「山形県の不動産価格の予測」

キーワード:変数変換 線形回帰 ランダムフォレスト サポートベクターマシン 地方都市

1.はじめに

本論文では山形県における2008年から2023年の間に取引された中古マンションの価格予想モデルの構築を行う。

2.データの確認

本論文で扱うデータは全て不動産情報ライブラリから取得している。データの時期は取引時期が2008年から2023年までのデータを用いる。不動産価格は様々な種類があるが、本論文では中古マンションの価格に焦点をあて分析していく。まずは分析・予測の対象となる中古マンションの価格データの特性を調べていくことにする。また本論文では時系的な変化は考慮していない。

中古マンションの価格のヒストグラム、箱ひげ図、基本統計量を以下に示した。

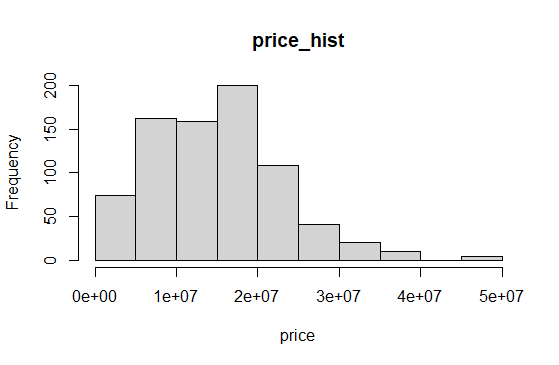


図1 中古マンションの価格のヒストグラム

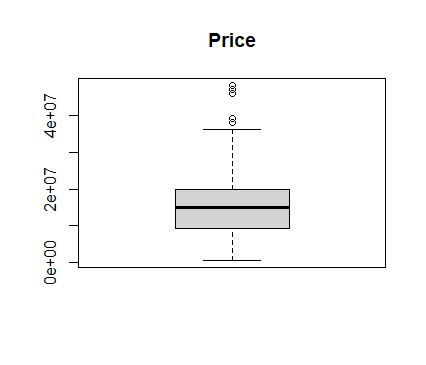


図2　中古マンション価格の箱ひげ図

表1 基本統計量

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 基準 | Min | 1st\_quart | Median | Mean | 3rd\_quat | max |
| 値 | 760,000 | 9,300,000 | 15,000,000 | 15,476,221 | 20,000,000 | 48,000,000 |

\*サンプルサイズ : 779

図1を見ると、500万円から2,000万円の範囲にデータはよく集まっていることがわかる。表1より、データの平均値が約1,547万円で、中央値が1,500万円、第1四分位数が930万円であり、第3四分位数が2,000万であることが示されており、このことからその傾向が見て取れる。

最大値と最小値に関してみると、表1より最も安い中古マンションの価格は76万円で、最も高い中古マンションの価格は4,800万円であることがわかる。図2より価格の大きい方外れ値が存在していることがわかり、それは図1からも読み取れるだろう。

中古マンション価格のような右に裾が厚いデータには、よく対数変換が用いられる。対数変換を用いることで正の歪度を持った中古マンション価格の分布に対して、歪度を減少させ、データをより正規分布に近づけることが可能である。これによって後の統計的推測及び統計的検定を行う際に都合がよくなり、また馬場・清水(2024)にもあるように、外れ値の探索がしやすくなるというメリットがある。よって中古マンション価格に対数変換を施したデータのヒストグラムと箱ひげ図を以下に図示した。

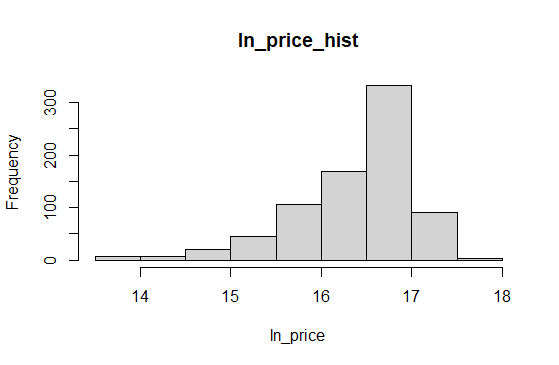


図3　対数変換した中古マンション価格のヒストグラム

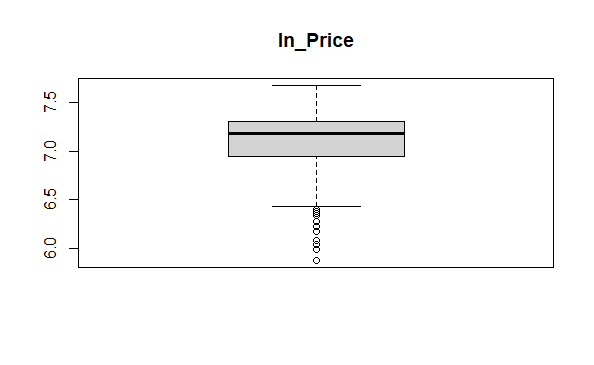


図4　対数変換した中古マンション価格の箱ひげ図

図3をみると、データは図1に比べて正規分布に近い値になったが、対数変換を施したことで逆に左に裾が厚くなってしまったことがわかる。実際に図4をみると、値の大きい方に外れ値は存在しなくなったが、一方で値の小さい方に外れ値が生じてしまっていることがわかるだろう。

