Analysis of Online Retail market and Consumer Behaviour Analysis

Piyush Daulat Dhawad
Student ID x21236933
Data Mining and Machine Learing - 1
National College of Ireland
x21236933@student.ncirl.ie

***Abstract*—Machine learning appears as a driving force behind innovation as online retail maintains its position as the industry leader in commerce. Innovations in demand forecasting, fraud detection, personalised shopping experiences, and customer service automation characterise this dynamic partnership. Large-scale datasets are analysed by machine learning algorithms, which then use the patterns they find to provide personalised product suggestions, improve inventory control, bolster security, and automate customer service.**

In this report we will under the impact of various feature on the decision of customer's buying decision in classification machine learning. Furthermore we will understand the impact of various variable on the price of a particular product in regression machine learning model.

I. Introduction

The world of commerce has changed dramatically in recent years, and online retail is now a major player. Customers' purchasing habits have changed as a result of the e-commerce platforms' accessibility, ease, and wide range of options. Online retail's success has been largely attributed to the emergence of e-commerce behemoths like Amazon, Alibaba, and others. These platforms have impacted the whole industry by establishing new benchmarks for product diversity, logistics, and customer service.

Online retailers are leveraging data analytics and artificial intelligence to offer personalized shopping experiences. From tailored product recommendations to customized marketing messages, personalization is key to engaging and retaining customers in a crowded digital marketplace.

Social commerce is the result of social media's incorporation with online retail. Direct purchasing is now made possible by social media sites like Facebook and Instagram, which conflate social interaction with shopping. Retailers may interact with customers in fresh and creative ways and reach a wider audience with the help of social commerce.

The global economy now includes the internet retail industry as a vital and dynamic component. Adaptability and creativity will be crucial for organisations operating in this sector as technology advances and consumer expectations change. The path ahead entails overcoming obstacles, grasping chances, and remaining aware of the dynamic demands and inclinations of the contemporary customer. Retail online is an experience, not just a transaction, and companies that recognise and embrace this paradigm will continue to influence how commerce develops in the future.

There are still plenty of interesting growth prospects in the internet retail space. Businesses can prosper by focusing on niche markets, going global, and using cutting-edge technologies. Using advances in artificial intelligence (AI), virtual reality (VR), and augmented reality (AR) can help merchants stand out from the competition and improve the overall shopping experience.

In this report we will take a look at Amazon's, HankyPanky's and Macys online retail business for inner wears and will apply various machine learning models to understand the consumer's buying decision and understand the impact of various variables on price of a product.

Following research questions will be answered with the help of this report:-

1. - Which is the most accurate and efficient classification and regression model to predict the purchase and the price of the product respectively?

2. - Which factors play huge part in the regression model of price?

3. - Which factors play part in the classification model for purchase of the product?

4. - What are the highest prices of the products in each dataset?
5. - What are the rating distribution of the each online retails

website?

II. RELATED WORK

A paper from R. Shi and C. Zhang discusses that forecasting the future sales of retail is important in terms of managing strategic planning, decision making and cost management. They have further citied that the some scholars delieve that non linear are better at predicting the price rather than a linear model. Technique such as ARIMA and BVAR are used to predict short term sales. While other groups of scholars uses machine learning and deep learning models where the variables have high correlation. R. Shi and C. Zhang have used XGBoost for regression machine learning and RandomForest for classification machine learning.

A paper produced by Y. Yang showcases the use of multivariate linear regression model which he has used to predict the consumption of Aero-Material. Here he has applied the linear regression model and used various statistical evaluation matrix to validate the model such as P test, F-test, t-test and residual analysis. He has further claimed that the model used is reproducible for various industries and various prediction problems.

R. Vyas and R. As produced a paper where they forecasted the seasonal sales of multinational retail store chain such as Amazon, Walmart etc. In their study they have found that 82 percent of their customers would buy higher in seasons like black Friday sales and other holiday season. They have transformed the data into the required data type such as they have introduced the datetime column and they have converted the categorical variable into numerical variable to suit the regression model. For evaluation of the model they have decided to take into account Weighted Mean Absolute Error and Root Mean Square Error. Furthermore they have taken into account ARMIA, Holt-Winter model in time series analysis and Linear Regression and random forest in Regression model. They have concluded that various models will bring their own characteristics and it depends on the business needs to choice various model.

