Report On Predicting Hotel Profitability Using Machine Learning Algorithm

Number of Words: 857

Purpose:

The purpose of this assignment is to predict the profitability of hotels using machine learning. A set of empirical data which includes geographical and socio-economic data about the locations and neighbourhood were provided.

Classification Models:

The predictive task that I have performed is Classification. Classification is one forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. It is a process of grouping objects into pre-set categories. There are few classification models which are commonly used they are

1. Naïve Bayes,
2. Decision Tree,
3. Random forest,
4. K means,
5. K nearest neighbour (KNN)
6. Support Vector Classifier (SVC).

Data Processing:

Data Pre-processing is one of the important steps in predictions. Data processing methods includes manual data processing, mechanical data processing. A data processing provides increased productivity and profits, better decisions and more accuracy. Data processing is done based upon the features giving in the data set. The essential data are to be to restored and the unwanted data can be eliminated if they are of no contribution for the prediction. The most important features which will be required for the prediction of the hotel profitability area distance, cost, surroundings, presence of offices and educational institutes. Based on the the importance of the respective features the data can been considered priority. Here in the data provided by the Hotel Manager they were many missing values. These null values were removed based upon the median of the entire column or the features in the given set.

Models:

After the processing of the data, three predictions were done using three models.

1. The support vector Classifier (SVC),
2. the Decision Tree (DT) and
3. K means

Firstly, The **DT algorithm** was used as it was easy to ready and interpret and also it was easy to prepare. The other reason for selecting DT is requires less data cleaning. Secondly, **Support Vector Classifier (SVC)** is one widely used machine learning algorithm which is able to split the data using more flexible boundaries. SVC was used considering the advantage, It is more efficient in high dimensional spaces and relatively memory efficient. **Naive Bayes** was used as the classifier performs better than other models with less training data if the assumption of independence of features holds. and it can handle both discrete and continuous data. It can be used to solve multi class prediction problem. The data set was split into training and testing set. I considered 70% as training and 30% testing. The output was calculated using the data set provided by using the all the above-mentioned models. This predicted output from the training set was then used along with test set provided. The predicted score was around 60% for SVC and Naïve Bayes, but Decision Tree produced **a predicted score of nearly 80%.** Out of all these three models, the predicted score for the Decision tree was the highest. Since **the predicted value for the Decision Tree model was high**, it was considered to predict profit or loss of the hotel to be opened in the proposed location.

Calculating the Annual Profit:

The Annual Profit can be predicted by the Regression method. Regression Method has been chosen for the very reason that it is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

Data Handling:

The process of data processing was a little trouble free as there were no missing values in the Dataset provided by them. The data set provided by them had few columns which contained non-numeric value, these categorical values were changed to dummy variables.

Models:

Among the various regression models used for prediction, I chose the following models based on the data set provided

- Linear Regression
- Lasso regression model
- Decision Tree Regressor

Firstly, **Linear Regression** is very simple to implement and perform on linearly separable datasets. One of the major advantages of Linear Regression is that overfitting can be reduced by regularization, So I performed the process of prediction using this.

Secondly, I used **Lasso Regression** Model for the very reason that it select features, by shrinking co-efficient towards zero and it avoids over fitting. In both the regression methods, the predicted score reached 60%.

Finally, I performed the **Decision Tree Regressor Model** as it is one of the quickest ways to identify relationships between variables and the most significant variable. Since it is a non-parametric method, it has no assumptions about classifier structure. This showed a predicted score of approximately 50%. I have tabulated my values of Prediction Score and the RMSE in the table below.

Table 1.1

Model	**Predicted Score(approx.)**	**RMSE**
Linear Regression	68%	0.6587
Lasso Regression	67%	0.6490
Decision Tree	20%	1080.446

I chose **Linear Regression Model** as it had the highest predicted score amongst the other. The **Annual Profit** is the **RMSE value** that was obtained after the process of Prediction.