2021年度　統計データ分析コンペティション

統計数理賞［高校生の部］

求められている住宅

森 颯太（香川県立高松商業高等学校）

論文の概要

空き家が増える原因を分析するために不動産価格予測ＡＩを作成し、住宅の条件を学習データとして用いることで、住宅の面積や築年数、最寄駅からの徒歩時間などの重要度が高いことを示した。結論として、住宅のリノベーションやリフォームに取り組むことによって空き家を減らすことを提案している。

論文審査会コメント

高校生がブラックボックス型の決定樹勾配ブースティング系AIを使うことはトレンディであり、極めてチャレンジングである。前半に記述統計で仮説検定風の分析を試みているのも好ましいが、AIによる傾向を地域ごとに比較したことや、一定水準の考察を行ったこと、欠測値補完も行ったことは、先端的事例であり高く評価できる。

求められている住宅

森　颯太

香川県立高松商業高等学校

１．研究のテーマと目的

近年、空き家の増加が問題視されている。総務省の平成 30 年住宅・土地統計調査によると、空き家率は 13.6%と過去最高である（図 1)。空き家の内訳をみると、一戸建てやマンションを含む共同住宅の割合がほとんどである（図 2)。空き家が増える原因としては、高齢者が転居し住む人がいなくなることや所有者を放置するといったものがある。空き家が増えると、景観の悪化や悪臭の発生、老朽化による倒壊、空き家内部での犯罪や放火、所有者不明による共同住宅管理組合への負担増などといった問題が起こる。

この問題について調べているうちに、現在求められている戸建住宅やマンションの特徴を分析し、空き家のリノベーションによって不動産的価値を高め、顧客に魅力的な価格設定をすることで販売促進につなげられないかという考えになった。そこで、統計データや AI を活用し、過去の不動産取引価格からどのような戸建住宅やマンションが求められているのか分析することを目的にして研究を行った。

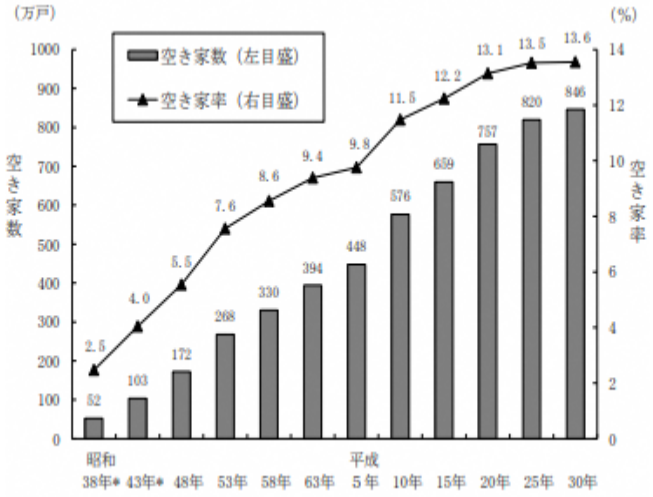

図 1　空き家数及び空き家率の推移 (1)

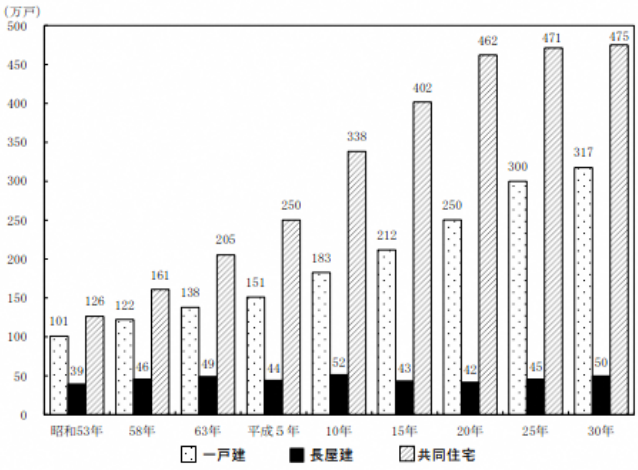

図 2　建て方別空き家数の推移 (1)

