PROJECT REPORT

Online Shoppers Intentions Using ML

INTRODUCTION

1.1 Overview :

Online shopping is the activity or action of buying products or services over the Internet. It means going online, landing on a seller's website, selecting something, and arranging for its delivery. The buyer either pays for the good or service online with a credit or debit card or upon delivery. The term does not only include buying things online but also searching for them online. In other words, I may have been engaged in online shopping but did not buy anything. We are going to predict whether the customer will buy the product or just go window shopping. Here, We will be using classification algorithms such as Logistic Regression, Random forest, & Clustering algorithm K-Means. We will train and test the data with these algorithms. From this, the best model is selected and saved in pkl format.

1.2 Purpose:

The main goal of this project is to design a machine learning classification system, that is able to predict an online shopper's intension(buy or not),based on the values of the given features.A number of different classification algorithms is tested, in order to pick the best one for the project.

2 LITERATURE SURVEY:

2.1Existing Problem

People often spend a lot of time browsing through online shopping websites, but the coversion rate into purchases is low. Determine the likelihood of purchase based on the given features in the datsaset. The dataset consists of feature vectors belonging to 12,330 online sessions. The purpose of this project is to identify user behaviour patterns to effectively understand features that influence the sales. Problem StatementThis projects aims to develop an online shopping for customers with the goal so that it is very easy to shop your loved things from a extensive number of online shopping sites available on the web. With the help of this you can carry out an online shopping from your home. Here is no compelling reason to go to the crowed stores or shopping

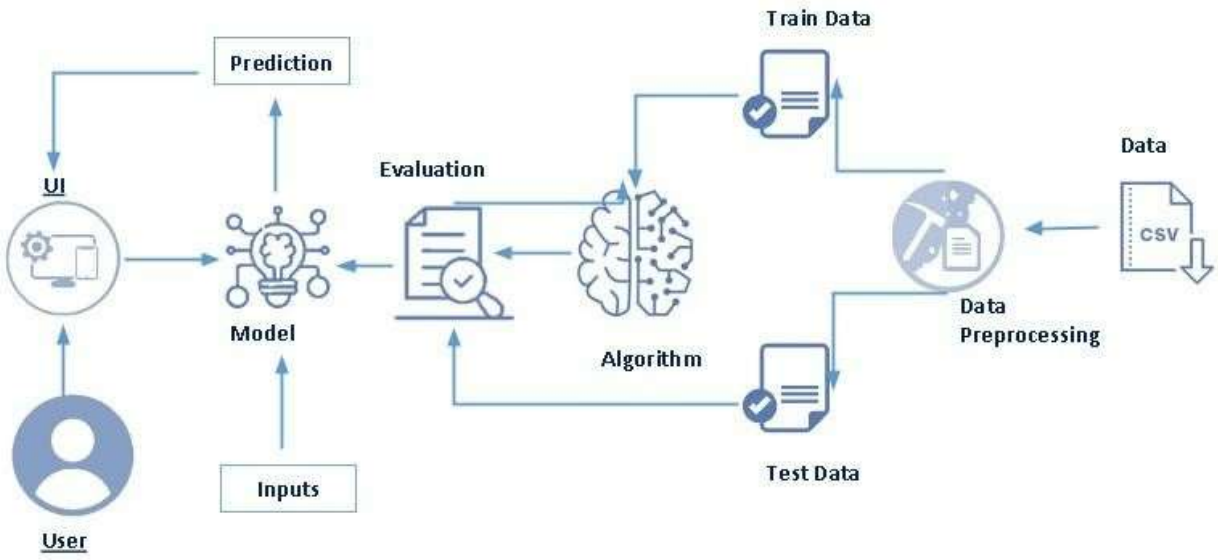
centers during festival seasons. You simpy require a PC or a laptop and one important payment sending option to shop online.To get to this online shopping system all the customers will need to have a email and password to login and proceed your shopping . The login credentials for an online shopping system are under high security and nobody will have the capacity to crack it easily. Upon successful login the customers can purchase a wide range of things such as mobiles, books, apparel, jewellery, infant care, gifts, tools, etc. can be dispatched using online shopping system. Not justthese, you can also purchase from outside nations by few clicks on your mouse. And of course you will get your requested ordered items at your door step. It is simple. You will pick your favourite items from variety of online shopping sites looking at cost and quality. No need to go physical shops with this you will have more time to spend with your family.It Just need a computer and a payment making options like net banking, credit card, debit card or paypal.Almost a wide range of things can be brought through online shopping system.You can purchase goods from foreign places from your bedroom and you will get your goods at your home.It is extremely secure. Customer service is accessible.

