

分析レポート

目的

- ① 参加したコンペで上位に入るために取り組んだことのまとめ
- ② ①の過程で作成した特徴量やモデルの精度に対する分析・考察

ページの構成

- 1 ページ：分析レポートの目的、ページ構成
- 2 ページ：コンペの詳細、上位に入るためにコンペで取り組んだこと
- 3 ページ：作成した特徴量について
- 4 ページ：モデルの精度に対する分析・考察（その1）
- 5 ページ：モデルの精度に対する分析・考察（その2）
- 6 ページ：モデルの精度に対する分析・考察（その2）
- 7 ページ：モデルの精度に対する分析・考察（その2）
- 8 ページ：モデルの精度に対する分析・考察（その2）
- 9 ページ：モデルの精度に対する分析・考察（その2）のまとめ、結論

参加するコンペについて

- ・【SOTA】マイナビ × SIGNATE Student Cup 2019: 賃貸物件の家賃予測
- ・回帰分析
- ・予測するものは賃料
- ・評価方法 : RMSE

データについて

データはコンペに与えられている学習用、検証用データを使用

学習用データ (train.csv) / 検証用データ (test.csv)

データ数 : 31,470行 / データ数 : 31,262行

0 id 1 **賃料 (目的変数)** 2 所在地 3 アクセス 4 間取り 5 築年数 築年数

6 方角 7 面積 8 所在階 物件自体の階数と物件がある建物の総階数 9 バス・トイレ 10 キッチン

11 放送・通信 通信設備の建てつけ等 12 室内設備 13 駐車場 駐車場の有無等

14 周辺環境 15 建築構造 16 契約期間

(IDと賃料以外は全てobject型)

①コンペで上位に入るために取り組んだこと (大まかな流れ)

データの理解、各説明変数の中身の確認、ドメイン知識を少し深める

↓

- ①目的変数に影響を与えそうなデータを可視化 (ここで外れ値も確認し、入力ミスと考えられた場合は修正)
 - ②①のデータをモデルが学習できる形に変換し特徴量を作成 (前処理・特徴エンジニアリング)
 - ③作成した特徴量を用いてモデル(LightGBMを使用)で交差検証しながらRMSEの変化を確認(RMSEが下がれば良い特徴量として使用)
- ①、②、③を繰り返すことで精度の良いモデルを作っていました

②作成した特徴量についてのまとめ

作成した全ての特徴量

- ・住所のデータから23区のどこに属するのかという区のデータ（ラベルエンコーディング）
- ・23区の平均地価、人口、その区への通勤者、飲食店の数
- ・trainデータの目的変数である賃料の統計量（平均値、標準偏差、最小値、最大値、中央値）を各区ごとに作成
- ・最寄駅までにかかる時間
- ・近くの建物（コンビニ、スーパー、学校）までにかかる時間
- ・賃貸物件の面積
- ・賃貸物件の方角（計8つ）
- ・賃貸物件の部屋の数（リビングは2部屋分と計算）
- ・賃貸物件の築年数
- ・賃貸物件の所在階と最高階
- ・賃貸物件の1部屋あたりの面積
- ・賃貸物件の設備されているインターネット、キッチン、バス関連の個数
- ・建物の建築構造（木造、コンクリートなど）

