# 不動産価格予測コンペにおける欠損値処理と特徴量エンジニアリング戦略

## 1. 欠損値処理の全体方針

**データの欠損パターン分析**: 提供されたデータを分析した結果、完全にランダムな欠損 (MCAR) は存在せず、欠損は何らかのパターンに基づいて発生しています[[1]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20MCAR%20%28%E5%AE%8C%E5%85%A8%E3%81%AB%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E8%A9%B2%E5%BD%93%E3%81%AA%E3%81%97%20,%E3%81%99%E3%81%B9%E3%81%A6%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AB%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B3%E3%81%8C%E3%81%82%E3%82%8B)。一部は他の変数に依存したランダム欠損 (MAR) であり、例えば最寄駅情報や都市計画情報の欠損は地方物件など条件付きで発生しています[[2]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20MAR%20%28%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E4%BB%96%E3%81%AE%E8%A6%B3%E6%B8%AC%E5%A4%89%E6%95%B0%E3%81%AB%E4%BE%9D%E5%AD%98%E3%81%99%E3%82%8B%E3%81%8C%E3%80%81%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%AF%E6%9D%A1%E4%BB%B6%E4%BB%98%E3%81%8D%E3%81%A7%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%3A%20,%2A%2ARemarks%2A%2A%3A%20%E7%89%B9%E8%A8%98%E4%BA%8B%E9%A0%85%E3%81%8C%E3%81%AA%E3%81%84%E9%80%9A%E5%B8%B8%E7%89%A9%E4%BB%B6)。一方、多くは構造的な欠損 (MNAR) で、建物が存在しない物件では建物関連項目が一斉に欠損、道路に面さない土地では道路幅が欠損する、といった具合です[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)。こうしたMNAR欠損は欠損そのものが重要な情報を含んでおり、単純な平均値・中央値補完は不適切です[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)。

**基本方針**: **「欠損が意味する情報を極力損なわない補完」**を行います。具体的には以下の戦略を取ります:

* **MAR（他変数に依存した欠損）**: 条件付きランダムな欠損は、各種グループ（物件種別やエリアなど）の中央値・最頻値で補完します。例えば「最寄駅までの時間」はエリアや物件種別ごとの中央値で補完し、極端な値は除去します[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)。こうすることでデータの代表的な値で穴埋めし、バイアスを抑えます。
* **MNAR（構造的欠損）**: 欠損自体が意味を持つ項目は、**欠損フラグの追加**や**専用カテゴリの付与**によって情報を保持します[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。例えば建物関連情報が欠損＝「建物なし」と解釈できるので、補完として数値を0やカテゴリを"No Building"に置きつつ、「建物が存在するか」のフラグも追加してモデルに欠損の事実を伝えます[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。
* **極端な欠損率の列**: 欠損率が90%以上で情報量の少ない特徴は、モデルの複雑化を防ぐため**削除**も検討します[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)。例えば備考欄（Remarks）は93.4%が欠損であり、有意なテキスト解析が困難なため基本的には使用しません（代わりに有無フラグのみ利用）[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)。

**モデル別の考慮**: LightGBMとニューラルネット(NN)では欠損への対処方法が異なります。

* **LightGBM**: 決定木ベースのモデルであり、欠損値をSplitの一分岐として自動処理できます[[8]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=)。そのため数値項目では欠損をそのまま残してもモデルが学習可能ですが、後述するように明示的な欠損フラグや適切な補完を施すことで精度向上が期待できます[[9]](https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm#:~:text=Firstly%2C%20the%20default%20is%20in,the%20predictive%20value%20of%20missingness)。LightGBMはカテゴリ変数の欠損も「欠損カテゴリ」として扱えるので、カテゴリ項目は'Unknown'等に補完して問題ありません。
* **ニューラルネットワーク**: 欠損値を直接扱えないため**事前にすべて補完**する必要があります。また欠損による情報損失を防ぐため、**欠損箇所を示すバイナリフラグ**を追加入力します。例えば、建物床面積が欠損の場合はTotalFloorArea=0に補完しつつ「建物有無フラグ=0」を追加します。このようにNNでも欠損が持つ意味（建物が無い、道路が無い等）を学習可能にします。

**一貫性と汎用性**: 訓練データとテストデータで欠損率にほとんど差がないことが確認されています（差は±0.15%以内[[10]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%AC%A0%E6%90%8D%E7%8E%87%E3%81%AE%E5%B7%AE%E7%95%B0%EF%BC%88Train%20vs%20Test%EF%BC%89%20%E3%81%99%E3%81%B9%E3%81%A6%E3%81%AE%E3%82%AB%E3%83%A9%E3%83%A0%E3%81%A7%E6%AC%A0%E6%90%8D%E7%8E%87%E3%81%AE%E5%B7%AE%E3%81%AF%C2%B10.15)）。したがって**同じ欠損値処理戦略をTrain/Test双方に適用**します[[11]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20TotalFloorArea%20%7C%2062.19,)。特定のデータセットに過剰適合しないよう、シンプルで再現性の高い補完方法（中央値補完や一律カテゴリなど）を選択し、外部データの情報も活用してテストデータの欠損補完に役立てます。

**外部データ活用**: 外部データの利用が許可されているため、欠損補完や特徴量強化に活用します。例えば**地理データ**（市区町村コードから緯度経度への変換）を用いて、物件の近隣環境や中心地からの距離を推定し、駅距離の欠損補完に役立てます。市区町村ごとの人口・地価公示などを取得し、地域の平均的な土地価格で欠損値を補完・検証するといったアプローチも可能です。外部データで補完する際も、モデルに欠損フラグを与えることで「外部データで補完した」という事実を認識させ、過信による誤差を防ぎます。

