JSDSS 日本ソーシャルデータサイエンス学会論文誌
第 8 巻　第 1 号（2024 年 3 月）

商品属性情報と価値推移を考慮した売価推定モデルの提案

三宅伸[†1]　大竹恒平[†2]　生田目崇[†3]

概要： 本研究は，リユースジュエリーを対象に，商品属性情報と中石の価格推移情報を用いた，売価推定モデルの提案を行ったものである．ジュエリー商品は，メインとなる宝石や周囲の装飾，ブランドなど，価格に対して影響を与える要素が多数あり，それらの多くはカテゴリカルデータとして記録されている．本研究では，Entity Embedding モデルによる売価推定モデルを構築し，その有効性の検証を行った．また，カテゴリカルデータに対して有用性が知られている XGboost モデルとの比較を行った結果，テストデータに対する汎化性能という点において，有効性の一端が確認された．

キーワード：売価推定，ニューラルネットワーク，Entity Embedding，XGboost，カテゴリカルデータ

1. はじめに

近年，インターネットの普及及びスマートフォンの利用拡大を背景に，電子商取引（Electronic Commerce; EC）の形態が多様化している．EC 市場の規模は年々右肩上がりに成長しており[1]，EC の市場規模が実店舗での市場規模に迫る図式となっている．中でも，価格比較サイトや EC におけるロコミ投稿の興隆や，インターネットオークションやフリマアプリ等の CtoC 型の二次流通市場も若年層を中心に成長している．

こうした EC を通じた購買が好まれる背景には，商品やサービスに関する比較・検討が容易であるという利便性に加え，実店舗よりも安く購入できるといった，消費者目線での価格優位性が大きな要因といえる．特に二次流通市場においては，商品に関連するトレンドや同一カテゴリの販売状況などにより，リアルタイムに価格が変化するため，消費者の価格感度もおのずと高くなると考えられる．以上の状況を踏まえると，二次流通市場において EC 事業者が販売を行う際には，適切な値付けが極めて重要な意思決定である．

本研究では，国内でリユースジュエリーの買取・販売を営む企業の実店舗・EC において販売された購買履歴データと商品属性データを対象として，機械学習による価格予測モデルの構築を試みる．ジュエリーに関する商品属性データは，メインとなる宝石や周囲の装飾，ブランドなど価格に対して影響を与える要素が多数あり，かつそれらはカテゴリカルな情報を有しているため，モデル構築の際にはデータのスパース性を考慮した適切な処理が必要である．

関連研究として，Yamaura et al. [2] はリユースジュエリーの販売価格の推定を商品仕様に関する変数と画像データの両方を用いたマルチモーダルモデルによって予測している．特徴としては，画像に関して畳み込みニューラルネットワークの一種である ResNet を用いた特徴量作成を行っている．ただし，変数はリングの重さや，宝石の大きさや種類といった限定された情報のみである．

また，ジュエリー同様コモディティ商品ではない一点ものであり，多数のカテゴリカル変数の属性情報を特徴として持つような，不動産データに対して価格予測を試みた研究がある．不動産に関する属性情報内には面積や最寄り駅からの時間距離など連続値も含むが，対象室内における様々な属性を表現するため，ケースバイケースのカテゴリ変数が多く含まれる．例えば，福井ら[2]のレインズに集約された成約データを用い，ニューラルネットワークによる不動産価格査定モデルの構築および，誤差逆伝搬法によるクラス分類を試みた研究や，加藤ら[3]の不動産ポータルサイトより取得した情報をカテゴリカル変数として用い，重回帰モデルによる土地価格の推定を試みた研究などがある．これらの研究結果より，カテゴリカル変数による属性情報を用いることで高精度な予測ができる可能性が示唆された．

本研究では，リユースジュエリーというドメインにおいて，カテゴリカル変数で与えられる属性情報に加え，各時点でのダイヤモンドの価格を外生変数として考慮した，売価推定モデルの提案を行う．本論文のデータの特徴としては，商品仕様に関して，宝石の品質などに関する詳細なデータが存在することであるが，多くのデータはスパースなカテゴリカルデータであり，過学習などが懸念される．そこで，本論文では，カテゴリカルデータについては埋め込み層によるカテゴリカル変数の次元圧縮を行う，Entity Embedding によって特徴を表現したモデルを構築し，ダイヤモンドリングのリユース売価の推定行う．

