商品の属性情報と価値推移を考慮した商品売価推定モデル
Price Prediction Model Using Attribute Information and Value Transitions

経営システム工学専攻　　三宅　伸
Cource of Industrail and Systems Engineering　Shin MIYAKE

1　研究背景と目的

本研究の背景は大きく分けて 3 つの要素から成る．1 つ目に，二次流通チャネルの拡大による小売市場への影響，2 つ目に機械学習技術の実務への応用，3 つ目に企業のマーケティングにおける顧客ニーズに適したサービス展開である．まず，1 つ目の二次流通チャネルの拡大による小売市場への影響について言及する．スマートフォンの普及を背景に，CtoC-二次流通市場の規模が拡大している [1]．これらが支持される理由として，価格設定における自由度の高さが考えられる．二次流通市場の台頭を起点に消費者はサービスと価格の関係に敏感となり，小売市場では今後最適な価格設定を行うことが重要である．次に機械学習技術の実務への応用について述べる．現在多くの企業にて，機械学習のアルゴリズムの導入及びサービス事例が見られる [2, 3]．特に主観に囚われない結果を導く機械学習の応用は今後マーケティングにおいて一層重要となり，より精度が高く，対象データに適したモデルの構築が必要である．3 つ目に企業のマーケティングにおける顧客ニーズに適したサービス展開について述べる．特に小売市場では売れ行きなどの商品の流動性を意識することが重要である．提示する商品価格と消費者との乖離が少ない場合，商品の流動性は早くなり，結果として売れ行きなどに現れると考えられる．そこで，企業の価格の設定においては，まずサービスへの質を上げることを意識し，この結果が消費者らが成す自発的購買頻度や全体的な商品流動性に現れるようなマーケティングを仕掛けることが非常に重要である．

本研究では，国内において実店舗と EC を販売チャネルに持つリユースジュエリー専門店の商品属性データと購買データを用い，機械学習による価格予測モデルの構築と生存時間分析による商品流動の検証を目的とする．特に商品属性データにはスパースなカテゴリカル変数であることが多いため，本研究では Entity Embedding を用いた価格予測を構築し，カテゴリカル変数の表現を豊にしている．また，モデルの比較対象として XGboost モデル [5] を利用し，2 つのモデルでの精度比較を行う．また，小売業界での実務においてモデルの妥当性として，精度の良し悪しのみならず価格による商品の流動性を考慮することが重要である．本分析では商品の流動性の観点から，生存時間分析を用い，予測価格と実際の価格における差分及び商品の販売期間と購買の関係性から，各モデルの評価を行う．

2　対象データ

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．また分析内では，対象企業で取り扱う全ての商品の中から，2018 年 01 月 01 日～2020 年 08 月 31 日の期間内に購入されたダイヤモンドリング 10,006 件に関する購買データ，商品属性データを用いる．加えて，商品の価格設定に影響を及ぼす外生変数として，世界最大のダイヤモンド販売機構が月毎に発表するダイヤモンド取引の価格指標を利用する．

2.1　商品の価格分布

ダイヤモンドリング商品で最も高額な商品は 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る．価格について，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 変換による変換を行った結果を用いる (図 1)．

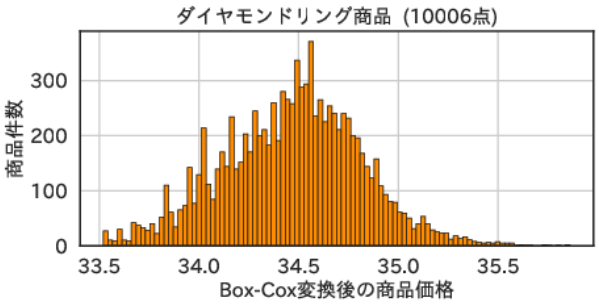

図 1: 商品の価格分布 (Box-Cox 変換後)

