Using machine learning technology to predict Etsy product attributes: classification and color recognition

Yibo Sun
School of Computing, Dublin City University, Ireland

***Abstract*—This study aims to explore the application of machine learning technology in predicting product attributes on the Etsy platform, including classification and color recognition. As a global online marketplace, Etsy's unique and creative products attract nearly 100 million creative buyers and 7.7 million sellers around the world. As the Etsy market expands rapidly, product classification and search efficiency have become key technical challenges. This study develops a machine learning-based model with the goal of accurately predicting a product's top-level category ID, bottom-level category ID, primary color ID, and secondary color ID. We improve the efficiency of the product classification and search functions of the Etsy platform by utilizing existing product data to train models. In terms of methodology, this study uses a variety of machine learning algorithms such as support vector machines and lightweight gradient boosting machines (LGBMClassifier) for model training. By comparing the performance of single feature prediction and comprehensive feature prediction, this study reveals the differences in the effectiveness of different models in processing text and color recognition tasks. Experimental results show that the model performs well in predicting top classification IDs, with an F1 score reaching 0.8259. However, in the color recognition task, the model's performance is weak, especially in the prediction of secondary color ID, the F1 score is only 0.3396, which hints at the complexity of color recognition and the limitations of existing models. In summary, although existing technologies show strong potential in some areas, challenges still exist when dealing with complex tasks such as color recognition. Future research needs to further explore advanced image processing technology and deep learning models, as well as optimize data preprocessing and feature engineering methods, so as to comprehensively improve the accuracy and efficiency of Etsy product attribute prediction.**

***Index Terms*—Etsy, machine learning, product attribute prediction, classification and color recognition, SVC, LGBMClassifier, multi-label classification, performance evaluation**

I. Introduction

Etsy is a global online marketplace focused on selling one-of-a-kind handcrafted and vintage treasures. It connects nearly 100 million creative buyers and 7.7 million sellers around the world, providing them with a platform to showcase and purchase hand-created and carefully curated goods. On Etsy, countless transactions and searches happen every second, covering a variety of categories from wall art to wedding-related merchandise. Such a large and diverse market not only brings huge business opportunities, but also huge challenges in managing and optimizing these product listings.

As the Etsy market expands rapidly, product classification and search efficiency have become key technical challenges. Currently, there are a wide variety of products on the market and they are updated rapidly, which requires Etsy to have an efficient mechanism to ensure that buyers can quickly and accurately find the products they want. In addition, correct color identification and classification not only improves user experience, but also increases the likelihood of transactions, as color and classification are important factors that influence buyers' purchasing decisions. Therefore, it becomes particularly important to develop a machine learning model that can automatically predict and classify the attributes of newly launched products.

Fig. 1. Etsy Wordcloud

This study aims to develop a machine learning-based model that can accurately predict the top-level category ID, bottom-level category ID, primary color ID, and secondary color ID of products on the Etsy platform. By using existing product data to train the model, we expect the model to show high accuracy on unseen test data, thereby supporting the product classification and search functions of the Etsy platform. In addition, this study will also explore the impact of different feature extraction technologies and model architectures on prediction performance, providing scientific basis and technical support for Etsy's future product recommendations and search engine optimization.

II. Related work

A. Existing research

In the field of e-commerce, machine learning is widely used in product classification and recommendation systems. For example, Boxin Du et al.(1) explores how to pre-train hypergraph structures through graph neural networks to enhance the efficiency and accuracy of product classification. Xiaojun Wu et al.(2) introduces the use of improved SMOTE and AdaBoost algorithms to predict user churn on e-commerce platforms. Although this is not directly about product classification, it involves the use of classification technology. Malik et al. (3)uses NLP and machine learning algorithms to optimize the e-commerce product recommendation system, showing the importance of text features in product recommendation.

Aryafar et al. (4)uses ensemble learning method to predict the click-through rate of promoted products on the Etsy platform, which indirectly involves the impact of product classification effect on click-through rate. Deniz et al. (5)uses machine learning to conduct multi-label classification of e-commerce customer reviews to help understand consumers' multi-dimensional evaluation of products. Finally, the Lynch et al. (6)extracts visual semantic features through deep learning and is used to improve the ranking task of modal learning. Although it focuses more on image processing, its method can inspire the development of product color recognition technology(7).