計23個

最終的に使用した特徴量

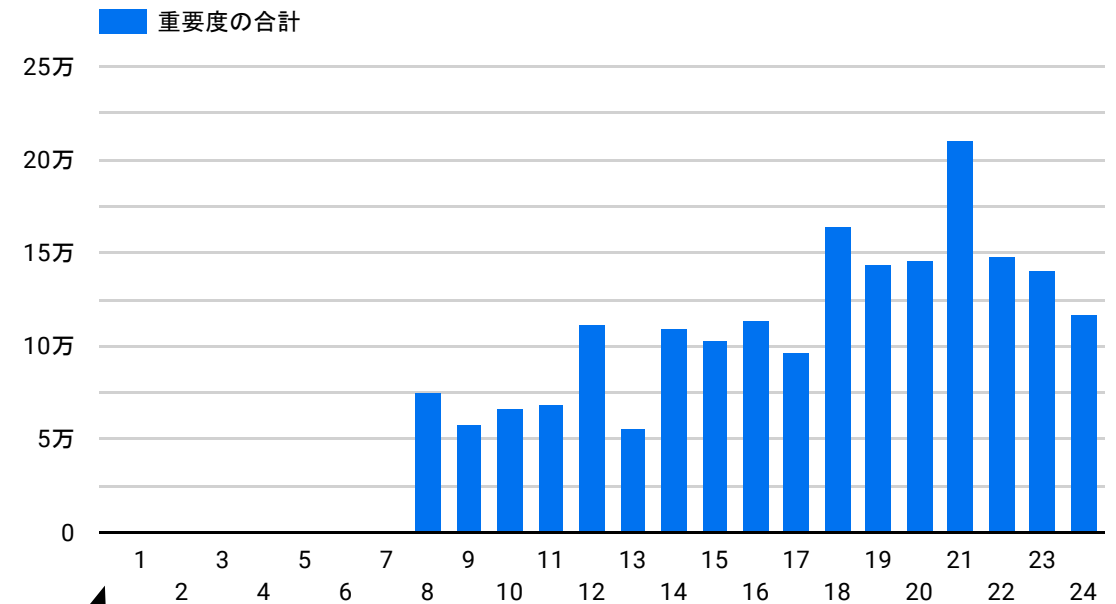
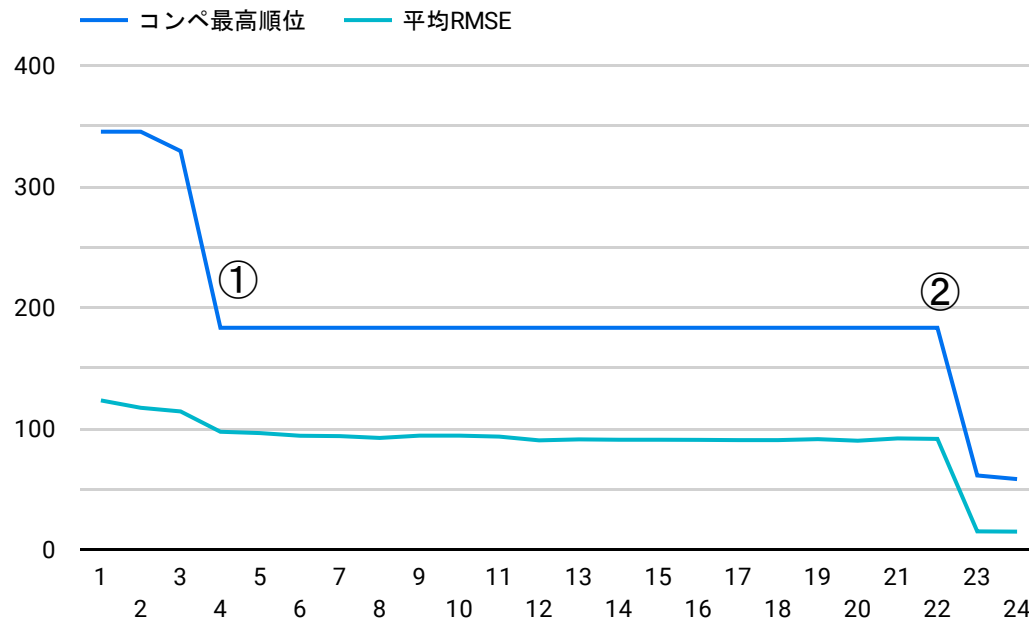
- ・住所のデータから23区のどこに属するのかという区のデータ（ラベルエンコーディング）
- ・23区の平均地価
- ・trainデータの目的変数である賃料の統計量（最小値、中央値）を各区ごとに作成
- ・最寄駅までにかかる時間
- ・賃貸物件の面積
- ・賃貸物件の方角（計8つ）
- ・賃貸物件の部屋の数（リビングは2部屋分と計算）
- ・賃貸物件の築年数
- ・賃貸物件の所在階と最高階
- ・賃貸物件の1部屋あたりの面積
- ・建物の建築構造（木造、コンクリートなど）

