CHAPTER 1

INTRODUCTION

1.1 Introduction

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way or a functional way.

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally this study outputs intelligent software systems. The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving. The intelligence is intangible. It is composed of reasoning, learning, problem solving, perception and linguistic intelligence.

Machine learning (ML) is the scientific study of algorithms that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

A bike is a two-wheeled motor vehicle. Bike design varies greatly to suit a range of different purposes: long distance travel, commuting, cruising, sport including racing, and off-

road riding. Research indicates that the cost of retaining a bike is less than attracting new ones. This is due to marketing costs required to appeal to new bikes. For this reason, together with the increase of competition it has become pivotal that the current bikes base is retained. Normally, bikes churn gradually and not abruptly. This means that by analyzing bikes historic buying patterns one can adopt a proactive approach in predicting churn. Since all transactions are inserted through POS and stored in databases, understanding bikes' needs and patters is possible as data is accessible.

According to, executives are dedicating marketing budgets to focus on bike retention campaigns. Various models designed to predict churn focus on statistical and renowned machine learning algorithms including Random Forest and Logistic Regression. This paper focuses on two aspects when predicting churn within the grocery retail industry. The first is based on the features which will be passed on to the model. Instead of using bikes buying trends to cluster the individuals, these values will be created as features and are passed to the model. Therefore, for each bike various features are created to allow the model to learn and identify patterns per individual. For this reason, two datasets are created to test and evaluate how data should be represented to predict churn. The second aspect is the implementation of the algorithms. The novelty of this method for extracting consumer purchasing behaviour.

1.2. Objective of Research

- To analyse the factors influencing the buying decision making. To analyse the perception of buyers towards different bikes.
- To obtain estimate whether a person buys a bike or not, based on gender, yearly income and age as inputs.
- To obtain accuracy based on previous data.

1.3. Problem Statement

Based on previous data we can predict whether a buyer can buy a bike or not. We considered a dataset which is categorized so we predict that it is a classification using machine learning. This categorized dataset consists of four columns named as gender, yearly income, age and bike buyer. Here we are using decision tree algorithm because when compared to other algorithms this algorithm get more accurate value.

1.4. Industry Profile

India production, bikes and parts is worth approximately 300 crores annually. Although far fewer bikes are sold each year, they account for about 75 percent of industry sales because they command much higher prices than bicycles.

Two-wheelers are one of the most versatile forms of transportation. The adaptive ability of a motorised two-wheeler can be characterised by its usage. Its use could vary from being used just for commute from point A to Point B. Quickly transport small packages of goods through the cramped bazaar streets. A fast ride to catch the school/college bus, or even a brief trip to purchase vegetables. The sheer pleasure of riding a motorbike with the wind blowing on your face, while on a pleasure ride is one of the most compelling reason to own a two-wheeler.

The Indian two-wheeler industry since its beginning, has evolved many folds in technology and, in the numbers being manufactured and produced. It has seen tremendous growth in about half a century, in comparison to other countries where two-wheelers are a major component of transportation.

CHAPTER 2

REVIEW OF LITERATURE

Applying statistical techniques and machine learning algorithms on available data may guide companies in identifying hidden trends and bike behavioural patterns. Implementing data-mining techniques to predict churn may give companies a competitive edge in improving the relationship with bikes. Using bike churn models which correctly classify churn, companies have added value.

Churn is a term used within the marketing field to indicate that a bike has moved to a competitor or has stopped transacting. Churn may be defined as bikes who have a high probability to stop transacting with the company or as described by churn may be identified when a bikes purchasing value falls beneath a threshold across a predefined period of time. Within the Grocery Retail Industry, the identification of the exact moment a bike will churn is hard to define. The output of this layer is sent to a max pooling layer.

CHAPTER 3

DATA COLLECTION

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. Before you head to the bike shop to pick up your new favourite mode of transportation, it's helpful to know all the little details that will make the bike buying process easier.

It is hard to know in advance, what kind of data will be helpful in future. We considered dataset which consists of previous years data about bike buyers. Using previous data we can easily predict present situation. A dataset consists of columns as Gender, Yearly Income, Age,Bike buyer. Here we consider Gender, Yearly Income, Age as independent variables and Bike buyer as dependent variable. The independent variables are given as inputs which are represented by variable x and dependent variable is given as output which is represented by variable y.

CHAPTER 4

METHODOLOGY

4.1 Exploratory Data Analysis

Bike buyer prediction project works under classification model. A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category to which a new data will fall under.