２．研究の方法と手順

最初にどのような地域に空き家が多い傾向があるかについて調べることとした。今回は、空き家が多くなっている原因として 3 つの仮説を立て、データ分析によって調べた。そして、地域別にどのような戸建住宅やマンションが求められているのかを調べるため AI に過去の不動産データを学習させ、不動産取引の成約価格予測 AI をつくる。AI が成約価格を予測する際には、取引で重要視されている項目を棒グラフに可視化し、分析した。

AI を使用することによって、多変量解析を統計分析の手段の 1 つとして活用することができる。AI は、人間が行うものと比べ莫大なデータ数を扱うことができる。構築には、LightGBM と呼ばれるオープンソースのモデル構築手法を使用している。LightGBM は、データ分析コンペティションで最上位の解法としても頻繁に用いられる非常に強力なモデルである。AI を実行する環境には、「Google Colaboratory」を使用する。データ分析に使用したデータを表 1 に、AI での戸建住宅・マンションの分析に使用した学習データを表 2 にまとめる。

表１　利用した教育用標準データセット

使用データ	出　　典	年　度
高齢単身世帯数 （65 歳以上の者１人）	総務省統計局「国勢調査報告」人口等基本集計	2015
15 歳未満人口	総務省統計局「国勢調査報告」人口等基本集計	2015
着工新設住宅戸数	国土交通省総合政策局「住宅着工統計」	2019

表２　AI に利用した学習データ

使用データ	出　　典	年　度
空き家数	平成 30 年住宅・土地統計調査　住宅数概数集計　結果の概要	2018
東京不動産取引情報	国土交通省　不動産取引価格情報ダウンロード(2)	
大阪不動産取引情報	国土交通省　不動産取引価格情報ダウンロード(2)	
神奈川不動産取引情報	国土交通省　不動産取引価格情報ダウンロード(2)	
兵庫不動産取引情報	国土交通省　不動産取引価格情報ダウンロード(2)	
香川不動産取引情報	国土交通省　不動産取引価格情報ダウンロード(2)	

３．データセットの加工

「東京不動産取引情報」と「大阪不動産取引情報」、「神奈川不動産取引情報」、「兵庫不動産取引情報」、「香川不動産取引情報」については、戸建住宅と中古マンションを別々に抽出し、欠損値の補完と必要であれば名称の変換を行い、学習に使用した。使用した変数を以下の表にまとめる。

表３　戸建住宅分析に使用した変数

使用変数	欠損値の補完方法と名称の変換、その他の処理
地区名（district）	そのまま使用
最寄駅（station）	AI に学習させるために（場所の名前）などを削除／欠損値を前後の値で補完
徒歩分（walk_min）	すべて分に換算／欠損値を平均値で補完
土地面積（Larea）	そのまま使用
土地形状（shape）	欠損値を前後の値で補完
延床面積（Farea）	欠損値を平均値で補完
建築年（year）	和暦表記を西暦表記に変換／欠損値を平均値で補完
建物構造（structure）	欠損値を前後の値で補完
成約価格	そのまま使用

表４　中古マンション分析に使用した変数

使用変数	欠損値の補完方法と名称の変換、そのほかの処理
地区名（district）	そのまま使用
最寄駅（station）	欠損値を前後の値で補完
徒歩分（walk_min）	すべて分に換算／欠損値を平均値で補完
間取り（type）	欠損値を前後の値で補完
専有面積（area）	そのまま使用
建築年（year）	和暦表記を西暦表記に変換
建物構造（structure）	欠損値を前後の値で補完
容積率（youseki）	欠損値を平均値で補完
成約価格	そのまま使用

表 3、4 の変数は AI に学習させる際に使用したものである。総データ数は約 40000 件とし、AI に学習させるための Train データ 35000 件、モデルの評価をするための Test データとして 5000 件、過学習を防ぐための Valid データとして 5000 件確保して学習を行った。そして、下にあるプログラム 1 が LightGBM により不動産過去データを学習させるプログラムで、プログラム 2 が要素別の重要度を判断するプログラムである。

```
X_trn, X_val, y_trn, y_val = train_test_split(X_train, y_train, test_size=5000, random_state=0)

lgb_dataset_trn = lgb.Dataset(X_trn, label=y_trn, categorical_feature='auto')
lgb_dataset_val = lgb.Dataset(X_val, label=y_val, categorical_feature='auto')

params = {
    'objective' : 'rmse',
    'learning_rate' : 0.1,
    'max_depth' :4,
}

model = lgb.train(
    params=params,
    train_set=lgb_dataset_trn,
    valid_sets=[lgb_dataset_val],
    num_boost_round=10000,
    early_stopping_rounds=100,
    verbose_eval=100
)
```

