Predictive Modeling of Product Attributes Using Machine Learning Techniques

Vaishnavi Kulkarni, 23266282
School of Computing, Dublin City University, Ireland
Email: vaishnavi.kulkarni2@mail.dcu.ie
CA684 Machine Learning Assignment - Spring 2024

***Abstract*—In the burgeoning field of e-commerce, accurately categorizing products and predicting their attributes, such as top and bottom categories and primary and secondary colors, is crucial for enhancing user experience, optimizing search engine functionality, and improving inventory management. This study explores machine learning's role in enhancing e-commerce by accurately predicting these attributes. This study has applied Naive Bayes, Logistic Regression, and Gradient Boosting models within a MultiOutputClassifier framework to address multi-label classification. The effectiveness of these models was assessed using the weighted F1 score, which highlighted their capabilities in predicting multiple attributes simultaneously. The results underscore the potential of these models to improve e-commerce operations by facilitating real-time product categorization, thus enhancing user experience and operational efficiency.**

I. Introduction

In the field of e-commerce, the precise categorization of products is pivotal for enhancing the user experience, optimizing search functionalities, and managing inventories effectively. As online marketplaces continue to grow, the challenge of efficiently categorizing and matching vast arrays of products becomes increasingly significant. Implementing machine learning techniques for automatic product categorization not only reduces the reliance on labor-intensive manual processes but also increases accuracy and efficiency [6].

Machine learning's role in text classification has evolved significantly with the advent of models that learn rich representations from large unlabeled text corpora before being fine-tuned on specific tasks. This methodology, derived from advances in Natural Language Processing (NLP), leverages techniques such as masked language modeling to pre-train models on extensive text data [4]. This approach has shown substantial benefits in tasks requiring an understanding of nuanced textual descriptions, particularly in low-resource settings.

Parallelly in the machine learning domain, the Gradient Boosting Classifier and the MultiOutputClassifier have been effectively utilized to tackle multi-label classification problems [5] and [18]. These classifiers are well-suited for scenarios where multiple attributes of a product, such as categories and colors, need to be predicted simultaneously, a common requirement in e-commerce platforms.

Furthermore, clustering techniques like K-means have been applied to enhance the feature selection process in various machine learning applications, with the Elbow method providing a data-driven approach to determining the optimal number of clusters [10] The application of these techniques in preprocessing ensures that models are trained on the most representative features, thereby improving prediction accuracy and model efficiency.

Building on this foundation, this paper investigates the application of machine learning techniques to predict multiple product attributes simultaneously, leveraging models such as Naive Bayes, Logistic Regression, and Gradient Boosting. Additionally, this paper explores the utility of K-means clustering and the Elbow method in refining the feature selection process for the Naive Bayes model. The K-means algorithm was utilized to cluster similar product titles, identifying inherent patterns and groupings that informed subsequent model training. The Elbow method further aided in determining the optimal number of clusters, thus enhancing the model's efficiency and precision in categorizing products based on their textual descriptions.

II. Related Work

The burgeoning e-commerce industry demands sophisticated tools for effective product categorization to enhance user experience and operational efficiency. Traditional rule-based classification systems have been largely supplanted by advanced machine learning techniques capable of managing complex and large datasets. These advancements have been critical as they not only improve accuracy but also automate and scale the processing of vast arrays of products [19]. As machine learning evolves, the focus has increasingly shifted towards models that can handle multiple labels per instance, essential for the multi-faceted nature of product data in online retail.

Significant progress in Natural Language Processing (NLP) has enabled more nuanced text analysis, crucial for extracting meaningful information from product descriptions. Embedding techniques that transform text into vector spaces have proven particularly effective, allowing classifiers to capture the subtleties and complexities of language used in product listings [13]. This methodological enhancement facilitates the extraction of relevant features for machine learning models, improving their predictive accuracy in classifying products into their respective categories and color schemes.

Moreover, the integration of distance metrics like the Jaccard index with robust classifiers such as XGBoost has refined

the ability of models to measure similarities between products, thus improving the granularity of product categorization [2]. The application of clustering techniques such as K-means and the Elbow method has also been instrumental. These methods help determine the optimal number of clusters, thereby enhancing feature selection and model efficiency by ensuring that classifiers are not overwhelmed by the high dimensionality of data [10].