N. Mohan and V. Jain cited various papers where they have found that SVM is very useful in predicting the disease which can be applied to predict weather the customer will buy a product or not. The SVM is trained used a split of data for training and testing. The model is trained on 4 different kernel i.e. Linear, Polynomial, RBF and Sigmoid. We can see that BRF model bought most efficient result i.e. 82 percent accuracy.

K. He and C. He have produced a paper where they have used Logistic Regression model and linear Regression model to predict the house sold and price of the house which can be used for retail sales and price of the retail business. They have cited that the classification will give 100 percent accuracy when considering the price and the locality of the house and can still give more than 85 percent accuracy when considering 2 council close to each other.

III. METHODOLOGY

There are various methodology available to approach the Machine Learning related questions. CRISP-DM and KDD are name of the few. Both the approaches are similar to each other with few differences. In this particular problem solving we will take a look at CRISP-DM method.
The Cross-Industry Standard Process for Data Mining, or CRISP-DM, is an established and extensively used technique that provides guidance for the full process of data mining or machine learning. It offers a thorough and organised framework that aids in the planning, carrying out, and assessment of data mining initiatives for businesses and data scientists. CRISP-DM is an open standard that was created by a group of professionals in the field.
The CRISP-DM methodology consists of following steps label=$\star$

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

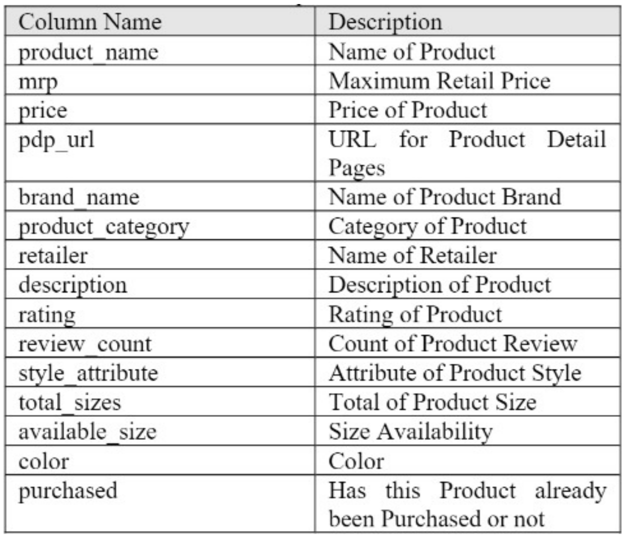
A. *Business Understanding*

In terms of Machine Learning and data mining business understanding comes in the initial phase of the project where the focus is on gaining a comprehensive understanding of the business problem and opportunity at hand. This stage is essential for coordinating technical efforts with the organization's overall objectives. The main objective during business understanding includes:-
I) Define Objective :- This step includes understanding of the business need and defining the objective according to the business need. In our example the objective can be defined as - What is the price bracket where the customer buy the product?
How can the business increase the price of a product without effecting the sale of a product?
Considering these are the business needs we can now clearly work on achieving these in the next step of the Business understanding
II) Translate Business goals into Analytical Goals:- Once the business needs are defined we can move on to defining the goals in terms on analytical goals to take the actionable insights from the given data. In our case our analytical goals can be translated as - Which is the most accurate and efficient classification and regression model to predict the purchase and the price of the product?
- What factors play huge part in regression model i.e. which factors affect the price the most?
- What variables play part in classification model i.e. which variable affect the purchase of the product?

There are further steps included into the business understanding such as identify the stakeholder, assessing the resource constrain, defining the success criteria, risk assessment and initial data exploration.

B. Data Understanding

After we clearly define the business role and analytical goal it is time to understand the data set that is provided. In this report we have total 3 datasets which are amazon.com, hankypanky.com and macys.com 's online retail store dataset. This data set included the sales of the inner wears.
The detailed contain of the data can be found in the table below.