2. データ概要

本研究では主にリユースジュエリーの買取・販売を行っている企業から提供いただいたデータ使用する．なお，本

†1 中央大学大学院
†2 東海大学
†3 中央大学 (連絡先：nama@kc.chuo-u.ac.jp)
投稿日：2023 年 12 月 14 日
採録日：2024 年 2 月 17 日

研究では，2018 年 1 月 1 日～2020 年 8 月 31 日の期間内に購入されたダイヤモンドリング商品 10,006 件に関する購買データ，商品属性データを用いる．商品属性が有する特徴としては，「主に中石，脇石，その他の要素によって商品が構成（図 1）され，それぞれの要素が最終的な価格決定に影響を与えること」，「リユース商品の場合，同様の商品は存在せず一点限りであること」，「メインとなる中石の市場価値が変動すること」などがある．特に価格設定に大きな影響を与える宝石は，中石と脇石部分に付属しており，脇石部分には付属していないものも含まれる．ダイヤモンドの品質評価基準は，一般に米国宝石学会が定めた国際基準が世界で採用されており，4C と呼ばれる Carat（カラット），Cut（カット），Color（カラー），Clarity（クラリティー）の 4 つの項目によって評価される[4].

また，商品の価格設定に影響を及ぼす外生変数として，

図 1　ダイヤモンドリングの構成要素

世界最大のダイヤモンド販売機構である IDEX が発表するダイヤモンド 1 カラットあたりにおけるダイヤモンド取引の月毎の価格指標 [5] を利用する．なお，業界全体のトレンドとしては，2018 年 1 月以降ダイヤモンドの価格は下落傾向（図 2）にあり，価格が下がる時期に購買件数が増加している．本研究で扱う商品属性を表 1 に示す．

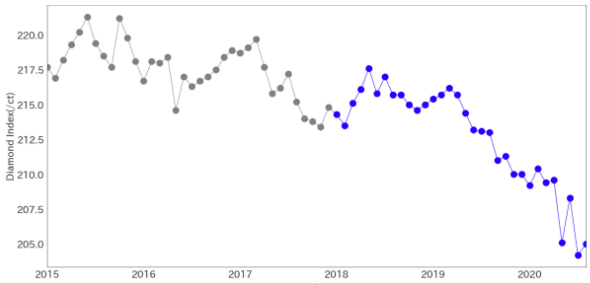

図 2　ダイヤモンド価格の推移（月毎）

価格に関する基礎集計の結果，最も高額な商品は 400 万円であり，全体商品 10,006 件のうち 90%の 8,956 件の商品は 10 万円以下で取引されていた．本研究では，商品の価格は上下の価格差が大きく，裾が広い分布をしていることを考慮し，以降扱う価格は Box-Cox 変換[6]を行った結果を用いる．

表 1　ダイヤモンドリングに関する商品属性
（括弧内の数値はラベル数を表す）

対象	カテゴリ変数	数値変数
中石	・カラット評価（5） ・カラースペック（5） ・クラリティスペック（5） ・輝きスペック（5） ・カラー（7）	・無色ダイヤモンド（数） ・有色ダイヤモンド（数） ・カラット（数）
脇石	・カラースペック（5） ・クラリティスペック（4） ・輝き評価（3） ・輝きスペック（5）	・無色ダイヤモンド（カラット数） ・有色ダイヤモンド（カラット数） ・ダイヤの有無
その他	・腕部分の金性（6） ・総コンディション（7） ・デザイン（23） ・ブランド（277）	・横幅 ・総重量 ・サイズ
価格遷移		・価格指標値

3. ダイヤモンドリングに関する売価推定モデルの構築

本節では，本研究で提案する Entity Embedding モデルに利用する変数及び提案モデルの概要を述べる．

Entity Embedding モデル[7]は，カテゴリデータに代表される質的なデータや，離散的な特徴を有するデータを，低次元のベクトルに変換し学習を行うモデルである．圧縮した次元を用いることで，カテゴリの関係性や特徴を効果的に学習することが期待される．図 3 に，本研究で用いる Entity Embedding モデルの概要図を示す．なお，モデル構築ならびに分析実行はすべて Python で行っている．

