

東京都の家賃予測分析

慶應義塾大学経済学部・久保将真

慶應義塾大学経済学部・黒岩達也

慶應義塾大学経済学部・萬代涼雅

要旨

本研究は、ランダムフォレスト回帰モデルを用いて東京都23区の家賃予測を行い、その予測精度と改善方法を検討することを目的としたものである。2022年と2024年のSUUMOの家賃データを取得し、モデルの構築と評価を行った。初期のモデルでは区ごとにデータを集約したため、地域ごとの家賃変動を反映できず、予測精度が低い結果となった。原因として、地域の多様性やパラメータ設定の問題が考えられたため、データを町単位に細分化し、さらにグリッドサーチによるハイパーパラメータの最適化を試みた。

第1章 序論

第1節 背景

東京23区では都市化の進展により、特定の地域で急激な人口増加やインフラ開発が進行し、地価が大きく変動している。この変動を適切に理解することは今後の居住、都市計画、不動産投資において重要である。一般的な家賃予測モデルは、過去の家賃データの時系列分析に依存することが多く、データが不足している地域や、突然の市場変動に対応しづらい。また時系列データが十分でない場合に適切な予測を立てるのが難しいという課題を抱えている。

第2節 研究目的

本研究では、より高い精度の家賃予測を可能にするため、後述するランダムフォレスト回帰モデルを用いた要因分析を行い、都市計画や不動産投資における意思決定を支援し、地域の持続可能な発展に寄与する。計量経済学の手法を実践的に活用することで、理論と実務の橋渡しの一助となることを目指す。

第2章 分析

第1節 分析手法

1.1使用モデル

ランダムフォレスト回帰モデルを用いて予測を行う。ランダムフォレスト回帰モデルでは、2022年のデータをもとにモデルを構築し、2024年の家賃を予測、実際の家賃との誤差率を求める。重回帰モデルでは、2024年データの8割をテスト用データとしてモデルを構築し、2024年データの2割を評価用データとして平均2乗誤差MSEと決定係数を算出する。

1.2使用データ

2022年のデータとして、2022年8月31日時点のSUUMOに記載されている東京都23区の賃貸物件5870件[1]をスクレイピングして取得する。同様に、2024年のデータとして、2024年10月31日時点の物件189537件[2]を使用する。

データとして、ターゲット変数である家賃のほか、区名、市町村名、築年数、階数、間取り、面積の項目を特徴量として抽出する。間取りに関しては、部屋数以外にL(リビング)、D(ダイニング)、K(キッチン)、S(サービスルーム)が記載されているものは、それぞれ0.5部屋分として計算する。

また、追加する外部指標として、区ごとの災害リスク[3]、人口および人口密度[4]、平均年収[5]、犯罪率[6]をそれぞれ指標化した特徴量を用い、多面的な予測を行う。

