CE802 Machine Learning Assignment

Pilot-Study Proposal

Word Count:712 words Reg no:2101139

Introduction

The aim of this pilot study is to find the feasibility of machine learning algorithms in finding whether a hotel if opened at a particular location will be profitable or not given the geographical and socio-economic data from that location and neighbourhood.

Machine learning is used in more and more domains these days because it can produce useful insights in real world scenarios when provided with appropriate and sufficient data. Another reason for the popularity of machine learning techniques is the wide availability of useful data. More data is collected each day and with more data, machine learning can provide better results.

The profitability of a hotel can depend on a variety of factors. Some of these are the population in the neighbourhood, the presence of relevant tourist places, pilgrimage centres in the location, the number and type of competing hotels in the neighbourhood, etc. So given the data about these factors for other hotels and whether they are profitable or not, will be helpful to predict whether a new hotel for which this data is available will be profitable or not.

Type of predictive task that must be performed

Since the data available specifically mentions whether it is a profitable hotel or not, we have labelled data. So, we will be able to perform supervised machine learning methods for this problem. The problem in hand is to estimate whether a hotel if opened in a new location will generate profit or loss. So basically, this is a classification problem, with the two classes being True (for profitable) and False (not profitable or loss).

Examples of Informative Features

Some examples of possibly informative features that can help in prediction are:

- Distance to nearest tourist destination or pilgrimage centre.
- If the tourist destination or pilgrimage centre is seasonal, then the number of days it is open in a year.
- Distance from the nearest town.
- Distance to the nearest IT Park (for customers attending conferences).
- Distance to nearest hospital.
- Distance from nearest airport.
- Distance from nearest railway station.
- Distance from nearest bus stand.
- Average arrival time of taxis to the hotel.
- Population in the neighbourhood.
- Distance to the nearest hotel/competition.
- Average price of a room per day in competing hotels in the neighbourhood.
- Average wage of hotel crew in the area (per month).
- Average wage of drivers in the area (per month).
- Average usable parking space (number of cars).
- Availability of water (a ratio of the number of litres available to the number of rooms available).
- Presence of an attraction that can be viewed from the room (like a backwater or lake).

Learning Procedures that can be used:

Since this is a classification problem, there are many classification algorithms we can use to predict the outcome. Some of them are Decision tree classifier, Random Forest Classifier, Support Vector Machine Classifier, K Nearest Neighbour Classifier, etc. The advantages of these methods is that they can be trained even with less amount of data but still provide good results. The advantage of Decision tree classifier is that it can give very good accuracy even without normalization or scaling and can even handle some missing values to an extent. Random Forest Classifier works by combining outputs of different decision trees. As the name suggests it is a "forest" of trees. So, it has all the advantages of decision tree classifier and can provide better accuracy in some cases. Support Vector Machine Classifier does not get affected much from overfitting, and is not affected much by outliers, so it is more robust even if the dataset has some bias. K Nearest Neighbour is a very simple but powerful algorithm and it explores features with complex relations.

Evaluating System Performance

The training data is split into 2 independent parts. The first part is for training and the second is for evaluation. While evaluating, we can get metrics like:

- Accuracy: the ratio of correct predictions to the total number of predictions
- Precision: TP/(TP+FP), a measure of quality
- Recall: TP/(TP+FN), a measure of quantity
- F1 score: 2*(precision*recall)/(precision+recall)

Where TP=True Positives

FP=False Positives

FN=False Negatives

By assessing these metrics, we can assess the performance of the model.

So, if the required data is provided, it is possible to predict whether a hotel will be profitable or not in a new location.

References

https://scikit-learn.org/

https://analyticsindiamag.com/7-types-classification-algorithms/

https://iq.opengenus.org/advantages-of-svm/

https://holypython.com/knn/k-nearest-neighbor-pros-cons/