

修士論文

飲食店用不動産物件の 賃料推定精度向上に向けた新モデルの提案と評価

鶴山 優季子

奈良先端科学技術大学院大学

先端科学技術研究科

情報理工学プログラム

主指導教員: 安本 慶一 教授

ユビキタスコンピューティングシステム研究室 (情報科学領域)

令和2年3月13日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士(工学)授与の要件として提出した修士論文である。

鶴山 優季子

審査委員：

安本 慶一 教授 (主指導教員)

中村 哲 教授 (副指導教員)

荒川 豊 客員教授 (副指導教員)

諏訪 博彦 特任准教授 (副指導教員)

飲食店用不動産物件の 賃料推定精度向上に向けた新モデルの提案と評価*

鶴山 優季子

内容梗概

飲食店向け不動産物件の賃料は、ベテランの営業職員が長年で培ってきた経験や勘などの暗黙知により決定されている。この手法では、賃料を決定する要因が明らかではなく、人によって価格が異なるといった課題や、新人営業職員への知識継承が難しいといった課題がある。

これらの課題を解決するため、荒川 [1] らは、ベテラン営業職員にインタビューを行い、飲食店向け不動産物件の賃料を決定している要因を抽出し、それらをベースとした飲食店向けの賃料推定モデルを提案している。荒川 [1] らにより抽出された要因としては、ベテラン営業職員による言語化が可能である顕在的情報と、ベテラン営業職員による言語化が困難である潜在的情報があり、顕在的情報の具体的な指標としては、坪数、居抜き、階数、駅徒歩時間、通行量、視認性、平均坪単価が得られている。また、具体的な指標が抽出されなかった潜在的情報としては、物件に付与されているキャッチコピーを用いており、キャッチコピー中の名詞および形容詞を抽出し、それらの単語が賃料に対し正負どちらにはたらくかを算出する手法を提案している。これらの特徴量を用いて構築した推定モデルは、決定係数が 0.738 となる結果が得られている。しかし、荒川 [1] らの構築したモデルでは、データ数や特徴量が少ないといった課題や、キャッチコピーの解析に関しては同一の単語のみ検出可能であり、曖昧さに欠ける課題や、文脈が考慮されていない課題がある。また、賃料推定モデルを実際の現場で用いるためには、より高い推定精度が求められている。

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 2 年 3 月 13 日.

本研究では、荒川らのモデル [1] の課題を解決すること、および推定精度を向上させることを目的とし、新たな飲食店向け不動産物件の賃料推定モデルを構築する。本研究では、賃料推定モデルの精度向上に関する以下の4つの仮説をたて、これらの検証を行う。

仮説1：データ数を追加することにより推定精度は向上する

仮説2：新たな特徴量を追加することにより推定精度は向上する

仮説3：特徴量の掛け合わせを考慮することにより推定精度は向上する

仮説4：キャッチコピーの文脈や曖昧さを考慮することにより推定精度は向上する

仮説1の検証の結果、データ数が100件の場合の解析結果とデータ数が483件の場合の解析結果を比較すると、決定係数が最大で11.5%上昇する結果が得られた。仮説2では、成約した年や業種など、10の特徴量を新たな特徴量として追加した。その検証の結果、仮説1の推定結果と比較すると、決定係数が最大で1.8%上昇する結果が得られた。仮説3では、坪数×居抜きや、駅徒歩時間×視認性などの特徴量を新たな特徴量として追加した。その検証の結果、仮説1の推定結果と比較すると、決定係数が最大で1.2%上昇する結果が得られた。仮説4では、キャッチコピーの文脈や曖昧さを考慮する手法として、Doc2Vecを用いる手法、およびWord2Vecを用いる手法の二種類の解析方法を用い、モデルの構築を行なった。その結果、仮説1の推定結果と比較すると、Doc2Vecの場合で決定係数が最大1.1%、Word2Vecの場合で決定係数が最大1.8%上昇する結果が得られた。

本研究では、これらの仮説を組み合わせることにより、さらなる推定精度の向上を図った。その結果、仮説1、仮説2、仮説4のWord2Vecの手法を組み合わせた場合に最も推定精度が高くなり、決定係数が最大で0.752となる結果が得られ、荒川 [1] らの手法よりも、1.9%向上した。

キーワード

機械学習, 賃料推定, データマイニング, 自然言語処理, Word2Vec, Doc2Vec

Proposal and Evaluation of a New Model for Improving Rent Estimation Accuracy of Real Estate for Restaurants*

Tsuruyama Yukiko

Abstract

The rent of real estate properties for restaurants is determined based on tacit knowledge such as the experience and intuition by experienced veteran sales man. In this method, the factors that determine rent are not clear, thus there are problems such as different prices for different people and difficulties in passing on knowledge to new employees.

To solve these issues, Arakawa et al.[1] interviewed veteran sales man to extracted the factors that determine the rent of real estate properties for restaurants, and based on these factors, they proposed a rent estimation model for restaurants. The factors that determined rent extracted by them are: explicit information that can be putting into words by veteran sales man and potential information that can not be putting into words by veteran sales man. As specific indices of the explicit information, a number of tsubo (floor space), inuki (means that goods and furniture are included), the number of floors, the walking time of the station, the traffic volume, visibility of restaurant, and the average of unit price were extracted. In addition, they proposed a method that uses the catch phrase as the potential information. As a method of analyzing catchphrases, they extracted the nouns and adjectives in the catchphrase, and calculated whether those words act on the rent positively or negatively. The estimation model constructed

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 13, .

using these features has a coefficient of determination of 0.738. However, the models constructed in the this research have problems such as the small number of data and features, and in the analysis of catchphrases the problem of lack of ambiguity because only the same word can be detected. In addition, in order to use the rent estimation model in actual environment, higher estimation accuracy is required.

In this study, we develop a rent estimation model for real estate properties for restaurants, aiming at solving the problems of previous research and improving the estimation accuracy. In this study, we make four hypotheses about the improvement of estimation accuracy and verify them. The four hypotheses are as follows:

First: Estimation accuracy improves by adding the number of data

Second: Estimation accuracy improves by adding new features

Third: Estimation accuracy is improved by considering multiplication of features

Fourth: Estimation accuracy improves by considering the ambiguity of catch phrase

As a result of the first hypothesis,, compared two dataset that the number of data was 100 and 483, the coefficient of determination increased by 11.5 % at the maximum. As a result of the second hypothesis, the coefficient of determination increased by a maximum of 1.8 % compared to the result of first hypothesis. As a result of the third hypothesis, the coefficient of determination increased by up to 1.2 % compared to the result of first hypothesis. In the fourth hypotheis, two methods were used to analyze the catch phrase, Doc2Vec and Word2Vec. Compared to the result of first hypothesis, the result showed that the coefficient of determination increased up to 1.1 % in Doc2Vec method and 1.8 % in Word2Vec method.

Based on the obtained results, the estimation accuracy was further improved by combining hypotheses. As a result, the estimation accuracy was highest when the method of Hypothesis 1, Hypothesis 2, and Hypothesis 4 (Word2Vec method) was combined, the coefficient of determination was 0.752 that was 1.8 % higher than Arakawa's model[1] .

Keywords:

Machine Learning, Data Mining, Rent Estimation, Natural Language Processing,
Word2Vec, Doc2Vec

目次

1. 序論	1
1.1 研究背景および先行研究	1
1.2 研究目的および課題に対するアプローチ	2
1.3 得られた結果	3
1.4 本論文の構成	4
2. 関連研究	5
2.1 住宅物件の価格推定に関する研究	5
2.2 飲食店向け不動産物件の賃料推定に関する研究	6
2.3 自然言語処理に関する研究	8
3. 賃料推定精度向上に向けた手法の提案	10
3.1 データ数の追加による推定精度の向上	10
3.2 新たな特徴量の追加による推定精度の向上	12
3.3 特徴量の掛け合わせの考慮による推定精度の向上	17
3.4 キャッチコピーの曖昧さの考慮による推定精度の向上	18
3.4.1 Doc2Vec 技術を用いたキャッチコピーの解析	19
3.4.2 Word2Vec 技術を用いたキャッチコピーの解析	19
4. 賃料推定モデルの構築および評価方法	21
4.1 賃料推定モデルの構築	21
4.2 賃料推定モデルの評価方法	21
5. 予備解析	23
5.1 ランダムフォレスト回帰のパラメータ	23
5.2 サポートベクター回帰のパラメータ	24
5.3 Doc2Vec のパラメータ	26
5.4 Word2Vec のパラメータ	28

6. 検証結果	30
6.1 データ数を追加した場合の検証結果	30
6.2 新たな特徴量を追加した場合の検証結果	32
6.3 特徴量の掛け合わせを考慮した場合の検証結果	35
6.4 キャッチコピーの文脈と曖昧さを考慮した場合の検証結果	39
6.4.1 Doc2Vec 技術による解析	40
6.4.2 Word2Vec 技術による解析	44
6.5 仮説を組み合わせた場合の検証結果	49
7. 考察	51
7.1 データ数を追加した場合の考察	51
7.2 新たな特徴量を追加した場合の考察	51
7.3 特徴量の掛け合わせを考慮した場合の考察	53
7.4 キャッチコピーの文脈や曖昧さを考慮した場合の考察	54
7.4.1 Doc2vec 技術による解析	54
7.4.2 Word2Vec 技術による解析	55
7.5 仮説の組み合わせによる推定精度の向上	55
8. 結論	59
謝辞	60
参考文献	61
研究業績	65

目 次

1	SECIモデルと荒川[1]らの研究アプローチ	7
2	年代とデータ数の関係	11
3	坪単価とデータ数の関係	11
4	階数と坪単価の関係	16
5	階数と坪単価の関係(補正後)	16
6	駅徒歩時間と坪単価の関係	17
7	キャッチコピーのモデルへの組み込みイメージ	19
8	Word2Vecによる単語のクラスタリングのイメージ	20
9	RFの木の数と推定精度	24
10	RFの木の深さと推定精度	24
11	SVRの γ と推定精度	25
12	SVRの ε と推定精度	26
13	単語出現回数の閾値と推定精度	27
14	windowサイズと推定精度	28
15	データ数と決定係数	31
16	データ数と平均二乗誤差	31
17	追加特徴量と決定係数上昇率	33
18	追加特徴量と平均二乗誤差低下率	33
19	抽出した特徴量を用いた場合の決定係数上昇率	34
20	抽出した特徴量を用いた場合の平均二乗誤差低下率	35
21	「坪数×他の要素」を追加した場合の推定精度比較	36
22	「居抜き×他の要素」を追加した場合の推定精度比較	36
23	「1/階数×他の要素」を追加した場合の推定精度比較	36
24	「1/駅徒歩時間×他の要素」を追加した場合の推定精度比較	37
25	「通行量×他の要素」を追加した場合の推定精度比較	37
26	「視認性×他の要素」を追加した場合の推定精度比較	37
27	「平均坪単価×他の要素」を追加した場合の推定精度比較	38
28	掛け合わせた特徴量によるモデル構築と決定係数上昇率	39

29	掛け合わせた特徴量によるモデル構築と平均二乗誤差低下率	39
30	Doc2Vec によるキャッチコピーのベクトル数と決定係数	41
31	Doc2Vec によるキャッチコピーのベクトル数と平均二乗誤差	41
32	抽出した品詞と決定係数上昇率 (LR)	42
33	抽出した品詞と平均二乗誤差低下率 (LR)	42
34	抽出した品詞と決定係数上昇率 (SVR)	43
35	抽出した品詞と平均二乗誤差低下率 (SVR)	43
36	追加クラスタ (0~35) と仮説 1 との推定精度比較	46
37	追加クラスタ (36~52) と仮説 1 との推定精度比較	47
38	クラスタの追加数と決定係数上昇率	48
39	クラスタの追加数と平均二乗誤差低下率	48
40	年代と平均坪単価の関係	53
41	構築したモデルの推定賃料と実績値 (賃料)	58

表目次

1	モデル構築に用いる全特徴量	15
2	クラスタリングした単語の特徴量化イメージと例	20
3	仮説 1 検証結果	32
4	新たな特徴量を追加した場合の推定精度および仮説 1 との推定精度比較	33
5	賃料推定に重要な特徴量の抽出結果	34
6	仮説 2 の検証結果および仮説 1 との推定精度比較	35
7	掛け合わせた特徴量を追加した場合の推定精度および仮説 1 との推定精度比較	38
8	仮説 4-1 の検証結果および仮説 1 との推定精度比較	44
9	クラスタ内の単語と推定精度相対誤差の正負	44
10	仮説 4-2 の検証結果および仮説 1 との推定精度比較	49
11	仮説の組み合わせと推定精度比較 (LR)	50

12	仮説の組み合わせと推定精度比較 (SVR)	50
13	仮説の組み合わせにより抽出された特徴量	57

1. 序論

1.1 研究背景および先行研究

機械学習の発展に伴い、不動産の分野においても物件の価格推定 [2, 3, 4, 5, 6] といった営業支援が広く行われている。しかし、飲食店向けの不動産に限ってみると、価格推定はほとんど行われておらず、未だにベテラン営業職員が長年で培った、経験や勘といった暗黙知に基づく価格の決定がされている。また、経験の浅い新人営業職員においても、ベテラン営業職員と同様に価格の決定が可能であることが望まれており、新人営業職員への知識の伝承は、先輩職員のもとで実際に業務を行う OJT (On-the-Job Training) などを通して行われてきた。しかし暗黙知による価格の決定では、価格を決定した具体的な要因を示すことができないため、人によって価格が異なるといった課題や、新人職員への知識の伝承が困難であるといった課題がある。

これらの課題を解決するため、荒川 [1] らは、飲食店向け不動産物件の賃料を決定付ける、根拠のある賃料推定モデルを提案している。飲食店向け不動産物件の価格は、ベテラン営業職員の暗黙知により決定されていることから、荒川 [1] らは野中 [7] らの SECI モデルをベースに、ベテラン営業職員の暗黙知に基づいた賃料推定モデルを提案している。SECI モデル [7] とは、暗黙知を表出化して形式知にし、さらにそれらを連結化することで、知識を概念として共有・伝承が可能になることを示したモデルである。このことから、ベテラン営業職員の暗黙知は、表出化して形式知にし、さらに連結化することで共有・伝承が可能となる。

荒川 [1] らは、飲食店向け不動産物件の賃料決定には、3つの要因があることを明らかにしている。一つ目は静的情報で、坪数や駅までの所要時間など、年数が経っても基本的には変化しない情報のことを指す。二つ目は動的情報で、通行量や視認性など、時間によって変化する情報のことを指す。三つ目は潜在的情報で、ベテラン営業職員による指標化が難しい情報のことを指す。荒川 [1] らは指標化の難しい潜在的情報として、物件ごとに与えられているキャッチコピーを用いる手法を提案している。キャッチコピーには物件の雰囲気や特徴が含まれており、それらの要素が賃料に影響を与えていると考えたためである。荒川 [1] らは、

キャッチコピーから名詞と形容詞を抽出し、それらの単語が賃料に対して正に働くか、負に働くかを算出する手法を提案している。

荒川 [1] らが構築した飲食店向け不動産物件の賃料推定モデルでは、決定係数が 0.738 となる結果が得られている。しかし、賃料推定モデルを実際の現場で用いるには、より高い推定精度が求められる。また、荒川 [1] らの推定モデルは、モデル構築に用いたデータ数や特徴量が少ないこと、さらに、キャッチコピーの解析では同一の単語のみ検出が可能であり、表記揺れなどに対応することができず、曖昧さに欠けるといったことや、文脈が考慮されていないといったことが課題としてあげられる。以上のことから、荒川 [1] らが構築した推定モデルは改善の余地があると言える。

1.2 研究目的および課題に対するアプローチ

本研究では、荒川 [1] らの課題を解決すること、および賃料推定精度を向上させることを目的とし、荒川 [1] らが構築した賃料推定モデルをベースに、新たな飲食店向け不動産物件の賃料推定モデルの構築を行う。本研究では推定精度向上に関する以下の仮説を立て、これらの検証を行う。

(1) データの追加による推定精度の向上

荒川 [1] らは、2014 年以前に契約が成立した 184 件の物件を対象とし、モデルの構築を行っていた。しかし、汎用的なモデルを構築するためにも、より多くのデータセットを用いる必要がある。本研究では「データ数の追加により推定精度は向上する」という仮説のもと、新たに 2014 年以降に契約が成立した物件を加え、検証を行う。本仮説を仮説 1 とする。

(2) 新たな特徴量の追加による推定精度の向上

荒川 [1] らは、坪数や駅徒歩時間、キャッチコピーなどの 7 つの特徴量を用い、モデルを構築していた。しかし、飲食店向け不動産物件ではあらゆる要因が複雑に絡み、賃料が決定されていることから、既存の特徴量以外にも、賃料推定に効果的な特徴量が存在すると考えられる。本研究では「新たな特徴量の追加により推定精度は向上する」という仮説のもと、新たに 10 の特徴量を追加し、検証を行う。本仮説を仮説 2 とする。また、既存の特徴量および新規の特徴量の中には、

賃料推定に重要でない特徴量も含まれていると考えられる。そのため、再帰的特徴量消去（Recursive Feature Elimination, RFE）を用い、賃料推定に重要でない特徴量を消去した上で、モデルの構築を行う。

(3) 特徴量の掛け合わせの考慮による推定精度の向上

荒川 [1] らは、通行量×視認性、坪数×平均坪単価のように、2要素を掛け合わせた特徴量を用い、モデルを構築していた。しかし、この2つの特徴量以外にも、要素同士を掛け合わせるにより、推定精度が向上する特徴量が存在する可能性がある。本研究では、「特徴量の掛け合わせを考慮することにより推定精度は向上する」という仮説のもと、掛け合わせた特徴量を仮説1にて構築したモデルに順に追加し、検証を行う。本仮説を仮説3とする。また本仮説においても、賃料推定に重要でない特徴量が含まれている可能性がある。そのため、RFEを用い、賃料推定に重要でない特徴量を消去した上で、モデルの構築を行う。

(4) キャッチコピーの文脈や曖昧さの考慮による推定精度向上

本研究では荒川 [1] らに引き続き、潜在的情報としてキャッチコピーを用いる。荒川 [1] らは、キャッチコピー中の名詞、形容詞を抽出し、それらの単語が賃料に対して正の影響を与えるか、負の影響を与えるかを算出している。しかし、この手法では同一の単語でのみ検出が可能であり、文脈や曖昧さを考慮できないなどの課題がある。そのため、本研究では「キャッチコピーの文脈や曖昧さの考慮により推定精度は向上する」という仮説のもと、キャッチコピーを自然言語処理の技術を用いて解析し、検証を行う。本仮説を仮説4とする。本研究ではキャッチコピーの解析手法を2種類提案する。一つ目の手法はDoc2Vec技術を用いる手法であり、「Doc2Vecを用いてキャッチコピーを解析することにより推定精度は向上する」という仮説のもと、検証を行う。本仮説を仮説4-1とする。二つ目の手法はWord2Vec技術を用いる手法であり、「Word2Vecを用いてキャッチコピーを解析することにより推定精度は向上する」という仮説のもと、検証を行う。本仮説を仮説4-2とする。

1.3 得られた結果

本研究では、推定精度向上に関する仮説を4つ立て、それらの検証を行った。

仮説1の検証の結果、データ数の追加により推定精度は向上する結果が得られ、100件の物件を対象とした解析結果と、全データである483件の物件を対象とした解析結果を比較すると、決定係数が最大で11.5%上昇する結果が得られた。

仮説2の検証の結果、年および地上階の特徴量を追加した場合に、推定精度が向上する結果が得られた。また、用いた全特徴量から、RFEにより重要でない特徴量を消去してモデルを構築した結果、仮説1の結果と比較して、決定係数が最大で1.8%上昇する結果が得られた。

仮説3の検証の結果、1つの特徴量を追加した場合に、推定精度が向上する結果が得られた。しかしながら、用いた全特徴量から、RFEにより重要でない特徴量を消去してモデルを構築した結果、仮説1の推定結果を上回る結果は得られなかった。