まず、各区の賃貸件数は以下の通りである。

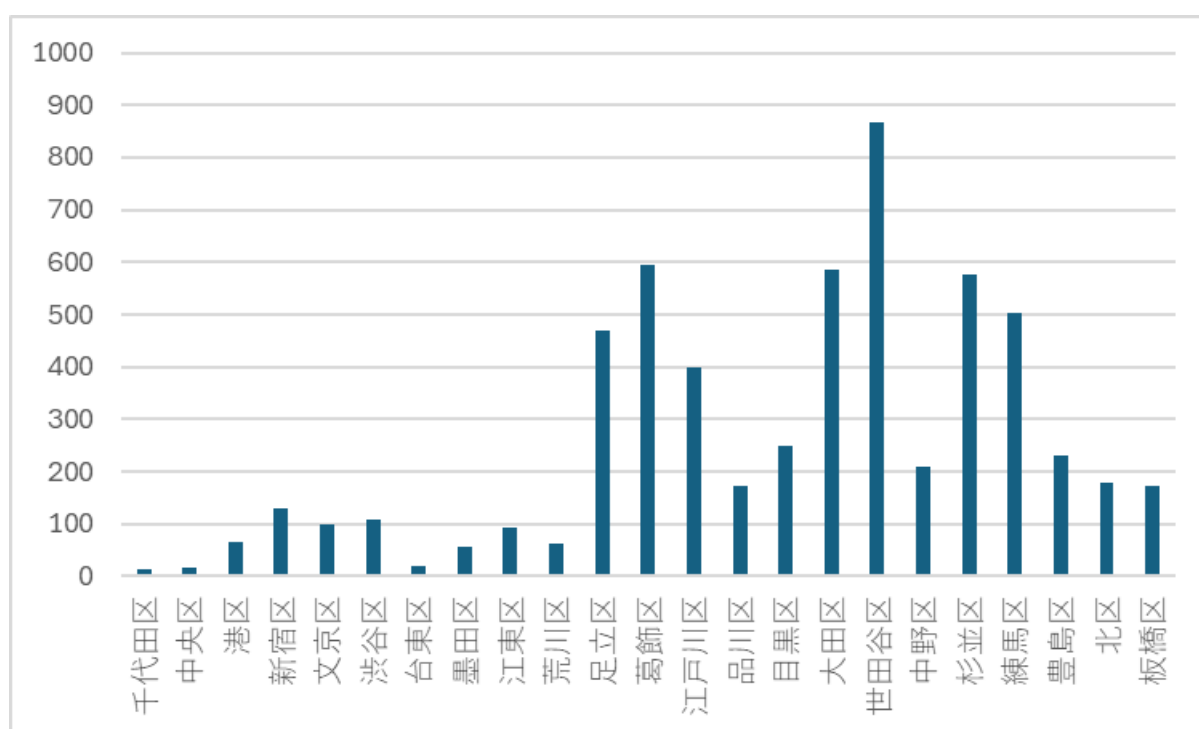


図1 2022年の物件数

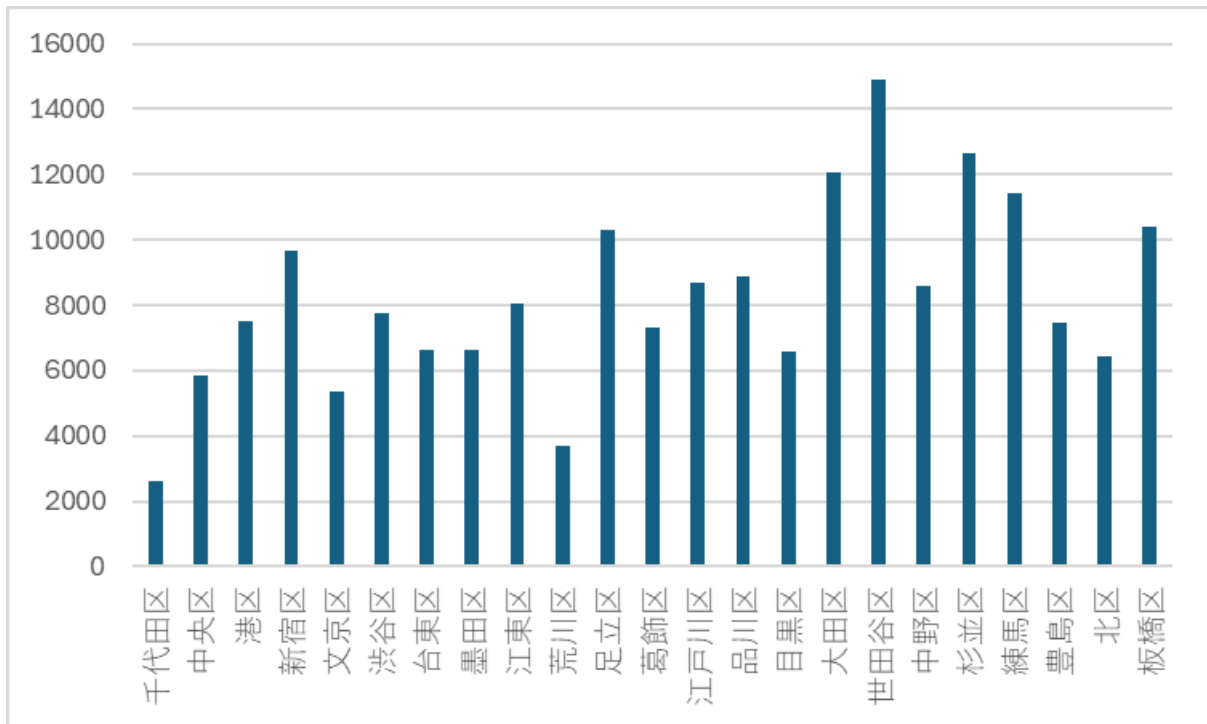


図2 2024年の物件数

第2節 各機械学習手法の概念

2.1 ランダムフォレスト回帰モデル

ランダムフォレスト回帰モデルは、複数の決定木を組み合わせて予測を行うアンサンブル学習の一種である。各決定木は、データの異なる部分サンプルで学習され、最終的な予測はすべての木の予測の平均を取ることで行われる。これにより、過学習を抑制し、予測精度を向上させる。

ランダムフォレスト回帰の基本的な流れは3段階に分けられる。まず、ブートストラップサンプリングを行い、データセットから複数の部分サンプルをランダムに生成する。続いて、各部分サンプルで決定木を構築する。決定木の分岐にはランダムに選ばれた特徴量の中で最適なものを使用する。最後に、新しいデータに対して各決定木で予測を行い、その平均値を最終的な予測値とする。

ランダムフォレストの特徴として、過学習しにくいことが挙げられる。個々の決定木が過学習しやすい場合でも、複数の木を平均化することで予測が安定しやすくなる。また、各決定木は独立して構築されるため、並列処理が可能になる。さらに、近接度マトリクスによって多段階の欠損値推定が行われるため、欠損値が含まれるデータを扱う際に問題が生じないため、従来の手法より予測精度が高く安定したモデルとして採用した。

なお、ランダムフォレスト回帰モデルでの予測は、各決定木の予測の平均を取ることで得られる。データポイント x に対する予測 y は次のように表せる。

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M T_m(x)$$

M:決定木の本数、 $T_m(x)$:m 番目の決定木による予測値
各決定木 $T_m(x)$ は、次のようにデータセットからランダムにサンプリングされたデータで構築される。

2.2 ハイパーパラメータチューニング