2.2　商品属性データ

ダイヤモンドリングは主に中石，脇石，その他の大別して 3 つの要素によって構成され，それぞれの要素の評価を考慮して価格が決まる．主に中石とはジュエリーの中央に付属している宝石のことを指し，脇石は中石の横に添えられた小さな宝石を指す．また，宝石の価格評価は主に Carat（カラット），Cut（カット），Color（カラー），Clarity（クラリティー）の 4 項目で行われ，本研究の利用データでもジュエリー内の宝石の評価値としてこれらのデータを含む．

2.2.1　中石のデータ概要

中石に関する属性情報には，中石部分に付属している石に関し，数値データではダイヤモンドのカラット数，無色ダイヤモンドの個数，有色ダイヤモンドの個数の 3 つを含む (図 2)．

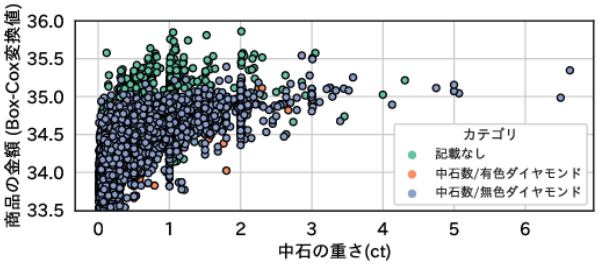

図 2: 中石　カラット数と価格分布

図 2 より，同じカラット数でも安価な商品や，逆に高額な商品など一概に金額とカラット数が比例関係でないことがわかる．また，分布の様子からは，多くの商品がカラット数が 1.0 や 2 .0 などの区切りの良い数値であることがわかる．これについては，新品商品とは異なるリユース市場の課題であり，商品を詳細な部分まで評価できないことに起因すると考えられる．このため，販売する際の中石のカラット数において許容される範囲で繰り上げもしくは切り捨てした値を商品項目に提示しているといったことが考えられる．中石に関するカテゴリ変数では中石のカット，カラースペック，クラリティ，輝きについて各 5 段階評価で表した変数，有色ダイヤの色について 7 色のフラグ変数が含まれる．中石に関する変数は，数値データが 3 変数と，カテゴリカル変数 5 変数の計 8 変数となる．

2.2.2　脇石のデータ概要

対象のダイヤモンドジュエリーでは脇石部分に宝石を含んでいない商品もあるため，脇石に関するデータでは数値変数とカテゴリカル変数が共にスパースとなる．数値変数では，脇石の無色ダイヤモンドのカラット数，有色ダイヤモンドのカラット数，ダイヤモンドの個数データが含まれる．特徴として，ダイヤモンドのカラット数が増えれば価格も上がる大まかな傾向はあるが，一概に比例関係にあるとは言えない．脇石のカテゴリカル変数には，ダイヤモンドのカラー，クラリティ，輝きにおける 5 段階評価と，企業オリジナルの輝きスペック評価が 3 段階の評価値が含まれている．脇石に関する変数は，数値データが 3 変数と，カテゴリカル変数 4 変数の計 7 変数となる．

2.2.3　その他データ概要

その他の変数としては，主にジュエリー本体への評価値を表す．数値変数としてはジュエリーの幅，総重量，サイズの 3 変数があり，特に欠損が多い幅の変数では k-menas 法を利用し，類似商品からデータの補完を行った．一方カテゴリカル変数にはジュエリーの金性 (プラチナや金) のフラグ値，6 段階の総コンディション評価，23 種類のジュエリーデザインにおけるフラグ値，277 種類のブランドに関するフラグ値となる．その他ジュエリーに関する変数は，数値データが 3 変数と，カテゴリカル変数 4 変数の計 7 変数となる．

2.3　ダイヤモンドの価格指標データ

ダイヤモンドの価格設定に対する外来変数として，本分析では世界最大のダイヤモンド販売機構が月毎に更新するダイヤモンド価格の評価指標を使用した．

3　価格予測モデルの構築

商品の属性データとダイヤモンド取引の価格指標データを利用し Entiry Embedding と XGboost による価格の予測モデルを構築する．なお，モデルの学習データとして 2018/01/01〜2020/03/31 に購入された商品のデータ 8,457 件を利用し，テスト用に 2020/04/01〜2020/08/01 に購入された商品 1,549 件のデータを利用する．