仮説4の検証の結果、仮説1との推定精度を比較すると、仮説4-1では決定係数が最大で1.1%、仮説4-2では決定係数が最大で1.8%上昇する結果が得られた。

本研究ではこれら4つの仮説を組み合わせ、さらなる推定精度の向上を図った。その結果、仮説1、仮説2、仮説4-2の手法を組み合わせることにより、決定係数が最大で0.752となる結果が得られた。この値は、荒川[1]らのモデルと比較すると1.9%の上昇であった。

1.4 本論文の構成

本論文の構成について述べる。第2章では本研究に関連する物件の価格推定などの研究や、自然言語処理に関する研究について述べ、第3章では賃料推定モデルの推定精度向上に向けた手法について述べる。第4章では賃料推定モデルの構築方法および評価方法について述べ、第5章ではモデルを構築する際のパラメータを決定するための予備解析について述べる。第6章にて第3章の結果について述べ、第7章にて考察を行い、第8章にてまとめとする。

2. 関連研究

本章では、住宅物件や飲食店向け物件の価格推定に関する研究、および自然言語処理に関する研究について述べる。

2.1 住宅物件の価格推定に関する研究

本項では、住宅物件の価格推定に関する研究 [2, 3, 4, 5, 6] について述べる。

Wang[2]らは、利便性に着目し、アメリカ国内の戸建て物件を対象に、利便性を組み込んだ価格推定モデルを提案している。測定の困難な利便性は、物件の郵便番号や所在地に含まれていると考え、これらの指標と一般的な住宅物件の指標を組み合わせ、ランダムフォレスト (Random Forest, RF) [8] により学習した結果、価格の推定精度が向上する結果が得られている。

Wu[3]らは、台湾での住宅選定に深く関わりのある風水に着目し、風水を考慮した不動産物件の価格推定を行なっている。モデルの構築には、バックプロパゲーションニューラルネットワーク (BPN)、ファジーニューラルネットワーク (FNN)、また彼らが独自に開発したハイブリッド遺伝子ベースのサポートベクター回帰 (HGA-SVR) からなる複数のアルゴリズムを用いて比較を行なっている。その結果、いずれの手法においても風水を考慮した場合に、推定精度が高くなることが得られている。

三浦 [4] らは、インターネット上にある膨大な情報から地域ごとの評判に関する変数を抽出し、不動産側が所有しているデータベース上の情報と組み合わせる方法を提案している。評判に関する変数は、インターネット上で地名を検索し、形態素解析により抽出されている。モデルの構築にはヘドニック分析法が用いられている。ヘドニック分析法とは、「ある商品の価格は、その商品を構成している属性によって決定する」という考え方のことであり、回帰分析を用いて分析を行う。三浦らの結果では、インターネット上の地域の評判を組み込んだ場合に、決定係数が高くなることが得られている。

Chiarazzo[5]らは、交通システムと地域ごとの環境の質が不動産の価格に影響を与えていると考え、人口ニューラルネットワーク (ANN) を用いた住宅物件の

価格推定モデルを構築している。特徴量としては、物件の立地や構造、駅やバス停などの交通に関する値、SO₂やNO_xなど汚染に関する値などが含まれている。その結果、全42要素ある属性のうち、8番目に環境汚染に関する属性が、15番目付近に交通に関する属性があげられることを明らかにしている。

Poursaeed[6]らは、アメリカ国内の住宅物件を対象に、坪数や部屋数といった指標に加え、画像データを用いて価格推定を行う手法を提案している。画像データは、リビング、寝室、キッチン、風呂場などに分類され、畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）を用い、各部屋ごとに高級度を算出している。その結果、画像データを用いた場合に、価格推定のエラー率が低下する結果が得られている。

以上の関連研究 [2, 3, 4, 5, 6] では、立地や間取りが価格のベースとなる住宅物件を対象とした価格推定モデルであり、排気設備や店前の通行量、周囲に競合店があるかなど、あらゆる要因が複雑に絡んでいる飲食店向けの不動産物件とは、価格を決定する指標が異なる。そのため、飲食店向け不動産物件の価格推定に特化した推定モデルを構築する必要がある。しかしながら、住宅物件向けの価格推定である以上の関連研究 [2, 3, 4, 5, 6] では、価格決定において一般的に使用される立地や間取りの情報に加え、利便性や風水、地域の評判、交通システム、環境、視覚情報など、それぞれの地域や考えに沿った特徴量や手法が取り入れられており、こういった価格推定モデルの構築方法は、飲食店向け物件の価格推定モデルにも取り入れることが可能である。以上のことを考慮に入れながら、飲食店向け不動産物件の賃料推定モデルを構築する。

2.2 飲食店向け不動産物件の賃料推定に関する研究

本項では、飲食店向け不動産物件の賃料推定に関する研究 [1] について述べる。

飲食店向け不動産物件の賃料決定は、ベテラン営業職員が長年で培ってきた、経験と勘といった暗黙知をベースとして行われてきた。しかしこの手法では、賃料を決定する具体的な要因が明らかではないため、職員によって賃料が異なるといった課題や、新人営業職員への知識伝承が困難などの課題がある。

荒川 [1] らは、以上の課題を解決するため、野中 [7] らの SECI モデルに基づいた飲食店向け不動産物件の価格推定モデルを提案している。SECI モデルと荒川 [1] らの研究アプローチを図 1 に示す。野中 [7] らによると、暗黙知は (1) 共同化 (暗黙知から暗黙知) → (2) 表出化 (暗黙知から形式知) → (3) 連結化 (形式知から形式知) → (4) 内面化 (形式知から暗黙知) → (1) 共同化といったサイクルを繰り返すことにより、伝承が可能であるとしている。ここで、暗黙知は言葉にできない知識のことを、形式知は言葉にできる知識を意味する。この場合、ベテラン営業職員の暗黙知は共同化したものを表出化して形式知にし、さらに連結化することで伝承が可能である。

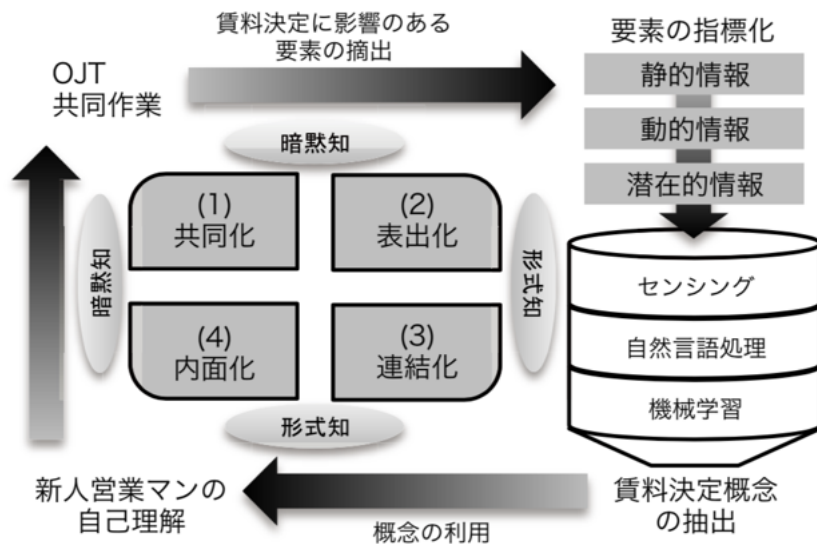


図 1 SECI モデルと荒川 [1] らの研究アプローチ

荒川 [1] らは、暗黙知を形式知として表出化するため、ベテラン営業職員に対しインタビューを行なっている。その結果、物件そのものの情報であり基本的に変わらない静的情報と、物件周辺の情報であり時間によって変化する動的情報が得られた。具体的な指標は、静的情報として坪数、居抜き、階数、駅徒歩時間、動的情報として通行量、視認性、平均坪単価が挙げられた。一方で、営業職員による言語化ができないものの賃料の決定に影響がある指標として、潜在的情報がある。具体的な指標が得られなかった潜在的情報としては、物件に与えられてい

るキャッチコピーを用いており、名詞および形容詞を抽出し、それらの単語が賃料に対して正の影響を及ぼすか、負の影響を及ぼすかを算出する手法を提案している。

以上の手法により構築された推定モデルは、決定係数が0.738となる結果が得られている。しかし、以上のモデルは、2014年以前に契約が成立した184件の物件のみを対象としており、データ数が少ないといった課題や、モデル構築に用いている特徴量数が少ない、また、キャッチコピーの解析方法においては、同一の単語のみ検出が可能であり、表記揺れに対応できないといった課題がある。

2.3 自然言語処理に関する研究

本研究では、自然言語処理技術を活用し、賃料推定モデルを構築する。そのため、本項では自然言語処理に関する研究について述べる。

自然言語処理分野の技術の発展に伴い、膨大なテキストデータの要約や翻訳、重要語の抽出などが広く行われており、その結果を経営戦略やファイナンスなど、様々な分野へ応用する研究がなされている [9, 10, 11, 12, 13]。例えば、Shimoda[14]らは、特定分野で世界的に高いシェアを持つ GNT (global nice top) 企業に共通する成功要因を理解するため、GNT 企業に関する自由記述データをテキストマイニングし、さまざまな産業の成功要因に関する情報を取得している。諏訪 [15]らは、インターネット株式掲示板の投稿内容分析を分析することで、株式指標の変動を予測できる可能性を示している。

膨大なテキストデータを解析する手法として、Word2Vec[16]と呼ばれる技術がある。Word2Vec[16]は、2013年にMikolovらによって提案された手法であり、単語をベクトル表現化することで、単語同士の意味の近さの算出や、単語同士の足し引きが可能となる。Zhang[17]らは、Word2Vecとサポートベクターマシーン (Support Vector Machine, SVM) を組み合わせ、衣料品に関するコメントが肯定的な意味を持つか、否定的な意味を持つかのクラスタリングを行っている。結果として、90%以上の精度で感情のクラスタリングが可能であることを示している。

Wang[18]らは、事前に分散表現化された単語を用い、それらとCNNを組み合わせ、短文のクラスタリングが可能なモデルを提案している。結果として、分散

表現化された単語を用いた場合に、クラスタリングの精度が向上することが得られている。

Xue[19]らは、Word2Vecを用い、中国でのソーシャルネットワークサービスであるweibo上の書き込みから、ユーザの感情をポジティブ、ネガティブ、ニュートラルに分類する手法を提案している。感情の分類には、ポジティブ、ネガティブな単語、程度を示す副詞、感嘆文、反語、感情を表すアイコンなどが学習され、感情の分類を行うための辞書を構築している。その結果、構築した辞書を用いた場合に、感情の分類が83%の精度で可能であることが示されている。

また、Word2Vecは単語をベクトル表現化することのできる技術であるが、それを文書レベルに拡張したものがDoc2Vec技術[20]である。Doc2Vecでは、類似した文書の算出や、文書同士の足し引きが可能となる。例として、Trieu[21]らは、英語で記述されたツイートにより学習されたDoc2Vecモデルを用い、ニュースのクラスタリングを行っている。その結果、学習したモデルを用いることにより、分類精度が向上することが得られている。

また、Doc2Vecを不動産の価格推定に用いた研究もある[22, 23, 24]。Vargas[22]らは、不動産物件の価格と物件に付与されているテキストデータの関係をXGBoostとDoc2Vecを用いて学習し、物件の価格が他の類似した物件の平均価格よりも高いか低いかを予測するモデルを提案している。その結果、テキストデータを組み込んだ場合に、推定精度がわずかに高くなることが示されている。

また、不動産物件に付与されたキャッチコピーをDoc2Vecにより数値化し、価格推定モデルに組み込んだ研究[23, 24]では、キャッチコピーを組み込んだ場合に、推定精度がわずかに高くなることが明らかとなっている。

以上の研究[22, 23, 24]では、テキストデータをDoc2Vecにより数値化し、それらを価格推定モデルに組み込むことで推定精度が向上する結果が得られている。本研究においてもテキストデータを自然言語処理により解析することで、推定精度の向上が見込めると考えられる。

3. 賃料推定精度向上に向けた手法の提案

本研究の目的は、荒川 [1] らの課題を解決すること、および賃料推定精度を向上させることである。これを踏まえ、本研究では以下の4つの仮説を立て、検証を行う。

3.1 データ数の追加による推定精度の向上

荒川 [1] らは、2014年以前に契約が成立した184件の物件を対象とし、解析を行っていた。しかし、より汎用的なモデルを構築するためにも、データ数を追加する必要がある。本項では、「データ数を追加することにより推定精度は向上する」という仮説を立て、これの検証を行う。本仮説を仮説1とする。具体的には、2014年以降に新たに契約が成立した物件を加え、全488件の物件データに対し、100件、200件、300件、400件、488件で解析を行い、推定精度を比較することで、仮説の検証を行う。また、追加データが既存のデータと同質とは限らないことを考慮し、 n ($n < 483$) 件の物件を対象にモデルを構築する場合は、(1) 483件から n 件をランダムに抽出、(2) n 件でモデルを構築、(3) 構築したモデルの推定精度を算出、の流れを10回繰り返し、算出された10の推定精度の平均値を、 n 件でモデルを構築した場合の推定精度とする。

本仮説でのモデル構築の際に用いる特徴量は、坪数、居抜き、階数、駅徒歩時間、通行量×視認性、坪数×平均坪単価の6つとする。これらは、荒川らが構築したモデル [1] にて用いられていた特徴量である。

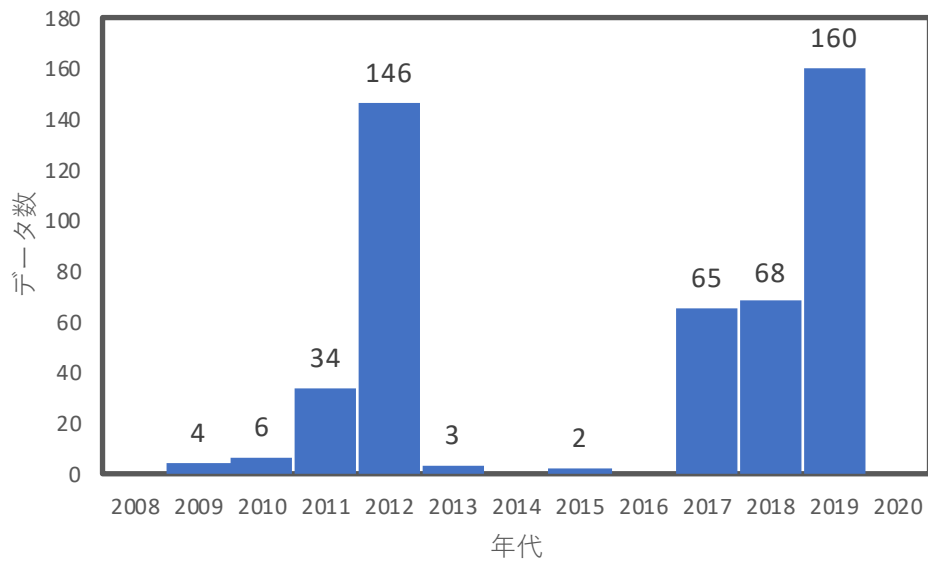


図2 年代とデータ数の関係

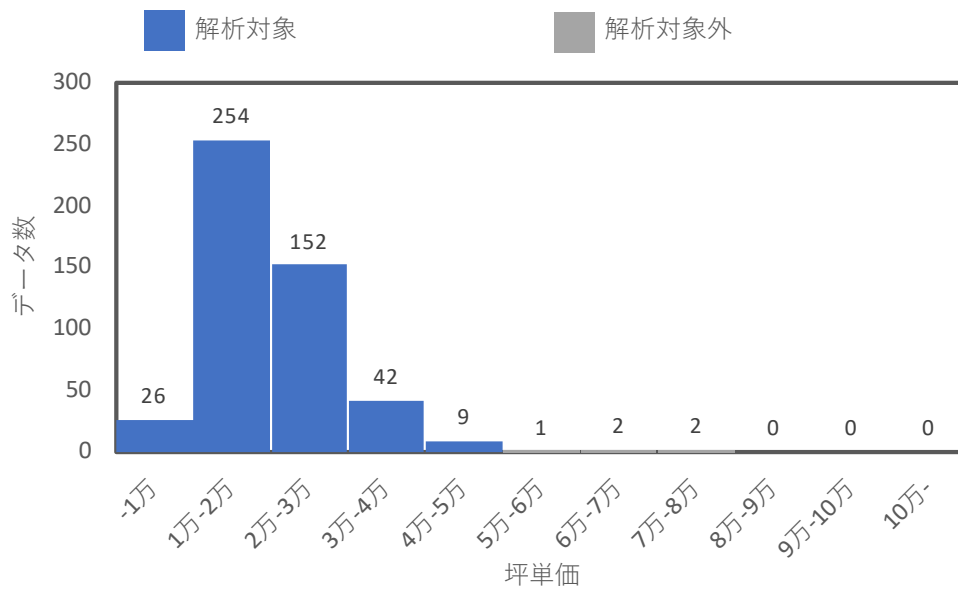


図3 坪単価とデータ数の関係

年代ごとの物件数は、図2に示す通りである。また、高額物件は特殊な物件で

ある可能性があるため、坪単価（＝賃料÷坪数）が 5 万円以上となる物件は、解析の対象外とした。そのため、解析対象となる物件数は 488 件中 483 件である。図 3 に坪単価とデータ数の分布を示す。

3.2 新たな特徴量の追加による推定精度の向上

荒川 [1] らは、静的情報として坪数、居抜き、階数、駅徒歩時間、動的情報として通行量×視認性、坪数×平均坪単価、潜在的情報としてキャッチコピーの計 7 つの特徴量を用いていた。しかし、飲食店向け不動産物件ではあらゆる要因が複雑に絡みながら賃料が決定されていることから、これら 7 つの特徴量以外にも賃料を推定する有効な特徴量があると考えられる。そのため本項では、「新たな特徴量を追加することにより推定精度は向上する」という仮説を立て、新たに 10 の特徴量を追加し、これの検証を行う。本仮説を仮説 2 とする。解析に用いる全特徴量を、表 1 に示す。

次に、表 1 に示した各特徴量について説明する。居抜きとは、物件の内装や調度品、空調設備などが付属されている物件のことであり、新規で店を始める人にとっては、内装にかかる工事費などを軽減できるメリットなどがある。本研究では、居抜き物件を 1、居抜きでない物件を 0 として指標化を行う。

本研究で解析対象となった物件の階数は、地下 1 階から 5 階に分布していることが確認できた。ここで、地下 1 階を -1、その他の階を順に 1 から 5 として指標化し、階数と坪単価の関係をグラフで示すと、図 4 のようになる。この図から、地下 1 階を除き、階数が上がるにつれ坪単価が低下していることが確認できる。このことから、地下 1 階を -1 として指標化するのではなく、2.5 として扱うこととする。このように指標化することで、全ての階が正の値に分布し、辻褄のあったグラフとなる。地下 1 階を 2.5 とした際のグラフを図 5 に示す。

通行量および視認性は、ベテラン営業職員により 5 段階にスコアリングすることで指標化を行っている。荒川 [1] らは、この 2 要素を掛け合わせた「通行量×視認性」の特徴量を用いている。これは、通行量が多いが視認されにくい、または視認されやすいが通行量が少ないといった場合に、物件の評価を低く見積もるためである。また、これらの指標には、ベテラン営業職員の主観的な評価が含ま

れる問題がある。しかしながら、これらの値を機械的に取得することができるようになれば、この問題を解決することができる。

平均坪単価は、各物件の坪単価を算出し、それらを最寄り駅ごとに平均した値としている。

坪数、居抜き、階数、駅徒歩時間、通行量、視認性、平均坪単価の特徴量は、荒川 [1] らによるベテラン営業職員へのインタビュー調査により、飲食店向け物件の賃料決定に有効な特徴量として抽出されたものである。本仮説では、これらの特徴量以外にも賃料推定に有効な特徴量があると考え、以下の 10 の特徴量を追加し、検証を行う。

年は、物件の契約が成立した年としている。荒川 [1] らは、2014 年以前に成約した物件を対象としていた。しかし本研究では、新たに 2019 年までに成約した物件も解析の対象とする。年が異なることによる賃料の影響を考慮するためにも、成約した年を新たな特徴量として追加する。

