Data_Mining_Final

by Ming Xiang Lee

Submission date: 25-Jan-2021 02:16PM (UTC+0800)
Submission ID: 1493822885
File name: Case_studies_Data_Mining_v2.pdf (1.39M)
Word count: 7545
Character count: 41187

Master of Data Science

Semester 1, 2020/2021

WQD7005 – DATA MINING

CASE STUDIES

Group Members:	**Matric Numbers:**
LU XIAN DING	17096993
MING XIANG LEE	s2021030
NICHOLAS LEE KAN	17201982

Topic: Machine Learning Model with Optimized Parameters for Ecommerce Product Classification

Abstract

In a new era of digitalization, a certain organization that still implements a traditional approach will not help the business to pave a digital way for the company to strategize efficient marketing strategies to increase revenue and target their customer's expectations of the product class. Therefore, to stay competitive in the market industry, a strong and consistent model is required.

To reduce the cost of manpower and many manual tasks, several machine learning algorithms are being applied to perform E-commerce product classification. For instance, the KNN model performing quite well in classifying household and fresh food products in an E-commerce store (Mathivanan. N., Ghani. N., Janor. R. 2019). Among all the algorithms that are used to classify the products, there is a parameter exist in the algorithm. We are going to figure out the optimized parameter to improve the accuracy of product classification. In the KNN model, we will investigate the optimal number of k-value that able to achieve the highest accuracy. In comparison to Neural language models, had much better performance in producing an embeddings word from a large size of quantities of product data. Convolutional Neural Network or CNN will achieve better results when comes to the detection of multiple similar objects even in different classes in an image. Out of the list of models to be compared, the L2AC classifier will accept a new class by adding a class. The experiment that would be discussed in this case study will tell a stronger performance in terms of baselines with the L2AC framework at the Literature Review's part. Given an algorithm that has been applied for the classification of the product, this case study will focus more on parameter optimization together with the algorithm for classification. Assuming the data

is structured without the need to contain any product's images for the process of analysis. In the future, the studies can be further up with image analytics by implementing the advanced convolutional model.

Keywords: Classification Algorithms, Support Vector Machine, Naïve Bayes, Random Forest, Decision Trees, Efficient Marketing, KNN.

1. Introduction

This main aspect of this case study is to understand how to improve further the current existing product classification by grouping it into multiple types of items and can target the correct group of customers for each item. Besides, there would additional aspect that we will consider which is the recommended technique that can be improved or suggested to help the organization to improve the sales and revenue, the purpose is to help in eliminating the old traditional approach being used which does not relevant and does not contribute much in improving the customer's experience and expectations of the product class.

1.1 Problem Statement

During this COVID-19 pandemic, there are tons of business planning to go through digitalization. The first step will be starting to sell product through E-commerce platforms. The E-commerce platform is full of different types of products. All the products need to be classified accordingly on the E-commerce website. This classification process requires a lot of manpower if doing it manually. Therefore, the main problems that exist in the e-commerce domain are time-consuming classifying the product to a correct category and high cost in hiring manpower to perform the classification manually. Inappropriate classification could potentially result in a recommendation engine nightmare wherein a customer has suggested a different product altogether (Nair. V.,

Malhotra. R., Mohapatra. S., Maknoor. N. 2018). Performing the product classification manually also exist the risk of human error.

1.2 Impact on the Problem Statement

It is always a challenge and time-consuming in the E-commerce sector to categorize large amounts of products into distinct product categories. It is time-consuming and cost consuming without automated models to classify the product based on the product attribute and consumer preference. Manual product classification gives the seller an impact, especially when uploading new products. Manual product labeling and tagging usually are labor-intensive, and maybe inaccurate and sub-optimal. The product mislabelling is increasing the risk for the seller to miss the sales opportunities as the buyer is not able to locate the desired products (Wirojwatanakul, P. & Wangperawong, A., 2019). Besides, as we stated previously, a recommendation system that is used to improve business sales may be affected if the classification is inappropriate. The malfunctioning of the recommender system in E-commerce gives the customers a poor shopping experience as the website keeps suggesting a wrong product to them.

1.3 Aim of The Study

The study aims to address this issue would be:

- To study existing techniques used for product classification especially for the e-commerce industry.
- To validate and compare the ability of different types of techniques and algorithms that have the highest accuracy in this case study.
- To validate the proposed techniques in product classification with suitable evaluation tools.

In addressing the main problem statement, machine learning algorithms stated in the aim of the study can effectively solve the problem. Various types of machine learning algorithms able to automate the classification job with high accuracy in a very short period.

1.4 Significance of The Study

It is a crucial step for marketers to implement an effective strategy on product categorization based on customer expectations and behaviors. This is essentially beneficial to e-commerce platforms such as Alibaba, Amazon, Lazada, and Shopee, as there will be thousands and millions of products from thousands of merchants (Li, M.Y., Kok, S., Tan, L., 2018). For a large scale of the product, the automated algorithm is important for an efficient manner of product categorization. With an efficient product classification, it eases user to locate the interesting item in a product catalog and make the purchase.

There would be a few positive impacts such as:

- Stakeholders will have a better advantage in improving their sales and revenue.
- Stakeholders will replace their existing business model with a better and improvised business analytics model.
- Studies and analysis between the relationship of the product's sales with marketing strategies can be conducted immediately to reduce the risk of churning out of the customers.

2. Literature Review

Below are 15 articles that have been compared in this case study. In the domain of E-commerce,

several directions can be investigated to further improve business revenue. The migration of business into the E-commerce platform created tons of digital data that can be useful to the business. By utilizing the sales data, we can predict customer behavior, segment the customer, etc. Apart from the study on the customer perspective, some studies focus on the internal part of the business which is cost optimization. There is research mention the importance of performing accurate e-commerce product classification without spending a large amount of cost by using the machine learning model. Accurate and quick product classification may help in building a good recommendation system. As there is already a lot of research study on the algorithm for product classification, in our case study, we are going to investigate further the optimized parameter for the machine learning algorithm in product classification.

Table 1 Literature Review Summary Table

Citation	Research Title	Problem Statement	Suggested Solution(s)	Significance (e.g. evaluated as performance)
[2]	An, J., Kwak, K. Jung, S.G., Salminen, J., Jansen.B., 2018.	1.Time and cost consuming to obtain reliable data to understand customer behavior via surveys. 2. Individual website data was focused in prior work for customer segmentation, however the data aggregation due to individual privacy preservation, causes in inferring customer attributes.	1.Behavioral and demographical customer segmentation using online social platforms. 2. Customer segments was isolated based on both behavior and demographics by using aggregated social media data for, then these two customer segments groupings were linked for a complete representation of the customer base.	1.NMF is more robust and effective comparing clustering methods and other matrix decomposition algorithms in looking for meaningful customer behavioral segments with a representative.
[14]	Singh, H. Neware, S., 2020.	1.Research objective: To divide the target market into subsets that share similar characteristics, needs, and priorities.	1.Segmentation, targeting, and positioning (STP) to identify and choose groups of potential customers 2.Predictive neural network and statistical analysis based on product reviews, products, buying patterns, viewing patterns, and time-based segments, and clustering techniques.	1.Feature extraction was performed with unigrams, bigrams, and trigrams calculation. 2.Parameters analysis to look for the most significant variable help to enhance the model prediction. 3.Predictive neural network approach can achieve accuracy which leads to better customer segmentation.

[17]	Tatjana, T., Verdenhofs, A., 2019.	1.Create customer segments based on predictive modeling by using big data available in an organization.	1.Combining 3 principles which are marketing, statistics, and Information Technology (IT) to generate a beneficial customer segmentation.	1.Merging the 3 principles created a new way of performing customer segmentation. 2.Limitation: This research only assumes discount offer is the main factor to impact customer decision, without considering other factors such as macroeconomic factors, different marketing messages, other sales promotions. 3.Further study: Can another type of business process benefit from segmentation performed with predictive modeling?
[19]	Xia, Y. , Levine,A. , Fabbrizio, P.D.G.D., Shinzato, K., Datta,A. , 2017.	1.Large scale categorization of millions of Japanese items into thirty-five product categories. 2.Long training time using Gradient Boosted Tree classifier.	1.Attention CNN (ACNN) model	1.ACCN model can reduce training time from 3 weeks to 3 days with 96% accuracy. 2.Future work: Experiment ensemble methods to exploit prediction errors from different models, to improve overall classification performance.

[16]	Umaashankar,V., Shanmugam,G., Prakash,A., 2019.	1.Organize product catalog and create product taxonomy to ease the user to locate the interesting item.	1.Establish a benchmark by comparing imaging classification and Attention-based Sequence model for predicting the category path.	1.Two benchmark models using classification and Attention-based Sequence approach to predict product taxonomy. 2.Future work: Avoid the generation of invalid category paths in the Seq-to-Seq model.
[3]	Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Shardoor, N., 2020.	1.The traditional approach of sales and marketing strategies incapable of helping the company to cope up with the pace of a very competitive market, as they did not have the information needed on customers' purchasing trends and patterns. 2.On the other hand, the company needs to manage its finances efficiently to maintain a positive cash flow.	1.The author suggested applying a few types of machine learning techniques such as clustering models to predict and measure the sales prediction. 2.Besides the author also suggested the usage of ANN techniques to understand more on the sales strategy. 3.The third one would be the Gradient Boost algorithm to exhibit maximum accuracy in picturizing the future sales transaction.	1.Few results evaluation technique has been carried out such as Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracies that determine the precision of results that are tabulated for each of the algorithms. 2.The results showed that the Random Forest Algorithm's accuracy is the highest at 93.53%.

[12]	Ristoski, P., Petrovski, P., Mika, P., & Paulheim, H., 2018.	1.The completeness of the product specifications and the taxonomies used for organizing the products differ across different e-shops. 2. Assign a set of classifications to deal with a set of labels from the product to the product.	1. Product data that are classified structured and labeled supervision for training feature capable of extracting attribute-value pairs from the typical textual of product's descriptions. 2. The model of Neural Language will produce a type of embeddings word from large quantities of data with the Microdata's technologies that significantly increase the performance for the feature of extraction model. 3. The usage of a Deep Convolutional Neural Network to produce an embeddings image from the product.	1. Conditional Random Fields with Continuous Features. 2. Application of Dictionary-Based Approach that builds a dictionary that contains the product attributes and also values that present in the product descriptions. 3. Text Embeddings with CRF that capable of handling type of dynamic text patterns in the descriptions of the product by enhancing the training of the CRF model.

