National College of Ireland

Sales and Logistics Analysis in E-Commerce using Machine Learning Models:UK

MSc Research Project
Masters in Computer Science Data analytics

Mysura Reddy Polam
StudentID: X21143323

School of computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mysura Reddy Polam		
Student ID:	X21143323		
Programme:	Masters In Computer Science Data Analytics	**Year:**	2022
Module:	MSc Research Project		
Supervisor:	Catherine Mulwa		
Submission Due Date:	15/12/2022		
Project Title:	Sales And Logistics Analysis in E-commerce using Machine Learning Models: UK		
Word Count:	**Page Count:**		

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: Polam Mysura Reddy

Date: 15/12/2022

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):	

Sales And Logistic Analysis In E-Commerce Using Machine Learning Models: UK

Mysura Reddy Polam
X21143323

Abstract

Machine learning Technology is perhaps the most advanced use of AI technology that can produce results and train entirely on its own without being supervised learning in this research. The profitability of an e-commerce business is obtained and comprehended it using a machine learning model. This study made the best use of open-source technologies such as SQL, Powerbi, and Python, by employing machine learning approaches for business data regression models analyzed. Gradient boosting and Bagging regression yielded results with a 94% accuracy and just minor errors. In forecasting the company's sales and logistical success the scope and diversity of business patterns are also recognized at distinct periods and regional zones, allowing business owners to make intelligent choices. Furthermore, previous relevant work findings and research requirements are discussed to fill gaps in research feature adjustments and models are now being developed.

1 Introduction

Prediction techniques in business particularly for online ecommerce market vendors, can play a crucial role in their infrastructure investments for inventories. If the expected outcomes are more beneficial to merchants and merchants, they will profit handsomely. Marketers might also look at customer thoughts all purchasing habits. For data mining methodologies, there are several offline and online tools available and also with data mining techniques, which is ready to aid the ERP model in projecting consumer behaviours and development outlook in evolving business clarifications to experience an understanding judgments. Essentially, the use of these artificial integrated data modelling approaches or tools yields results and previously unknown analysis. It conceals knowledge that experts cannot predict and that we cannot produce using conventional methods. Data extraction techniques that have previously been employed are unexpected. People are frequently oblivious of the impact of this, despite the fact it had the potential to be effective in creating valuable facts and insights from data sets. In generally, database entries are used, and this approach is termed as knowledge discovery from data sources(KDD)

1.1 Research Question

Many sectors have been looking for improved analysis for company improvement for decades. Many large corporations have already implemented many tools and databases using modern technology to pinpoint the company’s problem. However, numerous small and medium-sized businesses failed to do so due to the high cost of managing tools, networks, and other resources. Typically e-commerce companies have lost business value and revenue

due to their being unable to foresee sales patterns, and concerns such as logistics, geographical sales, consumer behavior, and product rating all have a significant influence.

Research Question:
To give the best solution for this scenario, how can machine learning techniques and business intelligence tools be used to help a company to increase sales and logistics performance with less time consumption and more economics?

Sub-Research Question:
To what extent can prediction of logistic performance with sales can help eCommerce business?

1.2 Research Objectives

The study of E-commerce companies using machine learning algorithms was created with references to numerous relevant articles. Most of the articles in this study field focused on customer segmentation and product categorization, however, the main aim of this report is to make predictions using machine learning algorithms, providing extra data visualization that is valuable for company stakeholders in understanding business volatility and gaining relevant information. Additionally, the sub-objective was investigated, as well as sales forecasting and analysis in other domains such as logistics department enhancements and regional sales trends. Figure1(source google) depicts an example of basic web almost utilization on mobile.

Figure 1: E-commerce.

To offer more detailed solutions to the research questions all of the implementation, assessment, and outcomes objectives described below are divided into two parts i.e. sales and logistics analysis

Obj1: A critical review of literature on Sales and Logistics Analysis .
Obj2: Experiment 1: Sales Analysis
Obj2(a): Exp1: Implement, evaluate and results of Linear Regression.
Obj2(b): Exp1: Implement, evaluate and results of Decision Tree Regressor.
Obj2(c): Exp1: Implement, evaluate and results of Gradient Boosting Regressor.
Obj2(d): Exp1: Implement, evaluate and results of Bagging Regressor.
Obj2(e): Exp1: Implement, evaluate and results of AdaBoost Regressor.

Obj2(f): Exp1: Implement, evaluate and results of KNeighbor Regressor.
Obj2(g): Exp1: Comparison of developed models.
Obj3: Experiment 2: Logistics Performance Analysis
Obj3(a): Exp 2: Implement evaluate and results of Linear Regression.
Obj3(b): Exp 2: Implement evaluate and results of Decision Tree Regressor.
Obj3(c): Exp 2: Implement evaluate and results of Gradient Boosting Regressor.
Obj3(d): Exp 2: Implement evaluate and results of Bagging Regressor.
Obj3(e): Exp 2: Implement evaluate and results of of AdaBoost Regressor.
Obj3(f): Exp 2: Implement, evaluate and results of KNeighbor Regressor.
Obj3(g): Exp 2: Comparison of developed models.
Obj3(h): Make use of open-source tools like SQL, Power Bi and Python for this research.

1.3 Roadmap

The rest of the technical study are organized as followed: chapter 2 includes an analysis of present literature in the performance of e-commerce development and associated concerns answered employing machine learning, implementing sales prediction and forecasting logistics. Based on previous research outcomes chapter 3 gives scientific technique accompanied by feature extraction and finally in chapter 4 the implementation evaluation and outcomes are presented.

2 Related Works on E-commerce

This section will go through relevant publications for calculating sales projections and other associated concerns in an e-commerce setting. although a study in this field is relatively limited several of the methodologies used in clinical investigations, marketing and sometimes even time series forecasting have proved useful to this research effort (2010 to 2022)

2.1 Critical Review of Sales-Related Issues in E-commerce

To far, very small research has been done in the subject of predictive modelling for sales pipelines in general and particularly in the win-propensity area(Yan et al.; 2015; Bohanec et al.; 2017a). As per (Bohanec et al.; 2017), the overwhelming majority of research shows limited evidence of strong corporate adoption. (Lawrence et al.; 2010) created ontarget and map solutions to find new sales prospects and match resources to them using likelihood modeling for IBM deployment. (Lawrence et al,; 2010) on the other hand, do not offer proof of which methods were employed nor do they disclose specifics on modelling solutions or how accuracy of the model was evaluated. In reality, appraisal is focused on assessing changes in tangentially connected measures such as channel and revenue growth. As well as goal fulfilment. Variations in these measures might be due to a variety of external sources other than the influence of inclination computing. D'Hean and Van Den Poel (2013) devised, A 3 phase strategy for sales client relationships the second of which included a propensity-baseded approach. The second stage of this methodology included logistic regression, decision trees and neural network approaches for determining whether a potential investment lead should turn into our sales prospects or not. Despite the writer's desire for a fully

autonomous prospect list creation solution the prototype list was physically inspected by corporate staff, sorting it into "good" and "poor" Leads, rather than inventing bespoke property criterion into the modelling process. Likewise, Duncan and Elkan (2015) created two greater tendency approaches, DQM and FMM, to transform marketing lead converting into a potentially by exploiting its win propensity, thereby accounting for not just leave converting propensity but also a sales opportunity when propensity. The researchers like this study use marketing automation data, emphasising the importance of huge historical data set, albeit not exactly measuring size, and trying to identify the very next marketing automation characteristics as vital towards the modelling: client business size, corporation selling price, geographic region.

Cyber crime is among the most typical issues that e-commerce company faces. Jay Nanduri and Yung-wen Liu offered her study to detect errors and fraud (Nanduri et al.; 2020), this article examines two machine learning strategies, one of which is called fraud islands the, other is a multi layer machine learning model. This study demonstrates that by applying a machine learning model, the probability of recognising fraudulent activity was increased significantly. However, despite its high analytical capacity, the approach can only detect template frauds, such as the fixed order of fraud and not normal ones.

Every commercial company must concentrate on actions that keep consumers faithful throughout the time. Maciej Ponndle and Jolanta Pondal (Ponndel & Pondal et al.; 2021) attempted to provide a way to boost advertising effectiveness using a machine learning approach. This article focuses on how to leverage client data in business to drive revenue (Hendra et al.; 2017). In this paper, the authors suggested a novel model that uses advanced data analysis and machine learning approaches to generate offers for clients that are relevant in commerce organisations. Client satisfaction and marketing strategies are prioritised since they are the most vital aspects of an e-commerce company.

Table 1: Summary of Related Works to Machine Learning Models

	Related Works- Sentimental analysis			
Author	Specific Features	Compared Results	Advantages	Limitations
Kourouklidis	Model driven engineering approach for monitoring machine learning models	72% results is achieved	Business problems were clearly solved.	No data preprocessing
Nanduri,	Fraud deduction, paper deals with two machine learning techniques one is fraud islands and a multi-layer machine learning model	69% accuracy is achieved	KDD methodology was properly followed. .	Limited with 2 models
Rong, L., Weibai	Neural Network with emotional tags	75% accuracy is obtained	Clear explanation how neural network used	Dataset is very small

2.2 Time-Series Forecasting

The time series technique is far more investigated in the domain of B2B and B2C sales plan, and it has witnessed a lot of noteworthy advances, extending standard ARIMA and regression approaches. Artificial neural network have been shown to beat ARIMA models in sales forecasting Tkac and Verner (2016) B2B technology, on the other hand ELM model was proven to be faster in marketing (Lu and Kao; 2014) and fast-paced B2C contexts, such as clothing stores (Yuet al,; 2011) (Xia et al,; 2012), more economical and much less computationally demanding than neural network models, better suited for actual forecasting. MARS-multivariate adaptive splines-was established as a better strategy for selecting suitable factors for forecast and delivered more high accuracy than neural network rear (Lu et al,; 2012). Certain recent period research advocate the application of an array of machine learning approaches to increase prediction and prediction accuracy. (Lu and Kao; 2016; Lu; 2014, Gurnani et al,;2017). According to the research combining ARIMA with xgboost makes it much more resilient to non linear data characteristics, size of data, patterns, and periodicity resulting in greater efficiency than SVM an neural network models (Gurnani et al.; 2017).

