Price Suggestion and Recommendation of Resale Products on E-commerce Websites

MSc Research Project
Data Analytics

Sanchit Pereira
Student ID: x18104002

School of Computing
National College of Ireland

Supervisor: Prof. Manuel Tova-Izquierdo

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sanchit Pereira
Student ID:	x18104002
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Prof. Manuel Tova-Izquierdo
Submission Due Date:	12/08/2019
Project Title:	Price Suggestion and Recommendation of Resale Products on E-commerce Websites
Word Count:	7591
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Price Suggestion and Recommendation of Resale Products on E-commerce Websites

Sanchit Pereira
x18104002

Abstract

E-commerce websites implements recommendation systems for recommending products to the customer and improve their sales. This can be effectively implemented if the website has new brand products, since new brand products have standardized pricing all over the e-commerce platform. But, there are also some e-commerce website which only sell resale products, the products sold on these websites are coming from sellers who have already used that product. So the pricing on these products are not standardized as prices of these products is decided by the seller of those resale products. Since only resale products are sold on these websites, such websites do not have all the products and therefore there is no recommendation system on these websites. The aim of this paper is to predict the prices of the product by considering the characteristics of the product and make the price standardized for all the products. After standardizing all the prices on the e-commerce website, a market basket based recommendation system is built for such e-commerce website. The algorithms used for predicting the prices are Light Gradient Boosting Machine(LGBM), Ridge Regression and Convolutional Neural Network (CNN). CNN outperformed LGBM and ridge regression for predicting the price, the root mean squared logarithmic error value (rmsle) of CNN was 0.44, while LGBM was at 0.53 and ridge regression was at 0.47. Apriori algorithm is used for recommendation system as association rules are implemented for market basket analysis. For generating the association rules the support value used is 0.001, while the value of confidence is 0.90.

1 Introduction

E-commerce is playing an important role in today's life. Why should one walk to a retail store when you can buy things while sitting at home? This mentality has made e-commerce popular. E-commerce companies make use of promotional activities like flash sale, discounted sales, etc. to attract more customers towards them. When a person visits their website for buying some products they recommend some other product which will be compatible to the product being purchased. This makes the customer aware about the other products on the e-commerce website. In this way, sales of the e-commerce companies are increased, this is the main reason why recommendation systems play an important role for an e-commerce websites. Also companies like Amazon, Alibaba which sell brand new products have standard pricing across all the products, so customers show interest in buying the recommended product as the pricing all across the other websites and retail stores is also same. This is not the case with the e-commerce websites selling resale products.

In companies like OLX, Mercari, etc. who sale resale products do not have standardized pricing across all products as the products being sold on their websites are not coming from a manufacturer but instead it is coming from a person who has already used it. When a person is selling a product he/she will always price the product on what he/she thinks is right. Due to this some of the products being sold on such resale websites are overpriced as the seller does not consider all the characteristics of the product. Looking from a customer's point of view, why will a customer purchase a product which is overpriced. For example, suppose a samsung galaxy s7 edge is being sold by the manufacturer at 1500 euros, while a six months used samsung galaxy s7 edge is being sold by a seller on OLX at 1400 euros. Now in this case the customer will always opt for a new brand samsung galaxy s7 edge as it is only 100 euros more than the resale samsung galaxy s7 edge. This creates a loss to the e-commerce companies selling resale products as they are not able to get their profit until the products on their website is being sold. Moreover, since such websites are not having standardized pricing they are not able to recommend products to the customers. To overcome these obstacles of the e-commerce websites selling resale products this paper implements a price prediction algorithm for standardizing the pricing of the resale products and a market basket based recommendation model which will recommend those standardized price products.

Coming to prediction algorithm this paper implements three models namely light gradient boosting machine (LGBM), ridge regression and convolutional neural network (CNN). The data used for predicting the price is coming from the website of kaggle which is availed by Mercari one of the leading e-commerce website in Japan. The data consists of six features which are as follows, name of the product, item condition id, brand name, categories, shipping and item description. Price is the target variable which will be predicted based on the above mentioned features. The data consists of 1482535 rows. After performing exploratory data analysis it was realized that item condition id and shipping barely play any role in predicting the price. Therefore name of the product, brand name, categories and item description were the key parameters for predicting the price. Name of the product, brand name, categories and item description were textual features, natural language processing techniques like count vectorizer, tfidfvectorizer, ngrams, text to sequences, etc were used to convert those textual data into numerical format. Since price is a continuous variable the metrics used for evaluation is root mean squared logarithmic error(rmsle), mean absolute error (mae) and r squared. Since the values of the price were skewed towards right, logarithm function was the best which could be applied to get those continuous values into normally distributed form. R squared metrics helps to know what percent of dependent variable is explained by the independent variables. The performance of CNN was better than LGBM and ridge regression for predicting the price. Now lets move to recommendation system.

The main reason for using market basket analysis for recommendation was that e-commerce companies selling resale products only have those products which are availed by the sellers and not by manufacturers that is why they can only recommend those products which they have in their stock. For example if a customer is interested in buying a laptop and if the e-commerce website have a laptop cover for resale then it can be recommended to the customer buying the laptop. Moreover due to lack of new brand products content based filtering(storing historical data of users for recommending future products) and collaborative filtering(recommending those products which have good reviews by other customers) won't work here. For recommendation, data from kaggle was used which was availed by instakart. The data had total of 6 files namely orders, products, department, aisles, order products prior, order products train. All these files had almost 40000 rows of data. There were a total of 8390 items in the dataset. Apriori algorithm was used for applying the association rules. The evaluation metrics used were support, confidence and lift.

1.1 Research Question

"How efficiently can machine learning optimize the sale of resale products using recommendation and prediction models?"

2 Related Work

2.1 Machine learning in the domain of E-Commerce

Author Greenstein-Messica and Rokach (2018) has implemented price prediction model for resale products on website of ebay. According to the author reputation of the seller, promotion indication and price of the product are important properties which can improve the efficiency of the prediction model. Author has used transactional data of six months from ebay for implementing the model. The model used for price prediction is content-aware price factorization (CAMF). The sellers reputation feature boosted the f1-score of the model to 84% which was pretty high as compared to the matrix factorization recommendation model implemented by author. As pricing in e-commerce is very dynamic, author Bauer and Jannach (2018) has implemented a model to optimize the pricing of sparse and noisy data of e-commerce. For implementing this the author has made use of two parameters, price changes over the years for a particular product and for related products. Bootstrap based confidence is combined with confidence Bayesian inference and kernel regression. The model proposed by the author was successful, the profit increased by 28.04% of revenue in just 4 months. On other hand, author (Dai et al.; 2018) has implemented a model for building customer's trust and increase sales of the e-commerce website. According to the author the decision of a customer of purchasing a product primarily depends on the reputation of the seller, followed by profile photos of the seller and stake of the transaction. Author Guo et al. (2018) has created a model for recommending products by considering the price and multi-category inter-purchase time. The author has used sequential pattern mining for understanding the customers change in interest for different products in a certain time interval. The author has divided the implementation in two stages, category search stage and product search stage. In category search stage, sequential pattern mining is done for all the categories the user is interested in and this is done for all the users. The prices of each categories are then modelled by matching the purchase sequence of the pattern for each category and for multiple categories inter purchase time interval. In product search at a recommended time a time factor is computed for a candidate product. In a similar way, a price factor is calculated for the candidate product. The matching time and matching price is then integrated to obtain a preference value for the user for that particular product. A fuzzy set theory is employed for getting the results of this implementation.

