FLIPKART PRODUCT ANALYSIS AND PREDICTION

Parvathy Vysakh
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0818539@mylambton.ca

Patricia Adolf
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0816792@mylambton.ca

Pooja Selby
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0821687@mylambton.ca

Suchithra Chandrasekharan
Artificial Intelligence & Machine Learning
Lambton College
Ontario, Canada
c0816811@mylambton.ca

***Abstract*— The online marketing world is constantly changing and evolving, including the online retailing industry, which continues to grow. The e-commerce market is one of the segments that is hugely affected by the changes that happen on the web. It's becoming more competitive, since new businesses and platforms are always appearing on the Internet; and it's getting more difficult for online market players to thrive and prosper. Ecommerce analytics is the process of gathering data from all areas that have an impact on your online store and using this information to understand the trends and the shift in consumers' behavior to make data-driven decisions that will drive more online sales.**

Keywords—NLP, product category, preprocessing, category tree, product, ML model, oversampling, undersampling

I. INTRODUCTION

Nowadays, people are spending more time on e-commerce websites for purchasing items, and the online platform almost entirely covers the global business site. For this reason, it is also common to read and understand the reviews for the products before purchasing them. Apparently, customers are more likely to buy a product if it has been given with positive reviews; therefore, analysing these customer review data is essential to make them more dynamic. Flipkart is the largest Indian Ecommerce site. Started in 2007 by two ex Amazon employees, Sachin Bansal and Binny Bansal, Flipkart started off by selling books and eventually moving on to other goods, such as clothing, electronics and other consumables. This project aims at predicting the category of a product that is usually available on e-commerce sites like Flipkart.

For example, when a customer researches through the catalogues of a product range, the search patterns are recorded and a persona is created for the customer so that when he returns to the website, the searching time is drastically reduced by showing the most relevant product that the customer could be interested in buying.

An analysis of 20,000 products of Flipkart. In this we analyse the various factors on which overall rating and product rating of any product depends.

Product Categorization for e-commerce businesses like Flipkart, Amazon, etc., is becoming extremely crucial these days. This is because the probability of a product being sold is mainly dependent on product categorisation. If a product is not in its correct category, there are very high chances that a customer might not find it, and hence, it will not get sold. The problem of Product Categorization can be categorized into both, Multiclass Classification and Multilabel Classification because it is not always necessary that a product can belong to only one category. This report includes observations and conclusions while training several Machine Learning models to solve the product categorization problem using the Multiclass Classification approach [1].

II. THE SUBJECTS OF THE STUDY

Data has been collected using web scrapping. Web scrapping is the automated procedure of extracting the large amount of data from websites. The data available on the websites which is unstructured can be converted to structured data using Web Scrapping. When you run the code for web scraping, a request is sent to the URL that you have mentioned. As a response to the request, the server sends the data and allows you to read the HTML or XML page. The code then, parses the HTML or XML page, finds the data and extracts it. The dataset in our project is a crawled dataset, taken as subset of a bigger dataset (more than 5.8 million products) that was created by extracting data from Flipkart.com, a leading Indian eCommerce store. This dataset has following fields: product_url, product_name, productcategorytree, pid, retail_price, discounted_price, image, isFKAdvantage_product, description, product_rating, overall_rating, brand and product_specifications.

This analysis will help the manager in order to increase their sales of the product. There are various factors on which overall rating of the product depends. However there is misconception that on reducing the price overall rating of the product increases customer are more satisfied with product. Reducing the Price or Discounted Price can increase the sale for limited period of time but it will not going to give long lasting effect on overall rating [2]. But the probability of a product being sold is mainly dependent on product categorization. If a product is not in its correct category, there are very high chances that a customer might not find it, and hence, it will not get sold.

