Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765863

PRODUCT PRICE SUGGESTION

Rida Shahwar
Post Graduate Scholar, Department of Computer Science, GITAM School of Technology
Hyderabad, India, rshshwar@gitam.in

Dr. Y. Md. Riyazuddin
Assistant Professor, Department of Computer Science, GITAM School of Technology
Hyderabad, India, rymd@gitam.edu

Abstract— Due to the pandemic's intensification, it was estimated that online sales will account for 18.1% of all retail sales globally in 2021. The days of physically checking items in stores, exploring multiple locations, and searching for the appropriate item in the right place are long gone. According to statistics, many store customers use cell phone devices to check prices. As a result, online retailers need to set the proper pricing for their products. Knowing how much a product is worth can take time and effort. Minute details greatly influence the price. Considering the volume of things being sold online, pricing products is increasingly challenging at scale. Seasonal changes in clothing pricing trends differ from those in electronic product prices, which depend on the technology and features of the product. Additionally, brand recognition significantly impacts product pricing [1].

The field of recommendation systems is rapidly developing in web services and e-commerce applications. Searching through various products takes a lot of time when purchasing online. A recommendation system helps expedite the discovery of an extensive range of goods that clients are interested in. As it is simple and dependable for a consumer to purchase online and locate the ideal solutions for them without any hassles, the use of this effective suggestion system is growing day by day. [2]

We are utilizing the Mercari Dataset, Japan's largest community-powered shopping app. We would develop an algorithm that automatically recommends the appropriate product prices. The vendor and the buyer will benefit from this in terms of pricing, which should result in more transactions—product descriptions provided by customers that include information on the product category, brand, and condition.

The suggested model includes dynamic pricing and recommended prices for the products. The following machine learning algorithms are used in this model: Decision Tree Regressor for tree-based models, Ridge Regression to reduce training error while balancing model complexity, Lasso Regression to reduce squared loss with l1 regularisation, and Light GBM, which has a quick learning time, high efficiency, good accuracy, and good compatibility with large data sets. To improve the efficiency Deep Learning Algorithm, simple RNN is used to compare. (Abstract)

Keywords— E-commerce, Dynamic pricing, Machine learning algorithms, Lasso regression, Ridge Regression, Decision Tree Regressor, Light GBM, Deep learning algorithm, RNN

I. INTRODUCTION

Consumers can deal with vendors directly through electronic commerce using a web browser and the Internet (EC). Customers can locate what they're looking for by going to the company's website or a shopper's search engine, which displays the accessibility of identical things at various e-vendors. Consumers can shop online using multiple devices, including desktop, laptop, tablet, and smartphone computers and technology. [2]

Online shops are typically open 24 hours a day. All day long, people can use the internet from their homes or places of employment. Therefore, customers are not required to visit a store. Text, images, and other multimedia files are required for product descriptions in online stores. Some internet retailers offer additional product details, such as directions, safety measures, testaments, or manufacturer descriptions, or they link to them. [3]

Some offer background details, suggestions, or how-to manuals to aid consumers in making purchasing decisions. In certain establishments, customers can even remark on or review the products.

Customers may even remark on or review products in some stores. Discovery shopping engines and online price comparison services can be used to find vendors for a specific item or service. Another benefit of shopping online is that it allows you to swiftly seek bargains for goods or services offered by numerous vendors.

Filtering, prioritizing, and effectively delivering important information are necessary to reduce information overload on the Internet, where the sheer volume of options can potentially be problematic for many Internet users. To offer users content and services, recommender systems scan massive amounts of dynamically created data. An intelligent clothing suggestion system was conceived and developed in this work to account for more individualized clothing needs. According to the user's preferences, a variety of recommended products are presented to them using Transfer Learning to extract rich information from the product photos and the cosine similarity technique. [4]

E-commerce websites use recommendation systems to increase sales by recommending products to customers. If the website sells things under a new brand, this can be done successfully because new brand products have uniform prices across the board for e-commerce platforms. But there are some as well. [3]

Things sold on e-commerce sites that solely sell products for resale come from sellers who have already utilized those products. As a result, the prices of these products are not standardized because the seller of the resale items sets them. These websites only include some of the products because they only sell products for resale; hence there is no recommendation system on these websites.

The purpose of this study is to normalize pricing across all items by predicting product prices based on the product's features.