プログラム 1　LightGBM の学習[4]

```
feature_importance = pd.DataFrame({
    'feature_name' : model.feature_name(),
    'importance' : model.feature_importance(importance_type='gain'),
})
feature_importance = feature_importance.sort_values('importance', ascending=False)

plt.figure(figsize = (6, 6))
sns.barplot(data=feature_importance, x='importance', y='feature_name')
plt.savefig('feature_importance.png')
```

プログラム 2　要素別の重要度[4]

4．データ分析の結果

4.1 高齢単身世帯数が多い地域ほど空き家が多いのではないか

高齢者が多い地域では、住居の持ち主が死亡してしまったり、介護施設に入ったりすることで空き家になってしまう。特に高齢者が単身で居住している世帯では、空き家となる可能性が高くなると考えこのような仮説を立てた。

これを検証するために、都道府県別の空き家数と高齢単身世帯数（65 歳以上の方が 1 人で住んでいる世帯の数）との相関関係を調べた。その結果、相関係数は 0.9832327 であり、正の相関がみられた（図 2）。つまり、高齢者が単身で居住している世帯が多い地域ほど、空き家数が多いといった傾向があることが分かった。したがって、この仮説は適切であった。

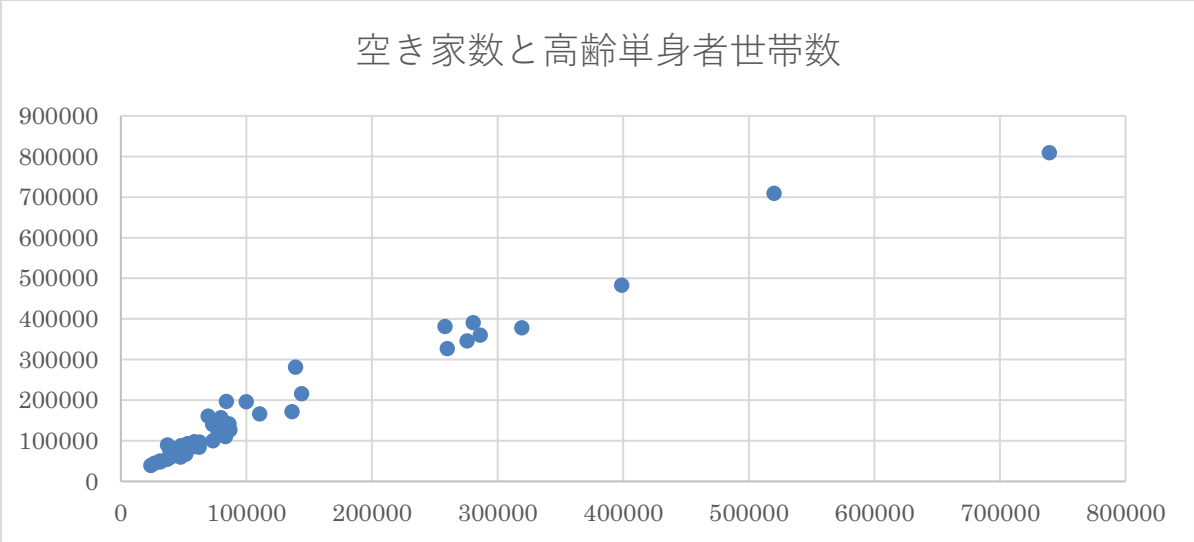

図 3　都道府県別の空き家数と高齢単身世帯数（縦軸：空き家数、横軸：高齢単身世帯数）

4.2 少子化が進んでいる地域ほど空き家が多いのではないか

少子化が進み、子どもの人数が少なくなると、あまり広い住居を必要としない世帯が増加し、部屋数の多い空き家が増えてしまうと考え、このような仮説を立てた。

これを検証するために、都道府県別の空き家数と 15 歳未満人口との相関関係を調べた。その結果、相関係数は 0.9636455 であり、正の相関がみられた（図 3）。つまり、15 歳未満の人口が多い地域であるほど空き家数が多い傾向があることが分かった。「少子化が進んでいる地域ほど空き家数が多いのではないか」という仮説は、否定された。