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 19999 non-null  object
 1   crawl_timestamp         19999 non-null  object
 2   product_url             19999 non-null  object
 3   product_name            19999 non-null  object
 4   product_category_tree   19999 non-null  object
 5   pid                     19999 non-null  object
 6   retail_price            19921 non-null  float64
 7   discounted_price        19921 non-null  float64
 8   image                   19996 non-null  object
 9   is_FK_Advantage_product 19999 non-null  bool
 10  description             19999 non-null  object
 11  product_rating          19999 non-null  object
 12  overall_rating          19999 non-null  object
 13  brand                   14135 non-null  object
 14  product_specifications  19985 non-null  object
dtypes: bool(1), float64(2), object(12)
```

Fig. 1. Dataset Description

The above dataset gives information about several products available on Flipkart and information about their category, price, product description, product specifications, etc. This dataset consists of 19999 rows and 15 columns. The labels in the dataset i.e. the Product Category in this case was given in the form of a tree. From a look at the dataset, it can be said that there are several non-ASCII characters in the dataset as well like bullets, characters from languages other than English, etc.

III. Data Extraction, Cleaning & Visualization

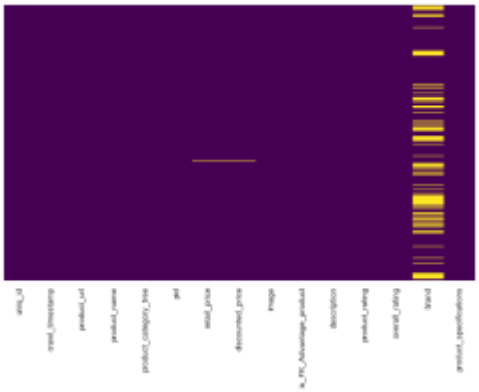
An in depth analysis of the dataset was done with the help of Bar Graphs, TSNE Visualizations, etc to get an idea about the most frequent unigrams in the Product Description, distribution of products and brands across the different Product Categories, analysis of the length of the description, etc. For Data Cleaning, Contraction Mapping, removal of custom stopwords, URLs, Tokenization and Lemmatization was done. The cleaning and prep section will consist of searching for missing data, looking for duplicate data, converting the time column to a pandas Timestamp object and sectioning off the timestamp column into a month column and a year column. Applying standard Data preparation techniques like checking null values, duplicate rows, removing unnecessary values, and text from rows in this research. Data consist of lots of missing values, nearly half of the year data was Missing. We can see that our data set have 12676 unique products and 20000 unique id's. Because of the clear imbalance in the dataset, balancing techniques like Oversampling and U ndersampling were performed on the dataset as well. These were then saved in the form of a CSV file [2].

IV. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [3].

A. *Dropping the unnecessary columns*

We will make sure that a unique id should be unique to remove duplicity. In the data exploration part, we will look at data types with visualization techniques and statistical techniques. The dataset and several of its hidden parameters were visualised (using libraries like seaborn, matplotlib, yello wbrick, etc). This then helped in data cleaning as several words from the Word Cloud were removed from the corpus as they did not contribute much in terms of Product Classification. As mentioned earlier, only description and product_category_tree are of utmost importance in predicting the category of a product, the rest of the columns which do not contribute any meaning to our problem statement are removed. However, some columns like product_name and brand are still kept for further visualisation of the data. On looking at the summary of the dataset, it was found out that the maximum number of NaN values occurred in the "brands" column, with some NaN values also existing in the "retail_price" and "discounted_price" columns as well. However, these rows were not dropped from the dataset. This is because the dataset already consists of a small number of entries (19,999 entries), and columns like "brands," "retail_price", and "discounted_price" do not contribute much in



providing us with those features that can help us predict the category of a product.

Fig. 2. Heatmap showing the distribution of all the Nan's throughout the data

B. *Lineplot of Products of a specific Brand*

The brand having maximum amount of products on Flipkart is Allure Auto with 469 many products From the lineplot below, we can see that most of the brands have less than 100 of their products on Flipkart while some brands have around 300 of their products on Flipkart.

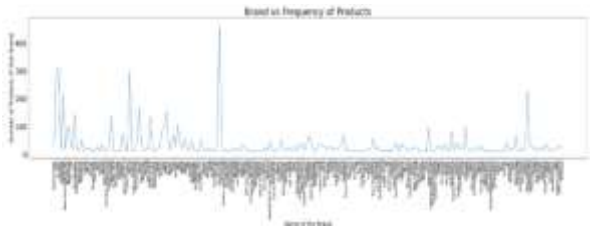


Fig. 3. Lineplot to visualise the frequency of all the products of a particular brand that is available on Flipkart.