Prediction models for the bike buyers were generated by Logistic Regression (LR), Support vector machine (SVM), Random Forest Classifier and K nearest neighbors (KNN),Decision Tree Algorithm. A comparison of these models was conducted to determine which method produced the best accuracy. Accuracy is only really useful when there are an even distribution of values in a data set.

The good news for us is in our data set they are nearly perfectly even. To assess the likelihood of bike buyer prediction, a predictive equation was developed using data from 700 cases. Above discussed algorithms, we get more accuracy in Decision Tree Algorithm. So,by using Decision tree algorithm we can predict whether a buyer buys bike or not.

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes.

There are two main types of Decision Trees:

1.CLASSIFICATION TREES(Yes/No types)

What we have seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.

2. REGRESSION TREES(Continuous data types)

Here the decision or the outcome variable is Continuous, e.g. a number like123.

4.1.1 Figures and tables

```
In [6]: dataset
```

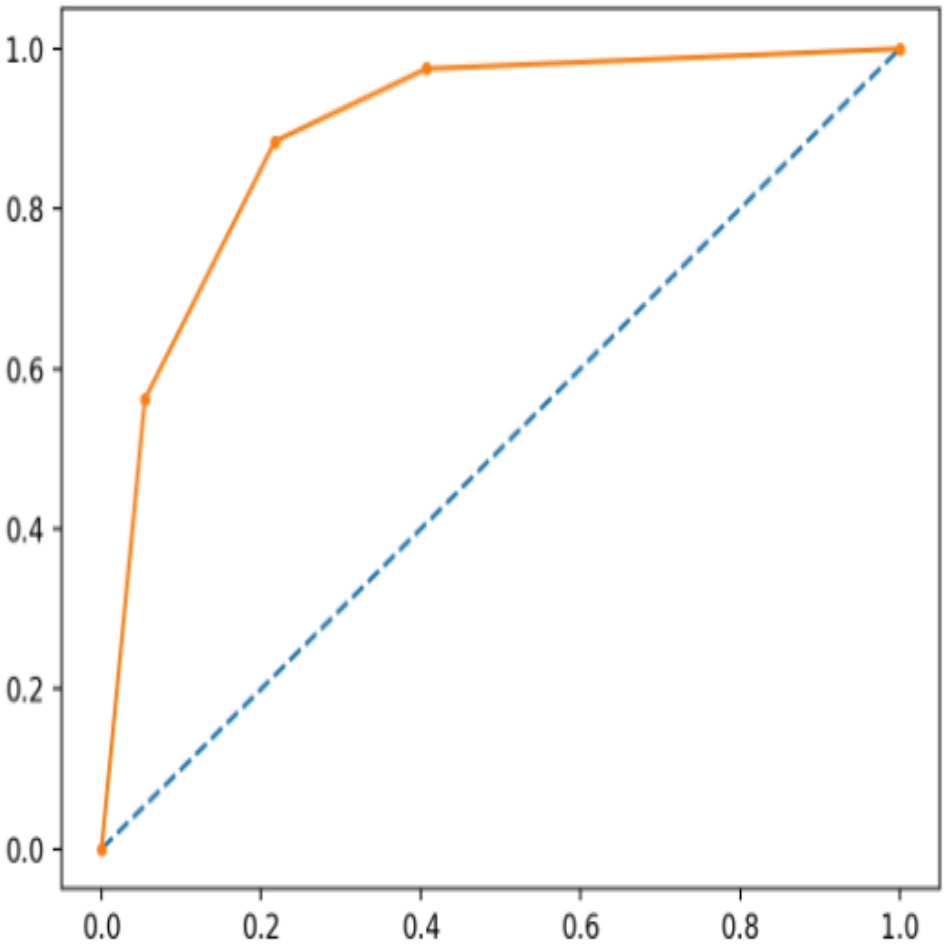
Out[6]:

	Gender	YearlyIncome	Age	BikeBuyer
0	1	90000	50	1
1	1	60000	51	1
2	1	60000	51	1
3	0	70000	49	1
4	0	80000	48	1
5	1	70000	51	1
6	0	70000	51	1
7	1	60000	52	1
8	0	60000	52	1
9	1	70000	52	1
10	0	70000	53	1
11	1	60000	53	1
12	0	100000	49	0
13	1	100000	48	0
14	0	100000	48	0
15	0	30000	38	1
16	1	30000	37	1
17	0	20000	72	1
18	1	30000	72	1
19	1	40000	39	0
20	1	40000	38	1

Table:1.1

```
AUC: 0.895
```

A plot of the ROC curve for the model is also created showing that the model has skill.