Predictive analysis is a process that analyzes historical and present-day data to extract information and forecast future events. It combines machine learning algorithms, data mining techniques, and statistical methodologies.

The goal of predictive analytics is to foretell actual product pricing. One of the key components of ai technology, known as machine learning, enables the creation of computer systems that can learn from the past without the need for programming in every situation. To assist automation with the fewest possible errors and minimize human labor, machine learning is seen to be urgently necessary nowadays.

We are utilizing the Mercari Dataset, Japan's largest community-powered shopping app. We would develop an algorithm that automatically recommends the appropriate product prices. The vendor and the buyer will benefit from this in terms of pricing, which should result in more transactions—product descriptions provided by customers that include information on the product category, brand, and condition. [3]

The suggested model includes dynamic pricing and recommended prices for the products. The following machine learning algorithms are used in this model: Decision Tree Regressor for tree-based models, Ridge Regression to reduce training error while balancing model complexity, Lasso Regression to reduce squared loss with l1 regularisation, and Light GBM, which has a quick learning time, high efficiency, good accuracy, and good compatibility with large data sets. A simple RNN, a Deep Learning Algorithm, is used to make the comparison to improve efficiency. Recurrent neural networks, or RNNs, include models based on LSTMs or GRUs. For data with extensive sequences, like those seen in our textual data, RNN models are used. [4]

II. RELATED WORK

While not a new idea, individual-level pricing discrimination is becoming more than just a theoretical possibility in the Online world. According to the industrial theory, demand pricing, also known as specific price discrimination, naturally increases a company's profitability because it enables it to capture a more significant portion of the purchaser's excess. According to observational data from recent merchandising tests with Online-based demand pricing, shoppers appear to abhor this strategy. In the present investigation, we investigate how the participation in a dynamic pricing event and the intention behind the pricing bias (i.e., whether a person is displayed with a more significant or a lower cost) affect the median levels of trust as well as the importance placed on the different dimensions of trust when determining overall trust. It does this by conceptualizing trust as having two dimensions: competence and goodwill. It is anticipated that trust levels will be lower and that benevolent trust will be accorded greater significance because growth pricing, such as dynamic pricing, is frequently thought to be unjust. The findings demonstrate that benevolence trust is given much higher value when forming a general perspective, and that mean benevolence trust is significantly lower (which results in a modest drop in total confidence). The direction-of-price-discrimination implications are generally not supported. [5]

The price prediction model for second-hand goods has been deployed on the eBay website by authors Greenstein-Messica and Rokach (2018). The price of the goods and the marketing indication, following the author's reputation as the seller, are significant factors that can increase the prediction model's effectiveness. The author used eBay transactional data from the previous six months for the model's implementation. Content-aware price factorization is the price prediction model (CAMF). Compared to the author's matrix factorization recommendation model, deployed before, the seller's reputation component increased the model's f1-score to 84%, a relatively high score. [6]

Given how volatile e-commerce pricing is, Bauer and Jannach (2018) created a method to optimize the price of sparse and noisy e-commerce data. The author accomplished this using two variables: price variations over time for a particular product and for related items. Bootstrap-based confidence is implemented in addition to kernel regression and Bayesian confidence inference. Because of the author's methods, profit increased by 28.04 percent of revenue in just four months. [7]

A model for product recommendations has been developed by authors Guo et al. (2018) that takes into account pricing and multi-category inter-purchase duration. The author employed sequential pattern mining to examine how customers' preferences for various products change over time. The author split the implementation process into two sections: the category search phase and the product search phase. For all categories the user has an interest in, sequential pattern mining is done throughout the category search phase, performed for all users. Next, By matching the trend of items purchased for every category and multiple categories throughout the intervals between purchases, the prices for each kind are calculated. The flow time of a potential product is computed during a proposed time in product recommendations. A price element is also determined for the suggested item similarly. The user's preference value for that specific product is then calculated by integrating the matching time and price. For this implementation, the results are obtained using a fuzzy set theory. [8]

A method that can forecast an item's price after a specified number of days has been devised. The user searches for the item on a particular online store, and the algorithm learns from the pricing history list to anticipate the price. For making predictions, linear regression models and polynomial regression models have been employed. [9]