地上階や地下階は、対象物件がビル等内であれば、そのビルの最上階および最下階とし、指標化を行っている。最下階に関しては、地下 1 階の場合は 1、地下 3 階の場合は 3 のように指標化を行う。図 4 に示した通り、坪単価は階数が上がるにつれ低下することが得られた。しかし、高層ビルの最上階にあるレストランなどは、高額物件であることが予想される。そのため、本研究ではこれらを新たな特徴量として追加し、モデルを構築する。

業種は、重飲食が可能かどうかや業種制限があるかを指標化している。ベテラン営業職員によると、重飲食や業種制限に関しては、店を選択する上で重要視する人も多く、本特徴量も賃料に影響を与えている可能性がある。

表 1 に示した特徴量のうち、「p-」と記載したものは、既存の特徴量のある値の範囲で指標化を行なった特徴量である。例として、p-居抜きは居抜き物件がどの職種に適しているかによって値が変わり、焼肉やお好み焼き店などの重飲食向きの物件は 0、重飲食向きでない物件は 3 といったように、より詳細な指標化を行っている。これらの特徴量は、値が大きければ賃料が高くなるように指標化されている。p-駅ランクは、平均坪単価をベースに指標化しており、p-合計は、「p-」と記載した特徴量の和をとっている。

本項では、これらの特徴量を仮説1にて構築したモデルに順に追加しモデルを構築することで、どの特徴量が賃料推定に有効であるかを検証する。また、既存の特徴量および新規の特徴量から、RFEにより重要でない特徴量を消去し、重要な特徴量のみでモデルを構築することで、推定精度の向上を図る。RFEとは特徴量の重み付けが可能なアルゴリズムを用いて特徴量の重要度を判断し、指定した特徴量数になるまで重要度の低い特徴量を消去する手法である。

RFEを用いるにあたり、特徴量選択の際のアルゴリズムを選定する。本研究では、重回帰分析 (Linear Regression, LR), ランダムフォレスト回帰 (Random Forest Regression, RF), XGBoost回帰 (eXtreme Gradient Boosting Regression, XGBR) の3つを対象に予備解析を行った。その結果、RFを用いて特徴量選択を行い、得られた特徴量を用いてモデルを構築した場合に、最も推定精度が高くなる結果が得られた。このことから、本研究ではRFEの特徴量選択のアルゴリズムとしてRFを用いることとする。

表 1 モデル構築に用いる全特徴量

	特徴量名	種類	既存 or 新規	型	とる値
1	坪数	静的情報	既存	float	
2	居抜き	静的情報	既存	int	0 or 1
3	階数	静的情報	既存	int	
4	駅徒歩時間	静的情報	既存	int	
5	通行量	動的情報	既存	float	0 ~ 5
6	視認性	動的情報	既存	float	0 ~ 5
7	平均坪単価	動的情報	既存	float	
8	キャッチコピー	潜在的情報	既存	float	
9	年	静的情報	新規	int	
10	地上階	静的情報	新規	int	
11	地下階	静的情報	新規	int	
12	業種	静的情報	新規	int	-2 ~ 3
13	p-坪数	静的情報	新規	int	0 ~ 5
14	p-居抜き	静的情報	新規	int	-3 ~ 3
15	p-階数	静的情報	新規	int	-3 ~ 3
16	p-駅徒歩時間	静的情報	新規	int	-5 ~ 5
17	p-駅ランク	動的情報	新規	int	0 ~ 20
18	p-合計	動的情報	新規	float	

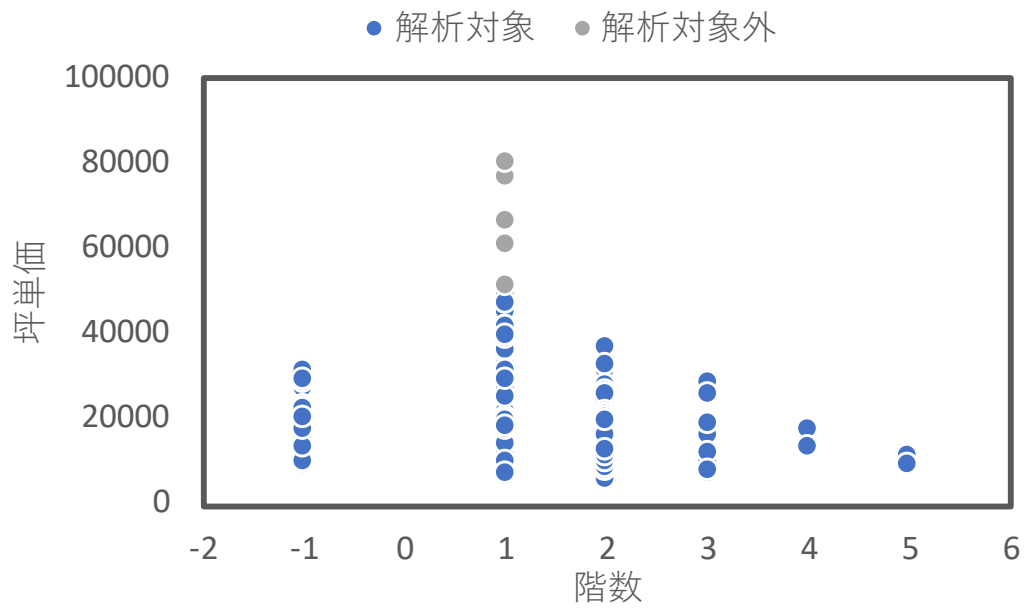


図4 階数と坪単価の関係

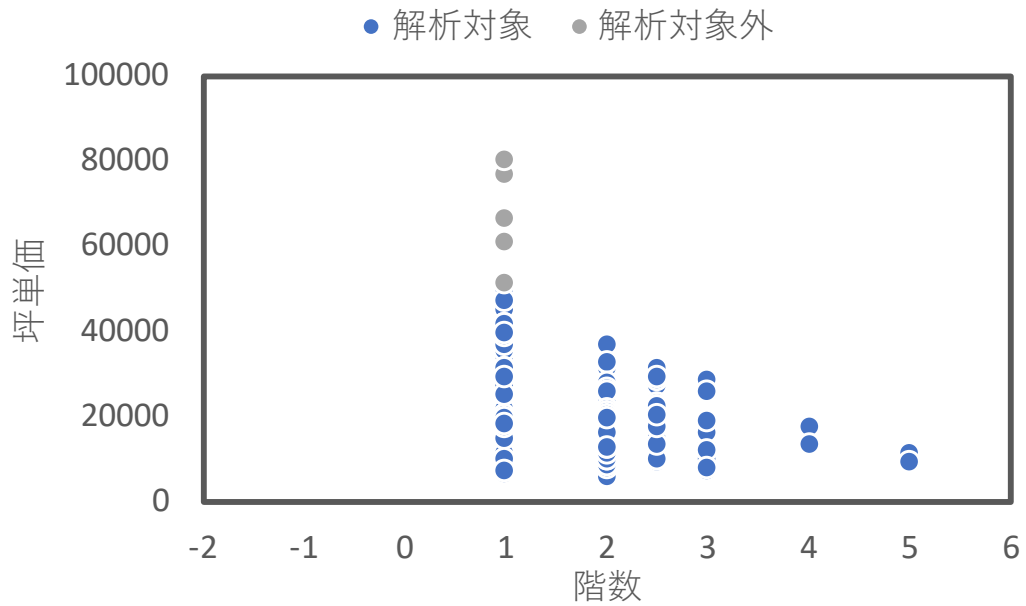


図5 階数と坪単価の関係 (補正後)

3.3 特徴量の掛け合わせの考慮による推定精度の向上

荒川らの研究 [1] は、通行量×視認性、坪数×平均坪単価といった特徴量を用い、モデルを構築している。しかし、これらの特徴量以外にも、坪数×駅徒歩時間や、視認性×平均坪単価など、要素同士を掛け合わせることで、推定精度が向上する特徴量が存在する可能性がある。本項では、「特徴量の掛け合わせを考慮することにより推定精度は向上する」という仮説を立て、これの検証を行う。本仮説を仮説3とする。本仮説の検証の際には、表1中の既存の特徴量から2要素を掛け合わせた特徴量を、新たな特徴量として追加する。

特徴量を掛け合わせる際に注意する点がある。それは、特徴量の値が大きくなることで賃料または坪単価が大きくなるもの、および、特徴量の値が大きくなることで賃料または坪単価が小さくなるものが存在することである。例として、図5に示したように、階数はその値が大きくなるにつれ、坪単価が小さくなることがわかる。階数以外にも、駅徒歩時間で同様の結果が得られた。駅徒歩時間と坪単価の関係を図6に示す。このことから、階数および駅徒歩時間は、逆数にした上で、他の特徴量と掛け合わせることにする。

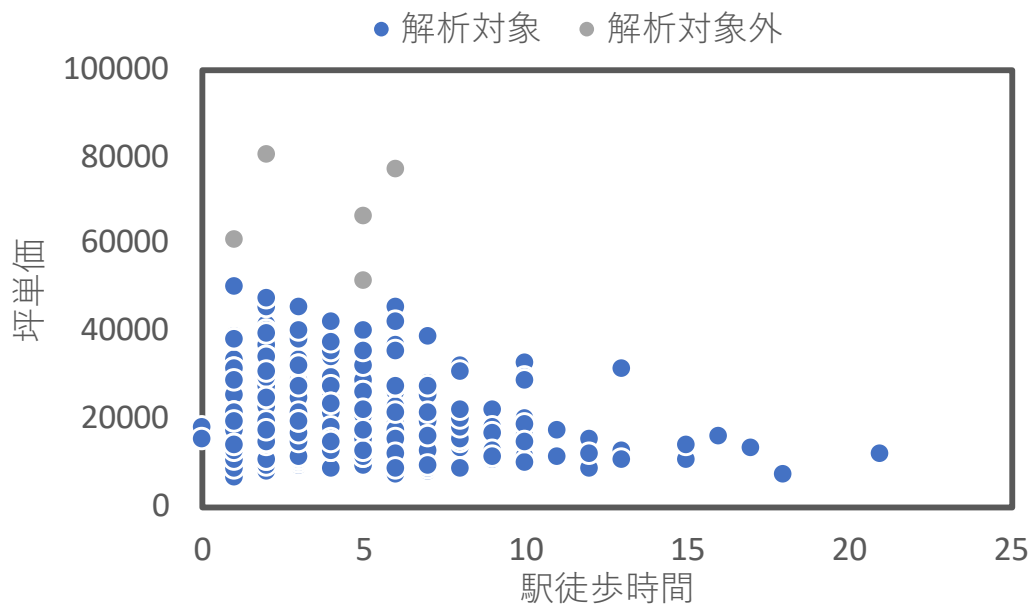


図6 駅徒歩時間と坪単価の関係

本仮説では、2要素を掛け合わせた特徴量を、仮説1にて構築したモデルに順に追加することにより、どの特徴量が賃料推定に有効かを検証する。その後、RFEにより、全特徴量から賃料推定に有効でない特徴量を消去することで、賃料推定に重要な特徴量を抽出し、推定精度の向上を図る。

3.4 キャッチコピーの曖昧さの考慮による推定精度の向上

荒川 [1] らは、キャッチコピー中の名詞および形容詞を抽出し、それらの単語が賃料に対して正に働くか負に働くかを算出する手法を提案し、モデルを構築していた。しかし、この手法では同一の単語でのみ検出可能であり、異なる人が記載したことによる表記揺れなどの曖昧さに対応することができない。具体的には、「レストラン」と「飲食店」といった、類似した意味を持つ単語は全く異なる単語として算出される問題や、「焼肉」や「焼き肉」、「居抜き」や「居抜物件」などの表記揺れに対応できないといった問題がある。こういった問題は、キャッチコピーのクリーニング作業にて取り除かれるべきノイズであるが、全ての表記揺れやノイズに対応するのは困難である。そこで本項では、こういった表記揺れに対応するため、「キャッチコピーの曖昧さを考慮することにより推定精度は向上する」という仮説を立て、これの検証を行う。本仮説を仮説4とする。

本研究ではキャッチコピーの曖昧さを考慮する手法として、2つの手法を提案する。一方は Doc2Vec 技術 [20] を用いる手法、一方は Word2Vec 技術 [16] を用いる手法である。Word2Vec [16] とは、ある単語を任意の数にベクトル化することのできる技術であり、コサイン類似度を求めることができることから、「王様」 - 「男」 + 「女」 = 「女王様」といった、単語の足し引きや類似単語の算出などが可能な技術である。また Doc2Vec [20] とは、単語レベルのベクトル化が可能であった Word2Vec 技術を、文書レベルまで拡張した技術であり、Word2Vec と同様に、文書の足し引きや類似した文書の検索などが可能となる。Doc2Vec および Word2Vec を用いた具体的な解析方法を以下に示す。

3.4.1 Doc2Vec 技術を用いたキャッチコピーの解析

キャッチコピーを解析する一つ目の手法として、Doc2Vec 技術 [20] を用いる。これを用いることにより、キャッチコピーの文脈を考慮することができ、また類似したキャッチコピーの算出もできることから、表記揺れにも対応することができると考えられる。本研究では、「Doc2Vec を用いてキャッチコピーを処理することにより推定精度は向上する」という仮説を立て、これの検証を行う。本仮説を仮説 4-1 とする。Doc2Vec によりベクトル化されたキャッチコピーは、図 7 に示すように静的情報や動的情報と同様にモデルへ組み込み、モデルを構築する。

Doc2Vec を用いるにあたり、キャッチコピーを学習する際のパラメータについては、5 章の予備解析により決定する。

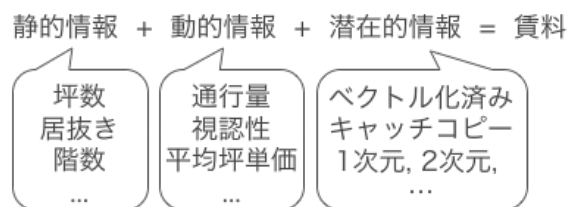


図 7 キャッチコピーのモデルへの組み込みイメージ

3.4.2 Word2Vec 技術を用いたキャッチコピーの解析

キャッチコピーを解析する二つ目の手法として、Word2Vec 技術 [16] を用いる。仮説 4-1 にて用いた Doc2Vec は、文書をベクトル化することのできる技術であり、キャッチコピーを文書として捉えることで、文脈や曖昧さを表現できると考えた。一方で、キャッチコピーに用いられている単語数は記号を除くと平均 12 単語であること、またキャッチコピーには独特な表現や言い回しが記載されていることを考慮すると、文書単位で学習することにより多数のノイズが含まれる可能性がある。そのため、文書単位ではなく単語単位で学習することも検討する。本項では、「Word2Vec を用いてキャッチコピーを処理することにより推定精度は向上する」という仮説を立て、これの検証を行う。本仮説を仮説 4-2 とする。

Word2Vecによりベクトル化された単語は、図8および表2のように前処理を行う。まず、キャッチコピーに含まれる単語をWord2Vecにより学習し、算出されたベクトルをもとに単語のクラスタリングを行う。そのイメージ図を図8に示す。その後、各物件のキャッチコピーに含まれる単語が、どのクラスタにいくつ含まれているかをカウントし、そのカウント数を各物件の特徴量とする。表2が本手法の例である。例えば、表2中の物件ID1では、「きれいな」と「物件」の単語がキャッチコピーに含まれているが、「きれいな」はクラスタ3、「物件」はクラスタ2に含まれているため、これらの特徴量の値がそれぞれ1となる。また物件ID3では、「大学」と「人通り」の二つの単語がクラスタ1に含まれているため、クラスタ1の値が2となる。得られた特徴量（図2中のクラスタ1、クラスタ2など）は、図7で示した式の潜在的情報として組み込み、モデルを構築する。このように処理することで、課題であった表記揺れにも対応することができ、賃料推定に重要であるクラスタや単語を抽出することが可能となる。

Word2Vecにおいても、学習する際のパラメータなどは、5章の予備解析により決定する。

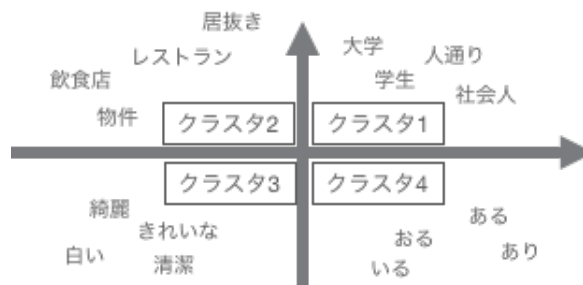


図8 Word2Vecによる単語のクラスタリングのイメージ

表2 クラスタリングした単語の特徴量化イメージと例

物件ID	キャッチコピー	特徴量 (クラスタ)			
		1	2	3	4
1	きれいな物件	0	1	1	0
2	人通りのあるきれいな物件	1	1	1	1
3	大学通り人通り多し	2	0	0	0
4	社会人の多いオフィス街	1	0	0	0

4. 賃料推定モデルの構築および評価方法

本章では、賃料推定モデルの構築方法および構築したモデルの評価方法について述べる。

4.1 賃料推定モデルの構築

賃料推定モデルの構築には、機械学習法を用いる。機械学習法としては、重回帰分析 (Linear Regression, LR), ランダムフォレスト回帰 (Random Forest Regression, RF) [8], サポートベクター回帰 (Support Vector Regression, SVR) [25], XGBoost 回帰 (eXtreme Gradient Boosting Regression, XGBR) の4つを用い、比較を行う。これらの学習モデルのパラメータについては、5章の予備解析により決定する。またモデルの汎用性を示すため、3-fold 交差検証法 (3-fold Cross-validation) を行う。n-fold 交差検証とは、データセットを n 個に分割し、 $n-1$ 個のデータセットで学習、残りの1つのデータセットでテストを行う。それをテストデータを変更しながら n 回繰り返し、それぞれの精度の平均値をとる方法である。今回は $n = 3$ とし、検証を行った。

推定モデルの構築に用いる特徴量は、3章にて述べた坪数、居抜き、階数、駅徒歩時間、通行量×視認性、坪数×平均坪単価をベースとして用い、仮説 2 にて、年、地上階、地下階、業種、p-坪数、p-居抜き、p-階数、p-駅徒歩時間、p-駅ランク、p-合計の10の特徴量を、仮説3では、仮説1にて用いた特徴量の掛け合わせを、仮説4では、Doc2Vec および Word2Vec によりベクトル化したキャッチコピーをそれぞれ追加し、モデルを構築する。

4.2 賃料推定モデルの評価方法

定量的な賃料推定モデルの評価方法としては、決定係数 R^2 と平均二乗誤差 $RMSE$ (Root Mean Squared Error) を用い、これらの値を比較することで行う。

決定係数の定義式を式(1)に示す。式中の y_i は実績値、 \hat{y}_i は推定値、 \bar{y} は実績値の平均値で、 n はデータ数を示している。ここで、実績値は実際の賃料のことを

示しており、ベテラン営業職員の暗黙知を含んでいることから、真値とすることに疑問が残る可能性がある。しかしながら、賃料は需要と供給により決定されることから、本研究では実績値を真値として扱う。決定係数はモデルの当てはまりの良さを示す評価値であり、最も良いスコアは1.0である。

平均二乗誤差の定義式を式(2)に示す。平均二乗誤差はモデルの当てはまりの悪さを示す評価値であり、値が大きくなるほど推定精度が悪いことを示す。したがって、最も良いスコアは0である。

$$R^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \hat{y}_i)^2} \quad (2)$$