2.3 Critical Review of Sentimental Analysis and Comparison

Product reviews are among the most significant components inside the e-commerce industry, as they allow the company to offer new items or Product reviews are among the most significant components inside the e-commerce industry as they allow the company to offer new items or delete products that customers dislike. Considering this as an challenge. Arwa S.M Alqahtani(Alqahtani; 2021) conducted a study on the Amazon opinions records and emotional categorization using machine learning algorithms. Meanwhile, Abdhallah Nyero and Joseph have predicted Internet goods purchases (Bada et al.; 2020). Comments were converted into vectors, which were then trained inside the model to assess outcomes. This article is based on client comments of her particular device, and the reviews when you added from a specific website, the set of data was quite small, and lastly. By using methods given in the article this paper can assist consumers in retaining the finest items.

Consumers tend to provide comments in the form of user ratings and publish it on the corporate website or on social networking sites for the public benefit because there is little contact amongst customers and firms in this he commerce sector. To comprehend client input, analyse it, and create better strategies based upon this. Emotional and qualitative methods of African cellular e-commerce applications was conducted by Tolulope Olagunju, Rita Orji, (Olagunju et al.; 2020) and Auon Haidar kazmi, Gautam Shroff (Kazmi et al.; 2016) used ML algorithms and lexicon based techniques to anticipate client behaviour. Inside this study, the outcomes of machine learning algorithms were contrasted, and various challenges faced by clients but just not associated with business management, such as marketing, logistics, product classification, and so on were analysed.

Comparable to the preceding example emotional analysis was done upon embedded device reviews in (rong et al.; 2021), where client opinions were originally separated into good and poor divisions, and this text was afterwards classified into phases and phrases to vector, vectors to phrases such data are fed into neural nets to develop the model. Convolutional neural networks are utilized to discover the link between the information set

and the subjective rating of consumer perceptions. One such research mostly concentrated on (Lu et al.; 2014) how to construct a neural network for emotion recognition and determine the relationship among and consumer behaviour. However, the report offers minimal information with narrow product assessments and fails to discover connections in other industry areas

Table 2: Summary of Related Works to sentimental analysis

	Related Works- Sentimental analysis			
Author	Specific Features	Compared Results	Advantages	Limitations
AlQahtani	Product sentiment analysis for amazon reviews	62% results is achieved	Usage of Vector models in sentimental analysis	Reviews where considered for product and dataset was very limited
Olagunju, T., Oyebode	Comparison of sentiment and thematic analysis	82% accuracy is achieved	Webscrapping the data & comparison of machine learning models and lexicon based approaches.	Paper didn't focus on business problems and results are not clearly mentioned
Rong, L., Weibai	Netural Network with emotional tags	60% accuracy is obtained	Usage of CNN with reviews	No clear details about dataset

2.4 Conclusion

The majority of sale prices activity revolves around time series analysis. Although investigation on sales propensity score is rare, several of the methodologies outlined in advertising and medical trials domains can be implemented. This involves using decision trees as well as group approaches like random forests and boosting. Data difficulties including, incomplete and inaccurate data, a large number of variables, the bulk of which are categorical and discontinuous, high dimensionality in tandem with the fluid nature of sales possibilities, and changes in components, such as market launch, were addressed. (Tang, L. And Xu, X. And Rangan, V et al.; 2017). Demand and an optimal machine learning approach that is both resistant to the stated data issues and easy sufficient mortal optimize and evolve. Related research found a substantial preference for commonly described and super learners. (Zhao et al.; 2016; Wang et al.; 2018), due to the stated data and pipeline issues, multi stage modelling designs including a number of techniques for aggregating to neural networks (D'Haen and Van Den Poel et al; 2013)are used. Many studies led in the creation of complicated modelling library (Lawrence et al.; 2010; Tang, L. And Xu, X. And Rangan, V et al.; 2017). This effort will look for a simpler option that is easy to adopt. The preceding studies do not show the usage of tailored models tracking which is the method that this scientific project will take.

3 Methodology Approach Used & Design Specifications

3.1 Introduction

KDD or CRISP-DM procedures are typically used in data mining, research but in this case, KDD works better since developed and delivered in the presentation layer is not applicable. Moreover, the goal for the research is really to add value to companies via the examination of sales and logistics performance logistic performance in e-commerce. Because not clearly linked to any company a customizer KDD technique is applied. Azevedo and Santos (2008)

and the appropriate research approach, which comprises of a two tier architecture, namely the business logic tier and the client tier are described.

3.2 Sales and Logistics Performance Methodology and Design Specification.

The following steps comprise the sales and logistics performance approach (refere to figure 2) for e-commerce (i) data source where sales and logistics data (which is added externally) are combined using SQL and fed into the mysql database,(ii) the data is transferred to Python for extracting features and the data sets logistics performance indicator and exploratory statistical analysis are performed using Python And powerbi, (iii) the final bundle data is then preprocessed and modified in accordance with the process modelling criteria,(iv) Linear Regression, Decision Tree Regression, Gradient Boosting Regressor, Bagging Regression, adaboost Regression and kneighbor Regressor and other ensemble models were trained (v) all models are assessed and evaluated using evaluation metrics.

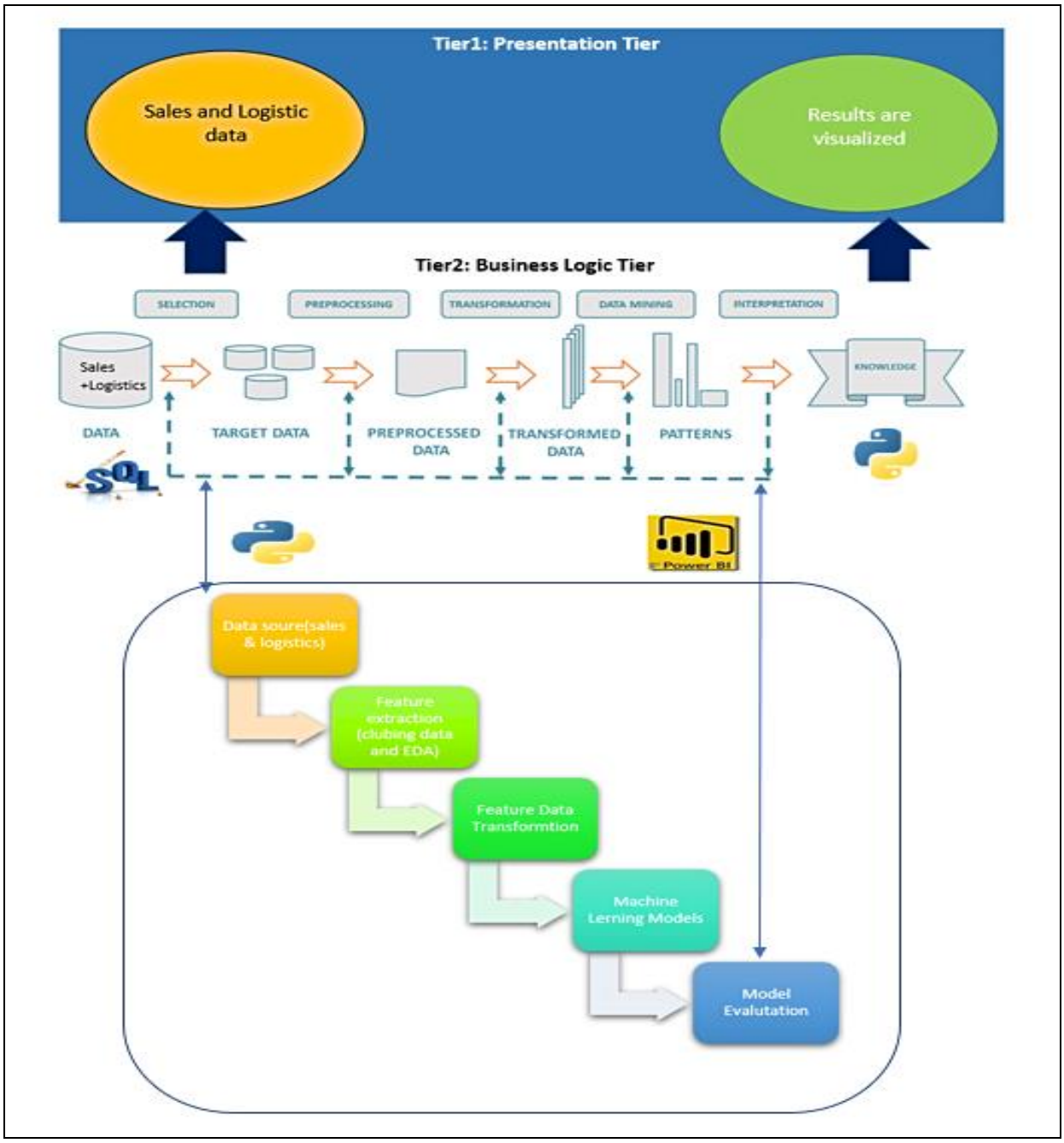

Figure 2: Sales and Logistics Methodology

The e-commerce sales and logistics performance prediction projects design process shown in figure 2 comprises of (i) client tier and (ii) business logic tier. Visuals are used to demonstrate the interpretations of regression models and exploratory data analysis (Python library and powerbi). In the business logic tier data analysis, extraction of features, processing and training of regression models are performed concluded by model assessment.