上記の結果を踏まえ、中古マンション価格のデータの扱いについて考えていく。中古マンション価格の中に取引価格が100万円以下という極端に小さい値をとるデータが存在しているが、今回の分析ではそのデータをそのまま扱う。なぜならば、馬場・清水(2022)において地方都市には非常に安い価格で取引されるような物件が多く存在し、それらの物件が不利な条件の重ね合わせによって生じていることが示されており、このことを踏まえると山形県の中古マンション価格のデータセットの中で見られた値が極端に小さいデータは、正常に市場の需要を反映して決まった値だと考えられ、それを外れ値として除外するのは望ましくないと考えたからである。また、値の大きなデータについても今回の分析では除外しない。価格の最高値は4,800万円であり、中古マンションの価格を考える上で明らかに異常なデータとは言えないと考えたからである。以上の理由から、中古マンション価格に対する異常値処理は本分析では一切行わない。データの異常値処理を行わないとき、図1と図3よりデータの分布を正規分布に近づけるという意味での対数変換を行う必要性はあまりないと考えられるため、中古マンション価格は対数変換を行わず、そのままの値を用いることにする。

以上のことから、本研究においては得られた中古マンション価格データの異常値処理は行わず、また対数変換も行なわずそのままのデータを用いて分析・予測を行っていく。

3. 説明変数の選択

3.1　欠損値処理

ここでは分析に用いる説明変数の選択について考える。用いる変数は不動産価格情報ライブラリから得られる各中古マンションに紐づいたデータから選択する。

後にも述べるが変数選択ではよく統計的検定や情報量基準などが用いられるが、そのような変数選択を行う前にデータの形や各変数データのサンプルサイズなどを確認し、一定の基準を満たさないようなデータについて統計分析を行う前に削除するという手続きを行う必要がある。以下ではその手続きを行った過程と結果を記載する。

まずは各説明変数の欠損値について集計を行った。表1の注釈にも記載したが、サンプルサイズが779しかないため、説明変数が欠損していることによるサンプルサイズの減少はできる限り小さくしたいという動機がある。よって各説明変数の欠損値を見て、サンプルサイズの減少をもたらす変数を事前に取り除くという作業を行う。

各説明変数の欠損値の数を集計し、欠損値の多いものから順に並べたものが以下のとおりである。

表2　欠損値の数(上位8項目までを掲載)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 項目 | 取引の  事情 | 今後の  利用 | 用途 | 最寄駅:  距離 | 改装の  有無 | 建ぺい率 | 容積率 | 建物の  構造 |
| 数値 | 766 | 389 | 298 | 223 | 180 | 143 | 143 | 133 |

表2より上位3項目はデータ数を大きく減らしてしまうため、仮に中古マンション価格に影響を与えているとしても削除するべきである。次に表2の下位5項目について考える。まず最寄り駅や改装の有無は中古マンション価格に重要な影響を及ぼしていると考えられるため多少サンプルサイズを犠牲にしても説明変数として導入するべき変数であると考えられる。建蔽率や容積率については、大事な要素であるが建ぺい率や容積率はその地域の都市計画に極めて依存している、つまり、説明変数の一つである「都市計画」と相関関係があるため、「都市計画」を用いるのであれば、少なくとも予測モデルにおいてこの2つの変数は不要であると考え除外した。他にも、建物の構造はこの後用いる築年数とある程度相関していると考えられたことなどから除外した。よって欠損値が多かった8つの変数のうち、最寄り駅までの距離と改装の有無の2つの変数残し、それ以外は除外することにする。また上記の手続きに加え、数値変換が難しいものも加えて削除した。

以上のような説明変数の整理を行った後、次は各中古マンション価格に紐づくデータ行の中で、一つでも欠損値が含まれていればその行を削除した。

3.2 データの変換

欠損値の処理後、文字列データを数値変換を行い、また変数同士の演算を行い新しい変数を作成した。具体的には以下の変数説明で述べる。作成したデータは、このファイルとともに提出したdata\_yamagataというCSVファイルに保存されている。