2.2 Proposed Solution

In the case of this report, I am trying to predict customer purchase intension which cannot be solved easily with traditional programming .So, I will use the Machine Learning algorithms to build a model or an algorithm that can help predict the outcome of customer purchase intention with as high a degree of accuracy as possible. There are two main Machine Learning problems. These are Supervised and Unsupervised. Here I have researching on what these two different types of Machines Learning model are trying to solve and to identify the most suitable Machine Learning model to the problem I am trying to solve in this report. Here, We will be using classification algorithms such as Logistic Regression, Random forest, & Clustering algorithm K-Means

3 THEORITICAL ANALYSIS:

3.1 Block diagram



3.2 HARDWARE / SOFTWARE DESIGNING:

Software Requirements

- Anaconda navigator (Jupiter)
- Python Packages
- Open Anaconda prompt as administrator

Hardware Requirement

- 2GB ran or above
- keyboard
- CPU
- Dual core processor

4 EXPERIMENTAL INVESTIGATIONS:

```
import pandas as pd
import numpy as np
```

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('fivethirtyeight')m
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import MinMaxScaler, LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.cluster import KMeans
from sklearn.metrics import classification_report,confusion_matrix
from sklearn.decomposition import PCA
from sklearn.model_selection import cross_val_score
import pickle

from sklearn.decomposition import PCA

df.head()

df.info()

df.shape

plt.figure(figsize=(11,11))
plt.subplot(121)
df['Revenue'].value_counts().plot(kind='pie',autopct='%.1f%%')
plt.subplot(122)
df['VisitorType'].value_counts().plot(kind='pie',autopct='%.1f%%')

plt.figure(figsize=(16,4))
plt.subplot(131)
plt.scatter(df['Administrative'],df['Administrative_Duration'],color='b')
plt.subplot(132)
plt.scatter(df['Informational'],df['Informational_Duration'],color='m')
plt.subplot(133)
plt.scatter(df['ProductRelated'],df['ProductRelated_Duration'],color='y')
```

```
"""Most of the visitors spending time on product related page"""

pd.crosstab(df['Revenue'],df['SpecialDay'])
pd.crosstab([df['Month'],df['VisitorType']],df['Revenue'])
df.describe(include='all')
df.isnull().sum()

#handling categorical features
le=LabelEncoder()
df['Month']=le.fit_transform(df['Month'])
df['VisitorType']=le.fit_transform(df['VisitorType'])
df['Weekend']=le.fit_transform(df['Weekend'])
df['Revenue']=le.fit_transform(df['Revenue'])

dfKMeans = df.drop('Revenue',axis=1)

scaler = MinMaxScaler()
scaled_df = scaler.fit_transform(df)

dfKmeans = pd.DataFrame(scaled_df,columns=df.columns)
dfKmeans.head()

#sum of square error (Elbow Method)
n_cluster = range(1,10,1)
sse = []
for i in n_cluster:
    k = KMeans(n_clusters=i)
    ypred = k.fit(scaled_df)
    sse.append(k.inertia_)

sse

plt.figure(figsize=(12,6))
plt.plot(n_cluster,sse,marker='.',markersize=40)

km = KMeans(n_clusters=4)
```

```
ypred = km.fit_predict(dfKmeans)
#dfKmeans['cluster']=ypred

pca = PCA(n_components=2)
dfPCA = pca.fit_transform(dfKmeans)
dfPCA

dfPCA = pd.DataFrame(dfPCA,columns=['PCA 1','PCA 2'])
dfPCA.head()

dfPCA['cluster']=ypred

plt.figure(figsize=(10,10))
sns.scatterplot(dfPCA['PCA 1'], dfPCA['PCA 2'],hue =
dfPCA['cluster'],palette=['blue','pink','orange','green'])
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color='black',s=300,marker='.'
,label='ce ntroid')
plt.legend()