Column Name	Description
product_name	Name of Product
mrp	Maximum Retail Price
price	Price of Product
pdp_url	URL for Product Detail Pages
brand_name	Name of Product Brand
product_category	Category of Product
retailer	Name of Retailer
description	Description of Product
rating	Rating of Product
review_count	Count of Product Review
style_attribute	Attribute of Product Style
total_sizes	Total of Product Size
available_size	Size Availability
color	Color
purchased	Has this Product already been Purchased or not

Fig. 1. TABLE TEMPLATE FOR RETAIL DATA

We can see from table contain we can see that there are many variable such as product name, price, rating, review count and purchased column which are useful for the machine learning model.

- Amazon.com Dataset
  On further inspection of the data we can see that there are more patterns which we can visualize.

  From Fig 2. we can see that the amazon product's price distribution the product are much more in the range of 55 USD and 68 USD. Futhermore we can see that there are far less products in the range of 10 USD to 30 USD. Therefore we can assume that the sales are far more in the range where the products listed are more i.e. in the range of 55 USD to 68 USD.
  From Fig 3. we can see that the amazon product's rating is more in the range of 4.2 to 4.5 on the scale of 5 which indicate high satisfaction of the end user in this case. Further to this observation we can say that even the price of the amazon's product is high the average rating of the products is high.

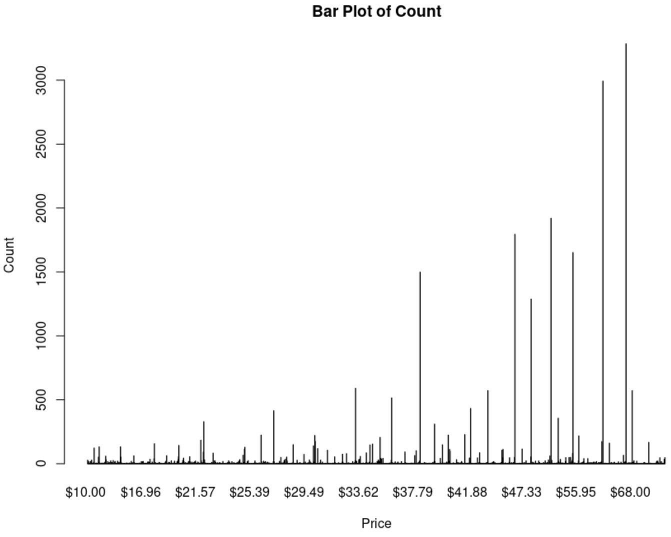


Fig. 2. AMAZON PRICE DISTRIBUTION

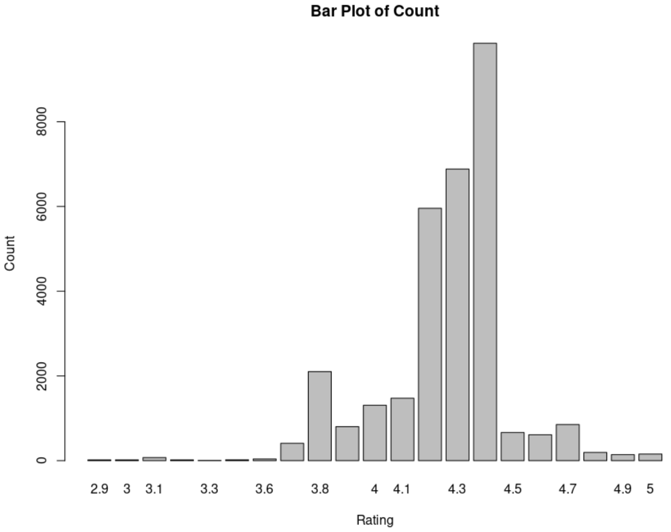


Fig. 3. AMAZON PRICE DISTRIBUTION

- HankyPanky.com and Macy.com
  If we carry similar observations onto HankyPanky product we can see a interesting pattern.
  If we observe fig 4 we can see that in case of Hanky Panky's products we see spike of population in range of 12 USD to 16 USD. Similary we can see that in case of Macy's product the price distribution can be seen more heavy in the range of 10 USD to 18 USD and some small spikes at around 25 USD and 37 USD.
  While this is true we can see that the review rating of Hanky Panky's is heavily populated towards 5 and same can be observed in case of Macy's product.
  Therefore an interesting observation can be made. Al-

