E-Commerce Product Price Prediction

Munmun Sangtani
Dept. of Electronics and Communication Engineering
Nirma University, Ahmedabad, India
20bec073@nirmauni.ac.in

Siddharth Agarwal
Dept. of Electronics and Communication Engineering
Nirma University, Ahmedabad, India
20bec119@nirmauni.ac.in

***Abstract*—Accurate pricing is crucial for e-commerce success, but predicting product prices at scale is challenging. We propose a novel machine learning-based approach to forecast prices using product attributes, historical pricing trends, and seller behavior. Our experiments demonstrate the effectiveness of our approach in predicting prices for diverse products, highlighting the potential of machine learning in improving pricing accuracy for e-commerce platforms.**

Index terms - Introduction, Theory, Methodology, Result and Analysis, Conclusion, Future Prospects and References.

I. Introduction

Pricing products at scale is a challenging task, particularly in the context of online marketplaces that host a vast array of products. This difficulty is further compounded by the fact that different product categories exhibit distinct pricing trends; for instance, clothing prices vary seasonally and are heavily influenced by brand names, while electronic product prices fluctuate based on product specifications. The challenge of pricing diverse products is well-recognized by Mercari, Japan's largest community-powered shopping app. The platform aims to provide pricing suggestions to its sellers, but this is a formidable task, as sellers are free to list any item or combination of items on Mercari's Marketplace, thereby adding to the complexity of pricing prediction models.

A. Dataset

- Train-id / test-id – id of the listing
- Name – the title of listing.
- Item-condition-id – the condition of the items by seller
- Item-condition-id – the condition of the items by seller
- Category-name – category of listing
- Brand-name – brand of listing
- Price – the price for the item sold. Our Target variable we want to predict, which is in USD.
- Shipping – 1 if shipping fee paid by seller and 0 by buyer
- Item-description – the full description of item.

II. Objective

We aim to develop a pricing algorithm that automatically suggests optimal prices for products listed on online platforms. By providing users with suggested prices based on product attributes and user-inputted text descriptions, enhance the buying and selling experience can be enhanced and the volume of transactions can be increased. This problem can be framed as a supervised learning task, where the objective is to predict prices based on various product features. The performance of the algorithm can be evaluated using the Root Mean Squared Error (RMSE) metric, which measures the accuracy of the model's price predictions. Additional user data, such as location, gender, and login information, could potentially improve the accuracy of the pricing algorithm.

III. Theory

A. Compressed Sparse Row (CSR) matrix

In many machine learning applications, including product price prediction, the input data can be high-dimensional and sparse, meaning that most of the features have zero values. To efficiently handle such data, we can use a sparse matrix representation known as a Compressed Sparse Row (CSR) matrix.

In this paper, we propose the use of CSR matrices to represent the product attributes in a product price prediction task. We start by preprocessing the dataset and converting it into a CSR matrix representation. Each row of the CSR matrix represents a product, and each column represents a product attribute. If a product has a non-zero value for an attribute, the value is stored in the CSR matrix, and otherwise, it is zero.

We then use LightGBM, a gradient boosting algorithm, to train a regression model on the CSR matrix. LightGBM can handle sparse input data efficiently and can scale to very large datasets.

To optimize the performance of the model, we tune hyperparameters such as the learning rate, number of iterations, maximum depth, and feature fraction. We also use cross-validation to evaluate the model's performance and prevent overfitting.

Our experimental results demonstrate that using CSR matrices can significantly improve the training and inference time of the model compared to using a dense matrix representation.

B. Root Mean Squared Logarithmic Error

RMSLE stands for Root Mean Squared Logarithmic Error. It is a performance metric used to evaluate the accuracy of models that predict values that have a wide range of magnitudes. RMSLE is particularly useful for models that predict values on a logarithmic scale, such as those used in pricing or rating predictions.

The formula for RMSLE is:

$$\sqrt{(1/n) * \sum((log(ypred + 1) - log(ytrue + 1))^2}$$

where y_pred is the predicted value, y_true is the true value, log() is the natural logarithm function, and n is the total number of predictions.

RMSLE is similar to RMSE (Root Mean Squared Error), but it applies the logarithmic transformation to both the predicted and true values before calculating the difference. This transformation ensures that errors in predicting smaller values are penalized more heavily than errors in predicting larger values, which is often desirable in real-world applications.

In RMSLE, a lower score indicates better model performance. A perfect score of 0 means that the model predicts all values exactly, while a score of 1 means that the model's predictions are off by a factor of e (the base of the natural logarithm).

C. Ridge Regression

Ridge regression is a linear regression method used for predicting a continuous outcome variable. It is a regularization technique that adds a penalty to the sum of squared coefficients of the regression model. This penalty term is called the L2 norm and helps to prevent overfitting of the model. In Ridge regression, the strength of the penalty is controlled by a tuning parameter called lambda (λ).