以上の方針に基づき、**「各列ごと」に具体的な欠損処理方法を設計**します。

## 2. 列ごとの欠損値処理方針

各特徴列（または列グループ）の欠損率・欠損パターンを踏まえ、以下のように対応します。

* **駅・アクセス情報**（最寄駅名・最寄駅までの時間）: 最寄駅名(NearestStation)の欠損率は約3%と小さく、郊外など「駅情報なし」を意味すると考えられます[[12]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E9%87%8D%E8%A6%81%E3%81%AA%E7%89%B9%E5%BE%B4%2A%2A%3A%20,%E7%A8%8B%E5%BA%A6%EF%BC%89)。欠損時は新たなカテゴリ"No Station"で補完し、**「駅なし」**という情報を保持します[[13]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,MinTime%E3%81%8B%E3%82%89%E6%8E%A8%E5%AE%9A%20or%20%E5%89%8A%E9%99%A4)。最寄駅までの所要時間は文字列(TimeToNearestStation)と数値(MinTimeToNearestStation, MaxTimeToNearestStation)で与えられていますが、MinTimeは欠損4%程度なので**物件種別別の中央値**で補完します[[14]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,MinTime%E3%81%8B%E3%82%89%E6%8E%A8%E5%AE%9A%20or%20%E5%89%8A%E9%99%A4)[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)（都市部と郊外で通勤時間が異なるため、単純な全体中央値ではなく種別・地域ごとに補完）。MaxTimeはMinTimeとの相関が高く、欠損時に直接推定するのが難しいため**特徴量から除外**します[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)。なお、駅情報系の欠損は価格に大きな偏りを与えない傾向があります（欠損時平均価格は約4,300万円で全体平均と同程度[[12]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E9%87%8D%E8%A6%81%E3%81%AA%E7%89%B9%E5%BE%B4%2A%2A%3A%20,%E7%A8%8B%E5%BA%A6%EF%BC%89)）が、モデルには「駅情報無し」を明示します。
* **地域特性**（周辺地域の用途: Region）: 欠損率31.5%と中程度ですが、林地・農地など**物件種別に依存した欠損**です[[15]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%E3%81%AETradePrice%E5%B9%B3%E5%9D%87%3A%20%C2%A547%2C753%2C803)。例えば林地は99.9%がRegion欠損で、住宅地では17.4%に留まります[[15]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%E3%81%AETradePrice%E5%B9%B3%E5%9D%87%3A%20%C2%A547%2C753%2C803)。これは林地等には「住宅地/商業地」といった区分が適用できないためです。そこでRegion欠損は**'Unknown'カテゴリで補完**し、モデルに「地域区分なし＝住宅地以外」を学習させます[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。林地・農地ではRegionが常にUnknownになりますが、物件種別Type自体でそれらは区別できるため、情報漏れにはなりません。またRegion欠損時の取引価格は非欠損時に比べ約23%低いため（約3,299万円 vs 4,775万円）[[15]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%E3%81%AETradePrice%E5%B9%B3%E5%9D%87%3A%20%C2%A547%2C753%2C803)、Unknownカテゴリを通じて価格へのネガティブな影響をモデルが把握できるようにします。
* **建物情報**（間取り、延床面積、建築年、構造、用途、将来の用途、リノベ有無）: 建物関連の多くの項目で欠損率が高く、**「建物が存在しない（土地のみ物件）」**で一斉に欠損するMNARパターンが見られます[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)。例えば間取り(FloorPlan)は72.3%欠損、延床面積(TotalFloorArea)62.2%欠損など高率で、欠損時の物件は平均価格がやや高い傾向にあります[[16]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E3%81%93%E3%82%8C%E3%82%89%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AE%E5%AD%98%E5%9C%A8%E3%81%97%E3%81%AA%E3%81%84%E5%9C%9F%E5%9C%B0%E3%80%8D%E3%82%92%E7%A4%BA%E3%81%99MNAR%EF%BC%88Not%20At%20Random%EF%BC%89%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)（建物なしの土地取引は更地として高額になるケースあり）。これらは次のように処理します:
* **間取り (FloorPlan)**: 欠損＝建物無しと推測されるため、文字列カテゴリを**"No Building"**という新カテゴリで補完します[[17]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。具体的なプランがある場合はそのままカテゴリ化（ユニーク値86種あり）し、欠損時はNo Buildingとします。加えて、**「建物有無フラグ」**を新設し、FloorPlanが欠損だったレコードではHasBuilding=0とします[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。
* **延床面積 (TotalFloorArea)**: 数値項目で62.2%が欠損します[[17]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物無し物件では0とみなせるため、**0で補完**します[[18]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AF0%E5%9F%8B%E3%82%81%EF%BC%88MNAR%E5%A4%89%E6%95%B0%EF%BC%89%20df,BuildingYear%27%5D.fillna%280%29)。0は本来取り得ない値（建物があれば最低1以上）なので、モデルに「欠損（建物なし）」を識別させる明示的な符号となります。LightGBMではNaNのままでも処理できますが、NNでは扱えないため0埋めに統一しています。併せて前述のHasBuildingフラグでフォローします。
* **建築年 (BuildingYear)**: 34.8%欠損[[19]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物が無いケースや築年不詳のケースです。欠損時は**0で補完**し[[18]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AF0%E5%9F%8B%E3%82%81%EF%BC%88MNAR%E5%A4%89%E6%95%B0%EF%BC%89%20df,BuildingYear%27%5D.fillna%280%29)、その後取引年Yearとの差分で**築年数(BuildingAge)を計算**します[[20]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1)。建築年が0（欠損補完値）の場合は築年数を-1などの異常値に設定し、モデルに「建物なし/築年不明」であることを伝えます[[21]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1)。この-1自体もNNではそのまま扱えますし、木モデルでは範囲外の値として分岐条件に利用できます。
* **構造 (Structure)**: 34.3%欠損[[19]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物なしの土地では構造も存在しないため、**"No Building"**というカテゴリで補完します[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物がある場合の代表的な構造（木造Wが54.7%など）もモデルに学習させます[[23]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2AStructure%EF%BC%88%E5%BB%BA%E7%89%A9%E6%A7%8B%E9%80%A0%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,B%EF%BC%88%E3%83%96%E3%83%AD%E3%83%83%E3%82%AF%E9%80%A0%EF%BC%89%3A%201.1)。
* **現在の用途 (Use)**: 35.7%欠損[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物が無い土地の場合、現在利用用途は空地でしょう。そこで欠損時は**"Vacant Land"**（空地）といったカテゴリで補完します[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。建物がある場合は住宅、事務所等の値が入っているため、それらはそのままカテゴリ化します。
* **将来の用途 (Purpose)**: 65.2%欠損[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。将来利用目的が未定の場合に欠損となっているようです。現在の用途Useと重複する情報でもあり[[24]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=%E3%82%AB%E3%83%A9%E3%83%A0%E5%90%8D%E3%81%AE%E3%81%A8%E3%81%8A%E3%82%8A%E3%80%81%E5%B0%86%E6%9D%A5%E7%9A%84%E3%81%AA%E5%88%A9%E7%94%A8%E4%BA%88%E5%AE%9A%E7%9B%AE%E7%9A%84%E3%81%8C%E8%A1%A8%E3%81%95%E3%82%8C%E3%81%A6%E3%81%84%E3%81%9F%E3%80%82%20%E6%9C%AA%E5%AE%9A)、欠損率も高いため**モデルから削除**を検討します[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)（実際にQiita分析ではこのカラムを除外している例があります[[24]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=%E3%82%AB%E3%83%A9%E3%83%A0%E5%90%8D%E3%81%AE%E3%81%A8%E3%81%8A%E3%82%8A%E3%80%81%E5%B0%86%E6%9D%A5%E7%9A%84%E3%81%AA%E5%88%A9%E7%94%A8%E4%BA%88%E5%AE%9A%E7%9B%AE%E7%9A%84%E3%81%8C%E8%A1%A8%E3%81%95%E3%82%8C%E3%81%A6%E3%81%84%E3%81%9F%E3%80%82%20%E6%9C%AA%E5%AE%9A)）。使用する場合でも'Unknown'で補完しますが、情報利得が低ければ削除します。
* **リノベーション有無 (Renovation)**: 73.9%欠損[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。新築か未記録の場合に欠損になっている可能性があります[[16]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E3%81%93%E3%82%8C%E3%82%89%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AE%E5%AD%98%E5%9C%A8%E3%81%97%E3%81%AA%E3%81%84%E5%9C%9F%E5%9C%B0%E3%80%8D%E3%82%92%E7%A4%BA%E3%81%99MNAR%EF%BC%88Not%20At%20Random%EF%BC%89%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)。欠損時は**"Unknown"**カテゴリで補完し[[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)、既存のYes/Noと合わせて3値（Yes/No/Unknown）で扱います。建物が存在しない物件ではリノベーションも不要なので欠損=Noとも解釈できますが、確証がないためUnknownとしています。

なお、建物関連の欠損は「建物無し」を示すため、**建物有無に関する派生特徴**を作成します[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。例えば**HasBuilding**(建物有り=1/無し=0)や、複数カラムにまたがる**欠損カウント**（後述の特徴量エンジニアリング参照）によって、建物情報が丸ごと無い物件をモデルが容易に識別できるようにします。

