1. コンペ成績

1. 入賞したコンペティション名

中古マンション価格予測2024夏の部

2. 最終スコア

MAE: 0.063858

3. 最終順位

1位

4. ユーザー名

yuto

2. 入賞者のバックグラウンド

1. 学歴/職歴

体調不良により大学を中退しました。その後、療養生活を経て遺跡発掘作業員として勤務し、現在はデータサイエンティストになるために勉強中です。

(現在はフルタイム勤務をコンスタントにこなせる体調です)

2. 上位入賞に役立った経験・知識

過去の中古マンション価格予測を含めた様々なコンペでのソリューションが役に立ちました。

また、それらのソリューションの基礎となる知識は参考書やudemyで学びました。

3. 参加したきっかけ・理由

AI・機械学習エンジニアの職種に転職するために参加しました。

4. コンペに費やした時間

2週間程費やしました。平均すると1日あたり2.5時間程費やしています。

3. ソリューション概要

1. 使用した実行環境・言語・ライブラリ・サービスなど

ローカル環境:

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:

2. 使用したモデル構築手法（LightGBM, CNN, etc）

LightGBM(目的変数:単位面積あたりの取引価格\_log):損失関数 l1

LightGBM(目的変数:取引価格(総額)\_log):損失関数 l1

Catboost(目的変数:単位面積あたりの取引価格\_log):損失関数 lq

SARIMAX(目的変数:四半期毎の取引価格(総額)\_logの平均):

3. 最も有効だった特徴量・モデル

最も有効だったモデル:Catboost

最も有効だった特徴量:建築年

4. モデルの学習にかかった時間

1回の学習に約2時間かかります。

※パラメータ最適化のアルゴリズムで約40時間のパラメータ最適化を行っています。

1. 特徴量選択・エンジニアリング
   1. 重要な特徴量
      * 10~20の重要な特徴量を示したimportance plot

参考資料の方に記載しました。

* + - ターゲットに対するpartial plots

欠損値があるデータに対してpartial plotsを使用できなかったのでSHAPを用意しました。また、計算コストが高く一部しか取得できていません。

参考資料(SHAP)の方に記載しました。

* + - 上記が難しい場合、重要な特徴量のリスト
  1. 特徴量選択の方法

特徴量を作成した際に短時間で学習できるLightGBMを作成し、精度が悪化する特徴量を使わないようにしました。

* 1. 重要な特徴量の作成方法

主に中古マンションと駅の経度緯度から作成できる特徴量が効果的でした。

作成手順を以下に記載します。

1. 国土地理院国土地理院APIから返ってきた住所「title」とAPIに送信した住所「address」の名寄せを行い取得住所の信頼性が無い or 地区名が欠損値を表す「住所不一致\_Flag」を作成する。
2. 国土地理院APIの情報を住所でメインテーブルに結合
3. 駅データの経度緯度の情報を駅名でメインテーブルに結合
4. 住所不一致\_Flagが1の場合に住所の経度緯度を駅の経度緯度で補間
5. 地区名の欠損値を「県名」「駅名」「駅からの徒歩時間を10分単位で丸めたもの」をキーにして補完
6. 住所の最頻値で駅名を補間
7. 住所経度緯度から一番近い駅で駅名を補間
8. 駅からの距離を算出する
9. Geohashの2~8レベルの作成と各レベルで集約特徴量を複数作成
   1. 使用した外部データ（使用が許可されている場合）

駅データ.jp,国土地理院APIを使用しました。

1. モデル構築手法
   1. 使用したモデル構築手法

上記の下処理~特徴量作成を行ったデータで「単位面積あたりの取引価格\_log」を目的変数にして、LightGBM,Catboostモデルを1つ作成しました。

また、目的変数を「取引価格（総額）\_log」のままでLightGBMモデルを1つ作成しました。

これら3モデルで共通してtrainデータをhold\_out(評価用データの比率:12%)でOptunaによるパラメータ最適化を行い、最も精度が高かったパラメータで8foldの交差検証モデルを作成しました。

最後にtestデータに対する時系列での価格変動を考慮して、四半期毎の「単位面積あたりの取引価格\_log」を予測するSARIMAモデルを作成しています。

* 1. アンサンブルを行った場合、モデルの重み付けの方法

決定木系の3モデルで交差検証をした際のtrainデータに対する予測結果を用いて、目的変数「取引価格（総額）\_log」と3モデルの予測結果のMAEが最も良くなるアンサンブル比率を求めました。(全ての比率を合計すると1になるようにしています)

このアンサンブル比率を求めるにはscipyをベースにしたアルゴリズムを構築しました。

1. その他、興味深い考察・発見
   1. ランキングスコアの推移と、非連続的なスコアアップをもたらした要因、あるいはスコアアップを妨げた要因（苦労した点）

ランキングスコアの推移:ランキングは時間を追うごとに上昇しました。

非連続的なスコアアップをもたらした要因:住所の経度緯度の取得結果が正しいかを判別するアルゴリズムの正確性を上げる事で、これらを元に作成された特徴量が精度向上に対する貢献度を上げていました。

スコアアップを妨げた要因:住所の経度緯度が正しいかを判別するアルゴリズムの正確性が低いと、精度が下がりました。可能な限り住所を名寄せする等の工夫が必要でした。

* 1. 勝因・上位入賞に当たり重要だったポイント
     + 1. 前回の優勝者様が行っていた「中古マンションと駅の経度緯度の取得」「欠損値補間」「geohash」が精度上昇に大きく寄与していました。
       2. 過去の入賞者様が行っていた特徴量エンジニアリングで一部のモデルの精度が上がりました。
       3. 各モデルのパラメータを最適化し、複数モデルのアンサンブル比率を調整するアルゴリズムを使用する事で更に精度を向上することができました。
       4. 四半期毎の全国の取引価格(総額)\_logの平均に対して時系列モデルを使用して2023年第二四半期から2023年第三四半期にかけて何%上昇するかを求めて、予測結果を定数倍する事で精度が上がりました。
  2. 重要だったEDAの結果

trainデータとtestデータの都道府県名の分布がそこまで変化がないことを確かめるEDAが役に立ちました。

この結果を元にtrainデータで交差検証をする際に、各foldで都道府県名の比率が等しくなるようにデータを分割したことも精度に寄与したと思います。

1. シンプルなモデル 精度は下がるが、よりシンプルなモデルについてご教示ください。
   1. 最終的な精度の90〜95％程度を達成する特徴量のリスト（10個未満に限定）
   2. 最も重要なモデル構築手法（アンサンブルせず、1つに限定）
   3. 上記2点を踏まえ、単純化したモデルのスコア
2. モデルの実行時間
   1. 学習にかかる時間

決定木系の3モデルがそれぞれ約2時間かかりました。8fold交差検証をする場合は2時間×8ほどの時間がかかります。

* 1. 推論にかかる時間

Catboostは数分でした。

LightGBMは1時間以上かかりました。

* 1. g.で示した、シンプルなモデルの学習にかかる時間
  2. g.で示した、シンプルなモデルの推論にかかる時間

1. 参考資料 参考にした論文、ウェブサイトなどの外部情報

[Nishika 中古マンション価格予測コンペ 2023夏の部 3位解法 #Python - Qiita](https://qiita.com/9en310/items/1ef8c957fd43b850162b)

[中古マンション価格予測2024春の部 | Nishika](https://competition.nishika.com/competitions/mansion_2024spring/topics/718?utm_source=twitter&utm_medium=social&utm_campaign=topic_share&utm_content=mansion_2024spring)