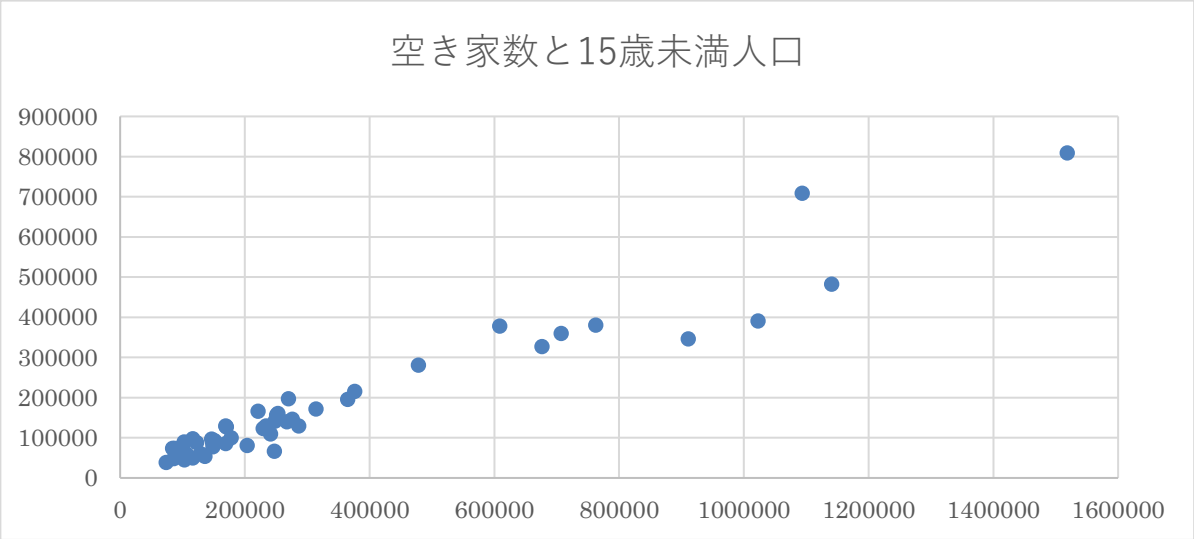

図 4　都道府県別の空き家数と 15 歳未満人口（縦軸：空き家数、横軸：15 歳未満人口）

4.3 新設住宅が多い地域ほど空き家数が多いのではないか

新設住宅の需要が多いと空き家の魅力が落ち、空き家のままとなってしまうことが多くなるのではないかと考え、このような仮説を立てた。

これを証明するために、都道府県ごとの空き家数と着工新設住宅戸数との相関関係を調べた。その結果、相関係数は 0.9441982 であり、正の相関がみられた（図 4）。つまり、新設住宅の着工が多い地域であるほど、空き家が多い傾向があることが分かった。したがって、「新設住宅が多い地域ほど空き家数が多いのではないか」という仮説は適切であった。