3.1　Embedding について

Embedding は主に自然言語処理の領域で利用されるニューラルネットワーク内の埋め込み層であり，複数の単語の生起行列を任意の次元数のベクトル行列に置き換える手法である．なお，Embedding 行列は以下のように定義される．

$$
Embedding\ Matrix = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{pmatrix}
$$

$$
\begin{cases} n : \text{単語数} \\ m : \text{任意の次元数 } (1 \sim (n-1)) \end{cases}
$$

ラベル化した単語の One-Hot 表現に対し，Embedding 行列の積を取ることで，各ラベルは任意の次元で定義されたベクトル表現を獲得できる．また現在，この手法は Entity Embedding [4] が提案されて以来，自然言語処理の領域のみならずカテゴリカル変数に対しても適用しうることが示された．本研究で扱うデータでは，多くのラベルを含むカテゴリカル変数を多く含むため，モデル内では Entity Embedding を利用して学習を行う．

3.2 Entity Embedding モデルの構築

本分析では，中石，脇石，その他に関する計 9 つの数値変数を 1 入力とし，残りの 13 個のカテゴリカル変数をそれぞれ別入力とした計 15 入力のニューラルネットワークを構築する．特に，カテゴリカル変数については入力後に Embeding 層を挟み，数値変数の入力と Embeding 後の出力を全結合層でまとめ，さらに 3 層の隠れ層を経由したからなる価格予測モデルを構築する (図 3)．

図 3: Entity Embedding モデル概要

設定したパラメータをまとめる (表 1)．利用データ内には特に欠損値等の外れ値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 正則パラメータ λ を全て 0.01 に設定している．学習過程では最大エポック数を 1000 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 とする．また学習過程において損失関数の改善が見られない場合に学習を停止させる Keras の EarlyStopping を導入しており，参照するエポック数は 15 回に設定している．以下に Entity Embedding による 10 エポック以降の損失関数推移を表す (図 4)．

表 1: Entity Embeddig モデル　パラメータ一覧

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
出力層ユニット	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
最適化関法	Adam

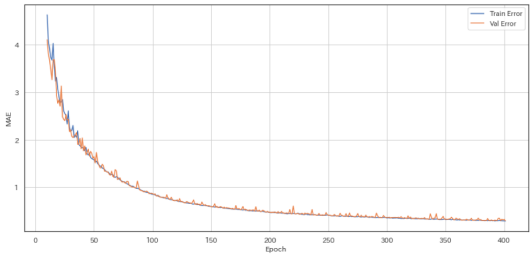

図 4: Entity Embedding モデル 損失関数

学習は 400 エポックで終了し，最終エポックでの loss 値は 0.115，val_loss 値は 0.125 で収束した．

3.3 XGboost モデルの構築

XGboost のパラメータ設定では探索法である GridSearch を用い，決定木の本数を 200，決定木の深さを 10，学習率を 0.3 に設定した．また，ブースティング回数は Entity Embedding での学習と同様の条件に設定した．以下に XGboost による 10 エポック以降の損失関数推移を表す (図 5)．

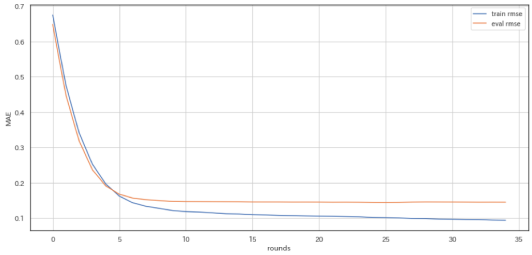

図 5: Entity Embedding モデル 損失関数

図 5 では学習回数が 43 回で終了し，最終エポックでの loss 値は 0.093，val_loss 値は 0.145 で収束した．以下に各モデルの予測結果を表す (図 6, 7)．