Evaluating the effectiveness of these complex models in a multi-label classification setting necessitates robust metrics. The F1 score, which balances precision and recall, is particularly valuable in scenarios where false positives and false negatives carry different costs, making it a preferred metric in studies requiring nuanced performance evaluation [3]. This metric is essential for assessing the performance of machine learning models in e-commerce product categorization, where accurate classification directly impacts customer satisfaction and business operations.

III. Methodology

A. Exploratory Data Analysis

In a comprehensive exploration of the dataset provided for the project, delved into both training and testing datasets formatted in Parquet, a columnar storage file format that offers optimized data compression and encoding schemes. Each dataset consisted of 362 files, reflecting a substantial volume of data typical in e-commerce platforms, which necessitates efficient processing and analysis techniques for optimal performance.

The training dataset comprised 229,624 entries, each uniquely representing a product with a detailed set of 26 attributes including categorical, textual, and image data, indicative of the complexity and variety inherent in e-commerce product data. Attributes ranged from basic identifiers like product ID to descriptive tags, categories, color information, and multimedia in the form of encoded images, underscoring the multimodal nature of the data. This rich dataset not only provided a profound basis for training our models but also posed challenges typical of high-dimensional data spaces such as sparsity and feature correlation [7].

Initial analyses focused on understanding the distribution and cardinality of key features. We identified 15 top categories and 2,609 bottom categories, illustrating the hierarchical categorization common in e-commerce settings, which aids in nuanced product discovery and classification [9]. The color attributes were also extensively categorized, with 19 primary and multiple secondary color identifications, enabling detailed product visual descriptions crucial for customer experience and satisfaction. Here we can set input as categorical data present in parquet files or image data. Figure 1 shows the sample of images present in the data set after completion of Exploratory Data Analysis.

B. Data Preprocessing and Feature Engineering

The preprocessing of data is a pivotal step in any machine learning pipeline, especially in complex domains such as e-commerce where data heterogeneity can significantly impact model performance. Here methodology included comprehensive preprocessing and feature engineering strategies designed to prepare the extensive dataset for subsequent modeling efficiently.

Initial steps involved splitting the original dataset into distinct training and validation sets to ensure model validation under unbiased conditions. This split was stratified based on the 'bottom category text' to maintain proportional representation of product categories across both datasets, ensuring that the model would be generalizable across different types of products [7]. The training set consisted of approximately 206,661 products, while the validation set contained about 22,963 products, providing a robust basis for training and subsequent performance evaluation.

Feature engineering focused on handling both categorical and textual data, recognizing the critical role of such features in predicting product categories and attributes. Categorical variables like 'type', 'room', and 'material' were processed using a combination of simple imputation for missing values and one-hot encoding to transform them into a format suitable for machine learning algorithms. This approach not only handled missing values effectively but also converted categorical attributes into a more informative binary matrix format, facilitating easier pattern recognition by the models.

Textual content from fields such as 'title', 'description', and 'tags' underwent transformation through TfidfVectorizer, which converts text to a matrix of TF-IDF features. This method is particularly effective in diminishing the impact of frequently occurring words that might otherwise overshadow the more meaningful content in text data [9]. By setting 'max features' to 1000, we aimed to capture the most relevant terms while maintaining manageable dimensionality, thus enhancing computational efficiency.

The preprocessing pipelines were integrated using a ColumnTransformer, which allowed different transformations for the specified types of data within the same workflow. This setup not only streamlined the preprocessing of heterogeneous data types but also ensured that all features were processed in a manner that maximizes their predictive potential before being fed into the predictive models.

C. Methods Overview

1) Naive Bayes using Multinomial Classifier: This study, employed a Multinomial Naive Bayes classifier, adept at text classification, to predict product categories based on their ID's and titles from an e-commerce dataset. We utilized a pipeline integrating Count Vectorization and TF-IDF Transformation, standard in natural language processing, to convert textual data into a numerical format conducive to machine learning [12].

The classifier's effectiveness was quantified using the F1 score, revealing its proficiency in identifying product categories, a vital step corroborated by the established precision-recall balance in model evaluation [14]. Furthermore, adapting the classifier to a MultiOutputClassifier framework enabled simultaneous predictions across various product attributes,

Fig. 1: Sample Images of products after Exploratory Data Analysis

showcasing the method's alignment with multi-label classification's benefits in complex data environments [1]. This approach established a foundational benchmark for the performance of more intricate algorithms.