表3は保存されているデータをまとめたものである。

表3　説明変数一覧表

|  |  |  |  |
| --- | --- | --- | --- |
| 変数 | 詳細 | データの型 | 出所 |
| Price | 中古マンションの価格 | 連続 | 不動産情報ライブラリ |
| Station | 物件から駅までの距離 | 順序変数(離散) | 不動産情報ライブラリ |
| room | 物件の部屋の数 | 順序変数(離散) | 不動産情報ライブラリ |
| region\_dummy | 山形市かそれ以外か | ダミー | 不動産情報ライブラリ |
| LandUse\_dummy | 土地の用途で分類 | ダミー | 不動産情報ライブラリ |
| Renov\_dummy | 改装を行ったかどうか | ダミー | 不動産情報ライブラリ |
| Age | 取引時点の築年数 | 連続 | 不動産情報ライブラリ |
| Squ | 物件の面積 | 連続 | 不動産情報ライブラリ |

3.1で欠損値を取り除くという操作を行ったことから、改めて分析対象となる中古マンションの価格に関して、ヒストグラムと箱ひげ図を提示する。

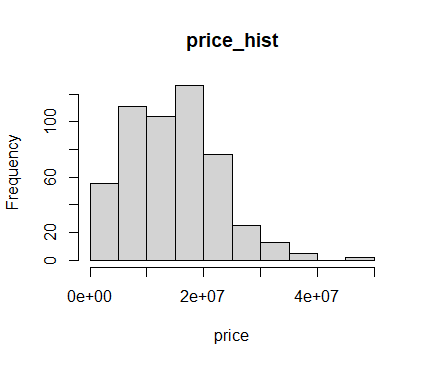


図5中古マンションの価格のヒストグラム

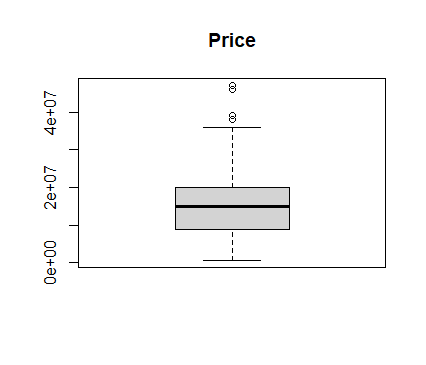


図6 中古マンションの価格のヒストグラム

次に各説明変数の具体的な説明と、中古マンション価格との関係性について記載する。

・Station

中古マンションから最寄り駅までの距離を表す。最寄り駅からの距離が長いほど電車を使う際に不便になることから価格が下がると考えられるだろう。不動産情報ライブラリから取得したデータは、中古マンションから駅までの時間が30分以下である場合はその値が連続値として得られるものであったが、30分以上になるような場合は、物件から駅までの時間が「30分から60分」という形でデータが集計されており、そのままのデータを当てはめることが出来ないという状態にあった。このようなことから本研究では駅までの時間を1から5までの順序変数に変換し。それを用いて分析・予測を行うことにした。変換は以下の基準に従ってグルーピングを行っており、値が大きいほど駅からの離れた場所に物件がある。このような変換を行うことで30分から60分のデータを数値的に扱うことが出来る。ただし回帰分析の結果の解釈には十分注意する必要があり、今回の場合であれば、回帰係数の値そのものにほとんど意味はなく、符号のみしか判断材料にならない。

表4　物件から駅までの距離におけるデータの変換

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 変換前 | 1~5(分) | 5~10分(分) | 10~20(分) | 20~30(分) | 30~60(分) |
| 変換後 | 1 | 2 | 3 | 4 | 5 |
| 総数 | 103 | 96 | 63 | 121 | 133 |

では変換したデータを用いて、中古マンション価格と物件から駅までの距離の関係性についてみていく。図7は、グループごとの物件価格の箱ひげ図である。図7を見ると、全体的な傾向としてやはり駅から離れた物件ほど安い価格がつくことが多いように見える。ただし、グループ1の中に駅からの距離が近いのにもかかわらず、最安値の物件があり、またグループ2は平均的にグループ3やグループ5よりも物件の価格が安くついていることがわかり、一概に駅からの距離によって物件価格が決まるとは言い難いだろう。

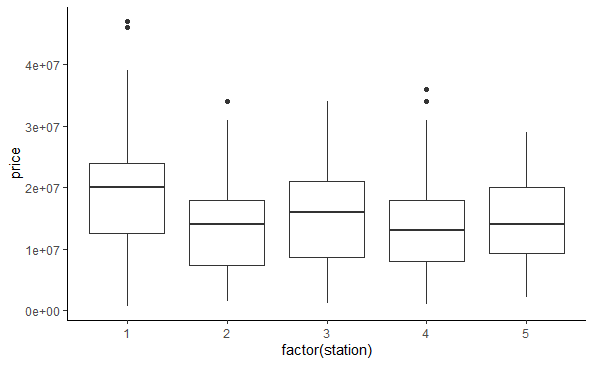


図7 グループごとの物件価格の箱ひげ図

・room

部屋の数は住宅価格に大きな影響を及ぼすと考えられる。部屋の数が多い方がより価格が高くつくと考えられるため、各中古マンションの独立した部屋の数に応じて表5のようにグルーピングした。