Consequently, we will be performing a series of preprocessing steps before calculating the words that are

trending. We implemented a preprocessing step for the strings as a unified function. Within the preprocessing we consider: conversion to lowercase, eliminate numeric values and words with one character, tokenization, lemmatization, stopword removal and forced casting into string types [4].

C. *Bar Graph of the most common words in Product Description*

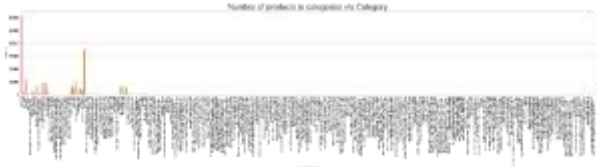
A bar graph of the 40 most frequent words occuring in the Product Description is made. This has helped us in adding some words to our stopwords list like shipping, delivery, flipkart, etc (which are then removed) as they do not have much meaning/contribution in the prediction of product category. Only the "description" column is considered for product

categorisation as much information was not available in the other columns, which could help us categorise a product. One approach that can also be taken for an even more detailed analysis of the dataset is by doing "named entity recognition." As some brands exclusively make products that belong to a particular category, they could be classified into their respective categories by combining the product description and brand. Thus, the other unnecessary columns were dropped. Some data points had their product description as a NaN value, and therefore, they could not be included in our dataset and were removed along with duplicate entries (entries having the same product description).

Fig. 4. Bargraph in decreasing sorted order for 40 most common words in the dataset

D. *Count of the different Product Categories in the Unfiltered Data*

To balance the dataset and increase the number of entries in a class, these redundant categories were clubbed into a single category. This somewhat helped in balancing the number of entries in a product category of the already small dataset. On plotting a graph of the number of products in each category vs. the Categories, it was observed that the dataset was largely



imbalanced with only very few columns dominating, and the rest is just noise. With the maximum number of products being around 5000 for the clothing category and other largely dominant categories having around hundreds of products, it was decided to drop those categories which have less than ten products. This was done to reduce the noise in our dataset, leading to increased accuracy of our model. After combining the redundant categories and removing the categories having less than ten products, the following was the graph obtained for the distribution of product across the categories:

Fig. 5. Bargraph showing noise in the dataset

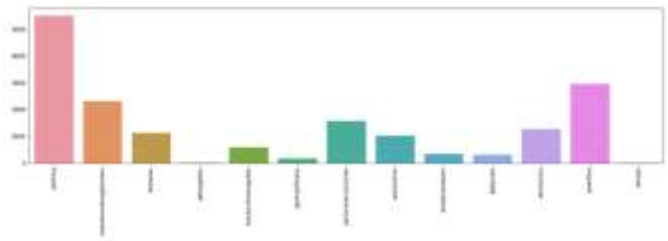


Fig. 6. Bargraph showing only those categories which have at least 10 products

E. *Combining the redundant categories*

From the above bargraph, we can see that there are some classes which can be grouped together like sunglasses and eyewear and home furnishing and home & kitchen. This will also help us in increasing the datapoints for each category in order to train a better model. With the dataset being already quite small, if any other subcategory was considered for classification, the number of entries in each category would have been quite small for some categories (example: sunglasses) while the other categories would have dominated (example: clothing and jewelry). This would have led to an even more imbalanced dataset which would have compromised the accuracy of our model on real-world data.

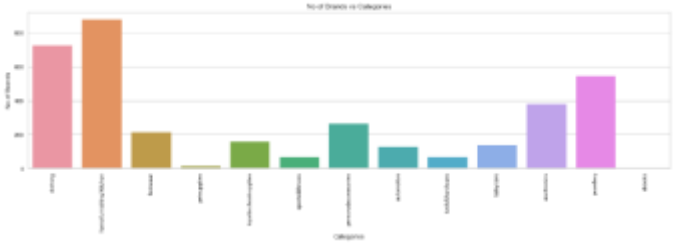


Fig. 7. Bargraph plotting the number of Brands per Category

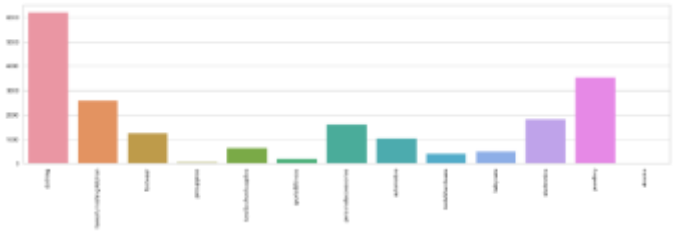


Fig. 8. Bargraph shown below shows the frequency of the product in each of the final classes (total classes = 13)

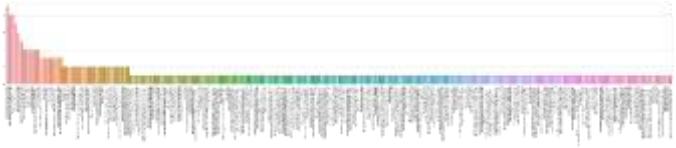


Fig. 9. Bargraph showing noise in the dataset

Fig. 10. Bargraph showing noise and primary categories in the dataset

V. Sentiment Analysis and Text Length Analysis

Sentiment analysis of the Product Description corresponding to each category. This is done by measuring the polarity of each product's description. Sentiment Analysis on the Product Description was also performed by calculating the polarity with the help of the TextBlob library in Python. From the Seaborn boxplot, it was observed that most of the descriptions across all the categories had positive/ neutral sentiments associated with them. This was also expected because the product description consisted more of factual data describing the product and not about its reviews.

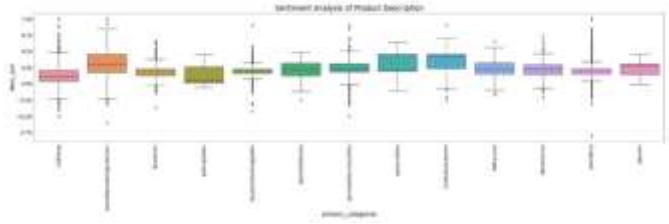