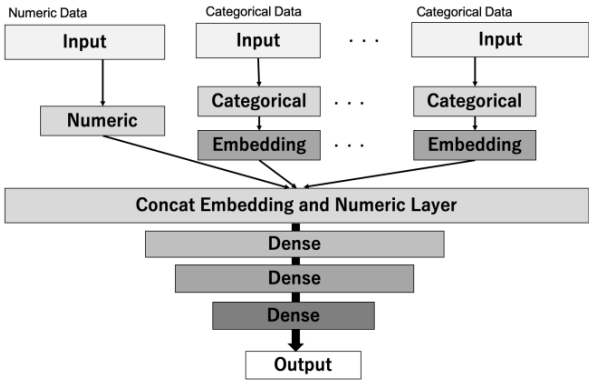

図 3　Entity Embedding モデルの概要図

3.1 変数の作成

Entity Embedding モデルでは，カテゴリカルデータを One-Hot に変換し，元のラベル数の次元から指定した次元への圧縮を行うが，圧縮する次元数については自ら設定する必要がある．そこで，本研究では最適な次元数を発見す

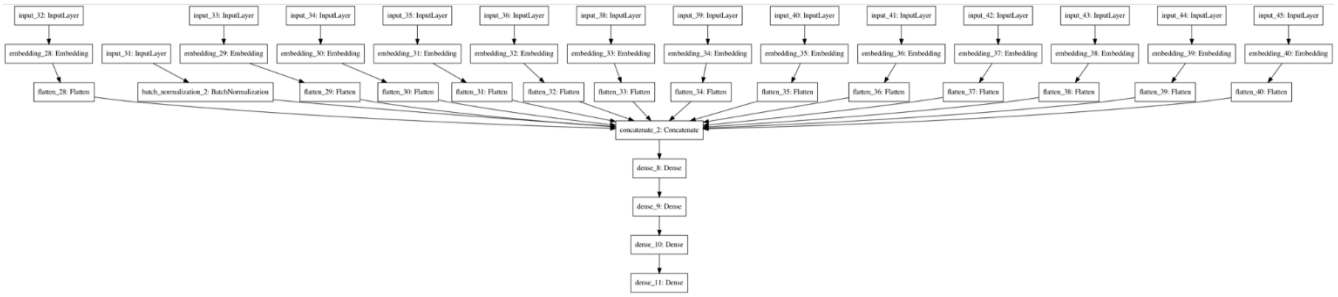

図 4　Entity Embedding を用いた売価推定モデルの概要図

ることを目的とした事前学習として，開発元である Google のチュートリアル [8] に従い，特定のカテゴリデータの次元数を 1 からカテゴリが有するラベル数−1 に変化させ，他のカテゴリ数をカテゴリが有するラベル数−1 に固定した上で，同じ構造のモデルによる損失関数の推移及び最終収束値から特定のカテゴリの次元数を設定した．このモデルは最終的な提案モデルと同じ売価価格予測を行うモデル構造であるが，各カテゴリ変数を圧縮する次元の設定により，最終的なモデルの精度がどれほど変化するのかを比較することを目的とする．これによりモデル内で各カテゴリ変数を最も高く表現力できる次元数の設定を行う．なお，次元数の設定時に利用するモデルでは，それぞれの入力及び Embedding 後にユニット数が 16 の隠れ層からなるモデルとし，最適化手法に Adam [9]，損失関数に平均絶対誤差 (mean absolute error; MAE)を用いるシンプルなニューラルネットワークモデルを構築する．活性化関数については，出力層のみ恒等関数を利用し，他では ReLU 関数を採用している．それぞれのモデルにおいて，学習及びテストに用いたデータ期間とサンプルサイズを表 2 に示す．

表 1　学習およびテストに用いたデータセットの概要

データセット	データ期間	サンプルサイズ
学習用データ	2018/1/1～2020/3/31	8,457
テスト用データ	2020/4/1～2020/8/1	1,549

例えば中石に関するカテゴリカル変数として，表 1 に示す通り，『カラット（5 ラベル）』が存在している．『カラット』の次元数を決定する際には，『カラット』以外の中石，脇石，その他に関するカテゴリ変数の Embedding 後の次元数をラベル数−1 で固定し，『カラット』の次元数を 1 から 4 まで変化させ，損失関数が最小となる次元数を特定する．同様の手順ですべてのカテゴリ変数に対して次元数を求める．各カテゴリ変数に対して損失関数が最小となる次元数を算出した結果を表 3 に示す．

