ASSIGNMENT: DESIGN AND APPLICATION OF A MACHINE LEARNING SYSTEM FOR A PRACTICAL PROBLEM

PILOT-STUDY PROPOSAL

(Word count:900)

submitted as part of the assignment

in

CE802 Machine Learning

by

Ganesh

Ravindran

(2106136)

Department of Mathematical Sciences
University of Essex
January 2021

Overview :

The uncertainty about profitability of a business while opening a new hotel or a new branch for an existing chain of hotels has ever been a problem for business owners This problem can be generalized to any industry regardless of their functionality such as opening a petrol station, opening a new mobile/computer showroom to almost any kind of businesses involving economic activities. This work serves as a guideline for using Machine learning which offers a wide range of tools, techniques, and frameworks, to address these challenges.

Identifying The Type Of Predictive Task:

In this project, the manager of a hotel chain is facing a problem that requires an automated decision: whether a new hotel opened in a given location will be profitable or not? (yes/no). In other words, the manager has a classification problem, where there are 2 categories eg: 1/0, yes/no, True/False, Positive/Negative etc..,

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data. In this proposal, a machine learning algorithm for classification will be suggested to successfully solve this problem.

Informative Features :

Features that are expected from the manager to provide which could possibly help the model to make better prediction.

Feature	Data Type	Brief Description
City	object	city where the hotel is being planned to open
State	object	state where the hotel is being planned to open
Country	object	Country where the hotel is being planned to open
Hotel size	ordinal	size of the hotel small s, medium m, large L
Marketing activities since a year	numerical int	1 if marketing was done 0 otherwise
Competion distance	float	distance of the competitor from the hotel
marketing activities interval	categorical	interval between marketing activities e.g. : Jan, Apr, Jul, Oct
marketing activities since two week	int	1 if marketing was done is last two weeks 0 otherwise
Sales	float	sales number every day
opening hours	numerical float	opening hours every day
customer count	int	number of customers per day
date	date	date when this data was recorded
yearly profit	float	profit every year

learning Procedures:

As stated previously, the hotel chain manager has a binary classification problem, where the output has two categories:

- 1 - New hotel opened in a given location will be profitable.
- 0 - New hotel opened in a given location will not be profitable.

The following machine learning procedures will be considered for the given binary classification task :

1. Cat boost Classifier :

Cat Boost is an algorithm for gradient boosting on decision trees. It gives Great quality without parameter tuning and supports Fast and scalable GPU version. This model was developed and open-sourced by Yandex using this would fetch better accuracy and Fast prediction. There are many situations in the real world where data changes over time. Cat Boost can perform very well in these situations by setting parameter has_time = True. These are the main reasons to consider cat boost classifier.

2. extreme Gradient Boosting classifier:

XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why XGBoost is also called a regularized form of GBM (Gradient Boosting Machine). It works well in small to the medium dataset and it is faster than Gradient Boosting. One of the main aspects of it is that It is designed to handle missing data with its in-build features and it also utilizes the power of parallel processing

3. Support vector machines:

SVM is comparatively less prone to outliers and it generally does not suffer the condition of overfitting and performs well when there is a clear indication of separation between classes which in the given case is a binary classification.

4. Random Forest Classifier:

Random Forest can automatically handle missing values and Normalising of data is not required as it uses a rule-based approach. It reduces overfitting in decision trees and helps to improve accuracy. Random Forest is usually robust to outliers and can handle them automatically.

5 . Decision tree classifier:

The decision tree classifier is less prone to outliers and it requires less effort for data preparation during pre-processing and one of the main features is that missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

6. Light Gradient Boosted Machine classifier:

The Faster training speed and higher efficiency with Lower memory usage and Support of parallel and GPU learning and also the Compatibility with Large Datasets (Futureproof) These characteristics of lightGBM makes it to my candidate algorithms.

Evaluation Metrics:

To evaluate and to avoid overfitting the data will first be divided into training and test of 80% and 20% respectively. The best model will be applied to the test features to evaluate the performance of the model which was trained with an 80% training set.

The best model will be used for evaluation of the model using model accuracy and confusion matrix which will help to get an idea of number of True positive, true negative, False positive and False negatives. Precision and recall can be calculated using the formulas :

- Precision = TruePositives / (TruePositives + FalsePositives)
- Recall = TruePositives / (TruePositives + FalseNegatives)

Also log loss will be used which is one of the most important classification metric based on probabilities.