* **土地情報**（土地形状、間口、方角）: 土地関連の情報も一部欠損があります。
* **土地形状 (LandShape)**: 31.7%欠損[[25]](file://file_0000000032cc6206941467b192be425b#:~:text=,or%20%E5%89%8A%E9%99%A4)。物件種別や測量難易度による欠損（MAR）で、「ほぼ整形」「不整形」等5種類のカテゴリーがあります[[26]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2ALandShape%EF%BC%88%E5%9C%9F%E5%9C%B0%E5%BD%A2%E7%8A%B6%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,%E3%81%9D%E3%81%AE%E4%BB%96%3A%204.4)。欠損時は**'Unknown'**で補完します[[25]](file://file_0000000032cc6206941467b192be425b#:~:text=,or%20%E5%89%8A%E9%99%A4)。元データ上、整形地か否かは価格に影響があるため、Unknownカテゴリによって「形状不明（＝測定困難な土地？）」を示します。
* **間口 (Frontage)**: 37.9%欠損[[25]](file://file_0000000032cc6206941467b192be425b#:~:text=,or%20%E5%89%8A%E9%99%A4)。道路に面する幅ですが、測定不可なケースや道路無しのケースで欠損しています（MAR寄り）[[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。欠損時の平均価格は非欠損時より35%低く、大きく不利な特徴です[[28]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%2A%2AFrontage%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%2A%2A%3A%20%E5%B9%B3%E5%9D%87%E4%BE%A1%E6%A0%BC%C2%A552%2C212%2C320%EF%BC%88%2B88%25%E9%AB%98%E3%81%84%EF%BC%89%E2%86%92%20%E6%95%B4%E5%BD%A2%E5%9C%B0%E3%83%BB%E5%84%AA%E8%89%AF%E7%89%A9%E4%BB%B6)。したがって**物件種別やエリアごとの中央値**で穴埋めし[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)、さらに**「Frontage欠損フラグ」**を追加します（LightGBMでは欠損値そのままでも良いですが、NNのために補完＋フラグにします）。こうすることで、例えば間口が欠損である土地（おそらく不整形で接道が難しい土地）であることをモデルが認識でき、価格低下要因として考慮できます。
* **方角 (Direction)**: 31.7%欠損[[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。土地が接する道路の方角ですが、道路が無かったり測定していない場合に欠損となります。欠損時は**'Unknown'**で補完します[[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。道路が無い場合、本来「方向」は存在しないため、Unknownは実質「適用不能」を意味します。これも後述の道路有無フラグで補足されます。方角自体（北向きか南向きか）は住宅価格に影響しうるため、既知の場合はOne-Hotエンコードなどでモデルに使います。
* **道路情報**（接道種類、道路幅）: 接道義務を満たさない物件では道路情報が欠損となり、価格に大きく影響します。
* **前面道路の種類 (Classification)**: 32.7%欠損[[29]](file://file_0000000032cc6206941467b192be425b#:~:text=,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。欠損は**「道路に面していない」**ことを示すMNARです[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。欠損時は新カテゴリ**"No Road"**で補完し、道路なし物件を明示します[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。その他、市区町村道・私道・国道など計5種類の値はOne-Hotエンコード可能です（私道か市道かで価格が変わるため）。
* **前面道路の幅 (Breadth)**: 33.2%欠損[[29]](file://file_0000000032cc6206941467b192be425b#:~:text=,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。これも道路無し物件で欠損（MNAR）します[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。欠損時は**0で補完**し[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)、**「接道幅無し」**を数値で表現します（幅0mというのは通常あり得ないため、欠損を示すのに都合の良い符号です）。併せて**道路アクセス有無フラグ**HasRoadAccessを作成し、Breadthが欠損（=0埋めされた）なら0、そうでなければ1とします[[31]](file://file_0000000032cc6206941467b192be425b#:~:text=df,Breadth%27%5D.notnull%28%29.astype%28int%29)。道路情報欠損は価格への負の影響が非常に大きく（欠損時平均価格は非欠損時より42%低い[[32]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E9%81%93%E8%B7%AF%E6%83%85%E5%A0%B1%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E4%BE%A1%E6%A0%BC%E3%81%AB%E5%A4%A7%E3%81%8D%E3%81%AA%E8%B2%A0%E3%81%AE%E5%BD%B1%E9%9F%BF%EF%BC%88%E6%8E%A5%E9%81%93%E7%BE%A9%E5%8B%99%E9%81%95%E5%8F%8D%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%EF%BC%89)）、モデルにはこの情報を確実に捉えさせます。
* **都市計画・法規制情報**（用途地域、建ぺい率、容積率）: いずれも欠損は少なめ（~4–6%）ですが、一部郊外物件で**「規制対象外」**を意味する欠損が発生しています[[33]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A)。
* **用途地域 (CityPlanning)**: 3.9%欠損[[34]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A)。市街化調整区域など都市計画外の地域で欠損になる場合があり[[35]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2ARemarks%2A%2A%3A%20%E7%89%B9%E8%A8%98%E4%BA%8B%E9%A0%85%E3%81%8C%E3%81%AA%E3%81%84%E9%80%9A%E5%B8%B8%E7%89%A9%E4%BB%B6)、欠損時は**"Outside City Planning"**（都市計画区域外）といったカテゴリで補完します[[34]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A)。20種類程度のカテゴリーがあり、これはOne-Hotエンコード可能な範囲です。
* **建ぺい率 (CoverageRatio)・容積率 (FloorAreaRatio)**: ともに5.9%欠損[[33]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A)。用途地域がない土地（規制なし）などで欠損します[[35]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2ARemarks%2A%2A%3A%20%E7%89%B9%E8%A8%98%E4%BA%8B%E9%A0%85%E3%81%8C%E3%81%AA%E3%81%84%E9%80%9A%E5%B8%B8%E7%89%A9%E4%BB%B6)。欠損時は**物件種別ごとの中央値**で補完します[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)。例えば郊外の広大な土地では平均的な値で穴埋めし、都市部の住宅地ではその中央値で埋めます。こうすることで、不明な規制値による極端な誤差を防ぎます。なお、規制そのものが無い場合、本来は非常に大きな値（例えば容積率1300%など）になり得ますが、そうした外れ値はモデルに扱わせず中央値程度で代用します。
* **備考 (Remarks)**: フリーテキストの備考欄で、**93.4%が欠損**しています[[36]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E5%89%8A%E9%99%A4%20or%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96)。多くの物件では備考が記載されず、一部だけ「瑕疵あり」「借地権付き」等の特殊情報が入ります[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)。欠損率が非常に高くテキスト解析も困難なため、**モデルからは基本的に除外**します[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)。ただし備考が**ある**物件は平均価格が10%ほど低く（問題物件である可能性）[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)、情報価値があります。そのため**備考の有無フラグ**だけは特徴量に加えます（Remarksが欠損なら0、値があれば1）。テキスト内容については、もし時間とリソースが許せば特定キーワード（例：「再建築不可」「事故」など）の有無を特徴量化することも検討しますが、今回は労力対効果からフラグのみに留めます。

以上を踏まえ、主な列の欠損補完戦略を以下の表にまとめます。

## 6. サマリー（欠損処理方針の表）

| **特徴量グループ** | **欠損率とパターン** | **補完・処理戦略** | **追加特徴量や備考** |
| --- | --- | --- | --- |
| **駅情報**: 最寄駅名・所要時間 | 欠損率3~5%（MAR）[[12]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E9%87%8D%E8%A6%81%E3%81%AA%E7%89%B9%E5%BE%B4%2A%2A%3A%20,%E7%A8%8B%E5%BA%A6%EF%BC%89) | 駅名欠損は"No Station"で補完[[13]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,MinTime%E3%81%8B%E3%82%89%E6%8E%A8%E5%AE%9A%20or%20%E5%89%8A%E9%99%A4)。時間は種別別中央値で補完[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)。MaxTimeは情報冗長のため削除[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)。 | 駅情報欠損は価格影響小。外部データで駅座標等を取得し距離計算も検討可。 |
| **地域**: Region (周辺地域用途) | 欠損率31.5%（MNAR: 種別依存）[[15]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%E3%81%AETradePrice%E5%B9%B3%E5%9D%87%3A%20%C2%A547%2C753%2C803) | 'Unknown'カテゴリで補完[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)（林地/農地など地域区分なし物件）。 | 欠損＝住宅地以外を示唆。Typeで林地等は区別済み。 |
| **建物情報**: FloorPlan, TotalFloorArea, BuildingYear, Structure, Use, Purpose, Renovation | FloorPlan 72%, FloorArea 62%, 等（MNAR: 建物なし）[[16]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E3%81%93%E3%82%8C%E3%82%89%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AE%E5%AD%98%E5%9C%A8%E3%81%97%E3%81%AA%E3%81%84%E5%9C%9F%E5%9C%B0%E3%80%8D%E3%82%92%E7%A4%BA%E3%81%99MNAR%EF%BC%88Not%20At%20Random%EF%BC%89%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84) | 間取り・構造等カテゴリは"No Building"/"Vacant Land"/"Unknown"で補完[[17]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。延床面積・建築年は0埋め（建物無しを示す）[[18]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AF0%E5%9F%8B%E3%82%81%EF%BC%88MNAR%E5%A4%89%E6%95%B0%EF%BC%89%20df,BuildingYear%27%5D.fillna%280%29)。Purposeは欠損多く削除検討[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)。Renovationは"Unknown"補完。 | 建物有無フラグ等を作成[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)し欠損パターンを特徴量化。BuildingYearとYearから築年数算出[[20]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1)。 |
| **土地情報**: LandShape, Frontage, Direction | LandShape 31.7%, Frontage 37.9%（MAR）、Direction 31.7%[[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)[[28]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%2A%2AFrontage%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%2A%2A%3A%20%E5%B9%B3%E5%9D%87%E4%BE%A1%E6%A0%BC%C2%A552%2C212%2C320%EF%BC%88%2B88%25%E9%AB%98%E3%81%84%EF%BC%89%E2%86%92%20%E6%95%B4%E5%BD%A2%E5%9C%B0%E3%83%BB%E5%84%AA%E8%89%AF%E7%89%A9%E4%BB%B6) | 土地形状・方角は'Unknown'で補完[[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4)。間口Frontageは種別別中央値で補完[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)し、欠損フラグ付与。 | Frontage欠損は価格-35%と相関[[28]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%2A%2AFrontage%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%2A%2A%3A%20%E5%B9%B3%E5%9D%87%E4%BE%A1%E6%A0%BC%C2%A552%2C212%2C320%EF%BC%88%2B88%25%E9%AB%98%E3%81%84%EF%BC%89%E2%86%92%20%E6%95%B4%E5%BD%A2%E5%9C%B0%E3%83%BB%E5%84%AA%E8%89%AF%E7%89%A9%E4%BB%B6)→フラグでモデルに伝達。 |
| **道路情報**: Classification, Breadth | 各約33%（MNAR: 道路なし）[[32]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E9%81%93%E8%B7%AF%E6%83%85%E5%A0%B1%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E4%BE%A1%E6%A0%BC%E3%81%AB%E5%A4%A7%E3%81%8D%E3%81%AA%E8%B2%A0%E3%81%AE%E5%BD%B1%E9%9F%BF%EF%BC%88%E6%8E%A5%E9%81%93%E7%BE%A9%E5%8B%99%E9%81%95%E5%8F%8D%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%EF%BC%89) | 道路種類欠損は"No Road"カテゴリ[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。幅員欠損は0埋め[[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4)。 | 道路有無フラグ (HasRoadAccess) 作成[[31]](file://file_0000000032cc6206941467b192be425b#:~:text=df,Breadth%27%5D.notnull%28%29.astype%28int%29)。欠損は価格大幅低下に直結[[32]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E9%81%93%E8%B7%AF%E6%83%85%E5%A0%B1%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E4%BE%A1%E6%A0%BC%E3%81%AB%E5%A4%A7%E3%81%8D%E3%81%AA%E8%B2%A0%E3%81%AE%E5%BD%B1%E9%9F%BF%EF%BC%88%E6%8E%A5%E9%81%93%E7%BE%A9%E5%8B%99%E9%81%95%E5%8F%8D%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%EF%BC%89)。 |
| **都市計画**: CityPlanning, CoverageRatio, FloorAreaRatio | ~4-6%（MAR: 区域外等） | 用途地域欠損は"Outside City Planning"等で補完[[34]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A)。割合系は種別別中央値補完[[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29)。 | 極端値は扱わず中央値で安定化。 |
| **備考**: Remarks (テキスト) | 93.4%（MAR: 特記事項なし）[[37]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,%E5%89%8A%E9%99%A4%20or%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96) | **列自体を削除**[[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E)（情報量極小）。代わりに欠損/非欠損フラグのみ利用[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)。 | 備考あり物件は価格低め傾向[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)→フラグで捕捉。テキスト解析は任意。 |

