Optimal Price Prediction For Online Retailers Using Machine Learning

Shardul Rajhans

Final Thesis Report

JUNE 2022

ABSTRACT

With the increasing amount of data for online e-commerce retailers, it becomes very difficult to predict the prices of each product that lists on the sites. There are several factors that majorly affects the price of the products like designs, brands, color, size, trends, season, etc. thus making the predictions difficult for the organizations. Also, with the increasing competition, it is important for an online retailer to retain their customers and attract several other customers with improving pricing strategies, improving recommendations, and maintaining the balance between sales, revenue, and profit. Dynamic pricing strategies like Segmented Pricing, Time-Based, Penetration, Competitive Pricing are heavily used and constantly improved along with monitoring the sales and revenue. This becomes a challenge to provide an accurate price that will be suitable for the customer and organization. This study aims on predicting the price of the product using textual information like product information and its characteristics. Also, competitive pricing methodology is used to analyze the market demand and provide suitable price using linear programming. Predictive analysis is used along with Natural Language Processing techniques for textual data processing. Also, machine learning models like Long Short-Term Memory, Gated Recurrent Units, Convolutional Neural Networks are analyzed based on the performance metrics like RMSLE, MAE and RMSE. CNN and LSTM has generated almost similar performance metrics. The CNN performed best with 7.2506 as RMSE and 0.2115 and RMLSE. However, the LSTM model has also similar performance metrics as compared with CNN. LSTM has 7.7766 as RMSE and 0.2113 as RMLSE. A dynamic range is calculated statistically using the competitors' price and predicted price using machine learning. A linear programming optimization based on linear relaxation is used to select the optimal price which maximizes the overall revenue. Using this modelling strategy, it will help the organizations to maintain the standard prices along with the dynamic approach to think for the discounts and coupon offers. Inventory management can be considered as future application of this study. This will again, increase doors for improving recommendation-based approaches.

TABLE OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PNR……	Promotion with No Recommendation
MCR……...	Minimum Cost Ratios
TF-IDF……	Term Frequency - Inverse Document Frequency
SGD……	Stochastic Gradient Descent
KNN……	kth Nearest Neighbour
SVM……	Support Vector Machines
LSVM…….	Linear Support Vector Machines
DT……	Decision Trees
ACC…….	Area Under the Curve
LDA…….	Linear Discriminant Analysis
LSA…….	Latent Semantic Analysis
RF…….	Random Forest
RNN…….	Recurrent Neural Networks
MLP…….	Multilayer Perceptron
LSTM…….	Long Short-Term Memory
NLP…….	Natural Language Processing
NB…….	Naive Bayes
EMD…….	Empirical Mode Decomposition
KRR…….	Kernel Ridge Regression
RVFL…….	Random Vector Functional Link
ELM…….	Elaboration Likelihood Model
BTS…….	Bureau of Transportation Statistics
FAA…….	Federation Aviation Administration
RAM…….	Read Only Memory
MSE…….	Mean Squared Error
MAE…….	Mean Absolute Error
LGBM/LightGBM……	Light Gradient Boosting Machine
MAPE……	Mean Average Percentage Error
CNN…….	Convolutional Neural Networks
ARIMA……	Autoregressive Integrated Moving Average
RMSE……	Root Mean Squared Error

RMSLE…… Root Mean Squared Logarithmic Error

EDA……. Exploratory Data Analytics

CSV……. Comma Separated Files

XGBoost…… Extreme Gradient Boosting Machines

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

1.1 Background of the study

The apparel industry is recognized as one of the most important industries that generates high revenue and contributes to the economy of the country. The main processes for clothing and garment manufacturing include designing, manufacturing, supply chain management and retailing. Apparel manufacturing is characterized by a wide range of product designs and input materials, variable production volumes, high demand for product variety and quality, extreme competitiveness. Being a wide range of designs and styles, the complexity of the manufacturing process increases (Nayak and Padhye, 2015). The retailers are responsible for delivering products to the end consumers.

According to (Güven and Şimşir, 2020), the ability of companies for the survival in the highly competitive market usually depends on their ability of incorporating machine learning techniques and making business decisions. Application of machine learning techniques like recommendations, demand prediction, sales forecasting, dynamic pricing, analysis of customer's lifetime value, etc. is the basic need to accommodate in the market. The forecasting of processes is not facilitated by changing customer needs and trends. For every wrong prediction made, it may result in devasting effects like decreased sales, reputation, and income loss, etc. For this reason, machine learning and predictive analytics is one of the most important inputs for companies to reach their short-term goals. Considering the requirement for completeness of data, it's possible to fill out the required missing data and ensure quality data integrity during the use of AI systems. With the increasing amount of size of the data, today, it is more reasonable and appreciable for AI systems to process the data by creating patterns. In addition, while classical AI methods require more statistical verifications like hypothesis by nature of the method, this does not apply to AI techniques.

With the Volume, Velocity and Veracity (the 3Vs), companies have access to all sorts of information about customer's experiences, financial transactions, inventory management, competitiveness of the marketplace. E-Commerce companies use Artificial Intelligence and Machine Learning Models for providing personalized services to the online shoppers (Myung and Yu, 2020) . Also, there is an increase in demand for the analytics amongst online retailers. Companies like Amazon, Alibaba, etc. make use of Predictive Analytics and Descriptive

Analytics to create Promotional Activities like discounted add flash sales, etc. so that large customers can get attracted.

However, every successful fashion e-commerce and retail brand has not only able to find their competitive advantages but also striving to set new targets, marketing, and branding techniques. Attributes like standard pricing, delivery, packaging, communication, grievance management have become quite important areas for focusing than the actual quality of the product. For establishing a successful online store, brands and organizations need to understand what the customer looks for: low pricing, huge discounts, quality products, unique story telling product descriptions, latest products, etc.

Predicting and standardizing the price of the product is a tough challenge since almost similar products have different specifications, quality, availability, colors, sizes, etc. and thus can have different prices. For example, one of the following bras costs $80 and the other one costs $20. Can we guess which one of them costs less and better?

Figure 1: Item Description for two different bras.

Although setting the price for product or an item is an old problem in e-commerce domain, there are a vast amount of pricing strategies that organisations can use depending upon the organisational goals. One organisation may target maximizing profitability for each item sold, where as the other organisation might need access to the new markets. (Narahari et al., 2005) At the same time, the technology is allowing sellers to collect detailed data about customers' buying habits, preferences, even spending limits, so they can customize their products and prices. Multiple factors like market competition, reputation, production values, distribution costs, locality, etc. play a key role for retailers to decide the pricing strategy. Using Artificial Intelligence can be a very efficient approach for the retailers as they can benefit from the predictive models to decide the best price considering the several factors and organisation's objectives.

Dynamic pricing is a pricing strategy used by businesses to assign flexible prices for products based on current market demands. (Dynamic Pricing: The Future Of Retail Business, 2016) Typically, the seller makes use of available data about the market to determine its own pricing strategy, such as its competitor's prices, preferences of customers, buying frequencies. There has been research related to Automated Dynamic Pricing with the assumption of a little complete information about the market. (Dasgupta and Das, 2000) Often, retailers dynamically alter the prices of their product in order to match their competitor's price. A general problem of competition-based pricing is that price wars can arise resulting cyclic pattern or downward spiral. (Bauer and Jannach, 2018) Using dynamic pricing and changing the prices of the product with no objective function in the mind may lead to suboptimal results.

The art of dynamic pricing which is also referred to as real-time pricing, revenue and yeild management. Custimizing the inventory goods after segmentation of the customers on the basis of choice of product and thus adjusting different prices to them is dynamic pricing. Artificial Intelligence and Machine Learning Algorithms should be able to effectively automate decisions for pricing in order to maximize profits, as they can perform decisions for pricing with the help of sophisticated predictions and calculations, by having all the available data into perspective, and changing their strategy of pricing for adapting to a dynamic environment. Moreover, dynamic pricing can be implemented in many ways such as Surge Pricing, Competitive Pricing, Penetration Pricing, Time-Based, Peak Pricing, etc. Organisations make use of data available for picking suitable dynamic pricing strategies which can be easily accomodated in the existing systems for increasing revenue. However, there are both advantages and disadvantages for implemeting dynamic pricing without prior considerations.

1.2 Problem Statement

This study presents the work with an online retailer, Victoria's Secret, as an example of how an online retailer can use its wealth of data, machine learning and predictive analytics strategies for optimizing pricing decisions daily. Victoria's Secret is in the online fashion sample sales industry, where discounts on the products are offered on daily basis and flash sales periodically on designer apparel and accessories. However, the movement in the fashion industry about customer decisions and purchase prediction is uncertain. Several competitors in the market like Hanky-Panky, Calvin Klein, Amazon, etc. also have brick-and-mortar online stores engaging in handling competition, attracting new customers and maintain quality customer base.

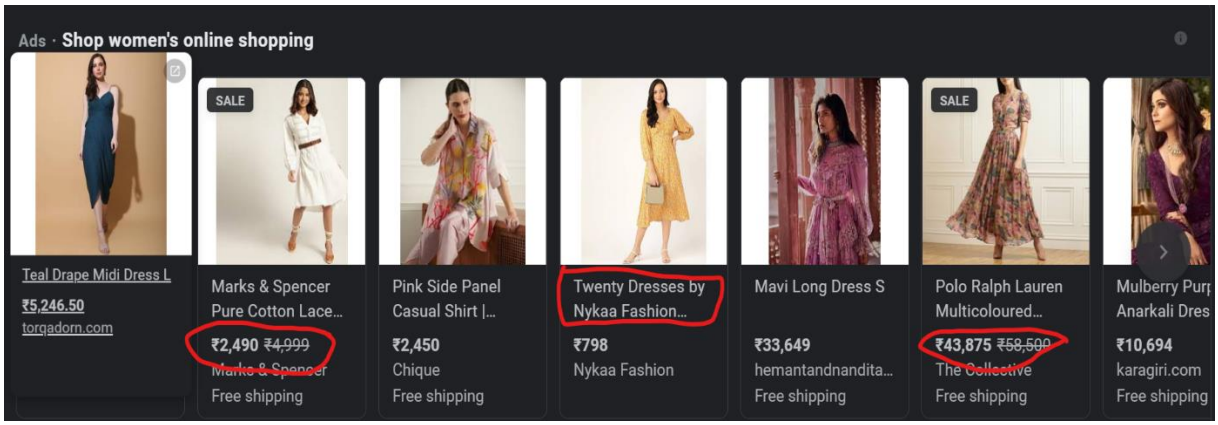


Figure 2: Discount and Item Description (ImageSource: https://google.com)

Upon visiting Victoria's Secret website (Figure2 is a sample image from google, the image from Victoria's Secret is not appropriate for display purpose), customer sees several categories of products to buy. Victoria's Secret also displays a short description as well as discount applied at the time of visit. Also, upon visiting the item, detailed information about the product, styles, available sizes, and colour, etc. are also displayed.

Generally, these online platforms for fashion and retail make no restriction on how the sellers set prices for their product listings on the site. Even with the increase in the data and number of products produced daily, it will be difficult for setting the price of the product based on the materials as well as production costs. However, as compared to the brand-new fashion items, (Han et al., 2020) most of the returned items (due to mismatch in sizes or misinterpretation of colours) listed on the platforms are unique for which price standardization and references can't be found easily, so it can be a great challenge for online retailers and sellers like Victoria's Secret to set strategy for reasonable pricing for their product listings. In other words, the online platforms will benefit a lot using their big data of various fashion items, can be used for providing valuable price suggestions for their fashion items. Also, identifying the mentality of online buyers and because customers rely heavily on the item descriptions as well as characteristics for making purchase decisions, this research signifies a safe assumption that the price of the products can be determined with the help of product characteristics. Moreover, this study aims at designing a price suggestion system first combined with price optimisation techniques for the online fashion items, which will provide effective price suggestions for product listed on the online platform with the help of characteristics of the product.

Also, only standardizing the prices of the product alone is not efficient in terms of competitive market to increase business revenue. To attract customers, organisations need to focus on retaining existing customers as well as attracting new customers by experimenting new

strategies or promotional activities. One of the key challenges that organisations face is setting the discount at individual product level as well as keeping the balance with revenue and sales. Since a large percentage of fashion items are sold on online platform and considering the growth of trends and variants in the products that arises daily, it becomes essential to research on product price optimisation along with attracting more customers and increasing the revenue and profit. Moreover, considering the competition and trends, setting the prices for fashion items too high will lose multiple customers and thus eventually a significant loss in revenue, however, setting the prices for fashion items too low will attract more customers but eventually this might also lead to revenue loss as more appropriate price can also be set for the product.

Considering these problems, this study proposed a two-fold approach for price standardisation and price optimisation. This study will focus on predicting the market price based on product description only and analysing how discount can be calculated from product characteristics using regression techniques. This study will use standard machine learning techniques and algorithms for predicting suitable price for fashion items. Also, this study will formulate price optimisation models for maximizing the business revenue combined with price standardization considering competition.

However, use of standard textual processing techniques like TF-IDF won't help as this study aims for predicting the price based on long textual descriptions, hence SOTA based approaches like LSTM, GRU, CNN will be honoured. Also, one of the key challenges for price standardization is only textual description might not lead to satisfied results hence a more focus will be given on model tuning perspective. The existing approaches for price optimisations and dynamic pricing based on regression trees is nonparametric structure and bucketing of prices on different categories. Hence, a customized linear programming-based approach can be used for optimization of revenue and predicting the prices of products at individual level. As this study will also consider the competitor's data, the prices predicted might lead to customer satisfaction and help in increasing the business revenue.

1.3. Research Questions

This study will be based on the following Research Questions and will try to understand the challenges, significance, methodology, performance, evaluations, and outcome.

1. How to determine the Standardized and Accurate Price for the products considering only product description and considering the characteristics?
2. How does the mrp and discount of the product depend on their characteristics?
3. Which machine learning techniques can be used to optimally predict the product price with the application of Feature Extraction and Engineering?
4. How to calculate the product prices optimally without losing customers and revenue considering the competition in the market?

1.4. Aim and Objectives

The primary goal of this research study is to build a machine learning model to predict the standardized and accurate prices of the products and a statistical and strategical Dynamic Pricing approach will be proposed considering the potential risks impacting the business and revenue. This study will identify the importance of the factors that decides the standardized price of the products and evaluate the performances of different machine learning models using Feature Engineering, Extraction, Natural Language Processing and Deep Learning Techniques. This study will be divided into two major steps. At first, standardized and accurate prices of the product can be predicted. A statistical Dynamic Pricing approach can be used to predict a suitable and optimal price for the product using the predicted standard price and optimization function.

The research objectives are formulated based on the aim of this study which are as follows:

- Perform Exploratory Data Analysis for different characteristics of the product for the better understanding of the price distribution and variations.
- Predict the standardized price of the product from only their textual description as well as considering the product characteristics.
- Analyze how the price and the percentage discount offered on the product varies based on the characteristics?
- Perform data processing and feature extraction to get the relevant information from the data.
- Compare the performances of different machine learning models for optimally predicting the price using different model evaluation rubrics.

- Predict the suitable and optimal price of the products based on optimization function and standardized price.

1.5. Significance of the Study

Prediction of the prices of the products is a tedious task as products may be similar and have almost negligible differences like description, color, size, material quality and demand. With the increase in data, Price Prediction gets very difficult considering the variety of products, competition, seasonal demand changes, location specific factors. Analyzing the factors affecting the product prices and optimizing the range of the prices is the most challenging tasks amongst the online retailers and determines the direction of the growth of the organization and business revenue.

From the literature review in chapter 2, there are different approaches that has already been used for predicting the appropriate price ranges like Demand-Based, Cost-Based, Competitor-Based. This research would be an extension to the existing studies and aims to a two-stage approach (Standardized Price Prediction using characteristics and Competitive-Pricing Analysis) which will be helpful for the organizations. With the two-fold dynamic approach, price standardization and dynamic pricing can be achieved which might lead to significant increase in the sales and revenue. Also, knowing the Price Ranges of the products can open new doors for the appropriate recommendations.

1.6. Scope of the Study

Considering the primary goal of this study, the scope of the study is limited to create a two-fold approach for the prediction of price range of the product. While there have been several studies and novel methods which considered multiple influencing factors, researchers are still conducting the experiments and try multiple permutation and combination of methods to achieve the business objectives. Even, there might be several challenges and multiple factors like inventories, seasonal variations, market ups and downs, etc. that might influence the pricing strategies of the organizations, organizations must constantly monitor, improve, and try different combinations of the pricing strategies (Yin and Han, 2021). Using the two-fold optimization approach for the product pricing, the study aims contribute to one of the pricing strategies and perform predictive analytics techniques to achieve the best results.

The scope of the dataset is limited to textual processing, feature extraction and engineering, exploratory data analysis. After applying preprocessing, the choice of the machine learning approach and model evaluation for the first stage prediction can be a challenging task. This study aims to use LSTM, GRU and CNN based models for extracting information and

predicting price. Advanced methods will be performed if these methods do not generate the desired needs. For the second stage optimization, the dynamic pricing approach is limited to use the linear programming approach for optimal price suggestion. Also, a price optimization function will be used to adjust the product price considering aggregated competitor's data. However, calculation of demand function is not in the scope of this study.

1.7 Structure of the study

The proposed research work has been distributed in the following sections or chapters: The literature review executed for this research work is described in the trailing chapter. Chapter 3 has the proposed methodology for price standardization and price optimisation. Also, the two-fold approach proposed in this study is described in detail in chapter 3. Moreover, chapter 4 describes the detailed stepwise implementation of the proposed machine learning models in details. The results, observations will be described in the chapter 5 and finally in chapter 6, the conclusions and future scope of the study will be proposed based on the gaps observed in the results and observations of the research. Hence, it can be concluded that the proposed research work has attempted to include a complete overview of product price optimization for the selected area of study beginning with background of the study and enlightening on the root causes of the price prediction strategies and dynamic pricing algorithms and finally providing best possible machine learning solutions to predict optimal product prices in advance.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

From past few decades, Retail and E-Commerce companies are generating terabytes of data. The data and information are collected through user logins, credit card transactions, IP addresses, etc. can be used by the organisations for growth of businesses and using personalization to retain and attract the customers. With the increase in the data, companies invest a lot to decide the strategies for pricing of the products. Organisations uses a lot of strategies for pricing like surge pricing, penetration pricing, dynamic pricing, time-based pricing. Further it is essential to identify the best strategy for the product pricing in order to increase the revenue, growth and customer base. Using available information of sales, products, inventories and business evaluations, suitable pricing strategy can be aimed with the help of data science and machine learning algorithms.

Although implementation of pricing strategies is already known to the world and dynamic pricing plays a lot of important roles towards business growth, there is always a learning curve and scope of research to implement it effectively as organisations invest a lot to identify the efficient and sophisticated algorithms for pricing strategies. In this section, a large literature will be reviewed to show the importance of data analytics and data science driven price predictions as well as dynamic pricing strategies. As the dataset used for this study is very recent and there is no related paper published, this section will focus on studies that addressed using of textual information and other characteristics to identify the pricing and some dynamic pricing problems. Literature Review will be conducted understanding the methodology and identifying the gaps and scope of improvement for the regression algorithms, natural language processing, feature extraction and model evaluation rubrics perspectives in particular followed by highlighting the characteristics and future work scope.

2.2 Predictive Analytics in E-Commerce and Retail

E-Commerce and Retail companies makes use of Artificial Intelligence, Machine Learning, Data Science, and mining to make predictions about product demands, analysis of sales, crime and fraud detection, promotion and recommendations, price predictions, maintenance, purchases and shopping habits, etc. for continuously monitoring the business growth and sales. According to the study conducted by (Kumar and L., 2018), the author explained various use cases of the predictive analytics in several domains like E-Commerce, Retail, Banking and Financial Services, Health, and Insurance domain. The author conducted a deep study and a

literature review on the processes, opportunities, modelling and applications in real-world. The author elaborated the use of predictive analytics in detecting fraud, reduction of risk, marketing campaign optimization, operational improvement, clinical decisions and support system, etc. Also, a deep overview widely used machine learning models like decision trees, regression models, artificial neural networks, Bayesian statistics, ensemble learning, gradient boost models, support vector machines, time series analysis, k-nearest neighbour has been stated with applications. On the other hand, the author (Chatterjee, 2019) conducted research on how organisations use Big Data for Personalization and revenue growth. The author explained the application of predictive analytics using big data by collecting data from various resources such as transactions, business activity, click stream, processing video and for customer segmentation to derive granular insights. This study is conducted to analyse the customer profiling, targeted advertisements, predictive analytics, price personalisation and strategies used for improving user experience. According to the study conducted by (Morsi, 2020), e-commerce has a significant growth seeking technological inventions and innovative capabilities dealing with huge amount of data to gain insights for better decisions. This study focuses on the use of predictive analytics in Egyptian E-Commerce domain firms and discussed the predictive analytics processes, predictive model techniques like regression, classification, clustering, association rules, etc. datasets used for predictive analytics to derive the potential insights for business growth.

On the other hand, (Niu et al., 2017) conducted a study exploring online customer search behaviour for Walmart.com. The author used machine learning techniques like Random Forests, Logistic Regression for developing computational models. The study suggests that, with the enhanced approaches and metrics, the customer's search behaviour predicts strong decisions about the purchasing decision. Also, research conducted by (Patil et al., 2018) provides a view on how Predictive Analytics can also for the identification of fraudulent transactions of credit cards and online net banking. An interface analytical framework has been provided in this study with Hadoop which reads the data effectively and informs analytical server for fraudulent transactions. The research also observed performance on big data by giving low risk and high customer satisfaction. Another example of customer satisfaction is explained in the research (He and Li, 2017), where a three-dimensional customer segmentation model is used based on CLV (Customer's Lifetime Value), Customer Activities. This research identified 10 groups of customers which hints the marketing strategies to maximize profit. On the other hand, the study of (Hwangbo et al., 2018) presented a recommendation system for e-

commerce fashion retailer. This study aimed at developing a recommendation system for a large Korean fashion company. The author used K-RecSys extending the item-based collaborative filtering algorithm combining online click data and offline sales data. The author concluded that show time and purchase intention are important in recommendation systems for fashion items. Also, the research conducted by (Zhao et al., 2015) highlights setting price discounts automatically when recommending a product keeping in mind that price can change customer's purchase intension. This study focused on optimising the discounts for identifying the potential customers and willingness to pay. The author suggested that Personalized promotion can be used to boost customer satisfaction as well as seller profit.

2.3 Promotional Strategies for Product Pricing using Predictive Analytics

Organisations conduct promotional activities and different other strategies are widely used for Pricing the products with the help of Predictive Analytics. According to (Peng et al., 2019), e-commerce companies like Amazon, Ru La La, Tmall, etc. uses flash sales for online business platforms with limited purchase time. This is one of the promotional techniques used by the retailer business firms which increases the revenue and generate maximum profit. To test the study hypotheses, the author collected data from wjx.com. PCA was used along with varimax rotation, SPSS, multiple regression. Price Value, Functional Value, Emotional Value and Social Value are the parameters used for customer segmentation. This research model fails to consider the social dissemination, personality, or brand effects. The diversity in consumers' choices and the type of products and product categories make shopping behaviour difficult to understand for large stores managers (Chen et al., 2021). The customized product recommendations and discounted prices enable these small-scale stores to have loyal customers (Heitz-Spahn, 2013). According to the author (Jiang et al., 2015), suitable discounts need to be provided for the products and recommendations should be provided for the non-discounted products. Using this efficient methodology, the sales can be recovered by recommending other non-discounted products. Three models are used by the author like OPR, PNR and MCR.

Moreover, the author (Gupta and Pathak, 2014) proposed a generic framework using machine learning models to improve right price purchase and not the cheapest on e-commerce platform. The author focused on inventory led e-commerce companies however, author further stated that the model can be extended to online marketplaces without inventories. Also, this paper proposed the adaptive pricing personalization and prediction of purchase intent as the future work of the study. On the other hand, the study conducted by (Peng et al., 2019) for the analysis

of the competitive dynamics and volatile demand for Airbnb and Uber. This study focused on how the organisations accommodates and respond the demand fluctuations. This study considers seasonal and counter-seasonal pricing, market conditions such as consumer compositions, supply electricity, taxes, etc. This research proposed that high-end hotels are quite sensitive to the changes in the Airbnb hosting costs.

Also, there are several areas where organisations use promotional activities for finishing up the remaining items in the inventories. An example of promotional activities using predictive analytics is the research by (Smith, 2015). This research analysed the policies used by different organisations like optimal clearance prices prediction, inventory management considering reduced assortments and seasonal changes. These policies have been implemented on major retail chains. On the other hand, several studies carried out on the effect of promotional activities and its impact on customer satisfaction. The research by (Pang et al., 2021) analyses Price Guarantee of the product where customers pre-orders new or to-be-released products. To encourage activities like pre-orders, retailers like Amazon uses promotional activities such as Price Guarantee. The author proposed strategies used in such promotional activities like advance selling, price matching and effective dynamic pricing. Also, in the research conducted by (Liu et al., 2021), participation in the strategic transformation in the e-commerce is analysed. This study proposed that the participation in strategic transformation boosted sales for merchants resulted in greater post-sales product and merchant selling seasonal products. Moreover, to identify the determinants of conversion rates for e-commerce sites, the research conducted by (di Fatta et al., 2018) proposes that to improve the conversion rates, either quality or promotional activities and avoiding mixing of both. This study conducted exploratory regression analysis for identification of possible determinants.

From the above literature review, it can be stated that organisation invest a lot in promotional activities to attract several customers and increase sales. Predictive analytics is used heavily with the help of data mining and machine learning algorithms to promote offers and discount on the products.

2.4 Data Analytics using Textual Information

The author (Ganame et al., 2017) proposed using of n-gram model to differentiate between fake and real news. For the identification of true and false news, the author proposed different sets of n-grams. Author used various features of the n-gram baseline established on words. Data pre-processing techniques like stemming and stop-words removal are applied. Term Frequency – Inverse Document Frequency (TFIDF) is used in this study for extracting textual features. The author used six Machine Learning algorithms: SGD, KNN, SVM, LSVM, and DT on 3 datasets available online. This study achieved an accuracy of 87.0% for the identification of fake news using n-gram and LSVM.

Moreover, the author (Eshan and Hasan, 2018) has implemented a machine learning model for abusive text detection in Bengali language. Different NLP techniques like Count Vectorizer, TF-IDF Vectorizer, n-grams are used for textual data conversion in numerical vectors. SVM, RF and Multinomial NB models were compared for the performance against each other. The study shows that features collected from TF-IDF Vectorizer were better as to Count Vectorizer while working with SVM.

The textual data from the documents can also be analysed for examining the keywords in corpus. The study conducted by (Qaiser and Ali, 2018) focused on how TF-IDF can be applied on number of documents. The author performed data processing on the data collected from websites. Strengths and weaknesses of TF-IDF is discussed. The author proposed TF-IDF works best for identification of keywords in multiple documents but also proposed that TF-IDF lacks to identify words with slight change in the tense and provides unexpected results. Also, it is unable to check the semantic of the text and analyses information only at lexical level. The author also proposed other algorithms like Decision Trees, Rule Based Classifiers, Neural Network Classifiers, SVM classifiers, etc.

Sentiment analysis can be performed on Twitter data for analysis of Frauds, Fake News, etc. The Author (Heidari et al., 2021) conducted an empirical study to analyse how social media bots can change society's perspective using machine learning methods. The author analysed the tweet's sentiment features and accuracy of machine learning models for social media bot detection. This study used Random Forest, SVM and Logistic Regression for classification, Neural Networks for sentiment analysis. However, this study proposed to explore more on feature engineering as well as tuning models for accurate sentiment analysis with the help of different machine learning approaches.

(Liu et al., 2018) conducted research for text classification based on improved TF-IDF algorithm. The author focused on automatic text classification using user's information from large amount of text. This paper addressed problems such as ignoring contextual semantic links and several vocabulary importance. The author used word2vec and improved TF-IDF algorithm. This study given less important for weight-based title inclusion as it provides high-level summary of the content. On the other hand, research by (Yahav et al., 2019) provides another view where data mining techniques applied on comments for user-generated content. The author used text pre-processing used followed by adjustments to the TF-IDF for solving the problem of bias that is created by higher correlation of the comments from social media. However, this study fails in terms of overweighting the bias and bias of adjustments itself. Moreover, the research by (Dadgar et al., 2016) focused on a novel text mining approach for news classification based on TF-IDF and SVM. The author aimed to classify news into different groups for the simplification of identification. The author performed text pre-processing, conducted feature extraction based on TF-IDF and classification using SVM. Although the results obtained on the dataset were fair enough, this study fails to consider synonyms and dependency between the words.