また、本研究では決定係数がどの程度上昇し、平均二乗誤差がどの程度低下したかを算出し、各仮説が推定精度に与える影響を確認する。そのため、決定係数がどの程度上昇したかを示す決定係数上昇率、平均二乗誤差がどの程度低下したかを示す平均二乗誤差低下率をそれぞれ式(3)、式(4)のように定義し、算出する。以下の式は、例として仮説2の仮説1に対する決定係数上昇率および平均二乗誤差低下率を示しており、添字が仮説の番号を表している。両の式とも、正の値をとる場合に推定精度が向上したことを示しており、この場合では仮説1よりも仮説2の推定精度が高くなったことを示す。

$$\text{決定係数上昇率} = \frac{R_2^2 - R_1^2}{R_1^2} \quad (3)$$

$$\text{平均二乗誤差低下率} = \frac{RMSE_1 - RMSE_2}{RMSE_1} \quad (4)$$

5. 予備解析

本章では、モデルの構築に用いる機械学習モデルのパラメータを決定するため、予備解析を行う。

5.1 ランダムフォレスト回帰のパラメータ

賃料推定モデルを構築するための機械学習法の一つとして、RF[8]の回帰分析を用いる。RF[8]は、2001年にBreimanによって提案された手法であり、回帰木とアンサンブル学習の組み合わせたアルゴリズムである。

RFのパラメータを決定するため、グリッドサーチを行う。グリッドサーチとは、機械学習法のパラメータをあらゆる組み合わせで試すことで、モデルの推定精度向上を図る手法である。RFでは、木の数および木の深さをグリッドサーチの対象とし、検証を行う。モデルの評価方法は、3-fold交差検証を行い、得られた決定係数の平均値を比較することで行う。

まず木の数を決定するため、木の深さ = 7に固定し、木の数を変化させ解析を行った。その結果を図9に示す。この図から、決定係数は木の数が15から安定することが確認された。

次に木の深さを決定するため、木の数 = 15に固定し、木の深さを変化させ解析を行った。その結果を図10に示す。この図から、決定係数は木の深さが10から安定することが確認された。

以上の結果から、RFでモデルを構築する際のパラメータは、木の深さ = 10、木の数 = 15と決定した。

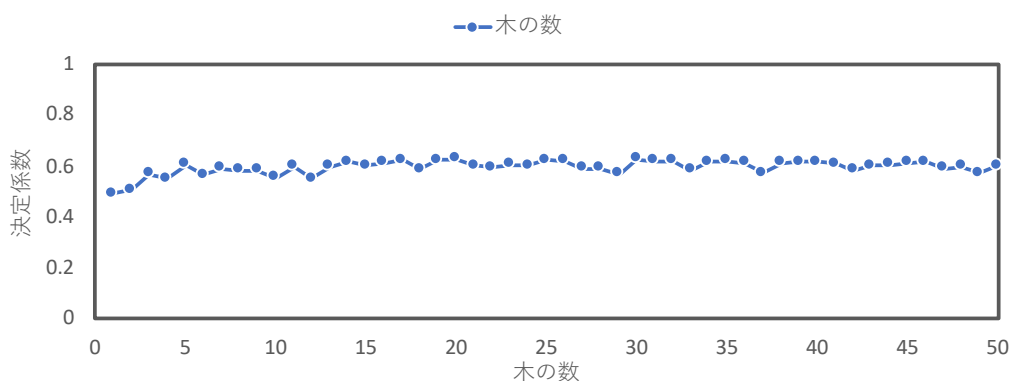


図9 RFの木の数と推定精度

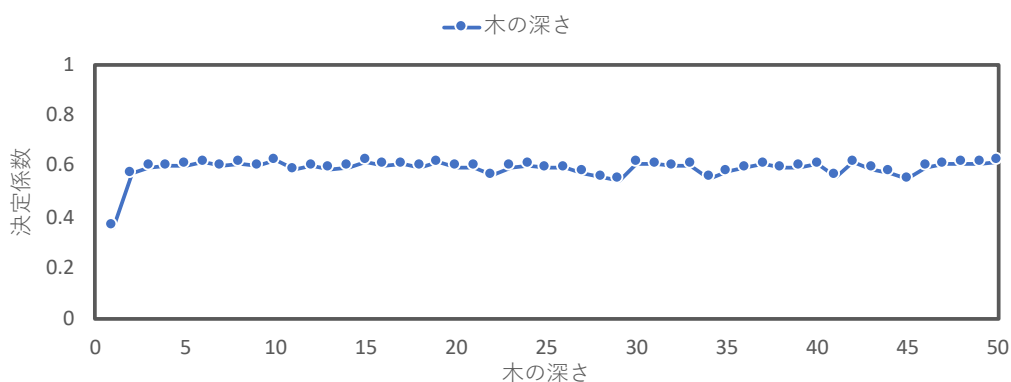


図10 RFの木の深さと推定精度

5.2 サポートベクター回帰のパラメータ

賃料推定モデルを構築するための機械学習法の一つとしてSVR[25]を用いる。SVMは、教師あり学習により2クラスの分類が可能なモデルを構築できる手法であり、SVRはSVMを回帰分析へ応用した手法である。

本研究では、SVRのカーネルとして、ガウシアンカーネルを用いることとした。ガウシアンカーネルを使用することにより、非線形回帰も可能となる。SVRでは、 C 、 γ 、 ε の3つのパラメータを、グリッドサーチの対象とする。モデル

の評価方法は、RFと同様に3-fold交差検証を行い、決定係数の平均値を比較することで行う。

まず γ を決定するため、 $\varepsilon = 0.01$ に固定し、 C 、 γ をそれぞれ変化させる。その結果を図11に示す。この図から、 $C = 10^8$ 、 $\gamma = 0.01$ で決定係数が最も高い値をとることが確認できた。

次に、 ε の値を決定するため、 $\gamma = 0.01$ と固定し、 $C > 10^6$ の範囲で変化させた。その結果を図12に示す。この結果から、高く安定した決定係数が得られているのは、 $C = 10^8$ であることがわかる。また、 $C = 10^8$ で最も決定係数が高くなったのは、 $\varepsilon = 10^{-6}$ の場合であった。

以上の結果から、SVRでモデルを構築する際のパラメータは、 $C = 10^8$ 、 $\gamma = 0.01$ 、 $\varepsilon = 10^{-6}$ と決定した。

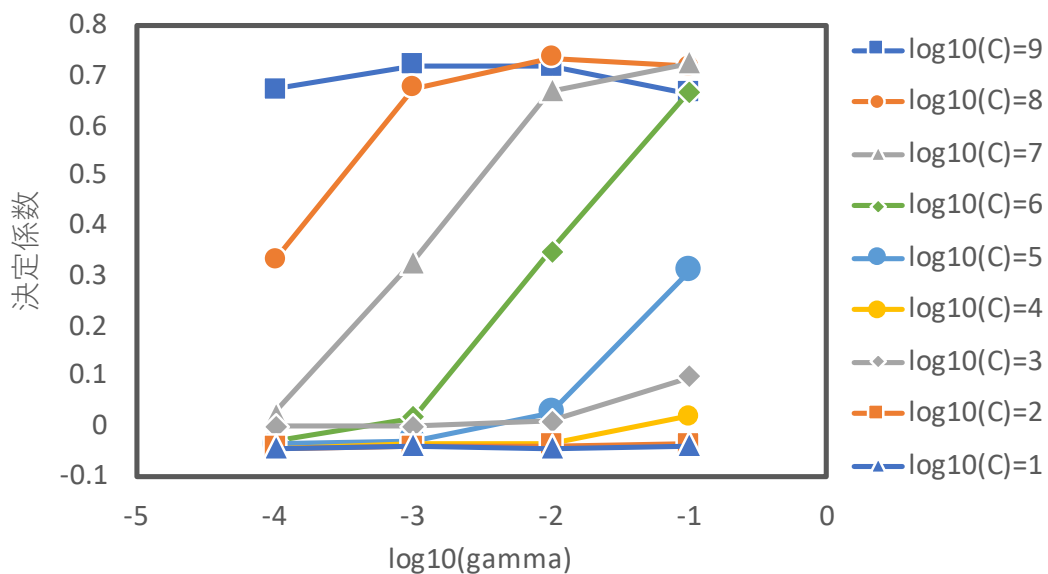


図11 SVRの γ と推定精度

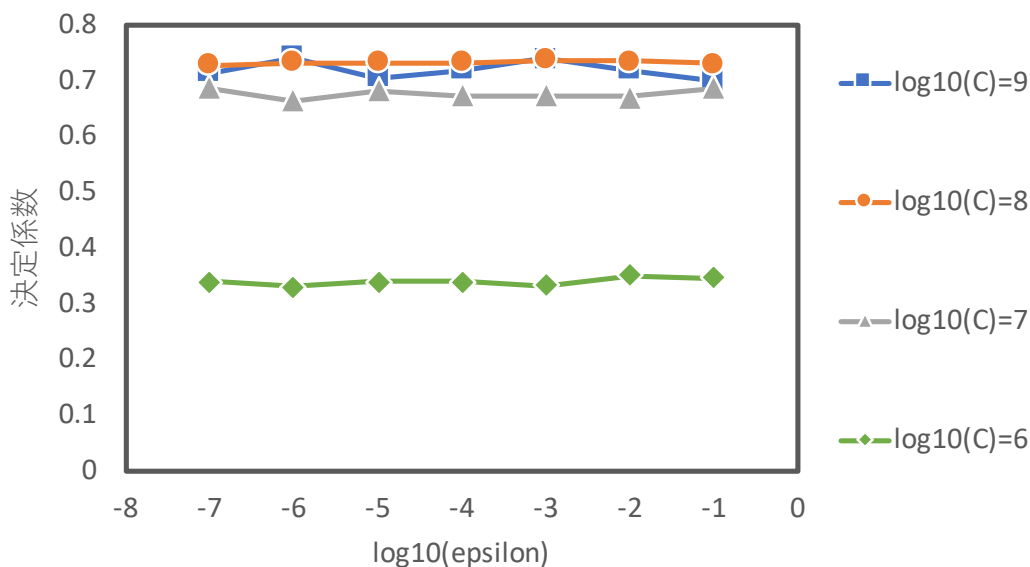


図 12 SVR の ϵ と推定精度

5.3 Doc2Vec のパラメータ

Doc2Vec でキャッチコピーを学習する際の前処理として、キャッチコピーの形態素解析を行った。本研究では、形態素解析エンジンとして、日本語に対応している MeCab[26] を用いた。MeCab[26] で使用可能な「mecab-ipadic-NEologd」は、週に 2 回以上更新される辞書であり、新語や固有表現に強いという特徴がある。本研究では、形態素解析の際の辞書として「mecab-ipadic-NEologd」を用い、キャッチコピーを学習する際のパラメータとしては、出現回数の低い単語を除く min-count と、学習対象の単語に対して考慮する前後の単語数である window サイズを、グリッドサーチの対象とした。

本項では、坪数、居抜き、階数、駅徒歩時間、通行量×視認性、坪数×平均坪単価の特徴量にベクトル化したキャッチコピーを用い、モデルを構築する。キャッチコピーの学習の際には、キャッチコピーから記号を除き、ベクトル数を 20 に固定した。また、モデルの構築には LR, RF, SVR, XGBR を用いた。

min-count を決定するため、window=5 に固定し、解析を行う。その結果を図

13に示す。この図から、LR、SVRにおいて、min-countによる決定係数の大きな差は得られなかったが、最も推定精度が高くなったのは、min-count=30の場合であった。

windowサイズを決定するため、min-count=30に固定し、解析を行った。その結果を図14に示す。この図から、LR、SVRで結果が安定することが確認できる。これら二つのモデルにおいて、windowサイズによる決定係数の大きな差は得られなかったが、最も推定精度が高くなったのは、window=5の場合であった。

以上の結果から、Doc2Vecによるキャッチコピーの学習を行う際のパラメータは、min-count=30、window=5と決定した。

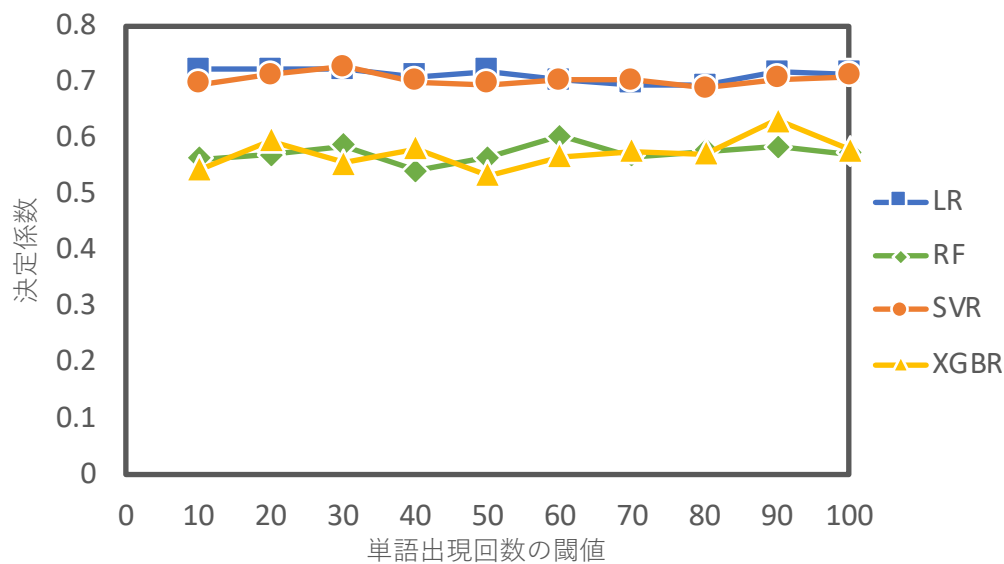


図13 単語出現回数の閾値と推定精度

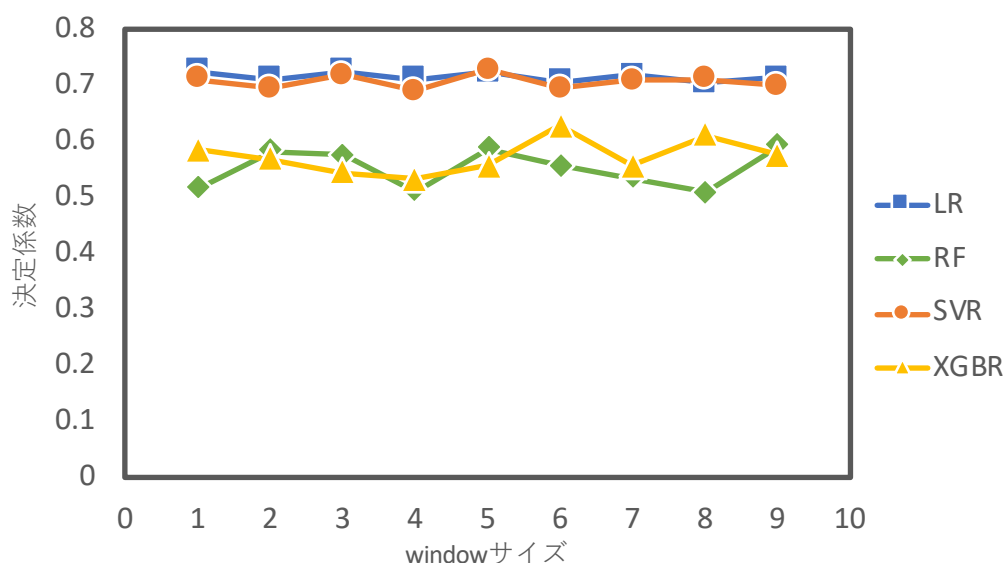


図 14 window サイズと推定精度

5.4 Word2Vec のパラメータ

キャッチコピーを Word2Vec により学習するための前処理として、形態素解析エンジンとして MeCab[26] を使い、辞書としては「mecab-ipadic-NEologd」を用いる。

Doc2Vec では、キャッチコピーを文書として捉え、学習を行った。しかしながら、本研究で用いるキャッチコピーは、駅名や道路名、商店街の名称などの固有名詞が多数含まれており、特に商店街の名称などのローカルな単語は、形態素解析の際の辞書に含まれていないことが多い。そのため、形態素解析の時点で誤解析が多数発生する可能性がある。また、こういった問題は、形態素解析の際の辞書に追加すれば解決するが、全ての地名、名称を追加することは現実的ではない。また、誤った形態素解析のまま Word2Vec により学習すると、異なる意味で使用していた単語が同じ意味で学習される可能性や、同じ意味で使用していた単語が異なる意味で学習される可能性がある。

以上のことから、Word2Vec ではキャッチコピーを学習するのではなく、インターネット上で公開されている、Wikipedia により学習済みの Word2Vec モデルを

使用し、単語のクラスタリングを行うこととした。これにより、形態素解析の誤りによる単語学習モデルへの影響はなく、単語のクラスタリングの際の影響もない。本研究にて用いた Word2Vec モデルのパラメータは、ベクトル数 = 50, window サイズ = 10, min-count = 5 である。

クラスタリング数を決定するため、本研究では X-means 法 [27] を用いる。X-means 法 [27] とは、2000 年に Pelleg らが提案した手法で、k-means 法の最適なクラスタ数 k を自動で決定することができるアルゴリズムである。キャッチコピーに含まれるノイズを除去するため、名詞、形容詞、および動詞を抽出し、X-means 法によりクラスタリングを行なった。その結果、クラスタリング数は 53 となる結果が得られた。これら 53 のクラスタを新たな特徴量として追加し、モデルを構築する。

6. 検証結果

本章では、3章の仮説の検証結果を示す。

6.1 データ数を追加した場合の検証結果

仮説1の検証結果を、図15および図16に示す。両図とも横軸にデータ数をとっており、縦軸に決定係数と平均二乗誤差をそれぞれ示している。また凡例は、モデルの構築に用いた機械学習法を示している。

これらの図を機械学習法で比較する。4つの機械学習法間に推定の差があるかを検証する、対応のある一要因の分散分析を行なったところ、LR-XGBR間で1%、SVR-XGBR間で5%有意であることが得られた。また図15および図16を確認すると、LRおよびSVRにおいて、決定係数が高く平均二乗誤差が低くなる結果が得られた。このことから、以降の推定モデル構築の際には、LRおよびSVRを用いることとする。

これらの図をデータ数で比較する。データ数が100件の場合の決定係数は、LRでは0.654、SVRでは0.662となる結果が得られた。一方で、データ数が483件の場合の決定係数は、LRでは0.729、SVRでは0.728となる結果が得られ、LRでは11.5%、SVRでは9.9%の推定精度向上が得られた。また平均二乗誤差においては、データ数が100件の場合、LRでは93954、SVRでは92965の結果に対し、データ数が483件の場合、LRでは86064、SVRでは86343の結果が得られ、LRでは8.4%、SVRでは7.1%の低下となった。以上の結果を表3にまとめる。

以降の仮説の検証では、仮説1で構築したモデルを推定精度の比較対象とする。つまり、LRでのモデル構築の場合、決定係数は0.729以上、平均二乗誤差は86064以下、SVRでのモデル構築の場合、決定係数は0.728以上、平均二乗誤差は86343以下であれば、推定精度が向上したことを示す。また、以降で示す推定精度の相対誤差は、仮説1の推定精度を基準として算出する。

