予測できません** ・学習データの範囲外の値（外挿）に対しては予測精度が低下する傾向があります。 Catboostは決定木ベースのモデルであり、学習データに基づいて分岐を作成するのが原因です。

複数モデルを使用した予測フロー



モデルのアンサンブルの効果

モデル	LightGBM	LightGBM	Catboost
損失関数	l1	l1	Lq:q=1.361
目的変数	単位面積あたりの取引価格_log	取引価格（総額）_log	単位面積あたりの取引価格_log

目的変数やモデルが異なることで、それぞれのモデルが異なる側面から価格を予測します。
統合することで、各モデルの弱点が補完され、予測精度が向上します。

SARIMAX

四半期ごとの過去データから、未来２期分の予測をします。
2023年第2四半期から第3四半期にかけての上昇率を算出し、将来の価格動向を示します。

予測精度

予測モデルの精度や信頼性を評価する指標としてはMAE(平均絶対誤差)を使用します。この値が小さいほど予測の誤差が少ないといえます。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

実測値	予測値	誤差	誤差の絶対値
100	90	10	10
120	120	0	0
80	50	30	30
110	140	-30	30
90	100	-10	10

誤差の絶対値の総和＝80
データの数＝5

平均

MAE：16

MAEは予測の偏りを検出できない点に注意
上記の例だと1データ誤差が30でも0でも「1データ辺りの誤差は16」と結論付けられます。

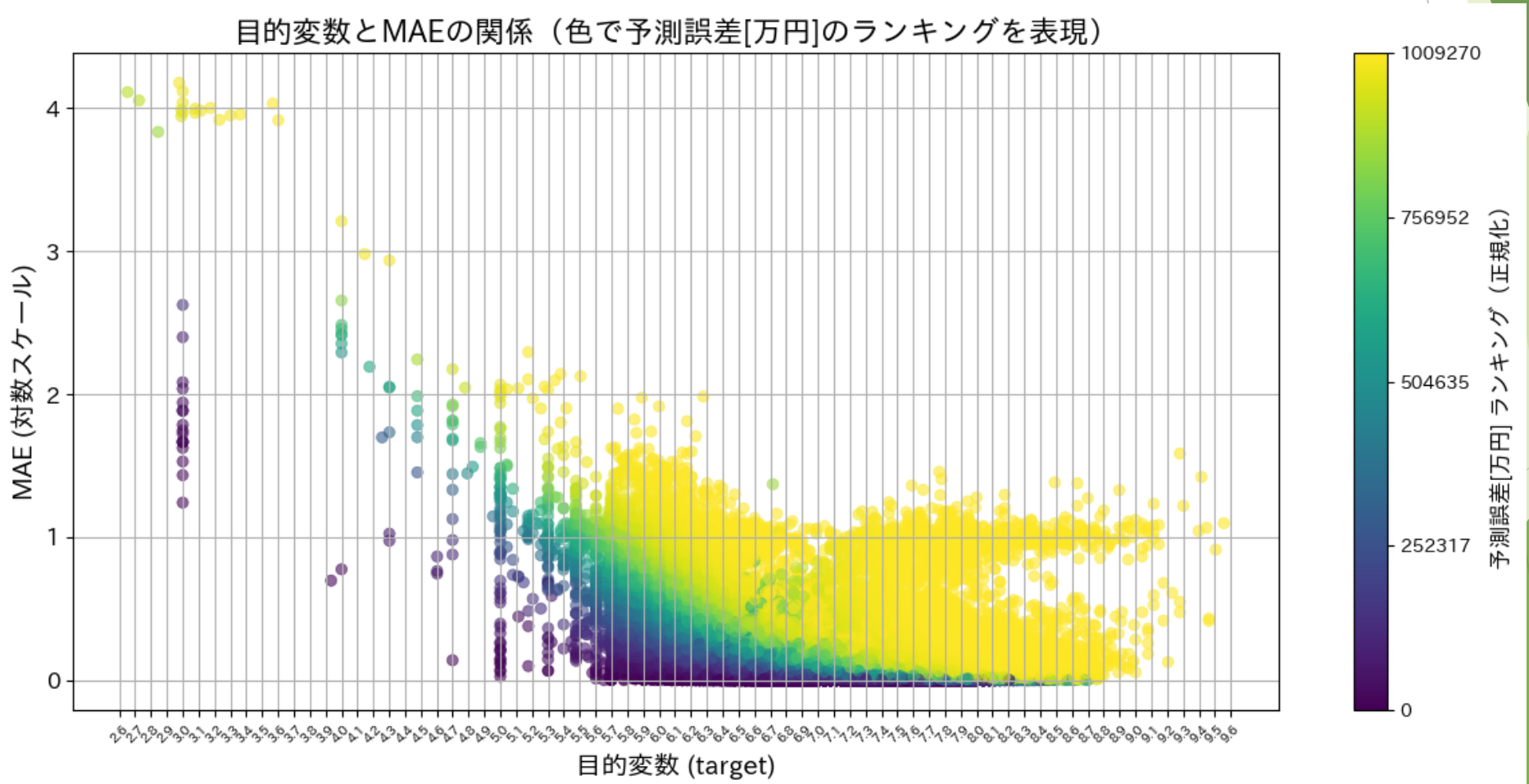
予測結果

未来の半年分の予測が1件あたり約0.063の誤差で予測できた

- 暫定評価(MAE)：0.065
- 最終評価(MAE)：0.063858 =>1位/172人

目的変数と平均絶対誤差（MAE）の関係

- この散布図は、モデルの予測誤差（MAE）と目的変数（target）の関係を示しています。
- 色は予測誤差（万円単位）のランキング（正規化）を表しており、黄色に近いほど誤差が大きく、紫に近いほど誤差が小さいことを意味します。



