Ahsanullah University of Science and Technology

Department of Computer Science & Engineering

Course No.	CSE 4108
Course Name	Artificial Intelligence Lab
Project	E-commerce customer's purchase prediction using their profile data

Submitted To:

Md. Siam Ansary	Tonmoy Hossain
Department of CSE, AUST	Department of CSE, AUST

Submitted By:

Name	Easinur Rashid Prantick
ID No.	17.02.04.030
Session	Fall - 2020
Section	A (A2)

Introduction

E-Commerce (Electronic Commerce) is the process of buying or selling products or services over the internet.
E-commerce marketing is nowadays the most strategic way to run any business globally. It is the method of making sales by creating and increasing awareness about an online store's product offerings and brand.

On the other hand an e-commerce platform is a software application that enables businesses to set up and manage an online store. The application comes with all the necessary tools required to market and sell the products.

Ecommerce marketers can leverage digital content, social media platforms, search engines, and email campaigns to attract visitors and promote purchases online.

A Brief Description of the Dataset

Name of the Dataset	Customer_Data.csv
File Format of the Dataset	.csv
Dimension of the Dataset	118 x 5
Number of Total Columns	5
Number of Total Rows	118
Number of Feature Columns	5
Name of Feature Columns	User_ID, Gender, Age, P_Code, C_Order
Number of Target Column	1
Name of Target Columns	C_Order

Dataset Description

Studying user data is nowadays the best strategic approach to any E-commerce business. In this dataset. It contains 5 features that can predict the taste of a user, the types of ID's of product he ordered. This dataset contains total 5 colums. They're User_ID, Gender, Age, P_Code, C_Order. In Gender colum the "Male" is lebelled as "1" and the "Female" is lebelled as "0". Besides when a user finally confirmed a purchase, In C_Order colum the confirmation is lebelled with "1" and the cancelation of order is lebelled with "0".

- Empty data: None
- Data type
  - Independent variable, X: integer type
  - Dependent variable, Y: integer type
- Label encode
  - Though dependent variable, Y is assigned as ’1’ and ’0’, where 1=high and 0=low.
- Splitting data set

- Data used for training = 70% (82 rows are used)
- Data used for testing = 30% (36 rows are used)

Description of the Models

1. **K Nearest Neighbor:** The first model is used in my project is KNN. For the training 75 percent data is used and the rest is used in the test set.
   K-Nearest Neighbors (KNN) algorithm is one such supervised learning method that can be used for classification and regression. Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. For example, classification of an animals as cat or dog, emails as spam or not.
   In classification, the prediction values are discrete values like 0 or 1, which relates to true or false. There can be multi-variant (more than one label) classifications as well. Whereas, regression is another type of problem, that requires prediction of continuous values. For example, if we want to predict the approximate value of a share in the stock market, we will have to use regression.
   KNN classifier is to classify unlabeled observations by assigning them to the class of the most similar labeled examples. Characteristics of observations are collected for both training and test dataset. For the purpose of displaying them on a two-dimension plot, only two characteristics are employed. In reality, there can be any number of predictors, and the example can be extended to incorporate any number of characteristics.

   Illustration of how k-nearest neighbors' algorithm works, there are two important concepts. One is the method to calculate the distance between two points. By default, the knn() function employs Euclidean distance which can be calculated with the following equation.
   There are also other methods to calculate distance such as Manhattan distance. Another concept is the parameter k which decides how many neighbors will be chosen for KNN algorithm. The appropriate choice of k has significant impact on the diagnostic performance of KNN algorithm. A large k reduces the impact of variance caused by random error, but runs the risk of ignoring small but important pattern. The key to choose an appropriate k value is to strike a balance between overfitting and underfitting. Some authors suggest to set k equal to the square root of the number of observations in the training dataset.

