Taxi Fare Prediction

ITS307 DATA ANALYTICS

BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY

(YEAR III, SEMESTER I)

RESEARCHER (S)

Nima (12200067)

Purna Kumar Limbu (12200075)

Sonam Wangchuk (12200084)

Tshering Dekar (12200091)

GUIDED BY

NIMA DEMA

Gyalpozhing College of Information Technology

Gyalpozhing, Mongar

1. Abstract

Taxi Fare Prediction is a web application that predicts the taxi fare in Bhutan. This application predicts the taxi fare based on the location (pick-up and drop off point), number of seats. distances traveled and based on fuel prices. This is handy and convenient for those who want to travel or reserve a taxi since it gives a taxi fare rate based on the fuel price. This application is available to all who reside in Bhutan.

Table of Contents

2. Introduction

2.1 Problem Statement

One of the most common problems faced while traveling in a taxi is the frequent change in the taxi fare often claimed by individual taxi drivers. To reduce any illegal taxi fare charged by drivers and to reduce the need to fully depend on drivers to know the fare, we propose this project to solve the above mentioned problem.

2.2 Aim

The aim of the project is to develop a web application where the system can predict taxi fare in the country.

2.3 Objectives of the Project: The objectives of this project are:

- To predict taxi fare in order to minimize the problem faced by the riders and drivers
- To analyze previous charging trends and analysis that data
- To implement a machine learning model that can predict the fare of a taxi.

2.4 Scope of the Project:

User Scope: The scope of this project is confined to Bhutan (The target users are people who avail taxi services while traveling within the country.)

System Scope: The system consists of factors such as fuel price, distance traveled, and pick up point, drop off point, number of seats in order to predict the approximate taxi fare.

2.5 Limitations

Our system has a limitation wherein there is more than one route available for the passenger to travel from one location to another, it is difficult to compute all the routes dynamically

2.6 Background Information

The taxicab makes a significant contribution to the accessibility of a city, providing a wide range of services across many different socio-geographic groups. One of the most important challenges affecting cities all over the world is the rising cost of taxis. To better understand the cab fares we are dealing with, we want to identify the main factors that affect them. We consequently made the decision to determine Bhutan's taxi fare projection. Due to the rapid increases of fuel prices, it affects both taxi driver and rider where it become difficult for them to travel as with the increase of fuel prices, the driver raises the fare of the taxi and the rider are not able to pay. This model highlights the prediction of taxi fare that happens in our country and therefore, our model will help both the driver and rider to predict taxi rate.

3. Related Work (Literature Review)

Fare and Duration Prediction: A Study of New York City Taxi Rides

According to this study, predicting fare and duration of a ride help passengers decide when is the optimal time to start their journey or help drivers decide which of the given two potential rides will be more profitable. In order to predict duration and fare, features such as pickup and drop-off coordinates, trip distance, start time, number of passengers, and a rate code detailing whether the standard rate or airport rate was applied (Antoniades et al., 2017).In this study, the data used are all the subsets of New York City Taxi and Limousine Commission's trip data which contains observations of around 1 billion taxi rides in New York City. The total data is split between yellow taxi (operates in Manhattan) and green taxi (operates in the outer areas of the city). The original dataset contains features as pickup and drop-off locations, longitude and latitude coordinates, time and date of pickup and drop-off, ride fare, tip amount, payment type, trip distance, and passenger count.

The study was done to estimate ride duration without real time data, by analyzing data collected from taxis. Through this estimation, it would help in making future predictions. With the objective to model and account for traffic in predictions, two additional features were calculated from the data: rides in an hour and average speed during the hour. Rides in an hour represent the number of started rides within the hour of each observation. The average speed represents the average speed of all those rides. It was found that the prediction result was fairly accurate. It was also found that more variables needed to be considered and modeled to further improve the accuracy. For example, the effect of location between pickup and drop-off points should be considered as well as the differences in driver's speed.

Similarly, for our project, we would like to use two features which are discussed in the above study: pickup and drop off points and distance traveled. However, we would like to use the location names of pickup and drop off point instead of using its longitude and latitude coordinates. Features such as payment type, tip amount, and average speed are not included in our project dataset. Instead of the number of passengers, we would use the number of seats as one of many features to predict the fare.

Taxi Fare Rate Classification Using Deep Network

As stated in this journal, it is difficult for individuals and organizations to estimate taxi trip fare using conditions such as time and day, which affects the traffic condition and starting location in a big city. The paper presents a comprehensive trip fare rate prediction based on time and location.

The taxi fare rate prediction was done using the following features:

- id: unique identifier for each trip

- Taxi id: unique identifier for each taxi
- Timestamp: Julian timestamp which identifies the trip start time
- Starting latitude: latitude coordinate of the pickup location
- Starting longitude: longitude coordinate of pick up location

The target label is Revenue class, which is categorical and reflects the taxi fare rate using five different classes: low, normal, medium, high and very high.