Line Plot of ROC Curve

Fig:1.1

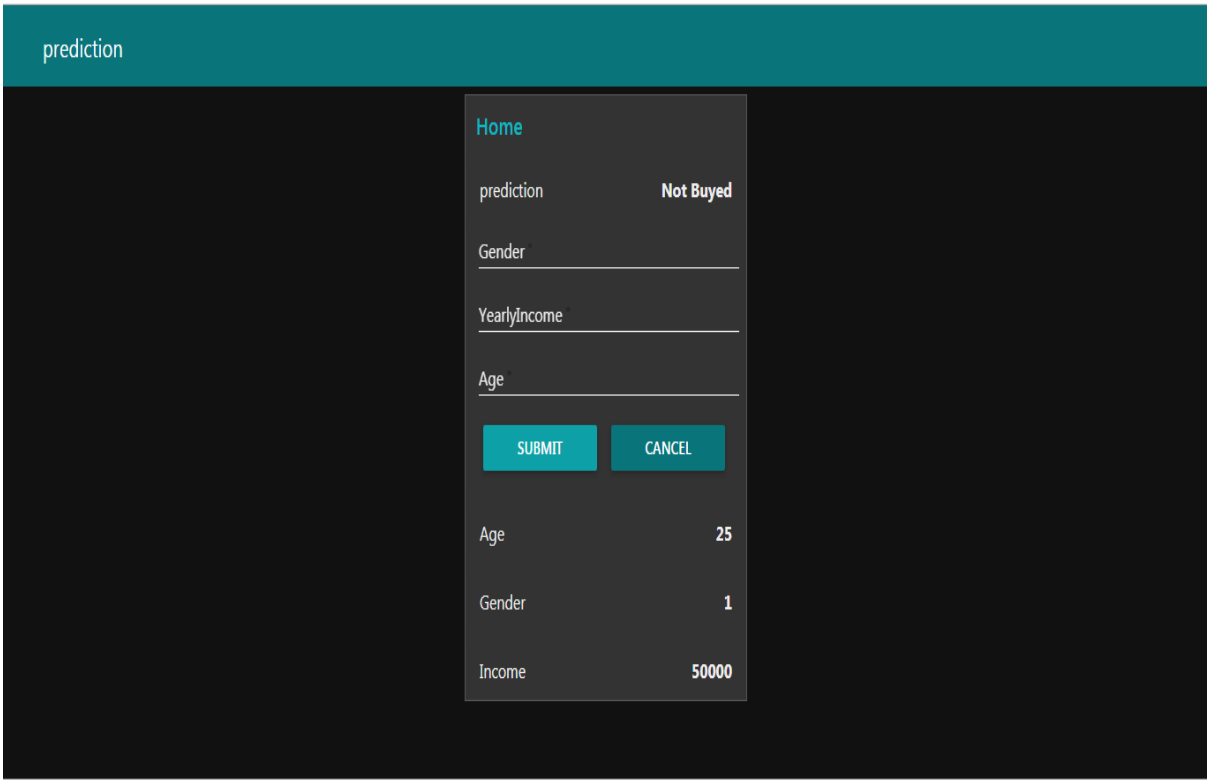

prediction
Home
prediction
Not Buyed
Gender
YearlyIncome
Age
SUBMIT
CANCEL
Age
25
Gender
1
Income
50000