平均絶対誤差（MAE）のヒートマップ分析

図1: 取引価格と築年数によるMAE

•このヒートマップは、物件の取引価格帯と築年数帯ごとのモデル予測誤差（MAE）の平均を可視化しています。

図2: 取引価格と最寄駅までの距離によるMAE

•このヒートマップは、物件の取引価格帯と最寄駅までの距離帯ごとのMAEを示しています。

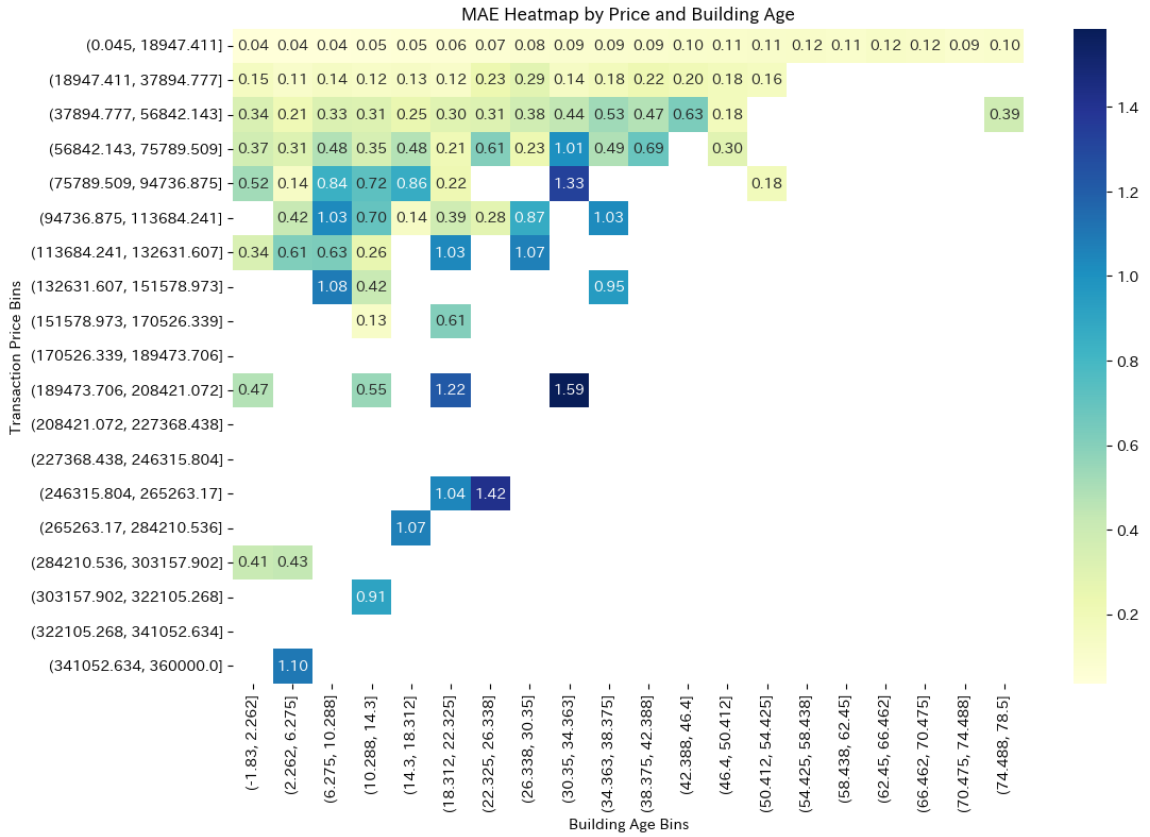


図1

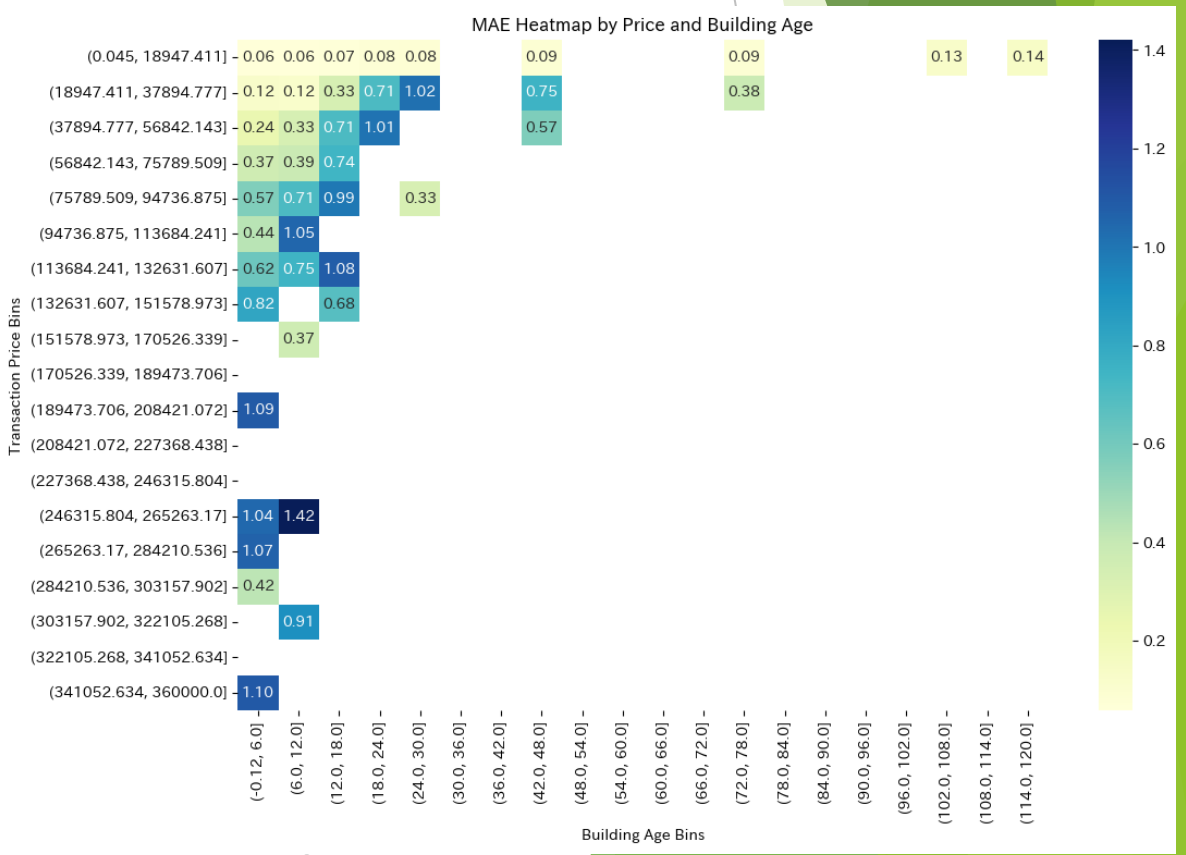


図2

特に効果があった特徴量

モデル	LightGBM	LightGBM	Catboost
損失関数	l1	l1	Lq:q=1.361
目的変数	単位面積あたりの取引価格_log	取引価格（総額）_log	単位面積あたりの取引価格_log
特徴量重要度 (上位11個)	最寄駅：名称 市区町村名 geohash_5 geohash_6 地区名 住所 geohash_7 geohash_4 建築年 築年数 取引時点	最寄駅：名称 geohash_6 geohash_5 市区町村名 地区名 住所 geohash_7 建築年 geohash_4 面積（㎡）－築年数 建築西暦年	建築年 改装 都道府県名 都市計画 旧耐震フラグ 取引時点 市区町村名 geohash_4 最寄駅：名称 購入までの築年数 geohash_5

※geohashについては次のページで補足

Geohash-レベル別範囲

Geohashは以下の範囲で世界地図をメッシュ状に区切った見た目になります。