Author Peng et al. (2019) explains how time influences a purchase intention of the customer. The author says that e-commerce giants like Amazon, Gilt, etc. employ flash sales on their websites offering high discounts for limited time period. This creates pressure on the customers to buy the products instead of regretting later. This is one of the promotional strategies employed by the e-commerce websites which increases the sales of their websites. To study this, data from China's e-commerce giant wjx.com was used. Principle Component Analysis(PCA) was implemented along with varimax rotation, SPSS and then for getting the output multiple regression was used. To understand the value of time pressure on the customers the author has used 4 parameters emotional value, functional value, price value and social value. The results show that for promotion of online sales social value and emotional value are two important parameters. The second finding of the author illustrated that time pressure and the correlation between social value, emotional value and purchase intention play a negative role in e-commerce environments. On the other hand using recommendation and pricing as the competitiveness parameter author Jiang et al. (2015) has redesigned promotion strategy for e-commerce. According to the author customers should be encouraged to buy products by giving some discount on them and at the same time other non-discounted products should be recommended to the customers. By doing this the loss which is occurred by giving discount can be recovered by recommending other non-discounted products. The parameters used by the author for implementing this are current price, product cost and reservation price. Three models are used by the author Online Promotion and Recommendation(OPR), Promotion with No Recommendation(PNR) and Minimum cost Ratios(MCR). Out of these OPR model outperformed the other two models by 27.7% and 10.3%.

2.2 Natural Language Processing

Author Eshan and Hasan (2017) has implemented a model for detecting abusive text in Bengali language. To deal with the textual data the author has implemented natural language processing techniques like countvectorizer, tfidfvectorizer, unigrams, bigrams and trigrams for converting the textual data in numerical format. After converting textual data in numerical format support vector machine, random forest and multinomial naive bayes models were implemented. The results showed that features acquired by tfidfvectorizer were better as compared to the features of countvectorizer while working with support vector machine with linear kernel. On the other hand author Tripathy et al. (2016) has made use of natural language processing technique to classify sentiments into positive, neutral and negative. Support vector machine, stochastic gradient descent, maximum entropy and naive bayes are used for modelling. The author has made use of combination of ngrams such as unigram, bigram, trigram, unigram+bigram, bigram+trigram and unigram+bigram+trigram. The output illustrated that lower value on 'n' in 'n-grams' gives better result than higher values. Which means that results obtained using unigrams and bigrams were excellent when compared to trigrams, fourgrams, etc.

Author Tripathy et al. (2015) has implemented text classification using machine learning classification algorithms namely naive bayes and support vector machine. These algorithms were used to classify the sentiments into positive and negative. For processing the text, author first converted the text into tokens, removed the stopwords and the punctuation marks and then used countvectorizer and tfidfvectorizer for converting the textual data into numerical form. The results showed that support vector machine (94%) outperformed naive bayes (89.5%) accuracy for classifying the sentiments. For classifying text three different feature extraction techniques were used by author Dzisevi and eok (2019) which are term frequency inverse document frequency(TF-IDF), linear discriminant analysis(LDA) and latent semantic analysis(LSA) over a neural network classifier. For implementing the model 10000 rows advertisement data was acquired from Lithuanian advertisement website for classifying into 20 categories. TF-IDF vectorizer outperformed the other two feature extraction techniques by helping the model to achieve a 91% accuracy score.

2.3 Prediction Algorithms

Author Chiang et al. (2018) has implemented ridge regression for big data to understand the computational speed of the algorithm. The data used for implementing the model is from bureau of transportation statistics and federal aviation administration which were pretty much large in size and consisted the data of each flight from year 2010 to 2015. Both of these dataset are merged to form a final dataset. Ridge regression is used for predicting the arrival and departure delay of the airlines. The ridge regression method proposed by author was successful in predicting the target variables with a mean squared error of 168,632.10 and mean absolute error of 394,368.89. The author concludes that ridge regression requires less memory, has faster computing speed and provides accurate results. Author Naik et al. (2018) has implemented a ridge regression model along with empirical mode decomposition (EMD) model for predicting the wind speed and power. The data used for predicting is from real wind farms. The speed of the wind needs to be predicted in a time span 10 minutes, 30 minutes, 1 hour and 3 hour. For each prediction the error rate of ridge regression was less as compared to the other models . The author concludes that ridge regression had the highest correlation coefficient factor. Apart from that the accuracy acquired by ridge regression was higher as compared to the other models.

Author Li et al. (2018) has implemented light gradient boosting machine(lgbm) for predicting the life expectancy of aircraft engines. The author says that engine is an essential part of the aircraft, it is necessary to do proper maintenance of airplane engine on time. Therefore a prediction model is build by the author so that the useful life of the aircraft engine can be estimated. The runtime of the turbofan and time window of row data are used as inputs for the model to capture the degradation information. The rmse value acquired by lgbm was 13.45 which was the least when compared to the other models used for prediction. The author concludes that lgbm is easy to interpret and it deals very well with data having high dimensional inputs. Author Chen et al. (2019) has implemented light gradient boosting machine (lgbm) for implementing for predicting the interaction of protein-protein. According to the author interaction of protein and protein plays an important role in a life of a cell such as signal transduction, transcriptional regulation and drug signal transduction. Experimental methods used for identifying protein-protein interaction are time consuming and costly, therefore author has implemented a machine learning approach to solve this problem. Firstly autocorrelation descriptor, pseudo amino acid composition, conjoint triad and local descriptors are used for extracting the features from the available information. Then elastic net is used for selecting appropriate features and the remaining redundant features are eliminated. The prediction accuracy received for identifying Mus musculus, Escherichia coli, Caenorhabditis elegans, Homo sapiens was 94.57%, 92.16%, 90.16%, 94.83% respectively which was a great achievement according to the author. On the other hand author Ju et al. (2019) has implemented light gradient boosting along with convolutional neural network for predicting the power of wind. The power of wind plays a very important role as various factors of electricity like the power system stability, quality of electric energy, wind power grind development, etc. These are the reasons why it is necessary for predicting wind power. Initially convolutional neural network is applied for extracting the features from the input data and then light gradient boosting is applied on the inputs along with CNN. The root mean squared error was on a lower side for CNN(1.752) as compared to LightGBM(1.848), while the mean absolute error was less for LightGBM(2.315) as compared to CNN(2.344). The author conlcudes that CNN was good in fetching data while LightGBM increases the robustness of the model.

Author Teng et al. (2019) has implemented convolutional neural network for estimating the target priority of air. A data of 900 units was used for training the neural network while 10 sets of data are used for testing. Since this is a regression problem least square error cost function is used for evaluating the model. The convolutional neural network implemented was of two layers the first was fully connected layer, 3 neural layers and the last one was output layer. The bias of all the layers in the CNN is initialized to zero, the learning rate was set to 0.001 with 1000 iterations. The absolute error obtained from CNN(0.0035) was pretty much low as compared to BP(0.0161) and PSO-SVM(0.0052) which was great. On the other hand author Petersen et al. (2019) has implemented convolutional neural network with LSTM for predicting travel time of the bus. Providing accurate and reliable results for public transport are essential so that attractive services can be offered to the customers in cities. The prediction was made for 15 minutes, 30 minutes and 45 minutes. The root mean squared error acquired was 2.66(15 minutes), 2.89(30 minutes) and 3.11(45 minutes) which was far more better as compared to other models. The overall root mean squared error acquired was 3.79, mean absolute error 3.02 and MAPE 5.61%. The author is convinced that this model of CNN and LSTM is preferable for predicting the price of public transport.