ハイパーパラメータチューニングは、機械学習モデルの性能を最適化するための重要なプロセスであり、パラメータを適切に調整することで、より精度の高い予測を実現できる。ランダムフォレストモデルには、以下のような主要なハイパーパラメータが存在し、それぞれがモデルの性能に影響を与える。

- ①n_estimators(決定木の本数)
- ②max_depth(決定木の深さの上限)
- ③min_samples_split(ノード分割のサンプルの最小)
- ④min_samples_leaf(リーフノードに必要なサンプル数の最小)

以上の4つのパラメータを最適化することで最も高いパフォーマンスが生まれるが、最適なパラメータを得るための明確な算定方法は存在しないため、各パラメータに網羅的に数値を当てはめて、そのモデルの精度を確認するほかない。そこで、本研究では、グリッドサーチという手法を導入した。グリッドサーチは、指定したパラメータの組み合わせを全て試行して最適なパラメータセットを見つける方法である。具体的には、各ハイパーパラメータに対して複数の候補値を設定し、それらの全ての組み合わせについてモデルを学習し、評価指標に基づいて最も良い性能を示す組み合わせを選択するというものである。

2.3 特徴量の重要度

ランダムフォレストモデル (RandomForestRegressor) が持つ feature_importances_ 属性を利用して特徴量の重要度を求める。

feature_importances_ は、各特徴量がどの程度モデルの予測に貢献したかを数値化したもので、ランダムフォレストがモデルを構築する際に計算されている。ランダムフォレストでは以下の方法で特徴量の重要度を決定している。

- ①不純度の減少: 各決定木が分岐を行う際に、どの特徴量がデータの分割を行っているかを確認し、その分岐がデータの不純度(MSE)をどれだけ減少させたかを記録する。
- ②全体の重要度の平均: ランダムフォレストでは多数の決定木が構築されるが、それぞれの木で得られた特徴量の重要度を平均して、全体の特徴量の重要度として算出する。

feature_importances_ は各特徴量の重要度を0から1の間で出力する。具体的な重要度の計算は、各特徴量が分岐に寄与した回数と不純度減少量の合計からなり、全体での割合として計算されている。