Stock price prediction is always considered as a challenging task. As per the study conducted by (Guo, 2020), market information is reflected by the current price and stock prices are affected instantly according to financial market. The main purpose of this study was to predict stock price considering the related articles from news websites. News headlines and textual data is analysed to perform textual processing and sentiment analysis which is combined the sentiment score using LSTM for prediction of closing price of future stocks and current return. The author compared the results from SVM, RF, RNN, MLP and LSTM and proposed the results showing LSTM performs better with smaller percentage error. Another research conducted by (Akita et al., 2016) proposed how newspaper articles affect the stock prices and deep learning can be used for the making the stock price prediction. The author proposed Paragraph Vectors (PV-DM), Bag of Words for Paragraph Vectors for textual information processing combined with numerical data using LSTM based approach and further Simple-RNN and SVR for the regression analysis. Moreover, the study of (Mohan et al., 2019) used deep learning techniques for stock price prediction using news sentiment analysis. This research used ARIMA, Facebook's Prophet and RNN/LSTM based approaches for processing textual information and sentiment analysis. The author termed this problem as hard because there are several challenges and assumptions need to consider. The models might not work well for low

or highly volatile stock prices. Also, author concluded that sentiment might not at all affect the stock prices based on news articles. These studies failed to include technical indices such as Moving Average and Convergence-Divergence for profit analysis.

For performing textual data analysis on different techniques of Natural Language Processing like Sentimental Analysis, different deep ensemble learning techniques received noticeable attention of researchers. According to research conducted by (Mohammadi and Shaverizade, 2021), different deep learning-based methods shown superb performance for sentiment analysis problem. The author proposed a novel approach for aspect-based sentiment analysis to extract opinions or attributes for topics and entities. This research was built on four different learning methods namely LSTM, BiLSTM, CNN and GRU. Also, outputs of these models are combined using stacking ensemble approach. Also, the research by (van Huynh et al., 2020) proposed deep neural networks models for determining suitable job for the person based on job descriptions. This research focused using deep neural networks like TextCNN, Bi-GRU-LSTM-CNN and Bi-GRU-CNN along with several pre-trained embeddings for IT job dataset. This study proposed to use LSTM variants for textual processing and improving performance.

On the other hand, a high complexity and low efficiency Chinese text sentiment analysis problem has been addressed by (Pan and Liang, 2020). The author extracted deep features from textual data and combined the sequence to learn features accurately. This research proposed Bi-GRU model based on Multi-Head and Self-Attention which assigns weight to word vectors, also focuses internal dependencies on sentences. The author proposed using attention mechanism instead of neural network to improve performance. Moreover, the research proposed by (Bansal et al., 2016) made use of news articles, plot summaries, blog posts, etc. for recommendation of new and unseen content. A deep recurrent neural networks model is leveraged to encode the text into a latent vector especially GRUs and collaborative filtering is applied for the recommendation purpose. The user proposed this model yields high accuracy and performance can be increased further using multi-task learning. However, this study proposed considering the suitable datasets with less sparsity for performance gain and thus improving high accuracy.

In addition, research conducted by (Zhou et al., 2019) proposed deep learning models of rental market dynamics on spatial and textual data from rental listings. Several machine learning and deep learning models were tested like Convolutional Neural Networks, Recurrent Neural Networks for price prediction. The author gained significant performance and accuracy over traditional textual processing algorithms and evaluated using performance metrics like RMSE

and MAE. However, performance metrics like RMLSE and MAPE are recommended penalizing the model. Moreover, (Mehtab and Sen, 2020) proposed a model for prediction of stock price using deep learning and natural language processing. For prediction of price movement, classification techniques are used whereas for predicting the closing price, various regression models have been used. LSTM based approach is used for predicting the stock closing price and the predictive model is augmented by integrating the sentiment analysis on data from twitter for correlation of public sentiment for stock and market prices. Also, the study conducted by (Li et al., 2019) proposed a textual data based crude oil price prediction using deep learning. The author used crude oil news headlines and price data, performed sentiment analysis and CNN based classification for feature extraction along with LDA topics to forecast the oil prices. The author used available news data for textual processing and sentiment analysis. This study proposed that textual features and financial features are complementary for producing more accurate results for oil price forecasting.

Moreover, prediction of price consisting of a textual data has also increased attention of researchers in the past few years. Research more relevant to this research for prediction of price with textual information is being conducted by (Han et al., 2020). This research proposed an intelligent system for prediction of price for second-hand items using both textual information and images. The author proposed a multi-model approach for prediction using textual and visual features along with other statistical items. This study used a binary classification model for identification of qualified items which can be used to predict the price. A customized loss function is used and for optimization of regression model for price prediction. This study used an indicator vector of length 32 to create a corpus of single words. Feature embedding operation is implemented for dimensionality reduction and values are updated to minimize the loss function. A softmax operation is applied on the weights along with processing visual information. This study proposed a more complicated and non-generalized approach with the help of loss function and optimization constraint were also domain specific. A deep understanding of mathematical equations for optimization and evaluation of model is required.

2.5 Data Analytics using Regression Techniques

The study conducted by (Naik et al., 2018), is analysed using Kernel Ridge Regression Model with EMD model for the prediction of Wind Speed and Power. The dataset referenced in this study is from Real Wind Farms. For each prediction at multiple intervals, the prediction error rate of KRR is less as compared to other models. It has been observed that Corelation

Conversion Factor and accuracy for the Ridge Regression is highest as compared to other models like RVFL and ELM.

On the other hand, the Author (Chiang et al., 2019) has implemented Ridge Regression for over Big Data to analyse the computation time, memory requirement, and the accuracy of the output. Referenced dataset for this study is from BTS and FAA and it's quite large (Split into RAM-accommodable subsets). As per the objective, to predict arrival and departure of flights. Proposed solution for regression was successful with a MSE of 168632.10 and MAE of 394368.89. The study concluded that Ridge Regression takes lesser memory, performs faster in terms of computational speed, and provided results are accurate which is a motivation to use Ridge Regression model for this work.

According to (Li et al., 2021), with large data and sufficient parameter tunning, it has been observed that the performance of Light Gradient Boosting Machine (LGBM) model is better than Random Forest model. This study aimed for high-accuracy price prediction model in the real estate investment market. Author used the price data of condominiums consisting of 63,093 records with 108 items. The author finally proposed the Price Prediction Model based on Light GBM with 8.349% Mean Average Percentage Error (MAPE) and high accuracy.

One of the studies related to Light Gradient Boosting Machines (LGBM) along with Convolutional Neural Networks (CNN) has implemented LGBM and CNN models for prediction of power of wind (Ju et al., 2019). In this study, for the sole purpose of feature extraction from the input data, CNN is applied followed with LGBM on the inputs. This study shows that RMSE observed for CNN based model is 2.315 and for LGBM model is 2.344. This research shows how CNN is good in fetching the data as compared to LGBM as it increases robustness of the model.

2.6 Dynamic Pricing Strategy: Surge Pricing

Many organisations like Uber, Amazon, Airbnb, etc. use the surge pricing where price predictions for the product are made based on season, time of the day, mobility in the market, demand, etc. According to the research conducted by (Zhang et al., 2017), dynamic pricing implemented by Uber, Lyft, etc have changed the demand-supply dynamics for fixed rate traditional market. The author created a review for understanding the underlying mechanism used by Uber in NYC. The research proposed how Surge Price Modifier works at different locations and Uber creates the categories of 27 classes and nominal fare, low surge, medium surge, high surge, and very high surge has been assigned prices for the rides. This research has used machine learning algorithms like Support Vector Machines (SVM), Discriminant Analysis

(DA), k-nearest neighbour (KNN), Probabilistic Classifiers like Logistic Regression and Naïve Bayes, Decision Trees (DT). This study analysed how and where the passengers are expected to pay the surged fare effectively. However, this study collected data based on specific time and over the time, the strategy used in the organisations updated rapidly. Also, the data considered should have been more generalized for better accuracy and proper predictions.

On the other hand, the author (Battifarano and Qian, 2019) also conducted a review for dynamic surge pricing for ride-sourcing companies like Uber, Lyft, etc. This research proposed how surge pricing predictions is useful information for drivers, riders, and companies on ride-hailing. This research explored the spatial-temporal characteristics between urban areas, traffic flow properties and surge multipliers. This study used a log-linear model with L1 regularization and coupled with pattern clustering. In this research, one key observation that might affect surge pricing models is the data collected from multiple sources and environments like traffics, weather, etc.

Also, the study conducted by (Mohamed et al., 2022) aimed for Price Prediction of Seasonal Items. This study used a dataset of a retailer who launched special sale for Christmas Event. In this study, Ridge Regression, SVM, RF and ARIMA are used evaluated against MSE, RMSE, R2 and MAPE. Also, this study proposed to consider hybrid machine learning ensemble models for improving the precision quality in the future. In addition, the research conducted by the author (Falode and Udomboso, 2021) proposed a machine learning model for efficient crude oil pricing strategy. The author analysed the implications of outbreak of Covid-19 for the entire value of crude oil industry. This study proposed a model for forecasting the prices of oils to minimise the risks associated with volatility. The author used a hybrid model with classical and machine learning techniques like autoregressive neural network. This research proposed the hybrid model with 20 to 100 neurons as the best performing model. However, this study fails to identify the extreme potential threats and seemingly unending recession with the help of hybrid model. In addition, (Fiat et al., 2018) performed a statistical analysis review for maximizing the social welfare that depends on taxicab and passengers' locations, valuations, distances. The author proposed two related models viz. a continuous model like War drop model and a discrete model for passenger-taxicab settings. The author proposed to compute the surge prices which maximizes the passenger-taxicab equilibrium social welfare. Also, the author proposed model capable for surge price computation in polynomial time.

In addition, a lot of research studies has been performed on promotional strategies for products as well as price optimization strategies for products which comes under surge pricing

techniques. (Greenstein-Messica and Rokach, 2020) proposed a machine learning based price optimization strategy for products having no price elasticity history. The author proposed how organisations can use a consistent pricing strategy for increasing customer trust and promoting just a small portion of catalogues each week. The author proposed a model using log-log demand model combined with non-linear gradient boosting machines for prediction of price elasticity impact for products with no elasticity information. Also, the research conducted by (Yan et al., 2020) analyses demand of ride-sourcing services and proposed machine learning based approach for pricing of ride-sourcing services. This research aims on accurately predicting the demand for efficient land-use, transportation, and policymaking. The author used trip level ride-sourcing data in Chicago. This research used Random Forest to estimate the zone-zone demand for services. The author proposed that Random Forest model has better fit over other traditional multiplicative models. It has been analysed socioeconomic and demographic variables contributed towards achieving higher prediction accuracy.

Moreover, several studies proposed machine learning models for prediction of product prices, demand and sales based on seasonal conditions. A comparative study conducted by (Ensafi et al., 2022) analysed how machine learning and neural networks can be used for prediction of sales for seasonal items. This research used a public dataset including sales history of a retail store and investigated the forecasting models for prediction of sales in future. This study analysed performance of Seasonal Autoregressive Integrated Moving Average (SARIMA), Triple Exponential Smoothing, Prophet, LSTM, CNN based models. The stacked LSTM performed better over other models. However, one of the limitations of this study was the skewness of the public dataset and choice of models were not in conjunction. On the other hand, the authors (Saha et al., 2021) conducted a survey on efficiency of app-based cab booking system with statistical analysis. This study analysed fleet utilization, recording of trips and bargaining facilities along with technologies like distributive algorithm, adaptive scheduling algorithm. Also, (Arismendi et al., 2016) proposed effect of seasonal volatility and effect on pricing for commodity. Although, this study proposed a statistical analysis for seasonal volatility for predicting price accuracy, the author analysed the competitiveness affects the pricing and several parameters to be taken in to account. In the next section, a literature review on dynamic competitive pricing strategies will be reviewed.

2.7 Dynamic Pricing Strategy: Competitive Pricing

Many organisations use competitors' data whenever available for assigning the prices to the products. A lot of research has been conducted on implementing the pricing strategy analysing the prices set by competitors. (Lin et al., 2020) investigated pricing and product-bundling strategies in e-commerce domain. Mixed-bundling strategies are also implemented in some of the platforms for better analysis of the pricing and revenue growth. (Heuer et al., 2015) conducted a review brand competition in fashion e-commerce. A unique dataset with more than 3.3 million observations provided by e-commerce company was used to estimate cross-price elasticities. This study shows how distinctiveness in fashion merchandise prohibits customers to take advantage with increased transparency in market e-commerce. The author also proposed asymmetric competition also exists between private and national brands. This study proposed markdown pricing strategies in the context of e-commerce. (Herliana et al., 2019) conducted a review on how customer loyalty in E-commerce competitive market for maintaining the customer satisfaction and the increase in affection for the product and service of the brand. With the increasing amount of competition, consistent implementation of the strategy considering several variables in e-commerce for customer loyalty is bound for building, developing and maintaining the customers in fashion e-commerce.

According to research conducted by (Bakir et al., 2018), the possibility to adjust the prices considering competitors' product prices is like the prediction of stock prices with adequate data. This paper proposed a forecasting time series model for prediction of phone prices in European market. This study analysed a comparison of machine learning models like LSTM, SVR, RNN, SVM, etc. Also, LSTM and RNN has given more accurate prediction for forecasting the next day's prices. This study fails in analysing the external time series data resulting in lowest RMSE. On the other hand, (Rai et al., 2019) conducted a comparative study for demand prediction in e-commerce using SOTA machine learning models. This study analysed effects of advertisements for determining the demand of the products for C2C e-commerce companies. Using the product description, images and context, deal probability of products is analysed. This research used SOTA models like ANN, Decision Tree ensembles, deep learning methods. However, this research fails in analysing the competing products, lost sales, feedback of customers for better result. Also, the author proposed the use of hybrid ensemble techniques for the demand prediction.

In addition, a several research has been conducted on dynamic competitive pricing using reinforcement learning considering several factors. Research conducted by (Kastius and

Schlosser, 2022) explains dynamic pricing considering competition using reinforcement learning. The author considered fixed price, undercutting price and two-bound strategies for deterministic approach. This research analysed performance of Deep Q-Networks (DQN) and Soft Actor Critic (SAC) for different market models like deterministic and non-deterministic competitor. This research proposed dynamic programming for optimal solution for tractable duopoly settings. Dimensionality can be a huge factor for oligopoly settings. This research stated SAC performs better than DQN and proposed reinforcement learning can be forced into collusion by most of the competitors without direct communication. On the other hand, (Lu et al., 2018) proposed a deep reinforcement learning based demand response algorithm for smart grid systems in energy management. This study formulated the problem statement as Markov Decision Process with Q-Learning being adopted for solving this problem statement. However, this study fails in optimising weighting factor between service provider's profit and customer's costs.

Moreover, (Yin and Han, 2021) proposed another deep reinforcement learning based approach for dynamic pricing strategy for E-Commerce platforms. This study analysed real-world dynamic pricing problems as Markov Decision Process. With the help of historical sales data and previous pricing actions, a pricing agent can be defined which interacts with the market environment and based on reward function, defined actions can be analysed for optimal pricing. This research extended the discrete pricing models to continuous pricing action space. Unknown demand function is also addressed using different reward functions. This study conducted field experiments on real-world data lasting for months and proposed the revenue conversion reward functions over revenue only and continuous action model over discrete. On the other hand, (Maestre et al., 2018) proposed a reinforcement learning based approach for fair dynamic pricing and maintaining the optimised revenue. This study proposed how wrong decisions may result in long-term revenue losses for the company. Also, this research was more focused on fairness between dynamic pricing strategy and revenue optimisation techniques. A complex environment is also considered to adopt the pricing policy and a trade-off between short-term and long-term goals of the organisation. A state-action-reward based reinforcement learning is used to analyse the fairness in dynamic pricing with revenue optimisation into consideration. This paper proposed reinforcement learning approaches takes several months for analysis of the action and reward-based approaches which can be optimised up to certain extent. On the other hand, (Ferreira et al., 2016) targeting Demand Forecasting and Price Optimization for Rue La La, shows pricing decisions taken by online retailers for monitoring wealth of the

organization. This study provided an approach using machine learning techniques for the prediction of future demand while identifying the gap of existing approach of Ru La La's pricing decision support tool. This study proposed a two-fold approach developing a demand prediction model and using this model as input into a price optimization model which resulted in increase in revenue. A regression-based approach for predicting the future demand and a LP Bound Theorem based statistical approach for price optimisations of product is considered. This study suggested future work in improving the overfitting problems and exploring less structured demand prediction models.

Analysing the literature review in the most popular pricing strategies like surge pricing and competitive pricing, the next section will focus on literature review for price optimization techniques and literature related to the problem statement.

2.8 Price Optimisation and Related Research Publications

Although the dataset used for this study is not used by other researchers, there is a lot of research on price optimisation and dynamic pricing already conducted. Research relevant to this study is the study conducted by (Kedia et al., 2020). This study proposed a price optimization framework for e-commerce fashion retailer called Myntra. This study proposed machine learning and optimisation techniques for optimal pricing for individual products. The study focused on demand prediction for the next day using sales data for past 7 days to maximize the revenue. To predict the acceptable change in price, a machine learning approach for prediction of price elasticity is proposed. For the final prediction of optimal price, a Linear Programming based approach is used. Also, this study compared performances of machine learning approaches like Linear Regression, Random Forest Regressor, XGBoost, MLP, ARIMA, LSTM, RNN for demand and price elasticity prediction. However, this study only made optimization on available parameters like clickstream data, product catalogues, price data, etc. A more research can be performed by analysis of market strategy environment like Competitor's Prices, Acceptable Ranges for price elasticity, etc.

Also, (Qu et al., 2017) performed analysis on price optimization for semi-luxury supermarkets with demand prediction. This study proposed a two-fold approach for formulation of pricing policies for semi-luxury supermarkets. The sales data from past 2.5 years retail stores has been used. This study proposed a Regression Trees and Random Forest based machine learning approach for weekly demand prediction. For price optimizations, branch-bound and branch-cut approaches were used which was further optimised by heuristic methods. The demand

prediction incorporated holidays, discounts, inventory, and regional factors for demand prediction. However, a more detailed analysis on price optimisation using statistical techniques like dynamic or linear programming could have been created better results.

However, price experimentation with dynamic pricing is always a crucial task for retailers as change in prices can affect the customer mentality and thus affects the revenue. (Cheung et al., 2014) performed research on dynamic pricing and demand prediction where experimentations were limited. As demand function was not known a priori, this research proposed price experimentation for demand learning. This research proposed the seller to make maximum M price changes for T time periods. This strategy incorporated O(log^(M)T)) of the algorithm. This research proposed a significant impact on revenue and bookings. On the other hand, the author (Besbes and Zeevi, 2009) addressed the same problem for adjusting prices in the finite sales timespan for revenue maximization. This study proposed to learn the demand function and optimised the prices based on learnt demand function. This study considered admissible business loss as lower bounds however which is not the optimal solution as organisations might not afford business losses even for specific timespan. In addition, (Ban and Keskin, 2021) proposed a model which can dynamically adjust the prices of the individual products by utilizing the information about customer's characteristics. This study predicted demand by learning the characteristics and observation of sales over T periods. This study proposed admissible policy of order s*sqrt(T). This study proposed the pricing strategy to be expecting revenue growth through experimentations. However, the time span for experimentation of six months and actual observations of revenue loss might lead to significant loss to the organisation which might not be improved over a specific period.

Also, (Ferreira et al., 2016) conducted research for Demand Forecasting and Price Optimization for e-commerce organisation named 'Rue La La', proposes pricing decisions taken by online retailers for monitoring wealth of the organization. This author proposed an approach using machine learning techniques for the prediction of future demand while identifying the gap of existing approach of Ru La La's pricing decision support tool. This study proposed a two-fold approach developing a demand prediction model and using this model as input into a price optimization model which resulted in increase in revenue. A regression-based approach for predicting the future demand and a LP Bound Theorem based statistical approach for price optimisations of product is considered. This study suggested future work in improving the overfitting problems and exploring less structured demand prediction models.

In addition, (Ban and Keskin, 2021)0 proposed a markdown and bundling pricing decisions. The author analysed three selling strategies for such as Markdown Pricing, Second Period Bundling and First Period Bundling. This research also proposed effects of strategic behaviour of customers on the pricing strategies. On the other hand, (Babar et al., 2015) proposed a demand elasticity prediction model in retail market. The author proposed demand elasticity problem as Markov Decision Problem and implemented machine learning techniques such as Q-learning for predicting the price. This study evaluated the demand elasticity which can be further extended by other techniques like Dynamic Pricing for predicting the prices of the product based on demand learning. Also, (Minga et al., 2003) proposed pricing algorithm based on bargaining in e-commerce. A demand sensitive model is analysed for price elasticity of demand changes per quantity, unit and gross margin. However, the key objective from this study was increase in demand decreases prices per unit of a good and increases the profit of margin.

Moreover, the author (Schlosser and Boissier, 2018) proposed a data driven approach for dynamic pricing considering competition. This study analysed the sales from behaviour of customers to the price reaction strategies. Also, state-of-the-art machine learning models for prediction of sales probabilities. A dynamic programming for computation of pricing strategies is also proposed based on number of competitors and unknown strategies analysis. In addition, (Singh and Dutta, 2015) analysed the dynamic price prediction techniques for amazon spot instances. The pricing is performed based on demand and supply of cloud resources across the world. The major challenge tackled in this study is predicting the spot price before placing the bids. A statistical dynamic programming approach is proposed which resulted in high accuracy for short-term and long-term forecasting. These studies contributed towards statistical analysis of pricing resulting in the revenue growth for organisations.

(Ye et al., 2018) proposed a customized model for dynamic pricing for Airbnb. This research proposed a price optimisation technique for helping hosts sharing their homes to set the optimal price. This research proposed a binary classification model for predicting booking probability, a regression model for optimal pricing using customised loss function, and additional personalization logic on top of that. This platform specific dynamic pricing made use of traditional supervised machine learning algorithms along with some customization in the loss functions. On the other hand, (Raju et al., 2003) proposed a RL based dynamic pricing for retail markets. This study used Q-learning and actor-critic algorithms for simulations. This is one of the promising ways in setting dynamic prices for multi-agent environments.

2.9 Discussion

This literature review finds it very interesting that a lot of research has already been conducted on Price Prediction for the Products in e-commerce and fashion retails along with different domains and integrated models with several other techniques like Dynamic Pricing, Price Optimisations, etc. Also, multiple organizations conduct experiments of different kinds of pricing the products which can be adopted effectively and with less real-world testing efforts. However, the explored areas and existing literature depicts the algorithms and has a lot of testing efforts before-handTable *1*: Summary of Literature Review along with Gaps. Table *1* gives a summary of the previous research along with their conclusions and few gaps. This table summarize an overview of the related research for textual information processing along with price optimization techniques. It's very clear from the literature review that there are several challenges like loss of information, misidentification, etc. for textual data processing. However, the price optimization algorithms referred mostly are either complex and inefficient with respect to time or requires a lot of data for generating insights.

One key challenge (Han et al., 2020) addressed is prediction of prices from textual and images, however the model is designed for the combination of textual and image information extraction. This motivates researchers to explore more on model tuning with considering other SOTA models like RNN, LSTM and GRU based models for accurately analysing the effect of textual information. However, the proposed research states null hypothesis as price can be predicted from the product's description extracting features from textual information. This research will focus to extend the existing SOTA models for price prediction using textual information. Also, it can be extended further to optimize and predict the efficient price dynamically. On the other hand, there is already a lot of research on supervised linear regression algorithms and ensembles techniques. Hence, this study will only extend the regression techniques for the prediction of discount.

Moreover, for optimising the product prices, several studies like (Yin and Han, 2021), (Ferreira et al., 2016), (Kedia et al., 2020) use reinforcement learning based techniques, Linear and Dynamic Programming based optimisations for dynamic pricing of the product. However, this research finds reinforcement learning based approaches as time consuming and requires access to a lot of data and real-world experimentations. On the other hand, the linear/dynamic programming-based approaches doesn't consider revolving market situations and competitive environment for making predictions about prices. Hence, this study finds that there is more

scope in improving existing approaches for price optimisations and will extend techniques for optimisation of product prices in competitive environment with available data.

Table 1: Summary of Literature Review along with Gaps.

Author	**Techniques Used**	**Challenges Faced**	**Remarks**
(Ganame et al., 2017)	TF-IDF, SGD, KNN, SVM, LSVM	The news data was quite large to process using TF-IDF.	87% accuracy using n-gram and LSVM.
(Eshan and Hasan, 2018)	GloVe, Count Vectorizer and TF-IDF	The data was in Bengali. Abusive text identification was challenging as the data couldn't generate correct vectors.	GloVe and TF-IDF was used along with CountVectorizer.
(Liu et al., 2018)	Text classification using Word2Vec and SVM	Dependency between the adjacent words and synonyms were lost.	Less importance for weigh-based title inclusion.
(Akita et al., 2016)	Textual Information and price prediction using Bag of words, LSTM, RNN, SVR	The dataset used was noisy and sparsy. Couldn't generate more information from Textual Information.	LSTM performed better with only slight margin of errors.
(Ju et al., 2019)	Wind power prediction using combined CNN and LGBM.	Lack of feature extraction.	CNN is good at fetching the metrics compared to LGBM.

(Ferreira et al., 2016)	Demand prediction and price optimization using RF and LP Bound Theorem	The algorithm used for price optimization was inefficient due to complexity.	A two-fold approach combined with Demand Prediction and Price Optimization was used.
(Kedia et al., 2020)	Price optimization framework for individual products.	RF, ARIMA, XGBoost, LSTM, RNN	The linear programming approach with linear relaxation optimization helped in generating more revenue.
(Han et al., 2019)	Price prediction for second-hand items using customized loss function.	Exponential complexity mathematical equations for optimizations.	A combination of Image and Textual Information using CNN performed better.

2.10 Summary

This section explored existing literature how organisations make use of predictive analytics techniques using machine learning algorithms and target business growth. A deep overview of existing literatures for textual information extraction is explored. This research identified some of the gaps in the existing studies and aims to explore more on SOTA models like RNN, LSTM and GRU. Also, existing literature about dynamic pricing techniques like surge and dynamic pricing, price optimisations for the product and related literature was reviewed. This study also identified some gaps and proposed area of research for price optimisations.

In the next section, methodology used for this research is explained in detail along with architecture, machine learning models and model evaluation techniques is also explained.

CHAPTER 3: RESEARCH METHODOLOGY

In this section, a detailed description of dataset that will be used for this study is provided. Further steps include key processes such as data pre-processing, transforming and feature engineering, proposed model and approach, machine learning techniques and comparing performances using model evaluation rubrics. For the second-stage, detailed description for evaluation of Dynamic Pricing will be provided.

3.1 Dataset Description

The dataset that is used in this study is related to Women's Innerwear and Swimwear products. The data is extracted from the popular retail sites via PromptCloud's data extraction solutions from June 2017 to July 2017 and is available publicly on Kaggle (Innerwear Data from Victoria's Secret and Others | Kaggle, 2017). This dataset has 600,000+ innerwear products information from popular retail sites such as Victoria's Secret, Amazon, Calvin Klein, Hanky Panky, etc. The overall size of the dataset is 530.25MB.

As this dataset has multiple files, the data from other retail sites can be used as the competitor's price. For the first stage of the machine learning, this study will mainly focus on the data from Victoria's Secret as this dataset has more information as well as variety of products. The datafile of Victoria's Secret is of 409 MB and has 453386 rows in total. The data is in CSV (Comma Separated Values) format. This file has 14 columns in total. The description of the relevant columns is provided below:

- **product_name:** Name of the Product
- **mrp:** Standard MRP of the product. This is the target variable.
- **price:** Discounted Price for the product.
- **product_category:** The category to which the product belongs.
- **description:** The detailed description of the product.
- **total_sizes:** The total sizes for the product.
- **available_size**: The available sizes in the inventory.
- **color**: The color of the prodct.