計13個

特徴量についての考察

- ・住所や面積、統計量などの各物件ずつでデータが異なり他データと比較可能なデータはモデルの精度向上に大きく貢献したと考えました。
- ・特に23区の住所や、統計量を特徴量として使用したことはコンペで順位を大きく上げるターニングポイントにもなりました。
- ・特徴量はあればあるだけよいのだと考えていた部分もありましたが、モデルの精度を高めよう（RMSEを小さくしよう）とした際に大雑把なデータや他の特徴量と似た傾向にあるデータなどはモデルの精度を下げる原因となっていました。そのため特徴量を作成するときは**①データの表している意味をしっかりと理解する②データの特徴をモデルが学習しやすい形で表現する**。この2つがとても重要だと再確認できました。

②モデルの精度に対する分析・考察（その1）



コンペへの提出回数

モデルの精度が向上したターニングポイント

POINT 1

①のポイントで**各データに23区という情報を特徴量として加えた**ことで順位が345位→183位へを上がり、RMSEが123→97へと下げることができた。23区という住所のデータはモデルの予測に良い特徴量として影響を与えていることがRMSEの変化から理解できる。住所の重要性はおおよそ予測はできたものの、23区という大まかな割り振りだけでここまでRMSEが変化したことは驚きであり、参考になった。

POINT 2

②のポイントで説明変数ではなく目的変数に焦点を当て、**目的変数を1平米あたりの賃料と変更した**ことで順位が183位→58位へを上がり、RMSEが97→14へと下げることができた。

(ただしRMSEが14というのは目的変数を1平米あたりと変更しているため1つ前のRMSEと相対的な評価ができないため良い数値なのかどうかかわらなかった、) 目的変数を小さくすることで大きな数値の目的変数を過小評価せず良いモデルが作れたのだと考える。

その他のポイント・考察

重要度の合計が増加すればRMSEも小さくなる（つまり精度の高いモデルが作れる）と考えていたが、23区ごとの目的変数の統計量を説明変数として加えた③のポイントで重要度の合計が大きく増加させた時のRMSEが一定であることから**重要度の合計とRMSEが相関関係を持っていない**と考えられる。

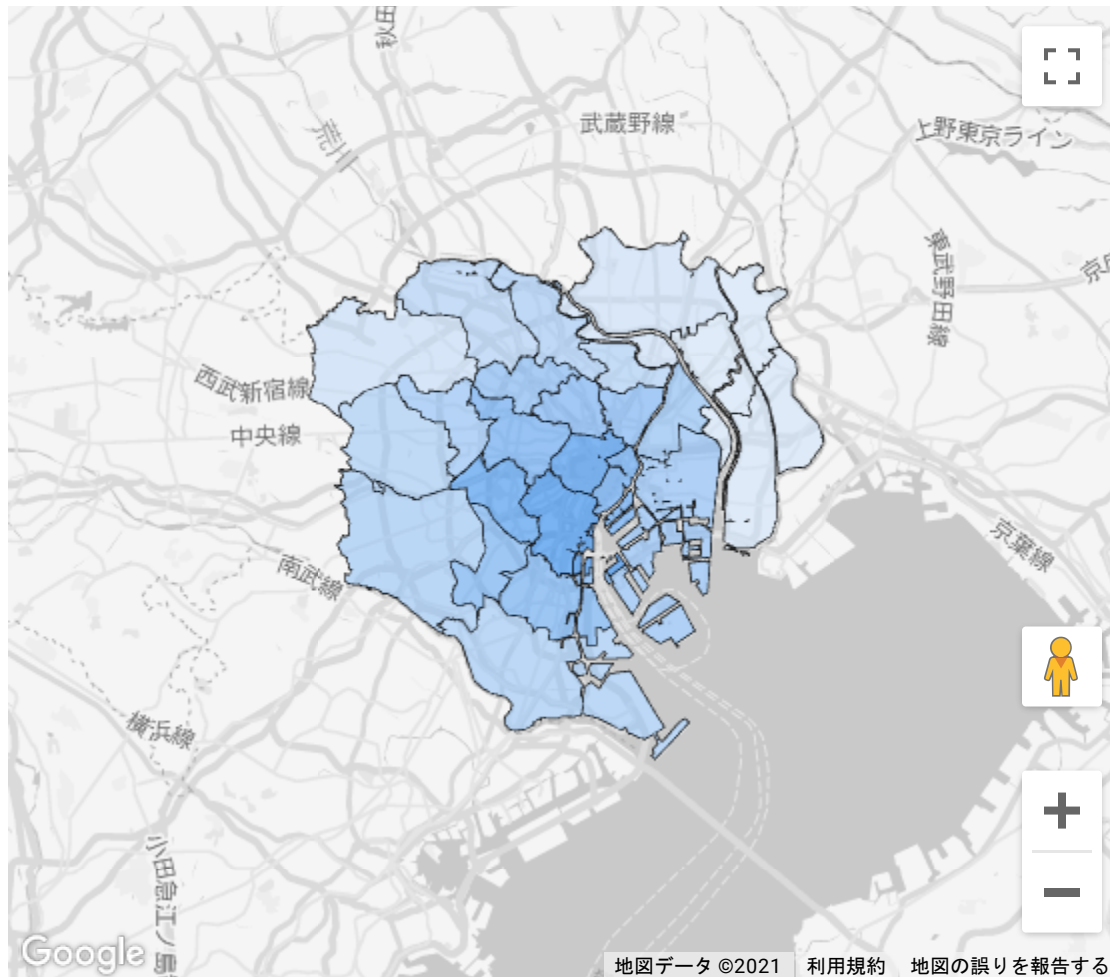
モデルの精度を視覚的に確認するためにtrainデータの実値（目的変数である賃料）、testデータからモデルが予測した賃料の予測値から