2.4 Market Basket Analysis

Author Kaur and Kang (2016) says that market basket analysis is a data mining technique used in various fields. Market basket analysis is implemented so that the buying patterns of the customers can be studied by the retailers which helps in decision-making. For implementing association rules author has made use of extended bakery dataset which consists of 20000 rows and 26 itemsets. The association rules are divided into two parts, upper association rules and lower association rules by keeping 20 as the threshold. The rules are evaluated using support, confidence and lift. On the other hand author Valle et al. (2018) has implemented minimum spanning trees within association rules. The author says that association rules gives large number of rules even for a small number of transactions which is unnecessary, so in order to get proper amount of rules the author has implemented minimum spanning trees. Minimum spanning trees improve the quality of market basket analysis. The data used for implementation is a correlation matrix between different product vectors. For calculating the quality these correlations of association rules are connected to the lift on which the association level between ancedent and consequence of the rule will be determined. As an output an undirected graph is obtained having N-1 edges and N number of nodes which has a connection path of minimum distance to the products.

For recommending product categories in a online supermarket author Fang et al. (2018) used association rules. In order to sell more products retailer sells products in a bundle. When too many products are purchased together the order is exempted from delivery charge which makes the customer happy as well as the retailer as more products are sold. For recommending appropriate products to the customer the author makes use of association rules. The dataset used in this paper consists of three columns product_id, customer_id and category_name. Since bundle needs to be recommended association rules are applied on the categories of the products and not on the products. For implementing association rules author has made use of apriori algorithm such that the value of support is kept at 0.1 and the confidence is adjusted at 0.01. On the other hand, author Cakir and Aras (2012) has implemented association rules for locating various products and adding it to the shopping cart based on different keywords entered by the customers. Apriori algorithm is used for implementing this association rules in C# software. Since association rules were based on a software the author made use of the following things: A text file was created for reading and writting the log files. Association rules were created by reading this log files by apriori.exe and a different text file was used for storing the results. Products with the highest probability value were displayed and added to the shopping cart.

3 Methodology

For completing this project CRISP-DM methodology was used. Each step of the methodology is divided into two sections prediction and recommendation as the data for both these models was different, therefore the pre-processing was also different. CRISP-DM has the following steps:

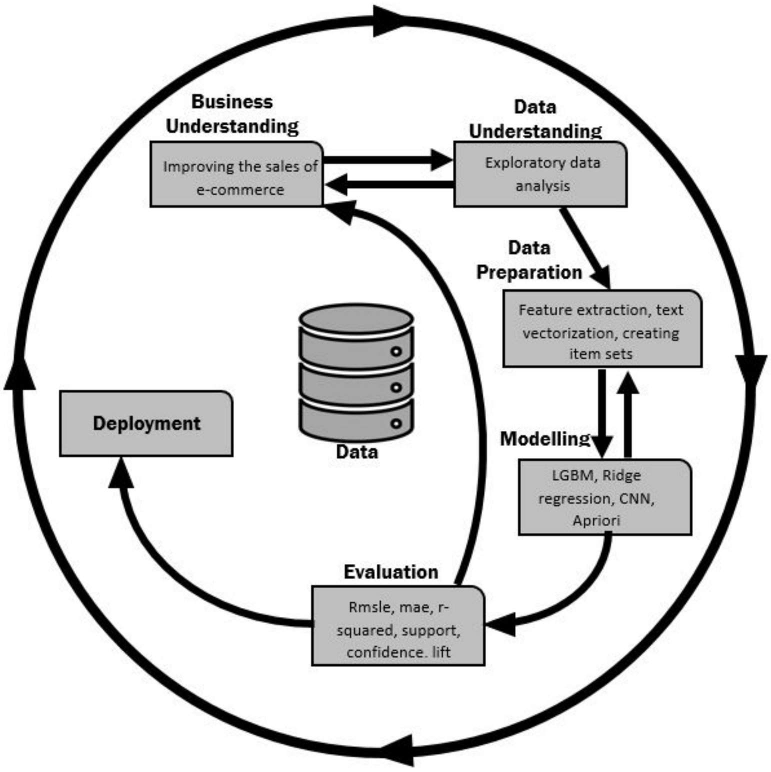

Business Understanding
Improving the sales of e-commerce
Data Understanding
Exploratory data analysis
Data Preparation
Feature extraction, text vectorization, creating item sets
Modelling
LGBM, Ridge regression, CNN, Apriori
Evaluation
Rmsle, mae, r-squared, support, confidence. lift
Deployment
Data

Figure 1: CRISP-DM

3.1 Business Understanding

- All the prices of the resale products on the e-commerce website will be standardized.
- The price which will be predicted will consider all the characteristics of the product and won't disappoint the seller.
- The standardized pricing will open new doors for recommendation.
- New resale products will be recommended to the customers.
- The above factors will optimize the sales of e-commerce website selling resale products.

3.2 Data Understanding

3.2.1 Prediction:

The dataset is downloaded from kaggle. This dataset is availed on kaggle by a website named Mercari which is a japanese e-commerce website which sells resale products. The data was downloaded from kaggle and was stored in google drive. The data consists of 1482535 rows and 7 columns. From google drive the data was pulled in google colab using python's library pydrive. The data set consist of the following 7 features: name of the product, item_condition_id, brand_name, categories, shipping, item_description and price of the product. The categories feature has 3 sub categories for each product such that the first subcategories is the main category followed by the department and the product category. The price feature consists of continuous value which ranges between 0-2009. The item_condition_id and shipping consist of categorical numeric values. The range of values in item_condition_id between 1-5, while shipping consists of only 1s and 0s. The 0 in shipping means that e-commerce website has not paid any money for shipping while 1 means that money has been paid by the e-commerce website for delivering that product.

3.2.2 Recommendation:

The dataset is downloaded from kaggle which was availed by Instakart. The recommendation system is built in R because there were many predefined libraries in R which made implementation easy to execute. The dataset is downloaded and loaded in R using read.csv function. The dataset consists of 6 files namely aisles, departments, order_products_prior, order_products_train, orders and products. The aisle file consists of two features aisle_id and aisle. The department file consists of two column department_id and department. Order_products_prior consist of 4 columns order_id, product_id, add_to_cart_order and reordered. Order_products_train consist of 4 columns order_id, product_id, add_to_cart_order and reordered. Order file consist of 7 files namely order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day and days_since_prior_order. Products file consist of product_id, product_name, aisle_id and department_id.

3.3 Data Preparation

3.3.1 Prediction

The data is converted into appropriate data types. Item_condition_id, shipping are converted to categorical data type. In categories column there are total three subcategories, these subcategories are separated by '/'. Using split function these categories are separated into three different columns. Categories and brand_name column had too many null values in it, brand_name had the highest number of null values. To fill these NA values, "missing" was added in place of NA in both the columns brand_name as well as categories. Then the data type of brand_name and categories was converted to category. For converting textual data into numbers natural language processing techniques were used such as tokenizing, punctuation removal, digits removal, stopwords removal, stemming, converting the upper case alphabets to lower case alphabets. After pre-processing the text, the text was converted to normalized form. Now this normalized text was supposed to be converted into numeric vectors. For converting the text into numeric vectors count vectorizer and tfidf vectorizer was used. Count vectorizer was applied on product name column and categories column while tfidf vectorizer was used on item_description column. Dummy variables were used on the item_condition_id and shipping column which encoded the two columns into 0s and 1s. Label binarizer was used on brand_name column for converting the textual data into numeric format. While for CNN text_to_sequence is used on name and item_description feature for converting the text to numeric format and for the other columns embedding were used, both of these function are present in keras library.