Victoria's Secret

product_name	mrp	price	brand_name	product_category	description	total_sizes	available_size	color
Body by Victoria Pe... 4% Victoria Sport Incre... 3% Other (423190) 93%	$10.50 11% $36.50 10% Other (358791) 79%	$10.50 10% $36.50 7% Other (374189) 83%	Victoria's Secret 76% Victoria's Secret Pi... 24%	Demi Bra 9% Push-Up Bra 8% Other (377069) 83%	Our fullest coverage... 4% An everyday fave wi... 3% Other (419667) 93%	["XS", "S", "M", "L", ... 19% ["32A", "32B", "32C... 12% Other (312196) 69%	S 7% XS 7% Other (392739) 87%	Black black Other (422778)
Very Sexy Strappy Lace Thong Panty	$14.50	$14.50	Victoria's Secret	Strappy Lace Thong Panty	Lots of cheek peek, pretty lace, a strappy back–this sexy panty is so not subtle. Allover lace with ...	["XS", "S", "M", "L", "XL"]	S	peach melba
Very Sexy Strappy Lace Thong Panty	$14.50	$14.50	Victoria's Secret	Strappy Lace Thong Panty	Lots of cheek peek, pretty lace, a strappy back–this sexy panty is so not subtle. Allover lace with ...	["XS", "S", "M", "L", "XL"]	S	black
Very Sexy Strappy Lace Thong Panty	$14.50	$14.50	Victoria's Secret	Strappy Lace Thong Panty	Lots of cheek peek, pretty lace, a strappy back–this sexy panty is so not subtle. Allover lace with ...	["XS", "S", "M", "L", "XL"]	S	plum dust
Very Sexy Strappy Lace Thong Panty	$14.50	$14.50	Victoria's Secret	Strappy Lace Thong Panty	Lots of cheek peek, pretty lace, a strappy back–this sexy panty is so not subtle. Allover lace with ...	["XS", "S", "M", "L", "XL"]	S	ensign blue

Figure 3: Overview of Dataset used in this study

3.2 Proposed Design Architecture

The proposed design architecture for this study is as follows:

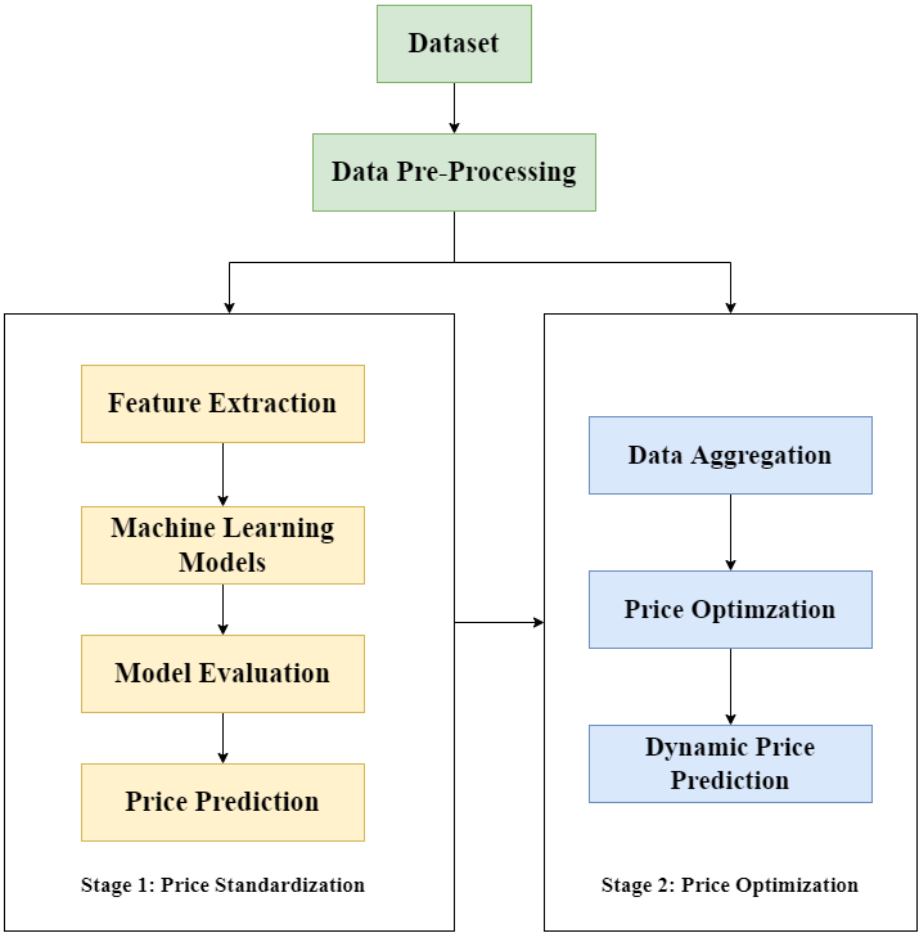


Figure 4: Proposed Design Architecture and Process Flow Diagram

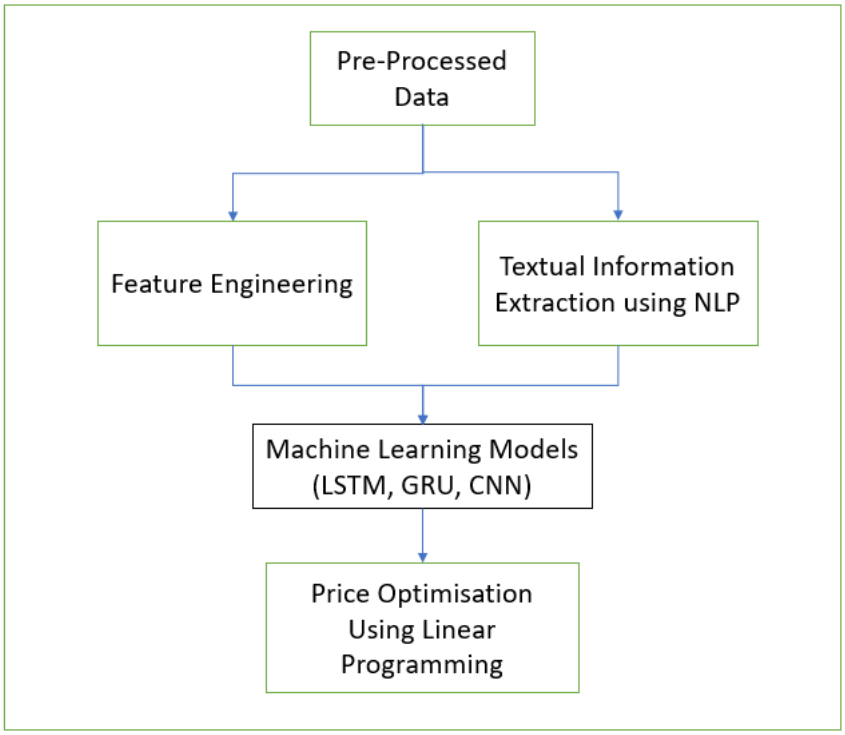


Figure 5: Proposed Architecture and Process Flow

The proposed design architecture that is used in this study is depicted in Figure *4*. A data pre-processing is required before we feed the data for the machine learning approaches. Considering the sparsity of the data, feature extraction will be the crucial step for derived attributes and dimensionality reduction. In the first phase referred as Price Standardization, machine learning models like LSTM, GRU and CNN are used to predict the standard price of the data from the characteristics as well as from the description based on the objective. In the second phase referred as Price Optimization, data aggregation is performed with the existing data with derived attributes and the data from other online retailers' prices used as competitor's data. Data is aggregated with the competitor's data using 'pandas' library. According to the goal of this study to predict the optimal price which will be contributed to optimize revenue, this is a standard optimisation problem whereas optimization function is generated along with the constraints. In addition, a statistically proven linear programming-based approach is used in this study for price optimizations. Organizations can use the pricing strategy conducted by this study for revenue and profit growth. The Figure *5* gives an overview of the steps that are performed in this study including machine learning models and price optimization strategy.

3.3 Data Pre-Processing

Data pre-processing is an important step for any machine learning based approaches as the data needs to be brought in the required format. Any inconsistencies in the data can lead to undesirable and unreliable results. Data Pre-processing is an essential step to bring the data in the required format. Data cleaning and pre-processing is used to remove the noise, missing values, etc. The columns like mrp, price, available and total sizes are brought in the appropriate data types. Imputation techniques are used for the columns having NA or NULL values. The missing data is replaced with mean or mode for numerical and "other" for the textual or categorical columns or even a decision to remove such data might be taken accordingly. Also, there are few columns like product_category which are categorical, thus needs to be encoded into numerical columns. For this, encoding techniques like Label Encoding is performed. In case the categorical columns have too many values, Binning is used to reduce the dimensionality of the categorical variables. After data cleaning and data pre-processing, feature extraction techniques will be used to bring the data into required format.

3.4 Feature Extraction and Engineering

Feature extraction is the crucial step for this study as the data is in raw format and the information needs to be preserved. Feature Extraction refers to the process of transforming raw data into the required numerical features which can be processed and simultaneously preserving the information in the original data. It has been observed that feature extraction yields better results instead of applying machine learning techniques directly. The feature extraction can be performed manually as well as using automation (with machine learning approaches). This study aims to extract manual and automated feature extraction based on the scope of the data.

Manual feature extraction helps in capturing the relevant information based on the background and domain understanding. For the dataset used in this study, the manual feature extraction technique is performed to derive the useful information like %discount offered, relative competing prices, color popularity, size popularity. % Discount is evaluated using the formula (1 – price/mrp). Size weight and size index is calculated as percentage of the total sizes or weighted mean percentage of the available styles. Relative competing prices the price that has been used by the competitors. It is calculated as the average of the average price (considering their mrp and discount) by different online retailers. Popularity index calculation is a challenge that can be calculated statistically using the data of revenue growth in the past years or using relevant statistical approach by considering the demand and supply of the organization along

with the revenue growth in the past years. Popularity Index is meant to capture how the demand changes with the price (Ferreira et al., 2016). For this study, popularity indexes are calculated using available information. The variation of price against the index column is observed and the percentage popularity indexes for size and color can be evaluated.

Automated features extraction uses specific algorithms or deep networks to extract required information automatically from the data. The dataset used for the study has multiple textual columns like description, product category, etc. Natural Language Processing techniques are required to convert the raw textual data into required format. Tokenizing, Punctuation Removal, Digits Removal, Stop-Words Removal, Digits Removal, Stemming, Case Conversion are used for processing the data. The processed data needs to be converted into the normalized format. This normalized text will be converted into numerical vectors. As per the study conducted by (Dzisevic and Sesok, 2019), GloVe embeddings yields better results and hence Glove Embedding is applied to product_name and product_category columns. Also, textual information processing as recommended by (Han et al., 2020) can also be targeted with CNN, LSTM and GRU based approaches is considered. This converted data is fed to machine learning models to yield better results.

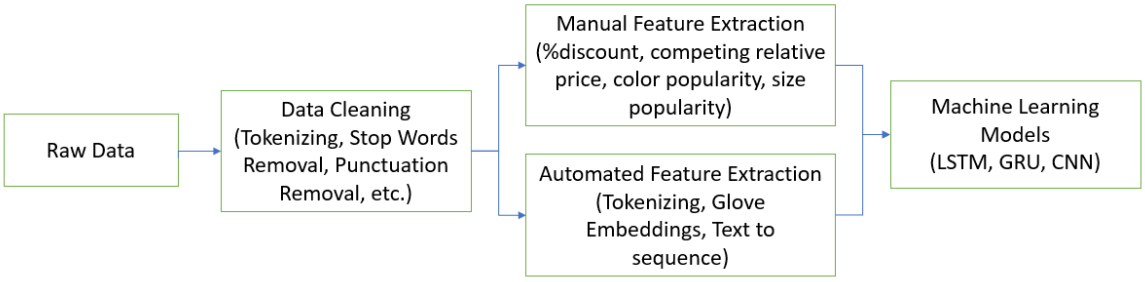


Figure 6: Process flow for Natural Language Processing Tasks

Glove Embeddings:

GloVe aka Global Vector is based on an unsupervised machine learning algorithm is used usually to obtain vector representations for words. On aggregated global word to word co-occurrence statistics, training is performed on a corpus, and the resulting representations provides insightful linear substructures for the word vector space (Pennington et al., 2014). Glove efficiently identifies the statistical information using training only on the non-zero elements in a word-to-word co-occurrence matrix, rather than using on individual context windows or on the entire sparse matrix in a large corpus. Glove uses vector representation for similar words in a corpus having the same meaning and is used widely in earlier researchers

with effective accuracies in NLP models. This study used Glove Embeddings for large product descriptions textual data for embeddings.

Convolutional Neural Networks (CNN):

Convolutional Neural Networks (CNN) are the most promising choice of neural networks in developing machine learning models. It performs very well specially in image classification and computer vision. Convolutional neural network is a special tool for modelling as it consists of features like detecting edges, corners, and various textures. It goes through every corner, vector, and dimension of the pixel matrix. Recently, the Convolutional Neural Network (CNN) has been adopted for the task of text classification and has shown quite successful results (Kim and Jeong, 2019). Using the information of the pixel matrix, CNN is quite sustainable to for data of matrix form. As description column is in the textual format, considering CNN for a textual layer is the similar idea of working with image. Textual data can be converted into a sequential data like the data in time series and interpreted as 1-D matrix. CNN can work with 1-Dimensional matrix convolutional layer and the modified data type. The key objectives to propose CNN as a machine learning algorithm for this study is as follows:

1. Even though CNN is usually used for image modelling, it can also learn from connection between words leading to a better choice for NLP based datasets.
2. It provided quite accurate results working with the data in matrix format and convolutional layers.
3. CNN works very well on high dimensional data there by providing good results and performance efficiency even of the textual dataset.

Long Short-Term Memory (LSTM):

Even though almost all the sequence prediction problems use LSTM based models for time series data, the power of LSTM allows learning and processing textual information. LSTM always has an edge over Feed-Forward Neural Networks and RNNs and performs better. The main reason behind the effectiveness is it stores learnt information for long durations of time and it selectively remembers the patterns. LSTM is a special kind of RNN and having capabilities like dependency learning for long-term of the data. LSTM has four layers interactively connected with each other which helps for long-term memory and providing better results. The information about the gates is as follows:

1. Learn Gate: Event and Short-Term Memory combined so that necessary information can be applied to the current input.
2. Forget Gate: This gate is useful for forgetting the information which is not useful.
3. Remember Gate: Long-Term Memory information, which is retained, Short-Term Memory and Events are combined in remember gate which updates LTM.
4. Use Gate: For predicting the output of the STM, the current even output of LTM, STM and event is used.

The key reason behind using the LSTM for this study is:

- LSTM stores the information about the states which gives a better accuracy over other sequence to sequence algorithms like RNN.
- With the use of states to remember or forget the information, LSTM always has an edge to make small modifications on the information.

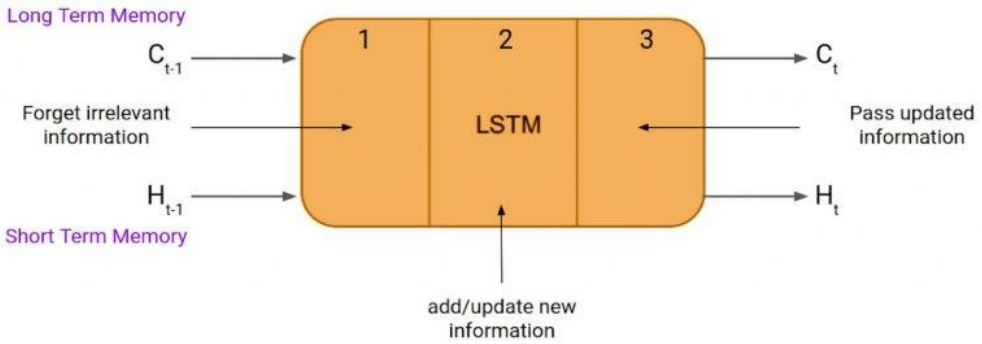


Figure 7: LSTM Architecture (ImageSource: https://google.com)

Gated Recurrent Unit (GRU):

Gated Recurrent Unit or GRU is an advancement over LSTM and standard Recurrent Neural Networks i.e., RNN. GRUs are quite similar as compared to Long Short-Term Memory. GRU also makes use of gates for controlling the flow of information. GRUs are relatively new in comparison with LSTM and almost have similar architectures.

GRU only has hidden state, and it doesn't have separate cell state. Due to this reason, GRUs are faster to train.

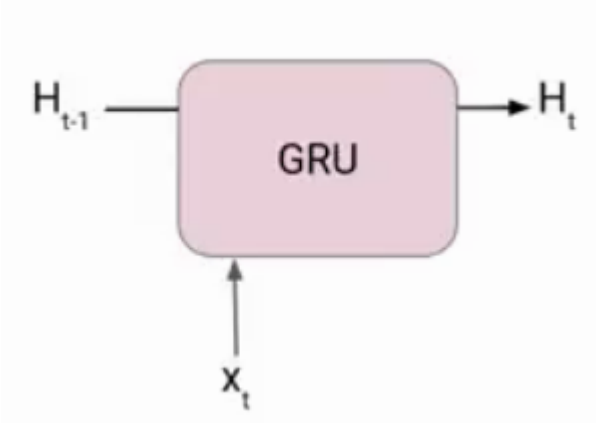


Figure 8: Gated Recurrent Unit Architecture. (ImageSource: https://google.com)

There are two gates in a GRUs as follows:

1. Reset Gate: Reset Gate primarily responsible for STM i.e., hidden state of the network.
2. Update Gate: Update Gate primarily responsible for LTM of the network.

The key reason to consider GRU for this study is:

1. GRU is also used to process textual information.
2. GRU has only two gates as compared to LSTM and works quite faster.

3.5 Exploratory Data Analysis

After data pre-processing and feature extraction, performing exploratory data analysis is essential to understand the existing behaviour of the data using different combinations. Univariate, Bivariate and Multivariate analysis is performed for tuning the data to understand the behaviour like Distribution Information, Outliers Detection, Multi Collinearity to test the hypothesis. Data is visualized using libraries like Matplotlib, Seaborn, etc.

3.6 Train-Test Split

Before passing the data to machine learning models, the data should be divided into train test split to evaluate the performance against the test model. The data of other online retailers available can also be used in the study. K-fold Cross Validation techniques is used with 10 folds to estimate the skill of the data on unseen data.

3.7 Model Evaluation Metrics

The model is evaluated against model evaluation metrics like RMSLE, MAE and RMSE.

RMSLE is the calculated as the logarithmic root mean squared difference of the predicted and actual value. It's quite like RMSE except for the case, it incurs larger penalty for the underestimation of the actual value than overestimation. As per the business case, predicting lesser value is less tolerable than the predicted the higher value as pretty much lesser value can lead to the downfall of the sales revenue and predicting very higher value can lead to the downfall of the profit. This makes RMSLE as the best choice for the model evaluation metric. It can be calculated as follows:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

Figure 9: RMSLE Formula (Image Source: https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-RMSLE-935c6cc1802a)

R-Square or R2 determines how good the model fits the dependent variables. It's evaluated unexplained variance divided by total variance. However, R2 the overfitting problem is not taken into consideration. This is introduced in Adjusted R2 because of its penalizing nature of additional independent variables.

MAE is quite equivalent to MSE with just the difference that takes sum of absolute value of error. MAE is the direct representation of the sum of error terms and treats all the error terms the same. MAE is used as an evaluation metric for this study to understand the performances of the model over all the available datafiles.

3.9 Stage-Two: Dynamic Price Prediction:

After prediction of standardized price in the first stage, this predicted price is used for the statistical calculation of the dynamic price. Data Aggregation is the first stage to be performed to have the data from other online retailers processed together.

The primary aim to use dynamic price prediction is to predict a suitable price range which the organizations can use to provide efficient discounts keeping the standardized price in mind.

After the dynamic price range calculation based on data aggregated from Competitor's data, the next step will be price optimisation. (Kedia et al., 2020) used a demand prediction model for demand for next day and a linear programming model was used for price optimisation. Also, (Ferreira et al., 2016) used a regression tree-based model for demand prediction and a LP Bound Theorem was used for price optimisation. Considering the scope of this study, a demand

prediction model is not used directly for analysing the demands directly, however, this study aims only for price optimisation based on competitive pricing, hence the algorithms used by this research will be extended as below:

The price range calculated from the Competitor's data can be used for the price optimisation. The discount provided by the organisation from the data can also be used for the price optimisation. With the help of Competitor's Price and Discount, as the key goal will be to maximize the revenue, the primary aim to keep the price as small as possible based on Competitor's Price. However, it won't always be less as this will turned out to be a cyclic dependency problem discussed by (Bauer and Jannach, 2018). Standardized Price calculated from stage1 of the architecture will also be considered. Also, the discount should be kept as small as possible, which will help in over discounting the products and keeping the organisation in profit. From 3 and 4, it turns out that a minimization of Competitor's Prices and Discount can give optimal results for pricing the products. Hence, a Linear Programming model is used which will minimize the optimization function and provided optimal and efficient solution to the problem.

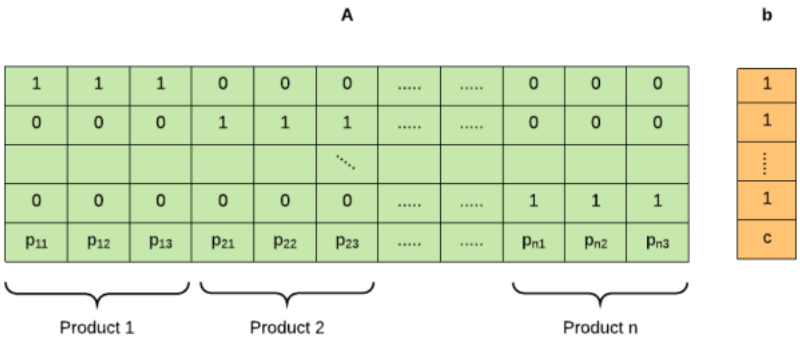


Figure 10: Linear Programming Optimization for price and demand. (Kedia et al., 2020)

The Figure *10* depicts the linear programming approach that will be used in this study which is extended from the research done by (Kedia et al., 2020). For each product, there will be three combinations of the prices which will be based on the competitor's price also. And for each price, there will be a demand which identifies the effectiveness of the price. As the demand increases, the price of the product should be modified accordingly to attract more customers and the overall revenue should also be optimized. For each product, each of the price-demand combination will be either selected or not selected, and the optimal combination will be analysed with the help of linear programming.

The objective function is to maximize the product of price and demand for each product to generate optimal revenue. The variable p denotes the price of the product, and the variable d

denotes the demand of the product. The variable x will contain the values of 0 and 1 which will be helpful in identifying which price-demand combination is selected out of three. This technique is called linear relaxation and reduce the time complexity from 3^N to linear. The constraints are provided in the equation, which explains the ranges of values of the x as well as the revenue generated will be the combination of the selected prices amongst the available price between standardized price and the competitor's price. This study used scipy's linprog library for linear optimization.

$$\boldsymbol{maximize} \sum_{k=1}^{3} \sum_{i=1}^{n} p_{ik} x_{ik} d_{ik}$$

Subject to the constraints:

$$\sum_{k=1}^{3} x_{ik} = 1 \; \forall \; i \; \in n$$

$$\sum_{k=1}^{3} \sum_{i=1}^{n} p_{ik} x_{ik} = c$$

$$0 \leq \; x_{ik} \; \leq 1$$

3.10 Expected Outcomes

As per the primary aims and goals of the study, following are the expected outcomes.

- A machine learning model that can be used to predict the prices of the product using the characteristics with evaluation rubrics.
- Price Optimization Function that can be useful for the organizations to provide discounts and analysis of business sales and revenue.

CHAPTER 4: ANALYSIS AND DESIGN

4.1 Introduction

This chapter contains detailed analysis, visualization, discussion related to the design and an overview of the methodologies used for the model development considered for this study. This chapter provides an overview of the dataset, the parameters considered during this study, followed by all the data cleaning and data preparation steps required for the further analysis. The data cleaning and data preparation is performed using elimination of unnecessary columns, transformation of categorical variables, scaling of numerical variables, removal of duplicates, etc. The strategies related to conversion of data types, binning for different categorical columns to reduce diversity, cleaning of textual data using regular expressions will be discussed. Univariate analysis will be performed for each of the variables considered for this study to understand the distribution of the required variables. Distribution of numeric variables using box plots, histograms and violin charts will be analysed. The identification of outliers and the strategy for treating the outliers will be discussed in detail. Moreover, strategies related to feature engineering and information extraction from other variables will be discussed in detail. The steps related to extraction of features from the important variables will be elaborated as there were few challenges during feature extractions.

The natural language processing techniques like stop word removal, punctuation removals and treatment for special characters will be performed followed by text to sequence along with glove embeddings using Keras and Tensorflow libraries. Exploratory data analysis will be performed to analyse and visualize the interdependency between variables and how each variable affects the distribution of the dataset. A detailed overview will be provided on splitting of the dataset into training and test dataset and parameters will be explained for the splitting strategy. Results will be interpreted using hypothesis testing and graphical visualization using matplotlib and seaborn.

Moreover, the strategy for optimal price prediction will be discussed for the stage 2. The preparation for the data for a particular brand and calculation of competitor's price strategy will be explained in brief to understand the suitability of the data for the application of Linear Programming. Also, the data will be visualized using Seaborn and Matplotlib to understand the effect of competitor's price on the products. A discussion related to data aggregation and strategies will be analysed further. A detailed summary of the analysis conducted, and the design will be discussed in the preceding sub-chapters.

4.2 Data Preparation

Even though the original data from Kaggle was in a structured format, the data was not clean or mature enough for direct analysis. Furthermore, the data is divided into many files, each including data on a different brand and its items. As a result, prior to univariate and multivariate analysis, a series of data cleaning, pre-processing, and data aggregation techniques were used to interpret and transform the data into the needed format while keeping the study's objectives in mind. This section will go over each of the cleaning and pre-processing stages, as well as the purpose and outcome.

4.2.1 Data Overview

The dataset used for this study is split across multiple files. The files include information about the products and particular brands. Figure *11* shows the information about brands and the number of products available. Victoria's Secret brand has highest number of products as compared to other brands. The dataset has 'product_name', 'product_category', 'color', and 'size' as a Candidate key. The information about the product 'mrp' and 'price' is distributed per each colour and each size. The dataset has 613,143 number of rows for each brand and with 14 columns.

	brand_name
victoria's secret	236606
b.tempt'd	21389
aerie	18258
calvin klein	12470
hanky panky	7627
us topshop	1021
vanity fair	595
nordstrom lingerie	191

Figure 11: Brand Information and Count of Product Items

4.2.2 Variable Elimination

The dataset has total 14 variables out of which, 'mrp' and 'price' are used as target variables. This dataset also has 'retailer', 'style_attributes', 'pdp_url' where 'retailer' value can be replaced with 'brand_name' and 'style_attributes' column doesn't have any information and is irrelevant for this study. The 'retailer' indicates the distributor of the products and for each brand, the value is 'unique'. This, 'retailer' duplicates the information of 'brand_name' and is dropped during analysis. On the other hand, 'style_attributes' doesn't have any information and 100% of the values are null, hence it is also dropped during analysis. The 'pdp_url' has links

for the images of the product. Based on the objective, this study only focused on textual information and price optimizations, so ‘pdp_url’ is also dropped during variable elimination. Also, there were few columns like ‘rating’, ‘review_count’ having null values more than 70% of the data, so imputation of null values is not possible. Thus, ‘review_count’ and ‘rating’ are also dropped. As a result of variable elimination, the dataset contains 9 relevant columns with 613,143 observations.

```
rating             134718
review_count       143151
style_attributes   240108
color                 253
dtype: int64
```

Figure 12: Counts of observations having null values

4.2.3 Variable Transformation

The datatype of all the 9 columns was ‘string’. So, every column needs an attention to be converted to specific formats for analysis and model building. The duplicates were dropped first to avoid repetition of values, and which might mislead the analysis. After dropping the duplicates, the dataset has 301026 observations. The ‘mrp’ and ‘price’ columns represented in the format of currency followed by value (e.g., $16). Hence, the currency needs to be brought into a single format like Dollar. A currency conversion factor is used for converting the price into dollars. The conversion rate for ‘₡’ and ‘rp’ is multiplied to convert into dollar. Moreover, there are two categorical columns viz. ‘product_category’ and ‘color’. The ‘product_category’ has 525 different categories and ‘color’ has 2558 different unique categories. Although, these categorical columns are spread across multiple categories, considering those categories without binning leads to devastating results during analysis as well as model building. Thus, in this study, similar items are grouped and a placed into a single category. The major disadvantage of this strategy is the loss in the diversity of the products that the brand has. To tackle this, feature engineering is performed to extract some features like ‘color_popularity’, ‘size_popularity’ which will be briefed in the later sections. Also, there were a lot of categories which might be treated as outliers with having a smaller number of counts in the whole column, these types of categories were placed into ‘other’ category. The Table *2* shows the number of categories after performing binning of the several categories available.