3.3 Specifications of Design

There are 4 levels altogether. TO start, dataset is downloaded from Kaggle data set and pre-process it using data discovery, irrelevant data, special characters, data encoding, and extraction of features. Following that, several subcategories are created in order to train the dataset .figure 3 depicts the research stream flow chart.The results are then evaluated using

metrics such as r2score, RMSE, MSE, and MAE. Finally, use tabular format to display the results such that they are clearly understood.

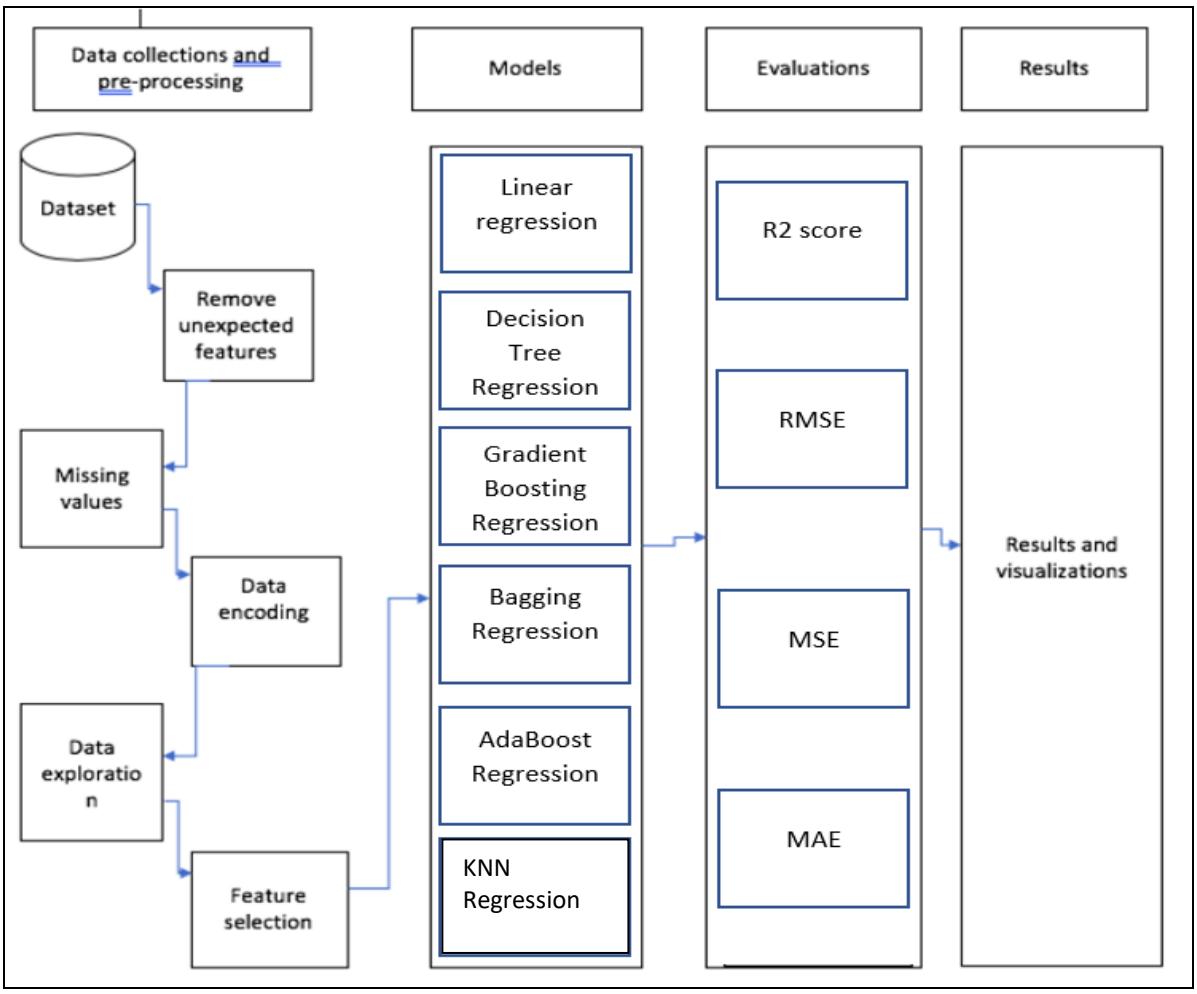

Figure 3:Design Process Flow diagram.

3.4 Conclusions

The sales and logistics methodologies have been changed to meet the objectives of this project, and a modified KDD approach is being utilized for this research. This approach is used to the project design phase and data is gathered from the Kaggle data source. The project framework is based on the two tier design. The following section implements, evaluates, and reports on models for predicting sales and logistic performance

4 Data Preprocessing, Implementation, Evaluation and Results of Sales and Logistics Performance Prediction

4.1 Introduction

This chapter discusses the implementation assessment and results of models used to forecast sales and logistical performance this section extensively explains the feature extraction and their selection in addition to the implementation to assess the models accuracy is utilized as a measure, and the RMSE, MAE, and MSE errors are used to determine which model is doing

well enough to forecast the implemented models are assessed in the last stage and the most accurate prediction is chosen.

4.2 Software Tools Used

SQL: SQL is utilised to govern the info within these database allowing users to get the information that they are seeking whenever it is needed. In this research, SQL is used to store and retrieve the pariticualr information from database and alter the data using SQL queries.

PowerBI: Microsoft power B tool is a set of tools, data adapters and application software that are used to acquire data from many sources, process it, and generate usable reports. the practise of combining data from several sources to generate rich, dynamic reports. Several dashboards were created in PowerBi after some data-processing in python. The csv file is exported to PowerBi from python to access the data.

Python- Python has many advantages for data analytics, such as many available libraries for different software like image processing, machine learning, data visualization. As a part of it python is used to build machine learning models in this research and also perform data pre-process and visualizations

4.3 Data collection and Pre-processing

4.3.1 Understanding the data and loading into the database

This is a transnational set of data containing all purchases for a UK based and licenced non store Internet retailer that occurred during December 2010 to December 2011. The firm primarily sells only one presents. The majority of the company's clients are resellers. The data set can be found on Kaggle as well as and the UCI machine learning repository.Variables in the dataset are Invoice No(transaction Number), Stock code(unique product code), Description(Product description), Quantity, Invoice date, Unit price, Customer ID, and country. This dataset has more than 5,40,000 records. This data is put into a mysql database to be stored, and some logistical data (requested days and supplied days) is added using SQL queries in mysql. This data is then transferred to Python using the psycopg2 module (with the hostname, database, username, password, and port id provided) for further processing.

4.3.2 Loading data into python

Using the psycopg2 package the data set is imported into Python from a PostgreSQL database. The data is then imported into a data frame using the pd.read.sql library. This package is employed by Python to produce SQL queries The data set is still being adjusted to better meet the needs of business representations, such as extracting the month and year from the invoice data and creating a new variable termed logistic performance, which is computed as seen below. Essentially, logistic performance demonstrates how successfully an order is delivered to a client within a specified time frame, this may assist stakeholders in determining where they fall short in terms of logistic performance.

$$\text{Logistic\_ performance} = (\text{requested\_days}/ \text{delivered\_days})*100$$

4.3.3 Data Pre-processing

Data pre-processing is the most important stage in big data, and it takes a large quantity of effort and time.Careful, preparation can result in the production of amazing algorithms.In this research, data preparation involves data exploration, excluding missing values, removing

unusual characteristics, data encoding and feature extraction. The purpose of data investigation is to help us better understand the data.

4.4 EDA (Exploratory Data Analysis) using Power Bi and python

Data Visualisation is accomplished both in powerbi and Python in which powerbi reports and dashboards can be displayed with most dashboards, focusing on sales and logistics performance and overall basic data visuals, whereas in Python visualisation is predicated on economic product and sales revenue, as described below.

Figure 4: EDA of Dataset

Fig 4 is the dashboard prepared in PowerBi, it tells about the overall summary of the dataset which was processed from python after adding a logistic performance variable to the original data. Coming to Fig 4 the dashboard is further divided into 5 parts where the 1st report explains the monthly sales of the company, the 2nd visual explains about top 5 exporting countries which has more sales , the 3rd visual represents about the least logistics which categorised by the number of days order requested and order got delivered, 4th tells about least exporting countries in terms of sales and finally 5th visual is the top 5 products with monthly sales.