[15]	Thobani, S., 2018.	1.There are a few problems that arise such as the customer wants to explore more options while minimizing their search efforts. 2.Besides, too many steps and hassle will reduce the customer's experience during the purchasing process. 3.Apart from that, the customers would need to receive continuous value after the purchasing, the seller would need to provide after-sales service to increase the customer's loyalty.	1.The author recommended the usage of a recommendation system on an e-Commerce website. 2.There are multiple approaches identified to the recommendation systems at such as Collaborative filtering, Content-based filtering, and Hybrid Recommender system. 3.Each recommendation system has functionalities to predict sales and analyze the rating given by the customer.	1. The significance of the studies focuses on a framework to accurately predict the customer needs and help companies build loyal relationships by targeting the right prospects at the right time on the rightmost relevant touchpoint to assist them in their purchase journeys. 2. ML model is used to optimize the weights of attributes while combining them by making use of the customer feedback about product relevance in the form of clicks, add to basket, purchase.
[13]	Safara, F., 2020.	1.The COVID-19 pandemic had affected all people daily life in various aspects. For instance, the behavior of consumer when shopping. 2.Research Objectives: I.To anticipate the consumer's behavior during a pandemic using machine learning methods. II.To figure out factors that contribute significantly in the online sales during COVID-19 pandemic.	1.The author uses statistical and machine learning approach to predict the consumer behavior. 2.The relationship of various factors are being studied. 3.Various kind of machine learning model is built in predicting consumer online shopping behavior.	1.Five classification models are examined and compared. The five algorithms are decision tree, support vector machine, sequential minimal optimization , naïve bayes, and ANN. 2. The predictions show better accuracy when bagging and boosting ae applied. It also aimed to reduce the bias and variance of a single estimate

[1]	Orogan, A., Onyekwelu, B., 2019.	1.Despite the complexity of building a customer behavior model, most customer behavior predictive models are quite simple. Many of the models are built without considering many important factors. Therefore, the predictions made by the model not really reliable. 2.Research Objective: To develop an association rule mining model to predict customer behavior using a typical online retail store for data collection and extract important trends from customer behavior data.	1.Apriori algorithm is used to figure out frequent itemset and generate association rules and patterns by using an online retail store dataset.	1.The research eventually figure out the frequent item sets, the consumer behaviors as well as strong association rules for an online store. 2.In this research the optimum rule generated with 0.1 and 0.2 as the support and confidence thresholds.
[20]	Xu, H., Liu, B., Shu, L., & Yu, P., 2019.	1. There are two challenges to solve the problem which is: - The solution for the model classifies the examples of seen classes into a respective class. Besides on when to detect or reject the examples of unseen classes. - How to include the new or unseen classes when data is enough without the need of retraining the model.	There is a meta-learning framework known as L2AC for the learning. This framework will be applied for product classification. Compare to traditional, the meta classifier can accept new classes by simply adding a new class.	1. Meta-classifier functioned as a core component for the L2AC. It is essential as a binary class as it takes the top of the k-nearest examples and determines whether an ex is grouped into the same class or otherwise.
[6]	Donati, L., Iotti, E., Mordonini, G., & Prati, A., 2019.	1. The problem arises when the traditional method is still the same which is generalization related to other domains such as fashion which is considered as hardly obtainable. 2. Too many feature extraction tasks such as logo recognition and relocate	1. Convolutional Neural Networks or CNN. 2. CNN features consisted of a pure deep learning algorithm used for classification, RoI which stands for Region of Interest. As a further	1. Extract the presence, location, and properties using the Stripes Recognition of the peculiar Adidas three-stripes 2. Segmentation of colors is an important extraction of the feature. It will be used to determine the contrast,

		them are the sub-problems of the detection task for the object. 3. The detection of the logo is the most vital task for data experts because the logo is the most important mark and symbol for their branded clothing. A strong and focused recognition will ensure the logo's presence and especially type and size is the key problem besides developing a marketing plan and the logic of the business.	enhancement for CNN's approach, the Regional Proposal Network or known as RPN added to the Fast R-CNN methodology.	differentiation between the stripes and background. Besides, the composition and patterns will be needed to be considered too. 3. The Deep Learning-Based Feature will be used to extract the prints, sleeves, or neck's shape. In this case, the availability of the labeled images would be sufficient for this technique.
[4]	Hang, W., & Banks, T. , 2019.	1.Pack classification seems becoming more important to the market research nowadays. This is known as the fundamental task in market research. Precisely classified all the brand, prize, type etc. can help the company to watch the market and make a good decision. 2.As As the data are generated at rapid speed, pack labeling accurately has been a challenge.	1.The authors suggested to use support vector machines (SVMs), neural networks, xgboost, and random forests in building an automated pack classification model.	1.This pare shows that the application of machine learning algorithms can give significant effect in market research audits. 2.Small dataset gives better performance when using simpler algorithms. For instances, decision tree, linear model, or SVM with the linear kernel. 3.Huge dataset gives better performance with more complex algorithms. For examples, neural network, SVM with a radial basis function kernel, or a boosting method. 4.There is a tradeoff between accuracy and training time and training resources by applying ensemble method.

[9]	Nair. V., Malhotra. R., Mohapatra. S., Maknoor. N., 2018.	1.Most retailers prefer to use the exciting environment of electronic commerce also known as e-commerce platforms in promoting their products and services. However, most of the products are assigned to the desired categories manually by the retailers. 2.Hence, this paper aims to compare the performance of supervised learning models in classifying various categories of e-commerce products.	1.They process the raw by using the text mining technique. The research had utilized the bag-of-word and correlation feature selection technique to perform data extraction and selection respectively. Then, five algorithms from the supervised learning model are being used for an experiment which are Naïve Bayes, KNN, Decision tree, SVM, and Random Forest.	1.KNN model gives the best accuracy compared to NB, DT, SVM, and Random Forest classification models. 2.For future work, the optimal number of neighbors (K) values of the KNN model can also be investigated in enhancing the performance of the model.
[8]	Mathivanan. N., Ghani. N., Janor. R., 2019.	1.The retailers need to focus on classifying the products correctly because this will bring a lot of negative impacts if the classification is done incorrectly. Either false positive or false negative in classification of products are going to affect the reputation of the retailers. Inappropriate classification of products will also destroy the recommendation system in E-commerce.	1.TFIDF is being used as a measure in the features selection stage. The classification algorithms are Random Forest, SVM, and Naïve Bayes.	1.In this research, the authors found out that products available online fairly rich in content, some with moderate content and some even lack of content. This causes the algorithm to give different results for different kinds of products. The authors mentioned that the successful part in this research is reduced human error and decrease the inappropriate classifications rate.
[5]	Chen, D., Sain, S. L., & Guo, K. , 2012.	1. In this research the authors aimed to study the customer behaviors in browsing the products' web pages, the customer behaviors in making a purchase, the response of customers towards the promotion made and classified the customers into different tiers.	1. RFM MODEL-BASED CLUSTERING ANALYSIS is used in this research. 2. The clustering analysis is being strengthen by applying decision tree into it. Decision tree is used to further	1. The authors mentioned that data preparation and the process of evaluating the machine learning model used the generous part of the time. These two processes are very essential as well and this is the reason most of the time will be spent in performing these two tasks.

			classified the customers in each cluster into different tiers internally.	2. They successfully classified the customers into segment within cluster. The authors suggested the customer buying patterns can be further study to better tackle the customers in generating greater revenue in business perspective.

3. Methodology

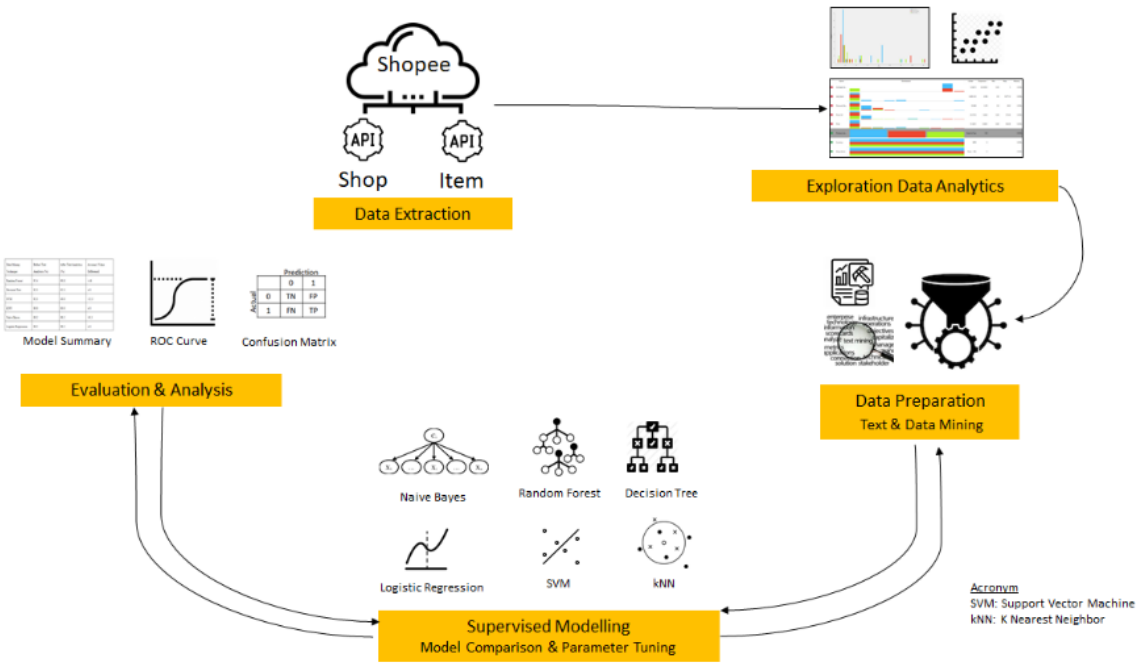

Figure 1 Data Mining Workflow

Figure above are the data mining workflow for the study, we are following the standard of Cross-industry standard process for data mining (CRISP-DM). The workflow starts with scraping data from Shopee Tesco web page, followed by Exploratory Data Analysis (EDA), text processing, and data cleaning. Once the data was processed and ready for modelling, few supervised models were tested for training purpose. Once the modelling was done, we evaluated the model performance using the evaluation matrix, confusion matrix, and the ROC analysis. The best model was then selected, and optimized by tuning the model parameters.

4. Data Modelling and Visualization

In this case study we will implementing our machine learning technique via Orange. The second case study is regarding Tesco FBS at which we will classifying items that are sold online by Tesco FBS in Shopee.