- ①23区ごとの1平米あたりの平均賃料（マップ）
- ②23区ごとの1部屋あたりの平均賃料（棒グラフ）
- ③23区ごとの最寄駅からの距離における平均賃料（棒グラフ）
- ④23区ごとの建物の築年数における平均賃料（棒グラフ）

計4つのグラフを作成し実際値と予測値のグラフで比較することでモデルの精度を視覚的に確認する。

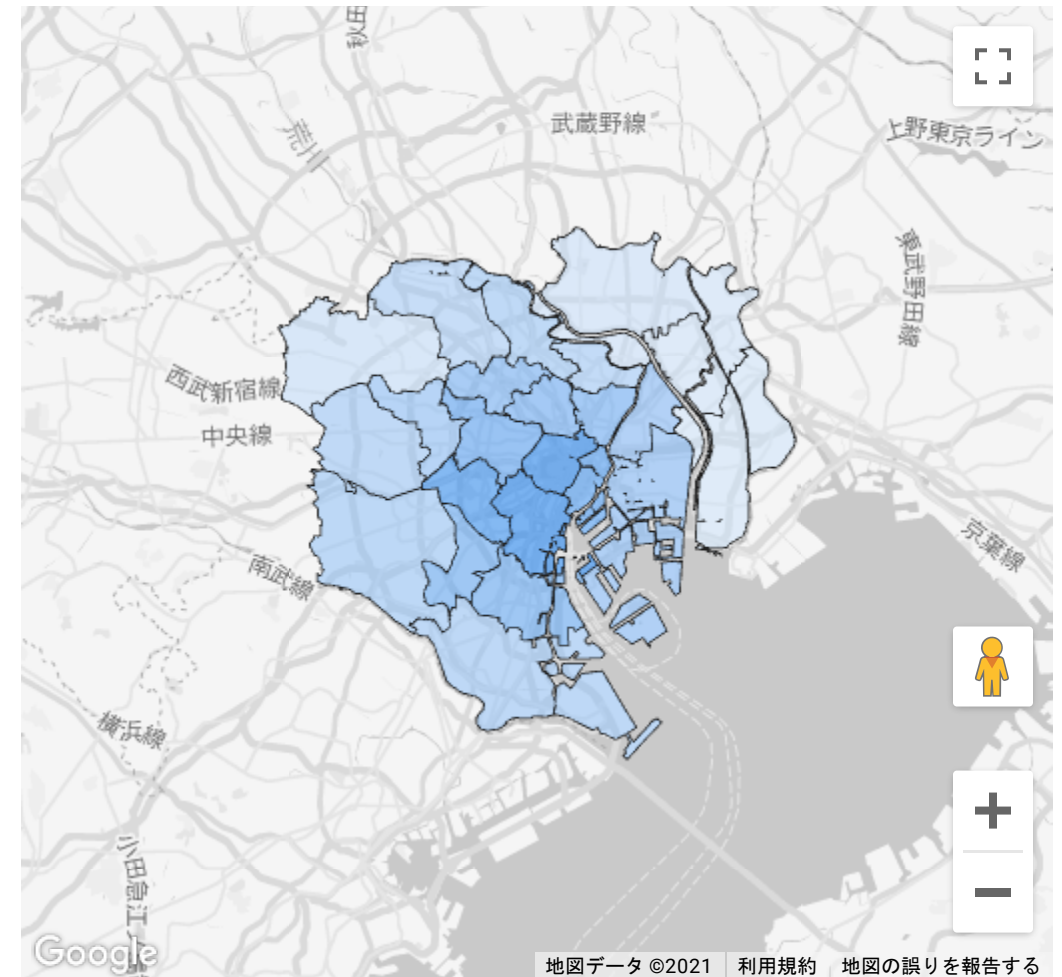
①23区の1平米あたりの平均賃料

実際値(trainデータの賃料)



1平米あたりの賃料 2,259.37  4,959.95

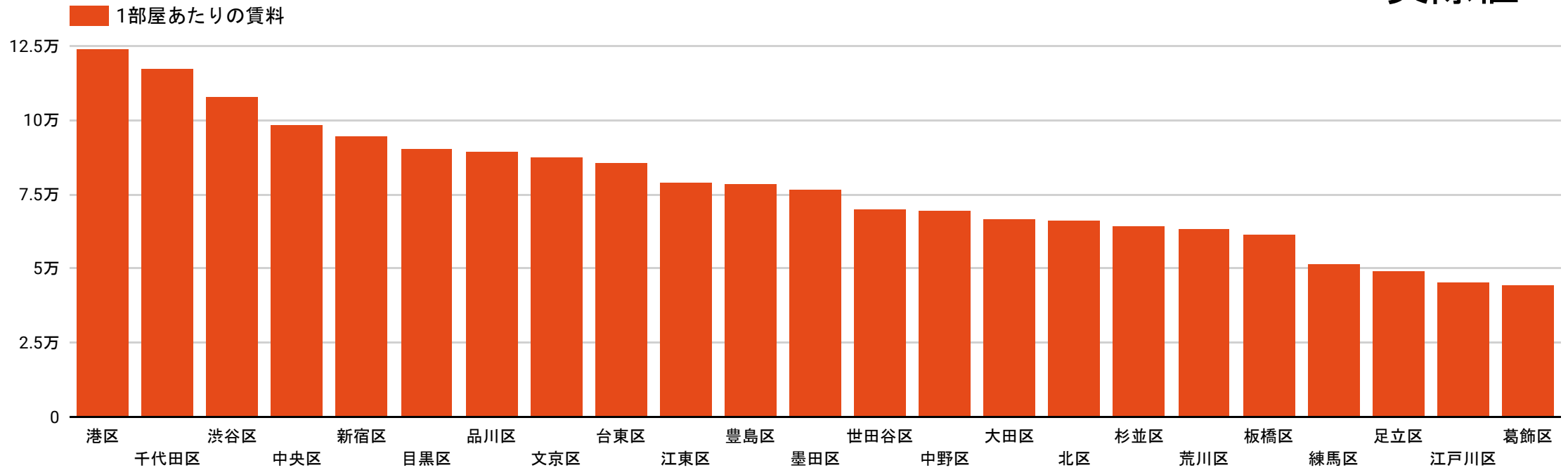
予測値 (testデータから予測した賃料)