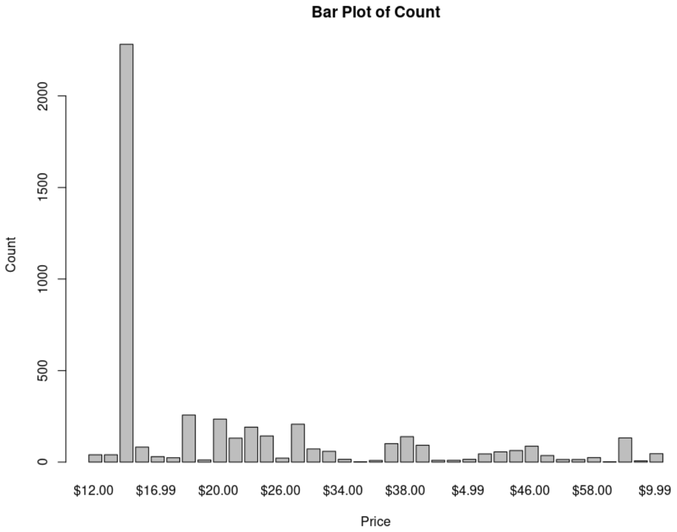


Fig. 4. HANKYPANKY PRICE DISTRIBUTION

though Amazon's product prices are higher than the Hanky Panky's and Macy's the review rating of the their product remains somewhat same.

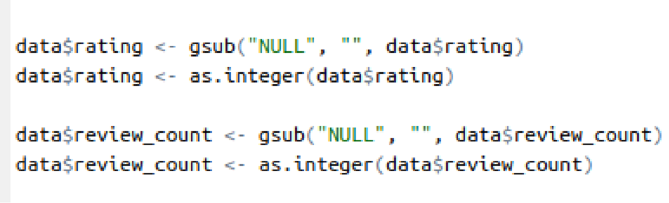
C. Data Preparation

Data preparation also known as data preprocessing is a crucial step in machine learning pipeline. It include below points:-

- Data Cleaning
- Data Transformation
- Handling Imbalance in data
- Dealing with Noisy data
- Feature Engineering
- Data Integration
- Handling Text and Categorical Data
- Data Splitting

In the step of data cleaning we identify and rectify errors, inconsistencies and inaccuracy in the dataset. The quality of data significantly impacts the performance and the reliability of the models. Data Cleaning includes various tasks such as handling missing values, Dealing with duplicates, outlier detection and treatment, consistent formatting, handling inconsistencies, dealing with irregular data.

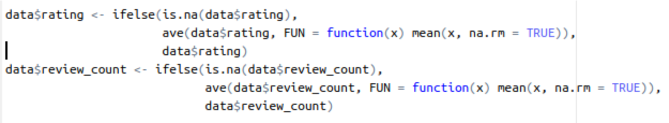
Data scientists and domain specialists frequently work together throughout the iterative process of data cleaning. Verifying that the machine learning models trained on the data are strong, dependable, and able to produce insightful findings or precise predictions is an essential step.



```
data$rating <- gsub("NULL", "", data$rating)
data$rating <- as.integer(data$rating)

data$review_count <- gsub("NULL", "", data$review_count)
data$review_count <- as.integer(data$review_count)
```

Fig. 5. Null Data Cleaning



```
data$rating <- ifelse(is.na(data$rating),
                      ave(data$rating, FUN = function(x) mean(x, na.rm = TRUE)),
                      data$rating)
data$review_count <- ifelse(is.na(data$review_count),
                            ave(data$review_count, FUN = function(x) mean(x, na.rm = TRUE)),
                            data$review_count)
```

Fig. 6. Mutating Null data with Average

In Fig. 5 in this particular code snippet we can see that there are some null values in the rating and the review count throughout the datasets which are needed to be handled. Since the Null is a data string the entire column in considered as a Character type. For implementing the machine learning models we can to be convert it to numeric or integer type data. First we will convert the 'NULL” character into black space ””. Once this is done we can mutate the empty field with the average or median of the entire variable.