3.3.2 Recommendation

The recommendation data had 6 six files in it. But for implementing market basket analysis we only need two columns the first column we need is order_id to know what products are purchased together and the second column in product_name to learn about the different name of the products which were brought in a single transaction. Now in the file named order_products_prior there were total 4 columns namely order_id, product_id, add_to_cart_order and reordered. From this we took order_id and product_id and stored it in a variable named mydata. Now using merge function all the columns from the products file were included in the my data variable. After merging we got a total of 5 columns which were product_id, order_id, product_name, aisle_id and department_id. Only order_id and product_name were kept in the dataframe rest all columns were deleted. This data was then split into order_id and product_name.

3.4 Modelling

3.4.1 Prediction

There are total three modelling algorithms used for prediction which are light gradient boosting(LGBM), ridge regression and convolutional neural network(CNN).

LGBM: LGBM is a gradient boosting based framework which used algorithms based on tree learning. LGBM grows vertically while other tree based alogithms grow horizontally [1]. LGBM is implemented in this paper for the following reasons [2]: higher efficiency and training speed is fast, memory usage is low, great accuracy, able to handle huge data set.

Ridge regression: Ridge regression is used for analyzing the data of multiple regression which is suffering from multicollinearity. When multicollinearity is faced the least squared estimates becomes unbiased and the variance becomes so huge that it deviates

[1] https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

[2] https://lightgbm.readthedocs.io/en/latest/

from the true values. When a degree of bias is added to the regression estimates, standard errors are reduced by ridge regression [3].

Convolutional Neural Network: CNN is a variant of multilayer perceptron and a part of deep neural networks class. In multilayer perceptron each layer of neuron is connected to the other neurons in the consecutive layer[4]. CNN is usually used for modelling the images but it is used in this paper for modelling text because as CNN learns through all the pixels in the image, similarly it can also learn different connection between words and can predict the price of the product.

3.4.2 Recommendation:

Apriori: Apriori algorithm is used for mining frequent itemsets and implementing association rules over a transaction database. It first looks for frequent items and keeps on extending it to bigger dataset until the item set appear to look sufficiently large in the database. It is generally used in the domain of market basket analysis by finding trends and determining association rules.

3.5 Model Evaluation

3.5.1 Prediction

There are three parameters for evaluating prediction model, which are root mean squared logarithmic error(rmsle), mean absolute error (mae) and r-squared.

3.5.2 Recommendation

For evaluating the recommendation model 3 parameters are used which are support, confidence and lift.

[3] `https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf`

[4] `https://en.wikipedia.org/wiki/Convolutional_neural_network`

4 Implementation

4.1 Design Architecture

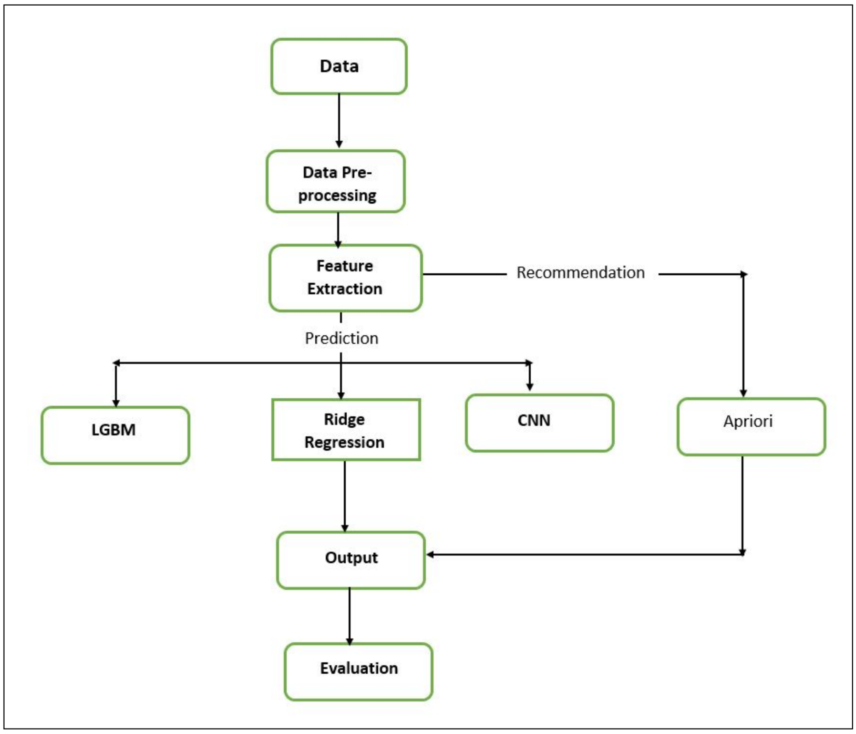


Figure 2: Design Architecture

4.2 Prediction

Starting with the prediction model, let's analyze the target variable which is the price of the product. According to mercari the price of resale products should lie between the price range of 3-2000 dollars, so we deleted the unnecessary rows from the data. The target value feature is completely skewed towards right as we can see in the figure 3.

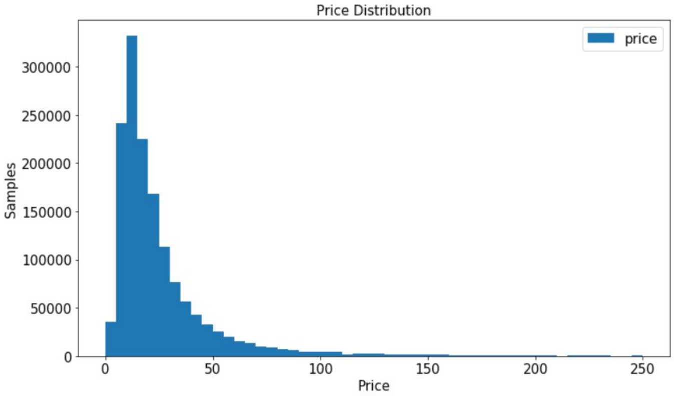


Figure 3: Price without logarithm

So in order to analyze the target variable we make use of logarithm function. Even after using the logarithm function still the data is skewed towards right as shown in figure 4.

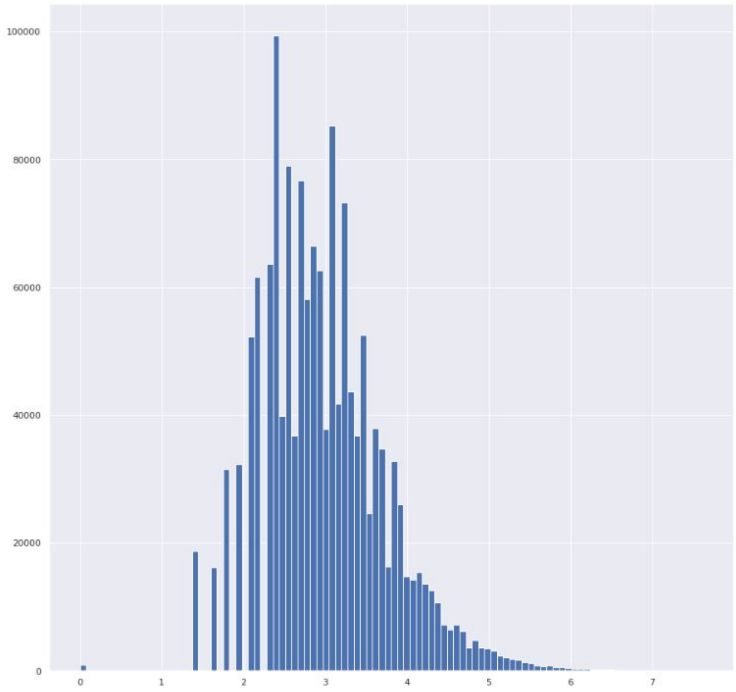
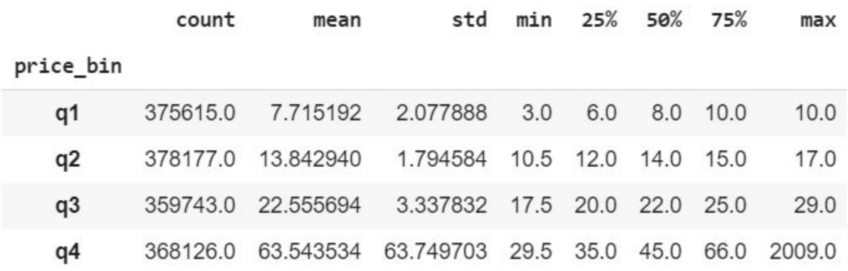