※上記以外、IDや取引年(Year)、四半期(Quarter)等は欠損がないためそのまま使用します。

## 3. 特徴量エンジニアリングの戦略

欠損値処理後のデータを基に、モデル学習に有用な新規特徴量の作成とカテゴリ変数のエンコード戦略を策定します。外部データの統合や地理情報の活用も含め、以下の方針で特徴量エンジニアリングを行います。

### カテゴリ変数のエンコード方針

* **高カードinality変数**（ユニーク値が非常に多いカテゴリ）:
* *最寄駅名 (NearestStation)*: ユニーク約1,743件[[38]](file://file_0000000032cc6206941467b192be425b#:~:text=,Target%20Encoding%E6%8E%A8%E5%A5%A8)と非常に多いため、**ターゲットエンコーディング**を採用します[[38]](file://file_0000000032cc6206941467b192be425b#:~:text=,Target%20Encoding%E6%8E%A8%E5%A5%A8)。各駅の平均取引価格を学習データで計算し、その期待値で駅名を数値化します（クロスバリデーション内で実施し過学習を防止）。これにより主要駅かローカル駅かといった影響を数値で表現できます。併せて、駅名から得られる外部情報（駅の所在地、路線数、ターミナル駅フラグなど）も取得できれば特徴量に追加します。例えば「山手線の駅かどうか」「急行停車駅か」「終電の本数」等は不動産価格に影響し得ますが、情報収集コストとの兼ね合いで限定的に検討します。
* *地区名 (DistrictName)*: ユニーク約12,436件[[39]](file://file_0000000032cc6206941467b192be425b#:~:text=,Label%20Encoding%E5%8F%AF)であり、住所の細かい字丁目レベルまで含まれるため、こちらも**ターゲットエンコーディング**が有効です[[39]](file://file_0000000032cc6206941467b192be425b#:~:text=,Label%20Encoding%E5%8F%AF)。各地区の平均価格や取引件数を計算し、平均価格（または平均ログ価格）を特徴量として与えます。これもCV内で算出し、テストデータには学習全体で得た平均を適用します。こうすることで、高級住宅街かどうか、地方の過疎地域か、といった地域性をモデルが把握できます。
* *間取り (FloorPlan)*: ユニーク86種[[40]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2AFloorPlan%EF%BC%88%E9%96%93%E5%8F%96%E3%82%8A%EF%BC%89%E3%81%AE%E4%B8%BB%E3%81%AA%E5%80%A4%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,%E3%83%A6%E3%83%8B%E3%83%BC%E3%82%AF%E5%80%A4%E6%95%B0%3A%2086)と比較的多く、One-Hotではやや次元が増えます。**ターゲットエンコーディング**または**頻度の低いカテゴリのまとめ**を検討します[[41]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding%E5%8F%AF)。例えば「1R/1K」「2LDK/3LDK以上」などグループ化しつつ主要なカテゴリのみOne-Hotにする方法もあります。また、間取り文字列をパースして**部屋数や間取り構成の数値特徴**に変換することも可能です（例: 部屋数=3, LDK有無=1, S(サービスルーム)有無=0 等）。これにより未出現の間取りにも対応しやすくなりますが、まずはカテゴリ補完+エンコードで対応します。建物が無い場合は"No Building"カテゴリで処理済みなのでエンコード対象外です。
* *市区町村コード／名称 (MunicipalityCode, Municipality)*: 市区町村は348通り程度[[42]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20or%20Target%20Encoding)。中程度のカーディナリティなので**Label Encoding（ラベルエンコード）**で整数に変換しLightGBMにはカテゴリ扱いで与えます[[42]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20or%20Target%20Encoding)。NNではOne-Hotだと348次元になり大きいため、**Embeddingレイヤー**で圧縮表現させるか、こちらもターゲットエンコーディングで平均価格などを与える方法があります。市区町村は都道府県と地区名の中間レベルの地域情報なので、都道府県と組み合わせて地域性を表現する特徴量（例: 東京都23区フラグ、政令指定都市フラグなど）を作るのも有効です。
* **低～中カードinality変数**（ユニーク値が少ないカテゴリ）:
* *物件種別 (Type)*: ユニーク約10種[[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding)（住宅地、宅地(建物有)、中古マンション、林地、農地 等）。**One-Hotエンコード**でダミー変数化します[[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding)。Typeは価格に直結する重要カテゴリで（例えば林地は極端に安い）、ツリー系モデルでは自動で分割されますが、NNではOne-Hotにより特定の種別に重みを持たせやすくします。
* *都道府県 (Prefecture)*: 関東7都県のみなので**One-Hotエンコード**します[[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding)。東京都・神奈川県など地域差を直接学習させます。都道府県は地理的広がりを大きく左右するため、後述の地理特徴とも組み合わせます。
* *構造 (Structure)*: ユニーク6種（木造W, RC造, 鉄骨S造など）[[23]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2AStructure%EF%BC%88%E5%BB%BA%E7%89%A9%E6%A7%8B%E9%80%A0%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,B%EF%BC%88%E3%83%96%E3%83%AD%E3%83%83%E3%82%AF%E9%80%A0%EF%BC%89%3A%201.1)。**One-Hotエンコード**可能です[[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding)。建物がある物件のみ値があります。構造は建物価格に影響（RC造マンションは木造戸建より高額等）するため有効です。
* *土地形状 (LandShape)*: ユニーク5種[[26]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2ALandShape%EF%BC%88%E5%9C%9F%E5%9C%B0%E5%BD%A2%E7%8A%B6%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,%E3%81%9D%E3%81%AE%E4%BB%96%3A%204.4)（ほぼ整形・長方形・不整形など）。**One-Hotエンコード**します[[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding)。Unknown補完した欠損分も含め、形状ごとの価格差（整形地は高値等）を捉えます。
* *方角 (Direction)*: ユニーク8方位程度。こちらも**One-Hotエンコード**します[[44]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding)（欠損はUnknownとして1カテゴリ追加）。日当たりや風水の観点から南向きが高めなどの傾向を学習できる可能性があります。
* *道路種類 (Classification)*: ユニーク5種（市道、私道、県道、国道、その他）。**One-Hotエンコード**します（No Road含め6種）。道路の公道/私道区分や幹線道路かどうかは評価額に影響するため、モデルに直接区別させます。
* *用途地域 (CityPlanning)*: ユニーク20種程度。One-Hotエンコード可能ですが、20次元とやや多いのでNNではEmbeddingにしてもよいでしょう[[45]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding%E5%8F%AF)。主要カテゴリ（第一種低層住居専用地域など）は件数も多く影響大なのでOne-Hotで細かく反映します。
* *リノベーション (Renovation)*: ユニーク2種（有/無）＋欠損(Unknown)。**One-Hotエンコード**でYes/No/Unknownをそれぞれ示すダミーを作ります。リノベ有無自体欠損多いですが、あれば中古物件価値向上要因になり得ます。
* **フラグ変数**: もともと数値で0/1が入っているフラグ類（例: AreaIsGreaterFlag(敷地2000㎡以上), TotalFloorAreaIsGreaterFlag(延床2000㎡以上), PrewarBuilding(戦前建築) 等）はそのまま**数値特徴量（0/1）**として扱います。これらは既に極端な条件を示す有用な特徴で、One-Hotと実質同じ意味なので追加処理は不要です。例えばPrewarBuilding=1なら築80年以上で建物価値が低い可能性が高く、価格予測に効いてくるでしょう。
* **ターゲットエンコーディングの留意点**: 高カードinality変数のターゲットエンコーディングは強力ですが、**リーケージ防止**が必須です[[46]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20,Type%E5%88%A5%E3%81%AE%E8%A3%9C%E5%AE%8C%E6%88%A6%E7%95%A5%E3%81%8C%E6%9C%80%E3%82%82%E5%8A%B9%E6%9E%9C%E7%9A%84%E3%81%AA%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)。学習時はK-fold CV内で各foldの平均価格を算出して他foldのサンプルに適用し、テストには全データで算出した値を用います。これにより過学習を避け、汎化性能を確保します[[47]](file://file_0000000032cc6206941467b192be425b#:~:text=%E9%AB%98%E3%82%AB%E3%83%BC%E3%83%87%E3%82%A3%E3%83%8A%E3%83%AA%E3%83%86%E3%82%A3%E5%A4%89%E6%95%B0%EF%BC%88%E3%83%A6%E3%83%8B%E3%83%BC%E3%82%AF%E5%80%A4%E3%81%8C%E5%A4%9A%E3%81%84%EF%BC%89%20,Hot%20Encoding%E5%8F%AF)。LightGBMとNN双方でこのエンコードを活用しますが、特にNNはカテゴリをEmbeddingで処理できるため、**ターゲットエンコード値とEmbeddingの併用**も考えられます。例えば最寄駅をEmbeddingしつつ、別途「駅平均価格」の数値特徴を与えることで、NNが線形効果と非線形効果の両面から駅情報を捉えられるようにします。