表5　住居の間取りおけるデータの変換

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 変換前 | 1room+L+D+K | 2room+L+D+K | 3room+L+D+K | それ以上 |
| 変換後 | 1 | 2 | 3 | 4 |
| 総数 | 39 | 124 | 271 | 82 |

　変換したデータについて、中古マンション価格と住居の間取りの関係性にについてみていく。以下の図8は、グループごとの物件価格の箱ひげ図である。図8を見ると、部屋の数が増えるほど物件価格が高くなるという明らかな傾向が見て取れる。

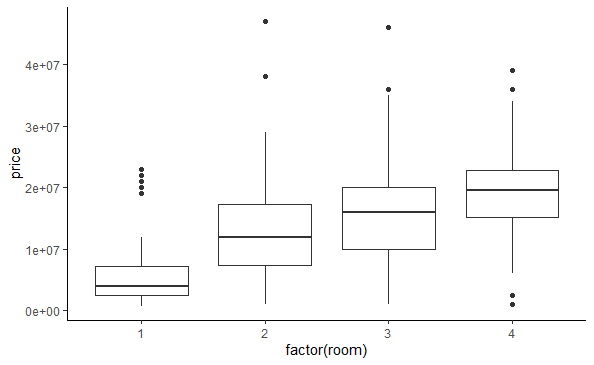


図8 グループごとの物件価格の箱ひげ図

・region\_dummy

地域差に注目した項を作成した。山形市における地域差に注目した理由は以下の山形市の人口動態の特徴による。下の表6を見ると、山形市は人口や世帯数が他の市と比較してとびぬけて多く、また人口密度も県内で最も高い。一方で2位以下の市町村は都市の規模としてあまり大きな差はないように思われる。鶴岡市は面積が大きい分人口が多いが、人口密度を見ると米沢市や天童市に劣っていることがわかる。

以上のことから山形市に居住することの需要は他の市町村のそれと比較して大きいと考えられるため、山形市を1、それ以外を0と置くようなダミー変数を用意し、モデルに組み込むことを考える。

表6　山形県の市町村における人口動態

|  |  |  |  |
| --- | --- | --- | --- |
| 都道府県 | 人口 | 世帯数 | 人口密度 |
| 山形市 | 240,441人 | 105,603戸 | 630.70人/km² |
| 鶴岡市 | 120,398人 | 49,436戸 | 88.37人/km² |
| 酒田市 | 97,395人 | 42,659戸 | 157.43人/km² |
| 米沢市 | 77,232人 | 33,605戸 | 141.06人/km² |
| 天童市 | 61,052人 | 22,867戸 | 536.55人/km² |

　実際にダミー変数を作り、それぞれのグループごとに中古マンション価格の箱ひげ図を描画したものが図9である。図９をみるとやはり山形市の方が他の地域と比較して価格が高い傾向があることが見て取れる。実際に統計的にグループの平均に関する差の検定を行った結果、2つのグループの平均に差はないという帰無仮説に対し、有意水準5％で棄却できことが明らかになった。以上のことから山形市にあるか否かは中古マンション価格を決める上で重要なファクターであり予測モデルに含める価値があるといえるだろう。

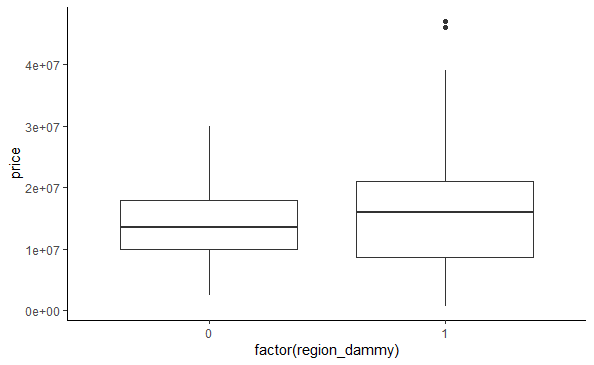


図9　グループごとの物件価格の箱ひげ図

・LandUse\_dummy

都市計画は、都市の発展を計画的に管理するための政策や戦略の総称であり、都市の健全な発展や住環境の改善を目的として実施されるものである。都市計画の中には、都市の秩序立った発展を目指し、用途地域という土地利用の区分が設けられており、各地域には特定の用途が割り当てられ、住宅、商業、工業、農業などの活動が許可または禁止される。