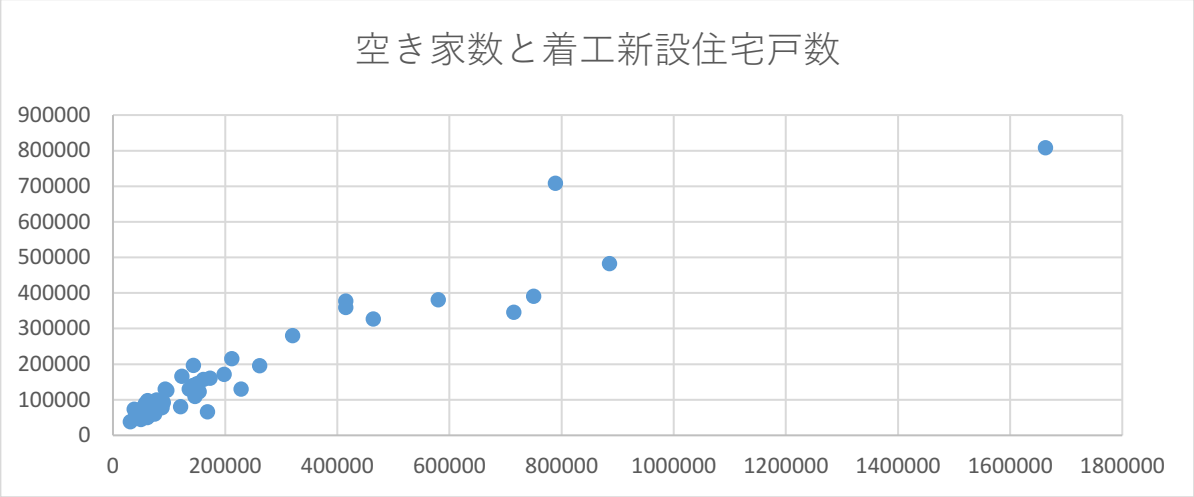

図 5　都道府県別の空き家数と着工新設住宅戸数（縦軸：空き家数、横軸：着工新設住宅数）

4.4 地域ごとに重要視される項目について

以上の結果から、高齢単身世帯数と着工新設住宅戸数が多い地域は空き家が多い傾向があることが分かった。地域差を比較するため、関東では東京都と神奈川県、関西では大阪府、兵庫県の 4 都府県を取り上げ、分析を行う。結果は以下のようになった。縦軸は項目を、横軸は重要度を表す値で示す。重要度を表す値とは、機械学習をする際「モデルの中でそれぞれ特徴量が何回使われているか」をもとに計算された値のことである。