Level	経度緯度方向の範囲(km)	緯度方向の範囲(km)
1	5000	5000
2	1250	1250
3	156.3	156.3
4	39.1	39.1
5	4.9	4.9
6	1.2	1.2
7	0.152	0.152
8	0.038	0.038
9	0.0048	0.0048
10	0.0012	0.0012
11	0.00015	0.00015
12	3.70E-05	3.70E-05

※上記の表は概算です。特に経度方向の範囲（km）は緯度に依存して変化するため、
実際の値と若干異なる場合があります。

考察

物件周辺の利便性が価格に大きな影響を与える

日用品・食料品店の存在

近隣に日用品や食料品を購入できる店舗があると、生活の利便性が高まります。これは不動産価格に正の影響を与える可能性があります。

その他の便利施設

医療機関、教育施設、公共交通機関なども、不動産価格に影響を与える重要な要素です。これらのデータを追加することで、予測精度が向上する可能性があります。

課題点・今後の施策

現状の課題

「大字」「小字」「郡」の正規化は成功したが、京都府の「通り名」までは正規化できていない。

今回は物件が属しているGeohashと駅からの距離が重要でしたので、この改善で精度向上が見込めます。

今後の改善策

物件周辺の商業施設データを導入します。これにより、生活利便性の評価が可能になります。

災害データの統合も有効になりそうです。安全性評価の精度向上につながります。