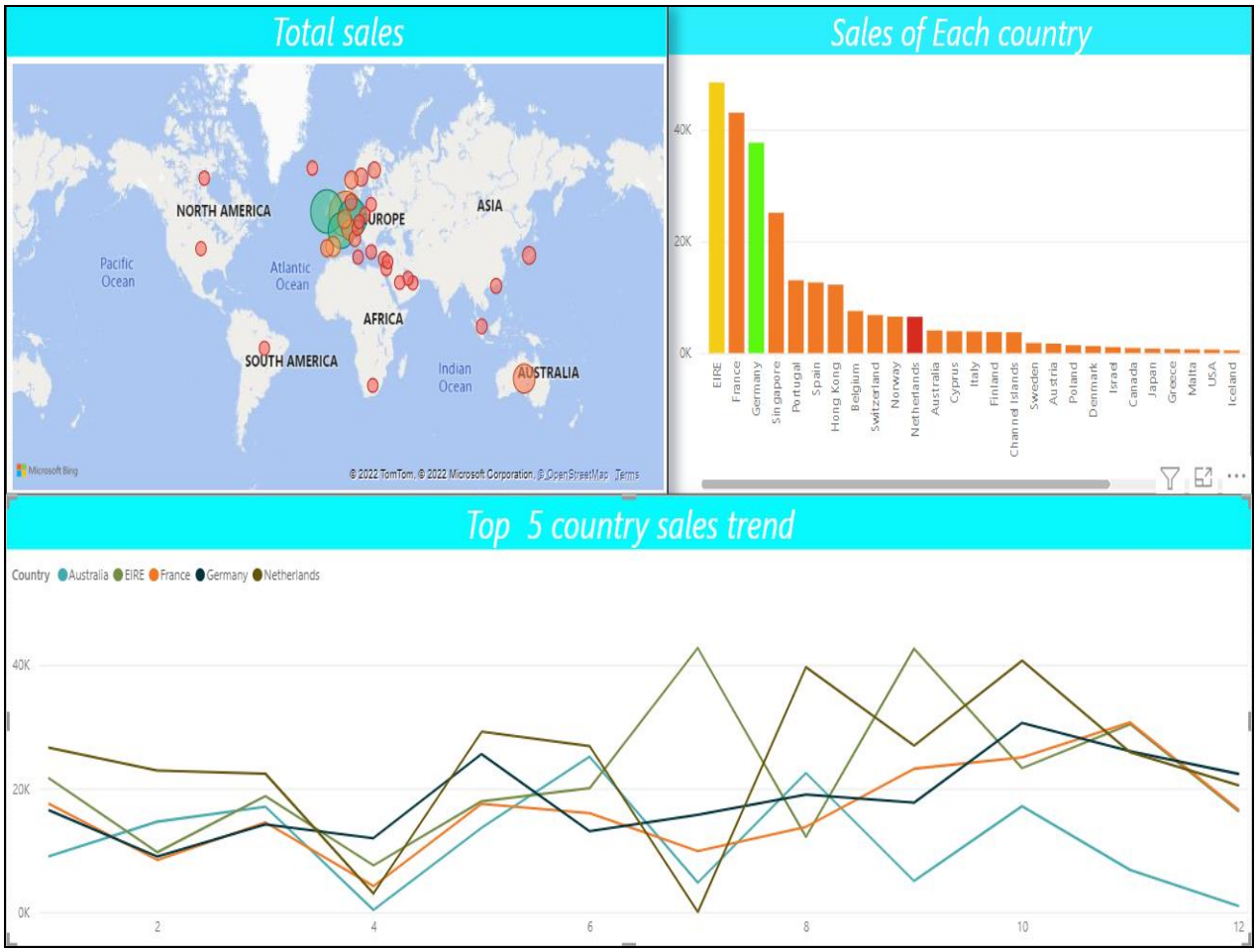

Figure 5: EDA of SALES data.

The 2nd dashboard Fig 5, explains everything about sales, where total sales is shown in world map, bar plot and trend of top 5 countries in 1st ,2nd and 3rd visuals respectively.

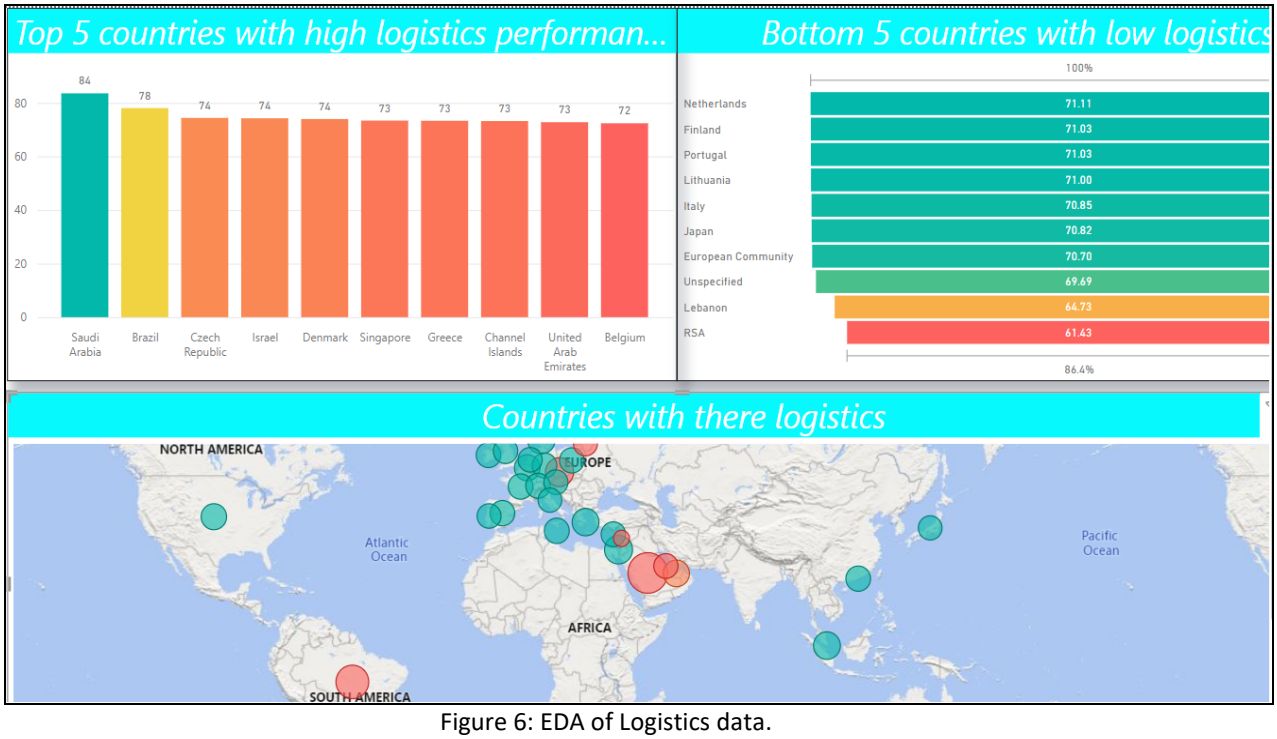

Figure 6: EDA of Logistics data.

The 3rd dashboard Fig 6 explains logistics, where the logistics performance of the top 5 countries and bottom 5 countries are visualized in bar chart and funnel chart respectively , and in world maps are represented, these are valued based on logistics performance.

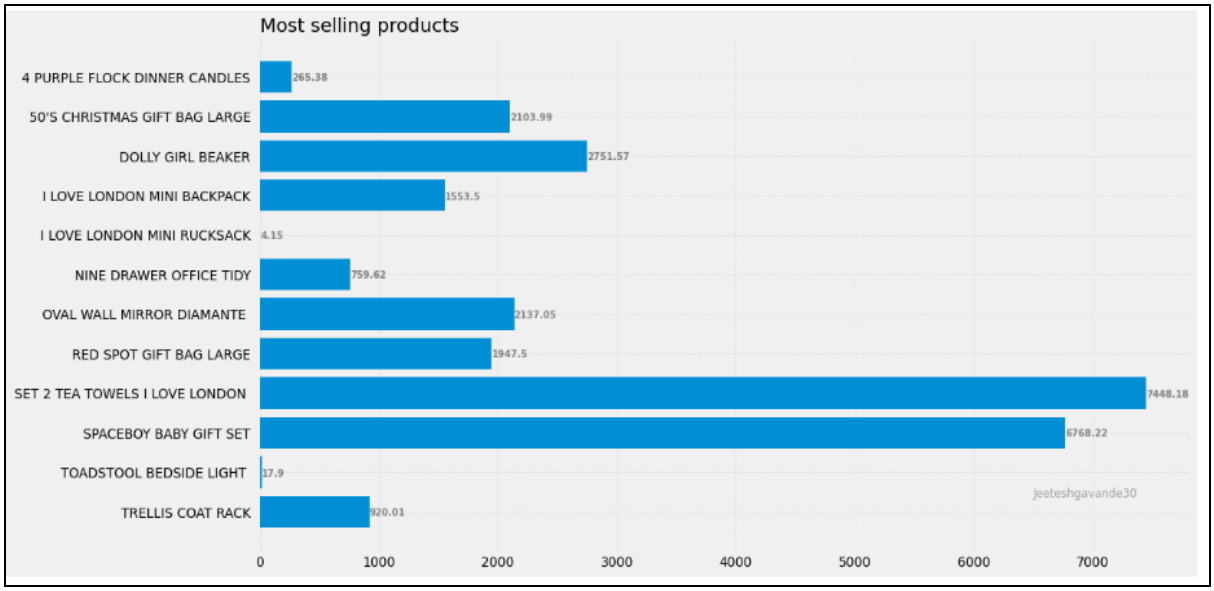

Figure 7: Most selling products

Fig 7, horizontal bar plot was created in python using the matplotlib library, where the plot is categorized based on the top 12 products and total sales made.