Fig. 11. Sentiment analysis of the Product Description corresponding to each category. This is done by measuring the polarity of each product's description.

For the discounted sales section, we are going to have to do some data manipulation beforehand. First, we are going to create a discounted percentage column by subtracting the discount price from the retail price and dividing that amount by the retail price [4].

Analysis of the length of the Product Description is done to help us get an idea about the minimum, maximum and average length of the same. This is done in order to decide whether we have to discard some datapoints having text length less than or greater to a threshold. There are discrepancies in the minimum length across all the categories. pet supplies and ebooks have a minimum length almost greater than 180 while clothing has the minimumlength.

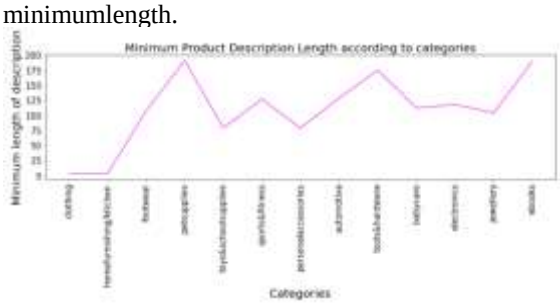
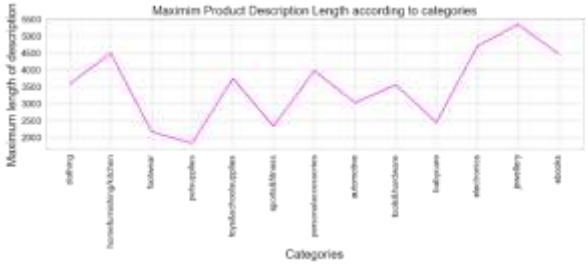


Fig. 12. Visualization of the Minimum description length across all the categories.

Almost all the product description lengths are greater than 2000 with jewelry category having the greatest length for product description (greater than 5000). The average length for all the categories lies more or less around 1000 words. I decided to not go with a certain minimum/maximum words threshold to prevent loss of information.

The minimum description length was 74, the maximum description length was 5309 and the average description length was 440. From the plotted graphs, we could see that there were discrepancies in the minimum length across the categories;



however, the average and maximum description lengths were very uniformly distributed. With the dataset already being quite small with fewer entries corresponding to each class, it was decided not to keep any minimum or maximum threshold and consider all the text lengths to prevent loss of useful information while training the Machine Learning models.

Fig. 13. Visualization of the Maximum description length across all the categories.

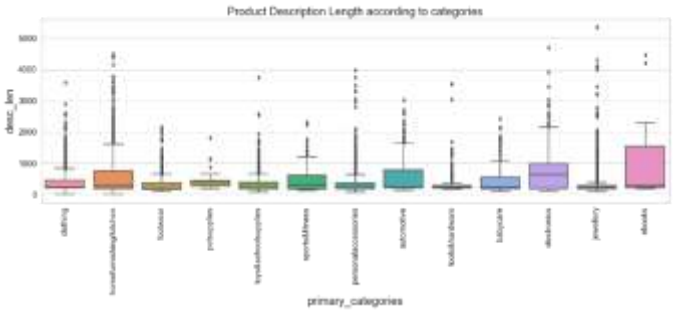


Fig. 14. Boxplot of the Average description length across all the categories

VI. Data Cleaning and Preprocessing

Data Cleaning and Preprocessing were done once and then cleaned, processed, and resampled datasets were saved in the form of a CSV file for later use while training and testing the

models. Character contraction is done to look at what percentage of the dataset is in English Characters. Punctuations, numbers, hyperlinks, etc. will all be removed during further cleaning of the dataset. The % of data consisting of only English Characters is 82.20710108917427.During this analysis of the characters, we can see that there are several emoticons, letters from different languages (Chinese), etc. We get all the words ending an apostrophe and single letter, all the words starting with a single letter and an apostrophe, and all the URLS and create a custom filter for these and . We remove all the stopwords, punctuations, and also add some words to our stopwords list like shipping, delivery, flipkart, etc. as they do not have much meaning/contribution in the prediction of product category in English. These are then removed from the corpus [5].

The following were the steps performed during data cleaning and preprocessing:

A. *Character Contraction:*

From one glance at the dataset, one could see that there are several emoticons, bullet points, numbers (denoting the efficiency or price of a product), etc in the dataset. Hence, contraction mapping was done to get an idea of what percentage of the dataset is comprised of only English language characters. Only 82.54% of the dataset comprised of English characters and on the further breakdown of the product description, it was found out that letters from other languages like Chinese, etc, emoticons, bullets, etc made up the rest of the dataset. Also, a custom contraction dictionary was created in order to map some of the words from the dataset to their useful meaning (those words were not simply discarded to prevent loss of information).

B. *Lowercasing, removal of custom stopwords, numbers, and punctuations:*

The entire product description was first converted into lowercase letters. Then a custom stopwords list was made which had stopwords from the English language that are available in the nltk package, punctuations available in the string package, and some extra words were included like "free shipping", "rs", "cash", "delivery", "product", etc. after drawing conclusions from the previously visualized Bar plots.The dataset also consisted of a lot of numbers (corresponding to the prices, product details, etc) and hence, these were also removed with the help of Regex as these numbers would not add much value in product categorization.