例えば、「商業地域」として割り当てられた土地であれば、その地域には商業施設やオフィスビルが集中し、「工業地域」として割り当てられた土地であれば、工業や倉庫の建設が集中する。一般に用途地域は、土地の利用可能性と利便性に大きな影響を与える。例えば、商業地域は商業施設や公共交通機関が充実しているため、利便性が高く、住宅価格が上昇する傾向がある。一方で、工業地域は公害などの問題が良く発生し、住宅環境としては好ましくないため、住宅価格は低くなる傾向がある。

　以上のことを踏まえて、用途地域ごとにグルーピングを行い、それをダミー変数として回帰モデルの中に組み込むことを考える。グルーピングは表7基準で行った。

表7 　用途地域に関するデータの変換

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 変換前 | 低層住居  専用地域 | 中高層住居専用地域 | 住居地域 | 商業地域 | 工業地域 |
| 変換後 | 1 | 2 | 3 | 4 | 5 |
| 総数 | 0 | 38 | 117 | 294 | 67 |

上の表7に基づいてグルーピングを行い、それぞれのグループごとに中古マンション価格の箱ひげ図を描画したものが図10である。図10をみると平均的には中高層住宅専用地域を除き中古マンション価格に大きな差はない。これは工業地域の中古マンション価格が低く、商業地域の方が価格が高くつくという直感に反しており、また中高層住宅専用地域の価格が他のものと比較して低いこともよくわからない。このようなことからこの変数はあまり分析に貢献しない、別の言い方をすれば予測にはあまり影響を与えない可能性が考えられる。

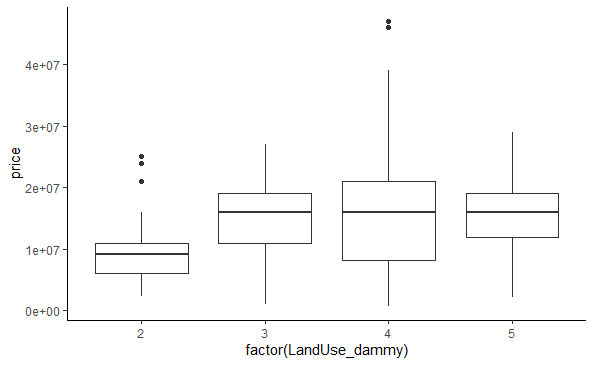


図10　グループごとの物件価格の箱ひげ図

・renov\_dummy

改装されたマンションはそうでないマンションより設備が新しくなり、清潔感も上昇するためマンションの値段が上昇することが考えられる。以上のことより、改装をしたものには1、未改装のものは0にわりあてるようなダミー変数を用意した。それぞれのグループにおける箱ひげ図を以下にプロットした。図11を見ると、改装していないほうが平均的に価格が高くついているという直感に反する結果が生じていることが見て取れる。

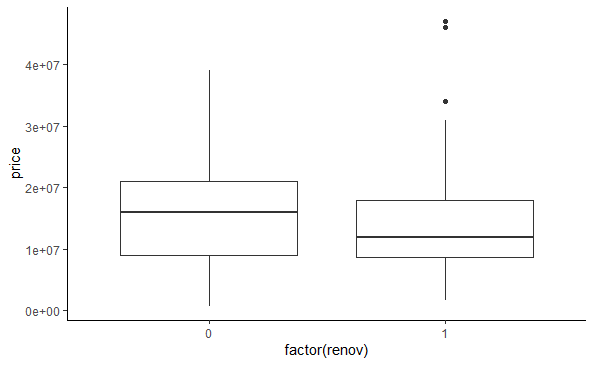


図11　グループごとの物件価格の箱ひげ図

データがうまく直感を反映しない結果になった理由として、改装なしのグループには、そもそも築年数があまりたっておらず、改装の必要のない物件が集まっていた可能性がある。そこで、15年以上の物件のみを対象にし、改装の有無によって価格が変わるか再度検証し、結果をプロットしたものが以下の図である。以下の図12を見ると、改装の必要がある築年数の物件に関しては、改装したものの方がそうでないものよりも高い価格がついていることがわかるだろう。統計的な検定を行っても、有意水準10％で2つのタイプにおける中古マンション価格の平均の差はないという仮説が棄却された。

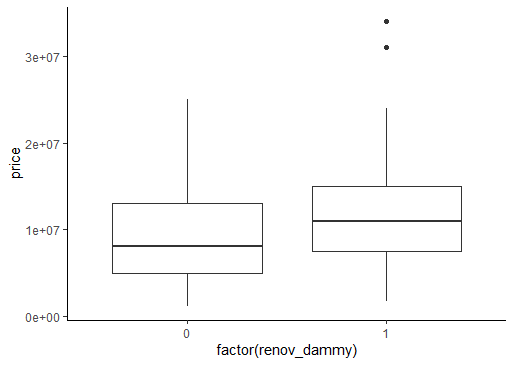


図12　対象を絞った下でのグループごとの物件価格の箱ひげ図

・Age

通常人間は新しいものを好むことから築年数が価格に作用していることが考えられるため、築年数を予測モデルに組み込むことを考える。築年数は不動産情報ライブラリには載っていなかったが、取引年データから建築年を引くことで求めた。築年数と中古マンション価格に関する関係を表したのが図13である。図13を見ると負の相関があることが見て取れ、やはり新しい中古マンションほど高い価格で取引されるということが示唆される。