### 集約・統計系特徴量の生成

* **欠損パターン特徴**: 欠損処理時に付与したフラグをさらに組み合わせ、**欠損の数を数えた特徴量**を作ります。例えば建物関連4項目（間取り・延床・築年・構造）の欠損数を合計したMissingBuildingInfoを定義し、0なら建物情報完備、4なら建物情報皆無という指標にします[[48]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20%E6%AC%A0%E6%90%8D%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B3%E3%81%8B%E3%82%89%E3%81%AE%E7%89%B9%E5%BE%B4%E9%87%8F%20,building_cols%5D.isnull%28%29.sum%28axis%3D1)。同様に道路関連4項目の欠損数MissingRoadInfoも作ります[[49]](file://file_0000000032cc6206941467b192be425b#:~:text=df)。これにより「建物が無い土地」「道路が無い土地」を数量的に表現でき、モデルがそれらの組み合わせ効果を捉えやすくなります。LightGBMなら欠損フラグ自体で分割を学習しますが、NNではこのような集計が効く可能性があります。
* **築年数・経過年**: 前述の通り取引年Yearと建築年BuildingYearから**築年数 (BuildingAge)**を計算します[[21]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1)。築年数が負値（-1）なら建物なし、0以上なら具体的な築年になります。築年数は経年劣化による価格減価を捉える重要な特徴です。ただし築年数が極端に大きい（戦前建築など）場合は一律古建築として扱われ、価格への影響は頭打ちになる可能性もあります。必要に応じて築年数をカテゴリにビニング（例: 新築0年、~10年、~30年、~50年、50年超）した特徴を作ることも検討します。
* **面積関連特徴**: 土地面積Areaと延床面積TotalFloorAreaから、**建蔽率・容積率の実現値**を計算します。例えば**延床面積率** = TotalFloorArea / Areaを算出し、FloorAreaRatioActualとして特徴量化します[[50]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20%E9%9D%A2%E7%A9%8D%E9%96%A2%E9%80%A3%E3%81%AE%E7%89%B9%E5%BE%B4%E9%87%8F%20%60%60%60python%20%E5%BB%B6%E5%BA%8A%E9%9D%A2%E7%A9%8D%E7%8E%87%EF%BC%88%E5%BB%BA%E7%89%A9%E3%81%AE%E9%9A%8E%E6%95%B0%E3%81%AE%E6%8C%87%E6%A8%99%EF%BC%89%20df,FloorAreaRatioActual%27%5D.fillna%280%29)。これは実際どれだけの容積率を使っているかを示し、例えば容積率の上限に対し余裕があるなら将来拡張可能で土地価値が高い、といった解釈ができます。建物が無い場合この値は0となります（欠損は0補完済み）[[51]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%B6%E5%BA%8A%E9%9D%A2%E7%A9%8D%E7%8E%87%EF%BC%88%E5%BB%BA%E7%89%A9%E3%81%AE%E9%9A%8E%E6%95%B0%E3%81%AE%E6%8C%87%E6%A8%99%EF%BC%89%20df,FloorAreaRatioActual%27%5D.fillna%280%29)。また、公的規制の容積率FloorAreaRatioとの比や差を取ることで「容積率余裕度」の特徴も作成可能です（例: UsageRatio = FloorAreaRatioActual / FloorAreaRatio）。同様に建ぺい率についても、土地面積に対する建物の一階部分面積（未提供）から計算できれば特徴にできますが、今回はデータ未提供のため割愛します。
* **地理的集約特徴**: 地域性を定量化するため、**エリアごとの統計量**を特徴に加えます（リークに注意しつつCV内算出）。例えば**「同じ市区町村内の取引件数」**や**「平均取引価格」**を計算して、その地域の人気度や価格水準をモデルに与えます。件数は物件周辺の流動性を示し、価格水準は立地グレードを示します。これらはターゲットエンコーディングに近いですが、価格以外の統計（件数や面積平均など）も含める点が異なります。また**最寄駅ごとの乗降客数**（外部データ）や**路線価**（公示地価の道路沿い評価）などが取得できれば、それらも駅・場所の評価指標として加えられます。外部の公的データを用いて**「この地域は地価が高い」**といった情報を暗に持たせることは、上位入賞者もよく行う手法です。
* **時間的特徴**: 取引年や四半期は価格動向のトレンドを反映します。そこで**四半期を表すダミー特徴**（Q1～Q4）や、年度をまたいだ**経済指標の外部データ**（例えばその年の金利や景気指数）を紐付けることも検討します。今回は評価指標がRMSLEであり、年次の価格変動（インフレ等）はある程度対数変換で吸収されますが、**年度トレンド**や**季節要因**はモデルに織り込んでおきます。具体的にはYearは連続値のままNNに入力し、LightGBMではそのまま使用（必要なら年ごとのOne-Hot）、Quarterは4種のOne-Hotにします。年が進むにつれ地価上昇傾向があるならモデルが重みを調整するでしょう。
* **テキストデータ/備考の活用**: 前述の通り備考テキストはフラグのみとしますが、**テキストマイニング**による特徴抽出も余裕があれば実施します。例えば「調停」「競売」「瑕疵」等のキーワード出現フラグを作れば、それ自体がマイナス要因として効く可能性があります（実際、備考あり物件は平均価格が約10%低い[[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7)ため、それを説明する文言が含まれているはずです）。ただし件数が少なく信頼性も低いため、上位入賞には直結しにくい細かな工夫となります。時間対効果を考慮し、本戦略では重点を置きません。
* **地理情報ベースの特徴**: 物件所在地をより定量化するため、外部の地理情報も統合します。例えば、市区町村コードから**緯度経度**を取得し、その地点から主要都心部（例: 東京駅）までの距離を計算したり、最寄駅の緯度経度との直線距離を計算したりします。時間が許せば地図上の座標で**クラスタリング**を行い「近隣エリアID」を作ることも有効です。K-meansで緯度経度をクラスタリングして得られたエリア区分をOne-Hotすると、価格帯の近いエリアをモデルが認識できます。また、地理的階層（都道府県-市区町村-地区）の中で、上位レベルの情報を下位に伝播させる特徴も考えられます（例: 市区町村平均価格を各地区に割り当てる）。外部からは**人口密度**や**世帯収入**など地域の社会経済指標を収集し、市区町村レベルでマージするのも一策です。これらは不動産価格に本質的な影響を持つため、もし取得可能であれば有力な追加特徴となります。