As for the features, more derived features were introduced to improve the performance of the system: time (morning, afternoon, evening), weekends, holiday, and distance from airport, distance from city center, and distance from one of the tourist spots in the given place. It was found that the time of day affects the taxi fare. For instance, the fare is lower in the afternoon than in the evening - the demand for taxis is greater in the evening. Weekends and holidays were also taken into consideration after observing that it also affects the fare rate.

Because the prediction system was based on deep learning with fewer features, it was concluded that the result was not optimal but was acceptable (Upadhyay & Lui, 2017).

As for our project, it will be based on structured machine learning with features such as pick up and drop off point, number of passengers, and fuel price at that particular time, and distance traveled. Unlike in the above paper, our project’s target will be regression in nature. Irrespective of the time, the taxi fare remains same in our country, therefore, time is not considered in our case.

Real Time Prediction of cab fare using machine learning

According to this study, for predicting the longer-term events predictive analysis uses data which is an archive. For capturing the trends which are important mathematical models are used from past data. The model then uses present data to predict the longer-term or to derive actions to require optical outcomes, tones of appreciation in recent time for predictive analytics thanks to development in support technology within areas of massive data in machine learning. Many industries use predictive analytics for making an accurate forecast like giving the amount of fare for the ride within the city. These resource planning are enabled by the forecast as an example, cab fare can be predicted more accurately. A lot of factors are taken into consideration for a taxi start-up company. This research work tries to know the patterns and use different methods for fare prediction. This research work is developed for predicting the cab fare amount within a certain city. The research work involves different steps like training, testing by using different variables like pickup, drop-off location for predicting cab fare.

An Automated Cost Prediction in Uber/Call Taxi Using Machine Learning Algorithm

According to the study, cab businesses such as Uber, Ola, Meru Cabs and others have sprung up in recent years and serve thousands of people across the world every day. It is now critical for them to correctly manage their data in order to come up with fresh business ideas and get the greatest outcomes. As a result, it becomes critical to precisely predict the fares. This paper compared all the fare details of specified cabs and predicted the lowest fare cab using linear regression method. In this paper, they implemented prediction models for the three models like Uber Go, Go Sedan and Uber Auto. Here deviation of the cab fares is also compared and using these data, build an application that can assist the users to select the cab with the determined benefits and lowest fare. In this model they use the machine learning technique of linear Regression model. The methodology and outcomes of this work can contribute to a more real-world demand. This application can improve the transport accessibility, reduce waiting time and reduce the transportation fare etc.

The Research on Planning of Taxi Sharing Route and Sharing Expenses

This journal article focuses on the issues of poor carrying rates, arbitrary route planning, and taxi rates. The focus of this work is on taxi sharing routes and the sharing expenditure model, where the target functions are the maximum carrying rate, the shortest driving distance, and the sharing expense of drivers. They take into account the issues with taxi capacity limitations, driving distance restrictions, passenger load limits, and fee issues. They utilize the passenger's pool to categorize passengers into distinct directions and different beginning locations, then they apply the price algorithm, station-supervised mutation, station fragment cross design, and championship selection technique to solve the model. They examine the taxi data in Lanzhou City through this research and experiment with different ways to present this information. The study's findings also demonstrate how the taxi sharing mode has clearly improved in terms of carrying rate, driving distance, and driving advantages when compared to daily no shared practice. They draw the conclusion that the sharing mode can be used in taxi sharing routes and that the sharing costs are fair and beneficial to both drivers and passengers.