Figure 4: Price with logarithm

As the price was skewed a binning column was added by considering the inter-quartile range of price such that values between 3 and 10 will lie in q1. The value greater than 2 and less than 17 will lie in q2 and so on. In bins q1, q2 and q3 things are fine but in q4 things really become weird. If we look at the standard deviation of q4 it is increased to 63.749703 which is pretty high as compared to the other three bins. Also the range of the q4 is quite large, which results in skewed data as seen in figure 5.



	count	mean	std	min	25%	50%	75%	max
price_bin								
q1	375615.0	7.715192	2.077888	3.0	6.0	8.0	10.0	10.0
q2	378177.0	13.842940	1.794584	10.5	12.0	14.0	15.0	17.0
q3	359743.0	22.555694	3.337832	17.5	20.0	22.0	25.0	29.0
q4	368126.0	63.543534	63.749703	29.5	35.0	45.0	66.0	2009.0

Figure 5: Distribution of price in binning

Coming to shipping variable it only contains 1s and 0s. 1 means that mercari had to pay for shipping while 0 means no shipping cost was incurred. From the data below it is clear the distribution of shipping. From the distribution it is clear the cost of shipping has no influence on the price of the product.

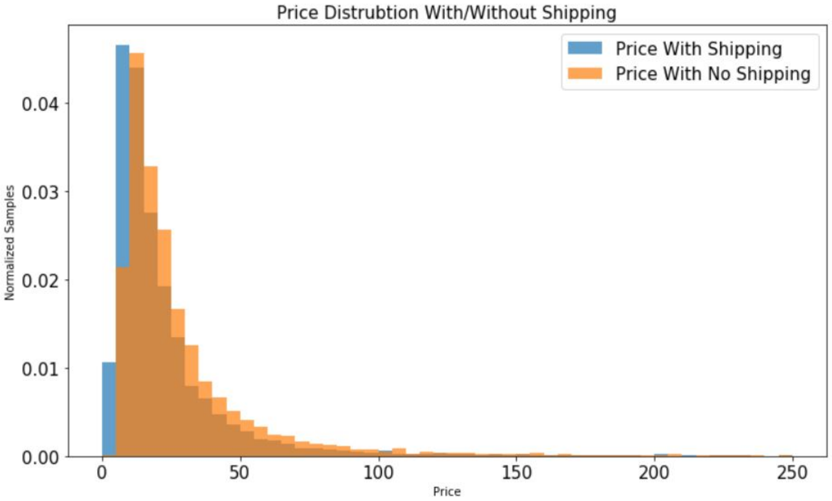


Figure 6: shipping

In brand_name feature there were total of 4809 unique values. Pink and Nike were the most sold branded products. For most of the products brand name was missing as shown in the figure.

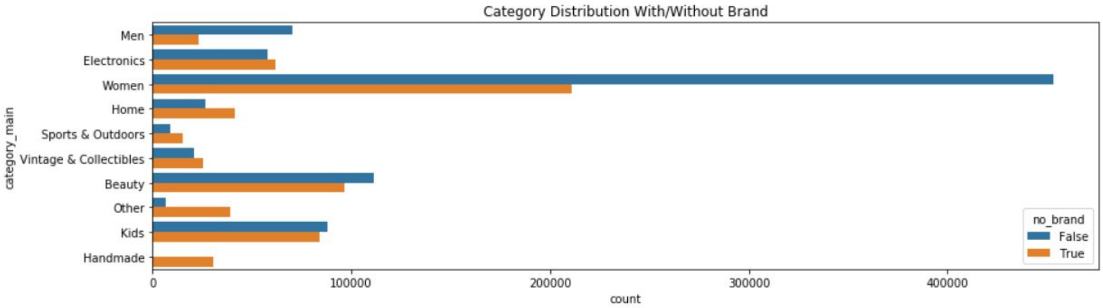


Figure 7: Brand Name

The category_name column has three subcategories which is separated by '/' delimiter. This category_name column was split into three different columns such that each column is separated with only one category. Products which are sold in the main category the most are from women category followed by beauty. Computers and tablets were the highest selling section in the first category. Laptops and notebooks were the highest selling in second section. All the missing values in categories and brand name were replaced by "missing".

Coming to the data item_condition_id and shipping were the only two numerical features present in the data. Now if we look at the correlation matrix in figure 8 we can see that both of these features hardly have any correlation with the target variable price. Which means that majority of the features required to predict the price lie inside the textual data. Hence, now the price prediction was a complete natural language processing problem.