One key factor influencing profitability in the cyber business sector is pricing, which is closely related to the company's sales. The development of a machine learning tool for pricing recommendations is required by the growth of online markets; this study's objective is to choose a machine learning model to develop a price recommendation tool for e-commerce businesses. On a dataset from an e-commerce company, Linear Regression, Random Forest, and LightGBM are applied to see how well models perform when used on data with various features and plenty of rows. The study also uses procedures to enhance the result of the models, such as parameter tuning, feature selection, or value removal. The results show that Light GBM outperforms Grid search CV regarding processing speed and prediction inaccuracy. [10]

To decrease the issue of information overload, which has the potential to cause problems for many Web users, it is vital to sift, prioritise, and efficiently communicate vital online information, where the number of options is overwhelming. Recommender systems, which filter through enormous volumes of dynamically produced data to give clients personalized content and services, tackle this problem. In this study, an intelligent clothing suggestion system was built and developed, considering increasingly individualized clothing requirements. The user is presented with various recommended products based on their selections by using Transfer Learning to extract rich information from the product photos and the cosine similarity technique. This was done on a web application for online shopping, where customers can select from various clothing options in our database. To help with tailored suggestions, only photographs with 80% or higher similarity to the user's chosen product are suggested to users. [11]

It is advised that using other image-based item information more effectively extracts representative visual elements from the photographs to produce price recommendations for used things with their pictures (e.g., category, brand). Then, we design a perception price suggestion module that uses the extracted feature representations along with some statistical significant improvements from the shopping platform as inputs. This module uses a binary classifier algorithm to determine whether a submitted component picture qualifies for price recommendation and a regression model that can provide price specific recommendations for items with qualified picture. Two distinct objective functions are suggested to jointly optimize the classification model and the regression model in accordance with the two requirements of the platform operator. According to the two needs of the platform operator, 2 different objective functions are presented to improve the classification algorithm and the regression method. It also provides a positive training technique for joint enhancement to improve the training of these two models. Our perception pricing estimation method has been put to the test extensively on a sizable real-world dataset. [12]

Price Forecasting helps sellers set accurate and reasonable prices for their used items using the images and textual information they provide to shopping websites. We specifically create a multi-modal price suggestion system that accepts the extracted visual and textual features as input along with some statistical item features gathered from the second-hand item shopping platform to determine whether the uploaded second-hand item's image and text are qualified for reasonable price suggestion with a binary classification model, and provide price recommendations for second-hand items with qualified image data and text descriptions with the system. To further enhance the regression model and provide pricing recommendations for used commodities, a customized loss function is developed. This can maximize the benefit for the owners and streamline online purchases. In order to evaluate the proposed pricing recommendation system more efficiently, we additionally create a series of criteria. Extensive testing on a huge validation set has shown the value of the proposed inter price recommender system. [13]

III. ABOUT DATASET

The dataset was obtained from Kaggle and consisted of two files: a learning file and a test file. 1.4 million plus entries are in the training file, and around 3 million are in the test file. The files are zipped and in the .tsv format, measuring 74MB and 280MB, respectively.

For the training dataset, each row represents a product offered and includes related data. Each column displays information related to the product. There are seven columns altogether, and they are as described below:

- Name: Shows the name of the recorded item.
- Item condition id: The state of the item as described by the vendor. The categories for this feature range from 1 to 5, with 1 being preferred to 5. It is an ordered categorical feature.
- Category name: This attribute keeps track of the category name. For example, "Women/Athletic Apparel/Pants, Tights, Leggings."
- Brand name: It provides the brand's title. Null values also exist.
- Price: The attribute we wish to forecast and the dependent variable is the price, specified in US dollars ($).
- Shipping: This feature(nominal) specifies who pays the shipping costs. If the vendor pays for shipping, it is a "1," and if the buyer does, it is a "0."
- Item description: This feature includes the item's complete information.

IV. PROPOSED METHODOLOGY

A. Exploratory Data Analysis

This step provides the opportunity to examine and analyze the dataset, which makes it crucial for resolving a machine-learning problem. This profound knowledge of the dataset is crucial for developing new characteristics to add to the machine learning model and understanding the already present characteristics. Now that we are aware of our motivations, let's begin EDA.

1) Depending on the price for the target variable: This distribution (Fig. 1) offers a number of benefits over the bare cost due to its resemblance to the Normal Distribution.

In contrast to the log-normal distribution, which has fewer data points for higher cost values, the data points in the Probability function are substantially more uniformly distributed over all ranges. It is advantageous in algorithms like linear regression, where the distribution of the dependent variables is assumed normal. Furthermore, Gaussian Distributed targeting variables also provide better performance for other models.