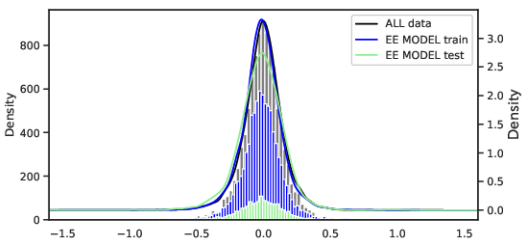

図 6: Entity Embedding モデル 価格予測結果

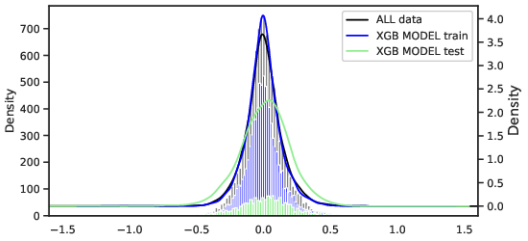

図 7: XGboost モデル 価格予測結果

テストデータに対する予測価格の結果では，XGboost が過学習を起こしているのに対し，Entity Embedding ではより高精度で予測を行うことができた．

3.3.1 生存時間分析

各モデルの妥当性評価として，イベントを購買と捉え，各商品の販売日数を生存時間とした生存時間分析を行う．購買行動において消費者が商品属性に対して感じる値ごろ感の感覚的尺度よりも実際の価格が安価である場合に購買発生が早く起きると考えられるため，各モデルにて予測した商品の価格に対し実際の価格が高い場合と低い場合の 2 群間にて生存時間の比較を行う．以下に Kaplan-Meier 法による Entity Embedding の各群における生存率の推移を表す (図 8)．

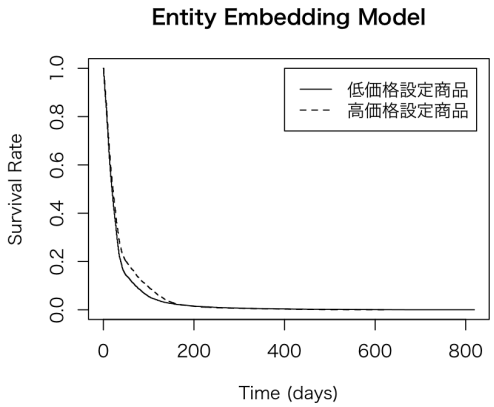

図 8: Entity Embedding モデル 生存曲線

結果では，ログランク検定により Entity Embedding のモデルのみ各群間に生存率の違いがあることがわかった．これによりモデルの価格設定において商品の特性を学習しつつ，早く売れる価格の予測が可能となった．XGboost の予測結果において，外れ値となった商品の特徴には頻出頻度の少ないブランドのデータが含まれており，スパースな特徴量をうまく表現できたことが Entity Embedding モデルの有用性に繋がったと考えられる．

4 まとめと今後の課題

本分析では商品属性データを利用して価格予測モデルを構築し，購買発生の速度にて妥当性の評価を行った．しかしファッションアイテムでは視覚による情報も重要であり，今後は商品の画像データ等をモデル内に利用することで，より消費者行動を考慮したモデル構築が可能であると考えられる．

参考文献

[1] 経済産業省 商務情報政策局 情報経済課,「平成 30 年度 我が国におけるデータ駆動型社会に係る基盤整備 (電子商取引に関する市場調査)」，pp. 28–37，`https://www.meti.go.jp/press/2019/05/20190516002/20190516002-1.pdf`，(2019).

[2] I. Portugal，P. Alencar，D. Cowan，“The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review”，*Expert Systems with Applications* Volume 97，pp. 205–227，(2018).

[3] 総務省,「平成 28 年度版　情報通信白書 IoT・ビッグデータ・AI ネットワーク とデータが想像する新たな価値 」，pp. 128–155，`https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n3100000.pdf`，(2016).

[4] C. Guo, F. Barkhahn, “Entity Embedding of Categorical Variables”，arXiv:1604.06737v1 [cs.LG] 22 (2016).

[5] T. Chen，C. Guestrin，“XGBoost: A Scalable Tree Boosting System”，*KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*，pp. 785–794 (2016).