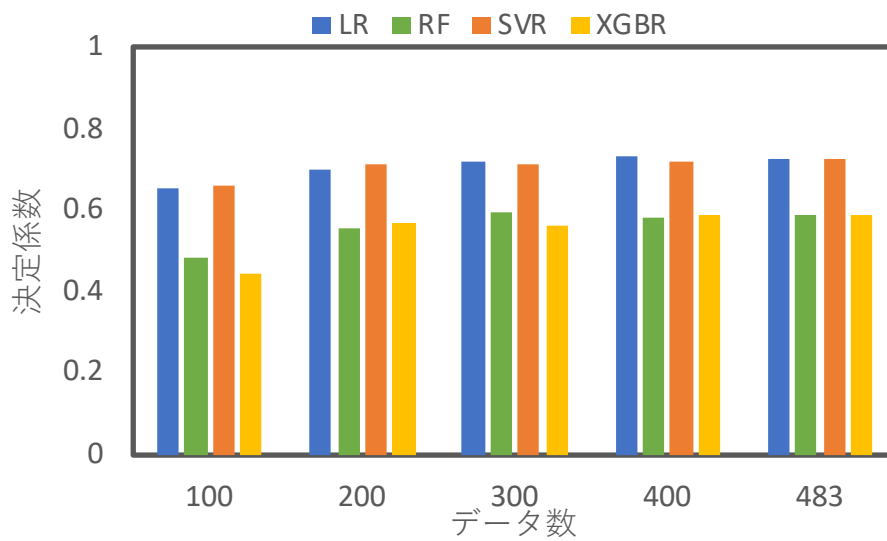


図 15 データ数と決定係数

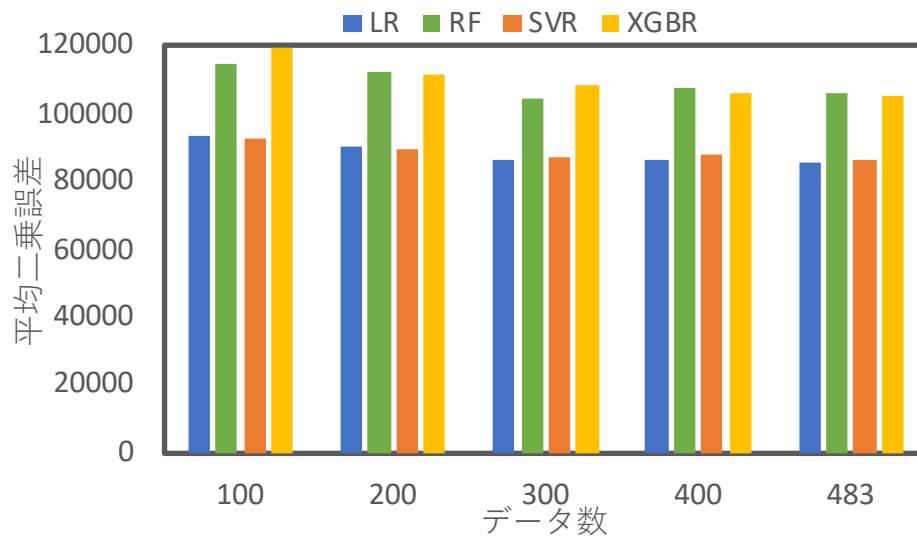


図 16 データ数と平均二乗誤差

表 3 仮説 1 検証結果

機械学習法	決定係数			平均二乗誤差		
	100 件	483 件	相対誤差	100 件	483 件	相対誤差
LR	0.654	0.729	+11.5 %	93954	86064	-8.4 %
SVR	0.662	0.728	+9.9 %	92965	86343	-7.1 %

6.2 新たな特徴量を追加した場合の検証結果

仮説 2 の検証結果を、図 17 および図 18 に示す。両図とも、横軸に追加した特徴量をとっており、縦軸はそれぞれ、仮説 1 の推定精度との相対誤差である決定係数上昇率と平均二乗誤差低下率をとっている。両図とも、正の値をとる場合に推定精度が向上したことを示している。

これらの図を特徴量で比較すると、年および地上階は、学習モデルによらず推定精度が向上することがわかる。また、地下階、p-居抜き、p-駅徒歩時は、学習モデルによらず推定精度が低下することがわかる。現時点で得られている具体的な推定精度を表 4 にまとめる。

次に、特徴量を順に追加するのではなく、キャッチコピーを除く既存・新規の 16 の特徴量から、RFE により重要でない特徴量を消去し、重要な特徴量のみで推定モデルの構築を行った。特徴量は最低 3 つを抽出し、最大で全特徴量を用いた。

RFE により抽出された特徴量を表 5 に示す。表 5 では、○で示した特徴量が抽出されたことを示している。また、これらの特徴量を用いモデルを構築した結果を、図 19 および図 20 に示す。

表 5 を確認すると、賃料推定に重要である上位 3 つの特徴量は、坪数、通行量×視認性、坪数×平均坪単価であることがわかる。それ以降は、p-合計、駅徒歩時間、地上階、年の順に、賃料推定に重要な特徴量が抽出された。また図 19 および図 20 を確認すると、LR では特徴量 12、SVR では特徴量 11 で、推定精度が最も高くなる結果が得られた。得られた推定精度を表 6 にてまとめる。



図 17 追加特徴量と決定係数上昇率

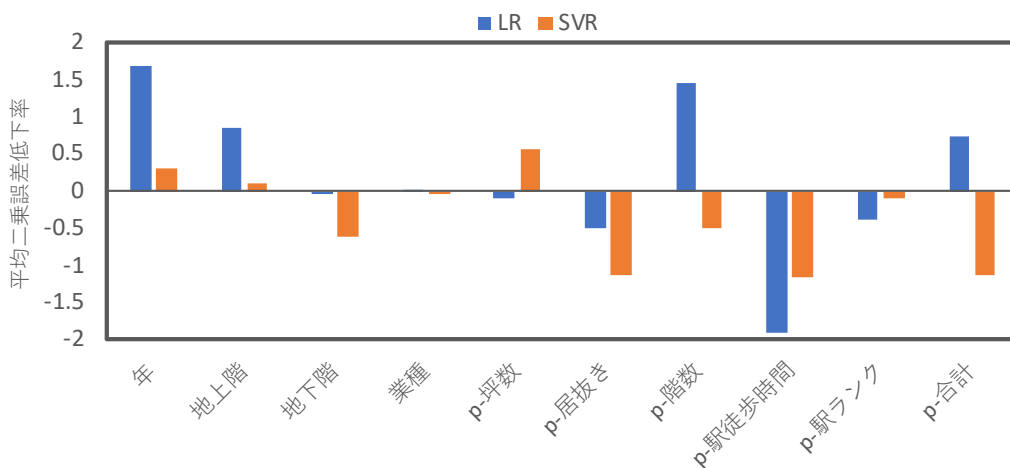


図 18 追加特徴量と平均二乗誤差低下率

表 4 新たな特徴量を追加した場合の推定精度および仮説 1 との推定精度比較

機械学習法	特徴量数	追加特徴量	決定係数		平均二乗誤差	
LR	7	年	0.738	+1.2 %	84599	-1.7 %
SVR	7	p-坪数	0.731	+0.4 %	85845	-0.6 %

表 5 賃料推定に重要な特徴量の抽出結果

特徴量	特徴量数													
	3	4	5	6	7	8	9	10	11	12	13	14	15	16
坪数	○	○	○	○	○	○	○	○	○	○	○	○	○	○
居抜き									○	○	○	○	○	○
階数							○	○	○	○	○	○	○	○
駅徒歩時間			○	○	○	○	○	○	○	○	○	○	○	○
通行量×視認性	○	○	○	○	○	○	○	○	○	○	○	○	○	○
坪数×平均坪単価	○	○	○	○	○	○	○	○	○	○	○	○	○	○
年					○	○	○	○	○	○	○	○	○	○
地上階				○	○	○	○	○	○	○	○	○	○	○
地下階						○	○	○	○	○	○	○	○	○
業種								○			○	○		○
p-坪数										○	○	○	○	○
p-居抜き												○	○	○
p-階数									○				○	○
p-駅徒歩時間													○	○
p-駅ランク										○	○	○	○	○
p-合計		○	○	○	○	○	○	○	○	○	○	○	○	○

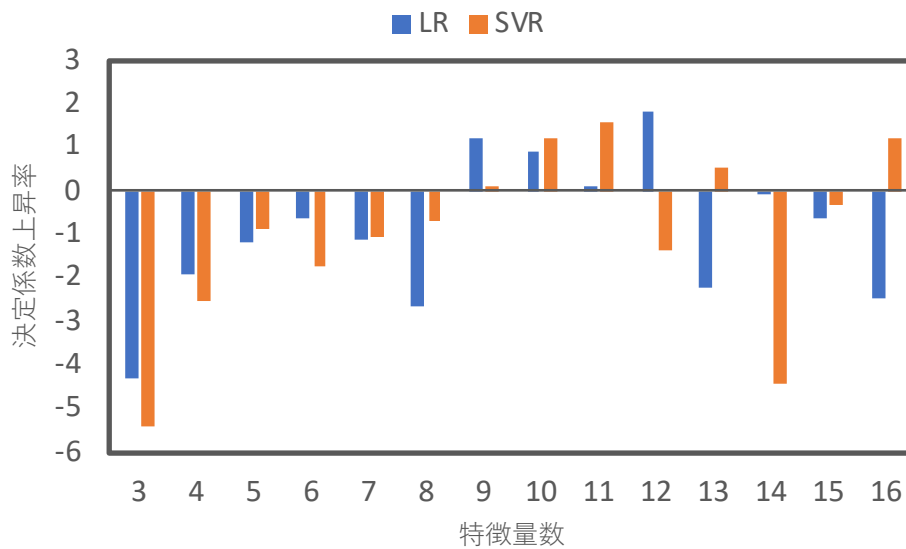


図 19 抽出した特徴量を用いた場合の決定係数上昇率

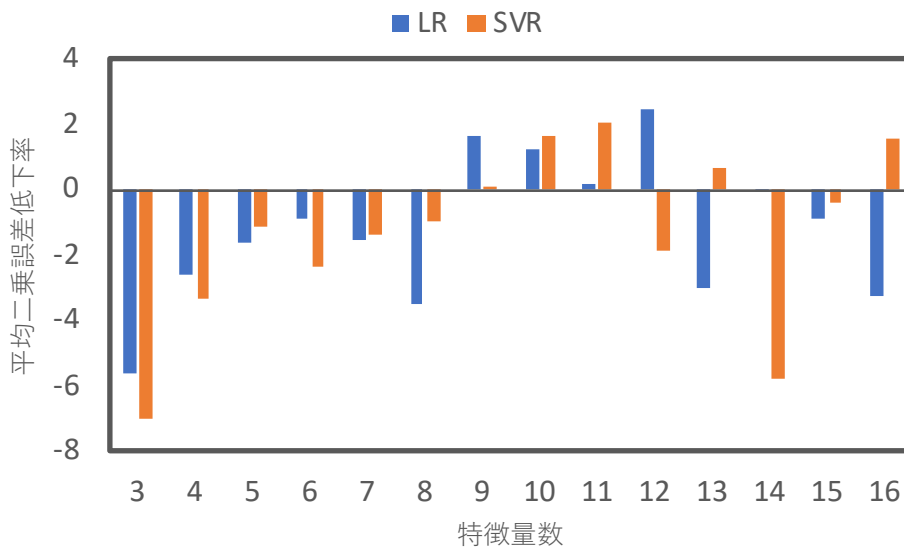


図 20 抽出した特徴量を用いた場合の平均二乗誤差低下率

表 6 仮説 2 の検証結果および仮説 1 との推定精度比較

機械学習法	特徴量数	決定係数		平均二乗誤差	
LR	12	0.743	+1.8 %	83917	-2.5 %
SVR	11	0.739	+1.6 %	84516	-2.1 %

6.3 特徴量の掛け合わせを考慮した場合の検証結果

仮説 3 の検証結果を、図 21 から図 27 に示す。いずれも横軸に追加した特徴量、縦軸に推定精度を示しており、左図が決定係数上昇率、右図が平均二乗誤差低下率を示している。これらの図から、いずれの特徴量も決定係数は最大で 1 % 程度の上昇にとどまっていることが確認でき、また掛け合わせる特徴量による推定精度の変化の傾向は確認できなかった。また、LR では 1/階数× 1/駅徒歩時間を追加した場合、SVR では居抜き× 1/階数を追加した場合に、最も推定精度が高くなることが確認された。現時点で得られている結果を表 7 に示す。

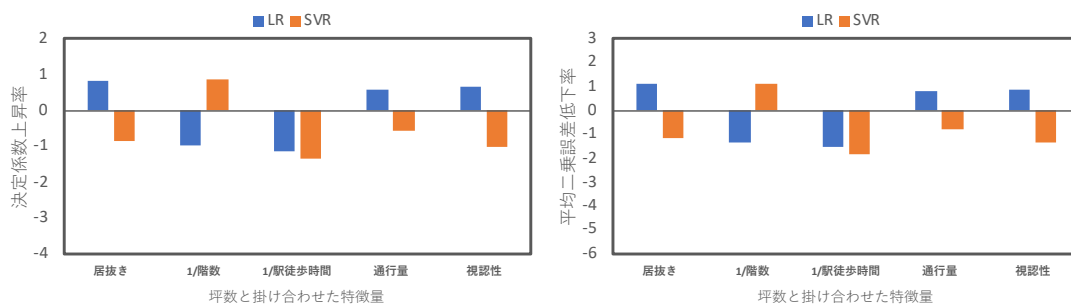


図 21 「坪数×他の要素」を追加した場合の推定精度比較

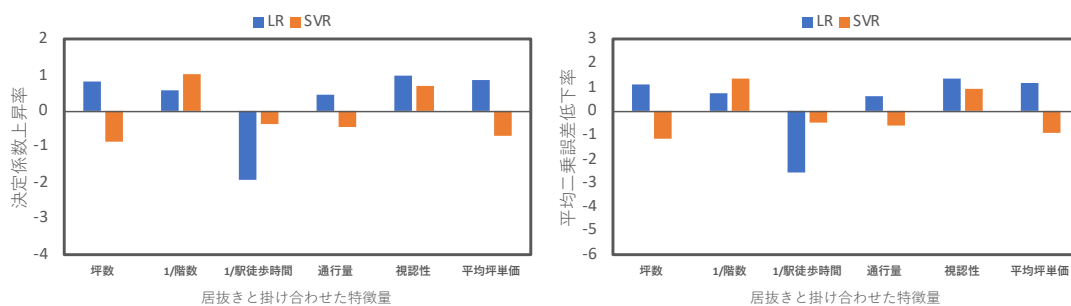


図 22 「居抜き×他の要素」を追加した場合の推定精度比較

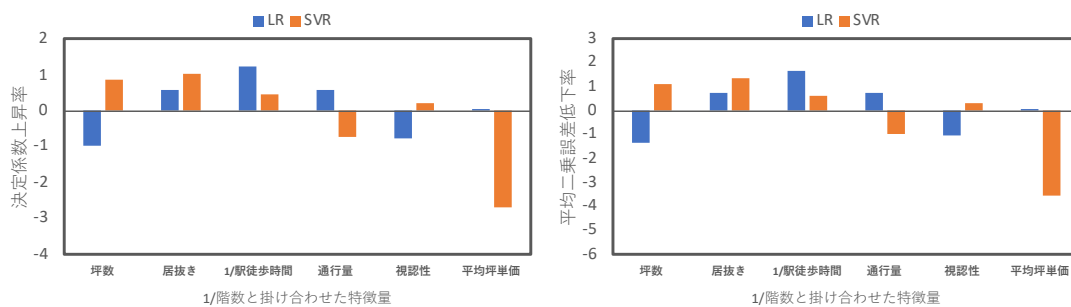


図 23 「1/階数×他の要素」を追加した場合の推定精度比較

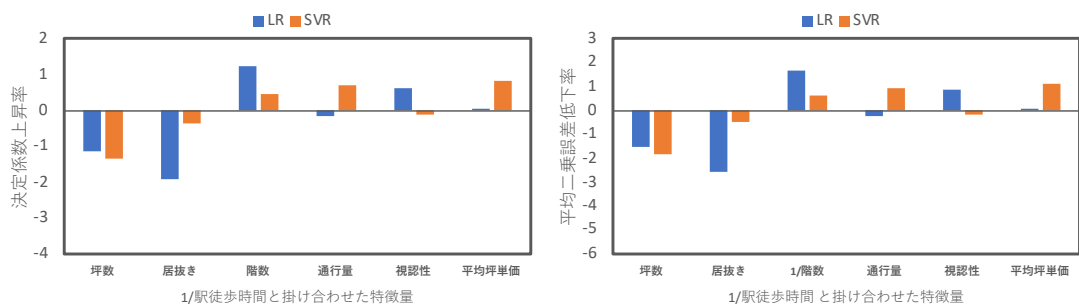


図 24 「1/駅徒歩時間×他の要素」を追加した場合の推定精度比較

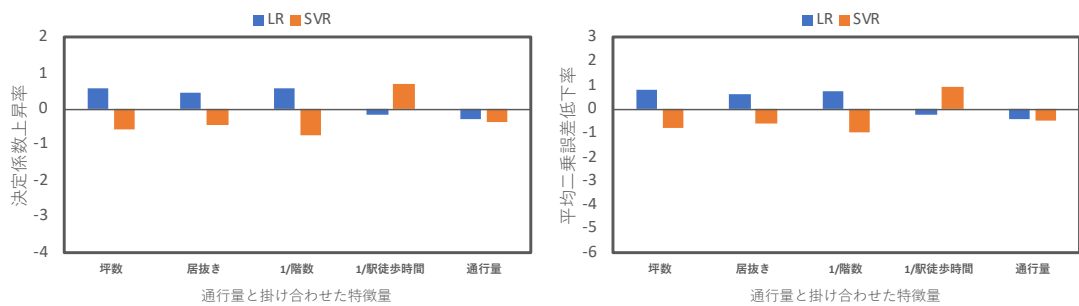


図 25 「通行量×他の要素」を追加した場合の推定精度比較

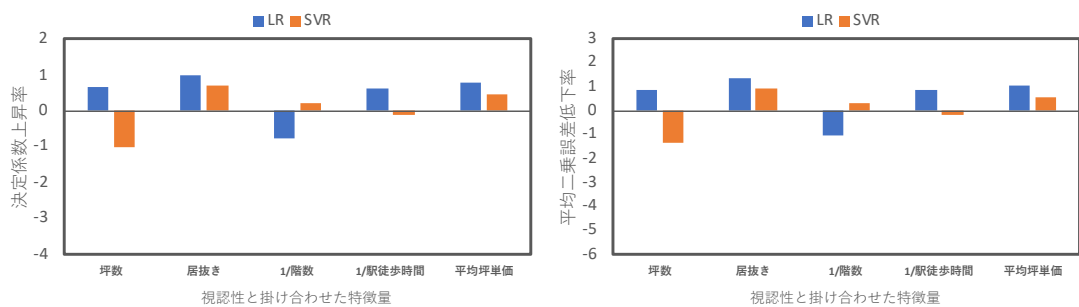


図 26 「視認性×他の要素」を追加した場合の推定精度比較

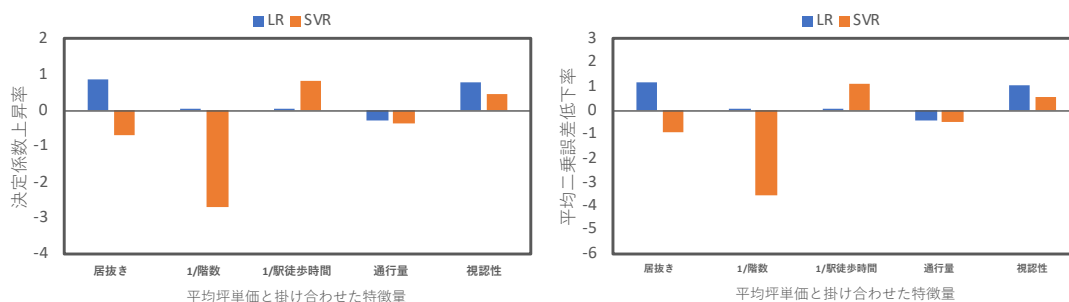


図 27 「平均坪単価×他の要素」を追加した場合の推定精度比較

表 7 掛け合わせた特徴量を追加した場合の推定精度および仮説 1 との推定精度比較

機械学習法	特徴量数	追加特徴量	決定係数		平均二乗誤差	
LR	7	1/階数× 1/駅徒歩時間	0.738	+1.2 %	84625	-1.7 %
SVR	7	居抜き× 1/階数	0.735	+1.0 %	85147	-1.4 %

次に、既存の特徴量および本仮説にて追加した新規特徴量から、RFEにより重要でない特徴量を消去し、抽出された特徴量を用いてモデルを構築した。その結果を図 28 および図 29 に示す。

これらの図は、横軸に抽出した特徴量数、縦軸に決定係数上昇率と平均二乗誤差低下率をそれぞれ示している。どちらの図も正の値をとれば、仮説 1 で構築したモデルよりも推定精度が向上したことを示す。また凡例は、モデル構築を行った機械学習法を示している。

