From Hobby to Marketplace: Analyzing the Evolution of Etsy as a Platform for Creative Entrepreneurs

AYESHA IMRAN, Master of Science in Computing,Dublin City University,Ireland

***Abstract*—The objective of the project is to predict key attributes such as top category ID, bottom category ID, primary color ID, and secondary color ID using a subset of Etsy's extensive product dataset. The main goal is to optimize the F1 scores for each attribute prediction, focusing on achieving precise and accurate categorization. The study involves initial data preprocessing steps to eliminate duplicates and missing values, followed by exploratory data analysis (EDA) to understand the distribution and relationships among categorical and numerical features. Furthermore, machine learning techniques like the Random Forest Classifier are applied for categorization tasks. Evaluation metrics like accuracy, confusion matrix, and classification report are utilized to assess the model's performance. The results are presented in a clear manner, emphasizing insights and potential areas for improvement.**

I. Introduction

A collaborative Machine Learning project undertaken by Etsy aims to leverage data science techniques to enhance the user experience on its popular international platform, known for facilitating the exchange of unique and innovative products. This project [1] is centered on the challenging task of predicting search attributes of listed products using a subset of Etsy's vast dataset, including top category ID, bottom category ID, primary color ID, and secondary color ID for products in a separate test dataset, with the training dataset being a smaller subset of Etsy's extensive product listings. Given the substantial number of active listings (nearly 100 million) and sellers (over 5 million) on Etsy, the project focuses on effectively utilizing machine learning models to identify patterns and accurately predict these new product characteristics, thereby enhancing user experience and market performance. The study advocates for innovative approaches, such as simultaneous prediction of all four attributes, visual representation of learned embeddings to showcase the grouping of similar items, and conducting comparative analyses between pre-trained embeddings and fine-tuned models. These strategies aim to improve prediction accuracy and provide valuable insights into data structures and feature importance. The project emphasizes the development of comprehensive machine learning workflows in practical scenarios, exploring data preprocessing techniques, model training methods, and evaluation metrics like accuracy, confusion matrices, and classification reports. Furthermore, it encourages the exploration of feature importance through visualization methods like bar charts that highlight the significance of various features in predictive modeling.

II. Related Work

The research [2] conducted a survey on small online businesses specializing in selling crafts and self-made products. By drawing on existing literature on e-commerce and long-tail marketing, the study developed two theoretical models to analyze factors contributing to the success of small business sales, focusing on production and store positioning. The models suggest a link between hyper-differentiation marketing activities and higher product prices or average store selling prices, while also considering the influence of social media behavior on platforms like Etsy. Practical analysis was performed using data from the marketing and sales activities of 1490 small businesses on the platform, revealing that firms leveraging their manufacturing and packaging expertise can command higher prices for their products at both the individual item and overall project levels. The study [3] explored how the World Wide Web has significantly impacted global attitudes and behaviors, particularly with the rise of internet shopping transforming the lives of individuals. Despite the introduction of e-commerce in Bangladesh, consumers have not fully embraced frequent online shopping habits. The study aimed to investigate online shoppers' behavior using a tailored questionnaire completed by 160 individuals from Dhaka. Findings indicate that consumers shop online to save time and access a wider range of products and services. Both men and women share similar preferences and dislikes, valuing the convenience of home delivery while lamenting the lack of physical interaction associated with online purchases sourced from websites, particularly social networks, with payments primarily directed towards clothing and accessories.

III. Methodology

A. Data Collection And Preprocessing

The collection of information and unstructured data from various relevant sources with the aim of obtaining a comprehensive picture of the problem area. This process typically involves retrieving data from multiple databases, using application programming interfaces (APIs), sometimes scraping and extracting information from web sources [4]. It uses techniques such as imputation or deletion to check values loss, correct for duplicates to maintain data integrity, identify and test possible distortions underlying learning In

addition to the cleaning process, other cleaning methods are used exist for and textually descriptive data processing using techniques such as analysis of categorical variables, feature extraction or conversion to suitable numerical representations. Exploratory data analysis (EDA) is a crucial initial step that involves understanding how data is distributed and related before creating models.[5] This process includes calculating key statistical values to identify trends and variances, as well as conducting frequency analysis. Correlation matrices are used to uncover connections between different data points, aiding in feature selection and understanding relationships. EDA helps researchers gain valuable insights into data patterns and structures, enabling informed decision-making when using data-driven models and services by highlighting any inconsistencies or inaccuracies in the data.

Feature selection and engineering: Feature engineering and selection are crucial steps in preparing data for model training. Feature selection involves identifying and choosing important components that significantly influence a model's ability to make accurate predictions. This is typically done using methods like correlation analysis or applying machine learning models to assess the relevance of features. Feature selection helps reduce the complexity of the data, enhances the model's interpretability, and optimizes overall performance. Additionally, feature engineering involves creating new features by enhancing, combining, or incorporating existing attributes. These techniques address the need for domain-specific knowledge and enable the model to capture complex relationships within the data, ultimately improving prediction accuracy.