4. Methodology

4.1 System overview

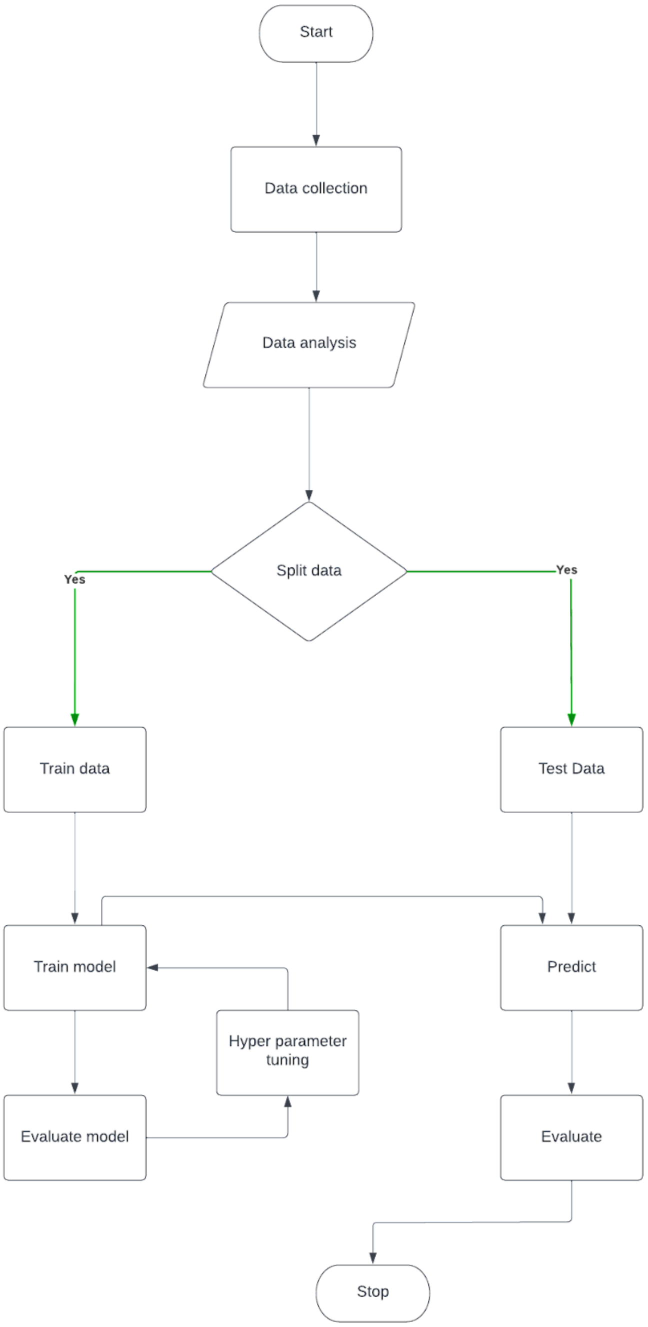

Figure 1: Work Flow

4.2 Algorithm

To train the model, we used the KNN algorithm. Among other algorithms, the KNN algorithm proved to be the best algorithm. Both training data and test data's accuracy were comparatively higher and close to each other. As for the parameter, we chose n_neighbors as 3 and obtained 91% accuracy score in train data and 93% accuracy score in test data. We tried changing the n_neighbors up to 10 and observed no considerable changes in the accuracy of the train data and test data, therefore, n_neighbors of 3 was kept unchanged. The KNN algorithm works on the similarity between new and stored data points (training points) and classifies the new test point into the most similar class among the available classes. The KNN algorithm is non-parametric, and it is called the lazy learning algorithm, meaning that it does not learn from the training dataset, but rather stores the training dataset. When classifying the new dataset (test data), it classifies the new data based on the value of k, where it uses the Euclidean distance to measure the distance between the new point and the stored training points. The new point is classified into a class with the maximum number of neighbors. The Euclidean distance function (Di) was applied to find the nearest neighbor in the feature vector

4.3 Dataset

The project uses the existing datasets acquired from Road Safety Transport and Authority (RSTA) to build the model. The existing datasets of taxi fare provided by RSTA will be used or analyzed for this project. This dataset has a mixture of both numerical and categorical features. Features include pickup point, destination point, distance (km), Fuel price, Fuel consumption and number of seats.

4.4 Evaluation Metric

1. Mean Squared Error (MSE): It is the average distance between the predicted and original values taken in the squares of predicted and original values. It is easier to calculate the gradient by which we can focus on more large errors rather than smaller errors.

$$\frac{1}{N} * \sum_{j=1}^{N} (Y_j - \hat{y}_j)^2$$

2. R-Squared (R2): is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \hat{y})^2}$$

5. **Result and Discussion**

- Initially, there were six features included for model training such as source, destination, distance, fuel price, fuel consumption and number of seats. The last feature was excluded because we found that it does not contribute to the fare prediction.
- Having tried multiple algorithms to produce good predictions, we found out that using KNN algorithm gives better accuracy compared to the rest of the algorithms.
- Our model's accuracy is 70 percent for now. The accuracy can be increased by adding more data.
- The user can go to the website to figure out the taxi fare. They must enter all the details (features) to get the output.

Experimental Setup

Programming language: Python
Platform for training model: Jupyter Notebook

Python libraries for data processing and data visualization
Pandas, Numpy, Sklearn, Matplotlib, Seaborn,

6. Conclusion

One of the most important challenges affecting cities all over the world is the rising cost of taxis. To better understand the cab fares we are dealing with, we want to identify the main factors that affect them. We consequently made the decision to determine Bhutan's taxi fare projection. As a result, our project was effective in identifying the machine learning techniques that are used with particular algorithms in testing the accuracy of selected prediction models and identifying which methodology has the highest precision in predicting taxi rate.

7. References

[1] R. Upadhyay and S. Lui, "Taxi Fare Rate Classification Using Deep Networks," September 2017.

[2] E. R. G, S. M, R. R. R, S. G. M, S. S. R and K. K, "An Automated Cost Prediction in Uber/Call Taxi Using Machine Learning Algorithm," in 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022.

[3] T. P. Jacob, A. Pravin, K. M. Prasad, G. T. Judgi and R. Rajakumar, "Real Time Prediction of Cab Fare Using Machine Learning," in 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022.

[4] X. Zhang, Q. Zhang, Z. Yuan, C. Wang and L. Zhang, "The Research on Planning of Taxi Sharing Route and Sharing Expenses," Mathematical Problems in Engineering, vol. 2020, p. 9, 14 Feb 2020.

[5] C. Antoniades, D. Fadav and A. F. A. Jr., "Fare and Duration Prediction: A Study of New York City Taxi Ride," pp. 1-6, December 16, 2016.