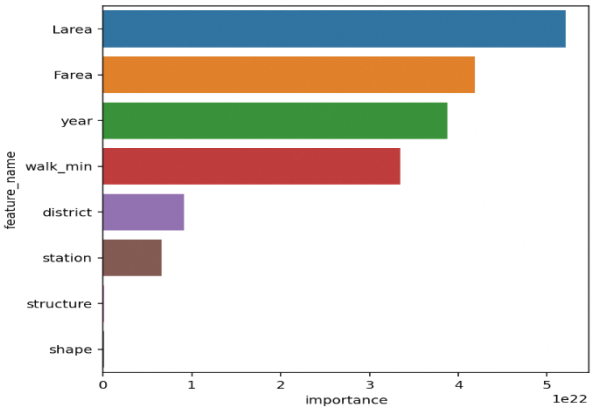

図 6　東京都　戸建住宅

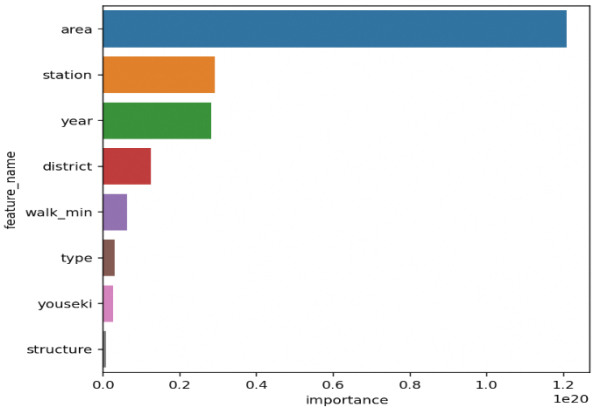

図 7　東京都　中古マンション

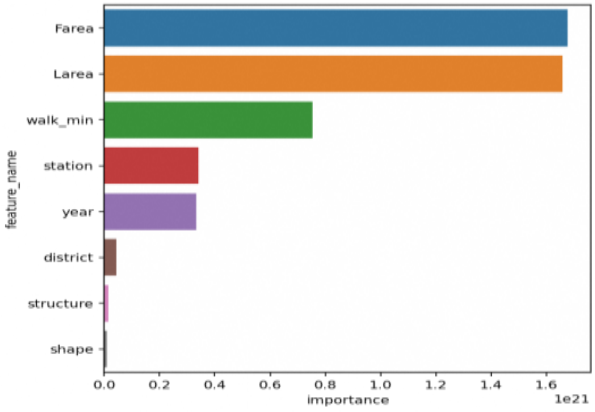

図 8　大阪府　戸建住宅

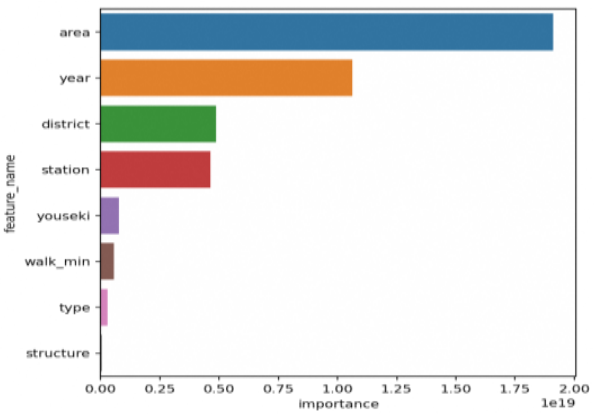

図 9　大阪府　中古マンション

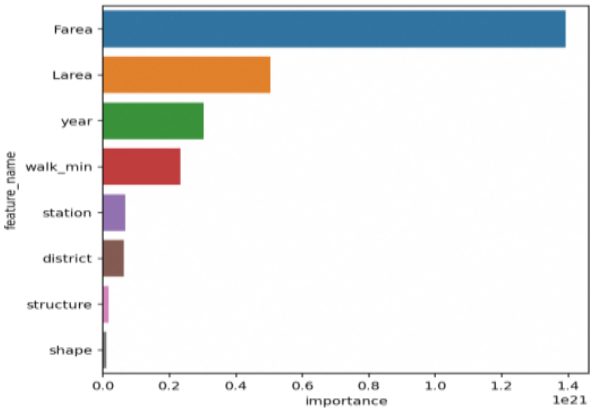

図 10　神奈川県　戸建住宅

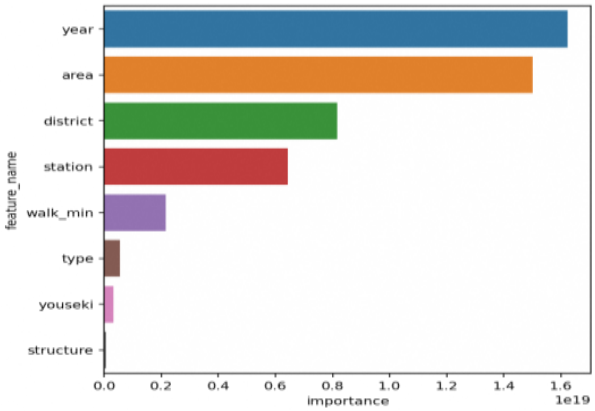

図 11　神奈川県　中古マンション

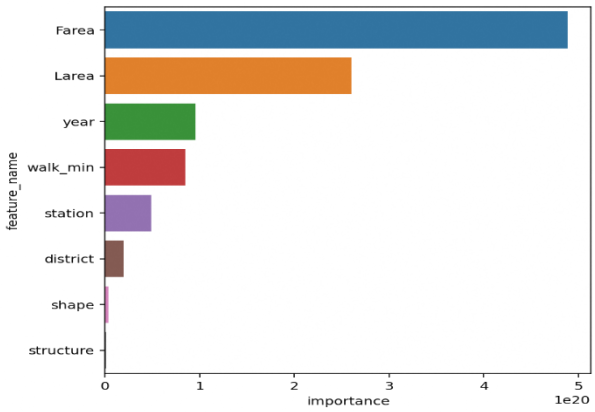

図 12　兵庫県　戸建住宅

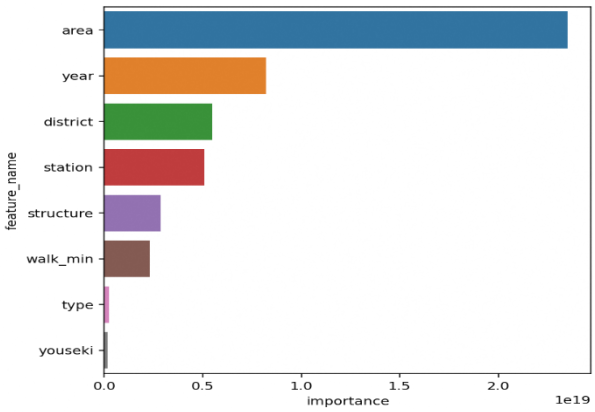

図 13　兵庫県　中古マンション

各県の結果を上位 5 位までを以下の表にまとめた。

表 5　戸建住宅・中古マンションの重要度（降順）

東京都		大阪府	
戸建住宅	中古マンション	戸建住宅	中古マンション
土地面積	専有面積	延床面積	専有面積
延床面積	最寄駅	土地面積	建築年
建築年	建築年	徒歩分※	地区名
徒歩分※	地区名	最寄駅	最寄駅
地区名	徒歩分	建築年	容積率

