Pilot Study

CE802 Assignment

Registration Number: 1804162

Word Count: 748

Introduction

This pilot-study proposal aims to investigate how machine learning procedures could be used to successfully predict whether a new hotel will be profitable or not; based on historical data of other hotels.

Identification of the Predictive Task

To determine if a hotel will be profitable or not, we will need to predict a specific class label for each hotel to properly categorize each hotel using the data provided. Therefore, the predictive task to be performed is a **classification** task. By identifying that only two possible classes exist ('True' and 'False'), we can further specify that the task to be completed is a **binary classification** task, as each hotel can only belong to one of two possible classes.

Possibly Informative Features

Whether or not a hotel will be profitable requires us to do so without knowledge of any possible sales. If we had this information, then a machine learning procedure would not be required. As we are unable to know how many customers or revenue the hotel is likely to make in the future, the features that we require are those that both provide financial information, as well as information that is likely to have a strong relationship to how many customers the hotel is likely to receive as the scale of the revenue the hotel receives is directly correlated to how many customers they get.

To get the best results for our predictions with this task, the following list contains informative features that should act to allow the system to correctly classify each hotel:

- Open Date:
  Hotels that have been open for longer are likely to be doing well, as they have been able to successfully become established.
- Location:
  If the hotel is in a busy city or tourism destination, then it is likely to see an increase of customers as there is a higher population of individuals that might stay in the hotel.
- Neighbourhood:
  If the hotel is in a neighbourhood with a higher socioeconomic status, then it is likely to be more desirable. This could also correlate to the price of staying at the hotel.
- Expenses:
  We need to know how much money the hotel is expected to be losing regularly.
- Available Room Types:
  Having access to variety of different rooms can attract a wider range of customers.
- Facilities:
  A hotel with a wider range of facilities is likely to attract a wider range of customers. However, these will inevitably increase the hotel's expenses. Different facilities could also have different associated costs, which could also increase revenue.
- Competitors:
  A hotel with less competitors is likely to see more potential revenue.
- Competitor Distance:
  Competitors that are located further from the hotel pose less of a threat in terms of loss of revenue for the hotel.
- Competitor Open Date:
  A competitor that not as established could pose less of a threat in terms of loss of revenue.

Learning Procedures to be Used

As we have identified the prediction task to be a binary classification task, the procedures selected to experiment are ones that are commonly used in such tasks.

- Decision Trees:
  An advantage to using decision trees is that it can establish clear relationships between features, whilst excluding the less important features by making splits using the most important features first.
- Support Vector Machines:
  These models can make use of a much more limited number of data samples to effectively classify data. More complex relationships can be made between features; however, these calculations can result in the algorithm taking longer.
- K-Nearest Neighbours:
  As the data we receive will be in relation to existing hotels, we want to compare the data for each of the new opening hotel with that of the existing hotel. Hotels that are similar to existing hotels are likely to be of the same class, which is where the KNN should perform well.

Evaluating System Performance

To evaluate the performance of the system before it is deployed, the training data will be split into two sets, with each set being comprised of 70% and 30% of the data respectively. As we don't want to evaluate the model using data that it has been trained on, splitting the data allows us to evaluate the model using unseen data by the model; whilst still having the labels belonging to each entry to compare against the values predicted using the test set features. This allows us to calculate an accuracy score for each model, which we will then compare to decide on the procedure to be used.