B. *Research gap*

Although the above literatures have made certain progress in their respective fields, they generally have some shortcomings. First of all, many studies focus on recommendation systems and user behavior analysis, rather than directly focusing on the identification and classification of product attributes. In particular, there are few applications in color recognition. Furthermore, there are insufficient specific descriptions of how to effectively process unstructured data and how to apply these methods in real-time systems, which limits the scope of practical applications of these techniques.

For example, although the Boxin Du et al.(1) and Lynch et al.(6) demonstrate advanced image and text processing techniques, they do not adequately describe the specific applications of how to quickly and accurately perform product classification and color recognition in actual e-commerce environments. In addition, most studies fail to fully consider the dynamics of real-time data streams, and the ability to process dynamically updated product information on e-commerce platforms needs to be improved.

Therefore, future research needs to focus more on how to directly apply these advanced machine learning technologies to product classification and color recognition on e-commerce platforms, especially in improving the flexibility and adaptability of processing unstructured data. And how to integrate these technologies into e-commerce systems in real time to improve user experience and operational efficiency.

III. Methodology

A. *Dataset description*

The data set used in this study contains product information on the Etsy platform. The data is stored in Parquet format and covers detailed descriptions and classification tags for various products. The data set is divided into training set and test set, which are used to train and evaluate the performance of the model. In the data preprocessing stage, we first read multiple Parquet files and then merged them into a single data frame to facilitate subsequent data processing and analysis.

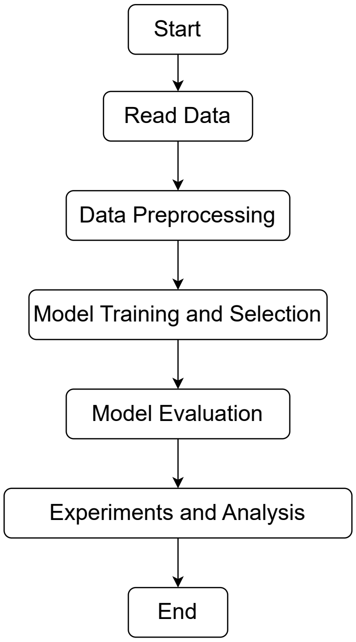

Fig. 2. Methodological flow chart

B. *Feature engineering*

The main task of feature engineering is to extract useful information from original text data for model learning. In this study, feature extraction includes the following steps: 1. Text vectorization: Use CountVectorizer to convert product titles into word frequency vectors. This step helps us convert text data into numerical data that the model can process. 2. TF-IDF conversion: Apply TfidfTransformer to further process the word frequency vector and convert it into TF-IDF format, which is the word frequency-inverse document frequency of the text. This helps to reduce the impact of common words on the model and enhance the model's ability to capture key information.

C. *Model selection and training*

In the model selection and training part, we used two different machine learning algorithms:

1. Support Vector Machine (SVC): SVC with linear kernel is selected for classification. This model is suitable for processing medium-sized data sets. During the training process, we adjusted the regularization parameter C to optimize the model to achieve better classification results.

2. Lightweight Gradient Boosting Machine (LGBMClassifier): For large-scale data processing, we chose LGBMClassi-

fier. This tree-based model improves performance by building multiple decision trees and is suitable for processing data sets with complex feature interactions. Model parameters such as the depth of the tree and the number of leaf nodes are carefully adjusted to optimize the model's learning efficiency and prediction accuracy.

These models are integrated into the Pipeline, realizing an automated process from data preprocessing to model training.

D. Assessment method

The performance evaluation of the model uses cross-validation and multi-index evaluation methods: 1. Data splitting: Randomly divide the data into a training set and a validation set to ensure the fairness and accuracy of the evaluation.

2. Performance Metrics: Evaluate model performance using metrics such as precision, recall, and F1 score. These indicators help us understand the performance of the model from multiple dimensions, especially the responsiveness to different categories in classification tasks.

IV. Experiments and results

A. Experimental setup

This study used machine learning technology to predict product attributes on the Etsy platform, including the highest level category ID, the lowest level category ID, primary color ID, and secondary color ID. The experiment was conducted in the Anaconda environment containing TensorFlow and LightGBM libraries, running on a computer system equipped with high-performance CPU. The experiments began with individual predictions for each feature and subsequently used an integrated model to predict all four attributes.

B. Results display

Table I shows the results of the experiment in detail, including key performance indicators such as training time, prediction time, and F1 score.