第3節 分析結果

第1節の1.1.使用モデルおよび1.2使用データを前提として、ランダムフォレスト分析を実行した。
(表3)また、モデルの精度について次のような結果が得られた。(表4)

区名	荒川区	葛飾区	江戸川区	台東区	中野区
平均二乗誤差 (MSE)	349,627,347.2	161,921,853.7	294,858,911.9	314,233,228.8	856,789,279.9
決定係数 (R ²)	0.907618	0.922783	0.870295	0.901482	0.822937
区名	足立区	北区	板橋区	文京区	目黒区
平均二乗誤差 (MSE)	216,974,402	286,506,151.3	260,866,378.1	959,968,839.6	2,851,368,071
決定係数 (R ²)	0.86733	0.898541	0.857492	0.906987	0.814935
区名	豊島区	墨田区	練馬区	中央区	千代田区
平均二乗誤差 (MSE)	927,842,799.7	285,446,416.6	252,363,487.2	1,164,168,453	3,615,145,911
決定係数 (R ²)	0.833362	0.879026	0.865427	0.879737	0.922273
区名	杉並区	渋谷区	品川区	江東区	港区
平均二乗誤差 (MSE)	361,220,378.7	4,575,025,488	1,033,596,576	425,235,758.4	71,667,774,499
決定係数 (R ²)	0.894824	0.867478	0.861027	0.881394	0.461619
区名	新宿区	世田谷区	大田区		
平均二乗誤差 (MSE)	916,918,307.3	1,185,507,297	975,197,031.7		
決定係数 (R ²)	0.889999	0.858275	0.837946		

表3 23区別のランダムフォレスト回帰による推定結果

評価指標	値
平均絶対誤差 (MAE)	52,570.48 円
平均二乗誤差 (MSE)	14,686,216,234.40 円 ²
平方根平均二乗誤差 (RMSE)	121,186.70 円
平均絶対誤差率 (MAPE)	40.78 %
決定係数 (R ²)	-0.19

表4 23区別のランダムフォレスト回帰モデルの精度

第3章 考察と今後の課題

第1節 考察

3.1 モデルの評価

誤差を示す4つの数値が軒並み悪く、特に決定係数に至っては負の値となってしまった。この低精度の要因は以下のように分析することができる。

要因① データ区分の粒度

今回、23区の家賃データを各区ごとに分けて家賃予測を行ったが、同一の区内でも家賃水準が大きく異なることが、予測精度を悪化させていると考えられる。例えば、大田区には田園調布などの高級住宅街が位置する一方、家賃水準の低い下町エリアも広がっていることから、区ごとにデータを細分すると、家賃の傾向が異なる地域どうしを一括りに扱ってしまうという問題に陥っているといえる。

要因② ランダムフォレストモデルのパラメータ設定

決定木の本数を100に設定したが、家賃データの数に比して、モデルが学習するパターンが少ないと考えられる。また決定木の深さの上限を10に設定したが、家賃データの複雑性を鑑みると、より深い設定が必要になると考えられる。

3.2 モデルの改善

前節で述べた予測精度を悪化させた可能性がある要因に対して、次のような処置を施した。

改善① 地域単位の細分化

各地域の家賃の傾向を正確に反映するために、区単位ではなく、町単位でデータを分けて、再度分析を行ったところ、次の結果が得られた。(表6)

評価指標	値
平均絶対誤差 (MAE)	42,117.25 円
平均二乗誤差 (MSE)	8,627,632,397.44 円 ²
平方根平均二乗誤差 (RMSE)	92,885.05 円
平均絶対誤差率 (MAPE)	32.24 %
決定係数 (R ²)	0.30

表6 市区町村別のランダムフォレスト回帰による推定結果

町レベルの細分化によって、家賃の予測精度が大幅に向上した。区レベルの細分化した際と比べて、全指標が改善しており、特に決定係数が0.3 となり、予測モデルとしてはいまだ不十分ではあるが、モデルの説明力がある程度認められる。地域ごとの特性を捉えることの重要性が確