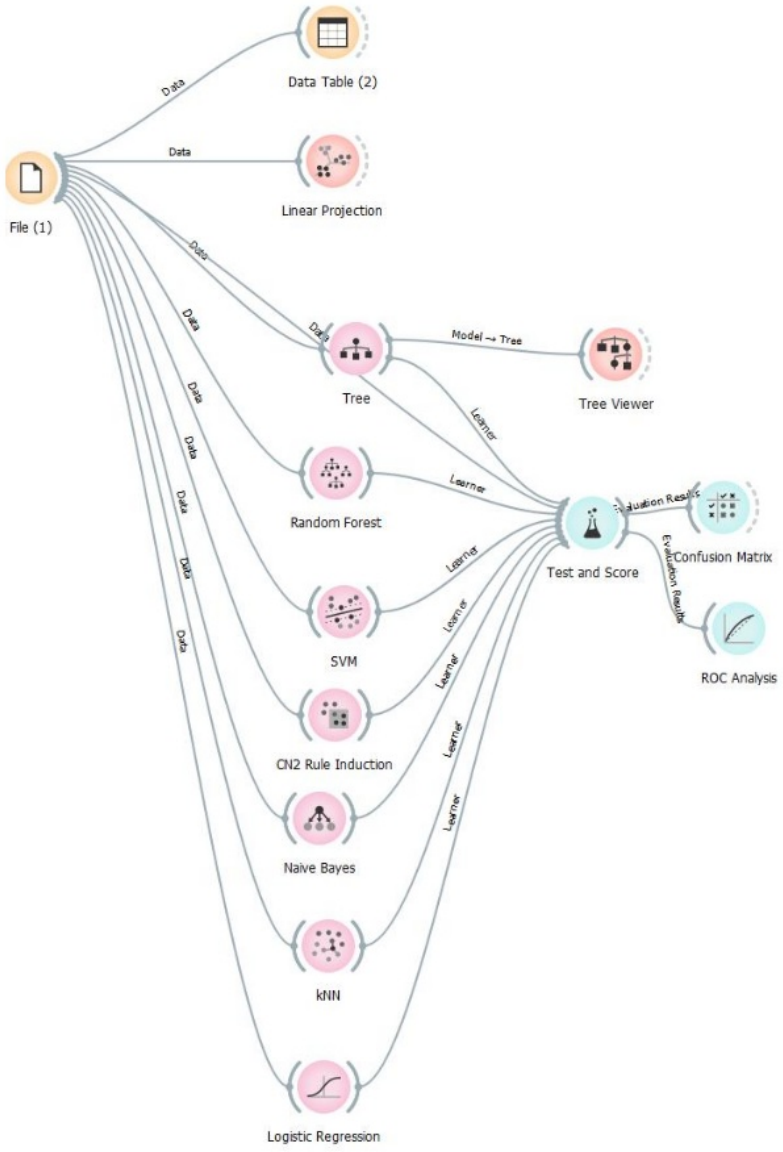

Figure 2 Modelling Workflow without Processing Text in Orange

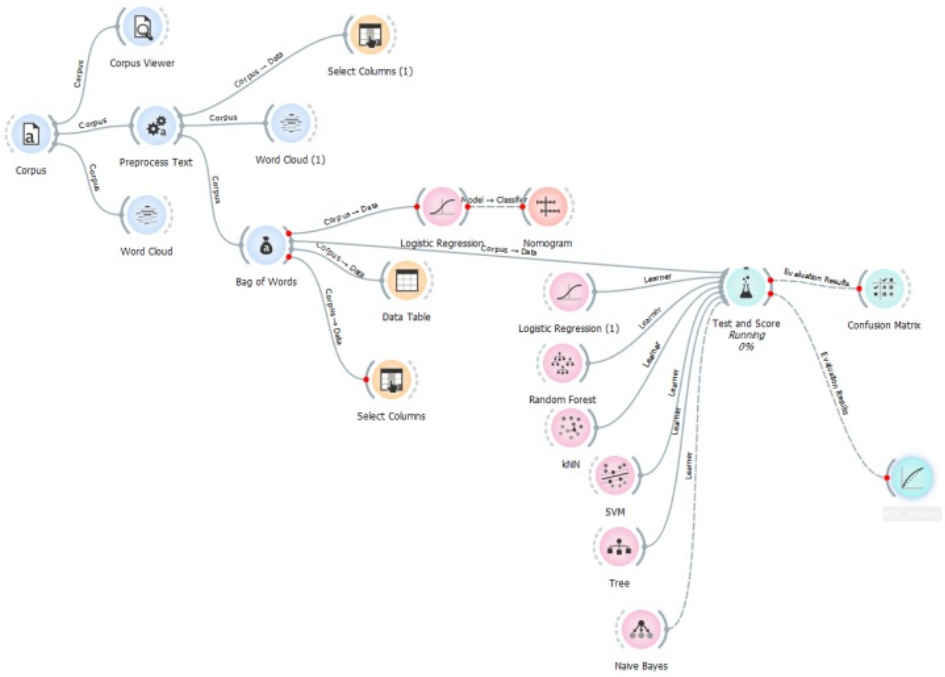

Figure 3 Modelling Workflow with Text Processing in Orange

This case study shall implement few data mining technique:

- KNN
- Naïve Bayesian
- Random Forest
- SVM
- Logistic Regression

Random Forest is an ensemble of a large number of the individual decision tree. Each tree would be a class prediction and the class with the most votes are the model's prediction. In this study, attributes Shop Name, Currency, Price after Discount, Price before Discount, Discount Rate, Unit Sold, and Average Rating are the features used in modeling prediction. Random Forest is more favorable than Decision Tree, as the model can train based on different subsets using different features for decision making. Naïve Bayes is a classification technique that is based on the Bayes' Theorem with an assumption of independence among predictors. Naïve Bayes considered few attributes such as Average Rating, Product Category, Shop Name, Product Name, Description, and Brand. Naïve Bayes is simple and easy to implement as it doesn't require so much training data, therefore it can be contributed a very high accuracy in comparison with Random Forest. KNN algorithm the k-Nearest Neighbors algorithm is one of the most commonly used machines learning algorithms for classification application. It classifies the new data according to the distance function, there are few types of distance functions being used in KNN algorithm, for instance, Euclidean distance, Manhattan distance, etc. A voting method is used to perform the classification. Each data point is classified to the nearest cluster. As you increase the number of nearest neighbors, the value of k, accuracy might increase. For example, if a new product will be classified according to the similarities of the product description, product name, price, etc. Another machine learning algorithm that often used for classification and regression issue is support vector machine (SVM). SVM is more commonly used to solve classification issues. In the SVM algorithm, the data points are plotted in n-dimensional space (n is number of features we have) with the value of each feature being the value of a particular coordinate. By finding the hyper-plane, classification can be performed. Ray, S. (2017). Introducing another machine learning algorithm which is logistic regression that also favorable to be used in solving classification problem. It used the concept of

probability in predicting each data point. Sigmoid function is a complex cost function that apply in the logistic regression model.

Let $Dataset\ A = \{x_1, \ldots, x_n\}$ of n data points $\forall n \leq N$
N is the total number of available features in Dataset A

Naïve Bayes algorithm:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

KNN algorithm:

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

SVM algorithm:

$$distance = \frac{y(Q^T X + b)}{\|w\|}$$

$$margin = \frac{1}{\|w\|}$$

Logistic Regression algorithm:

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & if \quad y = 1 \\ -log(1 - h_\theta(x)) & if \quad y = 0 \end{cases}$$

Random Forest:
Random forest randomly drawn a bootstrap sample B from the training data A. Construct decision tree from the drawn bootstrapped sample B. Combine the decision of all threes and find the highest votes as the class for the sample.

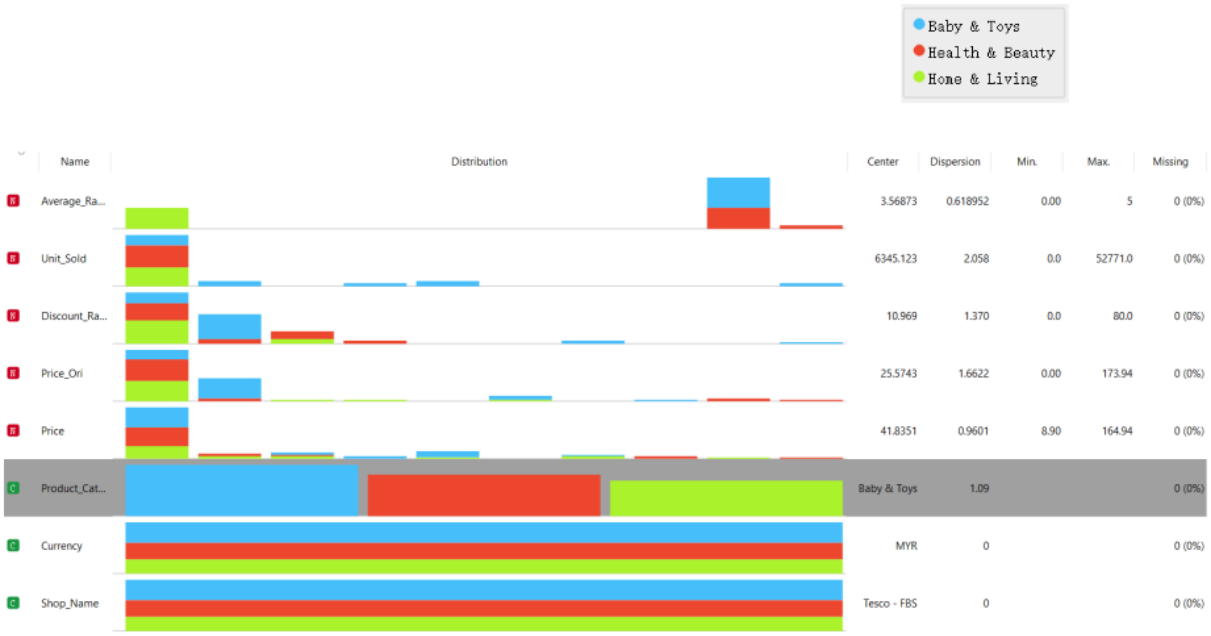

Figure 4 shows dataset used in this research was scraped from Shopee Tesco Shop. There are 10 attributes in this product dataset

The dataset consists of 65 instances, the dataset used is small. There are 5 numerical attributes and 5 categorical attributes. The figure above shows the distribution of 5 numerical attributes and 3 categorical attributes.

There are 2 more attributes which are product name, product description, and product brand being classified in meta-attributes. Therefore, we are going to test for the classification model without processed the 3 meta-attributes first. After that, we are going through another round of testing with processed meta-attributes into numerical attributes by text processing technique. The figure above shows that the product category distributes quite evenly which is our target attribute in this research.