TABLE I
Single model performance index table

Attribute	Training Time (s)	Prediction Time (s)	F1	Acc.
Top Category ID	2712.38	174.09	0.8231	0.8259
Bottom Category ID	606.411	0.630	0.5285	0.5441
Primary Color ID	5637.92	293.96	0.4482	0.4529
Secondary Color ID	162.91	2.93	0.3341	0.3396

Table II shows the results of the comprehensive model that simultaneously predicts four attributes.

TABLE II
Comprehensive model performance index table

Measurement	Value
Prediction Time	4325.69
Accuracy	0.4968
Precision	0.4914
F1 score for Top Category ID	0.7301
F1 score for Bottom Category ID	0.5240
F1 score for Primary Color ID	0.4338
F1 score for Secondary Color ID	0.2776

C. Result analysis

From the above results, we can see that the highest level category ID has the best prediction performance, with an F1 score of 0.8231, indicating that the model is relatively accurate and reliable in predicting this category. Predictions for the lowest level category ID also performed well, with an F1 score of 0.5285. However, the prediction of primary color ID and secondary color ID performed poorly, with F1 scores of 0.4482 and 0.3341 respectively. This may be due to the high diversity and complexity of color attributes, which makes it difficult for the model to capture enough features for effective classification. In addition, the low accuracy of secondary color IDs may also be related to the imbalance in the dataset, and the number of samples in some color categories may be too small, resulting in insufficient model training. Overall, better performance is obtained when predicting each attribute individually, but performance decreases when predicting all attributes combined. This suggests that further optimization of the model may be required, or algorithms more suitable for handling such multi-label classification tasks be explored. In addition, for the low performance of color recognition, future work can consider introducing more complex image processing techniques, or adding color-related features to improve prediction accuracy.

V. Conclusion

This study explores the performance of multiple models on different attribute classification tasks by applying machine learning technology to predict the attributes of products on the Etsy platform. Experimental results show that our model exhibits high accuracy and efficiency in predicting the highest-level category ID and the lowest-level category ID of a product. Especially for the prediction of the highest-level category ID, its F1 score reaches 0.8231, showing the strong potential of the model on such tasks. However, the prediction results of primary color ID and secondary color ID are unsatisfactory, which is mainly reflected in the lower F1 score and accuracy, which are only 0.4482 and 0.3341 respectively. This result hints at the complexity of the color recognition task and the limitations of current models in processing this type of data. Low performance in color recognition may stem from the model's failure to adequately capture key features for color discrimination, or from imbalance and diversity issues in the dataset itself. In addition, although the comprehensive model that predicts four attributes simultaneously provides operational convenience, its average accuracy and F1 score decreased, which may indicate a performance loss when a single model handles multiple tasks. This finding highlights the importance of selecting appropriate model configurations and optimization algorithms in practical applications. In summary, this study confirms the feasibility of using machine learning techniques for product attribute prediction and identifies the advantages and limitations of existing methods. Future work can consider introducing more advanced image processing technology and deep learning models to improve the accuracy of color recognition, while exploring more effective data preprocessing and feature engineering methods to deal with data imbalance problems. In addition, further research on how to optimize the model structure and training strategy to improve the performance of the model in a multi-task learning environment will also be the key to improving the performance of the product attribute prediction system.

References

[1] B. Du, C. Yuan, R. Barton, T. Neiman, and H. Tong, "Hypergraph pre-training with graph neural networks," 2021.

[2] X. Wu and S. Meng, "E-commerce customer churn prediction based on improved smote and adaboost," in *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, 2016, pp. 1–5.

[3] V. Malik, R. Mittal, and S. V. SIngh, "Epr-ml: E-commerce product recommendation using nlp and machine learning algorithm," in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 2022, pp. 1778–1783.

[4] K. Aryafar, D. Guillory, and L. Hong, "An ensemble-based approach to click-through rate prediction for promoted listings at etsy," in *Proceedings of the ADKDD'17*, ser. ADKDD'17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3124749.3124758

[5] E. Deniz, H. Erbay, and M. Coşar, "Multi-label classification of e-commerce customer reviews via machine learning," *Axioms*, vol. 11, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2075-1680/11/9/436

[6] C. Lynch, K. Aryafar, and J. Attenberg, "Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 541–548. [Online]. Available: https://doi.org/10.1145/2939672.2939728

[7] X. Xiahou and Y. Harada, "B2c e-commerce customer churn prediction based on k-means and svm," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, 2022. [Online]. Available: https://www.mdpi.com/0718-1876/17/2/24