#splliting dataset

x = df.drop('Revenue',axis=1)
y = df['Revenue']

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=10)

#logistic regression model
def logisticReg(x_train, x_test, y_train, y_test):
    lr = LogisticRegression()
    lr.fit(x_train,y_train)
    ypred = lr.predict(x_test)
    print('***LogisticRegression***')
    print('confusion matrix')
    print(confusion_matrix(y_test,ypred))
    print('classification report')
    print(classification_report(y_test,ypred))
```

```
def randomForest(x_train, x_test, y_train, y_test):
    rf = RandomForestClassifier()
    rf.fit(x_train,y_train)
    ypred = rf.predict(x_test)
    print('***RandomForestClassifier***')
    print('confusion matrix')
    print(confusion_matrix(y_test,ypred))
    print('Classification report')
    print(classification_report(y_test,ypred))

def compareModel(x_train, x_test, y_train, y_test):
    logisticReg(x_train, x_test, y_train, y_test)
    print('-'*100)
    randomForest(x_train, x_test, y_train, y_test)
compareModel(x_train, x_test, y_train, y_test)

rf = RandomForestClassifier()
rf.fit(x_train,y_train)
ypred = rf.predict(x_test)

cv = cross_val_score(rf,x,y,cv=5)
np.mean(cv)