これらの結果から、3 から 15 の特徴量を抽出した場合では、推定精度が低下する結果が得られ、仮説 1 を上回る推定精度は得られなかった。

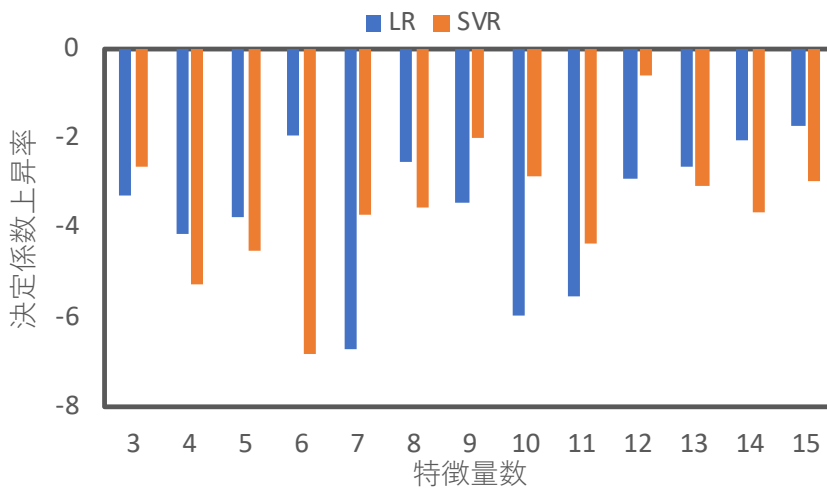


図 28 掛け合わせた特徴量によるモデル構築と決定係数上昇率

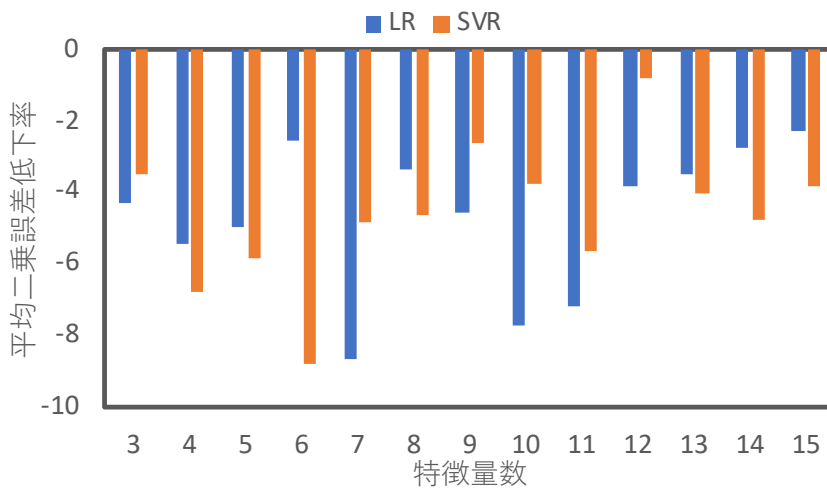


図 29 掛け合わせた特徴量によるモデル構築と平均二乗誤差低下率

6.4 キャッチコピーの文脈と曖昧さを考慮した場合の検証結果

本研究では仮説 4 にて、「キャッチコピーの文脈や曖昧さを考慮することにより推定精度は向上する」という仮説を立て、仮説 4-1 では Doc2Vec, 仮説 4-2 では

Word2Vec を用いてキャッチコピーを解析し、モデルの構築を行なった。以下にそれぞれの結果について述べる。

6.4.1 Doc2Vec 技術による解析

本節では、仮説 4-1 の検証結果について述べる。仮説 4-1 の検証結果を図 30 および図 31 に示す。これらの図は、横軸にキャッチコピーのベクトル数を取っており、縦軸にそれぞれ決定係数と平均二乗誤差を取っている。またノイズを除去するため、キャッチコピー中の括弧や句読点といった記号は除去した上でモデルを構築した。

これらの図をベクトル数で比較すると、ベクトル数の値が大きくなるにつれ、推定精度が低下する結果が得られた。このことから、ベクトル数を 1~5 と限定し、推定精度の比較を行う。

仮説 1 との推定精度の比較を、図 32~図 35 に示す。全ての図において、横軸にキャッチコピーのベクトル数、縦軸にそれぞれ決定係数上昇率と平均二乗誤差低下率をとっており、正の値をとれば、仮説 1 にて構築したモデルよりも高い推定精度が得られたことを示している。図 32 および図 33 は LR、図 34 および図 35 は SVR でモデルを構築しており、また凡例は、キャッチコピーから抽出した品詞を示している。キャッチコピーからの品詞の抽出は、「が」や「も」などの助詞、「そして」や「しかし」といった接続詞などの、物件の特徴を表さないと考えられる単語を、ノイズとして除去するために行った。

図 32 および図 33 を確認する。これらの図は、LR によりモデルを構築した場合の結果である。これらの図から、記号のみを除き、キャッチコピーを 2次元にベクトル化した際に、最も推定精度が高くなることが得られた。また、2次元にベクトル化した際は、品詞の抽出方法によらず、仮説 1 の推定精度を上回る結果となった。

図 34 および図 35 を確認する。これらの図は、SVR によりモデルを構築した場合の結果である。この結果から、名詞、形容詞、動詞を抽出し、キャッチコピーを 4次元にベクトル化した際に、最も推定精度が高くなることが得られた。

以上の結果を、表 8 にまとめる。

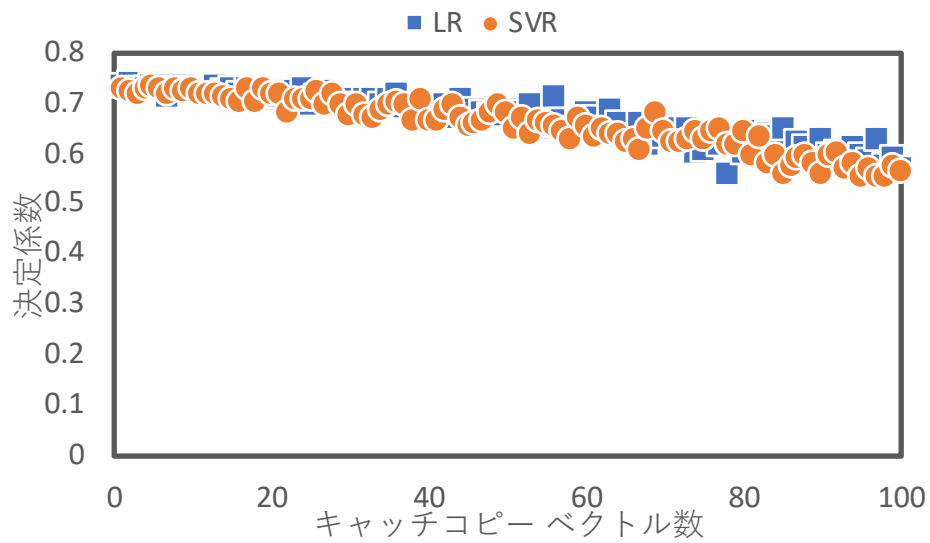


図 30 Doc2Vec によるキャッチコピーのベクトル数と決定係数

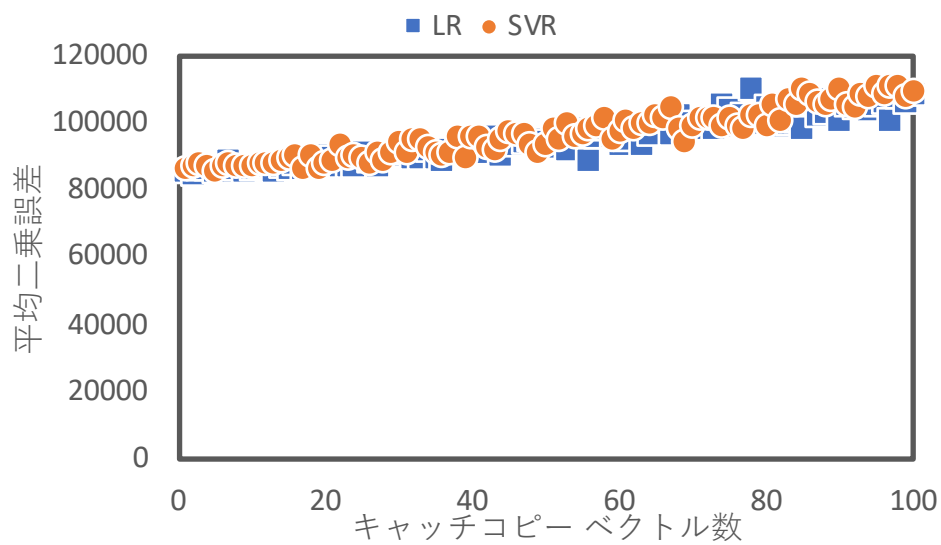


図 31 Doc2Vec によるキャッチコピーのベクトル数と平均二乗誤差

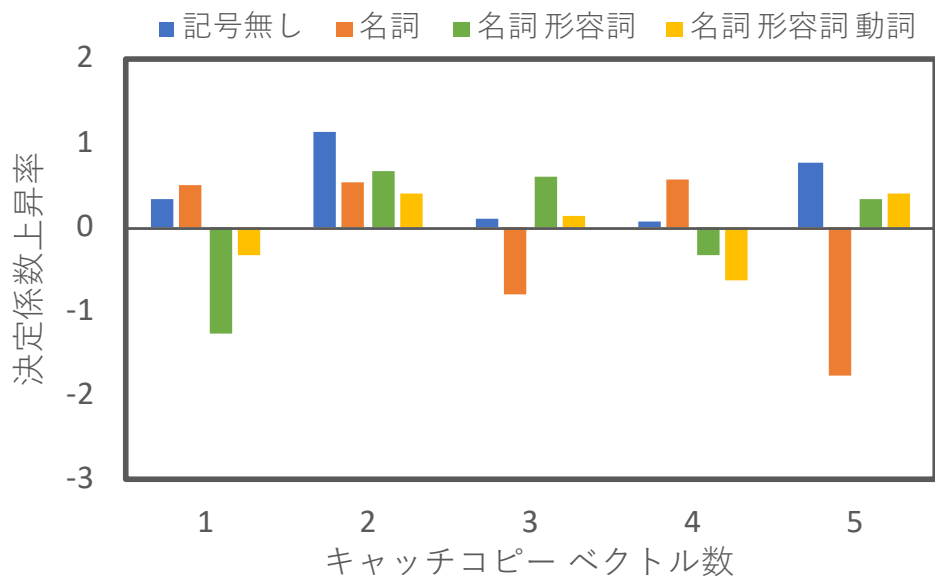


図 32 抽出した品詞と決定係数上昇率 (LR)

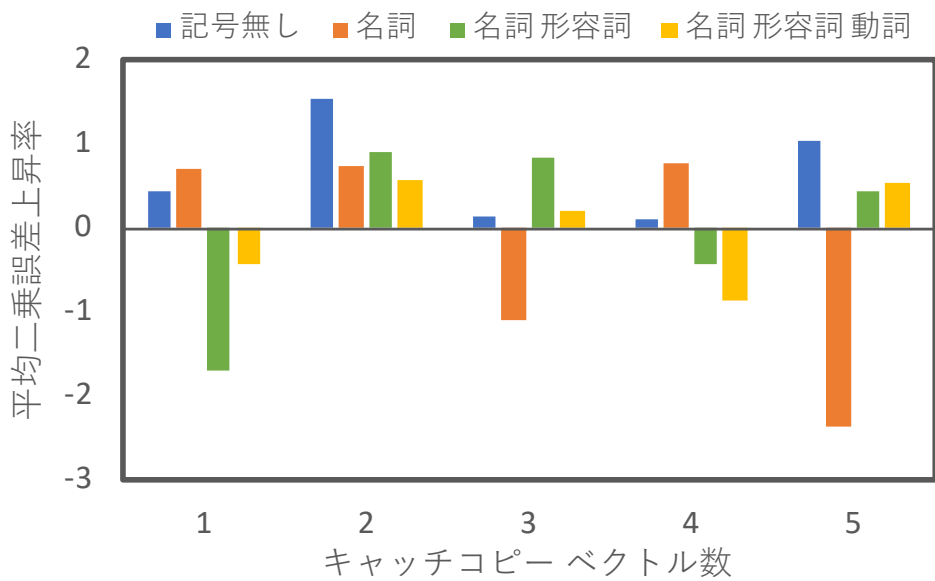


図 33 抽出した品詞と平均二乗誤差低下率 (LR)

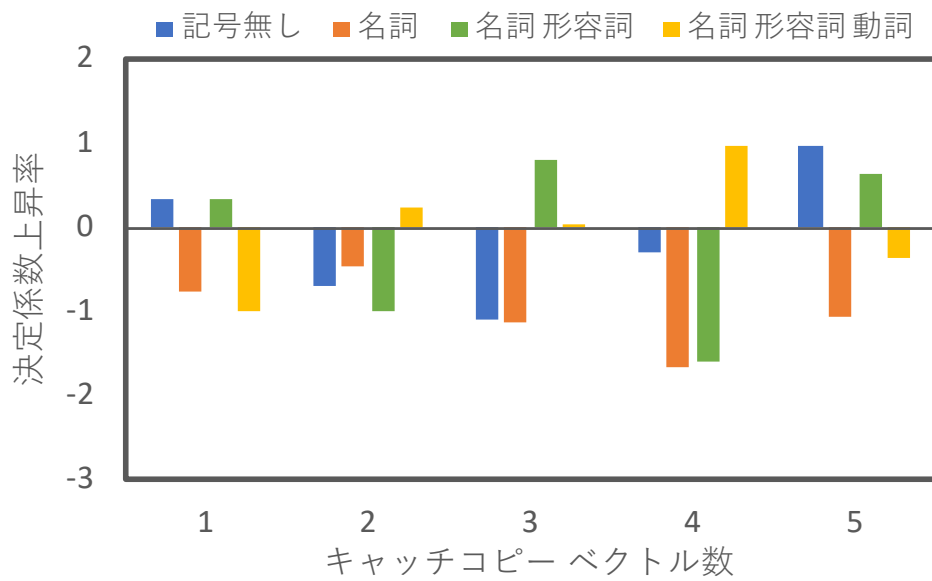


図 34 抽出した品詞と決定係数上昇率 (SVR)

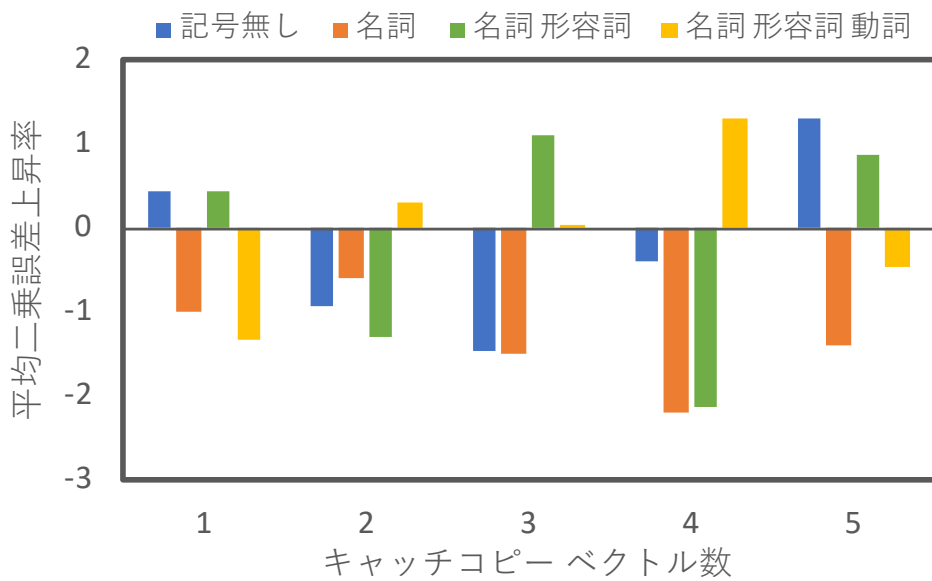


図 35 抽出した品詞と平均二乗誤差低下率 (SVR)

表 8 仮説 4-1 の検証結果および仮説 1 との推定精度比較

機械学習法	ベクトル数	抽出品詞	決定係数		平均二乗誤差	
LR	2	記号のみ除去	0.738	+1.1 %	84728	-1.6 %
SVR	4	名詞 形容詞 動詞	0.735	+1.0 %	85204	-1.3 %

6.4.2 Word2Vec 技術による解析

本節では、仮説 4-2 の検証結果について述べる。5 章の予備解析により、クラスタリング数は 53 と決定した。クラスタ中の単語を表 9 に示す。表中の+や-は、例えばクラスタ 0 を特徴量として追加し、LR にてモデルを構築した場合は、仮説 1 よりも推定精度が低下、SVR にてモデルを構築した場合は、仮説 1 よりも推定精度が向上したことを示している。

表 9 クラスタ内の単語と推定精度相対誤差の正負

クラスタ	単語数	単語の例	LR	SVR
0	14	採光 物件 駅前 交差点 交通量 大通り 明るい	-	+
1	97	フロアー 大門 アメ横 眺望 相互 イベント	+	-
2	64	専念 賑わい 観光 活気 アットホーム 街並み	+	+
3	105	最上階 吹き抜け 街道 名店街 高田馬場	-	+
4	83	滅多 空き 高層マンション 戸建て カジュアル	+	-
5	17	京葉道路 広告効果 利便性 日当たり 広告	-	-
6	112	鉄板焼き たこ焼 マクドナルド サミット	+	-
7	121	晴海通り 日光街道 歓楽街 団子坂 多摩堤通り	+	+
8	71	農大 花園 芝公園 区役所通り 歩道 自由が丘	+	-
9	120	商業地域 会館 伊勢丹 ゴールデン街	+	-
10	140	産業道路 初台 茅場街 葛西駅 中町	+	-
11	55	経年 旺盛 ラッシュ ランチライム 多忙 競合	+	+
12	110	跡地 ウッドデッキ 洋菓子店 洋食店 デザイン性	-	-
13	114	新装 新設 改定 拡張 整う 施す 螺旋階段 庭	-	-
14	86	鶴見駅 護国寺駅 東大前駅 千川駅 藤沢駅	-	+
15	95	明治神宮前 門前仲町駅 横浜市 若林駅	-	+
16	55	通路 コインパーキング 横断歩道 管理 出入り口	+	-
17	58	プレミアム イタリアンバー べーカリー	-	-

クラス	単語数	単語	LR	SVR
18	55	浅草通り アーケード 道玄坂 駅前通り	+	-
19	85	東池袋駅 落合駅 西新宿駅 東新宿駅	-	-
20	78	大門駅 浅草線 駅チカ 池上通り 表参道ヒルズ	+	-
21	21	フリー 自由自在 エレベーター 倉庫 シャッター	-	+
22	8	分割 メゾネット タイプ フリーレント	+	+
23	67	ビジネス デリバリー ビジネス街 五反田 町	-	+
24	65	集客力 拡大 独立 密着 きれいな 大手企業	+	+
25	36	大学 学生 密集 集まる 黒字 昼夜 平日 人気	-	-
26	12	路面店 路面 掲載 解放 中華料理店 らーめん	+	-
27	62	高級感 魅力 おしゃれ 神楽坂 西麻布	+	+
28	50	格安 解約 路地 必見 オススメ 老舗	-	-
29	11	残置 エアコン トイレ カラオケ	+	+
30	43	バス停 貴重 公園 便利 備品 中心地 直結	+	+
31	56	要町駅 八丁堀駅 高円寺駅 有楽町駅 早稲田駅	-	+
32	46	美容院 マッサージ 塾 自由通り スタジオ タイル	+	-
33	36	バス通り 繁華街 昭和通り 希少 一等地 道路	-	-
34	17	アクセス 徒歩圏内 駐車 交通 駐車場 スペース	-	+
35	9	面する 沿い 商店街 通り 1階ビル	+	-
36	49	浅草線 東急東横線 好立地 明治通り 丸ノ内線	-	+
37	30	新宿駅 神田駅 大塚駅 山手線 恵比寿駅 銀座線	-	-
38	6	新橋駅 六本木駅 スナック クラブ	-	+
39	14	徒歩 1分 2分 3分 4分 5分 6分 7分 JR	+	+
40	14	出店 チェーン オフィス エリア 大手	+	-
41	18	大型 集客 通行量 オフィス街 盛業 充実	+	-
42	18	ABC 会員 需要 限定 雰囲気 設備	-	-
43	4	開業 新規 営業	-	+
44	3	飲食店 駅 至近	-	+
45	12	無償 美麗 ダイニング 居抜き バー レストラン	+	-
46	9	角地 前面 宣伝 目立つ ガラス張り 効果	+	+
47	7	広い 間口 看板 視認性 抜群 良好	+	+
48	3	有 専用 階段	+	-
49	7	2 3 4 利用 利用可能 路線	+	-
50	15	相談 スケルトン 即時 新築 飲食 業種 条件	+	-
51	10	B1 2階 3階 一棟 付き	+	-
52	3	1 一括 貸し	+	-