認できたことから、さらに細かいレベルで細分化することも考えられるが、必要以上に細かくしすぎると、分散が大きくなり過学習のリスクが高まるので、別のアプローチにより、モデルの精度を向上させる。

改善② パラメータの最適化

ここまでの分析において、ランダムフォレストのパラメータの設定を①n_estimators=100(決定木の数)②max_depth=10(決定木の深さの上限)としてきたが、前述したとおり、この設定には明確な根拠がなく、改善の余地が残されていることから、グリッドサーチの手法を用いて、パラメータの最適化を行った。これにより、最適なパラメータが、①n_estimators=200(決定木の数)②max_depth=None(決定木の深さの上限)と計算された。

そして、この最適化されたパラメータを適用して、再度分析を実施した。モデルの評価結果は次のとおりである。(表7)

評価指標	値
平均絶対誤差 (MAE)	22,027.00 円
平均二乗誤差 (MSE)	14,912,818,786.36 円 ²
平方根平均二乗誤差 (RMSE)	122,118.05 円
平均絶対誤差率 (MAPE)	7.84 %
決定係数 (R ²)	0.70

表7 パラメータ調整後のランダムフォレスト回帰モデルの精度

決定係数が0.70、MAEが約22,000円、MAPEが7.84%と予測モデルとして現実的な数値が得られた。これは、決定木の数を増やし深さ制限を解除することで、モデルがデータの多様なパターンを学習できるようになったと考えられる。一方で、MSEおよびRMSEの数値が、悪化してしまった。これはいくつかのデータにおいて大きな誤差が発生し、二乗の形を取るMSEとRMSEに強い影響を与えたのだと考えられる。よって、この外れ値への影響への必要性が認められる。ただし、全体の精度に関しては先ほどよりも改善されているといえる。

改善③ 外部データの導入による特徴量の追加

家賃は、面積や立地といった物件そのものに関する特性だけではなく、犯罪率や平均年収などの外部要因にも一定の影響を受けると考えられる。また、複雑な関係を捉えるのが得意であり、多くの特徴量を扱うことが可能であるというランダムフォレスト分析の長所を活かす意味でも、本研究において、特徴量の追加を行った。新たな特徴量として、1平均年収、2犯罪率、3人口、4人口密度、5災害リスク、6最寄り駅からの時間、を追加したところ、次のような結果が得られた。(表8)(表9)

指標	値
平均絶対誤差 (MAE)	17,490.06 円
平均二乗誤差 (MSE)	2,825,607,104.42 円 ²
平方根平均二乗誤差 (RMSE)	53,156.44 円
平均絶対誤差率 (MAPE)	7.62 %
決定係数 (R ²)	0.91

表8 特徴量を追加した場合のランダムフォレスト回帰モデルの精度

決定係数が、0.91という高い値を示しているため、このランダムフォレストモデルが予測モデルとして実行可能な程度の信頼性を有しているといえます。また、誤差に関する数値に関しても、これまでと比べて、軒並み数値が改善されており、特にMAEがRMSEに対してあまり変動していないことから、大きな外れ値の影響が相対的に少ないモデルであると考えられ、先ほどのモデルの欠点が解消されたといえます。

特徴量	重要度
面積	0.701966
推定平均年収	0.067521
住所	0.056113
築年数	0.036963
犯罪率	0.033362
区名	0.026262
階数	0.025128
最寄駅からの時間	0.014913
間取り	0.011306
人口	0.010776
災害リスク	0.010575
人口密度	0.005114

表9 当該ランダムフォレスト回帰モデルの主要な特徴量

<重要な特徴量の分析>

・面積は、最も高い重要度(0.70)を示している。これは、家賃は通常、面積が大きくなるにつれて上昇するという至極当然のことを表しているが、重要度の高さが0.7という大きな値をとったことから、モデルが面積に依存しすぎていないか注意する必要がある。

・推定平均年収が2番目に高い重要度(0.067)となっており、地域の収入水準が家賃に強く影響することがわかる。不動産は通常人生の中で最も高い買い物であり、平均年収と家賃の高低に関連性があると考えられるため、この特徴量が重要であるという結果は、直感に沿ったものだといえる。