import pickle
pickle.dump(rf,open('model.pkl','wb'))
```

Logistic Regression Model

A function named logisticReg is created and train and test data are passed as the parameters. Inside the function, LogisticRegression() algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and

classification report is done.

Random Forest Model

A function named randomForest is created and train and test data are passed as the parameters. Inside the function, RandomForestClassifier algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

5 FLOWCHART

Technical Architecture:

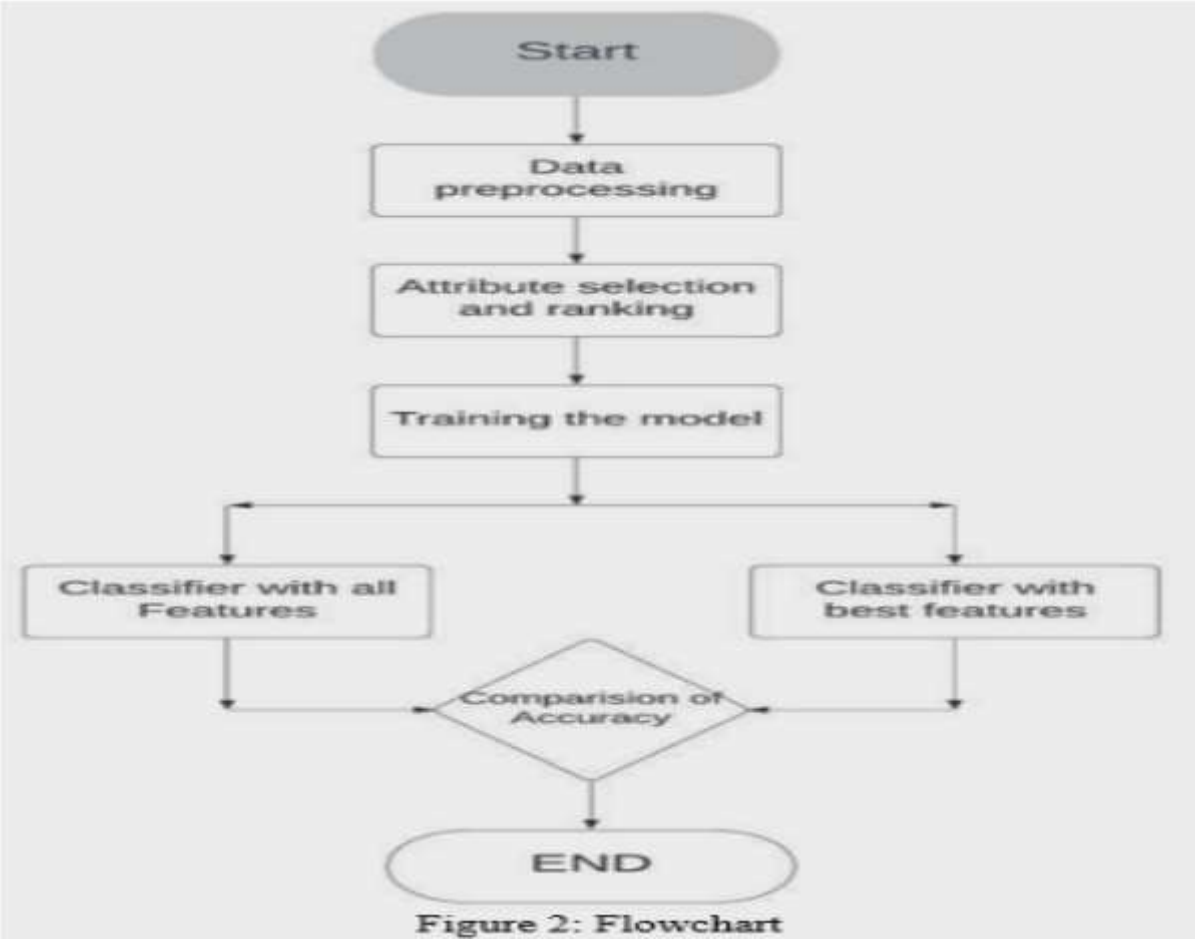
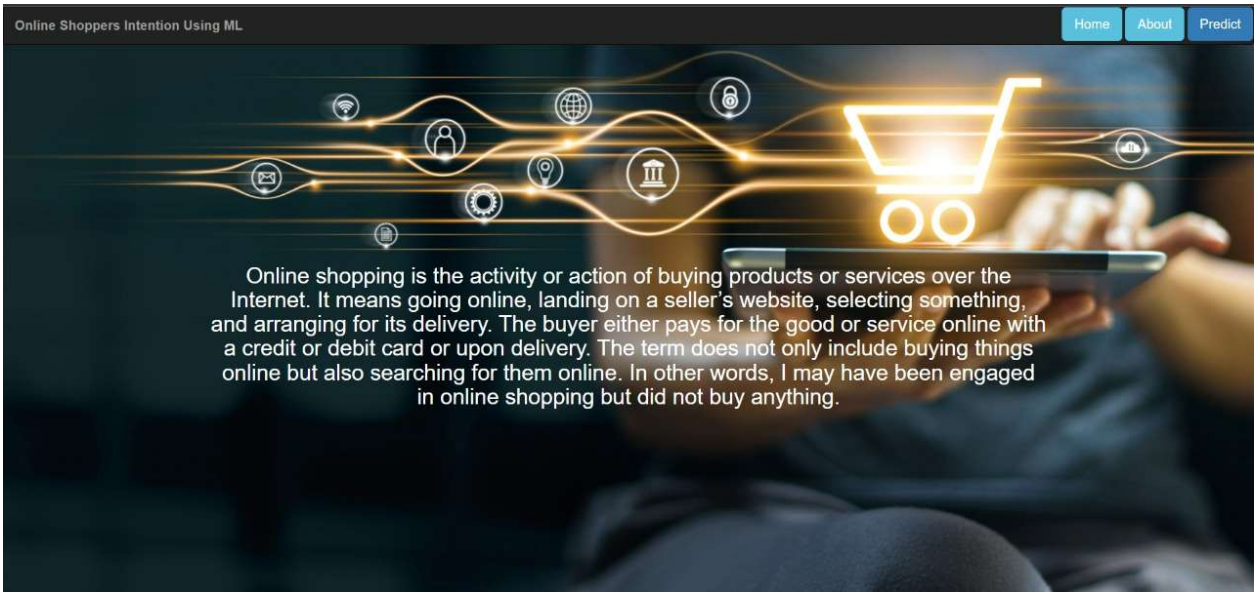
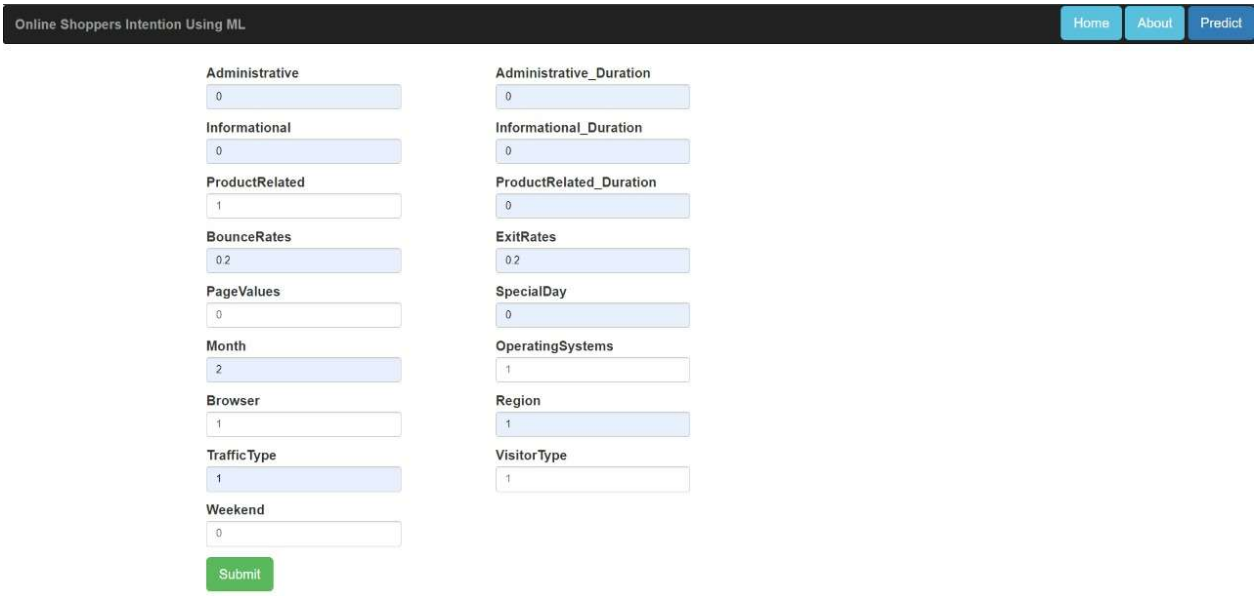
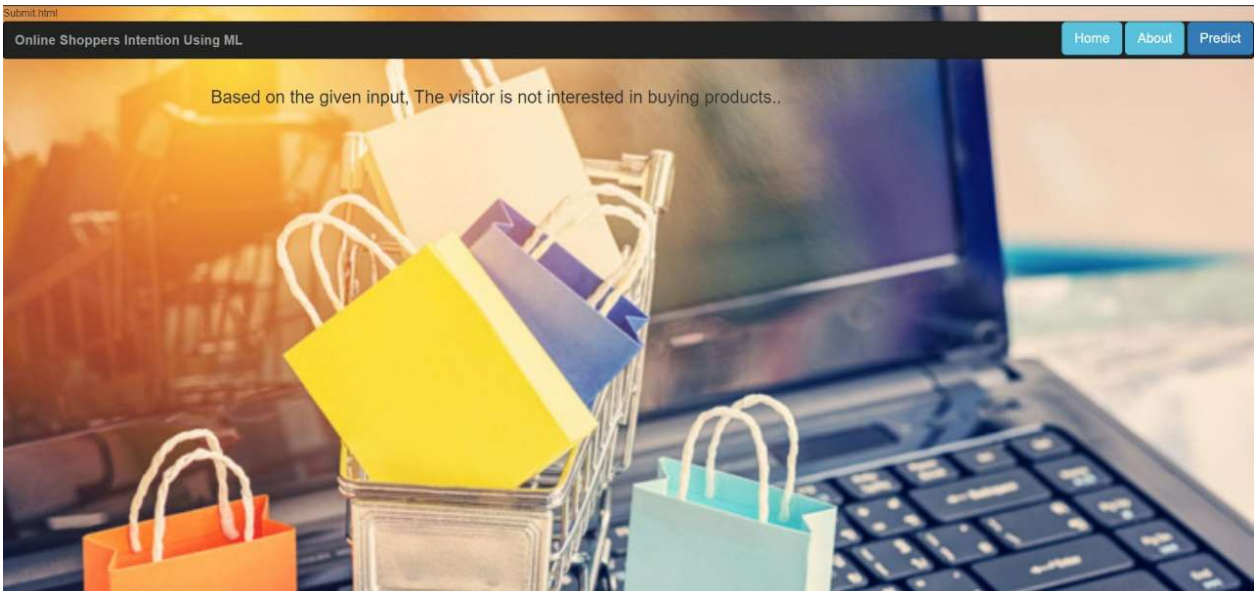


Figure 2: Flowchart

6 RESULT/OUTPUT




Online Shoppers Intention Using ML
Home
About
Predict
Administrative
0
Administrative_Duration
0
Informational
0
Informational_Duration
0
ProductRelated
1
ProductRelated_Duration
0
BounceRates
0.2
ExitRates
0.2
PageValues
0
SpecialDay
0
Month
2
OperatingSystems
1
Browser
1
Region
1
TrafficType
1
VisitorType
1
Weekend
0
Submit


Submit.html
Online Shoppers Intention Using ML
Home
About
Predict
Based on the given input, The visitor is not interested in buying products..

7.ADVANTAGES And DISADVANTAGES

Advantages:

1. Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3.Continuous Improvement

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4.Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5.Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

Disadvantages

1 Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2.Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3.Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4.High error-susceptibility