2) Logistic Regression: Leveraging the robustness of Logistic Regression for categorical data prediction, our study adapted this classic statistical method for the multifaceted task of classifying e-commerce product data. Embedded within a pipeline that incorporated text vectorization and TF-IDF weighting, our Logistic Regression model adeptly converted textual titles into a numerical matrix, a critical step in handling the nuanced language patterns of product descriptions. The use of a MultiOutputClassifier allowed the model to extend its predictive capabilities to multiple attributes concurrently, reflecting the complexity of real-world e-commerce scenarios [8].

The performance of our Logistic Regression model was rigorously evaluated using the F1 score to assess precision and recall, ensuring a balanced measure of model accuracy across various product attributes. Additionally, the model's classification effectiveness was visualized through confusion matrices, providing a clear graphical representation of the model's accuracy in predicting the correct product categories. These insights underscored the model's competency in text classification, cementing its role as a valuable tool in the predictive analysis of e-commerce data [15].

3) Gradient Boosting with XGBoost: In our study, we deployed the XGBoost algorithm, known for its effectiveness in handling complex multi-label datasets, which is essential for our e-commerce product categorization task. This advanced implementation of gradient boosting decision trees systematically refines models by focusing on errors from previous iterations, enhancing predictive accuracy for multiple outputs such as product categories and color attributes. We integrated XGBoost within a MultiOutputClassifier framework to allow simultaneous predictions across these varied attributes, aligning with the multi-dimensional demands of e-commerce data [2].

To ensure the robust performance of our model, we meticulously prepared and engineered features, transforming qualitative data such as product titles into a quantitative format suitable for XGBoost. This process involved encoding categorical variables and splitting the dataset into training and testing sets to evaluate the model's effectiveness accurately. Post-training, we assessed the model using classification reports and accuracy scores, which provided insights into its discriminative capabilities across different categories. Additionally, analyzing feature importances helped identify the most influential variables, guiding further model optimizations [7].

D. Word2Vec

Word2Vec, an innovative model designed to transform words into numerical representations, produces word embeddings where semantically similar words are encoded closer together in vector space. Leveraging the 'word2vec-google-news-300' model, we enriched the representation of product titles beyond simple one-hot encoding. This method allowed us to average word vectors in a document, resulting in a condensed feature vector per product title, well-suited for our machine learning models [16].

For this study, Word2Vec embeddings were pivotal in comparing the performance of a model using pre-trained embeddings with one that had been fine-tuned to our specific e-commerce dataset. The tailored fine-tuning of the model to our domain-specific dataset was instrumental in capturing the unique lexicon and subtleties of the e-commerce sector. The comparative performance was visually represented through a bar graph, highlighting the metrics across the pre-trained and baseline models. The graph revealed that while precision saw a marked improvement with pre-trained embeddings, the accuracy, recall, and F1 score were similarly matched, suggesting

the value of transfer learning and hinting at the potential for further refinement tailored to our dataset's particulars.

These observations indicate that the utilization of Word2Vec pre-trained embeddings can significantly enhance the precision of categorization in e-commerce platforms, allowing for a more nuanced and accurate representation of products. The comparative analysis underscores the benefits of embedding transfer learning, providing a foundation for improved performance in multi-class product categorization tasks.

E. Visualisations

1) UMAP using TF-IDF Embeddings: The UMAP (Uniform Manifold Approximation and Projection) visualization of TF-IDF embeddings serves as a powerful technique for distilling the high-dimensional data inherent in text analysis into a more comprehensible two-dimensional space. This reduction is achieved without significant loss of the relational structure of the data, as evidenced by the clusters in the plot. These clusters signify groups of products with semantically similar titles, suggesting a close thematic relationship among the items within each cluster. Moreover, the varied densities observed across the plot underscore the varying degrees of similarity between different product titles, with denser areas pointing to a higher concentration of similar items. Outlying points, starkly detached from the main clusters, likely represent unique or niche items that defy the common semantic threads seen in the larger groups. This graphical representation underscores the potential of UMAP in combination with TF-IDF to unveil the underlying patterns in product data, affirming its value in enhancing item categorization and recommendation systems in e-commerce settings.

In Fig 2. The scatter of points across the graph indicates clusters where product ID share common themes, suggesting a successful capture of semantic relationships by the TF-IDF embeddings. The dense central area highlights a significant overlap in thematic content, while the sparser fringes suggest unique or specialized products. Notably, the distinct outlier reflects an item with minimal relation to others, hinting at either a highly unique product or possible anomalies in data.