Table 2: Unique categories of variables

Sr. No.	Category Name	Values
1	product_category	bras, panties, bralettes, bodys, babydoll, slip, camisoles, onepiece, tops, rompoers, shorts, activewear, robe, suspenders, other
2	color	black, pink, blue, white, nude, grey, purple, red, green, yellow, brown, maroon, orange, multicolor, other

4.2.4 Information Extraction and Feature Engineering

Considering the objective of the study to predict the optimal price using product characteristics, out of the 9 columns, there are some columns like 'available_size', 'color', 'mrp' and 'price' which has more information rather than just using those as a categorical variable. This sub-chapter explains how information is extracted from these variables to calculate new variables which are '%discount', 'size_popularity', 'color_popularity'.

The calculation for '%discount' is pretty much straight forward. The 'price' variable signifies the actual price at which the product is sold and the 'mrp' variable signifies the original labelled price of the product. The discount offered is the absolute difference of original price and sold price divided by the original price. The percentage discount calculated can be used further to understand and analyse the pattern and the cases where the discount is offered.

$$\%discount = \frac{mrp - price}{mrp} * 100$$

On the other hand, the 'color' has wide number of categories available for a particular product. Thus, a simple binning into a popular category will evade the weightage and the advantage of wide varieties offered for a particular product. Thus, we need to consider a weighted component also to prioritize the importance of wide variety of a particular product available. The calculation of 'color_popularity' and 'size_popularity' is done in almost similar way. First, a 'size_index' is calculated to check the percentage of a particular size in total products. For each product and for a particular size, to calculate how many products are available, a 'size_weight' is calculated. Using 'size_weight' for each size, the 'size_popularity' is the average of all the 'size_weight' for each product. The calculation of 'color_popularity' is done in similar way. These features can be used to analyse the behaviour of variables with respect to the target and other variables. The general formula used for 'size_popularity' calculation is as follows:

$$sizeIndex = \frac{totalSizeCount}{totalNoOfProducts}$$

$$sizeWeight = sizeIndex * sizeCountForProduct$$

$$sizePopularity = \frac{sizeWeight}{NoOfItemsForProduct}$$

4.3 Univariate Analysis

To analyse and understand distribution of the values, spread of the data and anomalies in the data of for each variable, univariate analysis is carried out in this study. Pandas, Matplotlib, Seaborn, etc are the libraries that were used to conduct the univariate analysis. The data is visualized using several charts, graphs, and tables for understanding the distribution and relations between variables. This sub-chapter explains the distribution for each variable, the processing required as well as the importance of treating outliers and missing values.

1. 'mrp' and 'price':

These two columns resemble each other and have very high collinearity, hence 'mrp' is used as a target variable for this study. The 'price' column is used for calculation of '%discount' as discussed in the chapter 4.2.4. The Figure *13* shows the statistical characteristics of the 'mrp' and 'price' columns where mean value for price is 33.33 and mrp is 37.31, which shows that average 'mrp' is usually higher than the 'price' indicating the loss to the organisations. The minimum price and maximum price for the 'mrp' and 'price' are approaching to 0 to 615$ which is almost impossible cases in fashion industry. Also, 99% percentile value is 66.5 which is acceptable, hence all the data lies at 99% of the value is considered further in the study.

	price	mrp
count	301026.000000	301026.000000
mean	33.331434	37.319078
std	17.053397	15.983750
min	0.007716	0.007716
5%	9.990000	10.500000
25%	19.990000	29.500000
50%	34.500000	36.500000
75%	46.000000	49.500000
90%	56.500000	58.500000
95%	60.000000	62.500000
99%	66.500000	68.000000
max	615.000000	615.000000

Figure 13: Charactersistics of mrp and price

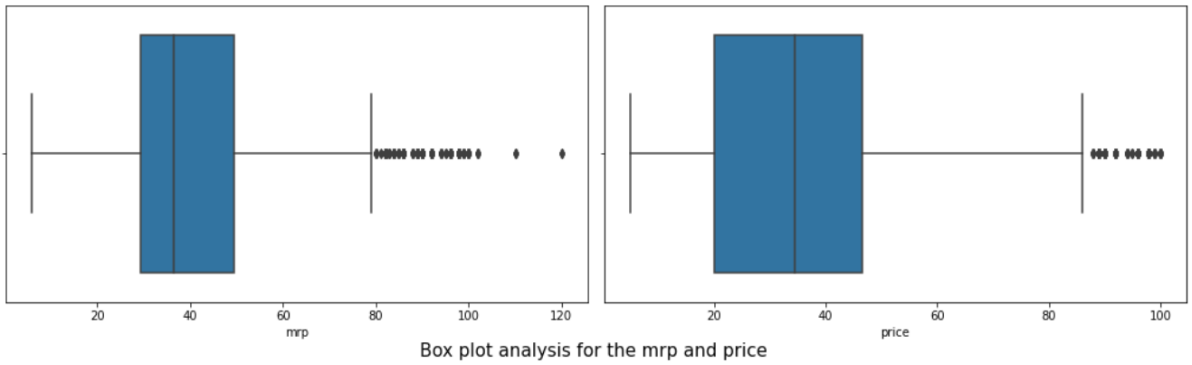


Figure 14: Box plots for mrp and price

Moreover, Figure *14* represents the box plots for mrp and price columns after treatment of outliers. The 25% of the data approximately has mrp around 30 and price around 20, where for 'price' can be seen a little bit on the higher side for 50%, but the overall 'price' is convergent mostly between 30 to 50. Thus, we can conclude that most of the products has price in the range of 30 to 50 and very few product costs more than 80$.

2. 'brand_name':

The dataset contains information about multiple brands and there are wide number of brands provided in the dataset. For few brands, very few numbers of records are available and such records are outliers for this study. Thus, such records having less than 50 records are dropped from the dataset. The 'victoria_secret' brand has most number of items and varieties followed by 'b.tempt'd'. This study used 'victoria's secret' data as primary dataset during the price optimization stage as the number of records are higher than any other brands. And the other brands data is used as a competitor's data which is used for price optimization.

brand_name	Total number of items	Number of unique products
aerie	18258	106
calvin klein	12470	443
b.tempt'd	21389	395
hanky panky	7627	852
victoria's secret	236606	593
vanity fair	595	43
nordstrom lingerie	191	23
us topshop	1021	281

Figure 15: Number of product items for each brand

Figure *15* explains how for each brand, there are different records for each size and color, and the unique records for each of the brand is elaborated in second column. Figure *16* shows a

visualization for each of the brands and their available counts of the products. Victoria Secret has the greatest number of observations whereas nordstorm lingerie has the least number of observations.

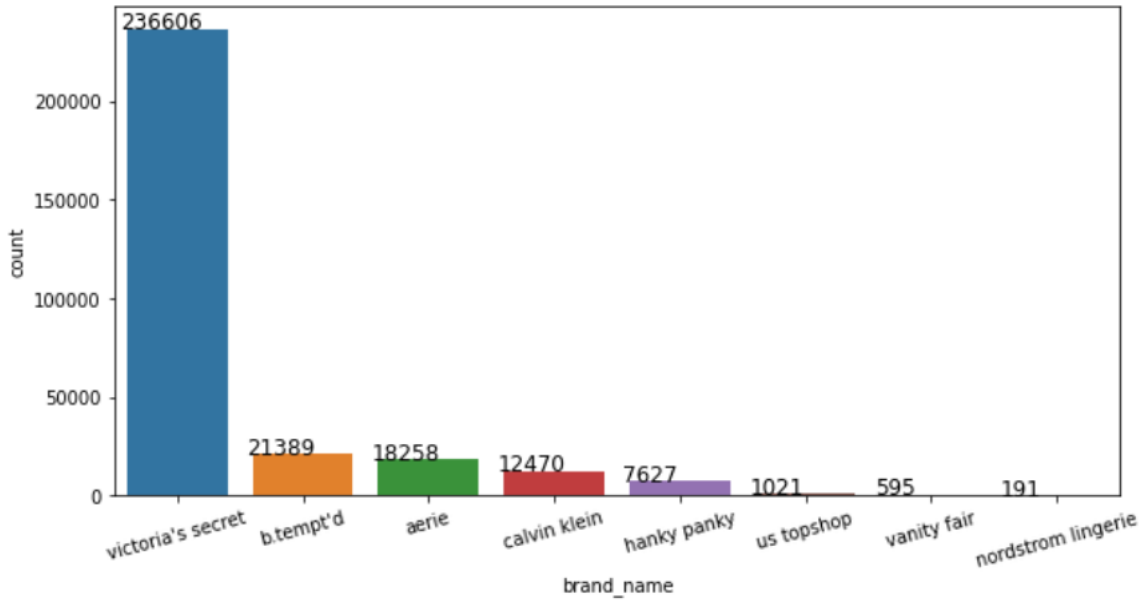


Figure 16: Bar chart for counts of product items

3. ‘product_category’:

The product category column is earlier processed to reduce the number of distinct categories for available and binned similar items into a single category. e.g., thongs, bodypants, bikini bottoms are grouped into a single category ‘panties. The Figure *17* explains the distribution of items and unique products in each category. The dataset has 214593 items for bras as compared to any other categories like panties, bralettes, etc. However, the number of unique items for bras are slightly less than panties. Which means that, there is a lot of variety for bras with respect to size and colour as compared to any other category. This also helps organisations to decide strategies related to pricing based on the popularity and demand. The more the variety, the more is demand and hence the pricing decisions like discounts, sales, campaign, etc. can be carried out at frequent intervals. The bralettes are specific categories of bras but are identified as special categories based on the number of available products. This study has identified it as a different category because bralettes has 26182 different items and binning those into a single ‘bras’ category would have caused data imbalance during cleaning stage itself. Also, there are few brands which also have information about different products like lace, bottles, beauty products, etc have been placed into ‘other’ category because it was not related to the objective for this study. Also, Victoria’s secret doesn’t have information about such irrelevant products, and this helped in stage 2 for price optimization stage where competitor’s price is considered.

product_category_wide	Total number of items	Number of unique products
bras	214593	701
panties	40095	1060
bralettes	26182	321
other	10036	311
bodys	1481	76
slip	1195	33
babydoll	1187	56
camisoles	1052	73
onepiece	829	11
tops	805	30
rompers	200	8
shorts	174	19
activewear	170	6
robe	118	8
suspenders	40	16

Figure 17: Number of product items for each category

4. ‘available_size’ and ‘total_sizes’:

A particular product usually has various number of sizes supported for each product and category depending upon the brand and colours. The dataset has ‘total_sizes’ columns which represents the supported sizes for a particular product. The size varies from XXS to XXL and 30A to 44H.

	Total number of items	Number of unique products
S	13728	444
M	13584	451
XS	13564	437
L	13243	443
34B	8466	135
32D	8452	132
34DD	8449	118
34C	8397	133
32DD	8356	114
34D	8311	130
36C	8174	130

Figure 18: Number of product items for each available size

The available size column represents price and mrp for each size. The available size is taken from total_sizes available thus the total_sizes column is repeated for a particular product. Hence, this study considers only 'available_size' column which is being further used to calculate 'size_popularity' as explained earlier. Figure *18* explains few of the sizes along with their counts and unique items. The 'S' size has more information available with 13728 records as compared to other sizes. However, 'M', 'XS', 'L' has almost similar number of records available so the overall popularity for the sizes and supported products are similar. We can conclude that every brand generates almost similar number of products for all the sizes in few exceptional cases. There are lot of supported sizes available in the dataset and the binning would have been performed but the few sizes may vary according to the dimensions and the binning would have caused a loss in the data quality. However, there are wide varieties of sizes for each product, this study also analysed the effect of sizes on the target variable in the later sections.

5. 'color':

The 'color' column represents the colour of a particular item. The dataset had wide number of supported colours including 'multicolor' products also. The 'color_group' is the column derived from the 'color' column after combining similar items to reduce categorical distributions.

color_group	Total number of items	Number of unique products
black	44185	1583
multicolor	40079	534
pink	31966	544
blue	29319	758
white	28361	806
nude	22712	491
grey	18375	426
other	16912	1110
purple	16574	367
red	12863	476
green	11808	262
yellow	8601	233
brown	6965	259
maroon	5140	88
orange	4297	177

Figure 19: The number of colors and unique items

Figure *19* represents the distribution of color column performed after binning. There are 44185 items for 'black' and the dataset has 1583 unique items for 'black' color. Also, the 'multicolor' category has second highest observations. As, there are a lot of categories and items available in the dataset, the effect of price over the color is analysed in the further sections of this study.

6. size_popularity and color_popularity:

The size popularity is derived from the weights of sizes whereas color popularity is derived from weights of colors for each particular item to reduce the effect of binning. These columns are required and used further for analysing the effect of colors and sizes on the prices of the products. The Figure *20* describes the color and size popularity for victoria's secret products. The values range from 0.07% to 258% and 440%. The more the value of the popularity, the more the product has varieties of items for both size and color perspective.

The calculation for size and color popularity explained in the section 4.2.4 is completely based on the mathematical as well as statistical interpretation of the weights for each product. The parameters are always helpful to understand the available variety for a particular item which is directly proportional to the observations.

	color_popularity	size_popularity
count	669.000000	669.000000
mean	0.108923	0.557278
std	0.176009	0.500050
min	0.000070	0.000070
25%	0.018442	0.203148
50%	0.060374	0.382548
75%	0.128570	0.754805
max	2.585449	4.400694

Figure 20: Description for size and color popularity

4.4 Treatment of missing values

After performing the univariate analysis, most of the variables in the dataset doesn't have missing values. However, there are few exceptional cases in which missing values exists like for variables like 'color' and 'category'. The records having missing values cannot be dropped as it will lead to very few data during training and testing the machine learning models. Also, for 'product_category', several similar categories were binned together except few cases where the data has information about products apart from lingerie like bottles, caps, etc. Thus, such records have been placed into 'other' category.

4.5 Multivariate Analysis (Bivariate Analysis)

Since exploration of one variable at a time in univariate analysis is helpful for analysing the distribution of the data, the relationship between two attributes or variables can only be carried out when two variables are analysed at the same time. This can be achieved after conducting multivariate analysis for identification of existence of an association between target variables and the dependent, the significance of the minute differences can be analysed through statistical tests. As the target variable is continuous and dependent variables are of categorical type, the bivariate analysis is limited to analyse the effect of a dependent variable to identify specific pattens. The bivariate analysis is conducted on complete dataset in this study and the preference is given to the target variable to analyse the patterns of the dependent variables.

1. mrp vs brand_name:

The target variable 'mrp' is analysed against the 'brand_name' in this study to understand how the price varies with each respect to the brand. The brand does affect the price of the products (The two variables are dependent). Figure *21* shows the box plot visualizations for the brands against their prices. It is very clear from the visualization that price distribution changes according to the brand. B.temp'd has most expensive products amongst the brands. The 50% value is approximately 60$. The 50% value for hanky panky, calvin klein, Victoria's secret, etc. is approximately 40$. Hanky Panky has the most distributed price amongst all the brands.

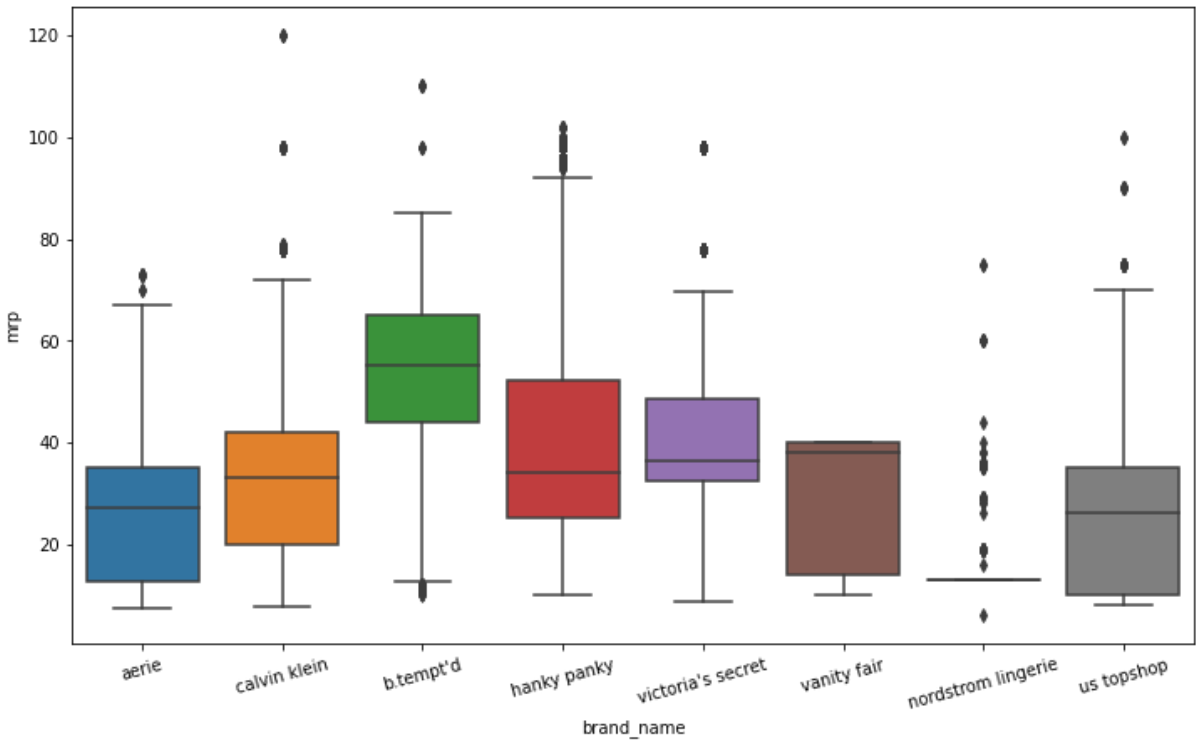


Figure 21: Price and Brand Name

2. mrp vs product_category:

The target variable ‘mrp’ is analysed against the ‘product_category’ in this study to understand how the price varies with each respect to the category of the product. The price of the products varies with respect to the kind of the product. Figure *22* shows the box plot visualizations for the categories against their prices. It is very clear from the visualization that price distribution changes according to the brand and there are certain price ranges for category. ‘bodys’ is the most expensive category amongst the other categories, also prices are distributed from lowest to highest. The 50% value is approximately 55$. The 50% value for bras, tops and slips, etc. is approximately 45$. Panties has the most distributed price amongst all the brands with lowest 50% price.

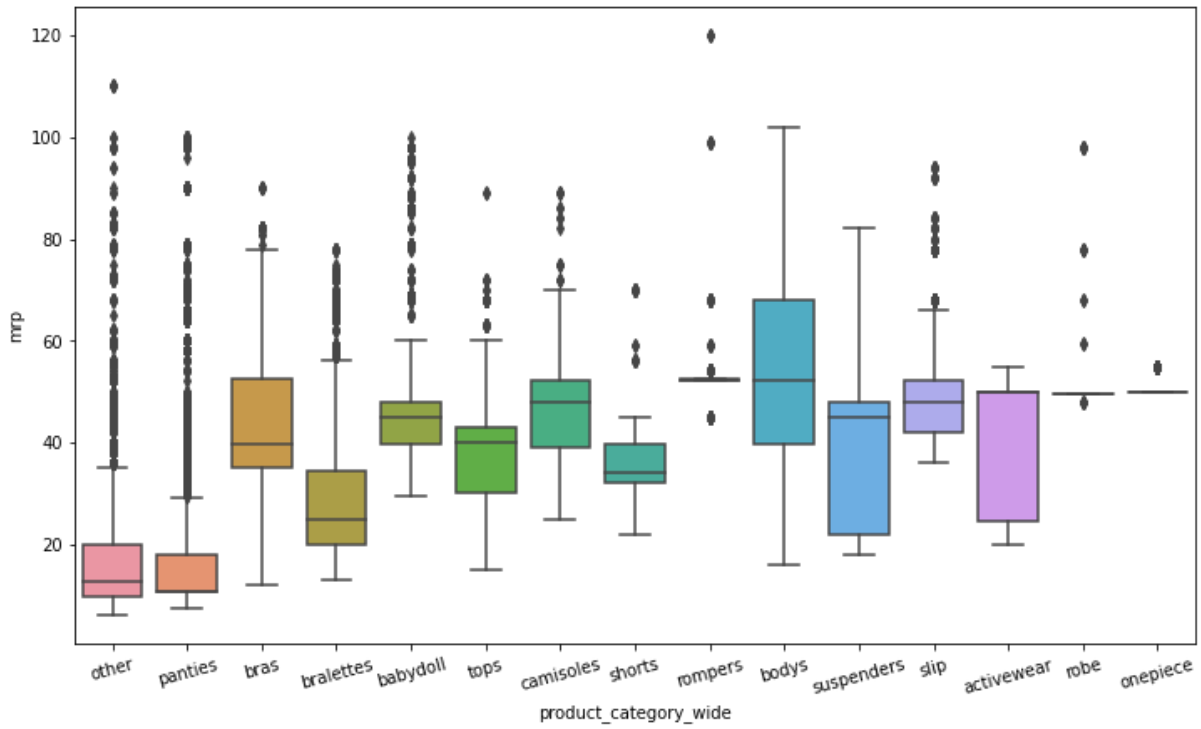


Figure 22: Price of items with respect of Category

3. mrp vs color_group:

The target variable ‘mrp’ is analysed against the ‘color_group’ in this study to understand how the price varies with each respect to the color of the product. However, it’s been observed that there is not many variations of the prices with respect to colors. Figure *23* shows the box plot visualizations for the colors against their prices. It is very clear from the visualization that price distribution remains similar according to the colors and there are certain price ranges for each color group. The ‘brown’, ‘maroon’, ‘pink’ and ‘nude’ are the most priced colors with respect

to the 50% of the data. The ‘multicolor’ coloured products appears to have the compact ran e of price varying from 20$ to 50$ as compared to the distributions of the other color groups.

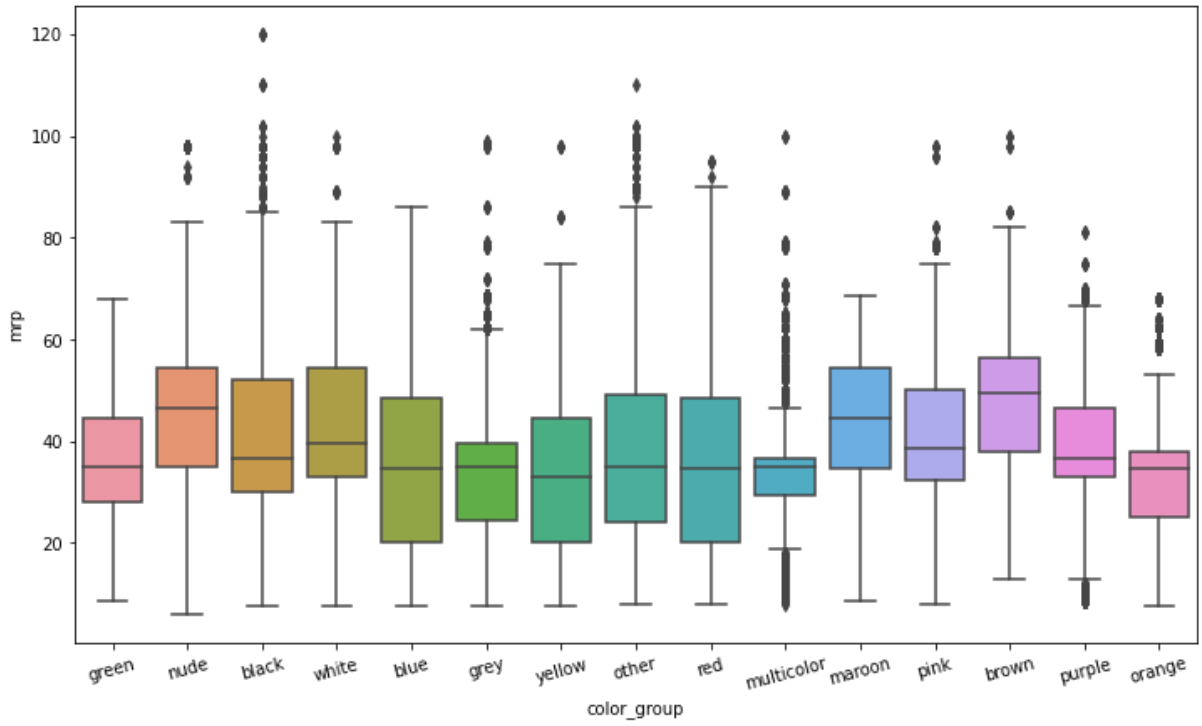


Figure 23: Price of product items with respect to colors

4. %discount vs brands_name:

The discount is the derived attribute using mrp and price which represents the percentage of the discounts offered on the product by organisations. It’s been observed that the discount varies from brand to brand, and it can be the organisational decisions to provide discounts to attract customers.

Figure *24* shows the average and the 95% distributions of the discounts offered by the brands. Aerie offers most discount around 20% as compared to other brands whereas hanky panky and nordstorm lingerie offers where less discounts below 5%. The average percentile discounts offered by the brands calvin klein, b.temp’d and victoria’s secret is almost similar approximate 10%. Also, the second chart shows the median discounts provided by the brands. The median value for aerie is around 30% whereas rest of the other brands has 0% median. That means the brands offers prices on few of the selected items instead of all the items.

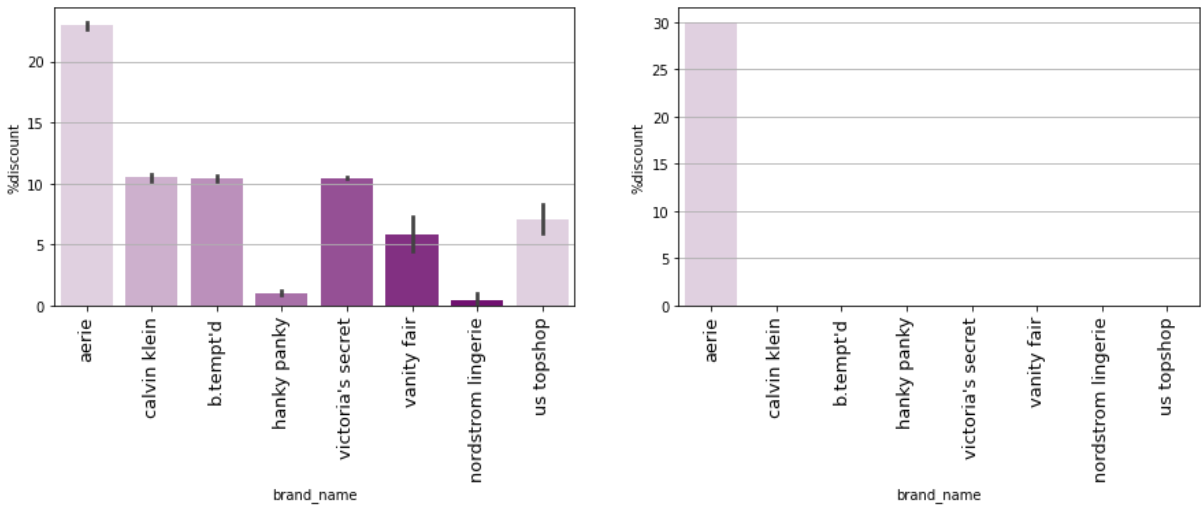


Figure 24: Distribution of Discounts offered by brands

5. %discount vs color and size:

It is important to understand whether the discount is offered specific color groups or specific sizes by the organisations as a part of strategic decisions. As there is a wide variety of color groups and sizes, only few and popular colors and sizes are analysed in this study against discounts. The Figure *25* shows the distributions of mean discounts against each color group and sizes. The primary observations are there is not many variations with respect to the popular available sizes for discounts except XL. The organisations don't prefer available_size while offering the discounts. Also, the average discounts offered in each of the color group is different which means the organisations provides discounts based on the color groups. The maroon category has fewer average discounts as compared to other colors. It was also observed in the section 3 that maroon is most priced color group also. The multi-coloured products have been provided with highest average discounts on the products.