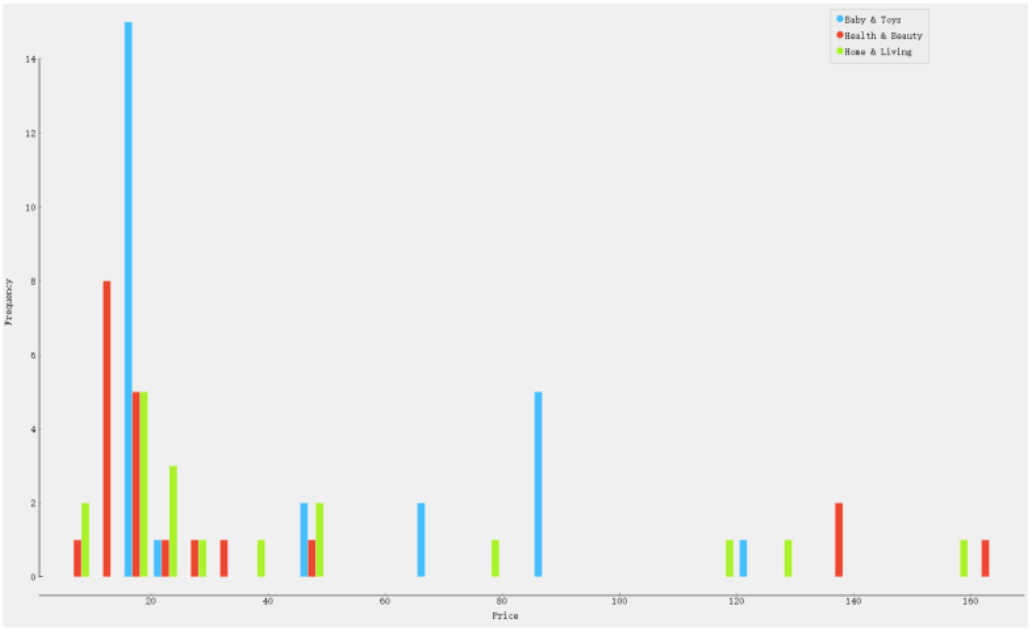

Figure 5 shows there are two attributes should be highlighted among the rest of the attributes

The first one is the price of the products. Most of the price of products saturated below RM40, especially Baby & Toys and Health & Beauty products. This attribute may not be a good feature in a product classification model.

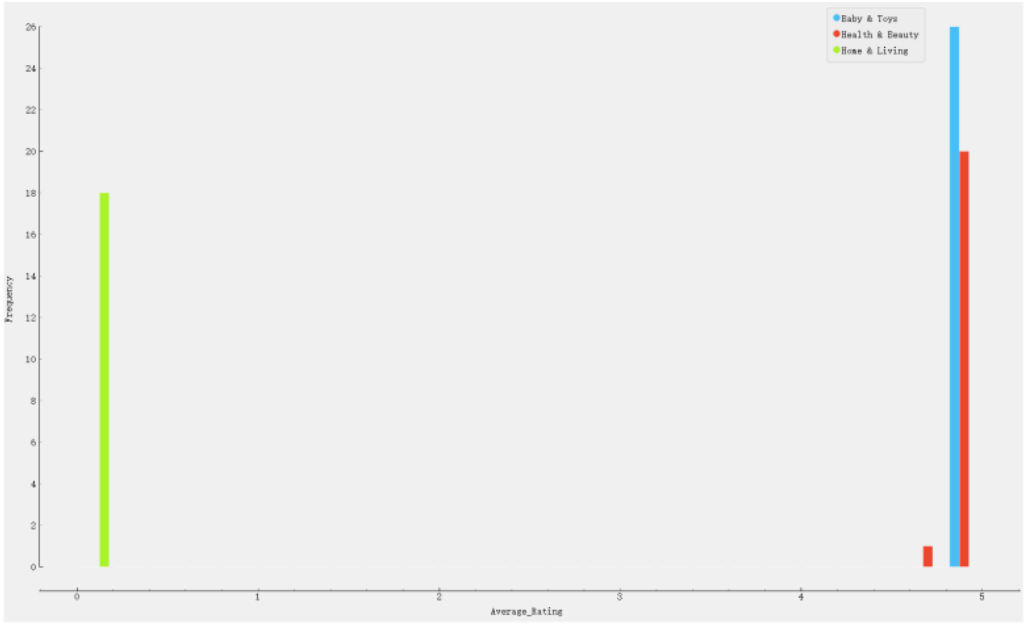

Figure 6 shows attribute is the average rating of the products

The distribution of the average rating of Home & Living products are relatively low, but Baby & Toys and Health & Beauty products are very high.

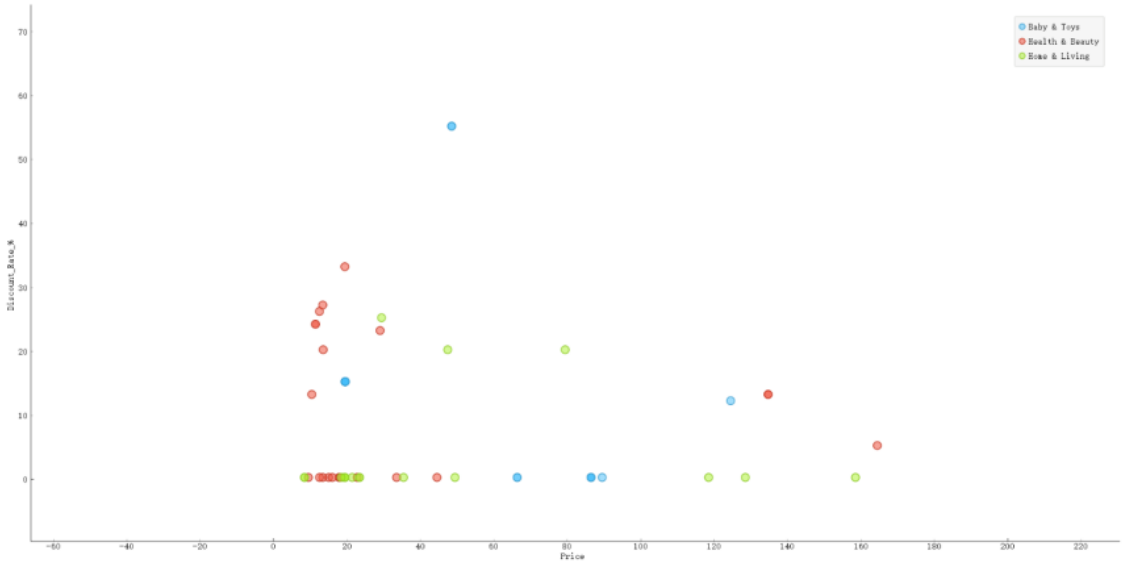

Figure 7 shows for numerical attributes, price and discount rate may be more appropriate to act as a feature in the classification model for new products compare with average ratings and units sold.

This is because new products would not have average ratings and units sold, thus, these two features are not supposed to be involved in the following model development process. From the scatter plot above we can see that Health & Beauty products are concentrated on the left side of the graph which means the price of Health & Beauty products are relatively low with various kinds of discount rates.

These numerical attributes may not enough to build a good classification model, we still need to preprocess the product names and product descriptions. These two attributes must be tokenized and convert into numerical attributes that can be consumed by the classification model.

5. Data Preprocessing on data

In this research, data was scraped from the Shopee website, the raw data was pre-processed before an input for modeling. From the data description and visualization, we observed there is a missing value for the "Discount Rate" column.

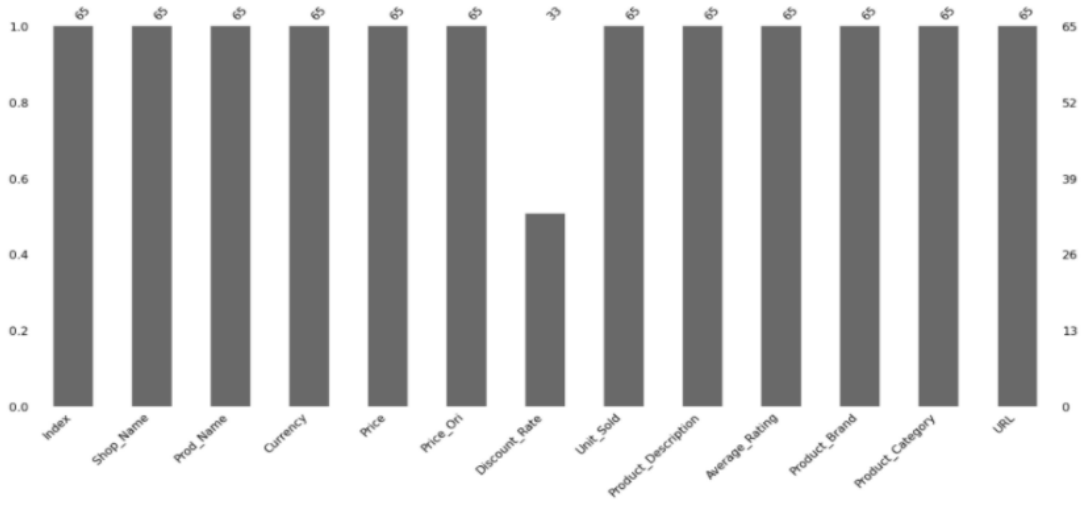

Figure 8 Missing Value Visualization Bar Chart

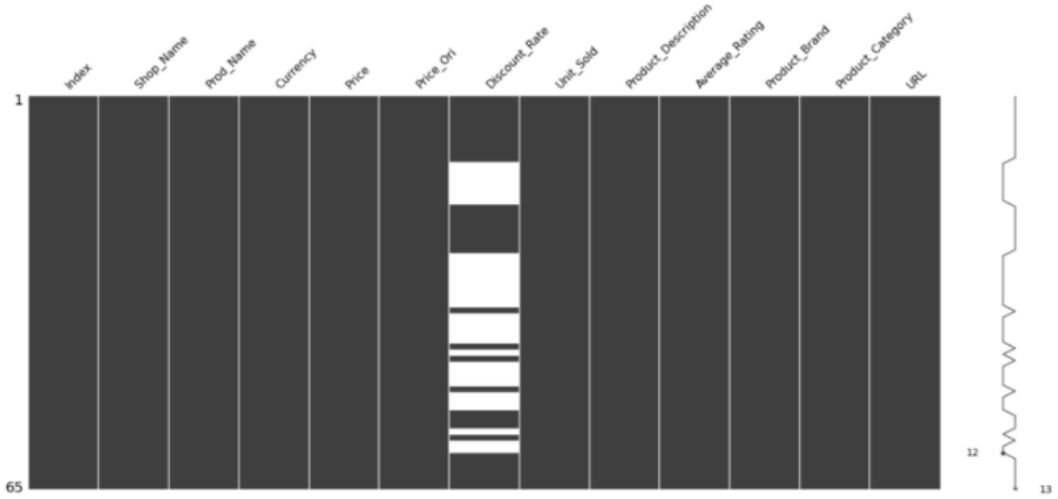

Figure 9 Missing Value Visualization Matrix

These missing values were then replaced with value 0 to remove any missing value during the analysis.