C. *Removal of hyperlinks, extra whitespaces, and non-ASCII characters:*

With the help of Regex, many of the hyperlinks (corresponding to the details of the product), extra whitespaces, and non-ASCII characters like emoticons, bullets, check-marks, etc were removed from the corpus.

D. *Tokenization and Lemmatization:*

Tokenization and Lemmatization were performed in a single step. WordNetLemmatizer available in the nltk package was used for Lemmatization. Initially, instead of lemmatization, stemming was used. But on further study, it was observed that it did not preserve the essence of the description and hence, did not give good results. Thus, the approach was then changed to lemmatization.

VII. VISUALISATION OF THE CLEANED DATASET

A. *Visualisation of the most common words:*

40 most common words are plotted in the form of a bargraph after removal of the unnecessary data. From the bargraph, we can clearly see that these are the words which actually can help us in identifying the particular category of a product [1].

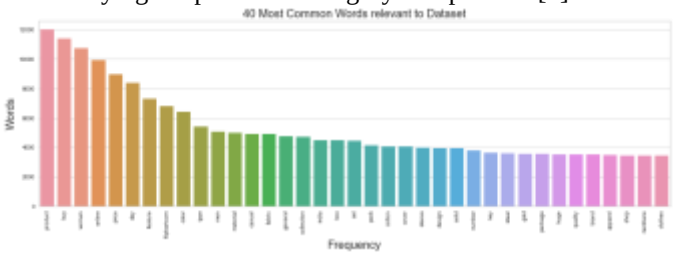


Fig. 15. 40 most common words are plotted in the form of a bargraph after removal of the unnecessary data

B. *Converting the Primary Category of all Noise datapoints into Noise label*

For all the rows in the dataframe where the Main Category is 0 (i.e. noise), then the primary_category label is updated to "noise". This is done to ease the process of plotting the TSNE graphs.

C. *Visualization of the Overlapping of classes using TSNE Plots*

TSNE Plots are plotted after creating the Product Description's vectors. This helps us in visualization by reducing the 13 dimensional plot into a 2 dimensional plot.

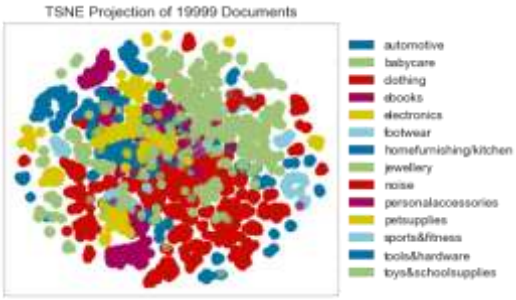


Fig. 16. TSNE Plots after creating the Product Description's vectors.

The preprocessed unbalanced dataset is then saved as a csv file for further training. The unbalanced dataset also contains noise in order to check the working of our model on a dataset having sufficient noise as well.

VIII. BALANCING THE DATASET (RESAMPLING)

We can clearly see that the products are distributed very non-uniformly throughout the different product categories. The dataset is very imbalanced with more than half of the dataset residing in clothing and jewelry categories. Two of the following dataset balancing techniques are adopted in order to

ensure that there is an almost equal frequency of products for each major product category (in order to train an unbiased model and increase the accuracy).Hence, it was essential to balance the dataset in order to achieve higher accuracy. In order to measure the performance of our model on several various kinds of datasets, the unbalanced dataset (consisting of 237 entries of noise data points as well) was also saved in the form of a CSV file to train and test the model.

A. Random Oversampling by Resampling:

Oversampling was done randomly to increase the data points in the minority classes i.e. all the product categories except clothing and jewelry. The final samples in each category (n_samples) were then decided to be n_samples = 3000. This value was chosen because it was close to the frequency of the majority classes as well (so that there is not much loss of useful data). This random oversampling was implemented by the resample function provided by the scikit-learn library. The data points that referred to "noise" in terms of product category were also removed as it was decided that the model accuracy could be improved by removing this noise. As this dataset was already cleaned before, the cleaned and balanced dataset was then saved as a CSV file for further use while training the ML model. Random oversampling of the minority classes (i.e. classes apart from clothing and jewelry) is implemented. Only the 13 major categories are considered and the 237 datapoints classified as noise are also removed from the dataset. This is done to help increase the accuracy of the model that will be trained further. Around 3000 samples for each category is considered in the final oversampled dataset.

B. Random Undersampling by Resampling:

Undersampling was done randomly to decrease the data points in the majority classes i.e. clothing and jewellery. The final samples in these categories (n_samples) were then decided to be n_samples = 1700. This value was chosen such that it was ensured that there is not much loss of useful information. However, there is a possibility that there might have been a loss of information for the "clothing" category as its data points were decreased from around 5300 to only 1700. The random undersampling was implemented by the resample function provided by the scikit-learn library. The data points that referred to "noise" in terms of product category were also removed as it was decided that the model accuracy could be improved by removing this noise. As this dataset was already cleaned before, the cleaned and balanced dataset was then saved as a CSV file for further use while training the ML models. Random undersampling of the majority classes (i.e. clothing and jewelry) is implemented. Only the 13 major categories are considered and the 237 datapoints classified as noise are also removed from the dataset. This is done to help increase the accuracy of the model that will be trained further. Around 1700 samples for clothing and jewelry categories are considered in the final undersampled dataset. This number is chosen by keeping the frequency of the other categories in mind while trying not to lose useful information during undersampling.

IX. Machine Learning Models for Product Categorization

Several machine learning models can be used to implement text classification methods to categorise a product into its correct category based on the product title/description. In this notebook, we have only used the Product Description as the main corpus. This is mainly because of the following two reasons:

- On analyzing the dataset, one can notice that most of the product descriptions somehow include the name of the brand in them. Hence, there wasn't a need to specifically concatenate the brand name along with the description.
- Furthermore, there was one limitation in trying to include the brand name along with the description. This is because there were a lot of NaN values present in the "brands" column and if we removed those data points, the already small data frame consisting of 20,000 entries would then directly shrink to around 12,700. This would have led to the loss of a lot of useful information that otherwise would have helped in increasing the accuracy of our model. Hence, only the product description column was considered for training the models.

The following Machine Learning algorithms were used to train a model and tested by evaluating several metrics like classification report, confusion matrix, accuracy score, etc.

- Linear Support Vector Machine
- Logistic Regression (Binary Classification Method)

Every single one of the above algorithms was tried on all 3 of the following datasets, we split the dataset into training and test parts, develop bag of words implementation, TF-IDF implementation, fit the training dataset to the model and develop evaluation metrics for the dataset [6]. In addition the following are also done:

- Imbalanced Dataset: This dataset consists of previously cleaned and preprocessed data points (with lowercasing, tokenization, lemmatization, etc). The data points mainly belonged to the 13 Primary Categories that were earlier mentioned. However, there was also some noise present in the dataset that was removed before training of the model. This was done because we thought that passing around only 237 rows that correspond to noise would not be much helpful in training a good model and might further lead to less effectiveness when it came to real-world data.
- Dataset Balanced using Oversampling Technique: The imbalanced dataset was balanced using the oversampling technique (implemented using resampling). Each of the classes had around 3000 samples in them. The data points have also been cleaned and preprocessed previously and they only belong to the 13 Primary Categories that were earlier mentioned and all the noise in the dataset was removed.
- Dataset Balanced using Undersampling Technique: The imbalanced dataset was balanced using the

undersampling technique (implemented using resampling). The majority classes (clothing and jewellery) were undersampled such that they only consist of 1700 samples. These data points have also been cleaned and preprocessed previously and they only belong to the 13 Primary Categories that were earlier mentioned (all the noise in the dataset was removed).

The following table summarizes the Validation Accuracies when the models were trained and tested on the above-mentioned datasets.

A. Encoding of the Product Classes:

In order to plot the ROC Curves and find the AUC score, there was a need to have a proper encoding for the 13 primary categories (in both directions). Hence, two of the following dictionaries are created to create a mapping. Helper dictionaries created which are later used to manipulate the testing output into suitable form before plotting the ROC Curves.