神奈川県		兵庫県	
戸建住宅	中古マンション	戸建住宅	中古マンション
延床面積	建築年	延床面積	専有面積
土地面積	専有面積	土地面積	建築年
建築年	地区名	建築年	地区名
徒歩分※	最寄駅	徒歩分※	地区名
最寄駅	徒歩分	最寄駅	建物構造

※最寄駅から徒歩でかかる時間

全体の共通点として、重要度が高い項目に延床面積や土地面積、専有面積など「面積」に関するものが多いことが挙げられる。このことから、どの都府県も広さを求めているということが分かった。また、地域によって差が出ると予測していたが、同じ地域内でも差が出たり、反対に異なる地域で一致したりするということも分かった。しかし、都府県ごとにみると重要視されている項目が少し似てはいるものの各地域の特徴がみられた。例えば、東京都の戸建住宅だけ土地面積が重要視されていたり、神奈川県の中古マンションだけ建築年が重要視されていたりする。また、戸建住宅では土地形状や建物構造、中古マンションでは間取りや建物構造、容積率はあまり重要視されない傾向があることが分かった。

4.5 都市部と地方にはどのような重要度の差が出るのか

都市部である東京都、神奈川県、大阪府、兵庫県と比べ、私の地元である香川県は住宅の重要度にどのような差が生まれるのかという疑問を持ったため、香川県の不動産取引データも AI に学習させ都市部との比較を行った。結果は以下のようになった。縦軸は項目を、横軸は重要度を表す値で示す。しかし、香川県はデータ数が約 10000 件と少なかったため、ほかの 4 県とは少し条件が異なる。

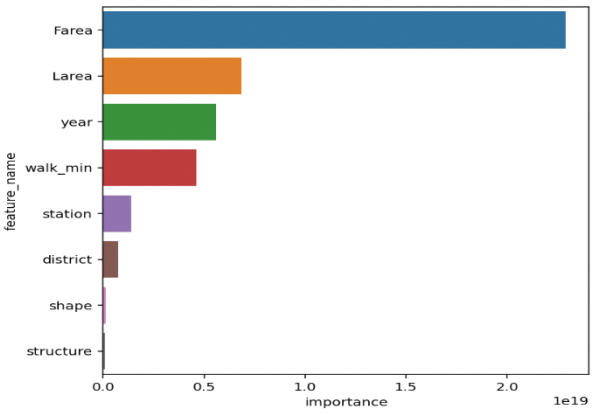

図 14　香川県　戸建住宅

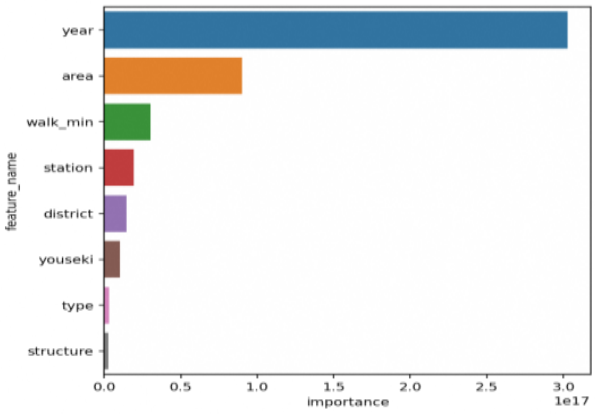

図 15　香川県　中古マンション

ここでも重要度が高い項目には面積に関するものが多かったことから、都市部と地方という違いによって住宅の重要度が大きく変わるわけではないことが分かった。住宅の重要度の分析をするときは、都市部や地方といった見方は必要がないこと、県ごとに少し差が出るという認識が必要であることがこの研究から判明した。