表 3 より，大半の変数においてカテゴリ数−1 以外の次元数が選択されていることがわかる．また，特に元の次元数が高いデザイン，ブランドについては大幅に次元圧縮がされていることがわかる．

なお，数値変数については，カラット数に関する変数を除き，元の数値を用いた．カラット数については，データの分布を確認したところ，切りの良いカラット数の箇所にデータ分布が集中していたため 6 区間で分割したデータ変換を行った．例えば中石のカラット数であれば，「0~0.33」，「0.33~0.66」，「0.66~2」，「2~3」，「3~3.5」，「3.5~7」の 6 区分とし，該当する区分に数値が入力として存在する．

表 2　各カテゴリ変数について次元圧縮の結果

対象	変数概要	元の次元数	Embedding 次元数
中石	カラット評価	5	3
	カラースペック	5	3
	クラリティスペック	5	2
	輝きスペック評価	5	3
	カラー	7	2
脇石	カラースペック	5	2
	クラリティスペック	4	3
	輝き評価	3	1
	輝きスペック	5	4
その他	腕部分の金性	6	5
	総コンディション	7	6
	デザイン	23	6
	ブランド	277	101

3.2　Entity Embedding モデルの構造

前節で述べたカテゴリカル変数，数値変数を用いた，Entity Embedding モデルを構築する．図 4 に，本研究で提案する Entity Embedding モデルの概要図を示す．提案モデルは，カテゴリカルデータをそれぞれ分割してモデルに入力する．なお，数値データは全てまとめて入力する．カテゴリカルデータは Embedding 層により各次元に圧縮した後，連続層の入力とともに全結合し，3 つの隠れ層を持つニューラルネットワークにより価格の予測を行う．表 4 に入力データの概要を，表 5 にモデルに用いたハイパーパラメータを示す．なお，モデルの学習用データセット，テスト用データセットについては，表 2 に示したものを用いた．

利用データ内には欠損値などの外れ値を持つサンプルが多数含まれているため，過学習等の防止策として結合層以降の各層の各ユニットでは，誤差を更新する際に評価関数に対しペナルティを加える L2 正則化パラメータを導入

し，そのパラメータλは全て 0.01 に設定した．学習過程では最大エポック数を 1,000 回，バッチサイズを 10，交差検証におけるバリデーション率を 0.1 とした．また，更新が進まなくなった場合に学習を停止させる Early Stopping を導入している. なお, 参照するエポック数は 15 回とした.

表 3　提案モデルにおける Input 一覧

Input 番号	データ概要	データ型
Input 1	数値データ全て	数値
Input 2	中石カラット評価	カテゴリ
Input 3	中石カラースペック	カテゴリ
Input 4	中石クラリティスペック	カテゴリ
Input 5	中石輝きスペック評価	カテゴリ
Input 6	中石カラー	カテゴリ
Input 7	脇石カラースペック	カテゴリ
Input 8	脇石クラリティ	カテゴリ
Input 9	脇石輝き評価	カテゴリ
Input 10	脇石輝きスペック	カテゴリ
Input 11	腕部分の金性	カテゴリ
Input 12	総コンディション	カテゴリ
Input 13	デザイン	カテゴリ
Input 14	ブランド	カテゴリ

表 4　モデルに用いたハイパーパラメータ

パラメータ	値
第 1 中間ユニット数	98
第 2 中間ユニット数	98
第 3 中間ユニット数	32
出力ユニット	1
第 1 中間層　活性化関数	ReLU
第 2 中間層　活性化関数	ReLU
第 3 中間層　活性化関数	ReLU
出力層　活性化関数	Linear
最適化手法	Adam

4.　モデルの評価

本節では提案モデルの予測精度について評価を行う．具体的には，提案モデルを用いた売価推定と，決定木をベースとする機械学習手法であり，カテゴリカルデータに対する有用性が知られている XGboost を比較モデルとし，売価推定の結果を比較する．なお，本研究では比較モデルにおける決定木の本数は 200，決定木の深さは 10，モデルの学習率は 0.3 と設定し，出力においては価格予測を目的として線形モデルを採用した．また，ブースティング回数は Entity Embedding のエポック数と同じ 1,000 回に設定し，各学習におけるバリデーション率を 0.1 とし，Entity Embedding と同様に評価関数であるバリデーションの評価 MAE 値が 15 回連続で改善されない場合に学習を停止する処理をした．なお，これらのハイパーパラメータの探索は sklearn の GridSearchCV を用いた．