Since we must assess the log of the predicted and factual prices anyhow to take into account our performance metric, it is favored to use log(price+1) as our target worth and calculate RMSE over it. Log(price+1) will therefore be our outcome variable. Some products have a cost of 0 dollars.

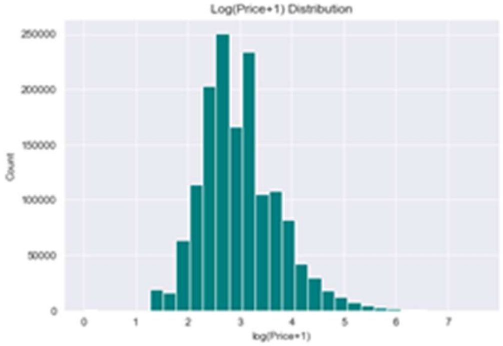


Fig.1. Log-Normal Distribution of price

B. Benchmark Model

Here, we are utilizing this model to assess the performance of our ML models. The comparison model we're using is a trouble-free standard model that determines the mean output based on the two attributes of "shipping" and "item condition." When we enter the shipment (0 or 1) and item condition id, the algorithm retrieves all the figures with comparable values (1 to 5). It computes the outcome by averaging these points.

This model's RSMLE value of 0.7267, which serves as a benchmark, was tested using the validation data. A model is underperforming if its metric is higher than this value.

C. Model 1

To train the initial model more quickly, we strive to make the output vector's dimension as small as possible. Thus, average Word2Vec is applied to text data, while Label Encoding is applied to categorical data.(Fig. 2)

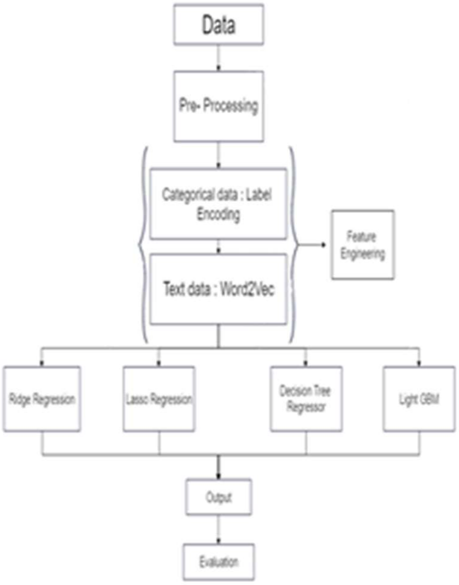


Fig. 2 : Model 1 Architecture

The following machine learning models are used:

a) Lasso Regression

This model is a linear one that uses l1 regularization to reduce squared loss. For this model, the Sklearn package is used. The model is trained using the optimal parameters after being hyperparameter tuned across a range of values.

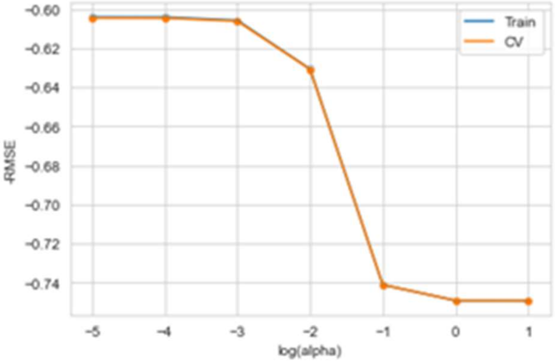


Fig. 3. : Performance of Lasso regression

The model's performance is calculated as 0.6037 on the validation set and 0.60476 on the test set.

b) Ridge Regression

Another linear model that lowers squared loss is the Ridge model, which also applies l2 regularization. (Fig. 4)

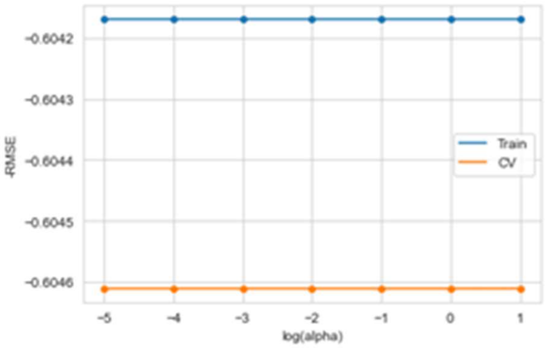


Fig. 4. : Performance of Ridge Regression

On validation data, this model's RMSLE is 0.60379, and on test data, it is 0.6048.

c) Decision Tree Regressor

As the name implies, the decision tree employs a tree-based model to forecast the results. Sklearn's Decision tree model is applied in this case.

