Ahsanullah University of Science and Technology
Department of Computer Science and Engineering

CSE4108: Artificial Intelligence Lab

Spring 2021

Project Report

Train Ticket Price Prediction

Lab Section: A2

Submitted To

Md. Siam Ansary

Department of CSE, AUST Department of CSE, AUST

Submitted By

Ameer Talha ID: 180104036

Sheikh MD Rezone Ullah ID: 180104037

Samin Ul Alam ID: 180104043

1. The description of the problem the group has worked on.

We have collected train schedule data from Bangladesh Railway E-sheba website. There are a total of 600 rows and 7 columns for predicting train ticket prices.

In this project, we can identify the predicted price that one wants to go from one source to another destination. We have solved this problem by implementing different types of Machine Learning algorithms. Suppose, anyone wants to go from Dhaka to Khulna but he/she does not know the price of the ticket and the classes of train. Our project will help him/her to have an idea about the price of the ticket and the class.

2.A brief description of the dataset:

At a Glance Overview

Name of the Dataset	Train Ticket Price Prediction
File Format of the Dataset	.csv
Dimension of the Dataset	600*7
Number of Total Columns	7
Number of Total Rows	600
Number of Feature Column	6
Name of Feature Columns	insert_date, source, destination, departure, arrival, train_class
Number of Target Column(s)	1
Name of Target Columns	fare(Taka)

Description:

The dataset has 7 columns and 600 rows. Of 7 columns, 6 columns are feature columns and they are insert_date, source, destination, departure, arrival, train_class . The 6th column is the target column which we are going to predict the value of and the name of that column is fare(Taka).

Below is a brief description of each columns:

Name of the Feature : insert_date

Description : The current date and time of our ordering tickets. Here, 3/1/22 10:30 AM is the insert_date.

Name of the Feature : source

Description : The location of our departure. Like we want to go from Dhaka to Sylhet. Here Dhaka is the source location.

Name of the Feature : destination

Description : The location of our arrival.Suppose, we want to visit Khulna from Dhaka. Here, Khulna is the destination.

Name of the Feature : departure

Description : The current date and time of our train departure. Here, 3/4/22 1:30 PM is the departure.

Name of the Feature : arrival

Description : The current date and time of our train arrival. Here, 3/4/22 9:30 PM is the arrival.

Name of the Feature : train_class

Description : Different types of train classes. Here, AC_S is the train_class.

3.Description of the used ML models.

1.Linear Regression:

70% data was used for training and 30% data was used for testing in this model.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

2.Gradient Boosting:

70% data was used for training and 30% data was used for testing in this model.

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

Gradient boosting algorithms can be used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

3.Random Forest:

70% data was used for training and 30% data was used for testing in this model.

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

4.Decision tree regression:

70% data was used for training and 30% data was used for testing in this model.

Decision trees build regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target.

5.SVR:

70% data was used for training and 30% data was used for testing in this model.

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already been requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration.

3.Performance Scores of Each Model :

Regression Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R2 Score
Linear Regression	80865.236	284.368	219.932	0.302
Gradient Boosting	9096.773	95.377	57.502	0.922
Support Vector Regression	108444.708	329.309	231.518	0.065
Decision tree regression	9075.511	95.265	30.372	0.922
Random Forest	8006.085	89.477	46.061	0.931

Discussion:

For MSE:
From the above figure it shows that the random forest model gives better scores compared to others.
For RMSE:
From the above figure it shows that the random forest model gives better scores compared to others.
For MAE:
From the above figure it shows that the decision tree regression model gives better scores compared to others.
For R2 Score:
From the above figure it shows that the random forest model gives better scores compared to others.

Now, we can come to the conclusion that, with the given performance metrics for the 5 regression models, we can say the Random Forest regression model is more suitable for this dataset. However, the decision tree regression model and gradient boosting model also scored much similarly beside Random forest. Since, the other two models have the least score in MSE, RMSE, MAE and R2 score. We can neglect them for this dataset.

Contribution :