2. **Support Vector Machine:** The second model is used in my project is SVM classification. For the training 75 percent data is used and the rest is used in the test set.
   SVM algorithm, aka the support vector machine algorithm, is linear. What makes the SVM algorithm stand out compared to other algorithms is that it can deal with classification problems using an SVM classifier and regression problems using an SVM repressor. Being a linear algorithm at its core can be imagined almost like a Linear or Logistic Regression. For example, an SVM classifier creates a line (plane or hyper-plane, depending upon the dimensionality of the data) in an N-dimensional

space to classify data points that belong to two separate classes. It is also noteworthy that the original SVM classifier had this objective and was originally designed to solve binary classification problems, however unlike, say, linear regression that uses the concept of line of best fit, which is the predictive line that gives the minimum Sum of Squared Error (if using OLS Regression), or Logistic Regression that uses Maximum Likelihood Estimation to find the best fitting sigmoid curve, Support Vector Machines uses the concept of Margins to come up with predictions.

Before understanding how the SVM algorithm works to solve classification and regression-based problems, it's important to appreciate the rich history. SVM was developed by Vladimir Vapnik in the 1970s. As the legend goes, it was developed as part of a bet where Vapnik envisaged that coming up with a decision boundary that tries to maximize the margin between the two classes will give great results and overcome the problem of overfitting. Everything changed, particularly in the '90s when the kernel method was introduced that made it possible to solve non-linear problems using SVM. This greatly affected the importance and development of neural networks for a while, as they were extremely complicated. At the same time, SVM was much simpler than them and still could solve non-linear classification problems with ease and better accuracy. In the present time, even with the advancement of Deep Learning and Neural Networks in general, the importance and reliance on SVM have not diminished, and it continues to enjoy praises and frequent use in numerous industries that involve machine learning in their functioning.

3. **Logistic Regression:** The third model is used in my project is LR classification. For the training 75 percent data is used and the rest is used in the test set.
   Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

   A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.

4. **Naive Bayes:** The fourth model is used in my project is Naïve Bayes classification. For the training 75 percent data is used and the rest is used in the test set.
   Naive Bayes is a machine learning model that is used for large volumes of data, even if you are working with data that has millions of data records the recommended approach is Naive Bayes. It gives very good results when it comes to NLP tasks such

as sentimental analysis. It is a fast and uncomplicated classification algorithm.
A classifier is a machine learning model segregating different objects on the basis of certain features of variables.

It is a kind of classifier that works on the Bayes theorem. Prediction of membership probabilities is made for every class such as the probability of data points associated with a particular class.
The class having maximum probability is appraised as the most suitable class. This is also referred to as Maximum A Posteriori (MAP).Naive Bayes classifiers conclude that all the variables or features are not related to each other. The Existence or absence of a variable does not impact the existence or absence of any other variable

Result Table

For the “Customer_Data.csv”, four models are used to justify and measures the result of the dataset. These models are

- K Nearest Neighbor
- Support Vector Machine
- Logistic Regression
- Naïve Bayes

As these four models are used to measure the results. To compare the results four performance metric score is executed. They are Accuracy, Recall, Precision, F1 score.

The result comparison table is given below.

Classifiers	Accuracy	Recall	Precision	F1 score
K Nearest Neighbor	**0.8333333333333334**	**0.5833333333333334**	**1.0**	**0.7368421052631579**
Support Vector Machine	**0.9666666666666667**	**1.0**	**0.9230769230769231**	**0.9600000000000001**
Logistic Regression	**0.9666666666666667**	**1.0**	**0.9230769230769231**	**0.9600000000000001**
Naïve Bayes	**0.8666666666666667**	**0.75**	**0.9**	**0.8181818181818182**

From the result table it is seen that **Support Vector Machine** and **Logistic Regression** Classification have gained the highest accuracy, which is **0.9666666666666667** or **96.7%**

Conclusion

After the model analysis, here in the **Support Vector Machine** and the **Logistic Regression** Classification have gained the highest accuracy, which is **0.9666666666666667** or **96.7%**. They both give the same result. The precision is 92%, it predicts that a user confirmed an order 92% of time as it is correct around 92% of the time. The recall is 100%. It gives measure that 100% of the time the model is able to identify the relevant data.

The other Models are also giving good accuracy. The accuracy of K Nearest Neighbor is 83.3% and the accuracy of Naïve Bayes id 86%.

As the accuracy for the **Support Vector Machine** and the **Logistic Regression** giving the highest accuracy, 96.7%. So, these two classification works well for my project.