4.5 Feature Selection

The data after visualization is then grouped by groupby function in python, where on country, year, month, and aggregated based on average logistics performance, and total sales made by the country. And the final dataset is checked again for correlation, where Principal Component Analysis (PCA) is applied for two variables i.e, sales (PC.Sales) and logistics(PC.Logistics), this data is added to final data frame (df). To reduce errors in the model building square root transformation is opted, whereas log transformation failed to reduce errors in this dataset.

This data is split into train and test set and evaluation metrics is carried on machine learning models. Note, to provide the solution for research question sub-research question, modelling was done twice, one based prediction of sales (research question) and other for predicting logistic performance(sub-research question).

4.6 Evaluation metrics

4.6.1 Mean Absolute Error(MAE):

It is the average of all the absolute discrepancies between the expected and actual values from the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$$

N is the total number of data points
Y_i =actual values
$\hat{Y}_i$ =predicted values

A small value of MAE signifies that the algorithm is excellent at forecasting, but a big MAE indicates that the model might struggle in specific areas, a MAE of 0 indicates that the model accurately predicts the outcomes.
The mean absolute error is measured using the same unit as that of the raw data and must only be evaluated with models with mistakes evaluated within the same units. The greater the MAE the more serious the mistake. It is resistant to outliers as a result by using absolute values MAE can cope with outliers.
A large inaccuracy does not override a large number of little errors in this case, and hence the output gives us with a generally impartial assessment of the way the model is functioning. As a result it fails to penalise the more serious mistake phrases.
Because MAE is not distinguishable we must use discrete optimization techniques such as stochastic gradient.

4.6.2 Mean Squared Error (MSE):

It refers to the sum of the square discrepancies in between observed and expected values the smaller the value the more accurate the regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

N=total Number of data points
Yi is actual values
Ŷi is the predicted values
It's measure is the square of the unit of the parameter. Mean squared error applies the squaring function to each mistake number to eliminate the value and penalise big errors. As we squared the error, the influence of greater mistakes becomes more evident than the impact of lesser errors, thus the model may now concentrate more heavily on the existing and potential errors. The significant cause this isn't particularly helpful is because if we make one particularly terrible forecast, the square will amplify the inaccuracy and may bias the measure towards overvaluing the models awfulness.
Model Evaluation Metrics in Machine Learning | by İrem Tanrıverdi | Analytics Vidhya | Medium

4.6.3 Root Mean Squared Error (RMSE):

The average root square divergence between both the true and forecasted values. the root mean square error is obtained by calculating the square root of MSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

We would like the RMSE value to be as small as possible the smaller RMSE number, forecasts the models forecast. A greater RMSE implies that the projected and actual values diverged significantly.
n is the total number of data points
yj is the actual values
ŷj is the predicted values

4.6.4 R^2 score, the coefficient of determination:

R-square describes how much the variation of one factor matches the variation of the other. in other terms, it determines the share of variable that is explained by the predictor variables in the independent variables

R-squared is a well known statistic for determining model correctness. It indicates how many of the measured values are to the best fit line produced by a regression method. A greater R-squared value denotes a better match this assists us in determining the link between the independent variable and the dependent variable. It is the sum of the squared divided by the entire amount of squares

$$R^2 = 1 - \frac{SSE}{SST}$$

SSE (sum of squared errors) is indeed the total of the squares of the difference between the real and anticipated values.

Yi is the target values, ŷi is the predicted values, and y-bar is the mean values, m is total no of data points

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

SST total sum of squares is indeed the total of the squares of her difference between the true value and it's mean.

$$SST = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

4.7 Conclusion of Implementation

The primary goal of implementation in this research study is to employ a variety of open source technologies that the industry may use. SQL, powerbi and Python are among the tools utilized here. The assessment measures for machine learning models are also thoroughly defined.

4.8 Experiment 1: Implementation of Sales Prediction Models

All machine learning models used in this study were created in Python using the sklearn module, which made it possible to import all machine learning model.The modeling, assessment, and findings were separated into two phases, as stated in the research aim. In other words, for a better understanding, machine learning modeling, assessment, and outcomes were performed on both sales and logistics.

The final gathered data required for modeling is recorded in a data frame, and the data is divided into training and testing. The split ratio was 80% in training and 20% in testing. Sklearn.model selection and train test split packages were utilized in this approach. For sales modelling, independent variables were “Country”, “Month”, “Year” and dependent variable is “Sale”. For logistics modelling, independent variables were “Country”, “Month”, “Year”, “PC.Sale” and dependent variable is “Logistics”.

4.8.1 Exp 1: Implementation, Evaluation And Results Of Linear Regression For Sales

In applying a linear equation to observationally data, linear regression can predict the connectiion between two variables. The term “variable” refers to one of two categories of variables: independent variables and dependent variables. Linear regression is a popular method for performing forecasting. The coefficient of correlation is an indicator of the connexion between two parameters. The coefficients value range is from -1 to +1. This coefficient indicates the strength of the reported data is correlation with two variables
Linear Regression Equation is given below
Y=a+bX

X is the independent variable, and plotted in x axis, Y is dependent variable plotted in y axis.

$$a = \frac{(\sum_Y)(\sum_{X^2}) - (\sum_X)(\sum_{XY})}{n(\sum_{x^2}) - (\sum_x)^2}$$

$$b = \frac{n(\sum_{XY}) - (\sum_X)(\sum_Y)}{n(\sum_{x^2}) - (\sum_x)^2}$$

Implementation of Linear regression for sales:
The "sklearn" module in Python is used to construct linear regression. LinearRegression is the function that is used to implement it (). It is trained using practice data. This model is developed with various feature combinations, including (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe containing all of these metrics.

Evaluation & Result of Linear regression for sales:
The accuracy of the model is around 5.9%, and the validation is nearly equal to the model's accuracy. This model's root mean square error (RMSE) is 111915, its mean square root (MSE) is 1255174522, and its mean absolute error is 59787. The square root transformation is used to reduce error and increase precision.

Implementation. Evaluation & result Linear regression of sales with square root transformation:

The data is transformed using the square root transformation and trained using the same linear regression. The model's accuracy is about 6% (shown in fig 8), with validation virtually equal to model accuracy. The model's root mean square error (RMSE) is 190.55, its mean square root (MSE) is 36310.94, and its mean absolute error (MAE) is 94.64. Before being added to the linear regression model, the values are kept in a dataframe. The regular linear regression model was outperformed by the square root transformation.

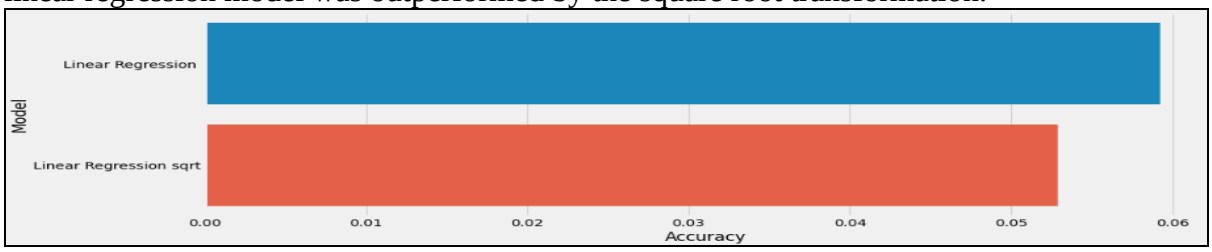

Figure 8: Accuracy of Logistic Models with and without transformation while prediciting sales

4.8.2 Exp 1: Implementation, Evaluation And Results Of Decision Tree Regression For Sales

Decision tree has a hierarchical system comparable to a computer engineering tree it has terminals that are linked together by edges a decision tree categorises data by inquiring at each point the decision tree takes these trays or characteristics and raises the proper questions at the appropriate stage or not to determine the extent to which the prediction/accuracy may be granted to the individual

Implementation of Decision Tree regression for sales:

Python's "sklearn" library is used to implement decision tree regression. DecisionTreeRegressor is the function used to implement it (). It has been trained using practice data. This model is constructed using numerous feature combinations, including (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe that includes all of these metrics.

Evaluation & Results of decision Tree Regression

The accuracy of the model is roughly 97%, and the validation is nearly equal to the model's correctness. This model's root mean square error (RMSE) is 17781.96, its mean square root (MSE) is 316198429, and its mean absolute error is 6201.5. The square root transformation is used to reduce error and increase precision.

Implementation, Evaluation & Results of Decision Tree Regression for sales with square root transformation:

The dataset is transformed with the square root transformation and trained with the same linear regression. The model's accuracy is around 91% (as shown in fig 9), and the validation is nearly equal to the model's accuracy. This model's root mean square error (RMSE) is 57.88, its mean square root (MSE) is 3350, and its mean absolute error (MAE) is 36. Before being added to the Decision Tree model, the values are stored in a dataframe. The Decision Tree regression model was outperformed by the square root transformation.