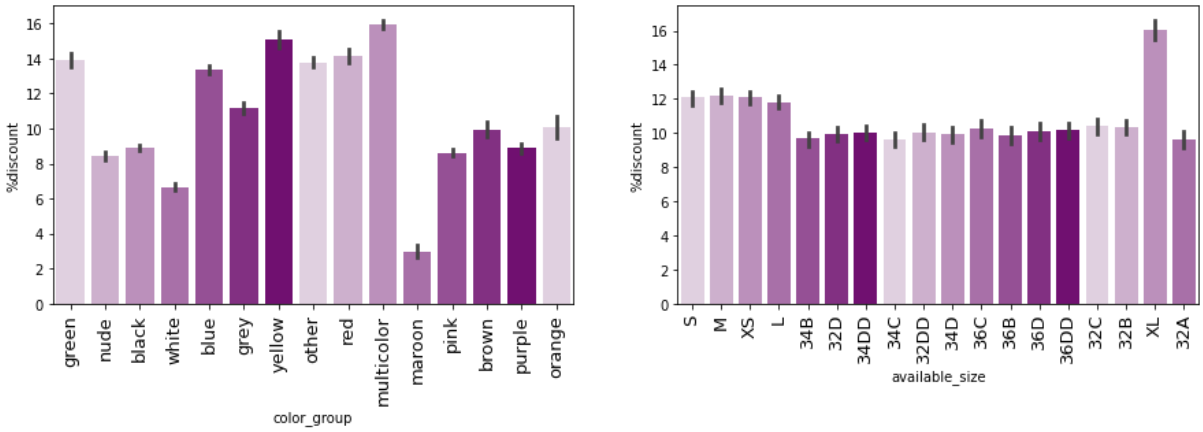


Figure 25: Discount vs Color and Size

4.6 Splitting of the original dataset:

In addition to the original dataset of victoria's secret with the imputed missing values, 5 additional data sets in .csv format were merged from the complete dataset. Data partitioning was done randomly to using train_test_split library in python. As there is no need to split the dataset based on certain conditions as well as data imbalance will not be the case considering the objective of the study for predicting and optimising the original price, the dataset is split into training as test dataset in a random behaviour. The dataset was imbalanced originally, after combining the data of all the brands and products, there were multiple records having 'product_name', 'category', size and 'color' as the candidate key. After considering only the unique product descriptions from the all the files, the size of the dataset reduced to approximately 6000. The dataset now has total 9 useful variables out of which 'mrp' is the continuous target variable, brand_name and the product category are the 2 categorical columns. However, to predict the price based on the product description and combining the models with other useful variables, the training dataset was used to train the model, generate a predictive model. Test dataset is then used to assess and evaluate performance of the model.

1) Complete Dataset Including Training and Test Data): 5982 records
2) Training Dataset: 3101 Records
3) Test Dataset: 345 Records

4.7 Data Pre-processing for Description and Product Name:

The dataset has 'product_name' and 'description' as textual columns. Based on objectives of the study to standardize the price using product characteristics such as 'description', 'category', 'brand_name', etc. the processing of the textual description of the products is carried out. The standard application domains involve techniques such as useful information extraction, translation of texts between languages or required parameters, written work summarization, automatic answering and inferring, and classification/clustering of documents (Otter et al., 2021). To process the textual data, NLP techniques like data stop words removal, punctuation removal, tokenizing, sequence padding, etc. were used in this study. There are a lot of studies conducted earlier on pre-processing techniques of textual information using natural language processing as discussed in the literature review. This study used few of common methods used by earlier researchers based on the requirements.

4.7.1 Stop Words Removal:

Words like ‘and’, ‘it’, ‘a’, ‘the’, ‘to’, etc. can be found in virtually every sentence in English-based documents. Secondly, these words make very poor index terms and have low discrimination value (Kaur and Kaur Buttar, 2018). Information carried out by such type of words is negligible. Also, the stop words or noise words have less candidature categorized as articles, prepositions, conjunction. Figure *26* shows a sample description from the dataset. There are few words which are repeating and can be ignored. Also, the words which occurred frequently like ‘demi’, ‘bra’, ‘made’, etc. would get some noise because of stop words. Because of high frequency, less predictivity, irrelevancy, etc. such words need to be filtered from the textual data before applying machine learning techniques to it. This study used stop word corpus from NLTK for removing stop words using python. Also, few custom words were added to the list generated to avoid misleading the models.

Figure 26: Frequently Used Words in the description

4.7.2 Punctuation Removal

Figure *27* shows a sample row from dataset for item description. The item description has textual information which can contain anything like postscripts, prefixes, punctuations like ‘,’,’:’ which are also not required for the analysis and can be removed. There are several studies conducted like the study conducted by author (Renault, 2020)believed that punctuations and emojis improves accuracy in prediction the sentiment of the word. But based on the objective of the study to predict the price, punctuations and unnecessary words needs to be filtered to remove noise as these will create unnecessary noise while building models. Hence, punctuations are also removed from the descriptions.

```
0    Introducing Everyday Loves™: Made with love. E...
0    soft cotton stretch fabric and a metallic logo...
0    An unlined demi cup bra featuring sheer, sexy ...
0    Say “buongiorno!” to this ladylike piece that ...
0    Hanky Panky Silky is the ideal fabric for unde...
0    The perfect amount of coverage in a subtle sil...
0    Lighter-than-air, full-cut Supima® cotton brie...
0    These feminine black knickers for maternity fe...
0    Lots of cheek peek, pretty lace, a strappy bac...
Name: description, dtype: object
```

Figure 27: A sample of Description from the dataset

4.7.3 Word Tokenizer:

Considering languages having many vocabulary words and a rich morphological system, the prominent trade-off with neural network language models is the size of the network (Noaman et al., 2018). Being most important step in the natural language processing, the process of splitting of words/phrase, paragraph, or sentence, etc. needs to be carried out in order to feed textual data to the machine learning models. The machine learning processes models in the numerical format for the prediction of target variables, text to sequence tokenization is used which converts the tokenized words into sequences of numbers in an array. This study used Tokenizer class from TensorFlow library for Tokenizing the words. This class vectorizes a text corpus, which turns each text data into a sequence of integers, where each integer can be the index of a word token in a dictionary, based on word count, based on TF-IDF, etc. Figure *28* shows a converted description into a sequence of description. The maximum length of the sequence generated from the dataset is 91. This sequence of textual information can now be fed to machine learning models which can predict the price based on the number of vectors available in the corpus. The prediction is completely based on the information available in the sequences.

clean_description	seq_description	seq_product_name
i do emblazoned hanky panky lace hipster swa...	[540, 541, 1760, 83, 90, 1, 98, 1201, 557, 296...	[540, 541, 70, 1, 175, 98, 283, 439, 561]
stretch signature lace fashions alluring doubl...	[34, 70, 1, 1035, 612, 542, 174, 415, 246, 109...	[2569, 2570, 1, 211, 397, 415, 246]
leopard spotted lace adds bold open gusset des...	[182, 691, 1, 416, 724, 211, 397, 135, 644, 49...	[2569, 2570, 211, 397, 1, 8]
scalloped trim adorned little blue bow flirts ...	[302, 75, 1406, 340, 285, 156, 2620, 521, 77, ...	[2571, 1, 426]
scalloped lace adorned little blue bow defines...	[302, 1, 1406, 340, 285, 156, 1990, 1407, 1109...	[2571, 1, 288]

Figure 28: Data Converted to Sequences of Integer Vectors.

4.7.4 Sequence Padding:

As all input vectors must be of the same size before feeding to a neural network, the techniques like sequence padding and truncation are applied (Lopez-del Rio et al., 2020). This study used keras library for sequence padding. The padding length needs a fine tuning before feeding to a machine learning models. As the maximum length of the sequence is 91, the 95% of all the sequences is 75. Hence, all the sequences are padded to the length of 75. The pre-sequence padding technique is used to pad the sequences with '0's at the start of the sequences. Also, the maximum length of the sequences for 'product_name' is 12, all the 'product_name' are padded to the length of 10.

4.8 Conversion of the Variables:

1. Normalization:

The main idea behind the normalization is the variables of different scales need not to contribute equally during model fitting, model learned functions and probably ending up in creating bias. Thus, to tackle this situation, scaling is performed for target variables. This study used Min Max Scaling for scaling the target variables in the trained and test data. As non-tree-based models are dependent on the normalization, scaling of the continuous is always preferred. Figure *29* shows the distribution of the target variable after scaling is performed. It's almost a standard distribution of the values where the primary values range lies between -0.50 to -0.25.

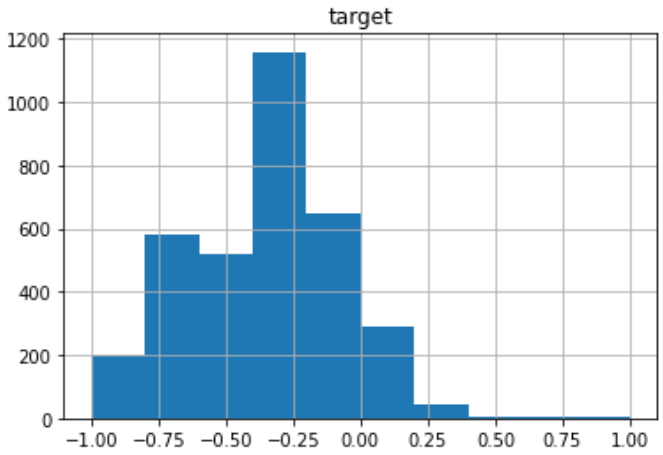


Figure 29: Distribution of target variable after Min-Max Scaling

2) Encoding:

For conversion of the categorical columns into numerical, encoding is required. Label Encoding is the process refers to converting labels into numerical form which then further used to convert them into a machine-readable form. This study used LabelEncoder from sklearn and then model is transformed into the numerical columns.

4.9 Hyper-Parameter Tuning for Models:

This study used hyper-parameter tuning before feeding the data to the machine learning models. Below are few of the techniques referenced for hyper-parameter tuning.

1. Cross Validations:

Several ML algorithms are in search of different patterns and trends. However, one algorithm isn't the best fit for all datasets. This study conducted a lot of experiments, to evaluate different ML algorithms and tuning their hyper parameters. Model validations using k-fold (10 folds) cross-validation was used on the aggregated data of different brands and products. Researche and studies conducted by (Liu and Sustik, 2021), reported that 10 is an optimal number of fields, resulting in optimal completion time as well as expected results. The estimates for accuracy achieved from various replications for k-fold cross-validation are usually highly dependent and correlated with the correlation being higher with the increase in number of folds (Wong and Yeh, 2020). The 10-fold cross-validation strategy partitions the entire aggregated dataset into 10 mutually exclusive subsets or folds randomly having the folds with the same number of records. These 10-fold cross-validation results were averaged to generate an estimate for determining the RMSE and RMSLE.

2. Tuning for Epochs:

As the gradient descent has learning rate, which is helpful for optimising the predictions. Similarly, epochs, iterations, batch sizes are the tuning parameters where the dataset size is huge, and it cannot be fed to the models at once. Also, passing the data to neural networks is not enough sometimes, the full dataset needs to be passed multiple times to the neural networks to optimise the learning rate (Roslidar et al., 2019). However, there are chances that the model overfits the data and perform poor on unseen data. Figure *30* shows how the data overfitting and underfitting problems and their causes. We need to always identify the optimal number of epochs so that the model doesn't overfit as well as underfit.

In this study, research has been conducted to analyse how the loss function and RMSE value on test dataset varies with varying the epochs count. The Table *3* shows the variation of loss and RMSE with respect to the epochs count. The loss decreases with increasing the epochs count but the RMSE on test data moves up and down. As the goal is to reduce the RMSE value, the optimal value for epochs is chosen as 50.

Table 3: Epochs Count and Loss Function

Epochs Count	Loss, RMSE
10	0.369, 13.289
50	0.03, 7.85
100	0.001, 10.64

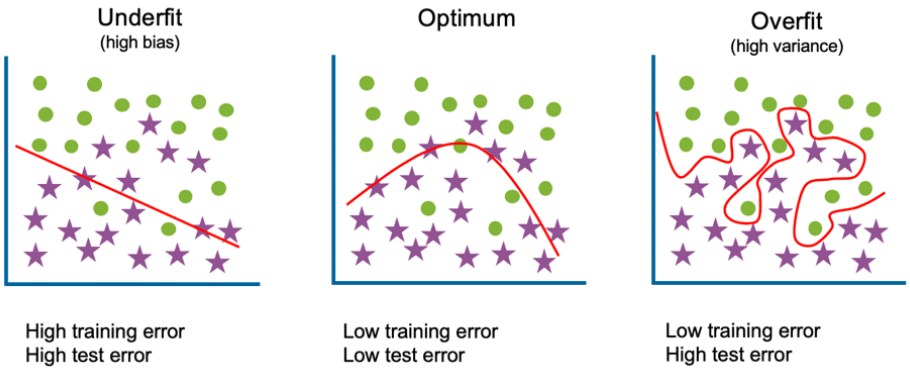


Figure 30: Model Fitting Issues: Source: (https://towardsdatascience.com/demystifying-model-training-tuning-f4e6b46e7307)

4.10 Data Preparation for Price Optimization using Linear Programming:

The idea of price optimization conducted earlier in the research by (Kedia et al., 2020) is extended in this study using Linear Programming. The data used for price optimization is for the brand Victoria's Secret because Victoria's Secret had 699 products with most of the varieties as compared with other brands. A competitor's price has been calculated with the help of the data from other brands. Now, having a price range (which is a standardized price of the product and competitor's price), the price can vary between the standardized price calculated from machine learning prediction and the competitor's price. The product price in that range is divided in three, to have the prices as p1, p2, p3 where p1/p3 can be standardized price and competitor's price. The p2 is always the average price of the predicted price and competitor's price. As the demand prediction is one of the toughest tasks, (Kedia et al., 2020) proposed ARIMA and LSTM based models for predicting the demand of the product using Sales Data. Based on the objective of the study and the dataset availability, this study used a dummy column (derived from %discount) to populate the demands for the product as d1, d2 and d3. As more is the discount, the price can be affected.

Now, using the values for prices and demands, the goal of the price optimization is to maximize the revenue along with providing some discounts to attract the customers. Hence, for each combination of the price, there will be three combinations which are p1d1, p2d2, p3d3. The optimal value from these three combinations can be chosen using Linear Programming to

maximize the revenue. Figure represents the price and demand combinations for each product. This study used linprog library available in the scipy to select the optimal price.

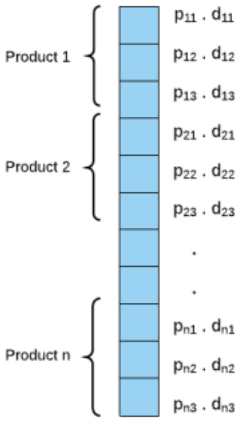


Figure 30: The price and demand combinations for each product.

4.11 Summary

Overall, the chapter 4 provided information about the dataset used in the study and how this data was utilised for conducting analysis and creating visualizations. The data set used in this study was obtained from Kaggle where there were 14 related columns and the target variable ‘mrp’ states the price of the product. There were multiple files in the dataset initially which was merged later for conducting analysis. There was a total of 613,143 records in the dataset initially. To prepare the data for analysis, data cleaning is performed first. The unrelated variables were eliminated, variable transformation is performed and the variables which were originally given in the category format was binned together to reduce the cardinality. Feature engineering is done to derive new variables like %discount, colour popularity and size popularity. Univariate analysis is performed on the dataset using visualizations like Box Plot, Count Plots, etc. After analysing the single columns, missing values are imputed accordingly. Multivariate analysis is conducted to understand the distributions, relations, co-linearity between the variables using visualizations. The data is then split into train and test data before machine learning model building stage. As there were approximately 6000 unique entries for product descriptions only, the dataset is split with 90% of the train data. However, the tuning is performed on train test split later using K-Fold cross validation techniques with 10 folds. As there are few columns in the dataset with textual information, natural language processing techniques were used viz. stop words removal, punctuation removal, tokenizing, text to sequence generation, sequence padding, etc. Min Max Scaling is performed on the target

variable for data normalization. Label Encoding is also used for encoding the categorical columns into numeric categories. A hyper-parameter tuning techniques like k-fold cross validation, efficient number of epochs selection, etc. is also performed to achieve optimal results. In short, an overview of the dataset along with visualizations, data cleaning, data pre-processing, exploratory data analysis, hyper parameter tuning was performed and analysed in detail. The next chapter will include the results, optimizations, visualization of the machine learning models output, performance of the models, validation, price optimization.

Moreover, the data preparation and approach for performing price optimization is also explained this section. The competitor's data is used to calculate the competitor's price. This competitor's price is used to calculate the price range in which the price optimization can be performed. This study modified the approach used by (Kedia et al., 2020) for optimizing the price of the products. This study used the scipy linprog library for maximizing the revenue. The results for price optimizations will be discussed in the later sections.

CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter will include all the findings of output results and analysis for the machine learning models used in this study. The results will be interpreted in detail along with the visualizations. The performance of all the models along with the parameters and optimizers will also be discussed in this chapter.

Moreover, the findings on the performance and the evaluation of machine learning models and its results will also be elaborated. To explain further the RMSLE, RMSE metrics values which were obtained from the construction of the GRU, LSTM, CNN based models will be displayed and interpreted. Thus, each machine learning model along with the optimizers 1. Stochastic Gradient Descent 2. Adam 3. AdaGrad will be compared and evaluated. With the three optimizers along with the above three models there will be 9 RMSLE and RMSE matrix that will be interpreted in this chapter. Amongst all the performance measures evaluated using with the help of metrics, will be analysed, and compared against all the machine learning models used in this study. The model with the best overall performance will be selected as the final predictive model to predict and standardize the price. This study will also highlight the approach for selecting the best model using performance metrics.

In addition, a detailed approach for price optimization in this stage 2 used for this study will also be discussed along with the results. This Linear Programming based price optimization algorithm which was earlier used in the research conducted by (Kedia et al., 2020) is interpreted in this study also compared with the other existing approach with pros and cons. Finally, using the Linear Programming for Price Optimization, a optimal predicted price which maximizes the revenue will also be displayed and interpreted in this section.

5.2 Evaluation of Machine Learning Models and Results:

Each of the machine learning models was compared based on their suitable optimizers, epochs, iterations, etc. and the best overall performance considering the performance metrics was chosen as the best prediction model for the standardized prices of the products. The optimizers and models were evaluated based on performance metrics like RMSLE, RMSE, Loss Percentage, etc.

For evaluation and standardizing the price of a product, it's very important to predict the optimal price in the same category and then optimizations on the price can be applied. Thus, the errors should be reduced in prediction of the price using the measures like RMSE, RMSLE.

As the goal is to reduce the RMSE and RMSLE values, a greater the value of RMSE and RMSLE, the worst the performance of the model. Hence, the RMSE and RMSLE values should be lower. So, the model performance is evaluated which minimizes the RMSE and RMSLE values in this study. The following Figure *31* explains the machine learning flow diagram where the data is fed to the input layer first, followed by embedding layer. The embedding is performed using standard Glove Embedding provided by Stanford. The LSTM/GRU/CNN layer is where the actual model is applied. The models are concatenated using Dense layers along with Batch Normalization. The output layer is combined to predict the prices of the product using linear activation.

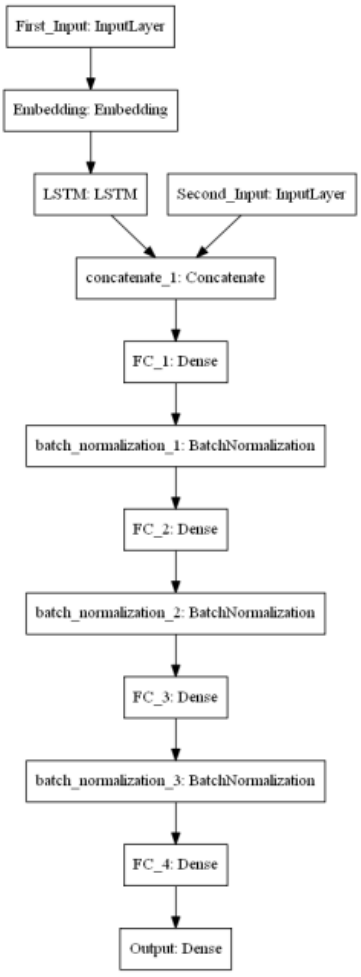


Figure 31: Machine Learning Models Flow Diagrams

5.2.1 Gated Recurrent Units based models:

The results for GRU are compared using the RMSE and RMLSE values for each of the optimizers. First, the GRU model is trained for only 'product_description' considering the objectives. A similar GRU model is trained considering the characteristics and derived attributes. The loss function based on MSE values is visualized for each of the optimisers.

5.2.1.1 GRU with only product description:

To predict the price of the product using only description, GRU with the input layer, embedding layer, GRU layer, main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different optimizers like adagrad, adam, sgd. Figure *32* shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *4* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 14.4821 and RMLSE as 0.3009.

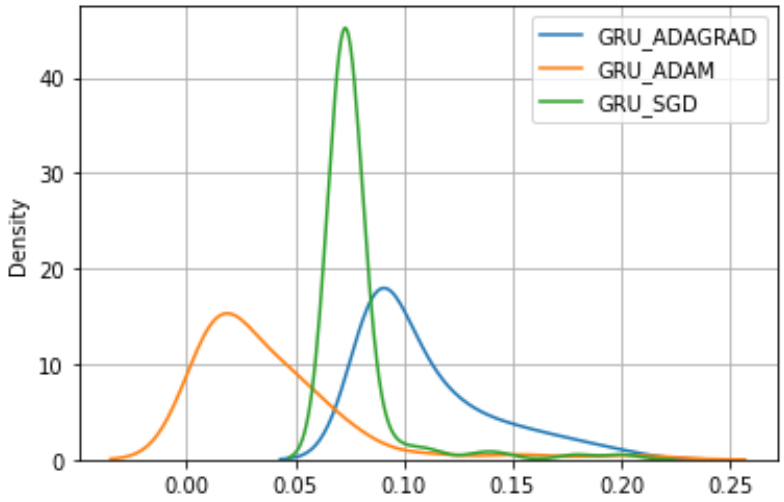


Figure 32: Loss Function Distribution against different optimizers for GRU with only product description.

Table 4: Performance comparison for GRU with only product description

Optimizer	RMSE	RMLSE
Adam	14.4821	0.3009
SGD	27.9451	0.5858
Adagrad	27.8693	0.6327

5.2.1.2 GRU with all the characteristics:

To predict the price of the product using all the characteristics of the product along with derived attributes, GRU with the input layer, embedding layer, GRU layer, main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different optimizers like adagrad, adam, sgd. Figure *33* shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *5* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 8.8067 and RMLSE as 0.2251.

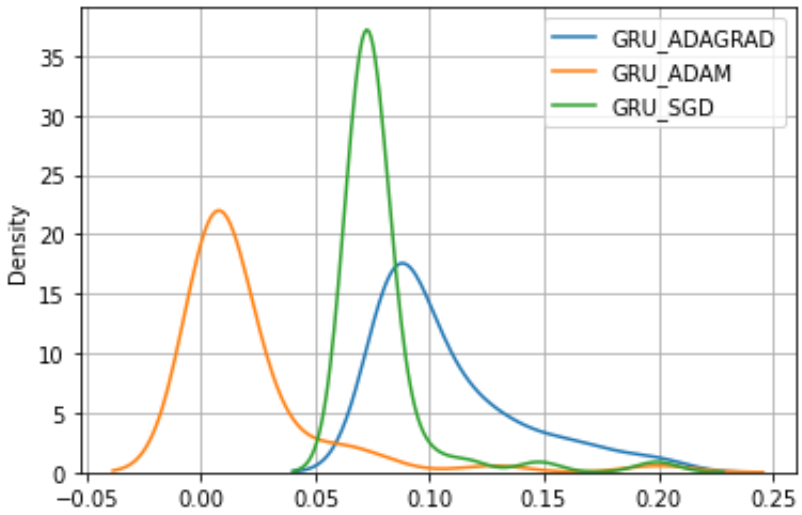


Figure 33: Loss Function Distribution against different optimizers for GRU considering the product characteristics

Table 5: Performance comparison for GRU considering all the product characteristics

Optimizer	RMSE	RMLSE
Adam	8.8067	0.2251
SGD	19.7921	0.5524
Adagrad	20.3184	0.6086

5.2.2. Long Short-Term Memory based models:

The results for LSTM are compared using the RMSE and RMLSE values for each of the optimizers. First, the LSTM model is trained for only ‘product_description’ considering the objectives. A similar LSTM model is trained considering the characteristics and derived attributes. The loss function based on MSE values is visualized for each of the optimisers.

5.2.2.1. LSTM with only product description:

To predict the price of the product considering only the product description, LSTM with the input layer, embedding layer, LSTM layer, main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different optimizers like adagrad, adam, sgd. Figure *34* shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *6* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 13.4699 and RMLSE as 0.2786.

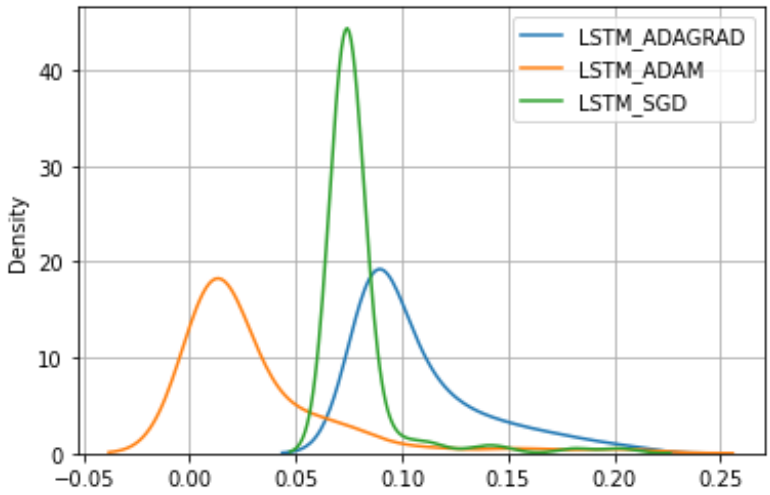


Figure 34: Loss Function Distribution against different optimizers for LSMT with only product description

Table 6: Performance comparison for LSTM with only product description.

Optimizer	RMSE	RMLSE
Adam	13.4699	0.2786
SGD	28.0606	0.5913
Adagrad	28.1460	0.6328

5.2.2.2. LSTM with all the characteristics:

To predict the price of the product using all the characteristics of the product along with derived attributes, LSTM with the input layer, embedding layer, LSTM layer, main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different optimizers like adagrad, adam, sgd. Figure

35 shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *7* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 7.7766 and RMLSE as 0.2113.