<重要度の低い特徴量>

人口密度や災害リスクの重要度は比較的小さい値が得られた。これは、これらの特徴量が家賃に直接的に大きな影響を及ぼす要素ではない可能性を示しており、人口密度は生活の利便性や混雑具合には影響するものの、家賃設定の直接的な要因とはなりにくいことが考えられる。同様に、災害リスクも住む地域の選定に関わるものの、家賃の変動要因としては小さいことが示唆される。

第2節 今後の課題

以下の4点が今後の課題として考えられる。

① 追加の特徴量

考察において、特徴量の追加が精度向上に寄与していることが示された。予測変動率が大きくばらついている地域の精度を向上させるためには、同様に他の決定要因(物件の状態、周辺の治安や施設、人気度など)を新たに特徴量として追加、検討することが必要である。

② 他モデルとの比較

今回はランダムフォレスト回帰モデルのみを扱ったが、同じ特徴量を扱い他モデルと比較することで説明力の明示が可能である。

③ データの拡充

今回は2022年と2024年のデータのみを扱ったが、より多くの年度あるいは最新のデータを収集することで、モデルの汎用性と精度を高めることが期待される。

④ ホットスポット特定

チューニングや特徴量の再検討により、決定係数0.91と高い説明力が示されたため、2024年のデータに対し同じものを使用することで、さらに1年後の家賃予測が可能である。このとき特に大きな上昇が見込まれるホットスポットを特定し地図上に可視化することで、鮮明な市場理解をサポートする。

使用データ出典

[1]株式会社リクルート(2022)「東京都のエリアから探す賃貸一戸建ての賃貸住宅[賃貸マンション・アパート]情報」 2024年10月31日閲覧

https://suumo.jp/jj/chintai/ichiran/FR301FC001/?ar=030&bs=040&ta=13&sc=13101&sc=13102&sc=13103&sc=13104&sc=13105&sc=13113&sc=13106&sc=13107&sc=13108&sc=13118&sc=13121&sc=13122&sc=13123&sc=13109&sc=13110&sc=13111&sc=13112&sc=13114&sc=13115&sc=13120&sc=13116&sc=13117&sc=13119&cb=0.0&ct=99999999&et=99999999&ts=3&cn=99999999&mb=0&mt=99999999&shkr1=03&shkr2=03&shkr3=03&shkr4=03&fw2=&srch_navi=1

[2]株式会社リクルート(2024)「東京都のエリアから探す賃貸一戸建ての賃貸住宅[賃貸マンション・アパート]情報」 2024年10月31日閲覧

https://suumo.jp/jj/chintai/ichiran/FR301FC001/?ar=030&bs=040&ta=13&sc=13101&sc=13102&sc=13103&sc=13104&sc=13105&sc=13113&sc=13106&sc=13107&sc=13108&sc=13118&sc=13121&sc=13122&sc=13123&sc=13109&sc=13110&sc=13111&sc=13112&sc=13114&sc=13115&sc=13120&sc=13116&sc=13117&sc=13119&sd=1&cb=0.0&ct=99999999&et=99999999&ts=1&cn=99999999&mb=0&mt=99999999&shkr1=03&shkr2=03&shkr3=03&shkr4=03&fw2=&srch_navi=1

[3]SAIBOU PARK MAGAZINE(2023.3.24)「東京23区の地震危険度ランキング最新版！首都直下地震のリスクが高いのは？」 2024年10月31日閲覧

<https://magazine.saiboupark.jp/technology/85/>

[4]帝国書院(2024.4.2)「人口・人口密度 [2023年]: 東京都」 2024年10月31日閲覧

<https://www.teikokushoin.co.jp/statistics/prefecture/detail/13/data/2/>

[5]マイナビニュース(2022.8.1)「東京23区の平均年収ランキングを発表! 年収の高い区、低い区の特徴とは?」 2024年10月31日閲覧

<https://news.mynavi.jp/article/20220801-2411175/>

[6]manabu不動産投資(2024.3.18)「【最新版】東京23区の治安ランキング！犯罪発生率が低い安全な区は？」 2024年10月31日閲覧

<https://manabu.orixbank.co.jp/archives/332>