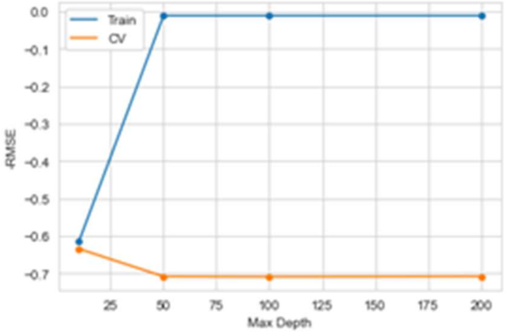


Fig. 5. : Performance of Decision Tree Regressor

The test data's RMSLE is 0.63648, while the validation data's RMSLE is 0.6311. When compared to more traditional models, the decision tree might perform better.

d) Light GBM

Light GBM models have ensemble model features and are easier to train; hence they are appropriate for this problem.

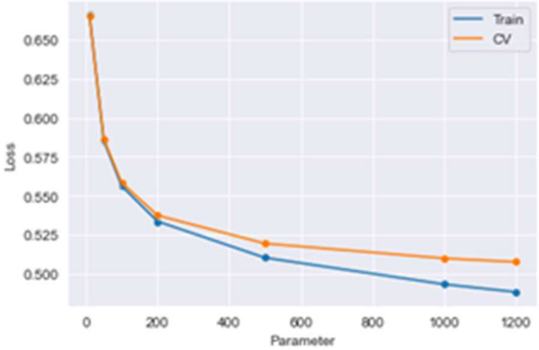


Fig. 6. : Performance of LightGBM

This model's performance on test data is 0.50050, while its performance on validation data is 0.49929.

D. Model 2

In the second method, we use Tfidf for text data and One Hot Encoding for categorical data.

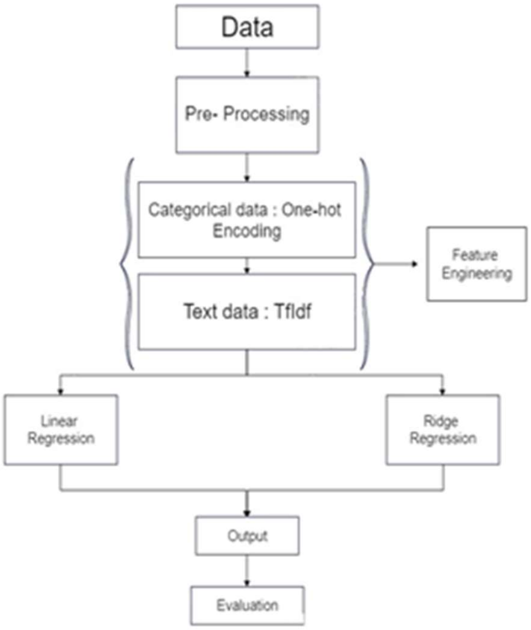


Fig. 7. : Model 2 Architecture

The machine learning models here we are using are:

a) Linear Regression

The squared loss is reduced by this use of linear regression. Sklearn's version of linear regression is used in this case.

A key figure of 0.4621 for test data and 0.4620 for validation data are produced by this straightforward model.

b) Ridge Regression

The Ridge Model, which includes tfidf and one hot encoding, is a hyperparameter tuned on several constraints.

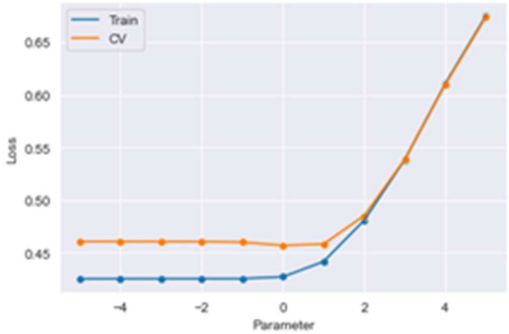


Fig. 8. Performance of Ridge Regression(Model 2)

We obtain the RMSLE of 0.4581 on Validation set and 0.45831 on Test set by training on the best parameters.