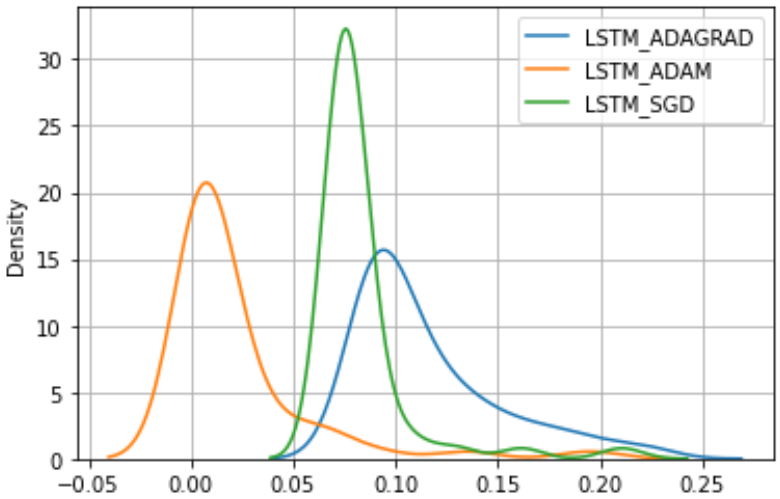


Figure 35: Loss Function Distribution against different optimizers for LSTM considering the product characteristics

Table 7: Performance comparison for LSTM considering all the characteristics of the product.

Optimizer	**RMSE**	**RMLSE**
Adam	7.7766	0.2113
SGD	20.1760	0.5645
Adagrad	20.9685	0.6274

5.2.3 Convolutional Neural Network based models:

The results for CNN are compared using the RMSE and RMLSE values for each of the optimizers. First, the CNN model is trained for only ‘product_description’ considering the objectives. A similar CNN model is trained considering the characteristics and derived attributes. The loss function based on MSE values is visualized for each of the optimisers.

5.2.3.1 CNN with only product description:

To predict the price of the product considering only the product description, CNN with the input layer, embedding layer, CNN layers (Conv1D, GlobalMaxPooling, etc.), main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different optimizers like adagrad, adam, sgd. Figure *36* shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *8* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 12.2505 and RMLSE as 0.2543.

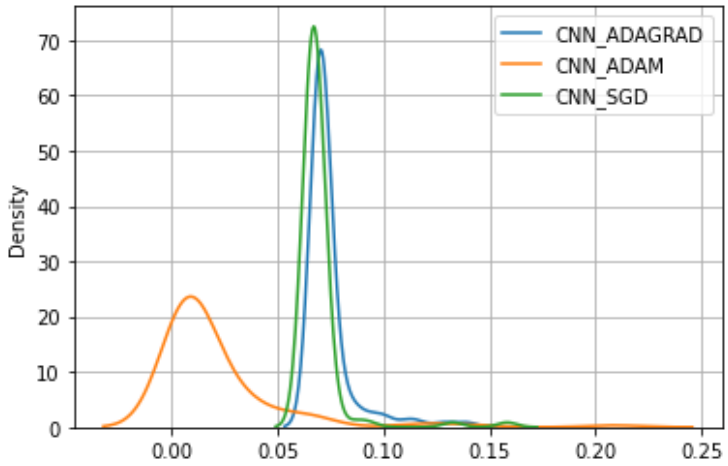


Figure 36: Loss Function Distribution against different optimizers for CNN with only product description

Table 8: Performance comparison for CNN with only product description

Optimizer	RMSE	RMLSE
Adam	12.2505	0.2543
SGD	27.4906	0.5579
Adagrad	20.9685	0.6274

5.2.3.2 CNN with all the characteristics:

To predict the price of the product considering all the characteristics of the product along with derived attributes, CNN with the input layer, embedding layer, CNN layers (Conv1D, GlobalMaxPooling, etc.), main layer and output layer with linear activation was used. The loss is evaluated using MSE with the help of a histogram. The model is evaluated against different

optimizers like adagrad, adam, sgd. Figure *37* shows the distribution plot for the loss functions. The density of adm is more near to 0 as compared with other optimizers like adagrad and sgs. Also, the Table *9* shows the RMSE and RMLSE values for different optimizers. Adam performed better as compared to other optimizers with RMSE as 7.2506 and RMLSE as 0.2115.

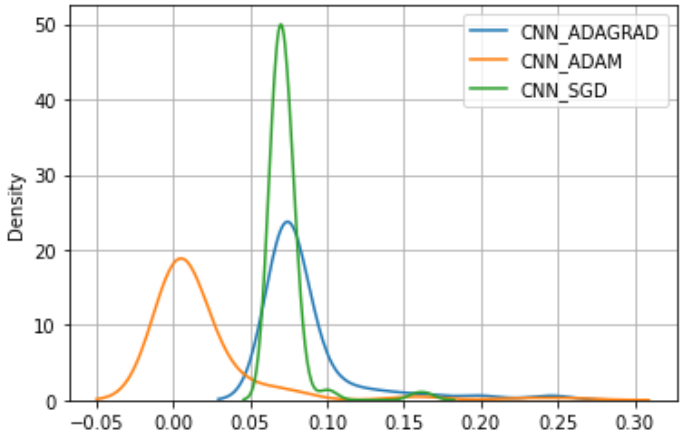


Figure 37: Loss Function Distribution against different optimizers for CNN considering the product characteristics

Table 9: Performance comparison for CNN considering all the product characteristics.

Optimizer	RMSE	RMLSE
Adam	7.2506	0.2115
SGD	19.6744	0.5479
Adagrad	20.4521	0.5761

5.3 Comparison of Performances of Machine Learning Models:

As per the analysis from previous section, it has been observed that almost all the machine learning models performed well for ‘adam’ as an optimizer. Also, the performance of the models with considering all the available product characteristics is better as compared to performance of the models considering only product description. Even though the dataset has approximately 6000 rows for the unique items, the models performed well according to the expectations and performance metrics. Table *10* shows the performance metrics for the models considering all the characteristics using ‘adam’ as an optimizer. The CNN performed best with 7.2506 as RMSE and 0.2115 and RMLSE. However, the LSTM model has also similar performance metrics as compared with CNN. LSTM has 7.7766 as RMSE and 0.2113 as RMLSE. There is very minute difference between CNN and LSTM with respect to RMSE and

RMLSE values. GRU has comparatively lower performance as compared with the other models. Thus, CNN was used for predicting and standardizing the price of the products.

Table 10: Performance Metrics Comparison Between Models

Model	RMSE	RMLSE
GRU	8.8067	0.2251
LSTM	7.7766	0.2113
CNN	7.2506	0.2115

5.4 Price Optimization using Linear Programming:

After data preparation for price optimizations, the figure 38 refers to the constraints that will be required to optimize the price of the product. Linprog module from the scipy library is used to predict the optimal prices of the products. As for each product, there will be 3 prices out of which one price will be chosen after linear programming which will be optimal in generating revenue. The first constraint refers to the price selection between three prices. Whereas the variable x is used as a flag which will have unique value 0 or 1 indicating whether the price is selected or not. This technique is called linear relaxations and is optimal to evaluate the price of the product linearly. Also, the Figure *38* refers to the optimizations results, indicating there were 9 iterations performed to select the optimal price. Also, the optimal sum for all the p1.d1 is evaluated as 832750.31. The x-values selects the optimal value between p1, p2 and p3. Also, the Figure *39* shows the sample dataset after evaluation of the optimal price using Linear Programming. This price always lies between competitor's price and the predicted price using machine learning assuring the optimal revenue overall and customer satisfaction.

$$\max \sum_{k=1}^{3} \sum_{i=1}^{n} p_{ik} x_{ik} d_{ik}$$

Subject To the constraints

$$\sum_{k=1}^{3} x_{ik} = 1 \; \forall \; i \in n$$

$$\sum_{k=1}^{3} \sum_{i=1}^{n} p_{ik} x_{ik} = c$$

$$0 \leq x_{ik} \leq 1$$

Figure 38: Constraints used for price optimization. Source: https://arxiv.org/pdf/2007.05216.pdf

```
Optimal value: 832750.31
x values: [1.00000000e+00 8.98212567e-13 2.90322731e-13 ... 3.28850963e-13
 1.16183054e-12 1.00000000e+00]
Number of iterations performed: 9
Status: Optimization terminated successfully.
```

Figure 39: Optimization Results after applying Linear Programming

	product_name	product_category_group	mrp	competitor_price	optimal_price
0	Allover Lace from Cotton Lingerie NEW! Dotted ...	panties	10.500000	32.500000	32.500000
1	Allover Lace from Cotton Lingerie NEW! Dotted ...	panties	10.500000	32.500000	10.500000
2	Body by Victoria Cheekini Panty	panties	14.500000	17.954545	16.227272
3	Body by Victoria Daisy Lace Slip	slip	52.000000	52.000000	52.000000
4	Body by Victoria Demi Bra	bras	50.617021	40.590782	40.590782
5	Body by Victoria Demi Bra	bras	45.924802	40.590782	40.590782
6	Body by Victoria Demi Bra	bras	52.979409	40.590782	40.590782
7	Body by Victoria Flutter Bandeau	bras	34.500000	40.590782	37.545391
8	Body by Victoria Front-Close Unlined Bralette	bralettes	34.948718	34.948718	34.948718
9	Body by Victoria High-neck Halter Bralette	bralettes	34.500000	34.500000	34.500000
10	Body by Victoria High-neck Halter Bralette	bralettes	34.500000	34.500000	34.500000

Figure 40: Sample data after price optimization.

5.5 Summary

The results, interpretation of models, and analysis of the whole study and their findings of the machine learning models were interpreted in this chapter. The model flow diagram and the implementation elaborated the steps taken for the building the models as well as interpreting the results in the output layer. Also, the standardized price of the products is also derived using the model with best performance. Based on this, the standardized price can be used to make informed choice on the price optimization using linear programming which will optimize the overall revenue and allow organisations to make dynamic decisions like discounts, sales, coupons, etc.

The performance metrics of the machine learning models which have been constructed using the training data sets with different optimizer techniques were interpreted in terms of their RMSE, RMLSE and MSE values. The performance table was displayed for each optimizer and the machine learning models were evaluated and the results showed that across all the 3 optimizers, ‘adam’ achieve the best performance. Also, machine learning models were evaluated for products with only descriptions and products with considering all the characteristics as well as derived attributes. The results showed that, products with only description performed poor as compared to products with all the characteristics. Moreover,

the performance of the different machine learning models like LSTM, GRU and CNN were evaluated using RMSE, RMLSE, MSE, etc. The results showed that CNN and LSTM are almost equally performed w.r.t RMSE and RMLSE values. The model based on CNN was used further for prediction of the standardized price. In addition, the standardized price was used further, and competitor's price was evaluated to choose an optimal price using Linear Programming approach. The linear programming model took 9 iterations to arrive to the optimal solutions. Finally, the results were interpreted and an expected outcome for optimal price of the products were displayed at last. Thus, a deep analysis of the machine learning models, and linear programming techniques were conducted in this chapter.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

In this chapter, the importance, overall workflow used for this research will be discussed in brief, followed by the results, conclusions and recommendations based on the results will be drawn. The discussion and conclusion will summarize the steps and methods performed such as data selection, data pre-processing and cleaning, data aggregation, parameter tuning, etc. Moreover, the results obtained from the several steps which include univariate and bivariate analysis, visualizations, development of the machine learning models and the model validation using the test dataset will be summarized. In addition, the conclusions will be drawn to portray whether the goal and aim of this study has been achieved or not. Under the contribution of this research, any new observations and findings from this study which are different from previous researchers and literature works on the similar dataset will be elaborated. The future recommendations and suggestions to researchers will include implementation of different methodologies and machine learning models on the data, considering the type of the data used or looking into the study from an eagle view to recommend ways for the improvement of the proposed model and methodology or determining a better model used in this study.

6.2 Discussion and Conclusion

This research was conducted on the publicly available dataset which was referred from Kaggle (Innerwear Dataset from Victoria's Secret and Others) consisted of information about the products and their basic characteristics like category, size, color, etc. Also, there were information about different brands available in the different files which were aggregated later. The dataset consisted of information of total of 613,143 items and approximately 6000 unique products along with 14 columns. Data pre-processing and transformation was carried out to bring the data in the required format. Data cleaning strategies like variable elimination, variable transformation, etc. were carried out. Information extraction and feature engineering was performed to derive the attributes like color popularity, size popularity, %discount, etc. Then, univariate analysis was conducted on important columns, and the results were interpreted to identify the distribution of the data as well as missing values and outliers' treatment. Multivariate analysis along with visualizations is also carried out to identify the hidden patterns in the dataset between the variables. The data pre-processing techniques on the textual columns was carried out using standard natural language techniques like stop words removal, punctuation removal, word tokenizers, sequence padding, etc. The columns were then converted using normalization, label encoding. The dataset is then split into training and test

using train test split along with K-Fold cross validation techniques with 10 folds. In addition, a hyper parameter tuning like identification of the epochs count was also conducted for improving the model better. Moreover, the data is also prepared for the price optimization stage, where the competitor's file is evaluated using the 7 other brands information in the dataset.

Next, the machine learning models used in this study viz. LSTM, GRU and CNN were evaluated, and the performance is compared against each other. The results were analysed using the performance metrics MSE, RMSE and RMLSE. According to the objective of the study, the model is prepared and evaluated using with the data consisting of only product description and the data consisting other characteristics along with the derived attributes. Also, the models were compared based on the different optimizers like 'adam', 'adagrad', 'stochastic gradient descent'. A combination of performance metrics was displayed to analyse the performance of all the models against each other. This study identified that models with 'adam' performed best amongst all other optimizers. Also, the model considering all the characteristics were always performed better instead of using the product description only. The CNN model performed best with 7.2506 as RMSE and 0.2115 and RMLSE and the similar performance is observed for LSTM with 7.7766 as RMSE and 0.2113. Thus, CNN model was used to predict the price of the product. In addition, an approach for the price optimization using linear programming was discussed to optimize the predicted price considering the maximization of revenue with few constraints. The price is optimized against the competitor's price and the optimization took around 9 iterations to evaluate the optimal price. With this approach, the organisations can make statistically proven strategical decisions on product prices using the optimal price to provide discounts, coupons, sales, etc. while the price standardization and price optimization together are helpful in accurately improve the revenue.

6.3 Contribution and Importance of the research

This research for optimal price prediction for fashion retailers yielded two results: standardizing the price using the product description and characteristics, price optimization using the statistical linear programming approach. The combination of price standardization and price optimization were not used in the earlier studies also, an attempt for predicting the price based on the product description and optimizing the price using linear programming which was used by previous researchers were combined to propose a novel approach for the price optimization. The CNN based approach to predict the prices of the products using the description was rarely used in the earlier research and this study proved the highest accuracy and efficiency of the CNN-LSTM-GRU based models for price predictions using product characteristics. Also,

combining the price prediction with the price optimization helps in maximizing the revenue which was statistically proven using the linear programming-based approach.

6.4 Limitations and Future Recommendations

Although the models performed well considering the characteristics of the products, the price prediction based only on the product descriptions needs a lot of data for accurately identifying the patterns in the description as well as the descriptions needs to be suitable and well-written for accurately identifying the prices. This study recommends the researchers to try the machine learning approach on other datasets also with large number of the valid descriptions available.

As CNN is good at handling high-dimensional data, it would be an interesting way to combine the layers of LSTM-GRU along with CNN to increase the accuracy in future. However, this won't guarantee in accurately predicting the price or improvement in accuracy.

The price optimization did a great job in optimally predicting the price, however several other approaches like dynamic programming, global optimization, inventory management, combination with recommendation systems, etc. can be tried in future. Also, optimally predicting the price is sometimes not enough, the approach can also be merged with the analysis of Customer's Lifetime Value, clickstream events, seasonal changes, etc.

REFERENCES:

Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K., (2016) Deep learning for stock prediction using numerical and textual information. In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*.

Anon (2016) Dynamic Pricing: The Future Of Retail Business. *Business Logistics in Modern Management*, 00.

Arismendi, J.C., Back, J., Prokopczuk, M., Paschke, R. and Rudolf, M., (2016) Seasonal Stochastic Volatility: Implications for the pricing of commodity options. *Journal of Banking and Finance*, 66.

Babar, M., Nguyen, P.H., Cuk, V. and Kamphuis, I.G., (2015) The development of demand elasticity model for demand response in the retail market environment. In: *2015 IEEE Eindhoven PowerTech, PowerTech 2015*.

Bakir, H., Chniti, G. and Zaher, H., (2018) E-Commerce price forecasting using LSTM neural networks. *International Journal of Machine Learning and Computing*, 82.

Ban, G.Y. and Keskin, N.B., (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 679.

Bansal, T., Belanger, D. and McCallum, A., (2016) Ask the GRU: Multi-task learning for deep text recommendations. In: *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*.

Battifarano, M. and Qian, Z. (Sean), (2019) Predicting real-time surge pricing of ride-sourcing companies. *Transportation Research Part C: Emerging Technologies*, 107.

Bauer, J. and Jannach, D., (2018) Optimal pricing in e-commerce based on sparse and noisy data. *Decision Support Systems*, 106.

Besbes, O. and Zeevi, A., (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 576.

Chatterjee, S., (2019) Big Data Analytics in e-Commerce: Understanding Personalisaion. In: *2019 2nd International Workshop on Advances in Social Sciences (IWASS 2019)*. [online] Francis Academic Press, UK, pp.201–208. Available at: https://www.webofproceedings.org/proceedings_series/ESSP/IWASS 2019/SS06029.pdf.

Chen, S.S., Choubey, B. and Singh, V., (2021) A neural network based price sensitive recommender model to predict customer choices based on price effect. *Journal of Retailing and Consumer Services*, 61.

Cheung, W.C., Simchi-Levi, D. and Wang, H., (2014) Dynamic Pricing and Demand Learning with Limited Price Experimentation. *SSRN Electronic Journal*.

Chiang, W., Liu, X., Zhang, T. and Yang, B., (2019) A Study of Exact Ridge Regression for Big Data. In: *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*.

Dadgar, S.M.H., Araghi, M.S. and Farahani, M.M., (2016) A novel text mining approach based on TF-IDF and support vector machine for news classification. In: *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*.

Dasgupta, P. and Das, R., (2000) Dynamic pricing with limited competitor information in a multi-agent economy. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Dzisevic, R. and Sesok, D., (2019) Text Classification using Different Feature Extraction Approaches. In: *2019 Open Conference of Electrical, Electronic and Information Sciences, eStream 2019 - Proceedings*.

Ensafi, Y., Amin, S.H., Zhang, G. and Shah, B., (2022) Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, 21, p.100058.

Eshan, S.C. and Hasan, M.S., (2018) An application of machine learning to detect abusive Bengali text. In: *20th International Conference of Computer and Information Technology, ICCIT 2017*.

Falode, O. and Udomboso, C., (2021) Efficient crude oil pricing using a machine learning approach. In: *Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition 2021, NAIC 2021*.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X. and Zeng, W., (2019) Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, 225.

di Fatta, D., Patton, D. and Viglia, G., (2018) The determinants of conversion rates in SME e-commerce websites. *Journal of Retailing and Consumer Services*, 41.

Ferreira, K.J., Lee, B.H.A. and Simchi-Levi, D., (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing and Service Operations Management*, 181.

Fiat, A., Mansour, Y. and Shultz, L., (2018) Flow Equilibria via Online Surge Pricing.

Ganame, K., Allaire, M., Zagdene, G. and Boudar, O., (2017) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. *First International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 10618.

Greenstein-Messica, A. and Rokach, L., (2020) Machine learning and operation research based method for promotion optimization of products with no price elasticity history. *Electronic Commerce Research and Applications*, 40.

Guo, Y., (2020) Stock Price Prediction Based on LSTM Neural Network: The Effectiveness of News Sentiment Analysis. In: *Proceedings - 2020 2nd International Conference on Economic Management and Model Engineering, ICEMME 2020*.

Gupta, R. and Pathak, C., (2014) A machine learning framework for predicting purchase by online customers based on dynamic pricing. In: *Procedia Computer Science*.

Güven, İ. and Şimşir, F., (2020) Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Computers and Industrial Engineering*, 147.

Han, L., Guo, L., Yin, Z., Tang, M., Xia, Z. and Jin, R., (2019) Vision-based price suggestion for online second-hand items. In: *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*.

Han, L., Yin, Z., Xia, Z., Tang, M. and Jin, R., (2020) Price Suggestion for Online Second-hand Items with Texts and Images. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*.

He, X. and Li, C., (2017) The Research and Application of Customer Segmentation on E-Commerce Websites. In: *Proceedings - 2016 International Conference on Digital Home, ICDH 2016*.

Heidari, M., Jones, J.H.J. and Uzuner, O., (2021) An empirical study of machine learning algorithms for social media bot detection. In: *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*.

Heitz-Spahn, S., (2013) Cross-channel free-riding consumer behavior in a multichannel environment: An investigation of shopping motives, sociodemographics and product categories. *Journal of Retailing and Consumer Services*, 206.

Herliana, S., Aina, Q., Aliya, Q.H. and Lawiyah, N., (2019) Customer loyalty factors strategy at E-Commerce Hijab Business: Frequency analysis method. *Academy of Entrepreneurship Journal*, 253.

Heuer, D., Brettel, M. and Kemper, J., (2015) Brand competition in fashion e-commerce. *Electronic Commerce Research and Applications*, 146.

van Huynh, T., van Nguyen, K., Nguyen, N.L.T. and Nguyen, A.G.T., (2020) Job Prediction: From Deep Neural Network Models to Applications. In: *Proceedings - 2020 RIVF International Conference on Computing and Communication Technologies, RIVF 2020*.

Hwangbo, H., Kim, Y.S. and Cha, K.J., (2018) Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28.

Jiang, Y., Shang, J., Liu, Y. and May, J., (2015) Redesigning promotion strategy for e-commerce competitiveness through pricing and recommendation. *International Journal of Production Economics*, 167.

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H. and Rehman, M.U., (2019) A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*, 7.

Kastius, A. and Schlosser, R., (2022) Dynamic pricing under competition using reinforcement learning. *Journal of Revenue and Pricing Management*, 211.

Kaur, J. and Kaur Buttar, P., (2018) A Systematic Review on Stopword Removal Algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, April.

Kedia, S., Jain, S. and Sharma, A., (2020) Price Optimization in Fashion E-commerce.

Kim, H. and Jeong, Y.S., (2019) Sentiment classification using Convolutional Neural Networks. *Applied Sciences (Switzerland)*, 911.

Kumar, V. and L., M., (2018) Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, 1821.

Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., Yang, Y., Gao, K., Huang, Z. and Peng, J., (2018) A Light Gradient Boosting Machine for Remainning Useful Life Estimation of Aircraft Engines. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*.

Li, T., Akiyama, T. and Wei, L., (2021) Constructing a highly accurate price prediction model in real estate investment using LightGBM.

Li, X., Shang, W. and Wang, S., (2019) Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 354.

Lin, X., Zhou, Y.W., Xie, W., Zhong, Y. and Cao, B., (2020) Pricing and Product-bundling Strategies for E-commerce Platforms with Competition. *European Journal of Operational Research*, 2833.

Liu, C. and Sustik, M.A., (2021) Elasticity Based Demand Forecasting and Price Optimization for Online Retail.

Liu, C.Z., Sheng, Y.X., Wei, Z.Q. and Yang, Y.Q., (2018) Research of Text Classification Based on Improved TF-IDF Algorithm. In: *2018 IEEE International Conference of Intelligent Robotic and Control Engineering, IRCE 2018*.

Liu, X., Zhou, Y.W., Shen, Y., Ge, C. and Jiang, J., (2021) Zooming in the impacts of merchants' participation in transformation from online flash sale to mixed sale e-commerce platform. *Information and Management*, 582.

Lopez-del Rio, A., Martin, M., Perera-Lluna, A. and Saidi, R., (2020) Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. *Scientific Reports*, 101.

Lu, R., Hong, S.H. and Zhang, X., (2018) A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Applied Energy*, 220.

Maestre, R., Duque, J., Rubio, A. and Arevalo, J., (2018) Reinforcement learning for fair dynamic pricing. In: *Advances in Intelligent Systems and Computing*.

Mehtab, S. and Sen, J., (2020) A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. *SSRN Electronic Journal*.

Minga, L.M., Feng, Y.Q. and Li, Y.J., (2003) Dynamic pricing: Ecommerce - Oriented price setting algorithm. In: *International Conference on Machine Learning and Cybernetics*.

Mohamed, M.A., El-Henawy, I.M. and Salah, A., (2022) Price prediction of seasonal items using machine learning and statistical methods. *Computers, Materials and Continua*, 702.

Mohammadi, A. and Shaverizade, A., (2021) Ensemble deep learning for aspect-based sentiment analysis. *International Journal of Nonlinear Analysis and Applications*, 12Special Issue.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D.C., (2019) Stock price prediction using news sentiment analysis. In: *Proceedings - 5th IEEE International Conference on Big Data Service and Applications, BigDataService 2019, Workshop on Big Data in Water Resources, Environment, and Hydraulic Engineering and Workshop on Medical, Healthcare, Using Big Data Technologies*.

Morsi, S., (2020) A Predictive Analytics Model for E-commerce Sales Transactions to Support Decision Making: A Case Study. *International Journal of Computer and Information Technology(2279-0764)*, 91.

Myung, R. and Yu, H., (2020) Performance prediction for convolutional neural network on spark cluster. *Electronics (Switzerland)*, 99.

Naik, J., Bisoi, R. and Dash, P.K., (2018) Prediction interval forecasting of wind speed and wind power using modes decomposition based low rank multi-kernel ridge regression. *Renewable Energy*, 129.

Narahari, Y., Raju, C.V.L., Ravikumar, K. and Shah, S., (2005) Dynamic pricing models for electronic business. *Sadhana - Academy Proceedings in Engineering Sciences*, 302–3.

Nayak, R. and Padhye, R., (2015) Introduction: The apparel industry. The apparel industry. In: *Garment Manufacturing Technology*.

Niu, X., Li, C. and Yu, X., (2017) Predictive analytics of E-commerce search behavior for conversion. In: *AMCIS 2017 - America's Conference on Information Systems: A Tradition of Innovation*.

Noaman, H.M., Sarhan, S.S. and Rashwan, M.A.A., (2018) Enhancing recurrent neural network-based language models by word tokenization. *Human-centric Computing and Information Sciences*, 81.

Otter, D.W., Medina, J.R. and Kalita, J.K., (2021) A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 322.

Pan, Y. and Liang, M., (2020) Chinese Text Sentiment Analysis Based on BI-GRU and Self-attention. In: *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*.

Pang, Z., Xiao, W. and Zhao, X., (2021) Preorder price guarantee in e-commerce. *Manufacturing and Service Operations Management*, 231.

Patil, S., Nemade, V. and Soni, P.K., (2018) Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. In: *Procedia Computer Science*.

Peng, L., Zhang, W., Wang, X. and Liang, S., (2019) Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: Considering the impact of product involvement. *Information and Management*, 562.

Pennington, J., Socher, R. and Manning, C.D., (2014) GloVe: Global vectors for word representation. In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Qaiser, S. and Ali, R., (2018) Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 1811.

Qu, T., Zhang, J.H., Chan, F.T.S., Srivastava, R.S., Tiwari, M.K. and Park, W.Y., (2017) Demand prediction and price optimization for semi-luxury supermarket segment. *Computers and Industrial Engineering*, 113.

Rai, S., Gupta, A., Anand, A., Trivedi, A. and Bhadauria, S., (2019) Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods. In: *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*.

Raju, C.V.L., Narahari, Y. and Ravikumar, K., (2003) Reinforcement learning applications in dynamic pricing of retail markets. In: *Proceedings - IEEE International Conference on E-Commerce, CEC 2003*.

Ravikumar, S. and Saraf, P., (2020) Prediction of stock prices using machine learning (regression, classification) Algorithms. In: *2020 International Conference for Emerging Technology, INCET 2020*.

Renault, T., (2020) Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, 21–2.

Rokem, A. and Kay, K., (2021) Fractional ridge regression: A fast, interpretable reparameterization of ridge regression. *GigaScience*, 912.

Roslidar, R., Saddami, K., Arnia, F., Syukri, M. and Munadi, K., (2019) A study of fine-tuning CNN models based on thermal imaging for breast cancer classification. In: *Proceedings: CYBERNETICSCOM 2019 - 2019 IEEE International Conference on Cybernetics and Computational Intelligence: Towards a Smart and Human-Centered Cyber World*.

Saha, P., Guhathakurata, S., Saha, S., Chakraborty, A. and Banerjee, J.S., (2021) Application of Machine Learning in App-Based Cab Booking System: A Survey on Indian Scenario.

Schlosser, R. and Boissier, M., (2018) Dynamic pricing under competition on online marketplaces: A data-driven approach. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Singh, V.K. and Dutta, K., (2015) Dynamic price prediction for amazon spot instances. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Smith, S.A., (2015) Clearance pricing in retail chains. In: *International Series in Operations Research and Management Science*.