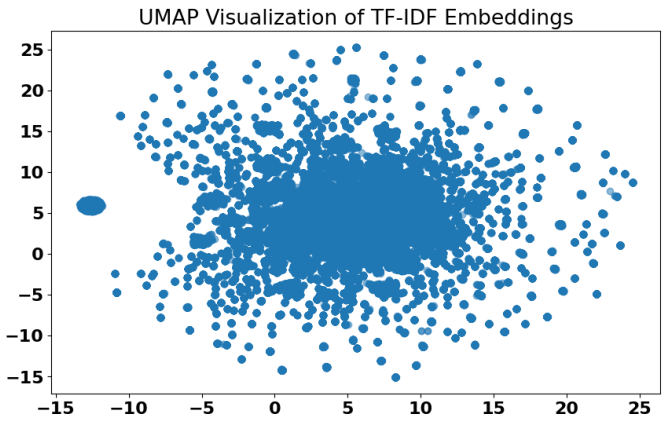

Fig. 2: UMAP Visualization of TF-IDF Embeddings

2) Elbow Method in K-Means Clustering: In k-means clustering, the Elbow Method is employed to determine the optimal number of clusters—an essential step to ensure the effectiveness of the clustering process. This method involves plotting the sum of squared distances from each point to its assigned cluster center, known as inertia, against the number of clusters. The resulting graph typically displays a distinct bend or 'elbow,' beyond which further increases in cluster count result in diminishing returns on cluster tightness [17].

For this study, the Elbow Method revealed that four clusters represent a point of inflection where increasing the number of clusters ceases to yield significant improvements in variance reduction. This finding is crucial as it suggests a natural grouping within the high-dimensional space of TF-IDF text embeddings, reducing complexity while retaining the intrinsic structure of the data. Utilizing this optimized cluster count, the products can be efficiently categorized, which is particularly advantageous for managing large inventories and improving customer search and recommendation systems. The utilization of the Elbow Method demonstrates a systematic approach to dimensionality reduction, which is vital in the preprocessing phase for machine learning tasks in e-commerce product categorization.

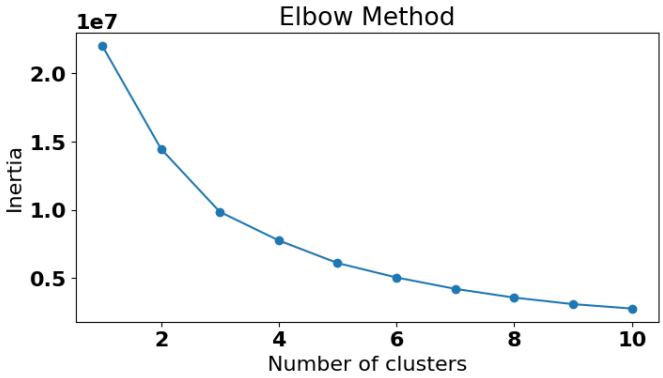

Fig. 3: Elbow Method

Fig 3. shows the within-cluster sum of squares (inertia) against the number of clusters (k). As k increases, the inertia decreases as the fit becomes more precise. However, the rate of decrease sharply diminishes after a certain point, forming an "elbow" in the graph. The location of this elbow typically indicates the most appropriate number of clusters for the dataset. In the provided graph, the elbow appears to be at k=4, suggesting that dividing the data into four clusters may yield a reasonable trade-off between cluster compactness and the number of clusters.

3) UMAP Visualization with K-Means Clustering: In the UMAP visualization enriched with K-Means clustering, the distinct color-coded clusters represent groupings of products based on similarity in their textual features, derived from TF-IDF embeddings. This technique, following the insights from the Elbow Method which suggested an optimal number of clusters, enhances our ability to discern natural groupings

	Naïve Bayes	Logistic Regression	Gradient Boosting
top_category_id	0.6413	0.7133	0.14
bottom_category_id	0.5033	0.1493	0.79
primary_color_id	0.2390	0.3378	0.065
secondary_color_id	0.1450	0.1672	0.131

TABLE I: F1-Scores for different models across attributes.

in the data. The visual representation indicates the presence of four predominant clusters, each potentially encapsulating a unique category of products with shared characteristics. Dense areas within clusters highlight commonalities in product titles, suggesting a high degree of similarity, while sparser regions may signify more distinct product niches. Moreover, outliers positioned away from the core clusters could denote unique items that do not closely align with mainstream product categories. The clarity provided by this visualization technique facilitates a deeper understanding of the product landscape, supporting improved categorization and retrieval in e-commerce settings.