図 5 に 10 エポック以降の提案モデルの損失関数の推移を，図 6 にブースティング回数 10 回以降の比較モデルの損失関数の推移を示す．どちらも収束していると判断できるが，提案モデルと比較した際，比較モデルの方が早い段階での学習ができていることがわかる．

次に，モデルの予測値から実際の価格値を引いた残差を用い，モデルの精度評価を行う．提案モデルおよび比較モデルの予測結果の残差のヒストグラムとカーネル密度分布を図 7，8 にそれぞれ示す.

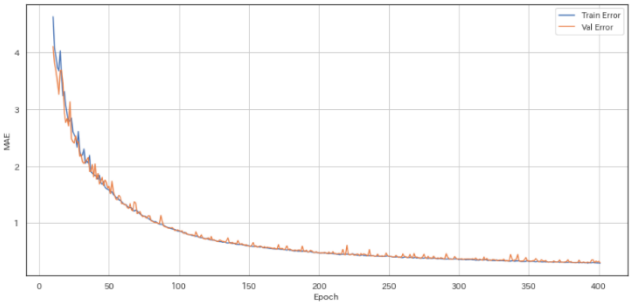

図 5　提案モデルにおける損失関数の推移

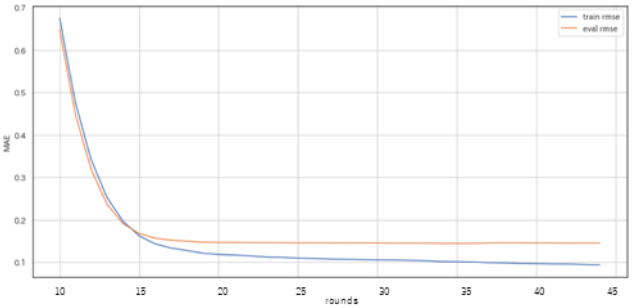

図 6　XGboost モデルにおける損失関数の推移

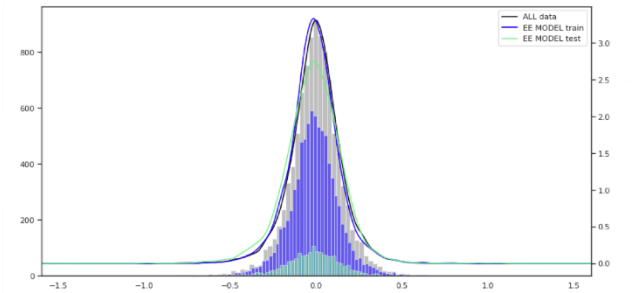

図 7　提案モデルにおける予測結果の残差

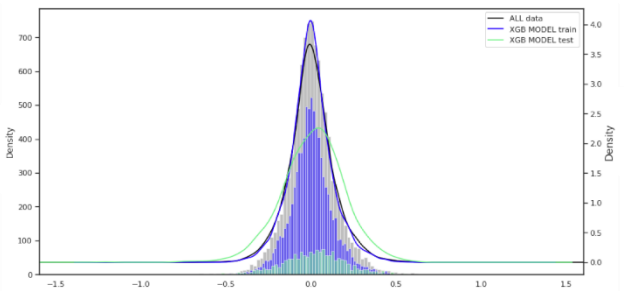

図 8　比較モデルにおける予測結果の残差

どちらのモデルもテストデータでは分散が大きく，平均

値の絶対値としては学習用モデルよりも高くなっているため，若干の過学習の傾向が見られた．図 7 より，提案モデルは比較モデルに比べて裾が広い残差分布となっているが，全体，学習用，テスト用すべてのデータセットについて，0 近辺を中心に正規分布に近い形状となっている．一方，図 8 より比較モデルの残差分布は，緑線で示したテスト用データについて，データの山が正の方向にずれており，分布の裾が厚いことがわかる．全体および，学習用のデータセットにおいては，提案モデル同様に，正規分布に近い形状となっていることから，比較モデルが過学習を起こしており，テスト用のデータでは価格を高く予測する傾向にあることがわかる．以上の結果より，テストデータへの汎化性能の観点から，提案モデルが比較モデルに対して有効であると考える．