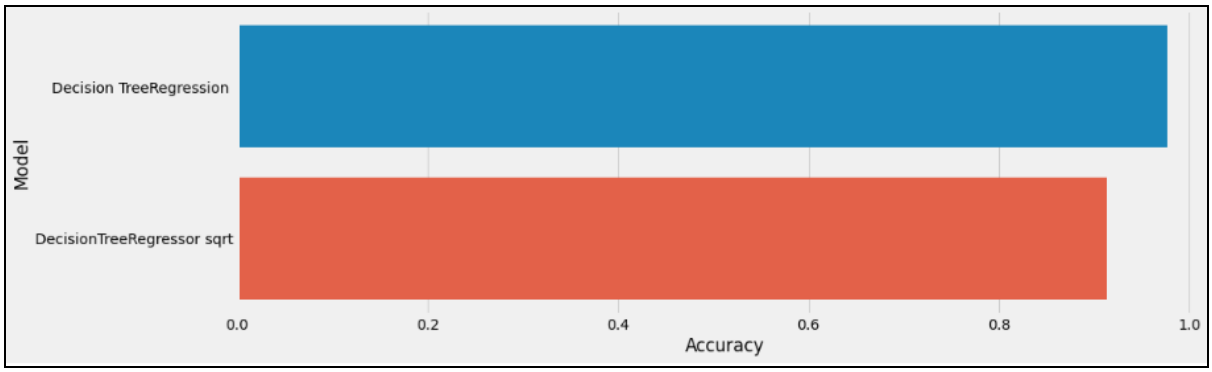

Figure 9: Accuracy of Decision Tree Regression Models with and without transformation while prediciting sales

4.8.3 Exp 1: Implementation, Evaluation and Results of Gradient Boosting Regression for sales:

Gradient boosting regression is a statistical approach used to investigate the connection among the number of variables (X&Y). Its analytical output reveals significant factors the influencing the variables Y. As well as the nature of the connection between such variables as well as the dependent variable.

Implementation of Gradient Boosting regression for sales:

The "sklearn" module is used to implement Gradient Boosting regression in Python. The function used to accomplish it is GradientBoostingRegressor (). It is trained on practice data. This model is constructed using numerous feature combinations, including (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe that includes all of these metrics.

Evaluation & Results of Gradient Boosting regression for sales:

The accuracy of the model is around 98%, with the validation is almost similar to model accuracy and root mean square error(RMSE) of this model is 17367.18, Mean square root(MSE) is 30161274, Mean absolute error for this model is 5742.45. In try to reduce error and increase accuracy the square root transformation is implemented.

Implementation, Evaluation, & Results of Gradient Boosting regression for sales with square root transformation:

The square root transformation is applied to the dataset and trained with same linear regression,The accuracy of the model is around 98%(shown in fig 10), with the validation is almost similar to model accuracy and root mean square error(RMSE) of this model is 47.32, Mean square root(MSE) is 2239 Mean absolute error(MAE) for this model is 47. The values are stored in a dataframe and appended to the Gradient Boosting regression model. The square root transformation performed better than normal Gradient Boosting regression model.

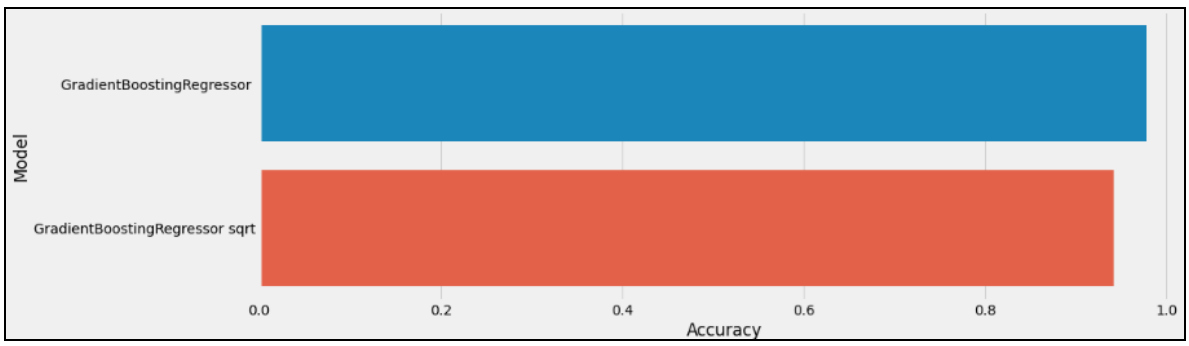

Figure 10: Accuracy of Gradient Regression Models with and without transformation while prediciting sales

4.8.4 Exp 1: Implementation, Evaluation And Results Of Bagging Regression For Sales:

A bagging regressor is an array recursive that fits base regresses to randomized portions of the entire data and then aggregates. Their individual forecasts (through voting or averaging) to generate a final result. A meta estimated of this type is often used to minimise the variability of a black box estimator (for example decision tree) by incorporating randomised into its building mechanism and then constructing ensemble from it.

Implementation of Bagging Regression for sales:
Python's "sklearn" package is used to perform tagging regression. BaggingRegressor is the function used to do this (). It has been trained using practice data. This model is built with a variety of feature combinations, such as (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe including all of these metrics.

Evaluation & Results of Bagging Regression for sales:
The model's accuracy is around 97%, and the validation is virtually equivalent to the model's accuracy. The model has a root mean square error (RMSE) of 16368, a mean square root (MSE) of 267924441, and a mean absolute error of 5730. 2. The square root transformation is used to decrease error and improve precision.

Implementation, Evaluation & results of bagging regression for sales with square root transformation:
The dataset is processed using the square root transformation and trained using the same linear regression algorithm. The model's accuracy is roughly 97% (as seen in fig 11), and validation is nearly as good. The root mean square error (RMSE) for this model is 50.72, and the mean square root (MSE) is 2570.8. This model's mean absolute error (MAE) is 31. The values are captured in a dataframe before being incorporated in the Bagging regression model. The square root transformation outperforms the standard Bagging regression model.

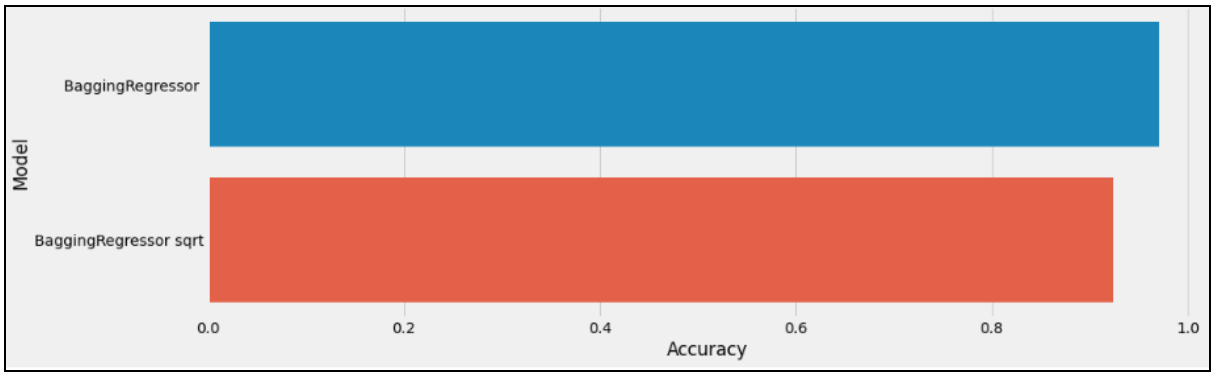

Figure 11: Accuracy of Bagging Regression Models with and without transformation while prediciting sales

4.8.5 Exp 1: Implementation, Evaluation and Results of AdaBoost Regression for sales

Adaboost also known as adaptive boosting is a machine learning approach that is utilised as an optimization learning. The most frequently Adaboost method his decision trees via one level which is decision trees only with one splitting. These trees are often referred to as decision's stumps. This method creates her model and provides weighting factor to all pieces

of data. It then applies higher weights to wrongly identified points. All variables with higher weights are weighed more heavily in the following method. It will keep training models still a lower error his reported.

Implementation of Ada Boost regression for sales:
The "sklearn" module is used to implement AdaBoost regression in Python. The AdaBoost Regressor function is used to implement it (). It is trained on practice data. This model is constructed using numerous feature combinations, including (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe that includes all of these metrics.

Evaluation & Results of AdaBoost Regression for sales:
The model's accuracy is approximately 98%, and the validation is practically identical to the model's accuracy. The root mean square error (RMSE) of this model is 11000, the mean square root (MSE) is 121011469, and the mean absolute error for this model is 5766.7. The square root transformation is used to minimize error and boost accuracy.

Implementation, Evaluation & results of AdaBoost Regression for sales with square root transformation:
The dataset is transformed with the square root transformation and trained with the same linear regression. The accuracy of the model is roughly 91% (as seen in fig 12), with validation nearly matching model accuracy. This model's root mean square error (RMSE) is 56.02, and its mean square root (MSE) is 3138.8. The mean absolute error (MAE) for this model is 37.35. Before being added to the AdaBoost regression model, the results are kept in a dataframe. The AdaBoost regression model was outperformed by the square root transformation.

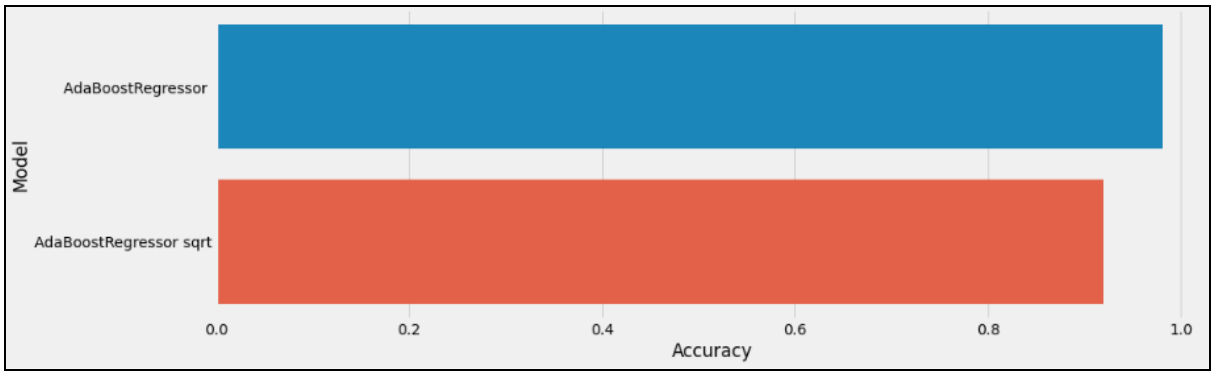

Figure 12: Accuracy of AdaBoosting Regression Models with and without transformation while prediciting sales

4.8.6 Exp 1: Implementation, Evaluation and results of KNeighbors Regression For sales:

KNN has really been utilised as a nonparametric approach in statistical estimates and pattern matching since the early 1970s. Methodology the averaging of the numerical targets of the K nearest neighbours is a basic execution of KNN regression. Another method is to take an inverse distance necessary to balance of the K closest neighbors. The very same distance measures are used in KNN regression as in KNN classifier.