IV. Evaluation

1) Evaluation Matrices: In the evaluation of machine learning models for predicting product attributes from textual data, F1 score serves as a critical metric due to its balanced consideration of both precision and recall. This is especially pertinent for e-commerce platforms where both false positives and false negatives can significantly impact user experience and business outcomes. Across the three algorithms—Naive Bayes, Logistic Regression, and Gradient Boosting via XGBoost—the focus was on maximizing the weighted F1 score for each class within the attributes of top category ID, bottom category ID, primary color ID, and secondary color ID.

The Multinomial Naive Bayes model demonstrated moderate effectiveness, with a weighted average F1 score of 0.6413, indicating a fair balance between precision and recall across various categories. Notably, the model showed higher precision in certain categories like ”clothing” and ”home and living,” which suggests that where the model predicts these categories, it is often correct. However, the lower recall in categories such as ”art and collectibles” and ”weddings” indicates a tendency to miss relevant items within these classes.

TABLE I and II accurately shows the comparison between F1-Scores of different models across different categories and Individual F1-Score of the particular model which gives us clear idea how these score can effect the dataset along with the model's accuracy.

Our findings revealed that while the Naive Bayes model offered a balanced F1 score of 0.6413, it was outperformed by Logistic Regression, which presented an improved F1 score of 0.7133 for the top category ID, suggesting a more robust generalization across product titles.

The lower scores for color prediction may reflect the more challenging nature of accurately associating color with the given text data.

Model	Naïve Bayes	Logistic Regression	Gradient Boosting
F1-Score	0.6413	0.8422	0.5124

TABLE II: Overall F1-Scores for different models.

The variance in F1 scores among the models and product attributes emphasizes the necessity of a tailored approach to model application, ensuring that the most suitable model is employed for each attribute. This targeted approach is particularly significant when applying these findings to unseen data, ensuring that the deployed model delivers the most accurate and reliable predictions to support e-commerce operations.

2) Jaccard Similarity: In assessing the performance of our classification models, we incorporated the Jaccard similarity score as a metric to evaluate the accuracy of predictions for multiple attributes simultaneously. The Jaccard score, particularly useful in multi-label classification scenarios, measures the similarity and diversity between the sets of predicted labels and the actual labels by calculating the size of the intersection divided by the size of the union of the label sets [11].

For our model, Jaccard scores were computed individually for each attribute—top category, bottom category, primary color, and secondary color. The scores reflect the degree to which the predicted labels match the true labels, with a focus on the presence and absence of each category in the predictions versus the actuals. Upon aggregating these scores across all attributes, the resulting Average Jaccard Similarity score for our model was 0.0917. This value indicates a modest level of agreement between the predicted and actual labels, suggesting specific areas where the model's performance could be enhanced. Although this score is not exceptionally high, it provides a crucial baseline from which improvements can be made, especially in refining the model to better capture the multi-dimensional traits of products in an e-commerce setting.

This evaluation step is vital as it not only quantifies the effectiveness of our predictive models in terms of similarity but also highlights the challenges in achieving high precision across multiple labels in a complex dataset. Future efforts might focus on optimizing feature selection and model parameters to improve this metric, thereby enhancing the overall accuracy and reliability of the model in real-world applications.

V. Conclusion

Our research navigated the complex domain of machine learning to enhance e-commerce product categorization, harnessing Naive Bayes, Logistic Regression, and Gradient Boosting. Logistic Regression emerged as the most effective, reflecting a superior balance of precision and recall in its F1 score. The clarity of UMAP visualizations affirmed the potential of this approach to discern subtle product relationships, while the elbow method critically informed cluster selection. This fusion of methodologies not only enriches the academic dialogue on classification challenges but also marks a leap forward in optimizing e-commerce search engines, promising a more intuitive online shopping experience.

REFERENCES

[1] CJC Burges. Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers, chapter geometric methods for feature selection and dimensional reduction: A guided tour, 2005.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[3] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[7] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[8] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[9] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[10] Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[11] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[12] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[15] Anand Rajaraman and Jeffrey D Ullman. *Mining of massive datasets*. Autoedicion, 2011.

[16] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[17] MA Syakur, B Khusnul Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, volume 336, page 012017. IOP Publishing, 2018.

[18] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[19] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.