次に，それぞれのモデルにおいて予測値と価格値が大きく外れたデータを基に，モデルの脆弱性を評価する．図 9 に，対象商品 10,006 件についての予測値と実際の価格の差分の箱ひげ図を示す．

図 9 より，特に提案モデルの予測においては若干大きく外れたケースがあることが伺える．テストデータに注目した場合，提案モデルはより高い精度で価格の推定を行うことができたが，全体のデータで当てはめると外れ値が多く，全てのケースを適切に予測できているわけではない．

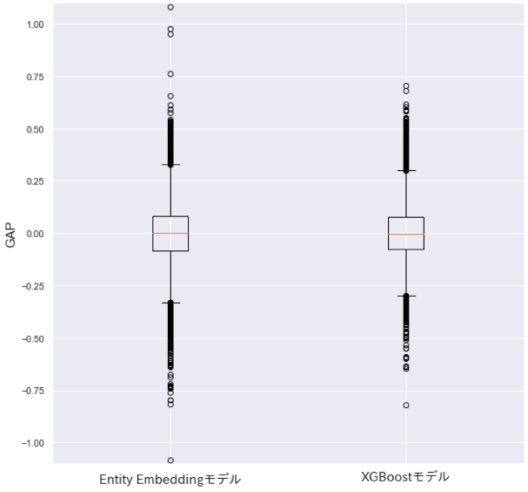

図 9　各モデルの差分分布と外れ値

各モデルの差分分布において，差分分布の下限を第一四分位数−1.5×四分位範囲，上限を第一四分位数+1.5×四分位範囲とし，この範囲を外れたデータを外れ値とした際の該当商品件数を表 6 に表す．

表 5　それぞれのモデルにおける外れ値の件数

モデル	負の外れ値（件数）		正の外れ値（件数）	
	学習	テスト	学習	テスト
提案モデル	168	52	147	36
比較モデル	88	66	139	78
Jaccard 係数	0.071	0.191	0.211	0.187

表 6 の結果から，これまでの結果と同様に提案モデルではテストデータにおいて外れ値が少なくなっているが，全体的に外れ値が負の領域で多いことがわかる．一方，比較モデルではテストデータにて，正負両方とも提案モデルよりも外れ値件数が多いことがわかる．学習データでは正の方向に外れ値が多いため，差分の絶対値は少ないが若干予測結果が高い値をとる傾向があると考えられる．

また，2 つのモデルにおける外れ値の集合についてその和集合に対する共通集合の比である Jaccard 係数の値に注目すると，学習データとテストデータにおける正の方向への外れ値が，両モデルにて比較的一致していることがわかる．

提案モデルにおいて外れ値となった実際の商品情報についての調査結果を踏まえ，価格を大きく外す要因について考察する．まず，提案モデルが負の方向に価格を外してしまう要因として，中石カテゴリに関する特徴量が考えられる．外れ値となった商品は，中石カットが FAIR 評価，中石カラースペックが C 評価，中石クラリティスペックが C 評価，中石輝きスペックが B 評価と，中石に関しては低く評価された商品であることがわかった．一方で脇石に関する評価においては A 評価となっている値が多く，モデルの学習においては中石に対する評価を優先的に価格評価で利用している傾向があると考えられる．

また，提案モデルが正の方向に価格を外してしまう要因としては，中石におけるダイヤの評価が全体の平均値に引っ張られてしまったことにあると考えられる．図 10 に，提案モデルが正の方向で外れ値的に評価した商品の価格と脇石の無色ダイヤモンドのカラット数分布を示す．

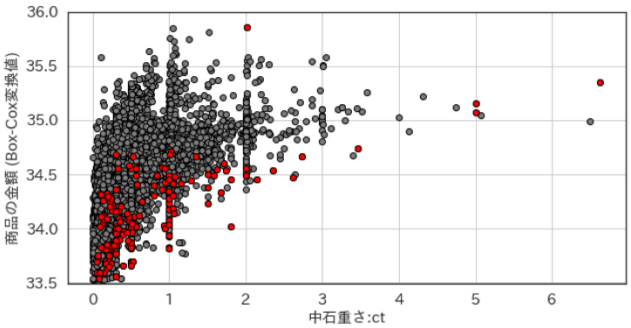

図 10　提案モデルにおける中石のカラット数と金額の分布

図 10 において赤点は正方向に外れ値となった商品であり，価格を高く予測してしまっている．これらの商品は分布の下部に位置していることがわかる．すなわち，中石のカラット数が低く，価格が安い商品について価格を高く見