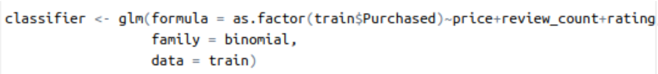
In Fig 6 in this particular code snippet we are exactly putting the average of the variable column to empty field which we have created in fig. 5 code snippet. This allows a homogeneous data variable.

After we have converted all the data type into required numeric data type we can split the data into training and testing data. Generally 70:30 data split is recommended. Which means 70 percent of the data should be labeled as Training data and 30 percent of the data should be labeled as Testing data. We can then train our model using training data and test the model and evaluate it using testing data.

D. Modeling

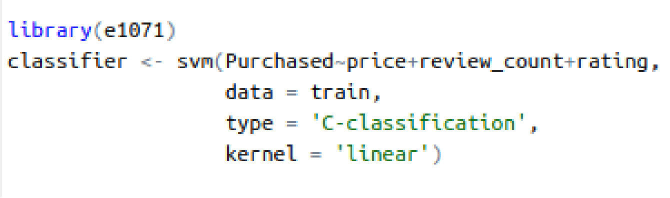
Modelling is the process of using data to create and train a mathematical or computational representation of a system or occurrence in the actual world. Creating a descriptive or predictive tool that can identify patterns in the data, make forecasts, or make educated decisions is the aim of modelling. An essential component in the entire data analytical workflow is modelling. In this particular report we will focus on making the predictive model for the given datasets. There are total of 3 datasets and total 5 models are applied. 3 are the classification type machine learning model and 2 are regression type machine learning model.

Fig.7 shows the implementation of logistic regression which is used to forecast the likelihood that an instance will fall into a specific class in binary classification. Logistic regression is not a regression algorithm, despite its name; it is a classification algorithm. It is extensively utilised in several disciplines, including as machine learning, economics, and medicine.



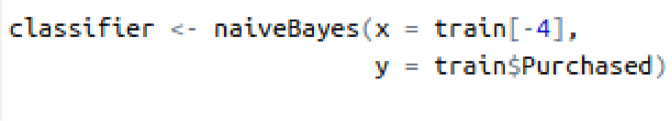
```
classifier <- glm(formula = as.factor(train$Purchased)~price+review_count+rating,
                  family = binomial,
                  data = train)
```

Fig. 7. Logistic Regression formula



```
library(e1071)
classifier <- svm(Purchased~price+review_count+rating,
                  data = train,
                  type = 'C-classification',
                  kernel = 'linear')
```

Fig. 8. Support Vector Machine Formula



```
classifier <- naiveBayes(x = train[-4],
                         y = train$Purchased)
```

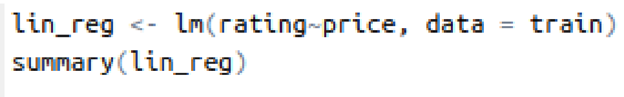
Fig. 9. Navie Bayse Formula

From fig. 7 we can see that we have given input to be predicted as ’Purchased’ which is a column of binary values therefore we have chosen it’s factor data type. Along side that we have taken ’price’, ’review count’ and ’rating’ since it is the most impacting variables to determine whether customer has purchased the product or not. The family of the classification is set as binomial and the data which is taken into account is training data.

Although logistic regression is quite popular for classification type modeling, it has some issues such as it assumes a linear relationship between the input feature and response variable. It assumes that the observations are independent of each other. Logistic regression is sensitive to the outliers.

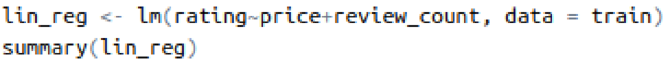
SVM which stands for Support Vector Machine is a supervised machine learning model which can be used for both classification and regression model. The objective of SVM is to obtain a hyperplane that best separates the data into different classes.