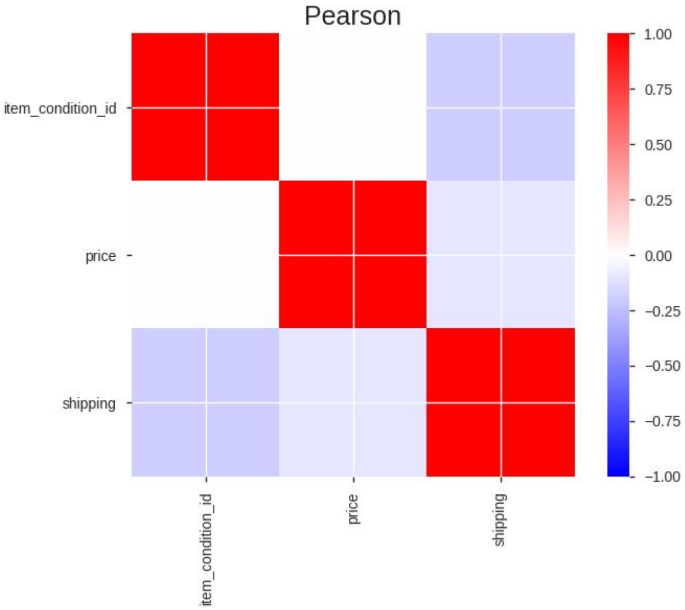


Figure 8: Correlation

So starting with natural language processing first all the data from name and item description was converted to tokens. After the punctuation inside the data were removed, all the digits were removed, stopwords were removed and all the uppercase letters were converted to lower case. All types of stemming was implemented on the data for bringing the data into its root form. For instance dancing will be converted to dance after applying stemming. After getting the data into the root form, the next thing is to convert the words into numbers. For doing this count vectorizer and tfidf vectorizer were used. There are lots of other textual feature extraction techniques like fast text, word2vec, etc. but count vectorizer and tfidf vectorizer are used because we wanted to avoid those techniques which are computationally heavy considering the hardware specification. Count vectorizer is applied on the product name and categories feature and tfidf vectorizer is applied on item description feature. Since product name and categories do not have much information in it those two features can be easily encoded by count vectorizer. For encoding item description, count vectorizer can also be used but since it contains a lot of information about the product, it is preferable to give proper weights to each word in the sentence. Also as described by Eshan and Hasan (2017) features extracted by tfidf vectorizer are far better than count vectorizer. While using tfidf vectorizer on item_description bigrams were used to see the capture the connection between two consecutive words. For converting the brand name labels into textual format label binarizer was used which is present in sklearn library. Item_condition_id contained categorical data in the range of 1-5 and shipping contained only binary data which was 0 and 1. Both of these features were encoded into dummy variables which encoded both of these variables in 0s and 1s and were later converted to array using csr_matrix. Since the remaining columns were encoded by count vectorizer and tfidf vectorizer the values of the other columns were also converted into range of 0-1. All these converted variables were clubbed together using hstack which combines the data horizontally. Functions like csr_matrix and hstack are imported from scipy.sparse library. After this all the data was converted into suitable form for pushing it into algorithms for predicting the price. This was the pre-processing used for ridge regression and LGBM, while the pre-processing used for CNN was a bit different. The natural language processing techniques used were minimal here, text was directly converted into numeric vectors using text_to_sequence which is found in keras library. For converting the other features into appropriate format, embedding function was used which again found in keras library. Since there are two layers in CNN the activation function used is relu. After applying relu function to the first two layers, the inputs of the data were down sampled using GlobalMaxPooling1D() function and this function was then concatenated to both the layers of CNN. Linear activation layer is used for output.

For splitting the data in training and testing k-fold cross validation techniques was used in ridge regression and LGBM, the data was split in 10 splits. For CNN, data was splitted in a ratio of 75:25 for training and testing, epoch was used for retraining the data. The size of epoch used is 3. As the data was quite big considering the hardware specification it is made sure that the algorithms which are applied on the data are not heavy on the CPU. The algorithms used for prediction are light gradient boosting(LGBM), ridge regression and CNN. The main reason for using LGBM is that as the name suggest **Light**GBM meaning it is computationally light on the CPU. Same goes with ridge regression, it requires less memory, good computing speed and gives accurate results as said by Chiang et al. (2018). CNN is usually used on images so that it captures the nearest pixel possible and gives accurate results. In this case the prices were solely dependent on the textual data as the numeric data item_condition_id and shipping were not of much use. This is the main reason why CNN is used in this paper so that it can learn through each textual word and then predict the price.

4.3 Recommendation

The recommendation model is implemented in R and not in python because there were lot of libraries available for implementing market basket analysis in R as compared to python. Therefore implementing in R was easier as compared to python. Coming to the data it contained five files namely order, products, order_products_prior, order_products_train, aisles and department. Features of these files are described in section 3. An order file contains the following features order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_orders. order_dow is orders done in day of the week, to make this more convenient to analyze we add a column which is named as day_week_name. The order_dow has categorical numbers between 0-6 such that 0 means sunday and 6 means saturday. This data is added to day_week_name column. Now lets look at the frequency of the orders over the week.

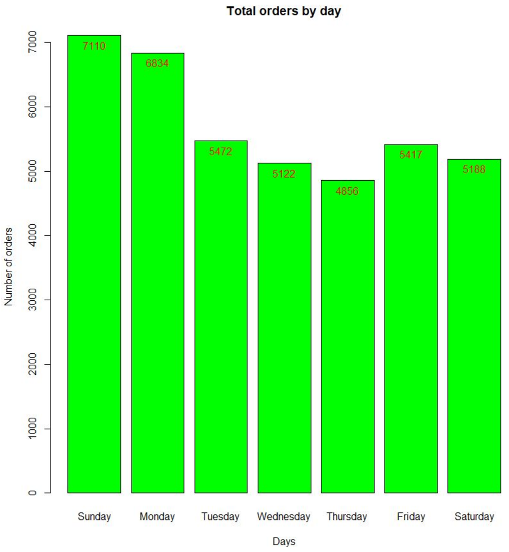


Figure 9: Orders over the week

From the graph below in figure 9 it is clear that maximum orders are placed during sundays and mondays, sunday being the busiest day. Now we need to analyze which time of day receives more number of orders. For doing this order_hour_of_day is visualized, from this we can see that hours between 9-17 are the busiest.

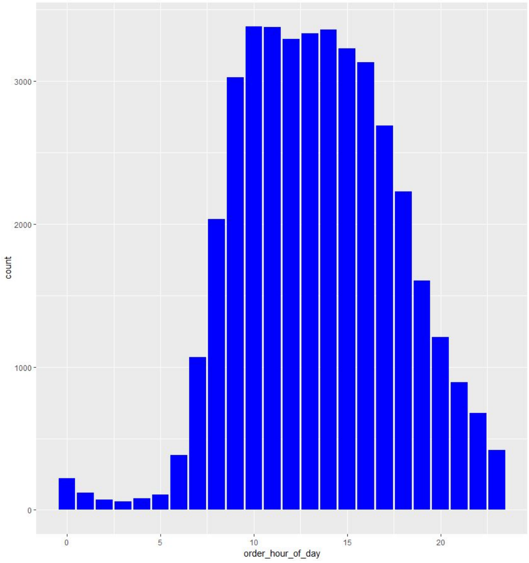


Figure 10: Busiest hours of the day

Another column created is products_department in which the data from products and departments file are merged together using merge function, in both of these files department_id was common. All the data present in each file is converted into appropriate data types. A new data frame is created named mydata. Initially, only order_id and product_id are added to this variable. Using merge function all the data from the products file are pulled in mydata variable. Now this mydata variable consists 5 features namely product_id, order_id, product_name, aisle_id and department_id. All the data in mydata variable is arranged according to order_id. For implementing market basket analysis only two features are needed namely order_id which contains the number of products purchased by each customer and product_name which contains the name of the products bought by each customer. All the remaining columns from the mydata variable are eliminated. Next the two columns in mydata variable order_id and product_name are split using split function. After looking at the summary of statistics it is observed that data is converted into sparse matrix which contains 3934 rows and 8390 columns. Banana is the most purchased product in the itemset which is present in 605 transactions. From the summary statistics it can be observed that a normal person purchases 7-8 products in one transaction. There are only two transactions which contains 51 products in it, which is considered to the the biggest transaction recorded in the dataset. Apriori algorithm is used for implementing the association rules. After implementing the algorithm a lot of rules were created based on the transactions. Not all of the rules established are unique some of the rules are redundant, we need to eliminate those rules. For eliminating those rules a variable was created named subset.matrix. Now the variables which contains all the rules are checked for subset, if the subset value is true, the variables are added to subset.matrix variable. The lower triangle of the subset matrix is set to NA and the column sums of those matrices are checked, if the value of the col sums is greater than 1 then it is added to a variable named redundant where the redundant rules are pruned.

5 Evaluation

5.1 Prediction

For evaluating the performance of LGBM and ridge regression three parameters are used namely root mean squared logarithmic error (rmsle), mean absolute error (mae) and r-squared. The range of values in the target variable is quite huge and owing to this root mean squared logarithmic error (rmsle) is used instead of root mean squared error (rmse). R-squared explains what percentage of the variance of the target variable is explained by the independent variables. The formulas for evaluating both of these parameters are as follows:

$$RMSLE^5 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(pi+1) - (log(ai+1))}$$

$$R - squared = 1 - \frac{squared\,error\,of\,regression\,line}{squared\,error\,of\,meanline}$$

5.1.1 Experiment 1: LGBM

The rmsle value obtained by using LGBM model is 0.538, mae was 0.4 and the r-squared value obtained is 0.48 as seen in the below table. The results obtained from this algorithm

[5] https://medium.com/@viveksrinivasan/how-to-finish-top-10-percentile-in-bike-sharing-demand-competition-in-kaggle-part-2-29e854aaab7d

was not good enough as compared to the other two algorithms. Each type of stemming was applied on the description feature of the data but it did not help much to improve the error value. The r-square value showed that is not even 50% of the variance in the dependent variable was explained by the independent variables.