以上の特徴量エンジニアリングにより、元データに存在しない視点（欠損パターン、時間・空間的要素、統計的集約）をモデルに提供し、予測精度の向上を図ります。

## 4. LightGBMとニューラルネットモデルでの特徴量活用の違い

提案した特徴量をLightGBMとNNのアンサンブルで活用するにあたり、各モデルの性質に応じてアプローチを調整します。

* **欠損値の扱い**: LightGBMは欠損値を自動処理できるため、数値項目では補完せずNaNのままでも学習可能です[[8]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=)。実際、多くの決定木系モデルでは欠損を一つの分岐条件として利用できます[[9]](https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm#:~:text=Firstly%2C%20the%20default%20is%20in,the%20predictive%20value%20of%20missingness)。しかし我々はNNとの共通前処理のため**全項目補完＋欠損フラグ付与**を採用しました。LightGBMにおいてこの戦略は、モデル自身が行う欠損分岐と同等かそれ以上の情報を提供します[[9]](https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm#:~:text=Firstly%2C%20the%20default%20is%20in,the%20predictive%20value%20of%20missingness)。例えば欠損フラグを与えることは、木が欠損を検知するために追加の分岐を使う必要を減らす効果があります[[9]](https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm#:~:text=Firstly%2C%20the%20default%20is%20in,the%20predictive%20value%20of%20missingness)。一方NNでは欠損そのものを扱えないため、**補完＋フラグ**は不可欠であり、LightGBMとNNで欠損処理方法を大きく変えることなく統一できています。
* **カテゴリ変数のエンコード**: LightGBMではカテゴリ変数を整数ラベルに変換してcategorical\_featureパラメータで指定すれば、アルゴリズム内部でカテゴリ分割を最適化します。したがって高カードinalityでもラベルエンコード＋ターゲットエンコードなどで対応可能です。一方、NNではカテゴリを直接数値として入力すると順序関係が無いのに数値の大小に意味を持たせてしまうため不適切です。NN側では**One-Hotエンコード**（低次元カテゴリ）や**Embedding層**（高次元カテゴリ）を用いて離散情報をベクトルに展開します。例えば最寄駅はEmbedding(1743→例えば32次元ベクトル)により近い駅同士の表現を学習できますし、都道府県はOne-Hot(7次元)で明示的に区別できます。LightGBMではOne-Hotにすると逆に木の深さを取ってしまう可能性があり、**ラベルエンコードのままカテゴリ扱い**にするほうが効率的です[[45]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding%E5%8F%AF)。モデル間でエンコード方法が異なるため、**特徴量の使い分け**として、LightGBM用の前処理とNN用の前処理を並行して行います。ただしターゲットエンコードの結果など数値特徴は両モデルで共通利用できます。例えば駅の平均価格といった特徴量は数値なのでLightGBMでもNNでも同じスケールのまま入力できます。
* **数値特徴のスケーリング**: LightGBMは決定木ベースであるため、特徴量のスケールに不変です。面積が平方メートル単位であろうと、対数変換しようと結果は同じです。しかしNNでは勾配学習の都合上**入力スケーリングが重要**です。標準化（平均0,標準偏差1）やMin-Max正規化により、全数値特徴をある程度揃えます。特に面積や価格など分布が歪んだものは対数変換も有効です[[52]](file://file_0000000032cc6206941467b192be425b#:~:text=TradePrice%EF%BC%88%E5%8F%96%E5%BC%95%E4%BE%A1%E6%A0%BC%EF%BC%89%20,%2A%2A%E5%AF%BE%E6%95%B0%E5%A4%89%E6%8F%9B%E5%BE%8C%E3%81%AE%E6%AD%AA%E5%BA%A6%2A%2A%3A%200.36%EF%BC%88%E3%81%BB%E3%81%BC%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83%EF%BC%89)。本タスクでは目的変数は対数変換済みでRMSLEを評価するため、モデルにはlog1p(価格)を学習させます[[53]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2A%E5%AF%BE%E6%95%B0%E5%A4%89%E6%8F%9B%E5%BE%8C%E3%81%AE%E6%AD%AA%E5%BA%A6%2A%2A%3A%200.36%EF%BC%88%E3%81%BB%E3%81%BC%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83%EF%BC%89)。これはLightGBMでもNNでも共通ですが、NNでは入力特徴（面積や築年数等）についても適宜対数やBox-Cox変換で正規に近づけ、さらに標準化して収束性を高めます。LightGBMでは生の値（と欠損フラグ）をそのまま使い、木が勝手に分割点を学習します。
* **外れ値・分布への耐性**: LightGBMは外れ値にロバストで、極端に大きな値でも葉に分岐させ重みを付けるだけです。一方NNは外れ値が損失を大きくし学習を不安定化させます。そこで、例えば土地面積が極端に大きいサンプル（林地で何万㎡など）は、NNではログ変換やクリッピングで影響を緩和します。価格についてはRMSLE評価のためログ変換必須ですが[[53]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2A%E5%AF%BE%E6%95%B0%E5%A4%89%E6%8F%9B%E5%BE%8C%E3%81%AE%E6%AD%AA%E5%BA%A6%2A%2A%3A%200.36%EF%BC%88%E3%81%BB%E3%81%BC%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83%EF%BC%89)、説明変数側も同様に**ロバストスケーリング**するのがNNでは肝要です[[54]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20,Type%E5%88%A5%E3%81%AE%E8%A3%9C%E5%AE%8C%E6%88%A6%E7%95%A5%E3%81%8C%E6%9C%80%E3%82%82%E5%8A%B9%E6%9E%9C%E7%9A%84%E3%81%AA%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)。またNNでは入力にDropoutを掛けたり正則化で過学習を防ぐ工夫も入れますが、LightGBMでは過度に外れ値へ適合しないよう木の深さやリーフ数に制限を設ける程度で対応します。
* **相互作用の扱い**: 決定木モデルは特徴量間の非線形な相互作用を自動で表現できます。例えば「駅遠 \* 都市部 \* 広大面積 \* 道路無し」という組み合わせが価格に効くなら、木構造がそれを表現します。NNでも十分な深さとユニットがあれば相互作用を学習できますが、シグモイドやReLUを通した連続変換の積み重ねになるため、過学習しやすい組み合わせや表現できない関数形もあります。そのため上記で作成した**集約特徴**（地域平均や欠損数など）は、NNにとっては一次関数で捉えやすい情報を事前に計算したものと言え、精度向上に寄与します。一方LightGBMにとってはこれら新特徴は既存特徴の組み合わせを人間が与えたものであり、木が自力で探せるパターンを提供している可能性があります。それでも計算コスト削減や過学習抑制のため、この**相互作用を明示した特徴量**は両モデルで積極的に使います。例えばMissingBuildingInfo=4（建物情報全欠）かつAreaIsGreaterFlag=1（広大土地）なら極めて低価格帯、といった条件はNNには明示的特徴が有効で、LightGBMでもそれを1回の分割でキャプチャできます。
* **特徴量の選択**: LightGBMでは特徴量重要度を算出し不要な特徴の影響を調べられます。NNでは明示的な重要度は得にくいため、アンサンブル全体として特徴を広く使いますが、NN訓練時には**過学習を招く冗長特徴はドロップ**する可能性も検討します。例えばOne-Hotエンコードで次元が増えたカテゴリは、効果が薄い場合NNではEmbeddingに変更する/次元圧縮する、など動的に調整します。LightGBMは冗長特徴があっても自動でスプリットに使わなければ影響が少ないですが、NNでは入力次元増加それ自体が負担になるためです。したがって**特徴量選択の閾値**はモデルにより異なり、最終的に双方のモデルで安定して効果がある特徴をエンsembleに残します。

## 5. 欠損がモデル精度に与える影響と軽減手法

**欠損値の影響**: 欠損が適切に処理されていないと、モデル精度に大きな悪影響を及ぼします。例えば、本データでは前面道路幅が欠損（=道路無し）だと価格が平均で約半分になるほど低下します[[32]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E9%81%93%E8%B7%AF%E6%83%85%E5%A0%B1%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E4%BE%A1%E6%A0%BC%E3%81%AB%E5%A4%A7%E3%81%8D%E3%81%AA%E8%B2%A0%E3%81%AE%E5%BD%B1%E9%9F%BF%EF%BC%88%E6%8E%A5%E9%81%93%E7%BE%A9%E5%8B%99%E9%81%95%E5%8F%8D%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%EF%BC%89)。もしこの欠損を単にデータから削除したり平均値で埋めたりすると、**「道路無しによる価格下落」**という重要な情報がモデルに伝わらず、高値と予測してしまう恐れがあります。同様に、建物関連欠損も土地のみ物件を示し、欠損時に価格が上振れするケース（更地で価値があるなど）が観察されています[[16]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E3%81%93%E3%82%8C%E3%82%89%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AE%E5%AD%98%E5%9C%A8%E3%81%97%E3%81%AA%E3%81%84%E5%9C%9F%E5%9C%B0%E3%80%8D%E3%82%92%E7%A4%BA%E3%81%99MNAR%EF%BC%88Not%20At%20Random%EF%BC%89%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)。これを無視すると、土地のみ物件を過小評価して予測する可能性があります。

**軽減手法**: 以上を踏まえ、以下の手法で欠損による精度低下を抑制・補正します。

* **適切な補完**: 前述の戦略通り、欠損の発生メカニズムに応じて最適な値で補完します。単なる平均/中央値ではなく、**意味の通る値**（0やUnknownカテゴリなど）で埋めることで、モデルが誤った情報を学習しないようにします。例えば延床面積が欠損なら0を入れることで、NNは「この物件は0㎡＝建物なし」と認識できますし、LightGBMは0を極端な値として扱えます。中央値埋めでは建物無し物件も平均的な広さがあるかのように見えてしまうため、意図的に外れ値を入れるのがポイントです[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)。
* **欠損フラグの利用**: 欠損に関連する追加フラグは、その欠損が持つ情報を直接モデルに渡します[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)[[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29)。例えばHasRoadAccess=0なら道路無し物件であり、モデルは大幅な減額要因と即座に認識できます。欠損フラグを用いることで、LightGBMでもNNでも欠損パターンによる価格変動を捉えやすくなり、精度向上に寄与します。特にNNでは欠損をそのまま表現できないため、フラグなしでは補完値0を「本当に数値が0」と誤解する可能性があります。フラグがあれば「これは本来欠損」という情報を追加で学習できます。
* **モデルの組み合わせ**: LightGBMは欠損処理能力が高く、欠損部分で独立したルールを構築できます。一方NNは補完値とフラグから連続的に影響度を学習します。両モデルをアンサンブルすることで、欠損に起因する予測誤差を**お互いに補完**できます[[55]](https://leonidasgorgo.medium.com/in-what-scenarios-would-you-prefer-using-tree-based-models-over-neural-networks-837dfd56aaa9#:~:text=In%20What%20Scenarios%20Would%20You,reducing%20the%20need%20for)。例えば道路無しの物件価格をLightGBMはうまく捉えてもNNが過小評価していれば、アンサンブル平均で補正されますし、その逆もしかりです。多様なモデルで欠損パターンを学習することは、ロバストな予測につながります。
* **バリデーションによる検証・調整**: 欠損値処理戦略が有効かどうかは、クロスバリデーションで確かめます[[56]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E8%A3%9C%E5%AE%8C%E6%96%B9%E6%B3%95%E3%81%AE%E6%A4%9C%E8%A8%BC%20,Type%E5%88%A5%E8%A3%9C%E5%AE%8C%20vs%20%E5%85%A8%E4%BD%93%E8%A3%9C%E5%AE%8C%E3%81%AE%E6%AF%94%E8%BC%83)。例えば欠損フラグを入れた場合と入れない場合でCVスコアを比較し、有意に改善するか確認します。もし効果が薄ければフラグを削減してモデルの自由度を増やす判断もあります。同様に、異なる補完方法（例えばFrontageを単純中央値 vs. 簡易予測モデルで推定補完）でスコア差を検証します[[56]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E8%A3%9C%E5%AE%8C%E6%96%B9%E6%B3%95%E3%81%AE%E6%A4%9C%E8%A8%BC%20,Type%E5%88%A5%E8%A3%9C%E5%AE%8C%20vs%20%E5%85%A8%E4%BD%93%E8%A3%9C%E5%AE%8C%E3%81%AE%E6%AF%94%E8%BC%83)。こうした検証により、**欠損処理による情報損失や過学習を監視**し最適化します。また、ホールドアウト検証で実際の未知データ（テスト相当）に対する頑健性も評価し、特定の欠損パターンに過適合していないか確認します。
* **リーケージ防止**: 欠損補完や特徴量生成の段階で、テストデータ情報やターゲット情報が漏れてはいけません[[46]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20,Type%E5%88%A5%E3%81%AE%E8%A3%9C%E5%AE%8C%E6%88%A6%E7%95%A5%E3%81%8C%E6%9C%80%E3%82%82%E5%8A%B9%E6%9E%9C%E7%9A%84%E3%81%AA%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84)。例えばターゲットエンコードや種別別中央値補完は**訓練データ内**で完結させ、テストには学習結果だけ適用します。特にターゲットエンコードは適切に行わないと欠損以上に大きなリーケージとなり、過大評価されたスコアに惑わされる恐れがあります。これもCVで分割単位を工夫（時間系列に依存するなら年代でブロック分割など）して対策します。