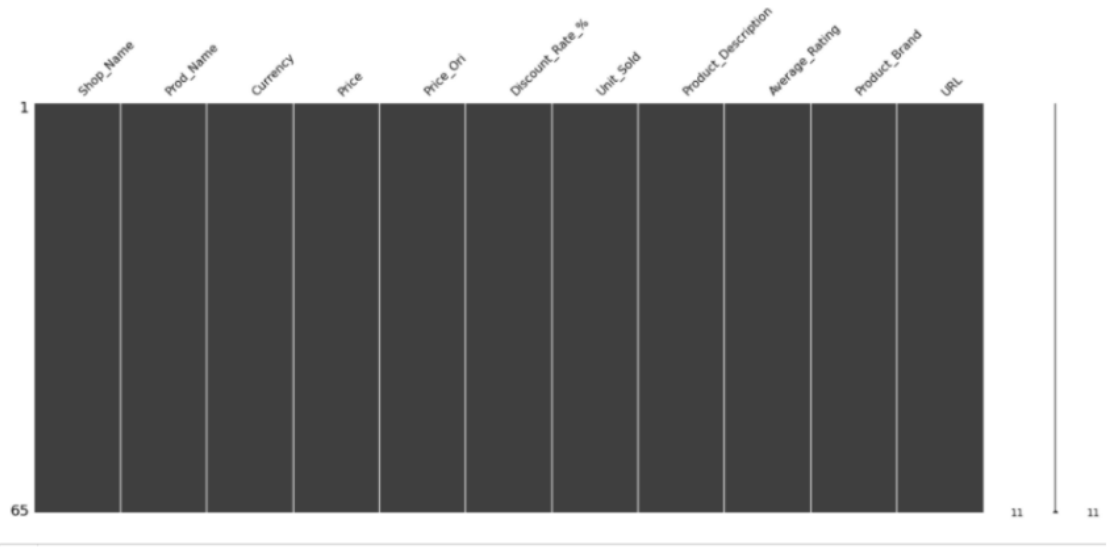

Figure 10 Missing Value Visualization Matrix after Data Cleaning

Besides removing missing values, data was also organized decently, target attribute which is the product category was defined before applying the classifier for prediction.

4.1 Text Preprocessing

Since there is the text found on raw data at column "Product Name", "Product Description", "Product Brand", and "URL", text preprocessing was performed. All the words were transformed into lower case. Tokenization was also performed to remove stop words and delimiters. After that, the words can be transformed into numerical representation by counting the term frequency in the data. These features will then be used for text classification.

Figure 11 Word Cloud before text preprocessing

Figure 12 Word Cloud after text preprocessing

From the word cloud, we could observe the delimiter, for example '(', ')' are the most dominant text before the text preprocessing, this will be misleading and causing low prediction during modeling. Thus, tokenization was performed, and after the text preprocessing, the dominant words are 'dutch', 'lady', '900g', 'tesco'.

6. Results and Discussion

In this research, since the input dataset is small with 65 observations, we are sampling our data using 5 folds stratified cross-validation. Different models have been tested and evaluated with the Area Under Curve (AUC), Classifier Accuracy (AC), F1, precision, and recall score.

To study the model optimization in product classification, we compared the modeling result with and without considering the product name as one of the features during prediction.

From the observation, overall models show better prediction after the text analytics, except SVM.

From the analysis, all the trained models listed in the table have a high Area Under Curve (AUC)

Model	AUC	CA	F1	Precision	Recall
kNN	0.929	0.800	0.791	0.797	0.800
Tree	0.940	0.923	0.924	0.925	0.923
SVM	0.990	0.923	0.923	0.938	0.923
Random Forest	0.987	0.954	0.954	0.955	0.954
Naive Bayes	0.977	0.892	0.892	0.906	0.892
Logistic Regression	0.980	0.985	0.985	0.985	0.985

Figure 13 Average Evaluation Result across Target Class before Text Analytics

Model	AUC	CA	F1	Precision	Recall
kNN	0.929	0.800	0.791	0.797	0.800
Tree	0.939	0.923	0.924	0.925	0.923
SVM	0.790	0.400	0.229	0.160	0.400
Random Forest	0.981	0.908	0.907	0.909	0.908
Naive Bayes	0.998	0.985	0.985	0.985	0.985
Logistic Regression	0.980	0.985	0.985	0.985	0.985

Figure 14 Average Evaluation Result across Target Class after Text Analytics

which indicates the models are good at prediction and able to distinguish between classes. From the average result, we observed that all models give reasonable accuracy value, let's start with figure 10, the evaluation result before the text analytics process. The Logistic Regression has the highest prediction accuracy which stands for at least 98.5% followed by Random Forest Model which collects 95.4% accuracy. Decision Tree shared

the same results with SVM which is 92.3% following by Naïve Bayes classified accurately at 89.2% and lastly kNN with the lowest 80%.

In figure 11, after we completed the text analytics process, we obtained major different results that we can share, we observed certain technique can sustain their accuracy value, some dropped significantly, and some technique surprisingly performed a very outstanding result of accuracy. Logistic Regression sustained the same accuracy results before and after the text analytics result, sharing the same results would be Naïve Bayes which garner an additional 9.3% of accuracy level, which resulted from both Logistic Regression and Naïve Bayes shared the same results at 98.5% of accuracy level. Following by a Decision Tree that sustained the same results before and after the text analytics process at level 92.3%. Random Forest dropped a few percentages of accuracy level after the text analytics process at 90.8%. KNN record the same level of accuracy level at 80% and lastly SVM recorded the lowest accuracy level at only 40%, a total of 52.3% decrement from the original value before the text analytics process.

A confusion matrix has been generated for each model to evaluate the model performance. The confusion matrix further supporting the score from the evaluation matrix. The Logistic Regression able to classify for all labels, except 1 item was misclassified as Health & Beauty category, while the actual category is Baby & Toys.

		Predicted			
		Baby & Toys	Health & Beauty	Home & Living	Σ
Actual	Baby & Toys	25	1	0	26
	Health & Beauty	0	21	0	21
	Home & Living	0	0	18	18
	Σ	25	22	18	65

Figure 15 Confusion Matrix of Logistic Regression before and after Text Analytics

The Logistic Regression shared the same results before and after the text analytics process. Logistic Regression predicted 25 items categorized as Baby & Toys correctly, 21 items as Health & Beauty and predicted 1 item wrongly as Baby & Toys, besides this technique classified 18 items correctly for the Home & Living category. This resulted in why Logistic Regression deserved a high prediction value even without a text analytic process.

		Predicted			
		Baby & Toys	Health & Beauty	Home & Living	Σ
Actual	Baby & Toys	24	2	0	26
	Health & Beauty	2	19	0	21
	Home & Living	0	1	17	18
	Σ	26	22	17	65

Figure 16 Confusion Matrix of Decision Tree before and after Text Analytics

The Decision Tree shared the same results before and after the text analytics process. Decision Tree predicted 24 items categorized as Baby & Toys correctly, 2 wrongly predicted and predicted 19 items as Health & Beauty correctly, and predicted 2 items wrongly as Baby & Toys and Home & Living. Besides

this technique classified 17 items correctly for the Home & Living category.

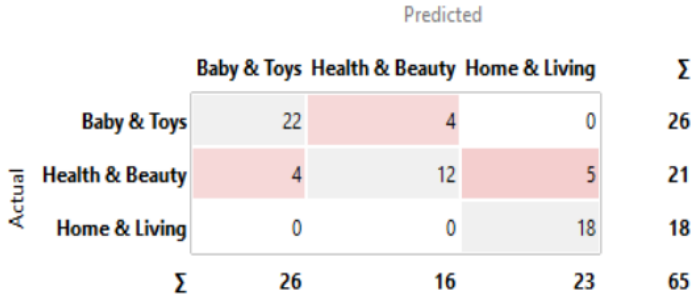

Figure 17 Confusion Matrix of KNN before and after Text Analytics

The KNN shared the same results before and after the text analytics process. KNN predicted 22 items categorized as Baby & Toys correctly, 4 wrongly predicted and predicted 12 items as Health & Beauty correctly and predicted 4 items wrongly as Baby & Toys. Besides this technique classified 18 items correctly for the Home & Living category and 5 wrongly predicted as Health & Beauty. The inconsistent prediction resulted from an average result for this technique.

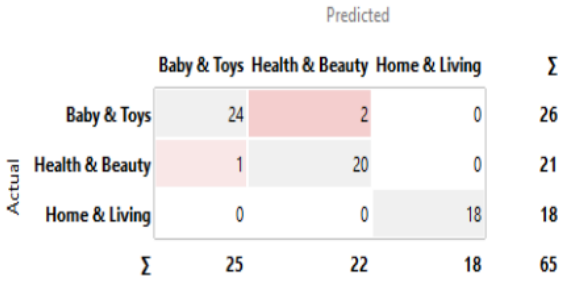

Figure 18 Confusion Matrix of Random Forest before Text Analytics

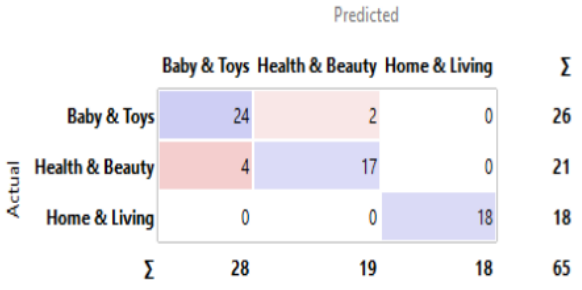

Figure 19 Confusion Matrix of Random Forest after Text Analytics

From the confusion matrix, it is proven that text analytics does not help in improving the classifier prediction.

Without the text features, the random forest model gives a better prediction accuracy.

The Random Forest showed a few percentages of decrement after the text analytic process. However, the model can predict correctly 24 on Baby & Toys, 17 on Health & Beauty, and 18 predicted correctly on Home & Living. However, the model performed much better without text analytics processes.

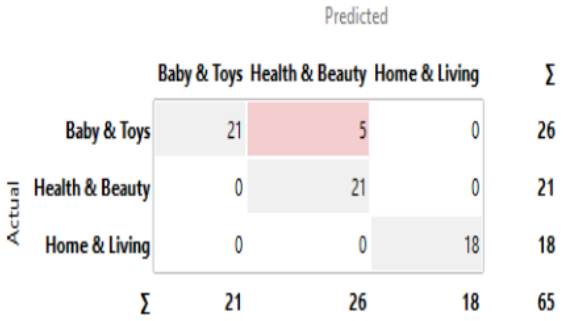

Figure 20 Confusion Matrix of SVM before Text Analytics

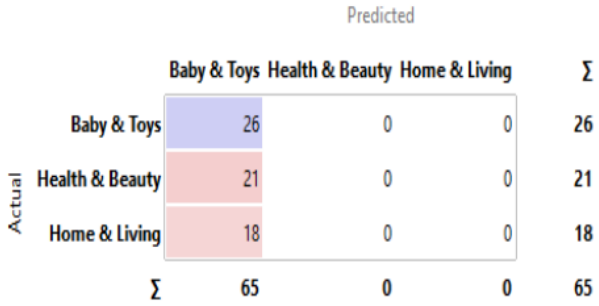

Figure 21 Confusion Matrix of SVM after Text Analytics

From the confusion matrix of SVM, it is proven that the text features are degrading the model prediction. The SVM performed better without text analytics with the performance of prediction shows 21 predicted correctly on Baby & Toys, 21 on Health & Beauty, and 18 on Home & Living which shows a very outstanding model performing on classifying the items accordingly. However, after the text analytics, it shows the prediction has dropped significantly with a total wrong prediction of 39 items, which 21 wrongly predicted on Health & Beauty and 18 wrongly predicted on Home & Living.