	No Stemming	Porter Stemmer	Lancaster Stemmer	Wordnet Lemmatizer
RMSLE	0.5389	0.5382	0.5384	0.5387
MAE	0.4098	0.4092	0.4094	0.4097
R-squared	0.4808	0.4821	0.4817	0.4813

5.1.2 Experiment 2: Ridge Regression

The rmsle, mae and r-squared value obtained by using ridge regression is a bit better as compared to LGBM model as shown below. By looking at the r-squared value it can be said that ridge regression is a better fit over LGBM. If neural network is not considered over this data, then ridge regression is one of the best fit.

	No Stemming	Porter Stemmer	Lancaster Stemmer	Wordnet Lemmatizer
RMSLE	0.4714	0.4721	0.4725	0.4716
MAE	0.3519	0.3525	0.3530	0.3522
R-squared	0.6028	0.6016	0.6009	0.6023

5.1.3 Experiment 3: Clubbing features

Since features name, brand_name ,category_name and item_description are completely textual. So all these columns were clubbed into one column and natural language processing technique is applied to see the error and the r-squared value. From the below table it is clear, that the error rate only increased and the r-squared value got less. Since the results were not good, so stemming was not applied further.

	LGBM	Ridge regression
RMSLE	0.5583	0.4845
MAE	0.4264	0.3617
R-squared	0.4427	0.5803

5.1.4 Experiment 4: Convolutional Neural Network

CNN being a neural network did a pretty good job for predicting the price of the products. The rmsle value of CNN is 0.4461, mae is 0.3345 and r-squared is 0.6460. Looking at the error and the r-squared value it is clear that it is the best fit when compared to the other two algorithms. Stemming is not implemented for CNN as it is clear from the above results of LGBM and ridge regression that it hardly helps to improve the error and r-squared value.

5.2 Recommendation

The evaluation measures used for judging the performance of the market basket analysis were support, confidence and lift. The formulas for all the three paramters are given below:

$$Support = \frac{Freq.ofitem}{Totalnumberoftransactions}$$

$$Confidence = \frac{supp(A \cup B)}{supp(A)}$$

$$Lift = \frac{supp(A \cup B)}{supp(A) * supp(B)}$$

5.2.1 Experiment 4: Apriori

Lower the size of the support, confidence and lift more rules are applied by the apriori algorithm. Just for sake of experimenting the value of support and confidence was kept very high. Support was kept at 0.001 while the value of confidence was kept at 90% still 25 association rules were acquired. Below is a output of market basket analysis, only top 5 rules are displayed to understand the output

	Rules	Support	Confidence	Lift
1	{Deli Fresh Honey Smoked Turkey Breast, 98% Fat Free, Gluten Free} =>{Banana}	0.001016777	1	6.502479
2	{Quaker Life Cinnamon Cereal} =>{Banana}	0.001016777	1	6.502479
3	{Mediterranee Strawberry Yogurt} =>{Coconut Yogurt}	0.001016777	1	786.800000
4	{Organic Chamomile with Lavender Herbal Tea Bags} =>{Bag of Organic Bananas}	0.001016777	1	7.713725
5	{Organic Cilantro,Organic Roma Tomato} =>{Organic Baby Spinach}	0.001016777	1	14.30545

Now here the value of support was kept at 0.001 and confidence was at 0.90 so rules was created considering the value and support applied. If we have a look at the above table all the rules created satisfy the specified value of support and confidence. To have a visual description of this association rules a directed graph is visualized.

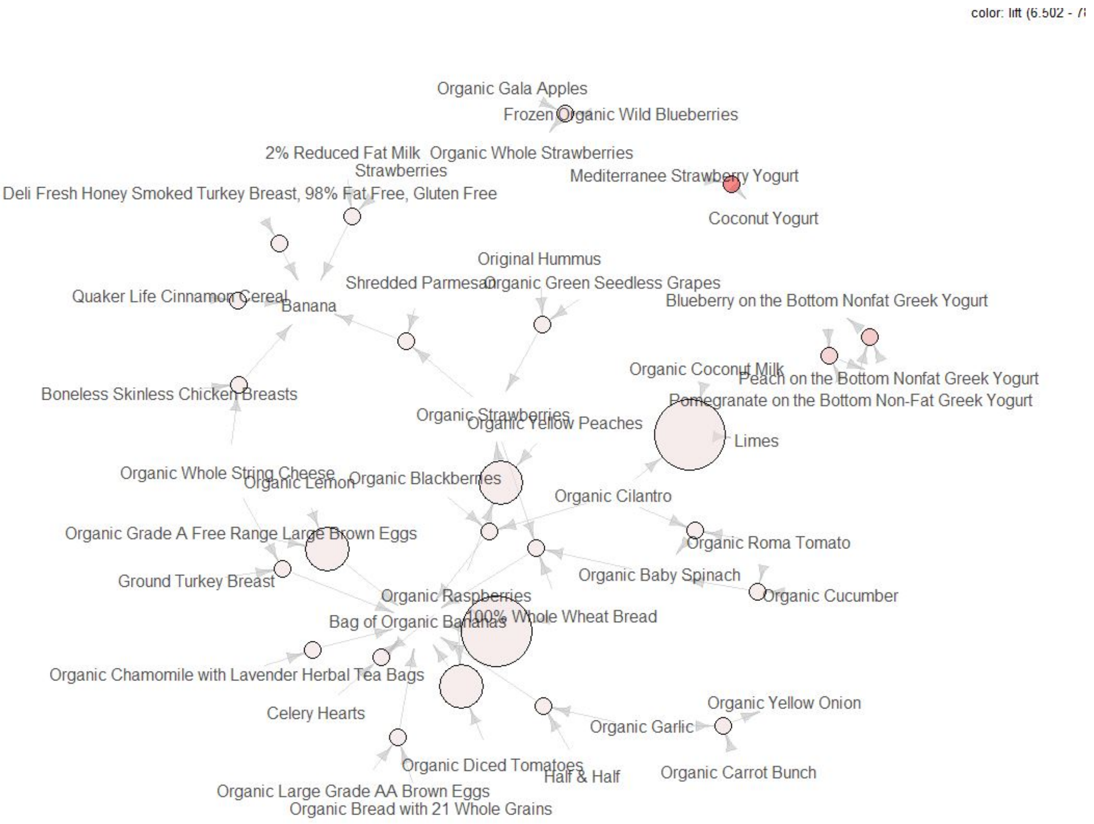

color: lift (6.502 - 7
Organic Gala Apples
Frozen Organic Wild Blueberries
2% Reduced Fat Milk
Organic Whole Strawberries
Strawberries
Mediterranee Strawberry Yogurt
Deli Fresh Honey Smoked Turkey Breast, 98% Fat Free, Gluten Free
Coconut Yogurt
Original Hummus
Shredded Parmesan
Organic Green Seedless Grapes
Quaker Life Cinnamon Cereal
Banana
Blueberry on the Bottom Nonfat Greek Yogurt
Organic Coconut Milk
Peach on the Bottom Nonfat Greek Yogurt
Boneless Skinless Chicken Breasts
Pomegranate on the Bottom Non-Fat Greek Yogurt
Organic Strawberries
Organic Yellow Peaches
Limes
Organic Whole String Cheese
Organic Lemon
Organic Blackberries
Organic Cilantro
Organic Grade A Free Range Large Brown Eggs
Organic Roma Tomato
Ground Turkey Breast
Organic Baby Spinach
Organic Raspberries
Organic Cucumber
Bag of Organic Bananas
100% Whole Wheat Bread
Organic Chamomile with Lavender Herbal Tea Bags
Celery Hearts
Organic Yellow Onion
Organic Garlic
Organic Diced Tomatoes
Half & Half
Organic Carrot Bunch
Organic Large Grade AA Brown Eggs
Organic Bread with 21 Whole Grains