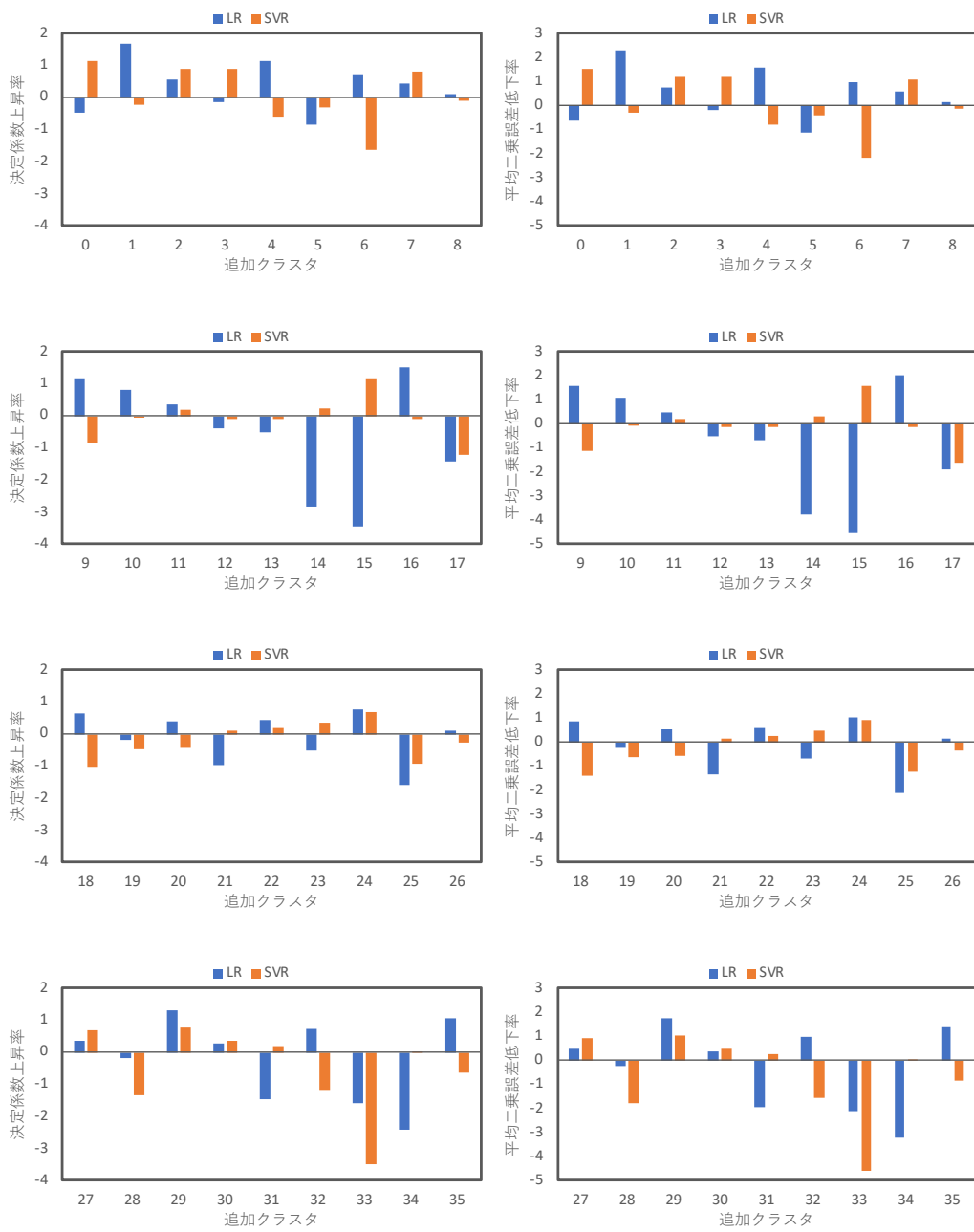


図 36 追加クラスタ (0~35) と仮説 1 との推定精度比較

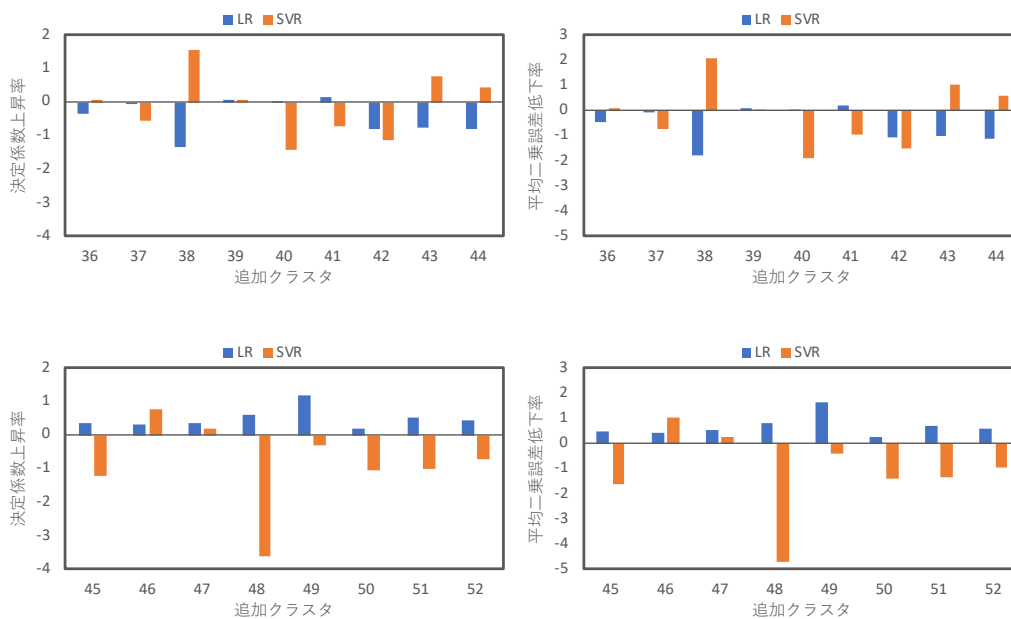


図 37 追加クラスタ（36～52）と仮説 1 との推定精度比較

表 9 を確認する。LR および SVR の両方のモデルで推定精度が向上したのは、クラスタ 2, 7, 11, 22, 24, 27, 29, 30, 39, 46, 47 であった。これらのクラスタに含まれる単語を確認すると、観光や集客力、活気といった集客が見込めるような単語や、歓楽街や中心地といった立地に関する単語、また、目立つ、看板、視認といった物件の視認性に関する単語が含まれることが確認できる。

図 36 および図 37 は、横軸に追加クラスタを、縦軸に仮説 1 との推定精度の相対誤差を示したものである。左図は決定係数上昇率、右図は平均二乗誤差低下率を示している。これらの図から、最も推定精度が高くなったのは、クラスタ 1 を追加し LR にてモデルを構築した場合、クラスタ 38 を追加し SVR にてモデルを構築した場合であった。

以上の結果は、仮説 1 で構築したモデルにクラスタを順に追加した結果であるが、仮説 1 で構築したモデルに複数のクラスタを追加することを考える。図 38 および図 39 は、仮説 1 のモデルに、RFE により抽出した 1～10 のクラスタを追加した場合の推定精度を示している。これらの図から、推定精度が最も高くなったのは、LR, SVR 共に、クラスタを 2 つ追加した場合であった。具体的な推定精

度は表 10 にまとめる。

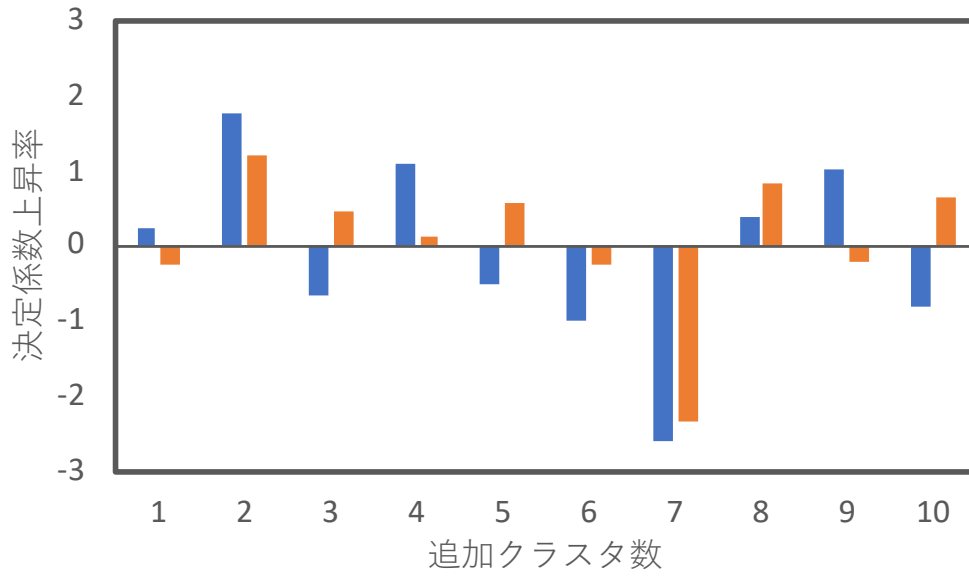


図 38 クラスターの追加数と決定係数上昇率

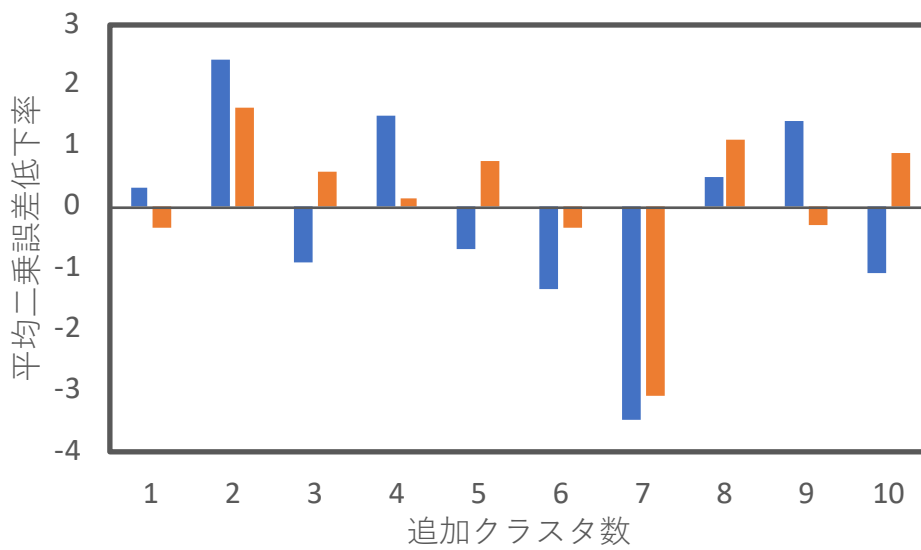


図 39 クラスターの追加数と平均二乗誤差低下率

表 10 仮説 4-2 の検証結果および仮説 1 との推定精度比較

機械学習法	追加クラス数	クラス	決定係数		平均二乗誤差	
			決定係数	変化率	平均二乗誤差	変化率
LR	2	35, 40	0.742	+1.8 %	83966	-2.4 %
SVR	2	35, 40	0.737	+1.2 %	84917	-1.7 %

6.5 仮説を組み合わせた場合の検証結果

本項では、以上で検証した仮説を組み合わせるにより、さらなる推定精度の向上を目指す。

本項のモデル構築に用いる特徴量に関して、例として仮説 2 と仮説 3 を組み合わせる場合、既存の特徴量と新規の特徴量を掛け合わせるにより、130 以上の特徴量が生成される。これらの特徴量の中には、賃料推定に重要でない特徴量が含まれる可能性がある。そのため、仮説を組み合わせた際の特徴量は、RFE により重要でない特徴量を消去し、3 から 15 の特徴量を用いてモデルを構築する。

仮説を組み合わせた場合の推定精度を、表 11 および表 12 に示す。表中の%で示した数値は、仮説 1 との相対誤差を示している。また表 11 は LR、表 12 は SVR によりモデルを構築した場合の結果である。

これらの表から、最も推定精度が高くなったのは、仮説 1、仮説 2、仮説 4-2 の組み合わせで、LR によりモデルを構築した場合であり、その決定係数は 0.752 となった。荒川 [1] らの推定モデルの決定係数は 0.738 であることから、本研究ではこれを上回る結果が得られた。

表 11 仮説の組み合わせと推定精度比較 (LR)

組み合わせた仮説	決定係数		平均二乗誤差	
1	0.729	-	86064	-
1 2	0.743	+1.8 %	83917	-2.5 %
1 3	0.717	-1.7 %	88035	+2.3 %
1 4-1	0.738	+1.1 %	84728	-1.6 %
1 4-2	0.742	+1.8 %	83966	-2.4 %
1 2 3	0.731	+0.2 %	85871	-0.2 %
1 2 4-1	0.751	+2.9 %	82609	-4.0 %
1 2 4-2	0.752	+3.1 %	82441	-4.2 %
1 3 4-1	0.728	-0.2 %	86253	+0.2 %
1 3 4-2	0.712	-2.4 %	88775	+3.2 %
1 2 3 4-1	0.741	+1.7 %	84129	-2.2 %
1 2 3 4-2	0.728	-0.2 %	86328	+0.3 %

表 12 仮説の組み合わせと推定精度比較 (SVR)

組み合わせた仮説	決定係数		平均二乗誤差	
1	0.728	-	86343	-
1 2	0.739	+1.6 %	84516	-2.1 %
1 3	0.723	-0.6 %	87021	+0.8 %
1 4-1	0.735	+1.0 %	85204	-1.3 %
1 4-2	0.737	+1.2 %	84917	-1.7 %
1 2 3	0.714	-2.0 %	88568	+2.6 %
1 2 4-1	0.743	+2.1 %	83838	-2.9 %
1 2 4-2	0.742	+2.0 %	84056	-2.6 %
1 3 4-1	0.729	+0.2 %	86070	-0.3 %
1 3 4-2	0.725	-0.4 %	86807	+0.5 %
1 2 3 4-1	0.732	+0.6 %	85634	-0.8 %
1 2 3 4-2	0.731	+0.4 %	85855	-0.6 %

7. 考察

本章では、仮説の検証結果の考察を行う。

7.1 データ数を追加した場合の考察

仮説1では、データ数を追加することで推定精度は向上するという仮説のもと、検証を行った。結果として、データ数を追加することで推定精度は向上する結果が得られた。

一方で、データ数を追加する際に必要となるのは、ベテラン営業職員による物件ごとの通行量および視認性の指標化である。これらの特徴量は、ベテラン営業職員による手作業のスコアリングであり、コストがかかるといった問題や個人差を含む問題がある。そのため、これらの特徴量の値を自動的に取得できるようになれば、これらの問題を解決することができる。

視認性を機械的に取得する手法を検討するため、荒川 [1] らはベテラン営業職員へインタビューを行い、どういった指標をセンシングすべきか議論を行っている。しかしながら、得られた指標とベテラン営業職員による主観評価が一致しないとの結果が得られている。そのため、本研究においても通行量の取得方法についてのみ考察する。

通行量を機械的に取得する手法として、GPS (Global Positioning System) がある。急速なスマートフォンの普及により、搭載されたセンサを活用する研究が広くされている。その中で、GPS を用いた歩行者の混雑度推定も行われており [28, 29]、本研究においても GPS を活用することにより、通行量の推定が可能であると考えられる。

7.2 新たな特徴量を追加した場合の考察

仮説2では、新たな特徴量の追加により推定精度は向上するという仮説のもと、検証を行った。新たな特徴量としては、年、地上階、地下階、p-坪数、p-居抜き、p-階数、p-駅徒歩時間、p-業種、p-駅ランク、p-合計を追加し、モデルを構築した。

その結果、年および地上階では、LR、SVRの学習モデルに依らず推定精度が向上した。一方で、地下階やp-居抜き、p-駅徒歩時間は学習モデルに依らず推定精度が低下する結果となった。

本研究では、荒川[1]らの課題であったデータ数の少なさを解決するため、2014年までに成約した物件に加え、新たに2019年までに成約した物件を追加し、モデルを構築した。荒川[1]らは2009年から2013年に成約した物件を扱っており、最大で4年の差があるが、本研究では2009年から2019年を対象としており、最大で10年の差がある。ここで、平均坪単価と年代の関係を示すと、図40のようになる。図中の近似式を確認すると、1年経つにつれ、平均坪単価が100円程度値上がりすることがわかる。これは物価の変化の影響が考えられ、本研究の場合、年代に最大10年の差があることから、年の特徴量が有効であったと考えられる。

また、階数に関しては図4で示したとおり、物件の階数が上がるにつれ、坪単価が低くなる結果が得られた。しかし、3章で述べたように、高層ビルの上層階にあるようなレストランでは、坪単価が高くなると考えられる。このことから、地上階の特徴量が有効であったと考えられる。

また、既存の特徴量および新規の特徴量を交え、RFEにより重要度の低い特徴量を消去し、モデルを構築したところ、LRでは坪数、居抜き、階数、駅徒歩時間、通行量×視認性、坪数×平均坪単価、年、地上階、地下階、p-坪数、p-駅ランク、p-合計の計12の特徴量を、SVRでは坪数、居抜き、階数、駅徒歩時間、通行量×視認性、坪数×平均坪単価、年、地上階、地下階、p-階数、p-合計の計11の特徴量を抽出した場合に、最も推定精度が高くなった。LR、SVR共に、既存の特徴量は全て抽出され、また年、地上階、地下階、p-合計の特徴量が抽出された。

また、p-合計の特徴量は、「p-」と表記された特徴量の和をとった指標であり、この値が大きければ、坪単価が大きい、最寄駅から近い、最寄駅が繁栄している、といった賃料が高くなる要因が多く含まれることを表す。そのため、賃料推定に重要な特徴量として抽出されたと考えられる。

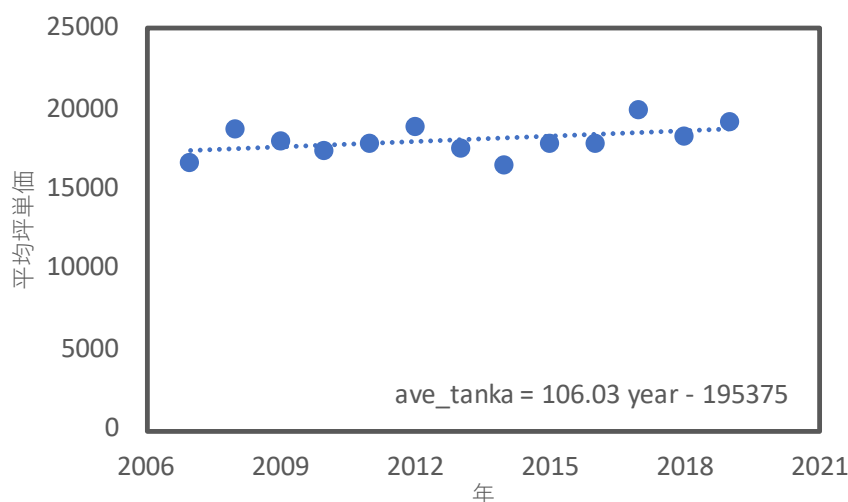


図 40 年代と平均坪単価の関係

7.3 特徴量の掛け合わせを考慮した場合の考察

仮説3では、特徴量の掛け合わせを考慮することにより推定精度は向上するという仮説のもと、検証を行った。用いた特徴量は既存の特徴量である坪数、居抜き、階数、駅徒歩時間、通行量、視認性、平均坪単価から2要素を掛け合わせた、坪数×居抜きや視認性×平均坪単価などである。これらの特徴量を仮説1にて構築したモデルに順に追加すると、いずれの特徴量も決定係数は最大で1%程度の上昇に止まり、また、どの特徴量を掛け合わせた場合に推定精度が向上するかなどの傾向は確認されなかった。さらに、これら全ての特徴量から、RFEにより賃料推定に重要でない特徴量を消去しモデルの構築を行ったところ、3から15の特徴量を抽出した場合は、仮説1の推定精度を上回る結果は得られなかった。