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

8.APPLICATIONS

customer level models that accurately predict buying patterns of customers, based on historical and current data. Or, you may want to determine the probability of a customer purchasing a product based on the price points. Analyzing such historical and current data and generating a model to Customer Relationship Management (CRM) solutions often require the creation of predict the future outcomes of a product/service is termed as Predictive Modeling.

9.CONCLUSION

The project aim was to build a solution that can predict customer purchase intention with as high an accuracy as possible. The highest accuracy It has managed to achieve was 89.9% accuracy . It was recorded as the highest accuracy by comparing and ranking the various model's performances with each other. It can be seen that Random Forest model performed the best among the various algorithms implemented. After testing the algorithm with various states of preprocessed data it can be concluded that different machine learning models would work better with different types of preprocessed data. I have shown the accuracy of the models based on different states of pre-processed at in the Results and Evaluation section. Comparing the results after the dataset had been cleaned for irregularities. The dataset was then transformed to suit the machine learning algorithms .

By the end of this project:

1. You'll be able to understand the problem to classify if it is a regression or a classification kind of problem.
2. You will be able to know how to pre-process/clean the data using different data preprocessing techniques.
3. You will be able to analyze or get insights into data through visualization.
4. Applying different algorithms according to the dataset and based on visualization.
5. You will be able to know how to build a web application using the Flask framework.

10.FUTURE SCOPE

The customer purchase intention prediction can be combined to the ecommerce website's product recommendation system. Further work can be done to see if recommending products based on customer intension could have an impact on increases sales or not. I.e. Does recommending discounts and special deals to customer who have no intention to purchase choose to buy instead? This can also be used to recommend more expensive or higher quality products to a customer if they already have a strong intention to purchase a similar product. As it is likely they would be open to consider other more expense alternative product . The Artificial Neural Network model I built has potential to be Hyperparameter optimized further. Perhaps, in the future if there are more hardware capabilities then one can perform a RandomSearch , GridSearch or Bayesian Optimization with granular parameter intervals

to obtain best hyperparameter results. There is also the potential of implementing reinforcement learning to solve this problem and researching how Reinforcement Learning solution compares to Artificial Neural Networks and other algorithms used in this report. It would also be useful to collect and perform the research on a larger dataset by which algorithm performance can be measured better because of better training with a larger dataset.