表 6　香川県の重要項目（降順）

香川県	
戸建住宅	中古マンション
延床面積	建築年
土地面積	専有面積
建築年	徒歩分※
徒歩分※	最寄駅
最寄駅	地区名

※最寄駅からかかる時間

4.6 価格予測 AI の性能

作成した価格予測 AI の性能がどれくらいなのか検証した結果、以下のようになった。AI は成約価格を¥22,414,581 と判断し、実際の販売価格との誤差は、¥85,419（-0.38%）となり、AI の有効を示す結果となった。

中古マンション　物件 A	
所在地	千代田区東神田
交通	岩本町駅　徒歩 6 分
間取り	1K
専有面積	22 ㎡
築年月	2005 年 5 月
構造	RC
容積率	100%
販売価格	¥22,500,000

```
data = data.append(pd.DataFrame.from_dict({
    "地区名":["東神田"],
    "最寄駅":["岩本"],
    "徒歩分":[6],
    "専有面積":[22],
    "間取り":["1K"],
    "建物構造":["RC"],
    "建築年":[2005],
    "容積率":[100],
    "成約価格":[-1],
})).reset_index(drop=True)
X = create_feature(data)
model.predict(X[-1:])
```

```
array([22414581.87524405])
```

図 16　AI による予想成約価格 (3)

５．結果の解釈

5.1 考察

4 の結果から、都市部と地方といった差によって重要視される項目が大きく変わることはないと分かったため、ここでは全体を通して分かったことや、そこからどのようなリノベーション・リフォームする必要があるのかという考察を述べる。まず、重要視されやすい項目として延床面積と土地面積、専有面積があったことから、広さが求められる傾向が強いことが分かった。戸建住宅であれば土地面積は隣家との兼ね合いなどから、制約条件であることが多くあるため、ある程度土地の広さを確保できる空き家を優先し、延床面積が可能な限り広くなるようなリノベーション・リフォームをすることが必要だと考えられる。また、行政は区画整理を推進する政策を行うべきである。加えて、徒歩分や最寄駅が重要視されやすい傾向もみられた。これについても最寄駅は制約条件であるので、交通機関へのアクセスがしやすい空き家を優先することが必要であると考えられる。その他に重要視されやすい項目として建築年があったことから、新しさが求められる傾向があることが分かった。空き家は中古となるので新しさを求めることはできないが、綺麗さや建物の安全性を高めることは可能であるため、その点を重視したリノベーション・リフォームが必要であると考えられる。特に中古マンションで建築年が重要視されやすい傾向が強かったので、戸建住宅の空き家よりもさらに綺麗さや安全性を高める工夫が必要であると考えられる。反対に、どの地域でも建物構造や土地形状は重視されにくい項目である。

現在、親族から家を相続したものの、持て余している人や、リノベーション事業への参入を考えている企業などもある。それゆえ、今回の結果を考慮し、リノベーション・リフォームに取り組めば、各地の空き家も魅力が増すと考える。そのためには、その活動を支える政策が必要だと考える。空き家の販売のためにリノベーションをすると補助金が下りたり、空き家を活用し賃貸住宅経営をすると一部税金が控除されたりといった支援をすることで、空き家を活用する人が増え、空き家による様々な問題も減少するだろう。

5.2 本研究について

本研究では、AI を使用することによって多変量解析をすることができた。不動産価格予測 AI を作成し、分析を行った結果、AI を用いた分析を事業に活用することは有効であると考えた。しかし、本研究では不動産取引データを扱う際、欠損値を平均値や上下の値で補完した。したがって、学習の正確性が損なわれているところがあったと考えられる。本研究の AI を事業に活用するには、データの欠損値をなくし、特徴量を増やして精度の高い予測を可能にする必要があり、これが今後の研究課題である。

６．参考文献

・(1) 平成 30 年住宅・土地統計調査　住宅数概数集計　結果の概要
https://www.stat.go.jp/data/jyutaku/2018/pdf/g_gaiyou.pdf

・(2) 国土交通省　不動産取引価格情報ダウンロード
https://www.land.mlit.go.jp/webland/servlet/MainServlet

・(3) 不動産情報サイト　アットホームより
https://www.athome.co.jp/mansion/chuko/tokyo/chiyoda-city/list/

・(4)「土日で学べる「AI 自動化」プログラミング
著者名：李　天琦、秋山　卓也　タイトル：不動産価格「予測 AI」を作る
ページ：p72～ｐ84　発行：日経 BP Nikkei Business Publications,Inc.　発行年：2020 年