総じて、欠損値は単なるノイズではなく**有益な情報**である場合が多いので[[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF)、それを如何に損なわずモデルに渡すかが鍵です。本戦略では補完による情報歪みを最小限にしつつ、欠損自体を特徴として活用することで、欠損値のネガティブな影響を打ち消しプラスの情報源に変えることを目指しています。

[[1]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20MCAR%20%28%E5%AE%8C%E5%85%A8%E3%81%AB%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E8%A9%B2%E5%BD%93%E3%81%AA%E3%81%97%20,%E3%81%99%E3%81%B9%E3%81%A6%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AB%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B3%E3%81%8C%E3%81%82%E3%82%8B) [[2]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20MAR%20%28%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E4%BB%96%E3%81%AE%E8%A6%B3%E6%B8%AC%E5%A4%89%E6%95%B0%E3%81%AB%E4%BE%9D%E5%AD%98%E3%81%99%E3%82%8B%E3%81%8C%E3%80%81%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%AF%E6%9D%A1%E4%BB%B6%E4%BB%98%E3%81%8D%E3%81%A7%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%3A%20,%2A%2ARemarks%2A%2A%3A%20%E7%89%B9%E8%A8%98%E4%BA%8B%E9%A0%85%E3%81%8C%E3%81%AA%E3%81%84%E9%80%9A%E5%B8%B8%E7%89%A9%E4%BB%B6) [[3]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20MNAR%20%28%E9%9D%9E%E3%83%A9%E3%83%B3%E3%83%80%E3%83%A0%E3%81%AA%E6%AC%A0%E6%90%8D%29%20%E6%AC%A0%E6%90%8D%E8%87%AA%E4%BD%93%E3%81%8C%E6%83%85%E5%A0%B1%E3%82%92%E6%8C%81%E3%81%A4%E6%A7%8B%E9%80%A0%E7%9A%84%E6%AC%A0%E6%90%8D%3A%20,%2A%2ARenovation%2A%2A%3A%20%E6%96%B0%E7%AF%89%E3%81%BE%E3%81%9F%E3%81%AF%E5%9C%9F%E5%9C%B0%E3%81%AE%E3%81%BF) [[4]](file://file_0000000032cc6206941467b192be425b#:~:text=,col%5D.transform%28%20lambda%20x%3A%20x.fillna%28x.median%28%29%29) [[5]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A52%3A%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96%EF%BC%88MNAR%E5%A4%89%E6%95%B0%E3%81%AB%E6%8E%A8%E5%A5%A8%EF%BC%89%20%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AA%E3%81%97%E3%80%8D%E3%82%92%E6%84%8F%E5%91%B3%E3%81%99%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84%3A%20,Breadth%27%5D.notnull%28%29.astype%28int%29) [[6]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%88%A6%E7%95%A51%3A%20%E5%89%8A%E9%99%A4%20,%E6%AC%A0%E6%90%8D%E3%80%81%E4%BA%88%E6%B8%AC%E3%81%B8%E3%81%AE%E5%AF%84%E4%B8%8E%E4%B8%8D%E6%98%8E%20%E2%86%92%20%E5%89%8A%E9%99%A4%E6%A4%9C%E8%A8%8E) [[7]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E5%82%99%E8%80%83%E3%81%8C%E3%81%82%E3%82%8B%E7%89%A9%E4%BB%B6%E3%81%AF%E4%BD%95%E3%82%89%E3%81%8B%E3%81%AE%E5%95%8F%E9%A1%8C%E3%81%8C%E3%81%82%E3%82%8B%E5%8F%AF%E8%83%BD%E6%80%A7) [[10]](file://file_0000000032cc6206941467b192be425b#:~:text=%E6%AC%A0%E6%90%8D%E7%8E%87%E3%81%AE%E5%B7%AE%E7%95%B0%EF%BC%88Train%20vs%20Test%EF%BC%89%20%E3%81%99%E3%81%B9%E3%81%A6%E3%81%AE%E3%82%AB%E3%83%A9%E3%83%A0%E3%81%A7%E6%AC%A0%E6%90%8D%E7%8E%87%E3%81%AE%E5%B7%AE%E3%81%AF%C2%B10.15) [[11]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20TotalFloorArea%20%7C%2062.19,) [[12]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E9%87%8D%E8%A6%81%E3%81%AA%E7%89%B9%E5%BE%B4%2A%2A%3A%20,%E7%A8%8B%E5%BA%A6%EF%BC%89) [[13]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,MinTime%E3%81%8B%E3%82%89%E6%8E%A8%E5%AE%9A%20or%20%E5%89%8A%E9%99%A4) [[14]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,MinTime%E3%81%8B%E3%82%89%E6%8E%A8%E5%AE%9A%20or%20%E5%89%8A%E9%99%A4) [[15]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%E3%81%AETradePrice%E5%B9%B3%E5%9D%87%3A%20%C2%A547%2C753%2C803) [[16]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E3%81%93%E3%82%8C%E3%82%89%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E3%80%8C%E5%BB%BA%E7%89%A9%E3%81%AE%E5%AD%98%E5%9C%A8%E3%81%97%E3%81%AA%E3%81%84%E5%9C%9F%E5%9C%B0%E3%80%8D%E3%82%92%E7%A4%BA%E3%81%99MNAR%EF%BC%88Not%20At%20Random%EF%BC%89%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84) [[17]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4) [[18]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%BA%E7%89%A9%E9%96%A2%E9%80%A3%E3%81%AF0%E5%9F%8B%E3%82%81%EF%BC%88MNAR%E5%A4%89%E6%95%B0%EF%BC%89%20df,BuildingYear%27%5D.fillna%280%29) [[19]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4) [[20]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1) [[21]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E7%AF%89%E5%B9%B4%E6%95%B0%E3%81%AE%E8%A8%88%E7%AE%97%20%60%60%60python%20BuildingYear%E3%81%8B%E3%82%89%E7%AF%89%E5%B9%B4%E6%95%B0%E3%82%92%E8%A8%88%E7%AE%97%20df,1) [[22]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4) [[23]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2AStructure%EF%BC%88%E5%BB%BA%E7%89%A9%E6%A7%8B%E9%80%A0%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,B%EF%BC%88%E3%83%96%E3%83%AD%E3%83%83%E3%82%AF%E9%80%A0%EF%BC%89%3A%201.1) [[25]](file://file_0000000032cc6206941467b192be425b#:~:text=,or%20%E5%89%8A%E9%99%A4) [[26]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2ALandShape%EF%BC%88%E5%9C%9F%E5%9C%B0%E5%BD%A2%E7%8A%B6%EF%BC%89%E3%81%AE%E5%86%85%E8%A8%B3%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,%E3%81%9D%E3%81%AE%E4%BB%96%3A%204.4) [[27]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,or%20%E5%89%8A%E9%99%A4) [[28]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%2A%2AFrontage%E9%9D%9E%E6%AC%A0%E6%90%8D%E6%99%82%2A%2A%3A%20%E5%B9%B3%E5%9D%87%E4%BE%A1%E6%A0%BC%C2%A552%2C212%2C320%EF%BC%88%2B88%25%E9%AB%98%E3%81%84%EF%BC%89%E2%86%92%20%E6%95%B4%E5%BD%A2%E5%9C%B0%E3%83%BB%E5%84%AA%E8%89%AF%E7%89%A9%E4%BB%B6) [[29]](file://file_0000000032cc6206941467b192be425b#:~:text=,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4) [[30]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,0%E5%9F%8B%E3%82%81%20or%20%E4%B8%AD%E5%A4%AE%E5%80%A4) [[31]](file://file_0000000032cc6206941467b192be425b#:~:text=df,Breadth%27%5D.notnull%28%29.astype%28int%29) [[32]