Figure 11: Market basket analysis directed graph

5.3 Discussion

Starting with the prediction algorithms, the results obtained are satisfactory considering the target variable and the size of the data. The market of the e-commerce is pretty much dynamic as seller wants to sell the products at higher cost and the buyer wants to buy the product at less price. This is making the price variable skewed. It is the skewness of the target variable which is creating problems for the prediction algorithms. The two features item_condition_id and shipping hardly added any value for predicting the price. So the only hopes left for predicting the price was the textual data. Coming to the pre-processing part the pre-processing required for CNN was the least as words were directly converted to vectors using text_to_sequence method from keras. While for passing data in LGBM and ridge regression model the textual data had to go through tokenizing, punctuation removal, digit removal, stop words removal, etc. Both types of methodologies are used for extracting textual features the traditional count vectorizer, tfidf vectorizer used for LGBM and ridge regression models and the other is text to sequences which is from keras library used for CNN. Eventhough with less pre-processing CNN proved to be a better performer with respect to rmsle, mae and r-squared value. In contrast, the execution time required for CNN was more as compared to the other two algorithms. Since the data is huge CNN is able to learn different patterns of text which influences the price. If more data is appended the error value can further be reduced using CNN.

6 Conclusion and Future Work

In this paper there are two things implemented price prediction and product recommendation for e-commerce websites selling resale products. Considering the size of the data the algorithms used are such that are less intensive on the CPU. The price is predicted using three algorithms linear gradient boosting machine (LGBM), ridge regression and convolutional neural network (CNN). Standard nlp techniques were used for ridge regression and LGBM, while deep learning techniques from keras library were used for preprocessing data for CNN. Considering the skewness and the size, the results acquired were amazing. In terms of performance CNN was the clear winner over the other two algorithms. Convolutional neural network outperformed ridge regression by approximately 3% and LGBM by 9%.

Recommendation model is implemented using apriori algorithm. Association rules are used for mining the frequent item sets in the data. The only two meaningful features for creating association rules were order id and product name. These feaures were splitted into transactions and then association rules were applied by setting the support value of 0.001 and confidence value of 0.90. Since the value of the support and confidence was so high only 25 rules were generated. After implementing the algorithm there were some rules which were redundant and these rules were pruned later by using subset function which is described in implementation section.

In future work, the images of the products can be included as a feature for making the prediction model more robust. By including the images the condition of the product can be more accurately judged. Furthermore, the seller's profile can also be used as another feature for improving the performance of the model. On the other hand some other algorithm should be experimented for applying association rules for the recommendation system.

7 Acknowledgement

I would like to thank my supervisor Prof. Manuel Tova-Izquierdo for guiding me to achieve my proposed research project. He always solved my queries in one-to-one session as well as on emails. I would also like to show my gratitide towards my family and my friends for supporting me through out my MSc journey.

References

Bauer, J. and Jannach, D. (2018). Optimal pricing in e-commerce based on sparse and noisy data, *Decision Support Systems* **106**: 53 – 63.
URL: *http://www.sciencedirect.com/science/article/pii/S016792361730221X*

Cakir, O. and Aras, M. E. (2012). A recommendation engine by using association rules, *Procedia - Social and Behavioral Sciences* **62**: 452 – 456. World Conference on Business, Economics and Management (BEM-2012), May 46 2012, Antalya, Turkey.
URL: *http://www.sciencedirect.com/science/article/pii/S187704281203515X*

Chen, C., Zhang, Q., Ma, Q. and Yu, B. (2019). Lightgbm-ppi: Predicting protein-protein interactions through lightgbm with multi-information fusion, *Chemometrics*

and Intelligent Laboratory Systems **191**: 54 – 64.
URL: *http://www.sciencedirect.com/science/article/pii/S016974391930262X*

Chiang, W., Liu, X., Zhang, T. and Yang, B. (2018). A study of exact ridge regression for big data, *2018 IEEE International Conference on Big Data (Big Data)*, pp. 3821–3830.

Dai, Y. N., Viken, G., Joo, E. and Bente, G. (2018). Risk assessment in e-commerce: How sellers' photos, reputation scores, and the stake of a transaction influence buyers' purchase behavior and information processing, *Computers in Human Behavior* **84**: 342 – 351.
URL: *http://www.sciencedirect.com/science/article/pii/S0747563218300967*

Dzisevi, R. and eok, D. (2019). Text classification using different feature extraction approaches, *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pp. 1–4.

Eshan, S. C. and Hasan, M. S. (2017). An application of machine learning to detect abusive bengali text, *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6.

Fang, Y., Xiao, X., Wang, X. and Lan, H. (2018). Customized bundle recommendation by association rules of product categories for online supermarkets, *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pp. 472–475.

Greenstein-Messica, A. and Rokach, L. (2018). Personal price aware multi-seller recommender system: Evidence from ebay, *Knowledge-Based Systems* **150**: 14 – 26.
URL: *http://www.sciencedirect.com/science/article/pii/S0950705118300893*

Guo, J., Gao, Z., Liu, N. and Wu, Y. (2018). Recommend products with consideration of multi-category inter-purchase time and price, *Future Generation Computer Systems* **78**: 451 – 461.
URL: *http://www.sciencedirect.com/science/article/pii/S0167739X1730256X*

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**: 257 – 270.
URL: *http://www.sciencedirect.com/science/article/pii/S092552731500208X*

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**: 28309–28318.

Kaur, M. and Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining, *Procedia Computer Science* **85**: 78 – 85. International Conference on Computational Modelling and Security (CMS 2016).
URL: *http://www.sciencedirect.com/science/article/pii/S1877050916305208*

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, *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3562–3567.

Naik, J., Satapathy, P. and Dash, P. (2018). Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression, *Applied Soft Computing* **70**: 1167 – 1188.
URL: *http://www.sciencedirect.com/science/article/pii/S1568494617307251*

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 & Management* **56**(2): 317 – 328. Social Commerce and Social Media: Behaviors in the New Service Economy.
URL: *http://www.sciencedirect.com/science/article/pii/S0378720617305712*

Petersen, N. C., Rodrigues, F. and Pereira, F. C. (2019). Multi-output bus travel time prediction with convolutional lstm neural network, *Expert Systems with Applications* **120**: 426 – 435.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417418307486*

Teng, L., Guo, Q. and Gao, Y. (2019). Target priority estimation based on convolutional neural networks, *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 1967–1971.

Tripathy, A., Agrawal, A. and Rath, S. K. (2015). Classication of sentimental reviews using machine learning techniques, *Procedia Computer Science* **57**: 821 – 829. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).
URL: *http://www.sciencedirect.com/science/article/pii/S1877050915020529*

Tripathy, A., Agrawal, A. and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach, *Expert Systems with Applications* **57**: 117 – 126.
URL: *http://www.sciencedirect.com/science/article/pii/S095741741630118X*

Valle, M. A., Ruz, G. A. and Morrs, R. (2018). Market basket analysis: Complementing association rules with minimum spanning trees, *Expert Systems with Applications* **97**: 146 – 162.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417417308503*