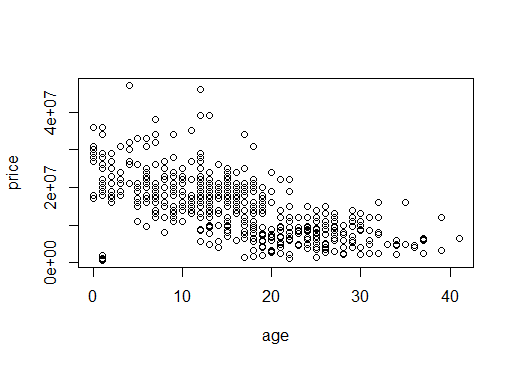


図13 築年数と価格の関連性

・Squ

中古マンションの大きさはその価格に大きな影響を与え、面積が大きい程価格が高騰することが考えられる。実際に中古マンションの面積と価格の関連性を図示したものが図14である。図14より、正の相関があることが見て取れ、この変数は予測に有効である可能性が高いといえるだろう。

　以上のようにして、説明変数の特徴量エンジニアリングと、中古マンション価格との相関関係について分析を行った。

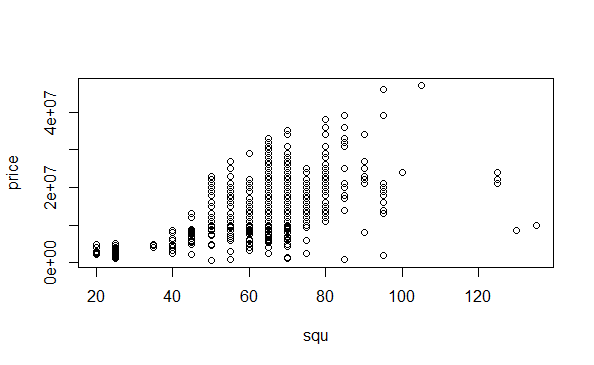


図14 中古マンションの面積と価格の関連性

4. 予測モデルの構築

　次に特徴量エンジニアリングした説明変数を用いて、モデルの構築及び予測を行っていく。モデルを構築するにあたりまずデータをランダムに8:2の比率で学習データとテストデータに分けた。学習データでモデルの構築を行い、テストデータで性能を評価する。

4.1　線形回帰モデル

3章で述べた説明変数を用いて最小二乗法を持ちいて線形回帰モデルの学習を行い、予測モデルを作成する。説明変数の選択は、学習モデルのAICを最小にするような説明変数を選ぶことにする。

上記の過程を経て得られた結果は以下のとおりである。

表8 線形回帰分析結果

|  |  |  |  |
| --- | --- | --- | --- |
| 説明変数 | 係数 | 標準誤差 | P値 |
| Intercept | 9,641,297 | 15,483,114 | 1.17\* |
| Station | -914,382 | 180,495 | 6.13\* |
| Squ | 202,388 | 17,229 |  |
| Age | -441,857 | 32,754 |  |
| Region\_dummy | 2,185,526 | 690,054 | 0.09550 |
| Renov\_dummy | 1,043,324 | 624,425 | 0.00165 |

修正R ²　0.5595

F値 107.7

以上の結果を見ると、中古マンション価格に影響を与えると思われた変数がほぼ採用され、修正R²が0.5595と大きくは価格の変動を説明するようなモデルを作成することが出来た。各変数のP値をみても大きく棄却されるような変数は含まれていないことがわかる。

多重共線性についてのテストも行った。結果は以下である。

表9 VIFの出力結果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 変数 | Station | Squ | age | renov\_dummy | region\_dummy |
| 値 | 1.060 | 1.164 | 1.250 | 1.076 | 1.044 |