Implementation of KNeighbor Regression:
The "sklearn" module is used to implement KNeighbor regression in Python. The function used to accomplish it is KNeighborRegressor (). It is trained on practice data. This model is constructed using numerous feature combinations, including (1) model validation score, (2) R2 score, (3) root mean square error (RMSE), (4) mean squared error (MSE), and (5) mean absolute error (MAE), as well as a dataframe that includes all of these metrics.

Evaluation & Result of KNeighbor Regression for sales:
The model is around 88% accurate, and the validation is nearly similar to the model's accuracy. This model's root mean square error (RMSE) is 50938, its mean square root (MSE) is 2594703152, and its mean absolute error is 84.8. The square root transformation is used to reduce error and increase precision.

Implementation, Evaluation & Results of KNN regression for sales with square root transformation:
The dataset is transformed with the square root transformation and trained with the same linear regression. The model's accuracy is around 25% (shown in fig 13), with validation nearly equal to model accuracy. The model's root mean square error (RMSE) is 168.6, and the mean square root (MSE) is 28436.8. The mean absolute error (MAE) for this model is 84.8. Before being added to the KNeighbor regression model, the values are kept in a dataframe. The KNeighbor regression model was outperformed by the square root transformation.

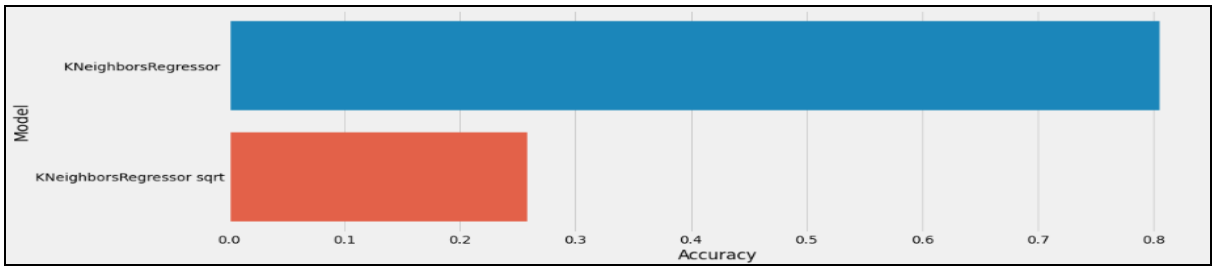

Figure 13: Accuracy of KNeighbor Regression Models with and without transformation while prediciting sales

4.9 Comparison of Implemented model for predicting sales:

The results and errors for regression models adopted to forecast sales, i.e., before and after data transformation, are given in Figure 15. The accuracy of each model is compared in Figure 14, and the errors of each model are plotted in a line chart in Figure 16. As shown, gradient boosting and AdaBoost regression fared well in terms of accuracy and error reduction, but linear regression performed poorly.

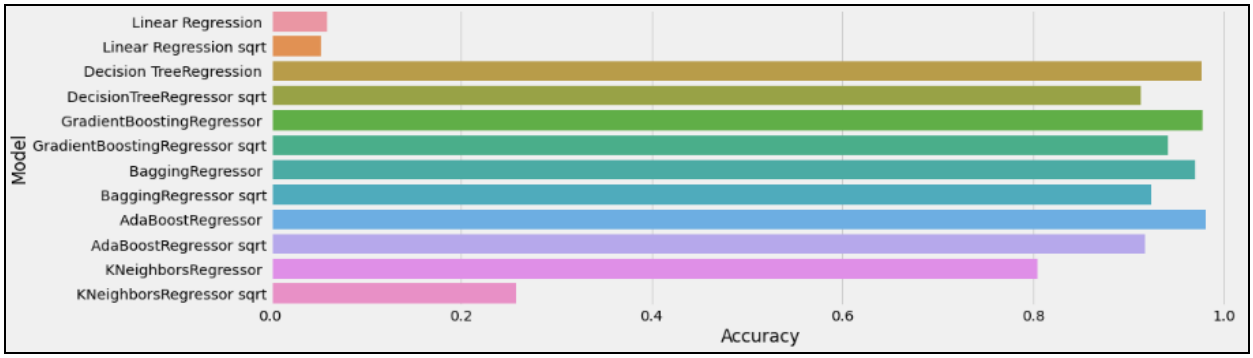

Figure 14: Accuracy results of all the models for predicting sales

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	59784.77	12525174522.17	111915.93	0.06
1	Linear Regression sqrt	94.65	36310.94	190.55	0.05
2	Decision TreeRegression	6201.57	316198429.53	17781.97	0.98
3	DecisionTreeRegressor sqrt	35.70	3326.37	57.67	0.91
4	GradientBoostingRegressor	5742.46	301619274.50	17367.19	0.98
5	GradientBoostingRegressor sqrt	26.22	2238.69	47.31	0.94
6	BaggingRegressor	6424.46	304778733.88	17457.91	0.98
7	BaggingRegressor sqrt	33.10	2829.38	53.19	0.93
8	AdaBoostRegressor	6523.99	188639727.58	13734.62	0.99
9	AdaBoostRegressor sqrt	37.13	3092.60	55.61	0.92
10	KNeighborsRegressor	13634.62	2594703152.12	50938.23	0.81
11	KNeighborsRegressor sqrt	84.88	28436.85	168.63	0.26

Figure 15: Results of Implemented Models to predict Sales.

Figure 16: Comparision of Errors based on implemented models

4.10 Experiment 2: Implementation Of Logistics Performance Prediction Models

At the first step of this procedure, the variables used to train and test the models for predicting sales are 'Country', 'Year', 'Month', and 'Sale' and for predicting logistics are 'Country', 'Year', 'Month', 'PC.SALE', 'Logistic Performance'.

4.10.1 Exp 2: Implementation, Evaluation & results of linear regression for logistics performance

The linear regressor implementation approach for predicting logistics performance is the same as linear regression for predicting sales. The accuracy of the model is around 48%, and the validation is nearly equal to the model's accuracy. The root mean square error (RMSE) of the model is 15.28, the mean square root (MSE) is 233.57, and the mean absolute error is 10.06. The measurements were stored in a dataframe.

4.10.2 Exp 2: Implementation, Evaluation & results of Decision Tree regression for Logistics

The Decision Tree regressor implementation procedure is the same as the Decision Tree regression implementation approach for sales prediction. The model's accuracy is about 98%, and the validation is practically identical to the model's accuracy. The model's root mean square error (RMSE) is 1.82, its mean square root (MSE) is 3.31, and its mean absolute error is 1.18. A dataframe was used to hold the measurements.

4.10.3 Exp 2: Implementation, Evaluation & results of Gradient Boosting regression for Logistics

Gradient Boosting regression for sales prediction is the same as Gradient Boosting regressor implementation for logistics performance prediction. The accuracy of the model is around 96%, and the validation is nearly similar to the model's accuracy. The root mean square error (RMSE) of the model is 1.82, the mean square root (MSE) is 3.32, and the mean absolute error is 1.14. The measurements were stored in a dataframe.

4.10.4 Exp 2: Implementation, Evaluation & results of Bagging regression for Logistics

The implementation of a bagging regressor for logistics performance prediction is the same as that of a bagging regression for sales prediction. The accuracy of the model is around 96%, and the validation is nearly similar to the model's accuracy. The root mean square error (RMSE) of the model is 3.8, the mean square root (MSE) is 14.716, and the mean absolute error is 1.64. The measurements were stored in a dataframe.

4.10.5 Exp 2: Implementation, Evaluation & results of AdaBoost regression for Logistics

The implementation of AdaBoost regressors for logistics performance prediction is the same as AdaBoost regression for sales prediction. The accuracy of the model is roughly 97%, and the validation is nearly equal to the model's accuracy. The root mean square error (RMSE) of the model is 3.3, the mean square root (MSE) is 11.38, and the mean absolute error is 2.9. The measurements were stored in a dataframe.

4.10.6 Exp 2: Implementation, Evaluation & results of KNeighbor regression for Logistics

The implementation of KNeighbor regressor for logistics performance prediction is the same as KNeighbor regression for sales prediction. The accuracy of the model is around 47%, and

the validation is nearly similar to the model's accuracy. This model's root mean square error (RMSE) is 17, its mean square root (MSE) is 313.38, and its mean absolute error is 12.14. The measurements were stored in a dataframe.

4.11 Comparison of Developed models for predicting logistics performance

After all models were applied and assessed, the results were appended to a single data frame, as previously mentioned, so that all metrics could be exhibited and compared at the same time, as seen in Fig 17 and Fig 18.