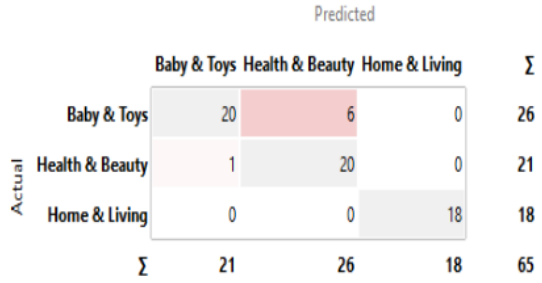

Figure 22 Confusion Matrix of Naive Bayes before Text Analytics

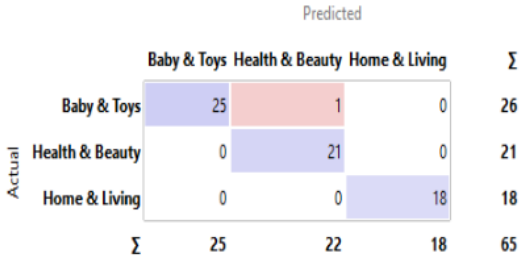

Figure 23 Confusion Matrix of Naive Bayes after Text Analytics

From the analysis result above, we conclude that text features are not necessarily improving the classifier model, it depends on the model algorithm. From the research above, text feature plays a significant role in the Naïve Bayes model, and it is improving the prediction accuracy by 0.093. It gives no difference in prediction accuracy to most of the models such as Logistic Regression, Decision Tree, and KNN models. However, the text features are degrading the results from Random Forest and SVM classifier.

The Naïve Bayes before the text analytics process able to predict 20 items correctly for both Baby & Toys and Health & Beauty and 18 items Home & Living. After the text analytics, the Naïve Bayes model able to pull another 5 more items into a correct zone for Baby & Toys, and 21 correctly predicted for Health & Beauty and 18 on Home & Living. Naïve Bayes performed much better after the text analytics process.

6.1 Results Evaluation

ROC Analysis:

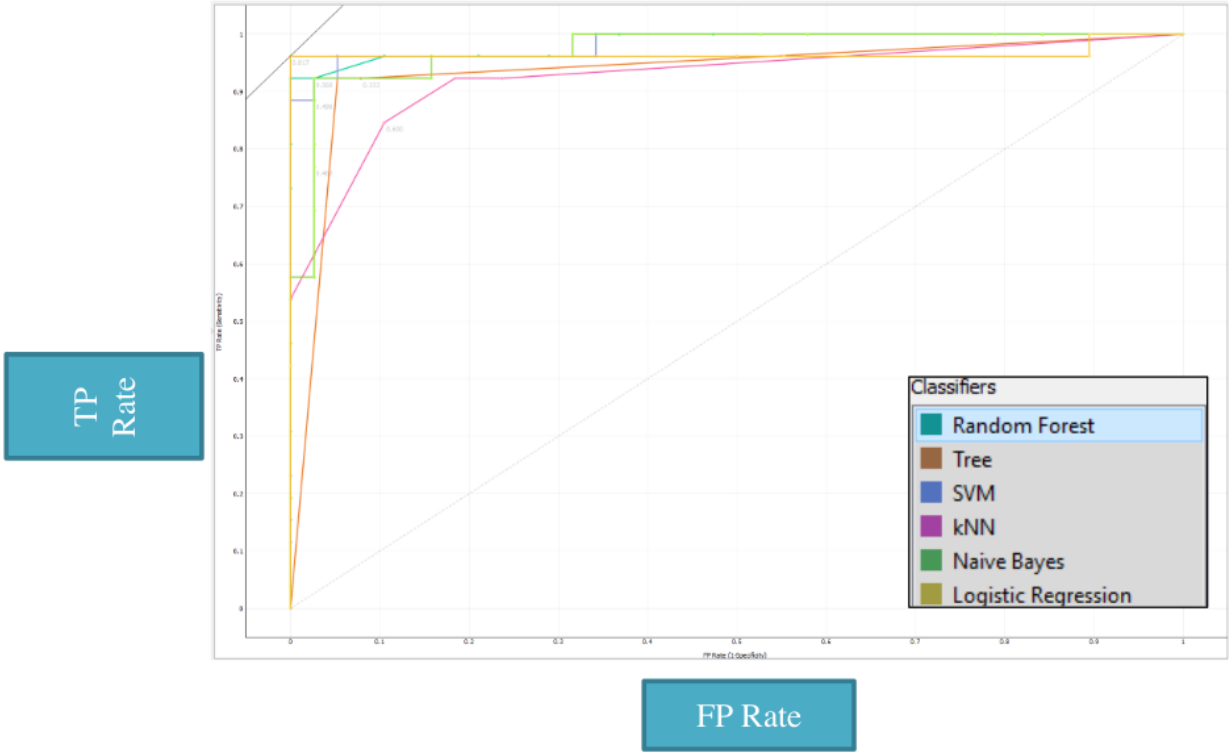

Figure 24 ROC Analysis with Target Class Baby & Toys before Text Analytic

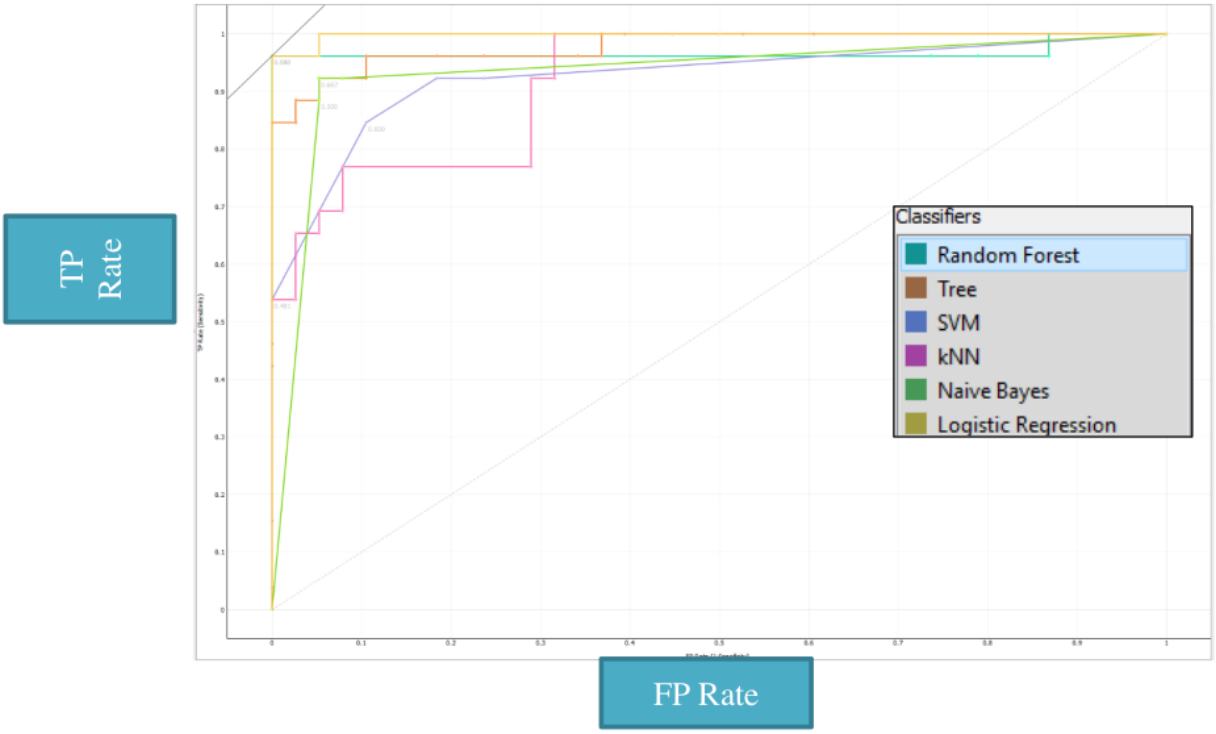

Figure 25 ROC Analysis with Target Class Baby & Toys after Text Analytics

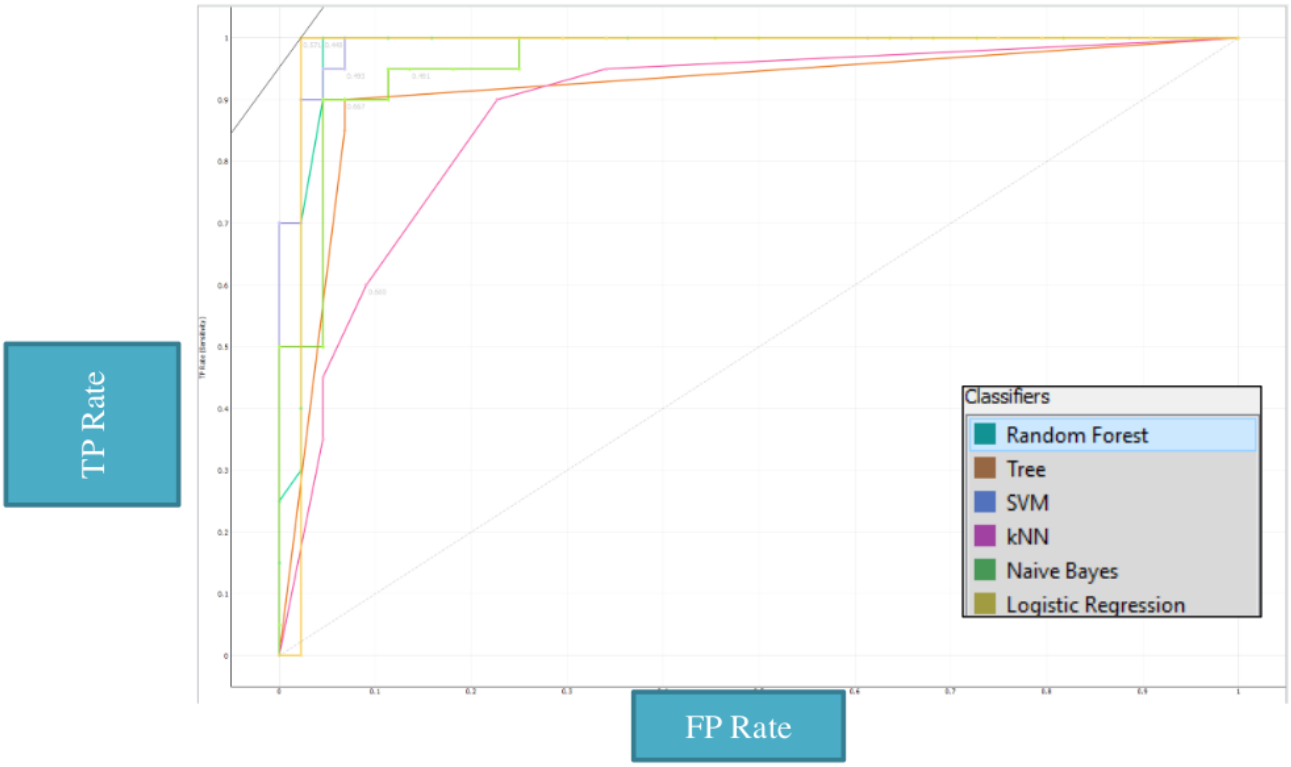

Figure 26 ROC Analysis with Target Class Health & Beauty before Text Analytics

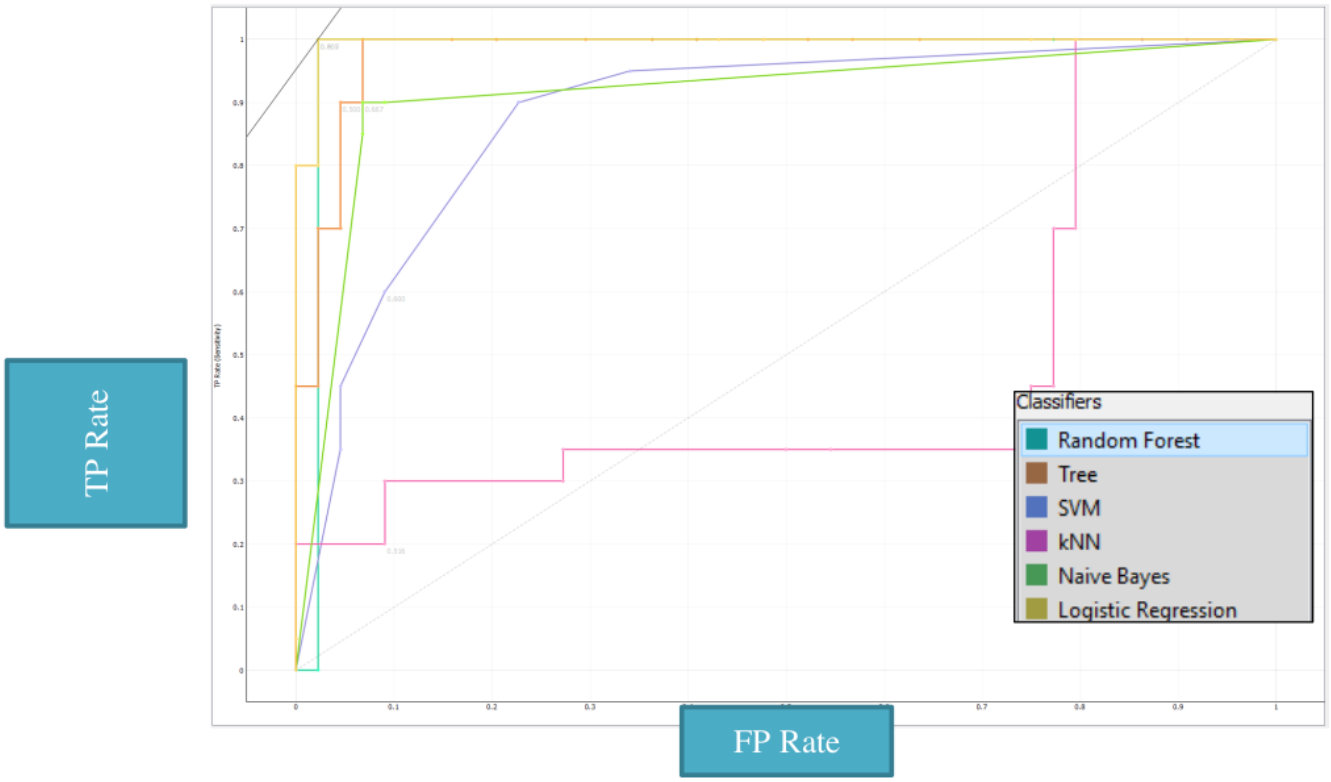

Figure 27 ROC Analysis with Target Class Health & Beauty after Text Analytics

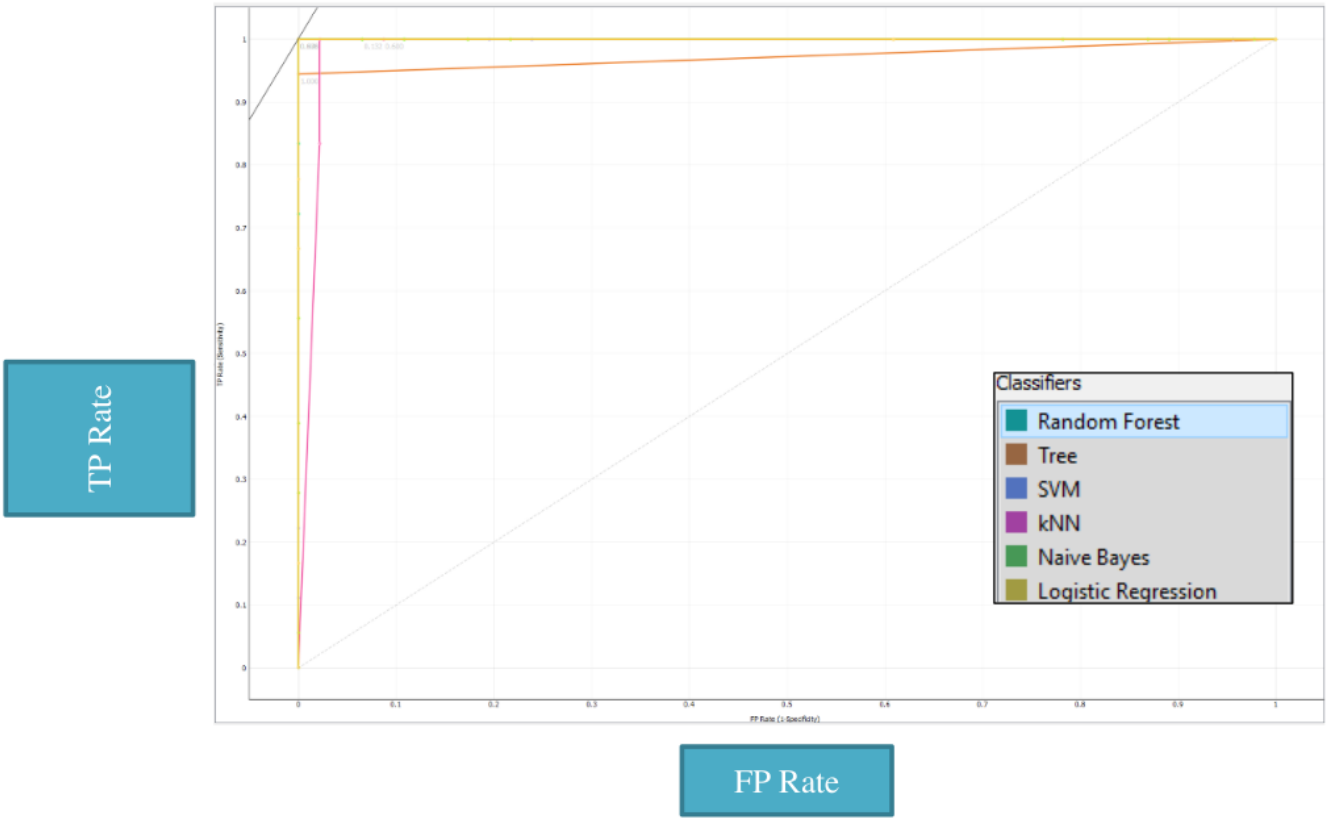

Figure 28 ROC Analysis with Target Class Home & Living before Text Analytics

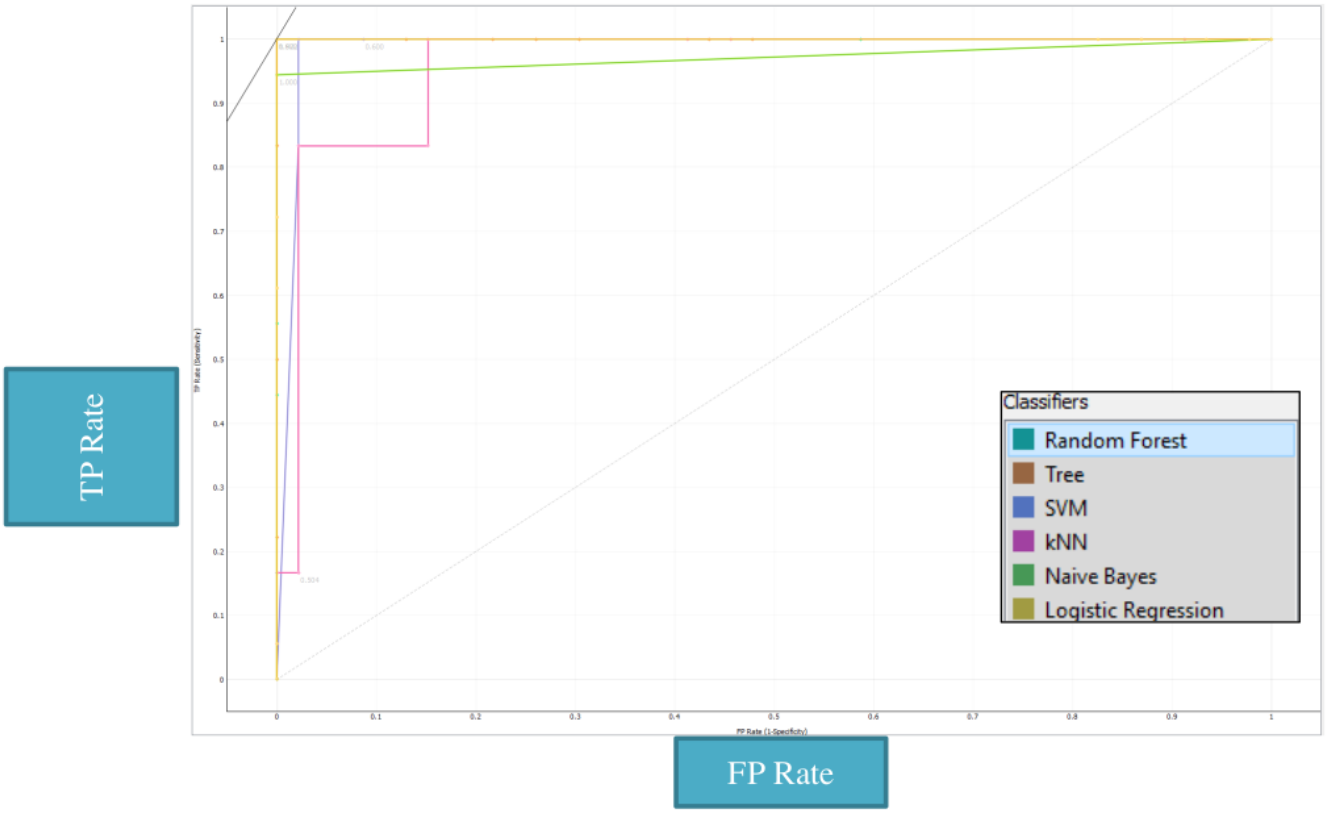

Figure 29 ROC Analysis with Target Class Home & Living after Text Analytics

From the ROC analysis graph above, the home & living give a consistent prediction, most of the models able to predict this category precisely. From the ROC analysis for models before text analytics, the AUC value is overall reasonable. While, the ROC analysis shows the result with the text analytics, SVM show lower ROC value especially when classifying Health & Beauty item.