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2A%E2%9A%A0%EF%B8%8F%20%E9%87%8D%E8%A6%81%E3%81%AA%E7%99%BA%E8%A6%8B%2A%2A%3A%20,%E9%81%93%E8%B7%AF%E6%83%85%E5%A0%B1%E3%81%AE%E6%AC%A0%E6%90%8D%E3%81%AF%E4%BE%A1%E6%A0%BC%E3%81%AB%E5%A4%A7%E3%81%8D%E3%81%AA%E8%B2%A0%E3%81%AE%E5%BD%B1%E9%9F%BF%EF%BC%88%E6%8E%A5%E9%81%93%E7%BE%A9%E5%8B%99%E9%81%95%E5%8F%8D%E3%81%AE%E5%8F%AF%E8%83%BD%E6%80%A7%EF%BC%89) [[33]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A) [[34]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E4%B8%AD%E5%A4%AE%E5%80%A4%20or%20Area%E5%88%A5%E6%8E%A8%E5%AE%9A) [[35]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2ARemarks%2A%2A%3A%20%E7%89%B9%E8%A8%98%E4%BA%8B%E9%A0%85%E3%81%8C%E3%81%AA%E3%81%84%E9%80%9A%E5%B8%B8%E7%89%A9%E4%BB%B6) [[36]](file://file_0000000032cc6206941467b192be425b#:~:text=,%E5%89%8A%E9%99%A4%20or%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96) [[37]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C%20,%E5%89%8A%E9%99%A4%20or%20%E3%83%95%E3%83%A9%E3%82%B0%E5%8C%96) [[38]](file://file_0000000032cc6206941467b192be425b#:~:text=,Target%20Encoding%E6%8E%A8%E5%A5%A8) [[39]](file://file_0000000032cc6206941467b192be425b#:~:text=,Label%20Encoding%E5%8F%AF) [[40]](file://file_0000000032cc6206941467b192be425b#:~:text=%2A%2AFloorPlan%EF%BC%88%E9%96%93%E5%8F%96%E3%82%8A%EF%BC%89%E3%81%AE%E4%B8%BB%E3%81%AA%E5%80%A4%2A%2A%EF%BC%88%E6%AC%A0%E6%90%8D%E3%81%AA%E3%81%97%E3%81%AE%E5%A0%B4%E5%90%88%EF%BC%89%3A%20,%E3%83%A6%E3%83%8B%E3%83%BC%E3%82%AF%E5%80%A4%E6%95%B0%3A%2086) [[41]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding%E5%8F%AF) [[42]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20or%20Target%20Encoding) [[43]](file://file_0000000032cc6206941467b192be425b#:~:text=%7C,Hot%20Encoding) [[44]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding) [[45]](file://file_0000000032cc6206941467b192be425b#:~:text=,Hot%20Encoding%E5%8F%AF) [[46]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20,Type%E5%88%A5%E3%81%AE%E8%A3%9C%E5%AE%8C%E6%88%A6%E7%95%A5%E3%81%8C%E6%9C%80%E3%82%82%E5%8A%B9%E6%9E%9C%E7%9A%84%E3%81%AA%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84) [[47]](file://file_0000000032cc6206941467b192be425b#:~:text=%E9%AB%98%E3%82%AB%E3%83%BC%E3%83%87%E3%82%A3%E3%83%8A%E3%83%AA%E3%83%86%E3%82%A3%E5%A4%89%E6%95%B0%EF%BC%88%E3%83%A6%E3%83%8B%E3%83%BC%E3%82%AF%E5%80%A4%E3%81%8C%E5%A4%9A%E3%81%84%EF%BC%89%20,Hot%20Encoding%E5%8F%AF) [[48]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20%E6%AC%A0%E6%90%8D%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B3%E3%81%8B%E3%82%89%E3%81%AE%E7%89%B9%E5%BE%B4%E9%87%8F%20,building_cols%5D.isnull%28%29.sum%28axis%3D1) [[49]](file://file_0000000032cc6206941467b192be425b#:~:text=df) [[50]](file://file_0000000032cc6206941467b192be425b#:~:text=3.%20%E9%9D%A2%E7%A9%8D%E9%96%A2%E9%80%A3%E3%81%AE%E7%89%B9%E5%BE%B4%E9%87%8F%20%60%60%60python%20%E5%BB%B6%E5%BA%8A%E9%9D%A2%E7%A9%8D%E7%8E%87%EF%BC%88%E5%BB%BA%E7%89%A9%E3%81%AE%E9%9A%8E%E6%95%B0%E3%81%AE%E6%8C%87%E6%A8%99%EF%BC%89%20df,FloorAreaRatioActual%27%5D.fillna%280%29) [[51]](file://file_0000000032cc6206941467b192be425b#:~:text=%E5%BB%B6%E5%BA%8A%E9%9D%A2%E7%A9%8D%E7%8E%87%EF%BC%88%E5%BB%BA%E7%89%A9%E3%81%AE%E9%9A%8E%E6%95%B0%E3%81%AE%E6%8C%87%E6%A8%99%EF%BC%89%20df,FloorAreaRatioActual%27%5D.fillna%280%29) [[52]](file://file_0000000032cc6206941467b192be425b#:~:text=TradePrice%EF%BC%88%E5%8F%96%E5%BC%95%E4%BE%A1%E6%A0%BC%EF%BC%89%20,%2A%2A%E5%AF%BE%E6%95%B0%E5%A4%89%E6%8F%9B%E5%BE%8C%E3%81%AE%E6%AD%AA%E5%BA%A6%2A%2A%3A%200.36%EF%BC%88%E3%81%BB%E3%81%BC%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83%EF%BC%89) [[53]](file://file_0000000032cc6206941467b192be425b#:~:text=,%2A%2A%E5%AF%BE%E6%95%B0%E5%A4%89%E6%8F%9B%E5%BE%8C%E3%81%AE%E6%AD%AA%E5%BA%A6%2A%2A%3A%200.36%EF%BC%88%E3%81%BB%E3%81%BC%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83%EF%BC%89) [[54]](file://file_0000000032cc6206941467b192be425b#:~:text=1.%20,Type%E5%88%A5%E3%81%AE%E8%A3%9C%E5%AE%8C%E6%88%A6%E7%95%A5%E3%81%8C%E6%9C%80%E3%82%82%E5%8A%B9%E6%9E%9C%E7%9A%84%E3%81%AA%E5%8F%AF%E8%83%BD%E6%80%A7%E3%81%8C%E9%AB%98%E3%81%84) [[56]](file://file_0000000032cc6206941467b192be425b#:~:text=2.%20%E8%A3%9C%E5%AE%8C%E6%96%B9%E6%B3%95%E3%81%AE%E6%A4%9C%E8%A8%BC%20,Type%E5%88%A5%E8%A3%9C%E5%AE%8C%20vs%20%E5%85%A8%E4%BD%93%E8%A3%9C%E5%AE%8C%E3%81%AE%E6%AF%94%E8%BC%83) dataset\_description\_for\_missing\_values.md

<file://file_0000000032cc6206941467b192be425b>

[[8]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=) [[24]](https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60#:~:text=%E3%82%AB%E3%83%A9%E3%83%A0%E5%90%8D%E3%81%AE%E3%81%A8%E3%81%8A%E3%82%8A%E3%80%81%E5%B0%86%E6%9D%A5%E7%9A%84%E3%81%AA%E5%88%A9%E7%94%A8%E4%BA%88%E5%AE%9A%E7%9B%AE%E7%9A%84%E3%81%8C%E8%A1%A8%E3%81%95%E3%82%8C%E3%81%A6%E3%81%84%E3%81%9F%E3%80%82%20%E6%9C%AA%E5%AE%9A) pythonによるデータ分析\_後編(不動産取引価格予測モデル\_モデル構築・分析・評価) #Python - Qiita

<https://qiita.com/teru-saitou/items/5292cd3f8c7ebe0f7c60>

[[9]](https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm#:~:text=Firstly%2C%20the%20default%20is%20in,the%20predictive%20value%20of%20missingness) machine learning - Can missing data imputations outperform default handling for LightGBM? - Cross Validated

<https://stats.stackexchange.com/questions/605474/can-missing-data-imputations-outperform-default-handling-for-lightgbm>

[[55]](https://leonidasgorgo.medium.com/in-what-scenarios-would-you-prefer-using-tree-based-models-over-neural-networks-837dfd56aaa9#:~:text=In%20What%20Scenarios%20Would%20You,reducing%20the%20need%20for) In What Scenarios Would You Prefer Using Tree-Based Models ...

<https://leonidasgorgo.medium.com/in-what-scenarios-would-you-prefer-using-tree-based-models-over-neural-networks-837dfd56aaa9>