E. Model 3

A resilient approach to problem-solving, deep learning algorithms use various layers between the input and output stages of the network. Additionally, the idea's functioning is made concrete by its similarity to the human brain.

Here, the DL model, simple RNN, is used. Recurrent neural networks, or RNNs, include LSTM- or GRU-based models. For data with extensive sequences, like those seen in our text data, RNN models are used (name and item description). As a result, these features utilize the GRU architecture.

Data must be vectorized before being sent to our network. Keras offers the embedding layer to data that is vectorized. However, it takes inputs in a specific format. First, we must tokenize our text data, which entails converting it into integers; they are simply the words' index values in a vocabulary acquired by fitting the train data.

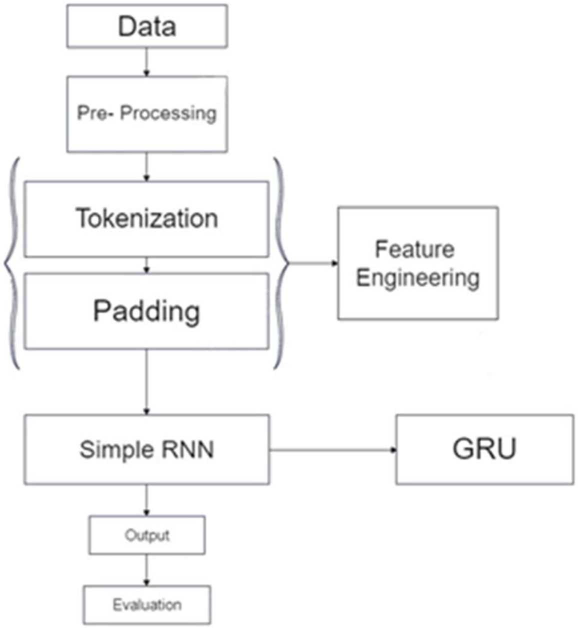


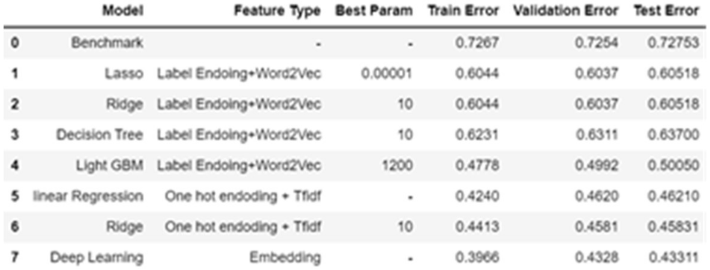
Fig. 9. Model 3 Architecture

When it comes to deep learning, the model's architecture is the most crucial component for improving performance. Model training is carried out using Adam Optimizer, with MSE as the loss and RMSE as the chosen measure. Model Checkpoint call-backs, a scheduler for the gain, and early stopping are also introduced.

The model is set up for 10 epochs; however, because of the early stopping call back, it is terminated after 4 epochs.

The performance metric for the deep learning model is 0.4331 for test data and 0.4328 for validation data.

V. MODEL COMPARISON



	Model	Feature Type	Best Param	Train Error	Validation Error	Test Error
0	Benchmark	-	-	0.7267	0.7254	0.72753
1	Lasso	Label Endoing+Word2Vec	0.00001	0.6044	0.6037	0.60518
2	Ridge	Label Endoing+Word2Vec	10	0.6044	0.6037	0.60518
3	Decision Tree	Label Endoing+Word2Vec	10	0.6231	0.6311	0.63700
4	Light GBM	Label Endoing+Word2Vec	1200	0.4778	0.4992	0.50050
5	linear Regression	One hot endoding + Tfidf	-	0.4240	0.4620	0.46210
6	Ridge	One hot endoding + Tfidf	10	0.4413	0.4581	0.45831
7	Deep Learning	Embedding	-	0.3966	0.4328	0.43311

Fig. 10. Model Comparision Table

- The Ridge Model, which combines One Hot Encoding and TFIDF, is superior than Models 1 and 2.
- Complex models, including For features with label encoding and word2vec, light GBM performs superior than straightforward linear algorithms like a lasso or ridge.

- However, Tdidf features increase proportionally, and linear models provide reasonable performance metric values.
- To improve the results, we perform deep learning in Model 3.
- Deep Learning turns out to be the most effective model.
- A simple RNN-based model performs better than machine learning methods, which provide good performance metrics.