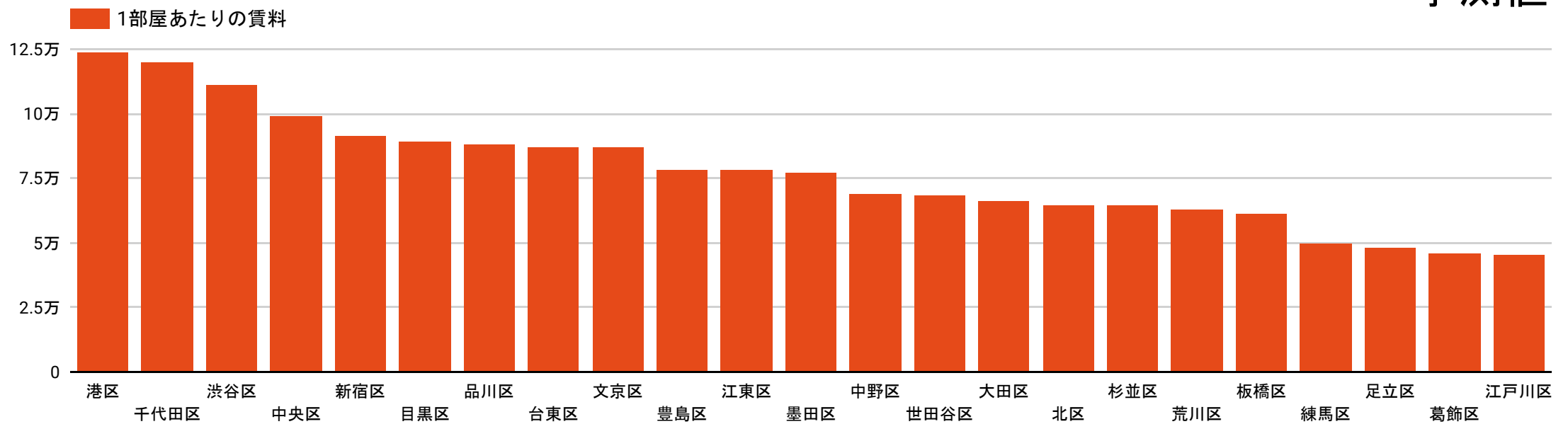
1平米あたりの賃料 2,278.6 4,956.86

②23区の1部屋あたりの平均賃料

実際値

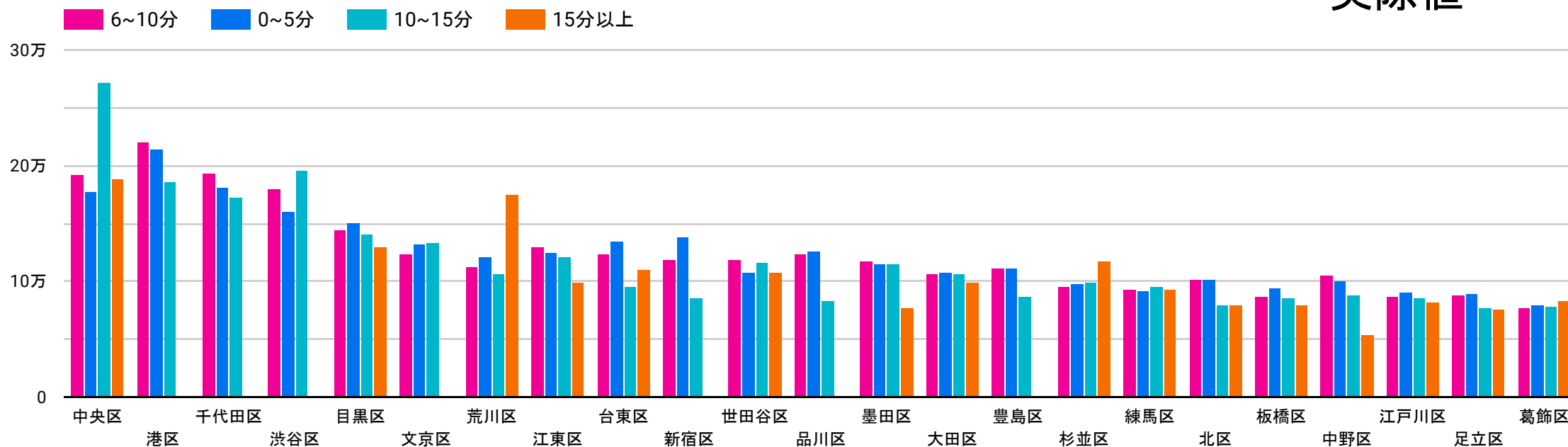


予測値

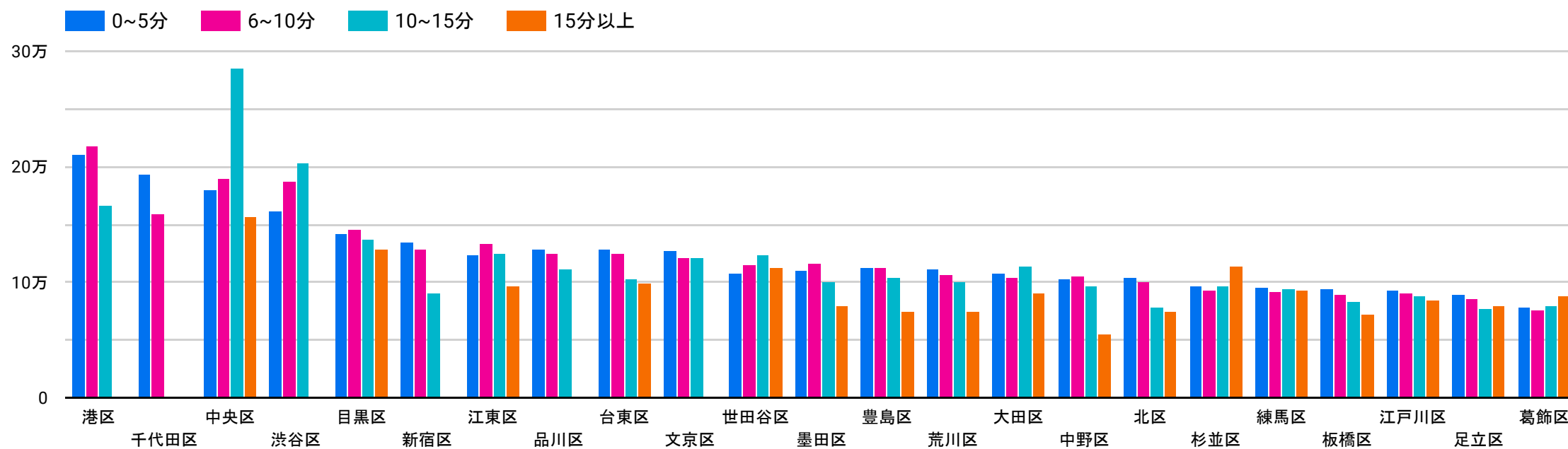


③23区の最寄駅からの距離における平均賃料

実際値

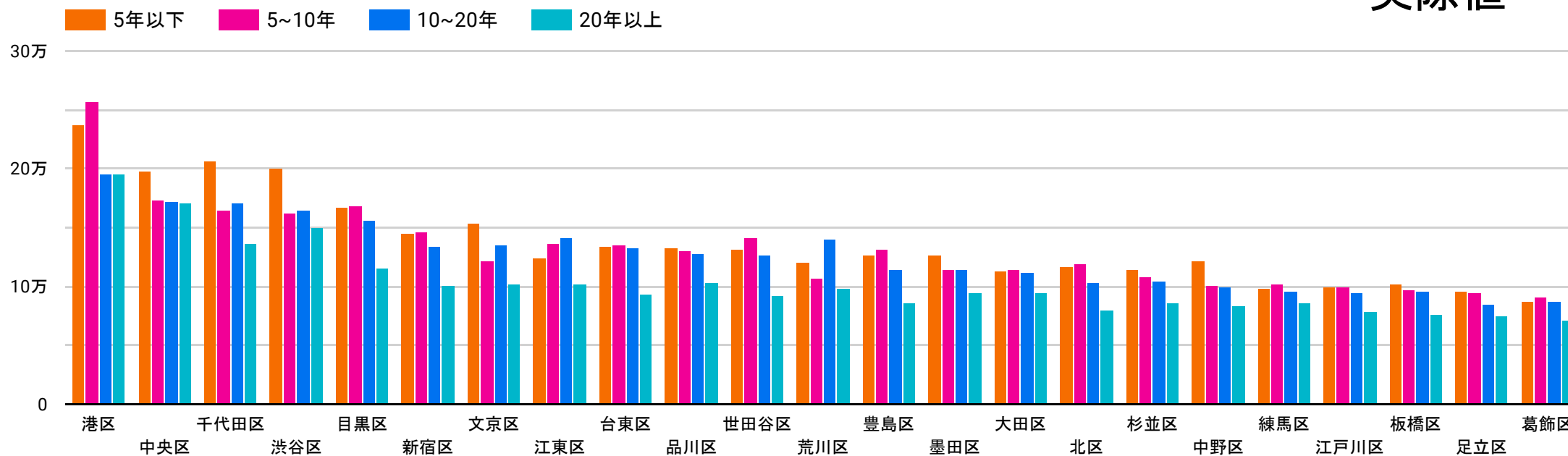


予測値

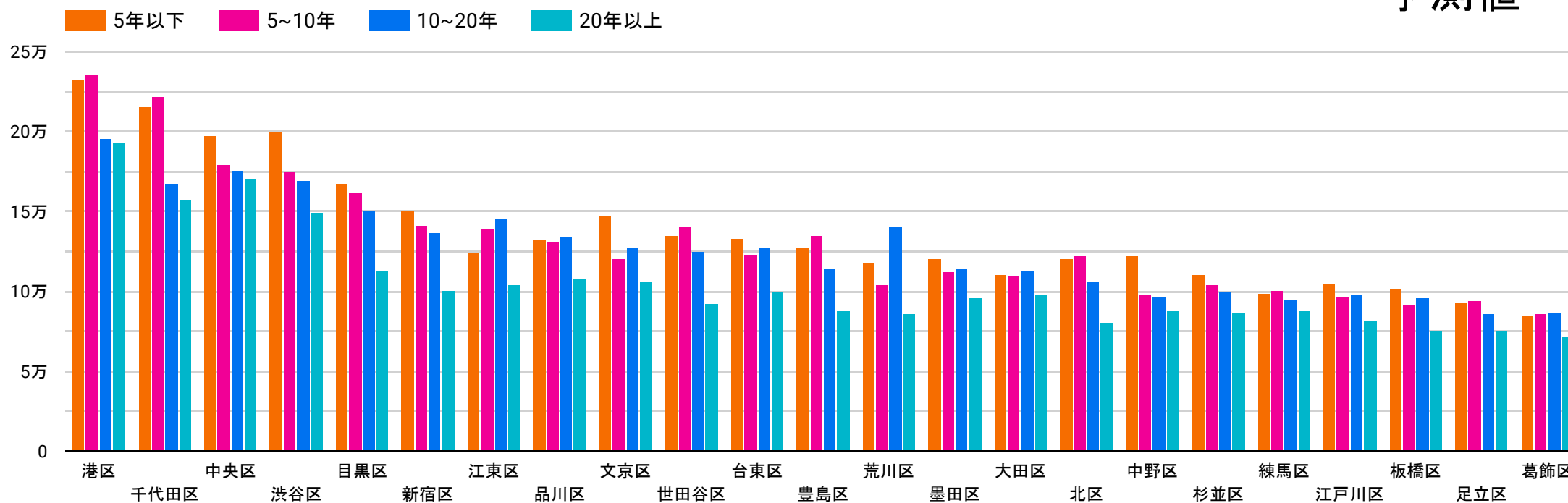


④23区の建物の築年数における平均賃料

実績値



予測値



モデルの精度に対する分析・考察（その２）のまとめ