Fig. 8 is a code snippet which shows the formula for SVM. R language does not natively support therefore we will be importing ’e1071’ library from ’e1071’ packages. We are again providing Purchase variable to be trained and given price, review count and rating. Further we are providing type as C-classification and kernel as linear. SVM supports linear, Polynomial, Radial Basis and Sigmoid kernels. K(x1, x2) = x1.x2 represents the Linear Kernel function. K(x1, x2) represents the value of kernel function for 2 input vectors x1 and x2. x1.x2 indicates the dot product of input vectors x1 and x2.



```
lin_reg <- lm(rating~price, data = train)
summary(lin_reg)
```

Fig. 10. Linear Regression Formula



```
lin_reg <- lm(rating~price+review_count, data = train)
summary(lin_reg)
```

Fig. 11. Multiple Linear Regression Formula

A dependent variable (goal) and one or more independent variables (features) are modelled using the statistical technique of linear regression. Finding the best-fitting linear connection between the values of the independent and dependent variables to explain variation in the dependent variable is the aim of linear regression. Linear regression assumes a linear relationship between the independent variables (features) and the dependent variable (target). Linear regression tries to minimize the difference between actual value and predicted values of the independent variable. This is done by defining the objective function, often referred as cost function. Fig. 10 code snippet is for linear regression where only one variable is considered to formulate the rating which is price. Fig. 11 code snippet is for multiple linear regression which takes into account price and review count variables to formulate rating.

E. Evaluation

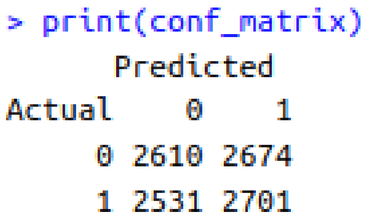
In machine learning models, evaluation entails determining how well a model performs and how well it can generalise to new, unobserved data. Determining the model’s expected performance in practical situations and spotting possible problems like overfitting or underfitting are the main objectives. Depending on the kind of work, a number of evaluation metrics and approaches are frequently applied (classification, regression, clustering, etc.). There are following ways to evaluate the performance of the model.

1) Evaluation Matrix:- For regression models parameters such as Mean Square Error, Mean Absolute Error and R-square are measured while for classification model parameters such as Accuracy, Precision, Recall, F1 Score and confusion matrix which shows true positive, true negative, false positive and false negative are measured.

2) Cross Validation:- To lessen the effect of dataset variability on model performance, apply k-fold cross-validation. Each subset of the dataset is used as the test set precisely once, and the model is trained and tested k times on the divided dataset.

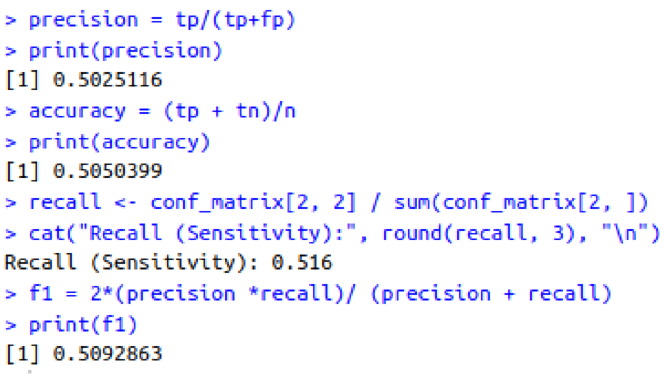
3) Area Under the Curve:- The region beneath the ROC curve. AUC is a single statistic that provides an overview of a binary classification model’s performance at various threshold settings.

Fig 12 shows the confusion matrix of the Support Vector Machine. Here we can see that there are 2610 true positive, 2674 false positive, 2531 false negative and 2701 true negative values which are being predicted. If we look at fig. 13 we can see that model have given around 50 percent



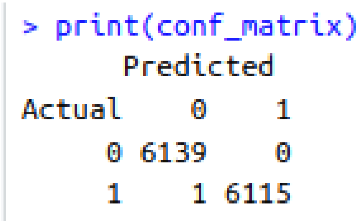
```
> print(conf_matrix)
         Predicted
Actual     0    1
      0 2610 2674
      1 2531 2701
```