RFEにより抽出された特徴量を確認したところ、賃料推定に最も重要である特徴量として、坪数×通行量、坪数×平均坪単価、1/階数×平均坪単価が抽出された。仮説2で抽出されたのは、坪数、坪数×平均坪単価、通行量×視認性であることから、坪数×平均坪単価は重要であることが確認できる。また、1/階数×平均坪単価に関しては、図4で確認したように、1階の物件と2階の物件とでは坪単価が異なることから、重要な特徴量として抽出されたと考えられる。さらに、

4つの特徴量を抽出した場合には、上記の3つの特徴量に加え、坪数×視認性が抽出された。以上のことから、賃料推定には坪数および平均坪単価が特に重要であると考えられる。

7.4 キャッチコピーの文脈や曖昧さを考慮した場合の考察

仮説4では、キャッチコピーの文脈や曖昧さを考慮することにより推定精度は向上するという仮説のもと、検証を行なった。キャッチコピーの文脈や曖昧さを考慮する手法として、仮説4-1ではDoc2Vecを用いる手法、仮説4-2ではWord2Vecを用いる手法を提案した。

7.4.1 Doc2vec 技術による解析

仮説4-1では、Doc2Vec技術を用いることにより推定精度は向上するという仮説のもと、検証を行った。まず、キャッチコピーのベクトル化を行うにあたり、ベクトルの次元数を決定する必要がある。そこで本研究では、キャッチコピーを1から100次元にベクトル化し、モデルを構築した。その結果、次元数が大きくなるにつれ、推定精度が低下する結果が得られた。これは、特徴量数が多すぎるにより学習データに特化したモデルが構築され、その結果、テストデータで推定精度が低下した可能性が考えられる。

一方で、ベクトル数が1から5であるキャッチコピーを用いモデルを構築した場合、決定係数は最大で1%程度の上昇に止まり、また抽出した品詞による推定精度の大きな差は得られなかった。物件の価格推定にDoc2Vecを用いた研究[22, 23]では、品詞を選別することによる推定精度への影響はほとんどないことが示されている。またキャッチコピーには、地名や駅名、ローカルな名称が多く含まれており、辞書にない単語は誤った形態素解析になる場合が多く、結果的に解析結果に影響を与えた可能性が考えられる。

7.4.2 Word2Vec 技術による解析

仮説 4-2 では、Word2Vec 技術を用いることにより推定精度は向上するという仮説のもと、検証を行った。Word2vec による解析の際には、Wikipedia により学習済みの Word2Vec モデルを用い、キャッチコピーに含まれる単語のクラスタリングを行った。クラスタリングの際には、最適なクラスタリング数を求めるため、X-means 法を用いた。その結果、最適なクラスタ数は 53 となった。これらのクラスタを新たな特徴量とし、仮説 1 にて構築したモデルに順に追加してモデルを構築した結果、集客が見込めそうな単語や、視認性に関する単語が賃料推定に有効であることが確認された。

また、RFE により賃料に重要でないクラスタを消去し、モデルを構築した結果、2 つのクラスタを抽出した場合に、最大 2 % 程度の決定係数の上昇が得られた。Doc2Vec では最大で 1 % 程度の上昇にとどまっていたが、キャッチコピーの記述が人により異なること、独特の表現や言い回しがあることなどを考えると、文書単位で考えるのではなく、単語単位で考えること、また類似語をクラスタリングすることで表記揺れに対応することで、推定精度が向上したと考えられる。

7.5 仮説の組み合わせによる推定精度の向上

本研究では、検証した 4 つの仮説を組み合わせることで、さらなる推定精度の向上を図った。その結果、最も推定精度が高くなったのは、仮説 1、仮説 2、仮説 4-2 の組み合わせで LR によりモデルを構築した場合であり、その決定係数は 0.752 であった。

本モデルの有効性を確認するため、各仮説間において分散分析および多重比較を行う。その結果、仮説 1 と仮説 2 間、および仮説 1 と仮説 4-1 間で 5 %、仮説 1 と仮説 3 間、仮説 1 と仮説 4-1 間、また仮説を組み合わせることで構築したモデルと仮説 1 間において 1 % 有意であることを確認した。また、仮説を組み合わせることでモデルを構築した場合に最も高い推定精度が得られたことから、本研究では本モデルを新たな賃料推定モデルとして提案する。

モデルの構築に用いた特徴量は、既存の特徴量から、坪数、階数、駅徒歩時間、

通行量×視認性，坪数×平均坪単価，新規の特徴量から，年，地上階，地下階，p-合計，キャッチコピーからは，Word2Vecにより学習し，53にクラスタリングしたうちの，クラスタ25，26，27，28，35，40，45，50が抽出された．表13に，抽出された特徴量，モデルを構築した際の特徴量の係数の正負，およびクラスタに含まれる単語を示す．

抽出されたクラスタに含まれている単語を確認する．係数が正となったのは，抽出された8のクラスタのうちの7つであった．クラスタ25には，大学や学生，人気など，通行量に関する単語が含まれており，また黒字といったポジティブなイメージを持つ単語も含まれていることが確認できる．クラスタ26および35では，路面店や面する，通り，1階など，視認性や階数に関する単語が含まれている．クラスタ27，45，および50では，高級感，美麗，新築といった単語が含まれており，物件の状態や雰囲気を表していると考えられる．またクラスタ40では，チェーンやオフィス，大手など，物件周辺の情報が含まれていると考えられる．こういった単語は，賃料が高くなる傾向にあることが確認できる．一方で，係数が負となったのはクラスタ28であった．これに含まれる単語を確認すると，格安といった賃料が低いことを示す直接的な単語や，解約，路地といった，賃料が低くなることが連想される単語も含まれていることが確認できる．

また，これらの特徴量を用い，モデルを構築した場合の分布図を図41に示す．この図は，横軸に推定賃料，縦軸に実績値である実際の賃料をとっており，実線は推定賃料＝賃料となる線，また凡例は，推定値が実績値の50%以内に収まっているかどうかを示している．この図から，本モデルは低価格物件を高く見積もる傾向があることが確認できる．抽出した17の特徴量一覧である表13を確認すると，賃料が高くなる要因として抽出されたのは，17のうちの13であるのに対し，賃料が低くなる要因として抽出されたのは，17のうちの4であった．さらに，8つのクラスタを確認すると，賃料が低くなる要因として抽出されたのは1つのみであった．このことから，本モデルは賃料を低く見積もるための要因が少ないことを確認した．

次に，図41中に示した1および2の推定結果について考察する．

図41中の1に関して，実績値である賃料は735,000円であるのに対し，推定値

は345,256円となり、大幅に低く見積もる結果となった。この物件の賃料が高くなる要因としては、主要な駅から徒歩1分であること、業種に関する制限が無いこと、駅前通りであること、看板の掲載範囲が広いこと、角地であること、などが挙げられた。

図41中の2に関して、実績値である賃料は535,000円であるのに対し、推定値は866,693円となり、大幅に高く見積もる結果となった。この物件の賃料が低くなる要因としては、主要な駅から徒歩6分であること、深夜の営業が不可であること、業種に関する制限があること、路地裏であること、看板の掲載範囲が狭いこと、などが挙げられた。

本研究で構築した賃料推定モデルにて用いた特徴量は、表13にて示した17の特徴量である。これらの特徴量から、駅徒歩時間、路地は抽出されていることが確認できるが、業種制限の有無に関する情報や、看板の掲載範囲に関する情報などは抽出されていないことが確認できる。従って、抽出されていないこれらの要素が、推定結果に影響を与えた可能性が考えられる。

表13 仮説の組み合わせにより抽出された特徴量

独立変数	係数の正負	単語の例
坪数	+	
階数	-	
駅徒歩時間	-	
通行量 × 視認性	+	
坪数 × 平均坪単価	+	
年	-	
地上階	+	
地下階	+	
p-合計	+	
クラスタ 25	+	大学 学生 密集 集まる 黒字 平日 人気
クラスタ 26	+	路面店 路面 掲載 中華料理店 らーめん
クラスタ 27	+	高級感 魅力 おしゃれ 神楽坂 西麻布
クラスタ 28	-	格安 解約 路地 必見 オススメ 老舗
クラスタ 35	+	面する 沿い 商店街 通り 1階ビル
クラスタ 40	+	出店 チェーン オフィス エリア 大手
クラスタ 45	+	無償 美麗 居抜き バー レストラン
クラスタ 50	+	相談 スケルトン 即時 新築 飲食 業種

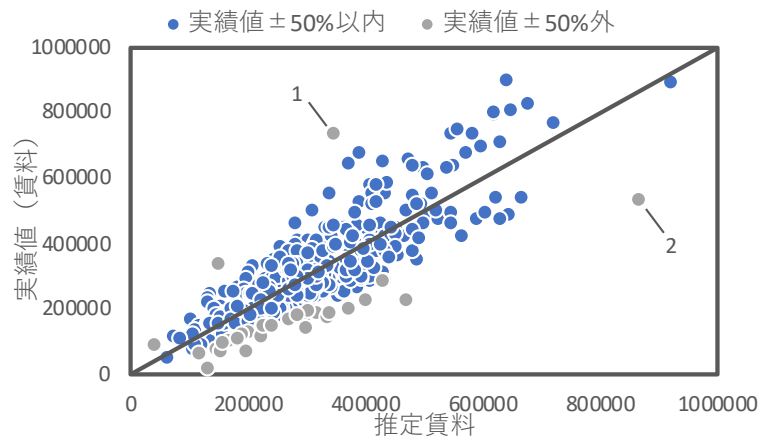


図 41 構築したモデルの推定賃料と実績値 (賃料)

8. 結論

本研究では荒川 [1] らの課題を解決すること、およびより高い推定精度を得ることを目的とし、新たな飲食店向け不動産物件の賃料推定モデルを構築した。荒川 [1] らの課題を解決するにあたり、本研究では推定精度向上に関する仮説を 4 つたて、検証を行った。

仮説 1 では、「データ数を追加することにより推定精度は向上する」という仮説を立て、検証を行った。その結果、データ数が 100 の場合とデータ数が 483 の場合の解析結果を比較すると、決定係数が最大で 11.5 % 上昇した。仮説 2 では、「新たな特徴量を追加することにより推定精度は向上する」という仮説を立て、検証を行った。その結果、仮説 1 の推定結果と比較すると、決定係数が最大で 1.8 % 上昇した。仮説 3 では、「特徴量の掛け合わせを考慮することにより推定精度は向上する」という仮説を立て、検証を行った。その結果、仮説 1 の推定結果と比較すると、決定係数が最大で 1.2 % 上昇した。仮説 4 では、「キャッチコピーの曖昧さを考慮することにより推定精度は向上する」という仮説を立て、検証を行った。キャッチコピーの曖昧さを考慮する手法として、本研究では仮説 4-1 にて Doc2Vec を用いる手法、仮説 4-2 にて Word2Vec を用いる手法を提案した。その結果、仮説 1 の結果と比較すると、仮説 4-1 の場合、決定係数は最大で 1.1 %、仮説 4-2 の場合、決定係数は最大で 1.8 % 上昇した。

また、これらの仮説を組み合わせることにより、さらなる推定精度の向上を図った。その結果、仮説 1、仮説 2、仮説 4-2 を組み合わせた場合に最も推定精度が高くなり、決定係数は 0.752 となる結果が得られた。結論として、本研究では荒川 [1] らの課題を解決し、かつ荒川 [1] らが提案した賃料推定モデルを上回る推定精度が得られた。

今後の課題は、大きく分けて二つある。一つ目は、通行量や視認性の機会的な取得方法について検討することである。二つ目は、構築したモデルを実際に使用してもらうことである。こうすることで推定精度の向上も図ることが可能であると考えられる。

謝辞

本研究を進めるにあたり、安本慶一 教授には、研究全般に関し、多大なるご指導・ご助言を賜りました。また、充実した研究環境の整備など、研究活動を手厚くご支援いただきました。感謝の意を表すとともに、心より厚く御礼申し上げます。

中村哲 教授には、ご多忙の中、論文審査委員を引き受けてくださった上で、副指導教官として様々なご助言をいただきました。本研究関連のプロジェクトでも、ご指導いただき、的確なアドバイスを賜りました。感謝の意を表すとともに、心より厚く御礼申し上げます。

荒川豊 客員教授には、ご多用にも関わらず論文の添削などで、的確なご指導およびご指摘をいただきました。また、本研究を進めるにあたり様々な視点から有用な知見やご助言を賜りました。感謝の意を表すとともに、心より厚く御礼申し上げます。

諏訪博彦 特任准教授には、本研究を進めるにあたり、使用するデータの分析を始めとした専門的な立場からのご指導・ご助言をいただきました。研究に関する相談にも丁寧に回答してくださったほか、研究発表の機会を多く設けてくださり、感謝の意を表すとともに、心より厚く御礼申し上げます。

藤本まなと 助教には、学生に近い立場から、論文執筆時の基礎や本研究に関するご助言を賜りました。感謝の意を表すとともに、心より厚く御礼申し上げます。

松田裕貴 助教には、論文の添削や、本研究を進めるにあたってのご助言などをいただきました。感謝の意を表すとともに、心より厚く御礼申し上げます。

金岡恵 事務補佐員、山内奈緒 事務補佐員には、学会や出張に関する事務処理を始め、研究生生活の様々な場面でご支援いただきましたこと、謹んで感謝申し上げます。

最後に、今日まで学生生活を様々な面から支えてくださった母を始め、家族に心より感謝申し上げます。

参考文献

- [1] 荒川周造, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一, 太田敏澄. 暗黙知センシングに基づいた飲食店向き不動産店舗の賃料推定. 人工知能学会全国大会論文集 第31回全国大会 (2017), pp. 1H2OS15a3–1H2OS15a3. 一般社団法人人工知能学会, 2017.
- [2] Changchun Wang and Hui Wu. A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences*, Vol. 6, No. 4, pp. 165–171, 2018.
- [3] Chih-Hung Wu, Chi-Hua Li, I-Ching Fang, Chin-Chia Hsu, Wei-Ting Lin, and Chia-Hsiang Wu. Hybrid genetic-based support vector regression with feng shui theory for appraising real estate price. In *2009 First Asian Conference on Intelligent Information and Database Systems*, pp. 295–300. IEEE, 2009.
- [4] Takafumi Miura and Yasuhi Asami. Hedonic analysis for estimation of condominium rent utilizing web information. *Procedia-Social and Behavioral Sciences*, Vol. 21, pp. 147–156, 2011.
- [5] Vincenza Chiarazzo, Leonardo Caggiani, Mario Marinelli, and Michele Otomanelli. A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia*, Vol. 3, pp. 810–817, 2014.
- [6] Omid Poursaeed, Tomáš Matera, and Serge Belongie. Vision-based real estate price estimation. *Machine Vision and Applications*, Vol. 29, No. 4, pp. 667–676, 2018.
- [7] Ikujiro Nonaka and Hirotaka Takeuchi. *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford university press, 1995.

- [8] Leo Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.
- [9] Mohammed Eltaher Hau-Wen Chang, Dongwon Lee and Jeongkyu Lee. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 111–118, 2012.
- [10] Toshizumi Ohta Hirohiko Suwa, Eiichi Umehara. Using the factor model to analyze internet bbs messages and stock returns. *SICE2011*, 2011.
- [11] P. D. Wyszocki. Cheap talk on the web: The determinants of postings on stock message boards. *Working paper, University of Michigan*, 1999.
- [12] R. Tumarkin and R.F. Whitelaw. News or noise? internet postings and stock prices. *Financial Analysts Journal*, Vol. 57, pp. 41–51, 2001.
- [13] W. Antiweiler and M.Z. Frank. Internet stock message boards and stock returns. *University of British Columbia Working Paper*, 2002.
- [14] Taro Yabuki Atsushi Shimoda. Success factors in global niche top companies: Analysis of free description data using text mining. *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 828–833, 2016.
- [15] 諏訪博彦, 梅原英一, 太田敏澄. インターネット株式掲示板の投稿内容分析に基づくファクターモデル構築の可能性. *人工知能学会論文誌*, Vol. 27, No. 6, pp. 376–383, 2012.
- [16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [17] Dongwen Zhang, Hua Xu, Zengcai Su, and Yunfeng Xu. Chinese comments sentiment classification based on word2vec and svmperf. *Expert Systems with Applications*, Vol. 42, No. 4, pp. 1857–1863, 2015.

- [18] Peng Wang, Jiaming Xu, Bo Xu, Chenglin Liu, Heng Zhang, Fangyuan Wang, and Hongwei Hao. Semantic clustering and convolutional neural network for short text categorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 352–357, 2015.
- [19] Bai Xue, Chen Fu, and Zhan Shaobin. A study on sentiment computing and classification of sina weibo with word2vec. In *2014 IEEE International Congress on Big Data*, pp. 358–363. IEEE, 2014.
- [20] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.
- [21] Lap Q Trieu, Huy Q Tran, and Minh-Triet Tran. News classification from social media using twitter-based doc2vec model and automatic query expansion. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*, pp. 460–467. ACM, 2017.
- [22] Vladimir Vargas-Calderón and Jorge E Camargo. A model for predicting price polarity of real estate properties using information of real estate market websites. *arXiv preprint arXiv:1911.08382*, 2019.
- [23] 鶴山優季子, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一ほか. Doc2vec を用いた飲食店向け不動産店舗賃料推定モデルの構築. *SIG-SAI*, Vol. 34, No. 3, pp. 1–6, 2019.
- [24] 鶴山優季子, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一ほか. 飲食店向け不動産物件の賃料推定手法の精度向上に向けた検討. マルチメディア, 分散, 協調とモバイル (DICOMO2019) シンポジウム 2019.
- [25] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pp. 155–161, 1997.

- [26] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. <http://mecab.sourceforge.jp>, 2006.
- [27] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, Vol. 1, pp. 727–734, 2000.
- [28] 米村淳, 大岸智彦, 井戸上彰, 小花貞夫ほか. スマートフォンを用いた人の混雑度推定手法の提案と評価. 研究報告モバイルコンピューティングとユビキタス通信 (MBL), Vol. 2013, No. 5, pp. 1–8, 2013.
- [29] 山本友理, 屋代智之, 重野寛, 岡田謙一ほか. 歩行者用道路上におけるリアルタイムな混雑情報の取得・提供手法. 情報処理学会研究報告モバイルコンピューティングとユビキタス通信 (MBL), Vol. 2004, No. 44 (2004-MBL-029), pp. 37–42, 2004.

研究業績

国内会議

1. 鶴山優季子, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一: “キャッチコピーを用いた飲食店用不動産賃料推定モデルの改善”, 2018 年度社会情報システム学シンポジウム (ISS).
2. 鶴山優季子, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一: “Doc2Vec を用いた飲食店向け不動産店舗賃料推定モデルの構築”, 2018 年度社会システム学と情報技術ウィーク (WSSIT).
3. 鶴山優季子, 曾根田悠介, 大坪敦, 菅田唯仁, 鳥越庸平, 党晨宇, 水本旭洋, 荒川豊, 安本慶一: “スマートホームにおける複合現実による家電操作システムの実装”, 2019 年度電子情報通信学会 総合大会 (IEICE).
4. 鶴山優季子, 諏訪博彦, 小川祐樹, 荒川豊, 安本慶一: “飲食店向け不動産物件の賃料推定手法の精度向上に向けた検討”, 2019 年度マルチメディア, 分散, 協調とモバイルシンポジウム (DICOMO).