As it increases, the magnitude of the coefficients decreases, which leads to a simpler model with lower variance but potentially higher bias. Ridge regression can be used for both simple and multiple linear regression problems and is particularly useful when the number of features in the data is high and there is a possibility of multicollinearity (correlation between the predictors).

$$L(w, D) + \lambda * ((w))^2$$

L2 Regularization Loss Function:

- L: aims for low training error
- Lambda: A scalar value that controls how weights are balanced
- W: weights are balanced against complexity

D. Light Gradient Boosting Algorithm

Light Gradient Boosting Machine (LightGBM) is a popular gradient boosting algorithm that has shown great success in various machine learning tasks, including product price prediction. In this paper, we propose the use of LightGBM for product price prediction.

The dataset used in this study consists of various product attributes such as brand, category, ratings, and reviews, along with the product prices. We first preprocess the data by encoding categorical features and normalizing numerical features.

We then use LightGBM to train a regression model that predicts the product prices. LightGBM uses a gradient-based approach to iteratively improve the model's predictions. It splits the data into small, hierarchical sets and fits a regression tree on each subset. The algorithm then combines the predictions from all the trees to create the final output.

To optimize the performance of the model, we tune hyperparameters such as the learning rate, number of iterations, maximum depth, and feature fraction. We use cross-validation to evaluate the model's performance and prevent overfitting.

Our experimental results demonstrate that LightGBM outperforms other commonly used regression algorithms such as linear regression, decision trees, and random forests. The proposed approach achieves a low mean absolute error and high R-squared value, indicating accurate price predictions.

IV. Literature Review

The present literature review covers few research papers that aim to predict prices in the e-commerce domain using machine learning techniques.

PriceCop – Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform employs Linear Regression and Least Squares Support Vector Machine to predict the next day price. The results show that the SVM model performs better with an accuracy of 84%, compared to 62% for the Linear Regression model. Another paper,E-Commerce Price Forecasting Using LSTM Neural Networks focuses on E-Commerce Price Forecasting using two machine learning models: Support Vector Regression (SVR) and Long Short-Term Memory (LSTM). The dataset used in this study is from the phone market. The results indicate that the LSTM model outperforms the SVR model with an RMSE of 23.64 euros compared to 33.43 euros for SVR. In E-commerce Price Suggestion Algorithm – A Machine Learning Application , three machine learning models: Linear Regression, Random Forest, and Light GBM were used. The study uses the dataset from the Meraci Online Platform. The results show that Light GBM performs best with an MAE of 7.63, while the other two models also perform well, with MAEs of 8.88 and 7.99 for Linear Regression and Random Forest, respectively.

The studies demonstrate the effectiveness of machine learning techniques in predicting e-commerce prices, with each study highlighting different models and datasets.

V. Methodology

A. Text Pre-Processing

Text data is highly unstructured and contains various forms of noise, making it unsuitable for analysis without preprocessing. Text pre-processing involves the cleaning and normalization of text data to remove noise and prepare it for analysis.

Terminologies for representation of text:

- Document: A piece of text. It can be a sentence, a paragraph or a full page report.
- Tokens: Many tokens form a word.
- Corpus: Collection of documents.
- Term Frequency(TF): A measure of how often a term occurs in a document.

- Inverse Document Frequency (IDF): Distribution of a term over a corpus.

To make text data more suitable for analysis, various preprocessing techniques can be applied:

- Stop Word Removal: Stop words are common words with little to no meaning, such as ”the” or ”to”.
- Bag of Words Representation: treats each word as a feature.
- TF-IDF (Term Frequency Inverse Document Frequency): A common value representation that gives more weight to words with lower frequencies. Rare words have more importance.
- N-Grams: Sequences of adjacent words, can also be used as terms to add more meaning. A word by itself may not have any value but as a pair, they add more value.
- Stemming/Lemmatization: Applied to convert words into their base forms for analysis.

B. *Data Preparation and Cleaning*

Data preparation for the dataset required pre processing of the text data. Missing values were replaced with NA. Preprocessing steps involved are:

- Handling Missing Values – Replaced the missing values with Nan.
- Lemmatization performed on item description
- Label Encoding – Turned categorical values into 0’s and 1’s
- Tokenization – Given a character sequence, tokenization is the task of chopping it up into pieces
- Scaling – Scaled the target variable (price)

Once the variables were cleaned and pre-processed, they were converted into a sparse matrix as there were thousands of word features. CSR matrix allows for better utilization of memory,, efficient row slicing, and fast matrix vector products.

RMSLE of two models was calculated, LGBM and Ridge regression.