Since the logistic regression has the best performance among all the supervised algorithm that we tested, we continue to tune its parameter to improve its performance. Two parameters can be tuned by using the orange tool which are regularization types and the strength of regularization (C). Two types of regularization are available to choose which are Lasso (L1) and Ridge (L2). By default, the orange tool helps us to choose L2 regularization and C=10 for the logistic regression model. The regularization technique helping us to avoid over-fitting by preventing the model to have a high value of the coefficient. When we tried to change the regularization type to L1, the accuracy of the model reduces to 96.9% from the original 98.5%. This may because of the limitation of lasso regularization, sometimes lasso is having difficulties in predicting some kinds of data. For instance, the model has greater number of predictors than the number of data points, lasso will automatically pick at most n predictors as non-zero, in this case of the predictors is important for that particular problem, then this action will give negative impacts on the performance of the model. The other scenario is lasso will select one of the two or more highly collinear variables randomly, and this obviously not giving a positive effect on the prediction result.

7. Conclusion

In this study, we have carried out six types of data mining technique to perform on building an efficient classification algorithm to ensure the objectives that we have set up:

1) Random Forest
2) Decision Tree
3) SVM
4) KNN
5) Naïve Bayes
6) Logistic Regression

The results of accuracy that we have obtained in our case study, as shown below:

Table 2 shows the comparison of accuracy results for each data mining technique.

Data Mining Technique	Before Text Analytics (%)	After Text Analytics (%)	Accuracy Value Difference
Random Forest	95.4	90.8	-4.6
Decision Tree	92.3	92.3	± 0
SVM	92.3	40.0	-52.3
KNN	80.0	80.0	± 0
Naïve Bayes	89.2	98.5	+9.3
Logistic Regression	98.5	98.5	± 0

In short, Logistic Regression and Naïve Bayes recorded the same accuracy level after the text analytics process, however, Logistic Regression does not improve after the text analytics process except for Naïve Bayes. The text analytics translates unstructured data into structured data. One of the reasons that text analytics work well with Naïve Bayes because of its advantage that can create fictitious input documents by characterizes the problem which can predict unforeseen data as generative data. Therefore, both of the models have been found successful in this case study. Next, SVM has been found to perform poorly because it is prone to over-fitting if the number of features is much greater than the number of samples. Besides, after the text mining, it has removed several key parameters that need to be set correctly to achieve the best classification results for any given problem. Therefore, SVM has been found performing unsuccessfully in this case study. What we realized from this case study is the tuning of the parameter of the model is essential to build a model with good accuracy. We can see that the orange tool helps us to choose suitable parameters for our logistic regression model but not the SVM model. Thus, we need to investigate carefully and testing the parameter that can be tuned for each model. For instance, we tested with lasso regularization replacing the ridge regularization in logistic regression which results in reducing the model accuracy.

The significance of the high accuracy level of Naïve Bayes and Logistic Regression helps to improve the existing method in classifying each item more accurately. The scenario of the machine learning technique perform outstandingly in this case study, will reflect on the Tesco FBS scenario to classifying items accordingly without spending so much time to assign each item to and spend unnecessary effort in categorizing each item. The technique is important to show that a good and high accuracy level of prediction will help in the business effort in accomplishing

their business functions by depending on the machine learning model that we have found out in this case study.

We also found out that the KNN algorithm that performing very well in classifying retail products from an online store (Nair. V., Malhotra. R., Mohapatra. S., Maknoor. N. 2018) not performing well in this Shopee dataset. We may conclude that different kinds of the dataset may need different kinds of an algorithm to well classifying the products. In our case, the Tesco dataset works better with Naïve Bayes and Logistic Regression model. Besides, the case study will play a significant role in other business supply chain to help automate several mundane tasks and allow the enterprises to focus on more strategic and impactful business activities. We suggest that future work can be done on studying the supervised with a large dataset and investigate image analytics using the advanced convolutional model to perform product classification.

Bibliography

[1] Adebola, O., Onyekwelu, B. & Orogun, A. (2019). Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach Computer Systems Performance Evaluation View project cellular network optimization View project Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach. International Journal of Innovative Research in Science, Engineering and Technology, Hauser 2007, 8391–8402. https://doi.org/10.15680/IJIRSET.2019.0808006

[2] An, J., Kwak, H., Jung, S. gyo, Salminen, J. & Jansen, B. J. (2018). Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. Social Network Analysis and Mining, 8(1). https://doi.org/10.1007/s13278-018-0531-0

[3] Bajaj, P., Ray, R., Shedge, S., Vidhate, S., Nikhilkumar, P. D. & Shardoor. (2020). SALES PREDICTION USING MACHINE LEARNING ALGORITHMS. International Research Journal of Engineering and Technology, 7(June), 3619–3625. https://www.irjet.net/archives/V7/i6/IRJET-V7I6676.pdf

[4] Banks, T. & Hang, W. (2019). Machine Learning Applied to Pack Classification. December. https://doi.org/10.1177/ToBeAssigned

[5] Chen, D., Sain, S. L. & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing and Customer Strategy Management, 19(3), 197–208. https://doi.org/10.1057/dbm.2012.17

[6] Donati, L., Iotti, E., Mordonini, G. & Prati, A. (2019). Fashion product classification through deep learning and computer vision. Applied Sciences (Switzerland), 9(7), 1–22. https://doi.org/10.3390/app9071385

[7] Joseph, S. R., Hlomani, H. & Letsholo, K. (2016). Data Mining Algorithms: An Overview. International Journal of Computers & Technology, 15(6), 6806–6813. https://doi.org/10.24297/ijct.v15i6.1615

[8] Muhammad, N., Mathivanan, N., Ghani, N. A. & Janor, R. M. (2019). E-Commerce Product Classification Using Supervised Learning Models. International Journal of Engineering Research & Technology, 8, 214–218.

[9] Nair, V., Mohapatra, S. K. & Malhotra, R. (2018). A Machine Learning Algorithm for Product Classification based on Unstructured Text Description. International Journal of Engineering Research & Technology, 7(06), 404–407.

[10] Nath, P. D., Das, S. K., Islam, F. N., Tahmid, K., Shanto, R. A. & Rahman, R. M. (2017). Classification of Product Rating Using Data Mining Techniques. Studies in Computational Intelligence, 710(March), 27–36. https://doi.org/10.1007/978-3-319-56660-3_3

[11] Ray, S. (2017). Understanding Support Vector Machine(SVM) algorithm from examples (along with code). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[12] Ristoski, P., Petrovski, P., Mika, P. & Paulheim, H. (2018). A machine learning approach for product matching and categorization. Semantic Web, 9(5), 707–728. https://doi.org/10.3233/sw-180300

[13] Safara, F. (2020). A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic. Computational Economics, 0123456789. https://doi.org/10.1007/s10614-020-10069-3

[14] Singh, H. & Neware, S. (2020). Improving customer segmentation in e-commerce using predictive neural network. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 2326–2331. https://doi.org/10.30534/ijatcse/2020/215922020

[15] Thobani, S. (2018). Improving E-Commerce Sales Using Machine Learning. Massachusetts Institute of Technology, 1(1988), 1–176.

[16] Umaashankar, V., Shanmugam, G. S. & Prakash, A. (2019). Atlas: A Dataset and Benchmark for E-commerce Clothing Product Categorization. ArXiv, 1–12.

[17] Verdenhofs, A. & Tambovceva, T. (2019). Evolution of Customer Segmentation in the Era of Big Data. Marketing and Management of Innovations, May, 238–243. https://doi.org/10.21272/mmi.2019.1-20

[18] Wirojwatanakul, P. & Wangperawong, A. (2019). Multi-Label Product Categorization Using Multi-Modal Fusion Models. ArXiv.

[19] Xia, Y., Levine, A., Das, P., Di Fabbrizio, G., Shinzato, K. & Datta, A. (2017). Large-scale categorization of Japanese product titles using neural attention models. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2(2016), 663–668. https://doi.org/10.18653/v1/e17-2105

[20] Xu, H., Liu, B., Shu, L. & Yu, P. (2019). Open-world Learning and Application to Product Classification Open-world Learning; Product Classification ACM Reference Format. Www, 3413–3419. http://delivery.acm.org/10.1145/3320000/3313644/p3413-xu.pdf?ip=133.19.43.2&id=3313644&acc=OPEN&key=D2341B890AD12BFE.E7C6D16F16E10784.4D4702B0C3E38B35.6D218144511F3437&__acm__=1558927726_147fa56db7b3d23f53b2f01029fdc845#URLTOKEN#

Data_Mining_Final

ORIGINALITY REPORT

5%	4%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hu Xu, Bing Liu, Lei Shu, P. Yu. "Open-world Learning and Application to Product Classification", The World Wide Web Conference on - WWW '19, 2019 Publication	1%
2	dspace-unipr.cineca.it Internet Source	1%
3	www.semantic-web-journal.net Internet Source	1%
4	www.irjet.net Internet Source	1%
5	doctorpenguin.com Internet Source	1%
6	Norsyela Muhammad Noor Mathivanan, Nor Azura Md.Ghani, Roziah Mohd Janor. "Performance Analysis of Supervised Learning Models for Product Title Classification", IAES International Journal of Artificial Intelligence (IJ-AI), 2019 Publication	1%

7 Submitted to Asia Pacific University College of Technology and Innovation (UCTI) 1%

Student Paper

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		