積もっている．このことから，モデルがカラット数の分布においてより平均的に近づくように価格の予測を行っていると推察することができる．この点については，比較モデルにおいても同様の傾向が確認された．中石は最も目に付くパーツであり，実務での価格設定においても評価のウェイトが高い．中石，脇石，その他の属性に対する学習ウェイトをパーツごとに変化させるなど，精度向上に向けて改良の余地があると考えられる．

5. まとめと今後の課題

本研究では，中古市場におけるリユースジュエリーの属性データおよび，ダイヤモンドの価格を外生変数として考慮した，売価推定モデルの構築を行った．具体的には，中石，脇石など，商品を構成する属性として多くのカテゴリカルデータを含むという特徴を加味し，Entity Embedding を用いたニューラルネットワークによる売価予測モデルを構築しその精度を評価した．提案モデルについて，カテゴリカルデータに対して有用性が知られている XGboost モデルとの比較を行った結果，テストデータに対する汎化性能という点において，有効性の一端が確認された．

本研究では，カテゴリカル変数の次元圧縮の際，次元数を変化させた際のモデルの評価関数により次元数を定めた．しかしながら，カテゴリカル変数同士の次元圧縮による相関等の関係性は考慮しておらず，次元数をより横断的に求める必要がある．

謝辞　本研究は JSPS 科研費 21K13385 の助成を受けたものです．

参考文献

[1] 経済産業省商務情報政策局, “令和４年度電子商取引に関する市場調査報告書” (2023).
[2] Y, Yamaura, N. Kanemaki and Y. Tsuboshita, “The Resale Price Prediction of Secondhand Jewelry Items Using a Multi-modal Deep Model with Iterative Co-Attention,” *Proceedings of KDD Workshop on AI for fashion, 2019*, 10 pages, (2019).
[3] 福井光，阪井一仁，南村忠敬，三尾順一，木下明弘，高田司郎，“レインズのニューラルネットワークを用いた不動産価格査定について，” 人工知能学会全国大会論文集，第 32 回全国大会，4A2-03，(2018).
[4] 加藤暢之，新妻弘崇，太田学, “重回帰分析による土地価格推定の一手法,” DEIM Forum 2018, vol. H5-3，(2018).
[5] Gemological Institute of America ウェブサイト, https://www.gia.edu/JP/daiamond（最終閲覧日：2023 年 12 月 13 日）
[6] G. E. P. Box and D. R. Cox, “An Analysis of Transformations,” *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 26, No. 2, pp. 211-252, (1964).
[7] Cheng Guo and Felix Berkhahn, “Entity Embeddings of Categorical Variables,” https://doi.org/10.48550/arXiv.1604.06737, (2016).
[8] Google for Developers, “Introducing TensorFlow Feature Columns ,” https://developers.googleblog.com/2017/11/introducing-tensorflow-feature-columns.html （最終閲覧日：2024 年 2 月 24 日）
[9] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” International Conference on Learning Representations (ICLR), pp. 1-15, (2015).

JSDSS 日本ソーシャルデータサイエンス学会論文誌
第 8 巻　第 1 号（2024 年 3 月）

Proposal of a Selling Price Estimation Model Using Product Attributes and Price Trends

Shin MIYAKE[†1]　Kohei OTAKE[†2]　Takashi NAMATAME[†3]

Abstract: This study proposes a sale price estimation model for reused jewelry using information on product attributes and jewelry price trends. Jewelry products have many factors that affect the price, such as the main gemstone, surrounding decorations, and brands, many of which are recorded as categorical data. In this study, we constructed a model based on the entity embedding model. A comparison with the XGboost model, one of the famous machine learning methods, which is known to be useful for categorical data, confirmed some of the effectiveness of the entity embedding model in terms of generalization performance for test data.

Keywords: Selling Price Estimation, Neural Networks, Entity Embedding, XGboost, Categorical Data

†1 Graduate School of Chuo University
†2 Tokai University
†3 Chuo University (Correspondence Author: nama@kc.chuo-u.ac.jp)a