PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE

DATA ANALYTICS MINI-PROJECT REPORT
ON

“TRIP HISTORY ANALYSIS ”

SUBMITTED BY

Atharva Shastri	41161
Prathamesh Thombre	41172
Prathamesh Sonawane	41166

Under the guidance of
Prof. Parag Jambhulkar

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2021-22

Contents

1 Problem Statement

Use trip history dataset that is from a bike sharing service in the United States. The data is provided quarter-wise from 2010 (Q4) onwards. Each file has 7 columns. Predict the class of user.

Make use of at least two classification algorithms and provide comparative analysis.

2 Abstract

Data Analytics is the process of analyzing data and drawing conclusion from it. One such dataset is the trip history of capital bikeshares, which logs the travel history of its riders. The dataset is available for each quarter after 2010. We select the first quarter of 2017 for our analysis. The main goal is predict the class of the user as Member or Casual. We inspect various algorithms to achieve this goal and compare their performance.

The bike sharing companies basically gather the data of different types of users and use it to track which members are taking the rental bikes and to which location. By using this data they can decide the to increase thebikes for particular route,give some discounts to regular members. Using this data we are performing analysis to find out member type of user.

3 Hardware and Software Requirements

3.1 Hardware Requirements

1. 500 GB HDD
2. 4GB RAM
3. Monitor
4. Keyboard

3.2 Software Requirements

1. 64 bit Open Source Operating System like Ubuntu 18.04
2. Python 3
3. Jupyter Notebook
4. Different Libraries
5. Libararies like sklearn, pandas, matplotlib

4 INTRODUCTION

The data includes:

1. Duration – Duration of trip
2. Start Date – Includes start date and time
3. End Date – Includes end date and time
4. Start Station – Includes starting station name and number
5. End Station – Includes ending station name and number
6. Bike Number – Includes ID number of bike used for the trip
7. Member Type – Indicates whether user was a ”registered” member (Annual Member, 30-Day Member or Day Key Member) or a ”casual” rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of “test” stations at warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it’s secure).

We perform one hot encoding of the remaing data fields and use various machine learning Classifier models with default parameters as our classificaton algorithm. We use train test split of 75-25 and report accuracy for each model %.

5 OBJECTIVE

- To analyse trip history dataset
- To predict the class of the user from given dataset..

6 Scope

We select only the 2010 file of records for our analysis. This is because the size of whole dataset makes it difficult to run the model on limited memory, currently we are making use of 115,000 records on the datasets to make our models work.

7 System Architecture

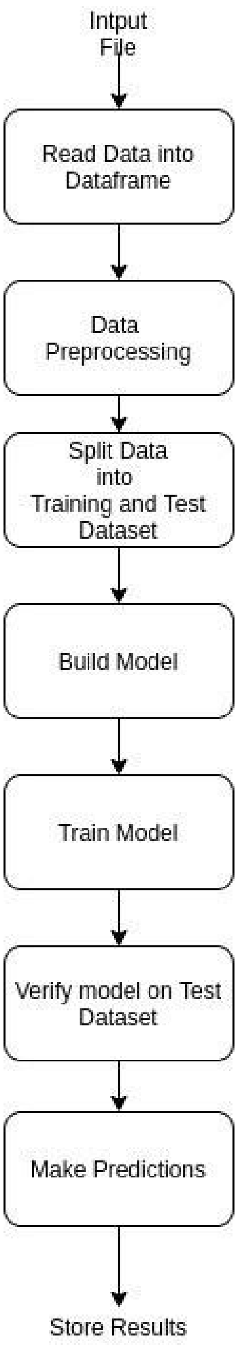

Figure 1: System Architecture

8 Algorithms

1. **Decision Tree :**

   (a) Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classificationproblems.

   (b) Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

   (c) We can represent any boolean function on discrete attributes using the decision tree. At the beginning, we consider the whole training set as the root.

   - Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
   - On the basis of attribute values records are distributed recursively.
   - We use statistical methods for ordering attributes as root or the internal node.

2. **K-Nearest Neighbour :**

   (a) Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data.

   (b) Classification is computed from a simple majority vote of the k nearest neighbours of each point

   (c) This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

   (d) Based on the value of k we get different accuracy so inorder to get good result with this algorithm the K value choosen should be correct.

   (e) KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

3. Random Forest

(a) Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

(b) The Working process can be explained in the below steps and diagram:

- Select random K data points from the training set.
- Build the decision trees associated with the selected data points (Subsets).
- Choose the number N for decision trees that you want to build.
- Repeat Step 1 2.

4. Support Vector Classifier :

(a) Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems

(b) The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

(c) This best decision boundary is called a hyperplane.

(d) SVM chooses the extreme points/vectors that help in creating the hyperplane.

(e) These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

9 Result

The Cross Validation Scores for Various models are:

Model	Cross-Validation Score
DecisionTree	0.7926225820962664
RandomForest	0.8625904010519395
KNeighborsClassifier	0.8468459116232395
SVC	0.8510329077130696

Table 1: Cross Validation Scores for vaious Models

We see that Random Forest Classifier gives the best score. We then use this model to perform training and testing of the model. After training, the model gives an accuracy of 86 %.

10 Conclusion

We used different classification algorithms and the results obtained were
The accuracy score achieved using Decision Tree is: 79.26 %
The accuracy score achieved using K-Nearset Neighbors is: 84.68 %
The accuracy score achieved using Random Forest is: 86.25 %
The accuracy score achieved using Support Vector is: 85.10 %

Based on the above results we can conclude that the accuracy is higher for Random Forest Classification algorithm and can further be increased by increasing the dataset size and a bit of preprocessing of data.