VI. Result and Analysis

Following observations were made about

- Price distribution:
  - The mean price in the dataset is 26 Dollars
  - The median price in the dataset is 17 Dollars
  - The max price in the dataset is 2000 Dollars
  - Due to the skewed dataset, the median price is a more reliable price to gauge off of.

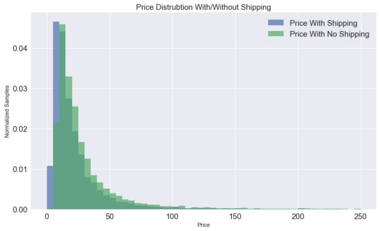

Fig. 1. Price distribution with or without shipping

- Top 20 Brand Distribution: Majority of the top brands are clothing brands and electronics. PINK and Victoria Secret are among the top 3 brands and are typically towards female customers.

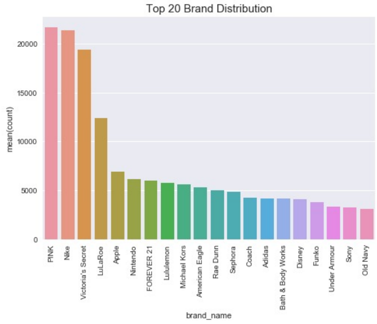

Fig. 2. Price distribution with or without shipping

- Main Category: It is seen that majority of the distribution is taken by women and beauty. They take 56% of the distribution. The prices are evenly distributed across all categories. The Men category the only one that averages out the most.

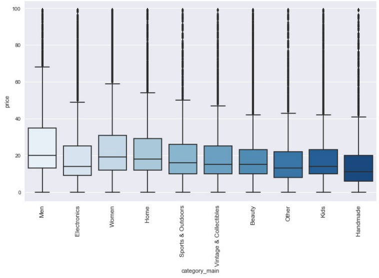

Fig. 3. Price distribution with or without shipping

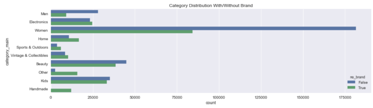

Fig. 4. Price distribution with or without brands

- Second category: It consists of electronics and female products.

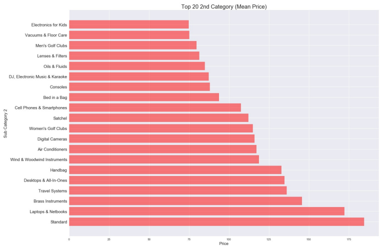

Fig. 5. Price distribution with or without shipping

- Effect of word count of item description on prices: It was observed that from about 0-300 words, there is a positive linear relationship. After that a gradual negative relationship was seen. It drops at about the 1000 word point. Overall, the word count did not have any significant role in deciding the price of an item.

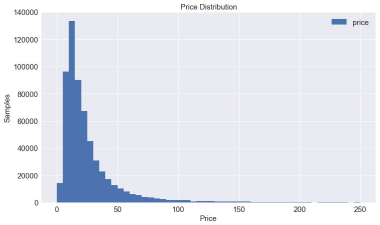

Fig. 6. Word count vs price

RMSLE for LGBM came out to be 0.54 and for Ridge regression, 0.48.

VII. Conclusion

In this paper, we have presented a comparative study of two machine learning algorithms, LightGBM and Ridge Regression, for e-commerce price prediction tasks. Additionally, we have evaluated their performance using the RMSLE metric, which is particularly useful for predicting logarithmic values.

Our experiments show that Ridge Regression outperforms LGBM in terms of RMSLE score, achieving a lower error in predicting product prices indicating that Ridge Regression is a more effective algorithm for e-commerce price prediction tasks.

Furthermore, we have shown that using RMSLE as a performance metric is essential for evaluating models that predict logarithmic values. The metric ensures that the model's errors in predicting smaller values are penalized more heavily, which is particularly relevant for e-commerce applications where accurate pricing can have a significant impact on business revenue.

VIII. Future Prospects

Predicting product prices is a crucial task in many industries, including e-commerce, retail, and finance. With the increasing availability of large-scale data and the advancements in machine learning techniques, the field of product price prediction is expected to grow in the future.

There are several directions in which future research on product price prediction could go:

Incorporation of external data: Incorporating external data, such as social media data and economic indicators, could enhance the performance of product price prediction models. These data sources could provide additional information that could improve the accuracy of the models.

Real-time price prediction: Real-time price prediction is becoming increasingly important in industries such as finance and e-commerce. Future research could focus on developing models that can predict prices in real-time, enabling businesses to make timely decisions.

Preparing Frontend for seller accessing: We can prepare a frontend through which seller can interact with our product and make predictions.

References

[1] https://lightgbm.readthedocs.io/en/v3.3.2/Features.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
[4] https://www.nltk.org/howto/stem.html
[5] https://seaborn.pydata.org/