Fig. 12. Support Vector Machine Confusion Matrix



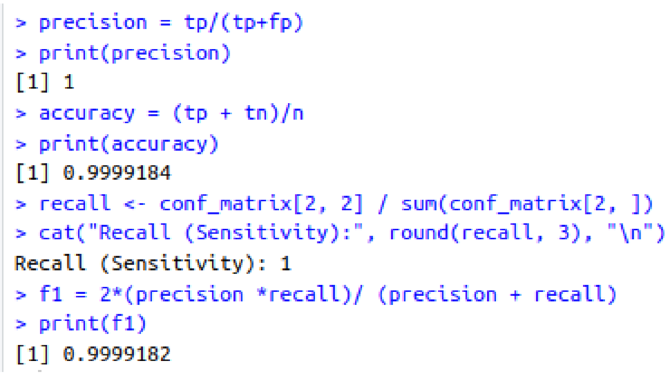
```
> precision = tp/(tp+fp)
> print(precision)
[1] 0.5025116
> accuracy = (tp + tn)/n
> print(accuracy)
[1] 0.5050399
> recall <- conf_matrix[2, 2] / sum(conf_matrix[2, ])
> cat("Recall (Sensitivity):", round(recall, 3), "\n")
Recall (Sensitivity): 0.516
> f1 = 2*(precision *recall)/ (precision + recall)
> print(f1)
[1] 0.5092863
```

Fig. 13. Support Vector Machine Evaluation Scores



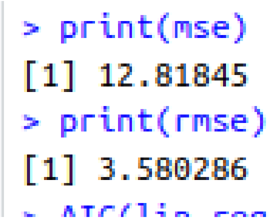
```
> print(conf_matrix)
         Predicted
Actual     0    1
      0 6139    0
      1    1 6115
```

Fig. 14. NavieBayes Confussion Matrix



```
> precision = tp/(tp+fp)
> print(precision)
[1] 1
> accuracy = (tp + tn)/n
> print(accuracy)
[1] 0.9999184
> recall <- conf_matrix[2, 2] / sum(conf_matrix[2, ])
> cat("Recall (Sensitivity):", round(recall, 3), "\n")
Recall (Sensitivity): 1
> f1 = 2*(precision *recall)/ (precision + recall)
> print(f1)
[1] 0.9999182
```

Fig. 15. Multiple Linear Regression Formula



```
> print(mse)
[1] 12.81845
> print(rmse)
[1] 3.580286
> AIC(lin_reg)
```

Fig. 16. Multiple Linear Regression Formula

precision and 50 percent accuracy. Recall is at 0.516 while f1 score is 0.509. By looking at these evaluation matrix can we declare that support vector machine is not a good fit for this particular dataset in classifying the whether the product has been purchased or not.

Fig 14 shows the confusion matrix produced by NavieBayes model. Here we can see that 6139 is ture positive, 6115 is true negative, 1 is false negative and 0 is false positive predicted values. Just by looking at the confusion matrix we can say that the model have performed extremely well in classification. Furhter to this we can see that the model has given 100 percent precision score, 99 percent accuracy. Although the result produced high fit of the model it is not recommended to make it in the deployment phase as the model is overfitting. A typical problem in machine learning is called overfitting, which occurs when a model learns the training set too well, capturing nuances and noise that are unique to the training set but may not translate well to newly discovered data.

REFERENCES

[1] R. Shi and C. Zhang, ”A study of sales forecasting in multinational retail companies: a feature extraction-machine learning-classification based forecasting framework,” 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2023, pp. 401-405, doi: 10.1109/ICSECE58870.2023.10263406.

[2] Y. Yang, ”Prediction and analysis of aero-material consumption based on multivariate linear regression model,” 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2018, pp. 628-632, doi: 10.1109/ICCCBDA.2018.8386591.

[3] R. Vyas and R. As, ”Seasonal Sales Prediction and Visualization for Walmart Retail Chain Using Time Series and Regression Analysis: A Comparative Study,” 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2022, pp. 1-6, doi: 10.1109/ICSTSN53084.2022.9761294.

[4] N. Mohan and V. Jain, ”Performance Analysis of Support Vector Machine in Diabetes Prediction,” 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1-3, doi: 10.1109/ICECA49313.2020.9297411.

[5] K. He and C. He, ”Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data,” 2021 IEEE International Conference on Computing (ICOCO), Kuala Lumpur, Malaysia, 2021, pp. 241-246, doi: 10.1109/ICOCO53166.2021.9673533.

[6] https://www.kaggle.com/datasets/PromptCloudHQ/innerwear-data-from-victorias-secret-and-others?select=hankypanky$_{c}om.csv$