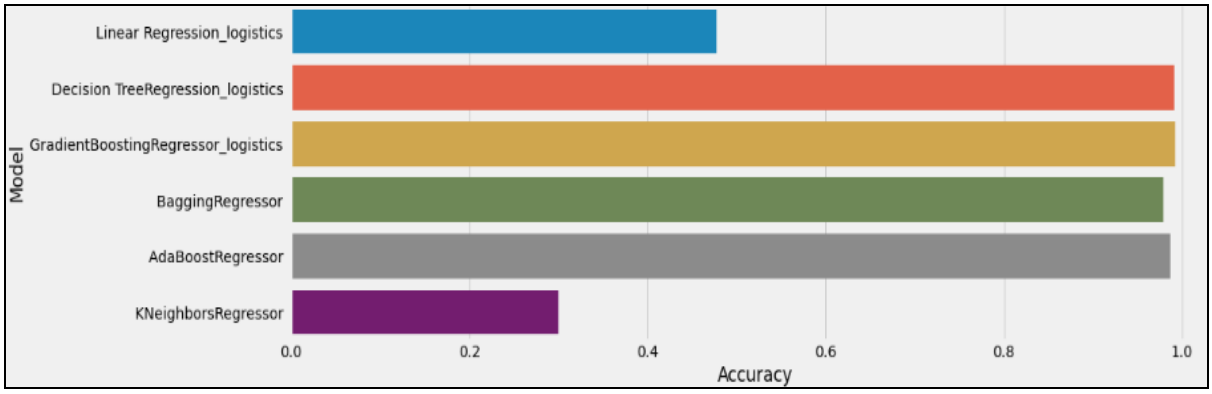

Figure 17: Accuracy results of all the models for predicting logistics performance

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression_logistics	10.06	233.57	15.28	0.48
1	Decision TreeRegression_logistics	1.19	3.32	1.82	0.99
2	GradientBoostingRegressor_logistics	1.14	3.34	1.83	0.99
3	BaggingRegressor	1.65	14.72	3.84	0.97
4	AdaBoostRegressor	2.93	11.38	3.37	0.97
5	KNeighborsRegressor	12.14	313.33	17.70	0.30

Figure 18: Results of Implemented Models to predict Logistics.

In addition to these models, model metrics were reviewed, as detailed in the configuration handbook.

4.12 Conclusion For Implementation, Evaluation And Results

The “Sklearn” package in Python is used to import all of the previously stated machine learning models. To decrease the models mistakes the square root transformation is used to predict sales. All of the modelling metrics were provided to data frames such that, in the ending, all of the metrics for each model could be appended to hand one data frame and the results could be viewed as shown in figures 15 and 18.

5 Conclusion And Future Works

5.1 Conclusion

The major objective of this study is to estimate sales and logistics for an ecommerce firm, as well as to employ a variety of open source technologies that may be used by small and medium sized businesses to assist them achieve productive outcomes. SQL Python And

powerbi were the tools employed in this study. The methodology began with downloading the data set and uploading it to SQL, where minor adjustments were made to the data set. This data is then retrieved in Python using the pscycpg2 package. The data set was pre processed in Python in the next stage, and data visualisation was done in powerbi Python. Python was used to implement the machine learning model 6, the models created are Linear Regression, Decision Tree Regression, Gradient Boosting Regression, Bagging Regression, Ada Boosting Regression, And Knn Regression.

Model training for sales prediction was done in two steps: first with original data and then with square root modification. PCA transformation has been used to sales factors for logistics prediction. In terms of outcomes Adaboosting, bagging regressor and gradient boosting did extremely well in terms of accuracy about 97%, but the error rate was too high to lower the error a square root transformation was done as explained in (4), the results show that Adaboost, bagging regression and gradient boosting regression have an accuracy of around 93% with very low errors, whereas for predicting logistics performance gradient boosting regression, bagging regression and decision tree regression have performed extremely well with an accuracy of around 97% and less error. Kneighbour regression and linear regression were the models that fared poorly in predicting both sales and logistics performance (4) other machine learning models were not included in the evaluation (4) and implementation stages but were shown in the configuration manual.

5.2 Future Work

Because the E-commerce business is so large in reality, I will attempt to do emotional analysis based on customer reviews on items in addition to predicting sales and logistical performance in this research. More transformation approaches will be enforced in the future to see whether mistakes can be decreased while creating machine learning models.

6 Acknowledgement

I’d like to thank Dr.Catherine Mulwa in particular for her supervision, advice, and support during the project. I’d want to thank my mother, father and sister for their faith in me

References

Bada, J.K., Nyero, A., Asianzu, E. and Namuwaya, H.A. (2020). *Predicting Online Purchase Intentions in Business Administration Graduate Students*. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/9144034/authors [Accessed 14 Dec. 2022].

Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, 71, pp.416–428. doi:10.1016/j.eswa.2016.11.010.

Bohanec, M., Robnik-Šikonja, M. and Kljajić Borštnar, M. (2017). Decision-making framework with double-loop learning through interpretable black-box machine learning

models. *Industrial Management & Data Systems*, 117(7), pp.1389–1406. doi:10.1108/imds-09-2016-0409.

Bohanec, M., Robnik-Sikonja, M. and Kljaji´c Borˇstnar, M. (2017c). Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting, Organizacija 50(3): 217–233. URL: http://organizacija.fov.uni-mb.si/index.php/organizacija/article/viewFile/780/117

D'Haen, J. and Van den Poel, D. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. *Industrial Marketing Management*, 42(4), pp.544–551. doi:10.1016/j.indmarman.2013.03.006.

Duncan, B.A. and Elkan, C.P. (2015). Probabilistic Modeling of a Sales Funnel to Prioritize Leads. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [online] doi:10.1145/2783258.2788578.

Gurnani, M. (2017). *(PDF) Forecasting of sales by using fusion of machine learning techniques*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/320653023_Forecasting_of_sales_by_using_fusion_of_machine_learning_techniques.

Hendra, Rini, E.S., Ginting, P. and Sembiring, B.K.F. (2017). Impact of eCommerce service quality, recovery service quality, and satisfaction in Indonesia. *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*. doi:10.1109/siet.2017.8304105.

Kazmi, A.H., Shroff, G. and Agarwal, P. (2016). *Generic Framework to Predict Repeat Behavior of Customers Using Their Transaction History*. [online] IEEE Xplore. doi:10.1109/WI.2016.0072.

Lawrence, R., Perlich, C., Rosset, S., Khabibrakhmanov, I., Mahatma, S., Weiss, S., Callahan, M., Collins, M., Ershov, A. and Kumar, S. (2010). Operations Research Improves Sales Force Productivity at IBM. *Interfaces*, [online] 40(1), pp.33–46. Available at: https://www.jstor.org/stable/40599236 [Accessed 14 Dec. 2022].

Lu, C.-J. (2014). Sales forecasting of computer products based on variable selection scheme and support vector regression. *Neurocomputing*, [online] 128, pp.491–499. doi:10.1016/j.neucom.2013.08.012.

Lu, C.-J. and Kao, L.-J. (2016). A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer

server. *Engineering Applications of Artificial Intelligence*, 55, pp.231–238. doi:10.1016/j.engappai.2016.06.015.

Lu, C.-J., Lee, T.-S. and Lian, C.-M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. *Decision Support Systems*, 54(1), pp.584–596. doi:10.1016/j.dss.2012.08.006.

Nanduri, J., Liu, Y.-W., Yang, K. & Jia, Y. (2020), Ecommerce fraud detection through fraud islands and multi-layer machine learning model, in K. Arai, S. Kapoor & R. Bhatia, eds, 'Advances in Information and Communication', Springer International Publishing, Cham, pp. 556–570.

Olagunju, T., Oyebode, O. and Orji, R. (2020). Exploring Key Issues Affecting African Mobile eCommerce Applications Using Sentiment and Thematic Analysis. *IEEE Access*, pp.1–1. doi:10.1109/access.2020.3000093.

Pondel, M. & Pondel, J. (2021), Machine learning solutions in retail ecommerce to increase marketing efficiency, in M. L. Owoc & M. Pondel, eds, 'Artificial Intelligence for Knowledge Management', Springer International Publishing, Cham, pp. 91–105.

Rong, L., Weibai, Z. and Debo, H. (2021). *Sentiment Analysis of Ecommerce Product Review Data Based on Deep Learning*. [online] IEEE Xplore. doi:10.1109/IMCEC51613.2021.9482223.

S. M. AlQahtani, A. (2021). Product Sentiment Analysis for Amazon Reviews. *International Journal of Computer Science and Information Technology*, 13(3), pp.15–30. doi:10.5121/ijcsit.2021.13302.

Tkáč, M. and Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, 38, pp.788–804. doi:10.1016/j.asoc.2015.09.040.

Xia, M., Zhang, Y., Weng, L. and Ye, X. (2012). Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. *Knowledge-Based Systems*, 36, pp.253–259. doi:10.1016/j.knosys.2012.07.002.

Xu, X., Tang, L. and Rangan, V. (2017). Hitting your number or not? A robust & intelligent sales forecast system. *2017 IEEE International Conference on Big Data (Big Data)*. doi:10.1109/bigdata.2017.8258355.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S. and Yang, X. (2015). On Machine Learning towards Predictive Sales Pipeline Analytics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). doi:10.1609/aaai.v29i1.9455.

Zhao, P., Su, X., Ge, T. and Fan, J. (2016). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials*, [online] 47, pp.85–92. doi:10.1016/j.cct.2015.12.012.