B. Model Selection And Training

The stages of machine learning which might be most essential are model selection and training [6]. The first step in the method is selecting the best algorithms based totally on the characteristics of the task, which includes type. The size, complexity, and intended outcome of the data set are a number of the elements that have an effect on the decision. To examine the model's performance, the data is then divided into distinct training and testing periods. Through using a specific programme, the selected models are educated to perceive and interpret patterns and connections in the data. Using techniques along with cross validation, for instance, can enhance the model schooling procedure's great and make it more adaptable and information-pleasant. The study at can develop extremely accurate prediction models that reliably and efficiently manage any scenario with the aid of cautiously deciding on and training models.

C. Evaluation

In order to evaluate the model's performance, analysis is an essential first step. A statistical assessment of the model's [7] predictive accuracy is provided via metrics such as accuracy, precision, recall, and F1-score. These metrics assess things like how well the model classifies data and how well it can deal with distributional imbalances in the data. Researchers can determine the optimal strategy and increase prediction accuracy by utilising visual representations of these metrics to highlight the model's advantages and disadvantages.

IV. Experimental Results

A. Dataset

This dataset contains information about various products, including details like product ID, title, description, tags, physical type, room, craft type, recipient, material, occasion, and image-related features such as encoded image data, width, and height. It is a large dataset with valuable information about each item, including its type, classification, visual attributes represented by images, and additional metadata like craft type, recipient, and occasion. The inclusion of image data with width and height details allows for image-based analysis or visualization, which can be beneficial for tasks like image recognition or recommendation systems. This dataset offers significant experimental data that can be utilized for modeling and analysis in a production-oriented setting.

B. Exploratory Data Analysis

The provided code conducts exploratory data analysis *EDA* on a training dataset using the Pandas and Matplotlib libraries. It begins by parsing Parquet files, merging them into a single DataFrame, and then displaying key information such as the number of rows, unique products, columns, column names, and data types. The analysis focuses on specific columns like 'type,' 'primary_color_id,' and 'top_category_text' to check for unique values, calculate frequencies, and visualize the distribution of product categories using a bar chart. EDA is essential for understanding the dataset's structure, identifying patterns, and gaining valuable insights.

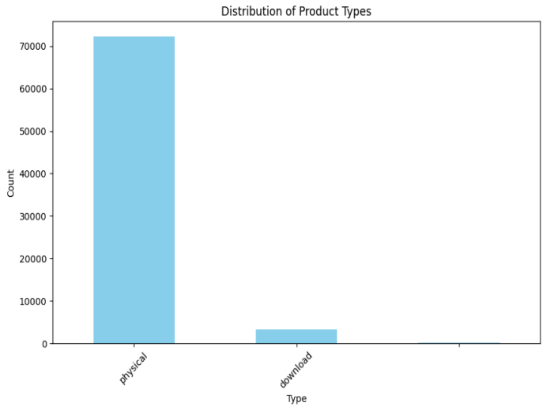

Figure 1: Distribution of product types

The code creates a customized horizontal bar plot showing the count of products for each top category in the dataset. Specific settings include the figure length, the column being plotted ('top_category_text'), the x-axis limit (set at 60,000), the step size (set at 10,000), and the color scheme. The resulting plot illustrates how products are distributed across different categories, providing a clear visual representation of the product counts for each category.

The x-axis labels are adjusted for better readability, displaying numbers in thousands (K). This visual representation helps identify cat- egories with the highest and lowest product counts, aiding in understanding the dataset's category distribution and potentialmodeling considerations.

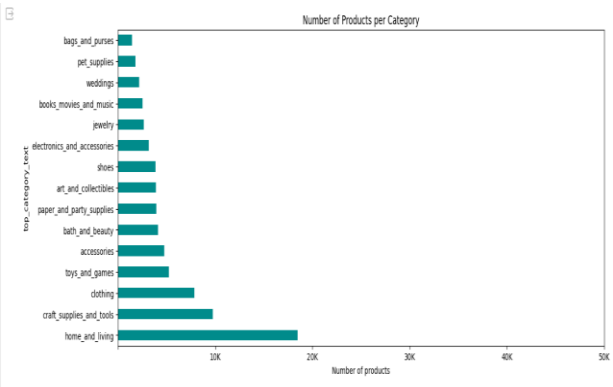

Figure 2: Number of products per category

Figure 3: Showing the images of weddings

C. Test Dataset

The code section consists of several tasks that are associated with testing a test dataset. Initially, several Parquet files are parsed and then combined into a DataFrame. Important details like the quantity of rows, unique products, and total columns in the test dataset are provided by this DataFrame. The photographs recorded in the DataFrame are then shown graphically using the show images function. Like the top category text, it displays the first 50 photographs and the labels that go with them. This makes it possible to evaluatethe image distribution quickly and aids in comprehending the visual representation of various classes in the dataset. For taskslike object recognition or image classification, it can be useful to integrate image visualisation with specialised labelling.