一般にVIFが5を超えた場合、多重共線性が疑われるが、上の結果より明らかな多重共線性はないといえる。

推定された線形回帰モデルを用いて、先に分割し残していたテストデータに当てはめ、学習モデルが未知データに対してどれほどの精度で予測できるかを検証する。

テスト結果を以下にプロットした。赤い線は45線を表し、この線の近辺にデータが散らばっていることが望ましい。

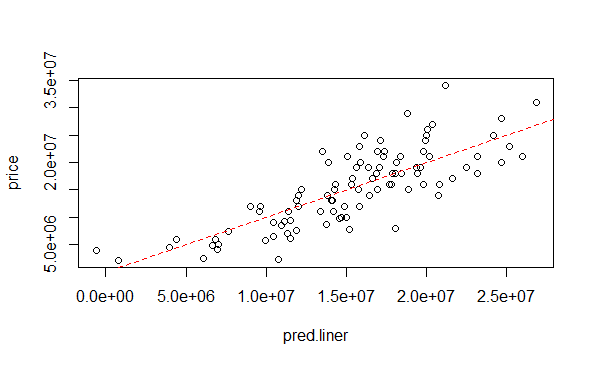


図14 中古マンションの面積と価格の関連性

図14を見ると、回帰線形モデルはデータの特徴をとらえており、全体的に良い予測を行っていることがわかる。

4.2 非線形モデルとの比較

これまでは線形回帰を用いてモデルを学習し、予測を行ってきた。推定された回帰モデルの性能を比較する一つの方法に、他の学習モデルを用いて、そのモデルから出力された予測値の精度と比較するという方法がある。以上のことから、本節においては説明変数は同一にし、学習モデルを変化させることでモデルの精度が線形回帰と比較してどのように変化するかを検証する。西・清水(2024)を参考にし、使用するモデルはランダムフォレストとサポートベクターマシン(SVM)である。

・ランダムフォレスト

ランダムフォレストは機械学習アルゴリズムの一種であり、複数の決定木と用いて分類・回帰問題を解く手法である。複数の決定木をバギングと呼ばれる手法を通じて組み合わせ、モデルを学習する。

ランダムフォレストを用いて得られた予測結果を以下にに示す。

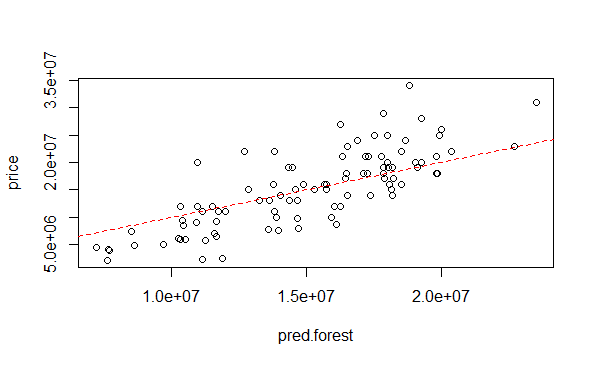


図15　ランダムフォレストによる予測

・サポートベクターマシン

　サポートベクターマシン(SVM)は機械学習の一種であり、分類や回帰問題を解く手法である。ランダムフォレストのように非線形なデータに当てはめることができる。

サポートベクターマシンを使って得られた予測結果を以下に示す。

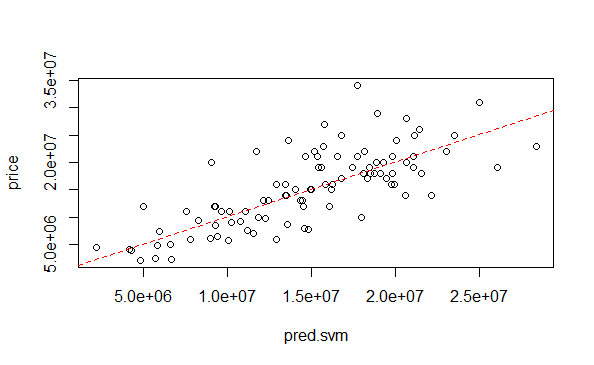


図16SVMによる予測

線形回帰との比較のためにRMSEの指標を用いる。各モデルにおけるRMSEの値は以下のとおりである。

表10各モデルにおけるRMSE

|  |  |  |  |
| --- | --- | --- | --- |
| モデル | 線形回帰 | ランダムフォレスト | SVM |
| RMSE | 4,347,662 | 4,457,415 | 4,031,618 |

以上の結果から線形回帰はランダムフォレストの予測パフォーマンスを上回っているが、

SVMの方がより良い予測を行っていることがわかる。ただしその差はあまり大きなものではなく、図14,15,16を見ても同じような予測を行っていることがわかるだろう。この結果を推察すると、住宅価格は真のモデルも線形に似た形をしているため、非線形モデルを当てはめても大きな予測の差は生まれなかったのではないかと思われる。

5. おわりに

　本論文では山形県の中古マンション価格の予測モデルの構築を行った。変数選択を行い、3つの予測モデルの構築を行った。本研究では不動産情報ライブラリーから得られるデータを利用し、それらのデータを可能な限り用いることを念頭に特徴量エンジニアリングを行った。追加研究として、不動産情報ライブラリーにある説明変数だけでなく、他の特徴量を増やすことなどや予測モデルを変更することなどが考えられるだろう。