```
category_mapping = {  0=: "homefurnishing/kitchen",
                      1=: "clothing",
                      2=: "jewellery",
                      3=: "personalaccessories",
                      4=: "electronics",
                      5=: "Footwear",
                      6=: "automotive",
                      7=: "toys&schoolsupplies",
                      8=: "tools&hardware",
                      9=: "babycare",
                      10 : "sports&fitness",
                      11 : "petsupplies",
                      12 : "ebooks"}

reverse_category_mapping = {"homefurnishing/kitchen":0,
                            "clothing":1,
                            "jewellery":2,
                            "personalaccessories":3,
                            "electronics":4,
                            "Footwear":5,
                            "automotive":6,
                            "toys&schoolsupplies":7,
                            "tools&hardware":8,
                            "babycare":9,
                            "sports&fitness":10,
                            "petsupplies":11,
                            "ebooks":12}
```

Fig. 17. Helper dictionaries

B. Linear Support Vector Machine

Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

- Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
- Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.

SVM Pros:

- Can handle large feature space
- Can handle non-linear feature interactions
- Do not rely on entire data

SVM Cons:

- Not very efficient with large number of observations
- It can be tricky to find appropriate kernel sometimes

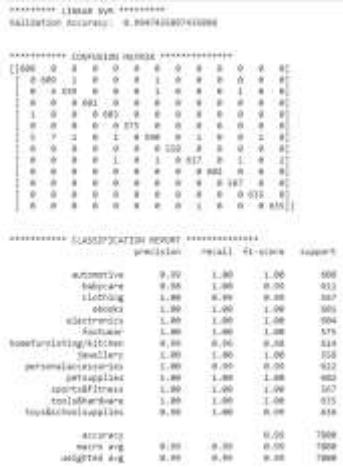


Fig. 18. Evaluation Metrics for Linear SVM

C. Logistic Regression (Binary Classification Method)

Logistic regression is a fundamental classification technique. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, and it's convenient for you to interpret the results. Although it's essentially a method for binary classification, it can also be applied to multiclass problems.

Logistic regression is a linear classifier, so you'll use a linear function $(\mathbf{x}) = b_0 + b_1x_1 + \cdots + b_rx_r$, also called the logit. The variables b_0, b_1, …, b_r are the estimators of the regression coefficients, which are also called the predicted weights or just coefficients.

The logistic regression function $(\mathbf{x})$ is the sigmoid function of $(\mathbf{x})$: $(\mathbf{x}) = 1 / (1 + \exp(-f(\mathbf{x}))$. As such, it's often close to either 0 or 1. The function $(\mathbf{x})$ is often interpreted as the predicted probability that the output for a given $\mathbf{x}$ is equal to 1. Therefore, $1 - (x)$ is the probability that the output is 0.

Logistic regression determines the best predicted weights b_0, b_1, …, b_r such that the function $p(\mathbf{x})$ is as close as possible to all actual responses y_i, $i = 1, \ldots, n$, where n is the number of observations. The process of calculating the best weights using available observations is called model training or fitting.

To get the best weights, you usually maximize the log-likelihood function (LLF) for all observations $i = 1, \ldots, n$. This method is called the maximum likelihood estimation and is represented by the equation $\text{LLF} = \Sigma_i(y_i \log(p(\mathbf{x}_i)) + (1 - y_i) \log(1 - p(\mathbf{x}_i)))$.

Logistic Regression Pros:

- Convenient probability scores for observations
- Efficient implementations available across tools
- Multi-collinearity is not really an issue and can be countered with L2 regularization to an extent
- Wide spread industry comfort for logistic regression solutions

Logistic Regression Cons:

- Doesn't perform well when feature space is too large
- Doesn't handle large number of categorical features/variables well
- Relies on transformations for non-linear features
- Relies on entire data

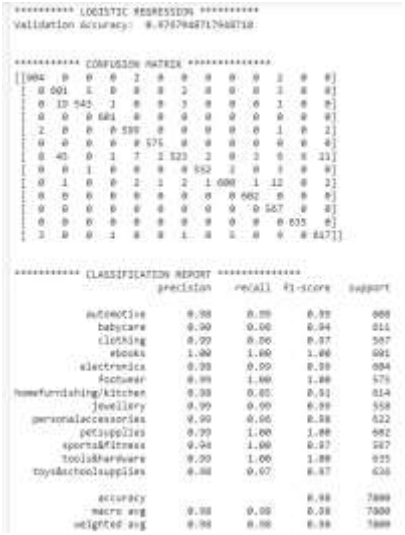


Fig. 19. Evaluation Metrics for Logistic Regression(Binary)

D. Plotting the ROC Curves for Multiclass Logistic Regression:

The ROC Curves are plotted and the corresponding AUC score for each of the categories has been plotted to get an idea about the accuracy of the model. Later, an aggregate AUC Score is also calculated which we can clearly see that Linear Support Vector Machine has consistently given the best accuracy across all three datasets. From the above table, we could see that the Validation accuracy of both the models are quite good. In order to further confirm this, ROC Curves were plotted and AUC Scores were calculated for the Multiclass Variant of Logistic Regression which had some interesting results. In the following table, we can see 4 out of the 13 ROC Curves that were plotted.an average for all the categories' One VS Rest ROC Curves.

The ROC Curves in the first row show that for some categories, our Logistic Regression Model has a lower True Positive Rate as compared to its False Positive Rate. Whereas in the second row, we can see that for categories like Sports & Fitness, Footwear, etc, our model works really well. However, in reality, it is crucial for our model to work well in classifying all categories and not only a few. The AUC Score for these models was also in the range of 60-70% which is not a quite good score.

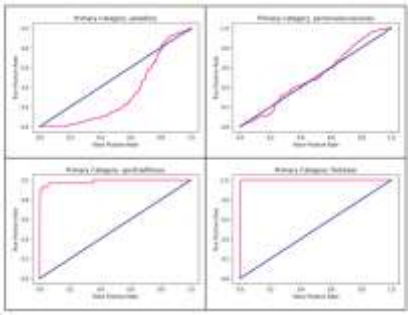


Fig. 20. Plotting the ROC Curves for 4 of the 13 main categories in our model

X. CONCLUSION

Analytics help eCommerce players in accurately predicting the demand for products. With this, the companies can effectively plan and manage their inventory to fulfill the customer demand on time and also appropriately coordinate with the merchant network for an effective supply chain management. In our project we performed analysis of pricing, product specification and brand and making predictions. Recommendation systems are high-priority feature in the success of e-commerce age.Data mining is the businesslike technique of expressing information from user data, with the correct utilization of algorithms we can amplify the performance and solve its problems. This paper concludes a step by step process of building a recommendation system with the help of various preprocessing, cleaning, machine learning modelling technique and evaluation measure matrices.

Linear Support Vector Machine has consistently given the best accuracy (99%) across all three datasets. ROC Curves were plotted and AUC Scores were calculated for the Multiclass Variant of Logistic Regression which had some interesting results. Logistic Regression Model has a lower True Positive Rate as compared to its False Positive Rate. Whereas in the second row, we can see that for categories like Sports & Fitness, Footwear, etc, our model works really well. However, in reality, it is crucial for our model to work well in classifying all categories and not only a few. The AUC Score for these models was also in the range of 60-70% which is a quite good score.

REFERENCES

[1] Tamvada, R. (2018, June 18). Recommendations-the machine-learnt way! Medium. Retrieved December 17, 2021, from https://tech.flipkart.com/e-commerce-recommendations-using-machine-learning-5002526e531a

[2] Recommendation system. Flipkart Tech Blog. (n.d.). Retrieved December 17, 2021, from https://tech.flipkart.com/tagged/recommendation-system

[3] www.jespublication. (n.d.). Retrieved December 17, 2021, from https://mlsoft.in/jespublication.com/upload/2020-20200313.pdf

[4] Scribd. (n.d.). HTTPS:WWW - ijarcce.com:upload:2016:march-16:IJARCCE 171. Scribd. Retrieved December 17, 2021, from https://www.scribd.com/document/476232517/https-www-ijarcce-com-upload-2016-march-16-IJARCCE-20171

[5] International Journal of Innovative Technology and Exploring Engineering (IJITEE). (2021, December 1). Retrieved December 17, 2021, from https://www.ijitee.org/

[6] Bhatnagar, P. (2021, August 4). Flipkart ecommerce product categories prediction. Medium. Retrieved December 17, 2021, from https://medium.com/@pmn.bhatnagar/flipkart-ecommerce-product-categories-prediction-684fe751593