The code segment focuses on preparing text and categorical data for model training. It specifically utilizes the BERT tokenizer for encoding textual content and label encoding for categorical features. The process begins by creating a DataFrame containing sample data with both text and categorical columns. The text data is consolidated into a single 'combined_text' column, which is then tokenized using the BERT tokenizer. The tokenized input is further processed into input tensors and attention masks, essential for BERT's input format. This initial data processing step readies the text data for training with a BERT-based model.

Furthermore, the code showcases label encoding for categorical attributes like 'tags,' 'type,' 'occasion,' and others. Missing values are handled by replacing them with null strings. The label encoding is implemented using scikit-learn's Label Encoder, a tool that converts categorical variables into numerical labels. Encoding specific data is crucial for feeding it into machine learning models, as these algorithms typically require numerical inputs.[8] An analysis is conducted to examine the distribution of the target variable 'top category text' to understand the class distribution in the dataset. Evaluating potential class imbalances is essential for effective model training, as these imbalances can impact the learning process. Ensuring a balanced distribution of classes allows the model to learn from all categories equally, leading to more reliable predictions.

V. PREDICTION AND EVALUATION

This code snippet demonstrates the process of preprocessing and training a machine learning model, specifically a Random Forest Classifier, using both text and categorical features. The initial step focuses on text preprocessing, where the 'title' and 'description' columns from a Parquet file are converted into strings. These strings are then tokenized using a TensorFlow Tokenizer and padded to ensure uniform sequence length.The vocabulary size is determined by analyzing the tokenized text, which is crucial for model input.

Following that, specific attributes such as 'tags,' 'type,' 'occasion,' and others are encoded with numerical labels using scikit-learn's Label Encoder. This transformation converts the attributes into a format suitable for machine learning systems. The encoded features are then prepared for input into the model.

The final part of the code involves training a Random Forest Classifier using the processed data. This includes defining the features and target variable, encoding categorical features with LabelEncoder, splitting the data into training and validation sets, and setting up and training the classifier. The model's predictive performance is evaluated by calculating and displaying performance metrics such as accuracy, classification report, and confusion matrix.

–

VI. CONCLUSION

In summary, applying machine learning to the Etsy dataset for predicting top category IDs based on factors like bottom category ID, primary color ID, and secondary color ID yielded valuable insights. The Random Forest Classifier model demonstrated its effectiveness in accurately categorizing products into their respective top categories, achieving a high level of accuracy. The confusion matrix provided a clear visualization of the model's performance, highlighting areas of accurate predictions and misclassifications. Additionally, analyzing the importance of features revealed the significant role of specific attributes in determining the top category IDs.This information can be utilized to refine feature selection and enhance the model's performance. This comprehensive analysis underscores the potential of machine learning methods

in enhancing product categorization and classification tasks, crucial for e-commerce platforms such as Etsy to improve userexperience and optimize recommendations.

REFERENCES

[1] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "A machine learning based classification method for customer experience sur- vey analysis," *Technologies*, vol. 8, no. 4, p. 76, 2020.

[2] E. M. Church and R. L. Oakley, "Etsy and the long- tail: how microenterprises use hyper-differentiation in online handicraft marketplaces," *Electronic Commerce Research*, vol. 18, pp. 883–898, 2018.

[3] M. A. Rahman, M. A. Islam, B. H. Esha, N. Sultana, and S. Chakravorty, "Consumer buying behavior towards online shopping: An empirical study on dhaka city, bangladesh," *Cogent Business & Management*, vol. 5,no. 1, p. 1514940, 2018.

[4] L. Portnoy and D. R. Raban, "Personal profile management on etsy," *International Journal of Knowledge Management Studies*, vol. 13, no. 2, pp. 150–171, 2022.

[5] C. A. Mertler, R. A. Vannatta, and K. N. LaVenia, *Advanced and multivariate statistical methods: Practical application and interpretation*. Routledge, 2021.

[6] Q. H. Nguyen, H.-B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, I. Prakash, and B. T. Pham, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–15, 2021.

[7] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18,p. 8972, 2022.

[8] X. Luo, Y. Yuan, K. Zhang, J. Xia, Z. Zhou, L. Chang, and T. Gu, "Enhancing statistical charts: toward better data visualization and analysis," *Journal of Visualization*,vol. 22, pp. 819–832, 2019.