- ・計４つのグラフからおおよそモデルが実際値と似た予測値を出力していることが確認できる
- ・①の23区の1平米あたりの平均賃料のグラフからは東京の中心に行くほど賃料が増加していることが視覚的に確認できるので、他県からの距離などを特徴量にすることも考えられた
- ・③の23区の最寄駅からの距離における平均賃料のグラフと④の23区の建物の築年数における平均賃料のグラフから、実際に比べて予測値が比較的高いことが確認できる。原因はわからなかったが予測値が平均的に賃料が高い原因がわかるとよりいっそう精度の高いモデルが作れると思われる。

分析レポートのまとめと今後の取り組み

今回はコンペで上位に入るために取り組んだこと、その過程で作成したモデルの精度とコンペの順位の推移、最終的に作成したモデルの予測値の精度などを分析レポートとしてまとめました。

コンペへ本格的に取り組んだのは初めてであり、特徴量作成の際に基礎的に足りていない部分（Pandas,Matplotlib）も学習ながらではありましたが多くのことを学べたと感じております。

そしてコンペで上位約１０％に入ることができたのは思っても見なかったです。

何よりもモデルの精度を上げるのは特徴エンジニアリングの部分であり、やはりそこに一番時間がかかりました。

23区の何区に属するのかという住所の情報を特徴量とした際に大幅にモデルの精度が良くなったが、その他に市、町の情報、または緯度経度の情報を特徴量とする事も考えついたが自分の能力ではDataFrameにうまく取り込むことができなかったので大きな反省点も残りました。