VI. CONCLUSION

Machine learning and deep learning are the next two buzzwords in this era of artificial intelligence (AI). AI is present everywhere; in particular, it is essential to human existence. Modern machine learning and deep learning algorithms can resolve any foreseeable problem.

The retail and e-commerce industries have grown to rely on machine learning and deep learning. The effectiveness of machine learning and artificial intelligence technologies has made their insights into enterprises essential for development and success.

For product pricing recommendations, we use the RMSLE (Root Mean Squared Logarithmic Error) measure. The RMSE gives significant errors a fair amount of weight. Therefore, when big mistakes are undesirable, the RMSE should be more helpful. RMSE has the advantage of penalizing significant errors more, making it more suitable.

Above are the results according to which we have analysed that the Deep Learning Model is the best model with RSMLE of 0.43311.

In this project, we have gone through a case study in data science where we have comprehended the problem statement, performed exploratory data analysis, transformed features, selected ML models, performed random search along with hyperparameter tuning, and assessed them on the test set. After choosing the DL model, we performed tokenization and padding, applied RNN, and compared the overall performance.

VII. REFERENCES

[1] C. p. Rajan Gupta, "A Machine Learning Framework for Predicting Purchase by online customers nased on Dynamic Pricing," procedia Computer Science, vol. 36, pp. 599-605, 2014.

[2] A. R. A. R. J. H. C. S. K. K. A. Soumya Wadhwa, "Personalizing Item Recommendation via Price Understanding," Wadhwa, et al, p. 8, 2020.

[3] A. R. Chada, "Strategic Pricing of Used Products for e-Commerce Sites," 06 September 2019. [Online]. [Accessed 18 December 2022].

[4] P. S. M. Anupama Namburu, "Product pricing solutions using hybrid machine learning algorithm," Innovations in Systems and Software- Springer, pp. 1614-5054, 2022.

[5] G.-C. C. Saul Abraham, "Pricing Recommendation by Applying Statistical Modeling Techniques," 28 June 2017. [Online]. [Accessed 3 January 2023].

[6] S. Pereira, "Price Suggestion and Recommendation of Resale Products on," 10 August 2019. [Online]. [Accessed 08 november 2022].

[7] B. Su, "An Application of Recurrent Neural Network: Prediction of Items' price by description," Proquest, p. 43, 2018.

[8] C. Liu, "Prediction and Analysis of Artwork Price Based on Deep Neural Network," Scientific Programming, p. 10 pages, 2022.

[9] O. F. L. Ellen Garbarino, "Dynamic Pricing in Internet Retail: Effects on consumer Trust," Wiley InterScience, pp. 495-513, 2003.

[10] L. R. Asnat Greenstein- Messica, "Personal price aware multi-seller recommender system: Evidence from eBay," Knowledge Based Systems, pp. Volume 150 page 14-26, 2018.

[11] D. J. Josef Bauer, "Optimal pricing in e-commerce based on sparse and noisy data," Decision Support Systems, pp. Volume 106, Pages 53-63, 2018.

[12] Z. G. N. L. Y. W. Junpeng Guo, "Recommend products with consideration of multi-category inter-purchase time and price," Future Generation Computer Systems, pp. Volume 78, Part 1, Pages 451-461, 2018.

[13] S. K. B. Priyanka Banarjee, "Prediction of Price of products by Machine Learning Method through Web based System," International Journal of Trend in research and Development, pp. Volume 7(2) pages 149-152, 2020.

[14] U. N. T. T. A. D. V. N. Linh Nguyen Tran Quang, "E-commerce Price Suggestion Algorithm- a Machine Learning Application," in Proceedings of the 2nd Indian International Conference on Industrial Engineering and Operations Management , Warangal, Telangana, 2022.

[15] V. S. A. K. Akshit Tayade, "Deep Learning Based Product Recommendation System and its Applications," International Reasearch Journal of Engineering and Technology(IRJET), pp. Volume 08 Pages 1317-1323 , 2021.

[16] Z. Y. Z. X. L. G. M. T. R. J. Liang Han, "Vision-based Price Suggestion for Online Second-hand Items," Social computing and Image processing, vol. Session 4C, pp. 1988- 1996, 2019.

[17] Z. y. Z. X. M. T. R. J. Lian Han, "Price Suggestion for Online Second-hand Items with Texts and Images," Emerging Multimedia Applications, vol. Poster Session C2, pp. 2784-2792, 2020.