Fig:1.2

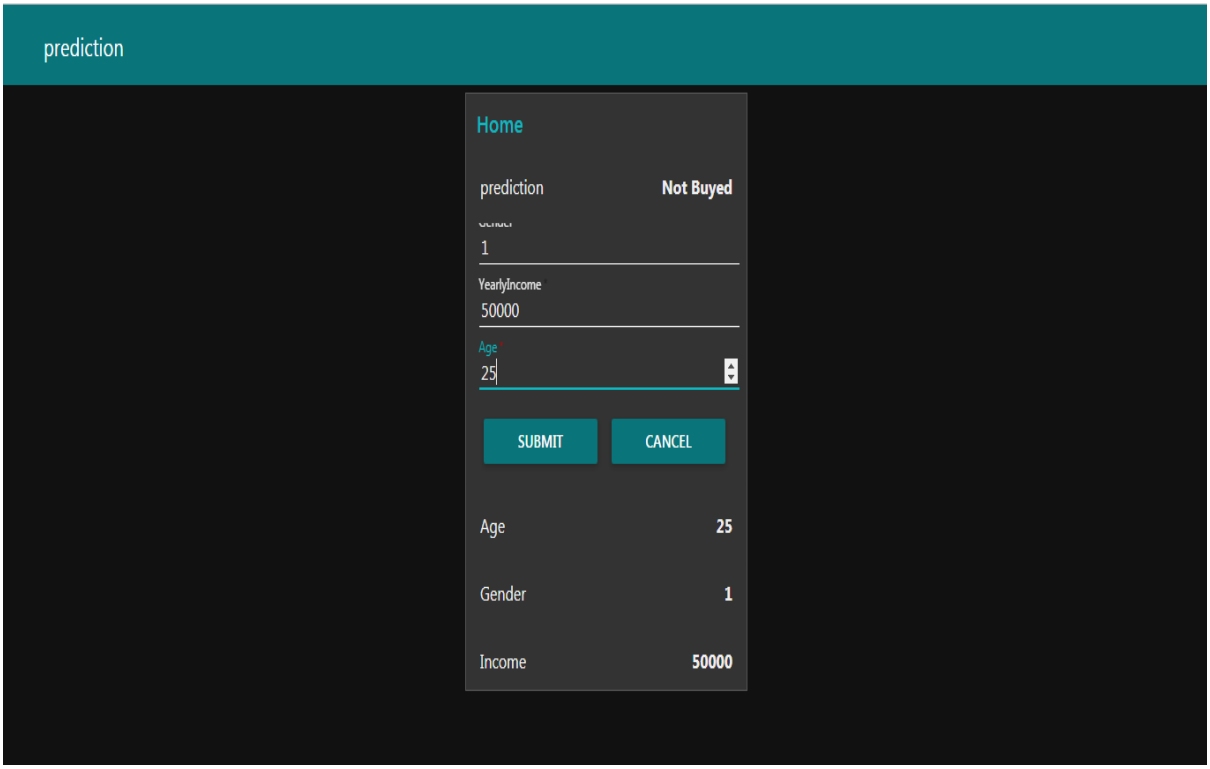

prediction
Home
prediction
Not Buyed
1
YearlyIncome
50000
Age
25
SUBMIT
CANCEL
Age
25
Gender
1
Income
50000

Fig:1.3

4.2 Statistical techniques and visualization

NUMPY

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem. NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using import numpy as np.

NumPy’s main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

PANDAS

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using import pandas as pd. New columns and rows can be easily added to the dataframe. In addition to the basic functionalities, pandas dataframe can be sorted by a particular column. Dataframes can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database.

MATPLOTLIB

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments. On the X array below we saying... include all items in the array from 0 to 2. On the y array below we are saying... just use the column in the array mapped to the 3rd row. The BikeBuyer column. We are using group by to view the distribution of values in our BikeBuyer column. Recall that this column is our target variable. It's that thing we are trying to predict.

Bike buyer prediction is a classification problem, we will import the DecisionTreeClassifier function from the sklearn library. Next, we will set the 'criterion' to 'entropy', which sets the measure for splitting the attribute to information gain. Accuracy is only really useful when there are an even distribution of values in a data set. This module for Node-RED contains a set of nodes which offer machine learning functionalities. Such nodes have a python core that takes advantage of common ML libraries such as SciKit-Learn and Tensorflow. Classification and outlier detection can be performed through the use of this package. These flows create a dataset, train a model and then evaluate it. Models, after training, can be use in real scenarios to make predictions. Flows and test datasets are available in the 'test' folder. Make sure that the paths specified inside nodes' configurations are correct before trying to execute the program.

4.3 Data modelling and visualization:

We consider dataset from https://github.com/xoraus/ML-BikeBuyersPrediction-/blob/master/BBC%20Data%20Set/BBC.csv and modified it into 700 rows and 4 columns. Imported libraries are numpy,pandas,matplotlib. NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools.

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science. A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program) and read it using a method called *read_csv* which can be found in the library.

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data. Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

Now we need to split our dataset into two sets—a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import *test_train_split* from *model_selection* library of scikit. The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data.

CHAPTER 5

REFERENCES

[1] www.kaggle.com

[2]www.github.com

CHAPTER 6

CONCLUSION

The aim of this research is to propose and implement a rule based system to predict the bike buyer from the collection of past data. This has been achieved by applying decision tree algorithm on previous dataset.