参考文献

西颯人・清水千弘(2024) 「不動産取引価格データによる予測モデルの構築 -計量経済学的接近法と機械学習の応用-」

馬場弘樹・清水千弘(2024) 「自治体の政策評価における計量経済的接近法 -不動産価格データに用いた政策評価の実践-」

馬場弘樹・清水千弘(2024)「自治体の政策評価における計量経済的接近法-不動産価格データを用いた政策評価の実践-」公益財団法人日本都市計画学会　都市計画論文集 Vol.57 No3.2022年10月

R コード

#packageをロード

library(tidyverse)

library(e1071)

library(readr)

library(dplyr)

library(stringr)

library(ggplot2)

library(stringi)

library(caTools)

library(randomForest)

library(rpart)

library(kernlab)

library(nnet)

library(car)

library(Metrics)

#データのimport

df <- read\_csv("data\_yamagata.csv",locale = locale(encoding = "UTF-8"))

#dfの型を確認

is.data.frame(df)

#データの数

ndf<-nrow(df)

#priceデータの確認

hist(df$price,main='price\_hist',xlab='price')

boxplot(df$price,main='Price')

summary(df$price)

#対数変換

ln\_price <- log(df$price)

hist(ln\_price,main='ln\_price\_hist',xlab='ln\_price')

boxplot(df$ln\_price,main='ln\_Price')

#staion

ggplot(df,aes(x=factor(station),y=price)) + geom\_boxplot() + theme\_classic()

#room

ggplot(df,aes(x=factor(room),y=price)) + geom\_boxplot() + theme\_classic()

#LandUse

ggplot(df,aes(x=factor(LandUse\_dammy),y=price)) + geom\_boxplot() + theme\_classic()

#region\_dammy

ggplot(df,aes(x=factor(region\_dammy),y=price)) + geom\_boxplot() + theme\_classic()

#region\_test

price\_yamagata <- df$price[df$region\_dammy == 1]

price\_other <- df$price[df$region\_dammy == 0]

t\_test\_result <- t.test(price\_yamagata, price\_other)

print(t\_test\_result)

#renovation\_dammy

ggplot(df,aes(x=factor(renov\_dammy),y=price)) + geom\_boxplot() + theme\_classic()

#extract data with over 15

df\_over15 <- df[df$age>15,]

ggplot(df\_over15,aes(x=factor(renov\_dammy),y=price))+geom\_boxplot()+theme\_classic()

# renovation\_dammy test

price\_renov <- df$price[df\_over15$renov\_dammy == 1]

price\_nonrenov <- df$price[df\_over15$renov\_dammy == 0]

t\_test\_result <- t.test(price\_renov, price\_nonrenov)

print(t\_test\_result)

#age

plot(df$age,df$price,xlab='age',ylab='price')

#squere

plot(df$squ,df$price,xlab='squ',ylab='price')

#シードの固定

set.seed(120)

# データをトレーニングセットとテストセットに分割

split <- sample.split(df$price, SplitRatio = 0.8)

df.train <- subset(df, split == TRUE)

df.test <- subset(df, split == FALSE)

#線形モデル

liner\_model <- lm(price ~ station + region\_dammy + squ + age + renov\_dammy + room + factor(LandUse\_dammy), data=df.train)

#結果を表示

summary(liner\_model)

#AICを最小に

step(liner\_model)

#線形モデル

liner\_model\_best <- lm(price ~ station + squ + age + renov\_dammy + region\_dammy

, data=df.train)

#結果を表示

summary(liner\_model\_best)

#多重共線性チェック

vif <- vif(liner\_model\_best)

vif

#予測

pred.liner<-predict(liner\_model\_best, newdata=df.test)

RMSE\_liner <- (mean((pred.liner-df.test$price)^2))^0.5

RMSE\_liner

#RMSE\_liner\_true <- rmse(df.test$price,pred.liner)

#RMSE\_liner\_true

plot(pred.liner,df.test$price,ylab='price')

abline(a = 0, b = 1, col = "red", lty = 2)

#ランダムフォレスト

classifier\_Random\_Forest <- randomForest(price ~ station + squ + age + renov\_dammy +region\_dammy,

data=df.train, ntree = 2300)

pred.forest<-predict(classifier\_Random\_Forest,newdata=df.test)

RMSE\_forest <- (mean((pred.forest-df.test$price)^2))^0.5

RMSE\_forest

plot(pred.forest,df.test$price,ylab='price')

abline(a = 0, b = 1, col = "red", lty = 2)

#svm

svm<-ksvm(price ~ station + squ + age + renov\_dammy +region\_dammy ,

data=df.train)

pred.svm<-predict(svm,newdata=df.test)

RMSE\_svm <- (mean((pred.svm-df.test$price)^2))^0.5

RMSE\_svm

plot(pred.svm,df.test$price,ylab='price')

abline(a = 0, b = 1, col = "red", lty = 2)

print(RMSE\_liner)

print(RMSE\_forest)

print(RMSE\_svm)