Wong, T.T. and Yeh, P.Y., (2020) Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, 328.

Yahav, I., Shehory, O. and Schwartz, D., (2019) Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering*, 313.

Yan, X., Liu, X. and Zhao, X., (2020) Using machine learning for direct demand modeling of ridesourcing services in Chicago. *Journal of Transport Geography*, 83.

Ye, P., Wu, C.H., Qian, J., Zhou, Y., Chen, J., de Mars, S., Yang, F. and Zhang, L., (2018) Customized regression model for Airbnb dynamic pricing. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

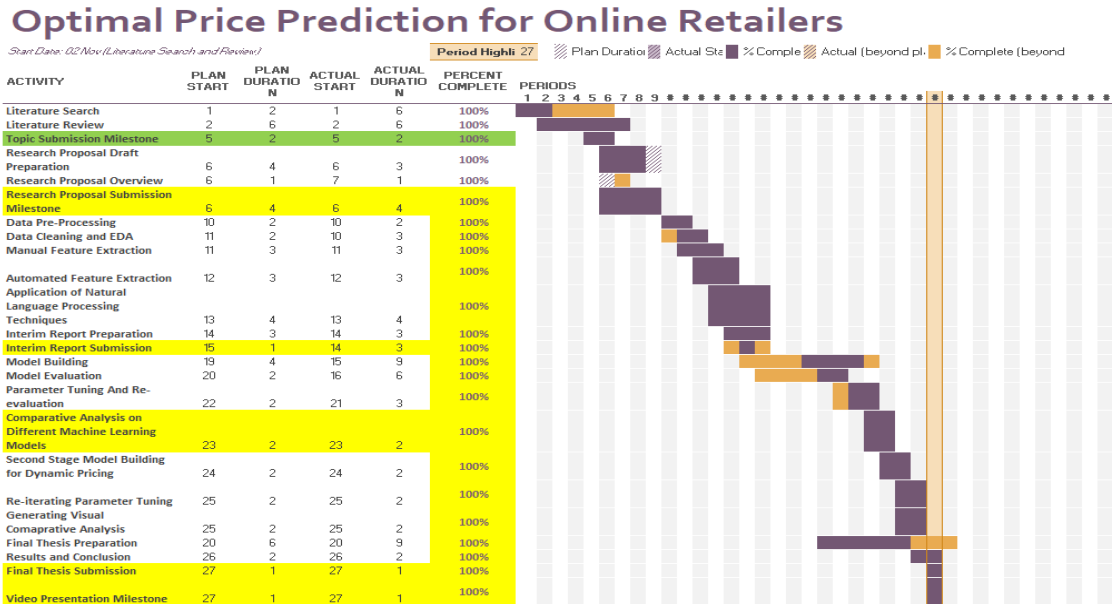
Yin, C. and Han, J., (2021) Dynamic pricing model of E-commerce platforms based on deep reinforcement learning. *CMES - Computer Modeling in Engineering and Sciences*, 1271.

Zhang, W., Kumar, D. and Ukkusuri, S. v., (2017) Exploring the dynamics of surge pricing in mobility-on-demand taxi services. In: *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*.

Zhao, Q., Zhang, Y., Friedman, D. and Tan, F., (2015) E-commerce recommendation with personalized promotion. In: *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*.

Zhou, X., Tong, W. and Li, D., (2019) Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. *ISPRS International Journal of Geo-Information*, 88.

APPENDIX A: RESEARCH PLAN



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Literature Search	1	2	1	6	100%
Literature Review	2	6	2	6	100%
Topic Submission Milestone	5	2	5	2	100%
Research Proposal Draft Preparation	6	4	6	3	100%
Research Proposal Overview	6	1	7	1	100%
Research Proposal Submission Milestone	6	4	6	4	100%
Data Pre-Processing	10	2	10	2	100%
Data Cleaning and EDA	11	2	10	3	100%
Manual Feature Extraction	11	3	11	3	100%
Automated Feature Extraction	12	3	12	3	100%
Application of Natural Language Processing Techniques	13	4	13	4	100%
Interim Report Preparation	14	3	14	3	100%
Interim Report Submission	15	1	14	3	100%
Model Building	19	4	15	9	100%
Model Evaluation	20	2	16	6	100%
Parameter Tuning And Re-evaluation	22	2	21	3	100%
Comparative Analysis on Different Machine Learning Models	23	2	23	2	100%
Second Stage Model Building for Dynamic Pricing	24	2	24	2	100%
Re-iterating Parameter Tuning	25	2	25	2	100%
Generating Visual Comaprative Analysis	25	2	25	2	100%
Final Thesis Preparation	20	6	20	9	100%
Results and Conclusion	26	2	26	2	100%
Final Thesis Submission	27	1	27	1	100%
Video Presentation Milestone	27	1	27	1	100%

Figure 41: Research Plan

APPENDIX B: RESEARCH PROPOSAL

OPTIMAL PRICE PREDICTION FOR ONLINE RETAILERS USING MACHINE LEARNING

SHARDUL RAJHANS

Research Proposal

OCTOBER 2021

Abstract

With the increasing amount of data for online e-commerce retailers, it becomes very difficult to predict the prices of each and every product that lists on the sites. There are a number of factors that majorly affects the price of the products like designs, brands, color, size, trends, season, etc. thus making the predictions difficult for the organizations. Also, with the increasing competition, it is important for an online retailer to retain their customers and attract several other customers with improving pricing strategies, improving recommendations and maintaining the balance between sales, revenue and profit. Dynamic pricing strategies like Segmented Pricing, Time-Based, Penetration, Competitive Pricing are heavily used and constantly improved along with monitoring the sales and revenue. This becomes a challenge to provide an accurate price that will be suitable for the customer and organization. This study aims on predicting the price of the product using its characteristics with minute differences. Also, competitive pricing methodology will be used to analyze the market demand and provide suitable price. Predictive analysis will be used along with Natural Language Processing techniques like TF-IDF Vectorization, Count Vectorization, etc for textual data processing. Also, machine learning models like Light Gradient Boosting Machines, Ridge Regression, Convolutional Neural Networks will be analyzed based on the performance metrics like RMSLE, MAE and R2/Adjusted R2. The Variable Importance can be helpful to the organization to analyze the importance of factors that might decide the standardized product pricing. A dynamic range will be calculated statistically using the competitors' data. Using this modelling strategy, it will help the organizations to maintain the standard prices along with the dynamic approach to think for the discounts and coupon offers. Inventory management can be considered as future application of this study. This will again, increase doors for improving recommendation-based approaches.

List of Figures

List of Abbreviations

PNR……	Promotion with No Recommendation
MCR……...	Minimum Cost Ratios
TF-IDF……	Term Frequency - Inverse Document Frequency
SGD……	Stochastic Gradient Descent
KNN……	kth Nearest Neighbour
SVM……	Support Vector Machines
LSVM…….	Linear Support Vector Machines
DT……	Decision Trees
ACC…….	Area Under the Curve
LDA…….	Linear Discriminant Analysis
LSA…….	Latent Semantic Analysis
RF…….	Random Forest
RNN…….	Recurrent Neural Networks
MLP…….	Multilayer Perceptron
LSTM……..	Long Short-Term Memory
NLP…….	Natural Language Processing
NB…….	Naive Bayes
EMD…….	Empirical Mode Decomposition
KRR…….	Kernel Ridge Regression
RVFL……	Random Vector Functional Link
ELM…….	Elaboration Likelihood Model
BTS…….	Bureau of Transportation Statistics
FAA…….	Federation Aviation Administration
RAM…….	Read Only Memory
MSE…….	Mean Squared Error
MAE…….	Mean Absolute Error
LGBM/LightGBM……	Light Gradient Boosting Machine
MAPE……	Mean Average Percentage Error
CNN…….	Convolutional Neural Networks
ARIMA……	Autoregressive Integrated Moving Average
RMSE……	Root Mean Squared Error

RMSLE…… Root Mean Squared Logarithmic Error

EDA……. Exploratory Data Analytics

CSV……. Comma Separated Files

XGBoost…… Extreme Gradient Boosting Machines

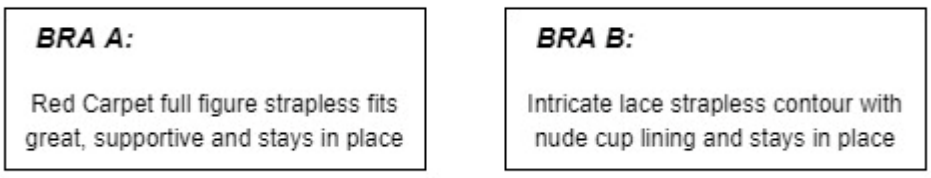
Table of Contents

1. Introduction and Background

The apparel industry is recognized as one of the most important industries that generates high revenue and contributes to the economy of the country. The main processes for clothing and garment manufacturing include: designing, manufacturing, supply chain management and retailing. Apparel manufacturing is characterized by a wide range of product designs and input materials, variable production volumes, high demand for product variety and quality, extreme competitiveness. Being a wide range of designs and styles, the complexity of the manufacturing process increases (Nayak and Padhye, 2015). The retailers are responsible for delivering products to the end consumers.

With the Volume, Velocity and Veracity (the 3Vs), companies have access to all sorts of information about customer's experiences, financial transactions, inventory management, competitiveness of the marketplace. E-Commerce companies use Artificial Intelligence and Machine Learning Models for providing personalized services to the online shoppers (Myung and Yu, 2020) . Also, there is an increase in demand for the analytics amongst online retailers. Companies like Amazon, Alibaba, etc. make use of Predictive Analytics and Descriptive Analytics to create Promotional Activities like discounted add flash sales, etc. so that large customers can get attracted.

Predicting and standardizing the price of the product is a tough challenge since almost similar products have different specifications, quality, availability, colors, sizes, etc. and thus can have different prices. For example, one of the following bras costs $80 and the other one costs $20. Can we guess which one of them costs less and better?



Although setting the price for product or an item is an old problem in e-commerce domain, there are a vast amount of pricing strategies that organisations can use depending upon the organisational goals. One organisation may target maximizing profitability for each item

sold, where as the other organisation might need access to the new markets. (Narahari et al., 2005) At the same time, the technology is allowing sellers to collect detailed data about customers' buying habits, preferences, even spending limits, so they can customize their products and prices. Multiple factors like market competition, reputation, production values, distribution costs, locality, etc. play a key role for retailers to decide the pricing strategy. Using Artificial Intelligence can be a very efficient approach for the retailers as they can benefit from the predictive models to decide the best price considering the several factors and organisation's objectives.

Dynamic pricing is a pricing strategy used by businesses to assign flexible prices for products based on current market demands. (Dynamic Pricing: The Future Of Retail Business, 2016) Typically, the seller makes use of available data about the market to determine its own pricing strategy, such as its competitor's prices, preferences of customers, buying frequencies. There has been research related to Automated Dynamic Pricing with the assumption of a little complete information about the market. (Dasgupta and Das, 2000) Often, retailers dynamically alter the prices of their product in order to match their competitor's price. A general problem of competition-based pricing is that price wars can arise resulting cyclic pattern or downward spiral. (Bauer and Jannach, 2018) Using dynamic pricing and changing the prices of the product with no objective function in the mind may lead to suboptimal results. Thus this study proposes a machine learning model on the Women's Innerwear Dataset to predict the optimal prices of the product considering different parameters. Also, this study recommends to use dynamic pricing ranges considering competitor's data jointly with the optimal pricing techniques.

2. Related Work and Problem Statement

In this section, recent related work will be reviewed to show the importance of data analytics and data science driven Price Predictions as well as Dynamic Pricing. As the dataset used for this study is very recent and there is no related paper published, this section will focus on studies that addressed Optimal Pricing and Dynamic Pricing problems, understanding the methodology from the regression, natural language processing, feature extraction and model evaluation rubrics perspectives in particular followed by highlighting the characteristics and future work scope. Finally, it will state a clear problem statement.

2.1 Business strategies for pricing used by e-commerce giants using Data Science

According to (Peng et al., 2019), e-commerce companies like Amazon, Ru La La, Tmall, etc. uses flash sales for online business platforms with limited purchase time. This is one of the promotional techniques used by the retailer business firms which increases the revenue and generate maximum profit. To test the study hypotheses, the author collected data from wjx.com. PCA was used along with varimax rotation, SPSS, multiple regression. Price Value, Functional Value, Emotional Value and Social Value are the parameters used for customer segmentation. For promoting the online sales, this study proposed social and emotional values as the important factors. This research model fails to consider the social dissemination, personality or brand effects.

The diversity in consumers' choices and the type of products and product categories make shopping behaviour difficult to understand for large stores managers (Chen et al., 2021). The customized product recommendations and discounted prices enable these small-scale stores to have loyal customers (Heitz-Spahn, 2013). According to the author (Jiang et al., 2015), suitable discounts need to be provided for the products and recommendations should be provided for the non-discounted products. Using this efficient methodology, the sales can be recovered by recommending other non-discounted products. Three models are used by the author like OPR, PNR and MCR.

Moreover, the author (Gupta and Pathak, 2014) proposed a generic framework using machine learning models to improve right price purchase and not the cheapest on e-commerce platform. The author focused on inventory led e-commerce companies however, author further stated that the model can be extended to online marketplaces without inventories. Also, this

paper proposed the adaptive pricing personalization and prediction of purchase intent as the future work of the study.

2.2 Related work in Natural Language Processing

The author (Ganame et al., 2017) proposed using of n-gram model to differentiate between fake and real news. For the identification of true and false news, the author proposed different sets of n-grams. Author used various features of the n-gram baseline established on words. Data pre-processing techniques like stemming, and stop-words removal are applied. Term Frequency – Inverse Document Frequency (TFIDF) is used in this study for extracting textual features. The author used six Machine Learning algorithms: SGD, KNN, SVM, LSVM, and DT on 3 datasets available online. This study achieved an accuracy of 87.0% for the identification of fake news using n-gram and LSVM.

For the identification of best methods which can capture textual features, the author (Dzisevic and Sesok, 2019) used three different text feature extraction approaches. TF-IDF algorithm along with its two modifications like LSA and LDA are used in this study for textual feature is used in the study using Dimensionality Reduction Techniques. The author proposed TF-IDF for extracting features outperforms other algorithms to achieve higher accuracy. The study also proposed TF-IDF along with LSA approach achieving almost equal accuracy.

Stock price prediction is always considered as a challenging task. As per the study conducted by (Guo, 2020), market information is reflected by the current price and stock prices are affected instantly according to financial market. The main purpose of this study was to predict stock price considering the related articles from news websites. News headlines and textual data is analysed to perform textual processing and sentiment analysis which is combined the sentiment score using LSTM for prediction of closing price of future stocks and current return. The author compared the results from SVM, RF, RNN, MLP and LSTM and proposed the results showing LSTM performs better with smaller percentage error.

Moreover, the author (Eshan and Hasan, 2018) has implemented a machine learning model for abusive text detection in Bengali language. Different NLP techniques like Count Vectorizer, TF-IDF Vectorizer, n-grams are used for textual data conversion in numerical vectors. SVM, RF and Multinomial NB models were compared for the performance against each other. The study shows that features collected from TF-IDF Vectorizer were better as to Count Vectorizer while working with SVM.

2.3 Related work on Prediction algorithms and Dynamic Pricing Strategies

The study conducted by (Naik et al., 2018), is analysed using Kernel Ridge Regression Model with EMD model for the prediction of Wind Speed and Power. The dataset referenced in this study is from Real Wind Farms. For each prediction at multiple intervals, the prediction error rate of KRR is less as compared to other models. It has been observed that Corelation Conversion Factor and accuracy for the Ridge Regression is highest as compared to other models like RVFL and ELM.

On the other hand, the Author (Chiang et al., 2019) has implemented Ridge Regression for over Big Data to analyse the computation time, memory requirement, and the accuracy of the output. Referenced dataset for this study is from BTS and FAA and it's quite large in size (Split into RAM-accommodable subsets). As per the objective, to predict arrival and departure of flights. Proposed solution for regression was successful with a MSE of 168632.10 and MAE of 394368.89. The study concluded that Ridge Regression takes lesser memory, performs faster in terms of computational speed and provided results are accurate which is a motivation to use Ridge Regression model for this work.

According to (Li et al., 2021), with large data and sufficient parameter tunning, it has been observed that the performance of Light Gradient Boosting Machine (LGBM) model is better than Random Forest model. This study aimed for high-accuracy price prediction model in the real estate investment market. Author used the price data of condominiums consisting 63,093 records with 108 items. The author finally proposed the Price Prediction Model based on Light GBM with 8.349% Mean Average Percentage Error (MAPE) and high accuracy.

One of the studies related to Light Gradient Boosting Machines (LGBM) along with Convolutional Neural Networks (CNN) has implemented LGBM and CNN models for prediction of power of wind (Ju et al., 2019). In this study, for the sole purpose of feature extraction from the input data, CNN is applied followed with LGBM on the inputs. This study shows that RMSE observed for CNN based model is 2.315 and for LGBM model is 2.344. This research shows how CNN is good in fetching the data as compared to LGBM as it increases robustness of the model.

The study conducted by (Ferreira et al., 2016) targeting Demand Forecasting and Price Optimization for Rue La La, shows pricing decisions taken by online retailers for monitoring

wealth of the organization. This study provided an approach using machine learning techniques for the prediction of future demand while identifying the gap of existing approach of Ru La La's pricing decision support tool. This study proposed a two-fold approach developing a demand prediction model and using this model as input into a price optimization model which resulted in increase in revenue. This study suggested future work in improving the overfitting problems and exploring less structured demand prediction models.

Also, the study conducted by (Mohamed et al., 2022) aimed for Price Prediction of Seasonal Items. This study used a dataset of a retailer who launched special sale for Christmas Event. In this study, Ridge Regression, SVM, RF and ARIMA are used evaluated against MSE, RMSE, R2 and MAPE. Also, this study proposed to consider hybrid machine learning ensemble models for improving the precision quality in the future.

2.4 Problem Statement

Examining the studies and related work in the Price Optimization and Dynamic Pricing, this study will focus on two major issues that will be a challenge for each online e-commerce retailers now-a-days if we compare the growth of data, demand and supply. This study will focus on two major issues viz Optimal Price Prediction (Price Standardization) of the Products and a statistical approach for the Dynamic Pricing (Penetration Pricing and Competitive Pricing Strategy) by using the input of Standard Price and analysing the Competitor's Data.

3. Research Questions

This study will be based on the following Research Questions and will try to understand the challenges, significance, methodology, performance, evaluations and outcome.

5. How to determine the Standardized and Accurate Price for the products considering their characteristics?
6. Which machine learning techniques can be used to optimally predict the product price with the application of Feature Extraction and Engineering?
7. How the Competitors' Data can be compared to determine the Dynamic Price Strategy of the products?

4. Aim and Objectives

The primary goal of this research study is to build a machine learning model to predict the standardized and accurate prices of the products and a statistical and strategical Dynamic Pricing approach will be proposed considering the potential risks impacting the business and revenue. This study will identify the importance of the factors that decides the standardized price of the products and evaluate the performances of different machine learning models using Feature Engineering, Extraction, Natural Language Processing and Deep Learning Techniques. This study will be divided into two major steps. At first, standardized and accurate prices of the product can be predicted. A statistical Dynamic Pricing approach can be used to predict a suitable price range for the product using the predicted standard price and competitors' prices.

The research objectives are formulated based on the aim of this study which are as follows:

- To perform Exploratory Data Analysis for different characteristics of the product for the better understanding of the price distribution and variations.
- To perform data processing and feature extraction to get the relevant information from the data.
- To compare the performance of different machine learning models for optimally predicting the price using different model evaluation rubrics.
- To predict the price ranges of the products based on competitors' data and predicted optimal price.

5. Significance of the Study

Prediction of the prices of the products is a tedious task as products may be similar and have almost negligible differences like description, color, size, material quality and demand. With the increase in data, Price Prediction gets very difficult considering the variety of products, competition, seasonal demand changes, location specific factors. Analyzing the factors affecting the product prices and optimizing the range of the prices is the most challenging tasks amongst the online retailers and also determines the direction of the growth of the organization and business revenue.

From the literature review, there are different approaches that has already been used for predicting the appropriate price ranges like Demand-Based, Cost-Based, Competitor-Based. This research would be an extension to the existing studies and aims to a two-stage approach (Standardized Price Prediction using characteristics and Competitive-Pricing Analysis) which will be helpful for the organizations. With the two-fold dynamic approach, price standardization and dynamic pricing can be achieved which might lead to significant increase in the sales and revenue. Also, knowing the Price Ranges of the products can open new doors for the appropriate recommendations.

6. Scope of the Study

Considering the primary goal of this study, the scope of the study is limited to create a two-fold approach for the prediction of price range of the product. While there have been several studies and novel methods which considered multiple influencing factors, researchers are still conducting the experiments and try multiple permutation and combination of methods to achieve the business objectives. Even, there might be several challenges and multiple factors like inventories, seasonal variations, market ups and downs, etc. that might influence the pricing strategies of the organizations, organizations have to constantly monitor, improve and try different combinations of the pricing strategies (Yin and Han, 2021). Using the two-fold optimization approach for the product pricing, the study aims contribute to one of the pricing strategies and perform predictive analytics techniques to achieve the best results.

The scope of the dataset is limited to textual processing, feature extraction and engineering, exploratory data analysis. After applying preprocessing, the choice of the machine learning approach and model evaluation for the first stage prediction can be a challenging task. Advanced methods will be performed if these methods do not generate the desired needs. For

the second stage optimization, the dynamic pricing approach is limited to use the weighted characteristics derived from the dataset and price range suggestion.

7. Research Methodology

In this section, a detailed description of dataset that will be used for this study is provided. Further steps include key processes such as data pre-processing, transforming and feature engineering, proposed model and approach, machine learning techniques and comparing performances using model evaluation rubrics. For the second-stage, detailed description for evaluation of Dynamic Pricing will be provide.

7.1 Dataset Description

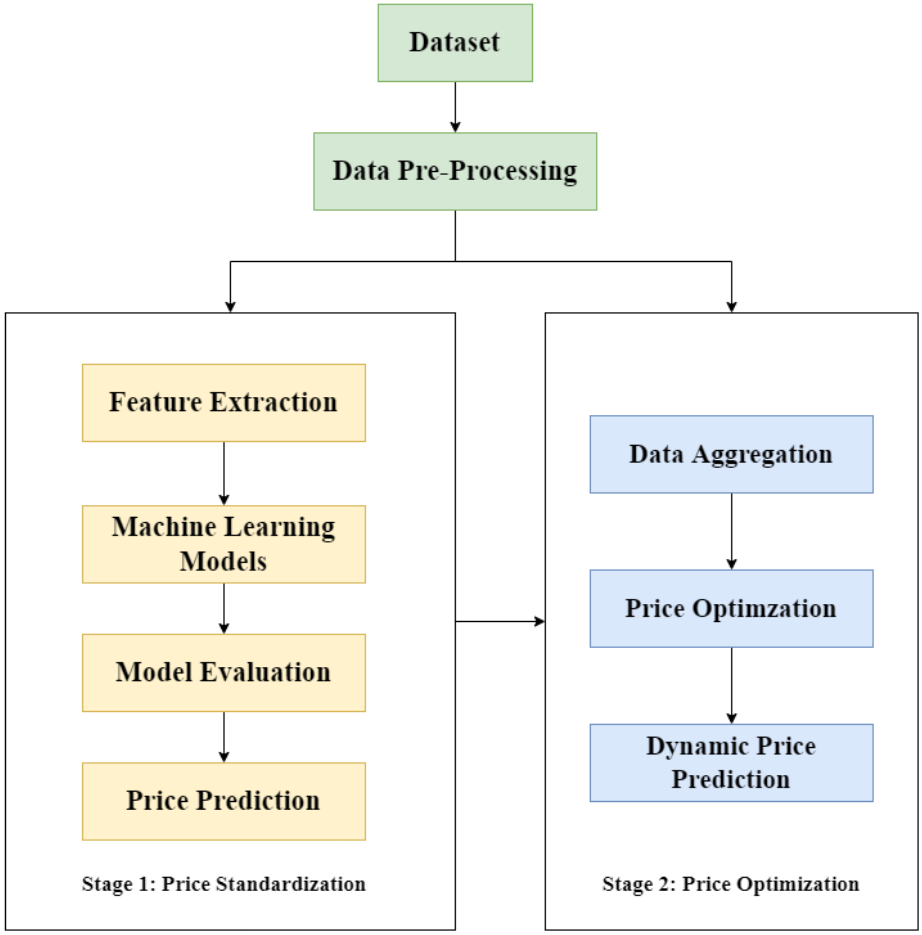
The dataset that will be used in this study is related to Women's Innerwear and Swimwear products. The data is extracted from the popular retail sites via PromptCloud's data extraction solutions from June, 2017 to July 2017 and is available publicly on Kaggle (Innerwear Data from Victoria's Secret and Others | Kaggle, 2017). This dataset has 600,000+ innerwear products information from popular retail sites such as Victoria's Secret, Amazon, Calvin Klein, Hanky Panky, etc. The overall size of the dataset is 530.25MB.

As this dataset has multiple files, the data from other retail sites can be used as the competitor's price. For the first stage of the machine learning, this study will mainly focus on the data from Victoria's Secret as this dataset has more information as well as variety of products. The datafile of Victoria's Secret is of 409 MB and has 453386 rows in total. The data is in CSV (Comma Separated Values) format. This file has 14 columns in total. The description of the relevant columns is provided below:

- **product_name:** Name of the Product
- **mrp:** Standard MRP of the product. This is the target variable.
- **price:** Discounted Price for the product.
- **product_category:** The category to which the product belongs.
- **description:** The detailed description of the product.
- **total_sizes:** The total sizes for the product.
- **available_size**: The available sizes in the inventory.
- **color**: The color of the prodct.

7.2 Proposed Design Architecture

The proposed design architecture for this study is as follows:



The proposed design architecture that will be used in this study is depicted in Figure 2. A data pre-processing is required before we feed the data for the machine learning approaches. Considering the sparsity of the data, feature extraction will be the crucial step for derived attributes and dimensionality reduction. In the first phase referred as Price Standardization, machine learning models can be used to predict the standard price of the data from the characteristics. In the second phase referred as Price Optimization, data aggregation will be performed with the existing data with derived attributes and the data from other online retailers' prices. A Variable Importance Matrix can be generated from this information and the price

range will be predicted for each of the products referred as Dynamic Price. Organizations can use the pricing strategy conducted by this study for revenue and profit growth.

7.3 Data Pre-Processing

Data pre-processing is an important step for any machine learning based approaches as the data needs to be brought in the required format. Any inconsistencies in the data can lead to undesirable and unreliable results. Data Pre-processing is an essential step to bring the data in the required format. Data cleaning and pre-processing will be used to remove the noise, missing values, etc. The columns like mrp, price, available and total sizes can be brought in the appropriate data types. Imputation techniques will be required for the columns having NA or NULL values. The missing data can be replaced with mean or mode for numerical and "MISSING" for the textual or categorical columns or even a decision to remove such data might be taken accordingly. Also, there are few columns like product_category which are categorical, thus needs to be encoded into numerical columns. For this, encoding techniques like One Hot Encoding can be performed. In case the categorical columns have too many values, Binning can be used to reduce the dimensionality. After data cleaning and data pre-processing, feature extraction techniques will be used to bring the data into required format.

