University of Essex

CE802 MACHINE LEARNING

Assignment 1: Design and Application of a Machine Learning System for a Practical Problem

Name: Ahmad Raza
Registration number: 2101194
PRID: RAZAA69707
Masters in Artificial Intelligence
2021-2022

Word count: 744
(Excluding the title page, background paragraph, headings, references)

Pilot Study Proposal

Background:

Gone are the days when different industries used to suffer huge losses because of their bad decisions and investments including both money and time. It's because those decisions were manual and based on human experience, so the chances of error were high because decisions could be biased. Now, times have changed and these companies invest a lot of money in analyzing their new projects beforehand so that huge losses can be avoided. All of this is possible due to machine learning algorithms that can be implemented in literally any industry provided that a reasonable amount of data is available from the past.

Predictive Tasks to be performed:

This project aims to find if a specific company's restaurant will make profit or loss if opened on a specific location. Different machine learning algorithms are going to be used for that purpose. Before choosing the algorithms, we have to decide which predictive task is going to be performed to get desired results. There is a wide range of predictive tasks available in machine learning such as classification, regression, clustering, etc. but we have to choose the task that best suits our available data and target requirements.

For this project, we will be performing classification tasks for following two reasons:

- Given data is labeled, so our tasks are narrowed down to supervised learning and options to perform clustering or any other unsupervised learning tasks have been ruled out.
- The target column has only two possible outcomes. i.e. if the hotel will make profit or loss. Therefore, we know that if the target column contains discrete values, **classification** is the predictive task that will be performed in order to predict those values. **Note:** If the values in the target column were not discrete, for example, instead of predicting if the hotel would make loss or profit, we were supposed to predict the amount of loss or profit, then we will be performing regression tasks.

Features acting as good predictors:

Now that, we have specified the predictive task we will be performing, next step is to identify the key features that will help our algorithm to learn and predict. Manager claims to have access to the data of previous hotels opened and their profits/losses. This data tends to contain their locations, neighborhoods and other geographical and economical features. That's exactly the kind of data we are looking for but assuming that the data has been collected from different sources, we have to define the important key features that will play a vital role in predicting the results, so that data provided to us is complete and robust. For this purpose, we have specified a few necessary features. **Disclaimer:** These are not the only features needed, but the most important ones.

Note: These features are collected from different online sources: [1], [2]

- **Competitors**: Number of competitor hotels present in the area.
- **Competition_Density**: How the competitor restaurants located in the area. For example, extremely dense, moderate density, less dense.
- **Average_weather**: Weather conditions in the area. For example: Extreme cold, cold, moderate, hot, extreme hot. This feature is going to determine the incoming traffic of the area.
- **tourist_attractions**: Number of tourist places in the area.
- **Industrial_density**: Industry and offices density in area.
- **Tourist_to_hotel_ratio**: Ratio of tourists visiting to the number of hotels accommodating them in the area.
- **Location**: Hotel distance from town center to see if it is easily accessible.
- **Hotel_type**: economical, luxury, etc.
- **Population_Employment_Ratio**: Employed individuals ratio w.r.t population in area.
- **Pricing:** Low or high w.r.t average pricing in area.
- **Pricing_Strategy:** Fixed or dynamic based on events, seasons, etc.
- **Occupancy:** Ratio of rooms rented divided by total rooms available at given time.
- **Customer_Satisfaction:** Ratio of satisfied customers.
- **Booking_approach:** Direct, third-party, on-arrival.

Learning Procedures:

As we decided in above part that we are going to perform the supervised learning-> classification so our learning tasks to be performed are shortlisted to only classification tasks that includes:

- Decision Tree Classifier
- K Nearest Neighbors
- Support Vector Machines Classifier
- Random Forest Classifier
- Naïve Bayes Classifier

We will be performing these tasks and compare the performance of the tasks to choose the one that is giving exceptional results.

System Evaluation before Deployment:

We are going to evaluate the performance of our system based on:

1. **Accuracy**

While considering the performance **Accuracy**, we will be considering the:

- Test Accuracy achieved on the model.
- Comparison of test accuracy with the average cross-validation accuracy to see if our model is not under-fitting/over-fitting.

2. **Precision:**

Only accuracy is not enough, another important factor for the evaluation performance is the precision that will determine the preciseness of the model for the predicted TRUEs. [4]

3. **Recall:**

It determines how many actual TRUE,s are predicted/missed by the algorithm. [4]

4. **F1:**

As we need to maintain a balance between Precision and recall therefore, we need F1 score that is a better evaluator than the precision and recall. [4]

All above metrics can be evaluated directly using the **classification report** functionality or they can be calculated manually from the **confusion matrix.**

Note: Accuracy, Precision, recall and F1 are not discussed in detail due to the word limit. Further details about them are present in the lecture notes as well as [4].

References:

[1] https://www.kaggle.com/c/restaurant-revenue-prediction/data

[2] https://www.linkedin.com/pulse/restaurant-location-planning-study-using-machine-learning-yimu-ding

[3] https://www.altexsoft.com/blog/business/hotel-revenue-management-solutions-best-practices-revenue-managers-role/

[4] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9