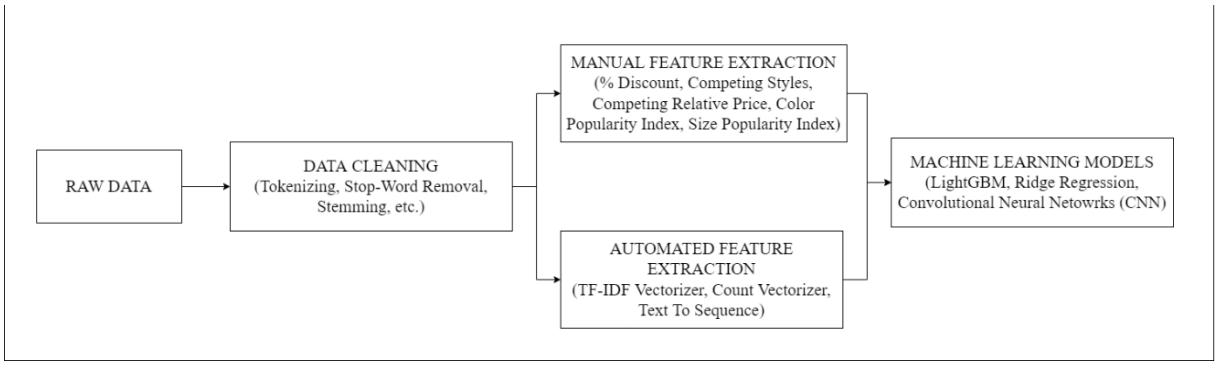
7.4 Feature Extraction and Engineering

Feature extraction will be the crucial step for this study as the data is in raw format and the information needs to be preserved. Feature Extraction refers to the process of transforming raw data into the required numerical features which can be processed and simultaneously preserving the information in the original data. It has been observed that feature extraction yields better results instead of applying machine learning techniques directly. The feature extraction can be performed manually as well as using automation (with machine learning approaches). This study aims to extract manual and automated feature extraction based on the scope of the data.

Manual feature extraction helps in capturing the relevant information based on the background and domain understanding. For the dataset used in this study, the manual feature extraction technique can be performed to derive the useful information like % discount offered, competing styles, relative competing prices, color popularity, size popularity, brand popularity. % discount can be evaluated using the formula (1 – price/mrp). Competing styles can be

calculated as percentage of the total sizes or weighted mean percentage of the available styles. Relative competing prices the price that has been used by the competitors. It can be calculated as the average of the average price (considering their mrp and discount) by different online retailers. Popularity index calculation is a challenge that can be calculated statistically using the data of revenue growth in the past years, or using relevant statistical approach by considering the demand and supply of the organization along with the revenue growth in the past years. Popularity Index is meant to capture how the demand changes with the price (Ferreira et al., 2016). For this study, popularity indexes are calculated using available information. The variation of price against the index column can be observed and the percentage popularity indexes for size and color can be evaluated.

Automated features extraction uses specific algorithms or deep networks to extract required information automatically from the data. The dataset used for the study has multiple textual columns like description, product category, etc. Natural Language Processing techniques are required to convert the raw textual data into required format. Tokenizing, Punctuation Removal, Digits Removal, Stop-Words Removal, Digits Removal, Stemming, Case Conversion can be used for processing the data. The processed data needs to be converted into the normalized format. This normalized text will be converted into numerical vectors. As per the study conducted by (Dzisevic and Sesok, 2019), TF-IDF vectorizer yields better results as compared to Count Vectorizer. Count Vectorizer can be applied to product_name and product_category columns and TF-IDF can be applied to convert the data into numerical vectors. This converted data can be fed to several machine learning models.



TF-IDF Vectorizer:

TF-IDF (Term Frequency – Inverse Document Frequency) is an algorithm used to transform text into a meaningful representation of numbers that can fit machine learning

algorithm for prediction. The TF i.e., Term Frequency gives the recurrence of the word in the whole document and can be calculated as Proportion of the number of times the word appears in the document with the total number of words. Whereas, the IDF (Inverse Data Frequency) is used to understand heaviness of uncommon words over total number of words. This study will be using TF-IDF Vectorizer because it is computationally very fast and provides best results as compared to other Vectorization techniques. Considering the hardware specifications and limitations and also according to the study conducted by (Dzisevic and Sesok, 2019), TF-IDF will give best results for the conversion of textual data into numerical vectors.

7.5 Exploratory Data Analysis

After data pre-processing and feature extraction, performing exploratory data analysis is essential to understand the existing behaviour of the data using different combinations. Univariate, Bivariate and Multivariate analysis can be performed for tuning the data and understand the behaviour like Distribution Information, Outliers Detection, Multi Collinearity to test the hypothesis. Data can be visualized using libraries like Matplotlib, Seaborn, etc.

7.6 Train-Test Split

Before passing the data to machine learning models, the data can be divided into train test split to evaluate the performance against the test model. The data of other online retailers available can also be used in the study. K-fold Cross Validation techniques can be used to estimate the skill of the data on unseen data.

7.7 Machine Learning Models

For building relationship between a dependent and one or more independent variables, machine learning models can be used. Regression analysis explains the effect of change of one target variable with respect to other dependent variables by keeping all the other characteristics constant (Ravikumar and Saraf, 2020).

After data pre-processing and feature engineering, the converted data in the required format can be fed to different machine learning models like LightGBM, Ridge Regression, Convolutional Neural Networks (CNN). The choice of machine learning algorithms is done on the basis of the Computation Time Complexity, Memory Usage, Ability to handle Big Data, Accuracy, Standard Regression Challenges like Multi-Collinearity, Data Suitability, Hardware Constraints, etc.

Light Gradient Boosting Machines (LGBM):

LGBM is the Decision Trees (DT) based ensemble algorithm and an improvised version of gradient learning. In each iteration, the residual is fitted by a negative gradient to learn a decision tree (Li et al., 2018). The growth of LGBM is vertical as compared to other tree-based algorithms. As compared to XGBoost model, speeding up of training process is achieved using LGBM as it used algorithms based on histograms, reduction in memory consumption and leaf-wise strategy with depth constraints (Fan et al., 2019).

To minimize the expected values of loss functions LGBM looks for the approximations. The key objectives to propose LGBM as a machine learning algorithm for this study are as follows:

- LGBM has high prediction accuracy, fast computational speed, minimizing overfitting issues.
- Working with high dimensional input, LGBM provides interpretable results and model is easy to understand.
- Being irreflexive towards noise and sparsity, LGBM can deal with redundant data and information.

Ridge Regression:

This study also proposes Ridge Regression for the machine learning model performance and comparison analysis. Ridge Regression is a machine learning algorithm where degree of bias is added to the regression estimates thus reducing the standard errors. This machine learning algorithm is preferred when the data suffer from multicollinearity. Following the regression equation used by Ridge Regression:

$$\underline{\mathbf{Y}} = \mathbf{X}\underline{\mathbf{B}} + \underline{\mathbf{e}}$$

where X and $\underline{\mathbf{Y}}$ are dependents and independent variables respectively. $\underline{\mathbf{B}}$ is the regression coefficients to be estimated and $\underline{\mathbf{e}}$ represents the residuals. (Rokem and Kay, 2021) Standardization of the data is performed before Ridge Regression as all Ridge Regression calculations are based on standardized variables. Ridge Regression uses corelation matrix of independent variables which are unbiased so that the population value is the expected value. Ridge Regression provided penalty terms for the regression coefficients which is helpful in adjusting the multicollinearity and variance-bias trade off. The key objectives to propose Ridge Regression as a machine learning algorithm for this study is as follows:

- Ridge Regression shrinks the parameters hence it is used to prevent multicollinearity.

- Model Complexity is reduced easily using Coefficient Shrinkage.

Convolutional Neural Networks (CNN):

Convolutional Neural Networks (CNN) are the most promising choice of neural networks in developing machine learning models. It performs very well specially in image classification and computer vision. Convolutional neural network is a special tool for modelling as it consists of features like detecting edges, corners and various textures. It goes through every corner, vector and dimension of the pixel matrix. Recently, the Convolutional Neural Network (CNN) has been adopted for the task of text classification and has shown quite successful results (Kim and Jeong, 2019). Using the information of the pixel matrix, CNN is quite sustainable to for data of matrix form. As description column is in the textual format, considering CNN for a textual layer is the similar idea of working with image. Textual data can be converted into a sequential data like the data in time series and interpreted as 1-D matrix. CNN can work with 1-Dimensional matrix convolutional layer and the modified data type. The key objectives to propose CNN as a machine learning algorithm for this study is as follows:

- Even though CNN is usually used for image modelling, it can also learn from connection between words leading to a better choice for NLP based datasets. It provided quite accurate results working with the data in matrix format and convolutional layers.

7.8 Model Evaluation Metrics

The model can be evaluated against model evaluation metrics like RMSLE, MAE and R2 /Adjusted R2.

RMSLE is the calculated as the logarithmic root mean squared difference of the predicted and actual value. It's quite similar to RMSE except for the case, it incurs larger penalty for the underestimation of the actual value than overestimation. As per the business case, predicting lesser value is less tolerable than the predicted the higher value as pretty much lesser value can lead to the downfall of the sales revenue and predicting very higher value can lead to the downfall of the profit. This makes RMSLE as the best choice for the model evaluation metric. It can be calculated as follows:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

R-Square or R2 determines how good the model fits the dependent variables. It's evaluated unexplained variance divided by total variance. However, R2 the overfitting problem

is not taken into consideration. This is introduced in Adjusted R2 because of its penalizing nature of additional independent variables.

MAE is quite equivalent to MSE with just the difference that takes sum of absolute value of error. MAE is the direct representation of the sum of error terms and treats all the error terms the same. MAE is used as an evaluation metric for this study to understand the performances of the model over all the available datafiles.

7.9 Stage-Two: Dynamic Price Prediction:

After prediction of standardized price in the first stage, this predicted price can be used for the statistical calculation of the dynamic price. Data Aggregation will be the first stage to be performed to have the data from other online retailers processed together.

The primary aim to use dynamic price prediction is to understand the Variable Importance of the Price Predictor's of the aggregated data and to predict a suitable price range which the organizations can use to provide efficient discounts keeping the standardized price in mind. Variable Importance is the evaluation of how effective the dependent variable is with respect to independent variables. Identification of such factors which contributes towards Price of the Product can be evaluated using Variable Importance from the machine learning model. According to the study conducted by (Ferreira et al., 2016) , the author used a regression trees model with bagging and identified the factors that can be responsible for the demand prediction of the product.

To identify the variable importance, this study will focus on the model built using gradient boosting (XGBoost). The benefit of using ensembles of decision tree methods like gradient boosting is Variable Importance can be easily identified from a trained predictive model. Variable Importance will predict the score and will indicate how useful and valuable a feature is in predicting price. For the price range calculation, based on the importance of Variables and competitor's price, a price range can be evaluated using the relative competitor's price and predicted price. Based on variable importance, a weightage can be assigned, and discounts can be provided.

7.10 Expected Outcomes

As per the primary aims and goals of the study, following are the expected outcomes.

- A machine learning model that can be used to predict the prices of the product using the characteristics with evaluation rubrics.

- Variable Importance and Price Range evaluations that can be useful for the organizations to provide discounts and analysis of business sales and revenue.

8. Requirements Resources

In this study, all the required computations tasks will be performed on the personal computer. Following are the resource requirements that will be minimal from the implementation perspective.

Hardware Specifications:

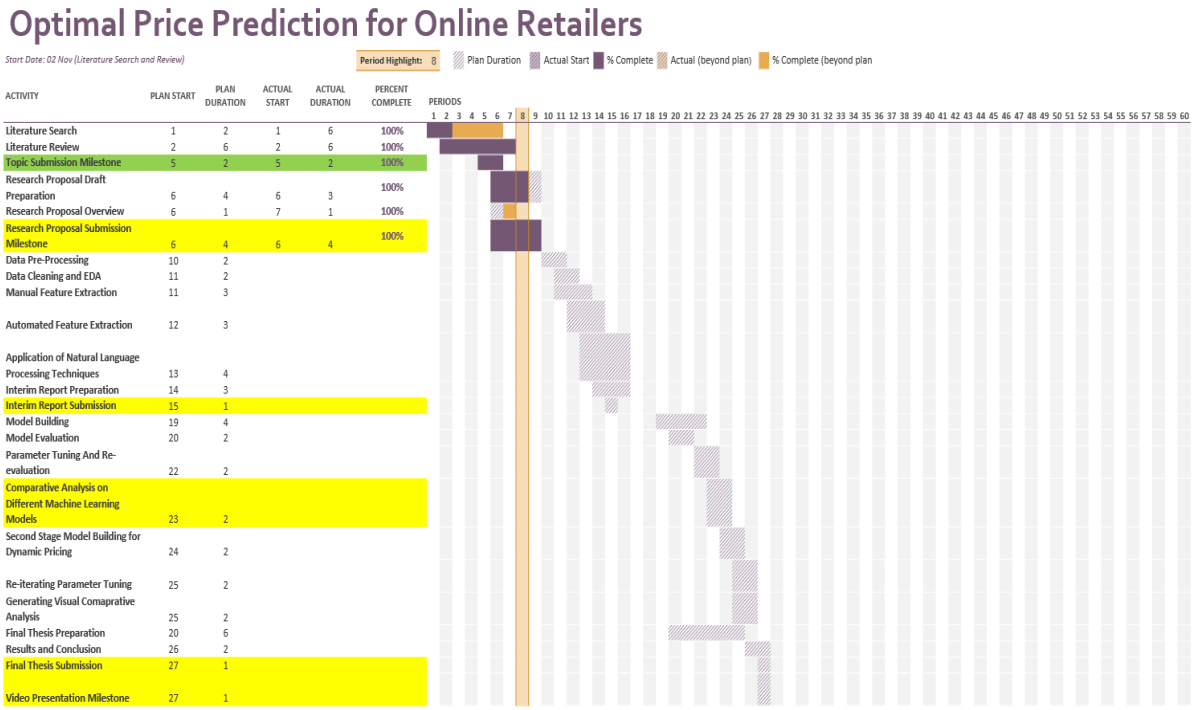
- Processor: Intel Core i5 with 8 GB DDR4 Memory.
- GPU: The processing speed of 1.6 GHz
- Operating System: Windows 10 64-bit

Software Resources:

The Python framework will be used to implement the Prediction model. For data processing, matplotlib, seaborn for EDA, and sklearn for feature extraction, scaling, predictive modelling, and assessment, keras library for performing CNN based operations, open-source libraries such as pandas and will be used.

9. Research Plan

The research plan for this study is been projected with the help of Gnatt chart which is shown below.



References:

Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K., (2016) Deep learning for stock prediction using numerical and textual information. In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*.

Anon (2016) Dynamic Pricing: The Future Of Retail Business. *Business Logistics in Modern Management*, 00.

Arismendi, J.C., Back, J., Prokopczuk, M., Paschke, R. and Rudolf, M., (2016) Seasonal Stochastic Volatility: Implications for the pricing of commodity options. *Journal of Banking and Finance*, 66.

Babar, M., Nguyen, P.H., Cuk, V. and Kamphuis, I.G., (2015) The development of demand elasticity model for demand response in the retail market environment. In: *2015 IEEE Eindhoven PowerTech, PowerTech 2015*.

Bakir, H., Chniti, G. and Zaher, H., (2018) E-Commerce price forecasting using LSTM neural networks. *International Journal of Machine Learning and Computing*, 82.

Ban, G.Y. and Keskin, N.B., (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 679.

Bansal, T., Belanger, D. and McCallum, A., (2016) Ask the GRU: Multi-task learning for deep text recommendations. In: *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*.

Battifarano, M. and Qian, Z. (Sean), (2019) Predicting real-time surge pricing of ride-sourcing companies. *Transportation Research Part C: Emerging Technologies*, 107.

Bauer, J. and Jannach, D., (2018) Optimal pricing in e-commerce based on sparse and noisy data. *Decision Support Systems*, 106.

Besbes, O. and Zeevi, A., (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 576.

Chatterjee, S., (2019) Big Data Analytics in e-Commerce: Understanding Personalisaion. In: *2019 2nd International Workshop on Advances in Social Sciences (IWASS 2019)*. [online] Francis Academic Press, UK, pp.201–208. Available at: https://www.webofproceedings.org/proceedings_series/ESSP/IWASS 2019/SS06029.pdf.

Chen, S.S., Choubey, B. and Singh, V., (2021) A neural network based price sensitive recommender model to predict customer choices based on price effect. *Journal of Retailing and Consumer Services*, 61.

Cheung, W.C., Simchi-Levi, D. and Wang, H., (2014) Dynamic Pricing and Demand Learning with Limited Price Experimentation. *SSRN Electronic Journal*.

Chiang, W., Liu, X., Zhang, T. and Yang, B., (2019) A Study of Exact Ridge Regression for Big Data. In: *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*.

Dadgar, S.M.H., Araghi, M.S. and Farahani, M.M., (2016) A novel text mining approach based on TF-IDF and support vector machine for news classification. In: *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*.

Dasgupta, P. and Das, R., (2000) Dynamic pricing with limited competitor information in a multi-agent economy. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Dzisevic, R. and Sesok, D., (2019) Text Classification using Different Feature Extraction Approaches. In: *2019 Open Conference of Electrical, Electronic and Information Sciences, eStream 2019 - Proceedings*.

Ensafi, Y., Amin, S.H., Zhang, G. and Shah, B., (2022) Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, 21, p.100058.

Eshan, S.C. and Hasan, M.S., (2018) An application of machine learning to detect abusive Bengali text. In: *20th International Conference of Computer and Information Technology, ICCIT 2017*.

Falode, O. and Udomboso, C., (2021) Efficient crude oil pricing using a machine learning approach. In: *Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition 2021, NAIC 2021*.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X. and Zeng, W., (2019) Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, 225.

di Fatta, D., Patton, D. and Viglia, G., (2018) The determinants of conversion rates in SME e-commerce websites. *Journal of Retailing and Consumer Services*, 41.

Ferreira, K.J., Lee, B.H.A. and Simchi-Levi, D., (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing and Service Operations Management*, 181.

Fiat, A., Mansour, Y. and Shultz, L., (2018) Flow Equilibria via Online Surge Pricing.

Ganame, K., Allaire, M., Zagdene, G. and Boudar, O., (2017) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. *First International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 10618.

Greenstein-Messica, A. and Rokach, L., (2020) Machine learning and operation research based method for promotion optimization of products with no price elasticity history. *Electronic Commerce Research and Applications*, 40.

Guo, Y., (2020) Stock Price Prediction Based on LSTM Neural Network: The Effectiveness of News Sentiment Analysis. In: *Proceedings - 2020 2nd International Conference on Economic Management and Model Engineering, ICEMME 2020*.

Gupta, R. and Pathak, C., (2014) A machine learning framework for predicting purchase by online customers based on dynamic pricing. In: *Procedia Computer Science*.

Güven, İ. and Şimşir, F., (2020) Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Computers and Industrial Engineering*, 147.

Han, L., Guo, L., Yin, Z., Tang, M., Xia, Z. and Jin, R., (2019) Vision-based price suggestion for online second-hand items. In: *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*.

Han, L., Yin, Z., Xia, Z., Tang, M. and Jin, R., (2020) Price Suggestion for Online Second-hand Items with Texts and Images. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*.

He, X. and Li, C., (2017) The Research and Application of Customer Segmentation on E-Commerce Websites. In: *Proceedings - 2016 International Conference on Digital Home, ICDH 2016*.

Heidari, M., Jones, J.H.J. and Uzuner, O., (2021) An empirical study of machine learning algorithms for social media bot detection. In: *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*.

Heitz-Spahn, S., (2013) Cross-channel free-riding consumer behavior in a multichannel environment: An investigation of shopping motives, sociodemographics and product categories. *Journal of Retailing and Consumer Services*, 206.

Herliana, S., Aina, Q., Aliya, Q.H. and Lawiyah, N., (2019) Customer loyalty factors strategy at E-Commerce Hijab Business: Frequency analysis method. *Academy of Entrepreneurship Journal*, 253.

Heuer, D., Brettel, M. and Kemper, J., (2015) Brand competition in fashion e-commerce. *Electronic Commerce Research and Applications*, 146.

van Huynh, T., van Nguyen, K., Nguyen, N.L.T. and Nguyen, A.G.T., (2020) Job Prediction: From Deep Neural Network Models to Applications. In: *Proceedings - 2020 RIVF International Conference on Computing and Communication Technologies, RIVF 2020*.

Hwangbo, H., Kim, Y.S. and Cha, K.J., (2018) Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28.

Jiang, Y., Shang, J., Liu, Y. and May, J., (2015) Redesigning promotion strategy for e-commerce competitiveness through pricing and recommendation. *International Journal of Production Economics*, 167.

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H. and Rehman, M.U., (2019) A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*, 7.

Kastius, A. and Schlosser, R., (2022) Dynamic pricing under competition using reinforcement learning. *Journal of Revenue and Pricing Management*, 211.

Kaur, J. and Kaur Buttar, P., (2018) A Systematic Review on Stopword Removal Algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, April.

Kedia, S., Jain, S. and Sharma, A., (2020) Price Optimization in Fashion E-commerce.

Kim, H. and Jeong, Y.S., (2019) Sentiment classification using Convolutional Neural Networks. *Applied Sciences (Switzerland)*, 911.

Kumar, V. and L., M., (2018) Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, 1821.

Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., Yang, Y., Gao, K., Huang, Z. and Peng, J., (2018) A Light Gradient Boosting Machine for Remainning Useful Life Estimation of Aircraft Engines. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*.

Li, T., Akiyama, T. and Wei, L., (2021) Constructing a highly accurate price prediction model in real estate investment using LightGBM.

Li, X., Shang, W. and Wang, S., (2019) Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 354.

Lin, X., Zhou, Y.W., Xie, W., Zhong, Y. and Cao, B., (2020) Pricing and Product-bundling Strategies for E-commerce Platforms with Competition. *European Journal of Operational Research*, 2833.

Liu, C. and Sustik, M.A., (2021) Elasticity Based Demand Forecasting and Price Optimization for Online Retail.

Liu, C.Z., Sheng, Y.X., Wei, Z.Q. and Yang, Y.Q., (2018) Research of Text Classification Based on Improved TF-IDF Algorithm. In: *2018 IEEE International Conference of Intelligent Robotic and Control Engineering, IRCE 2018*.

Liu, X., Zhou, Y.W., Shen, Y., Ge, C. and Jiang, J., (2021) Zooming in the impacts of merchants' participation in transformation from online flash sale to mixed sale e-commerce platform. *Information and Management*, 582.

Lopez-del Rio, A., Martin, M., Perera-Lluna, A. and Saidi, R., (2020) Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. *Scientific Reports*, 101.

Lu, R., Hong, S.H. and Zhang, X., (2018) A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Applied Energy*, 220.

Maestre, R., Duque, J., Rubio, A. and Arevalo, J., (2018) Reinforcement learning for fair dynamic pricing. In: *Advances in Intelligent Systems and Computing*.

Mehtab, S. and Sen, J., (2020) A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. *SSRN Electronic Journal*.

Minga, L.M., Feng, Y.Q. and Li, Y.J., (2003) Dynamic pricing: Ecommerce - Oriented price setting algorithm. In: *International Conference on Machine Learning and Cybernetics*.

Mohamed, M.A., El-Henawy, I.M. and Salah, A., (2022) Price prediction of seasonal items using machine learning and statistical methods. *Computers, Materials and Continua*, 702.

Mohammadi, A. and Shaverizade, A., (2021) Ensemble deep learning for aspect-based sentiment analysis. *International Journal of Nonlinear Analysis and Applications*, 12Special Issue.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D.C., (2019) Stock price prediction using news sentiment analysis. In: *Proceedings - 5th IEEE International Conference on Big Data Service and Applications, BigDataService 2019, Workshop on Big Data in Water Resources, Environment, and Hydraulic Engineering and Workshop on Medical, Healthcare, Using Big Data Technologies*.

Morsi, S., (2020) A Predictive Analytics Model for E-commerce Sales Transactions to Support Decision Making: A Case Study. *International Journal of Computer and Information Technology(2279-0764)*, 91.

Myung, R. and Yu, H., (2020) Performance prediction for convolutional neural network on spark cluster. *Electronics (Switzerland)*, 99.

Naik, J., Bisoi, R. and Dash, P.K., (2018) Prediction interval forecasting of wind speed and wind power using modes decomposition based low rank multi-kernel ridge regression. *Renewable Energy*, 129.

Narahari, Y., Raju, C.V.L., Ravikumar, K. and Shah, S., (2005) Dynamic pricing models for electronic business. *Sadhana - Academy Proceedings in Engineering Sciences*, 302–3.

Nayak, R. and Padhye, R., (2015) Introduction: The apparel industry. The apparel industry. In: *Garment Manufacturing Technology*.

Niu, X., Li, C. and Yu, X., (2017) Predictive analytics of E-commerce search behavior for conversion. In: *AMCIS 2017 - America's Conference on Information Systems: A Tradition of Innovation*.

Noaman, H.M., Sarhan, S.S. and Rashwan, M.A.A., (2018) Enhancing recurrent neural network-based language models by word tokenization. *Human-centric Computing and Information Sciences*, 81.

Otter, D.W., Medina, J.R. and Kalita, J.K., (2021) A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 322.

Pan, Y. and Liang, M., (2020) Chinese Text Sentiment Analysis Based on BI-GRU and Self-attention. In: *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*.

Pang, Z., Xiao, W. and Zhao, X., (2021) Preorder price guarantee in e-commerce. *Manufacturing and Service Operations Management*, 231.

Patil, S., Nemade, V. and Soni, P.K., (2018) Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. In: *Procedia Computer Science*.

Peng, L., Zhang, W., Wang, X. and Liang, S., (2019) Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: Considering the impact of product involvement. *Information and Management*, 562.

Pennington, J., Socher, R. and Manning, C.D., (2014) GloVe: Global vectors for word representation. In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Qaiser, S. and Ali, R., (2018) Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 1811.

Qu, T., Zhang, J.H., Chan, F.T.S., Srivastava, R.S., Tiwari, M.K. and Park, W.Y., (2017) Demand prediction and price optimization for semi-luxury supermarket segment. *Computers and Industrial Engineering*, 113.

Rai, S., Gupta, A., Anand, A., Trivedi, A. and Bhadauria, S., (2019) Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods. In: *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*.

Raju, C.V.L., Narahari, Y. and Ravikumar, K., (2003) Reinforcement learning applications in dynamic pricing of retail markets. In: *Proceedings - IEEE International Conference on E-Commerce, CEC 2003*.

Ravikumar, S. and Saraf, P., (2020) Prediction of stock prices using machine learning (regression, classification) Algorithms. In: *2020 International Conference for Emerging Technology, INCET 2020*.

Renault, T., (2020) Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, 21–2.

Rokem, A. and Kay, K., (2021) Fractional ridge regression: A fast, interpretable reparameterization of ridge regression. *GigaScience*, 912.

Roslidar, R., Saddami, K., Arnia, F., Syukri, M. and Munadi, K., (2019) A study of fine-tuning CNN models based on thermal imaging for breast cancer classification. In: *Proceedings: CYBERNETICSCOM 2019 - 2019 IEEE International Conference on Cybernetics and Computational Intelligence: Towards a Smart and Human-Centered Cyber World*.

Saha, P., Guhathakurata, S., Saha, S., Chakraborty, A. and Banerjee, J.S., (2021) Application of Machine Learning in App-Based Cab Booking System: A Survey on Indian Scenario.

Schlosser, R. and Boissier, M., (2018) Dynamic pricing under competition on online marketplaces: A data-driven approach. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Singh, V.K. and Dutta, K., (2015) Dynamic price prediction for amazon spot instances. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Smith, S.A., (2015) Clearance pricing in retail chains. In: *International Series in Operations Research and Management Science*.

Wong, T.T. and Yeh, P.Y., (2020) Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, 328.

Yahav, I., Shehory, O. and Schwartz, D., (2019) Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering*, 313.

Yan, X., Liu, X. and Zhao, X., (2020) Using machine learning for direct demand modeling of ridesourcing services in Chicago. *Journal of Transport Geography*, 83.

Ye, P., Wu, C.H., Qian, J., Zhou, Y., Chen, J., de Mars, S., Yang, F. and Zhang, L., (2018) Customized regression model for Airbnb dynamic pricing. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Yin, C. and Han, J., (2021) Dynamic pricing model of E-commerce platforms based on deep reinforcement learning. *CMES - Computer Modeling in Engineering and Sciences*, 1271.

Zhang, W., Kumar, D. and Ukkusuri, S. v., (2017) Exploring the dynamics of surge pricing in mobility-on-demand taxi services. In: *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*.

Zhao, Q., Zhang, Y., Friedman, D. and Tan, F., (2015) E-commerce recommendation with personalized promotion. In: *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*.

Zhou, X., Tong, W. and Li, D., (2019) Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. *ISPRS International Journal of Geo-Information*, 88.