SHAHRUKH – SHAHR78403

Tosco & Spency Customer Analysis – Pilot Study

Problem Statement & its Significance

The famous supermarket chain ”Tosco & Spency” would like to differentiate its offer and marketing strategy on the bases of two main classes of customers.

1. Some are prone to buy few expensive products (suppose class = 1)
2. Whereas, some usually buy many cheap products only. (suppose class = 0)

It is a very important usecase of machine learning techniques to predict the class of a new customer beforehand to pitch in the relevant marketing strategies only. Indeed, it will prove as a foundation to more complicated use cases and analysis of customers ahead in future using the power of machine learning.

Prerequesities

The manager has access to historical data of past customers and she has offered to provide me with information about several features of the customers (e.g. age, average income, nationality, etc.), and whether they usually buy few expensive products or not (ground truth variable).

Project Plan

This is certainly a binary classification problem where the predictive task is to find out the right class of a customer given a finite set of classes (2 classes, either 1 or 0).

Why there is no need to implement regression or clustering, because former is used for predicting continuous variables like amounts and the latter is used when we do not have past or historical data about ground truth variable of the problem and we try to find similarity between the customers to anticipate their similar behaviour and attributes.

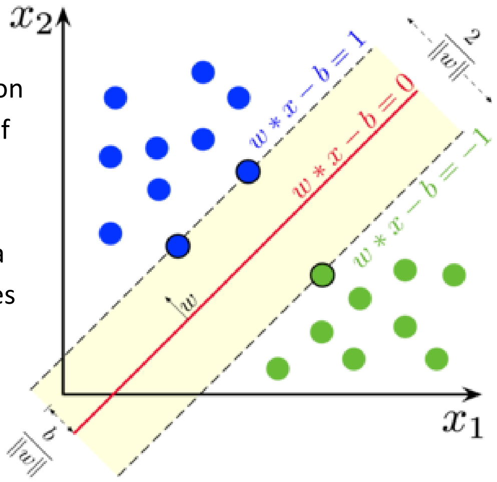
Given the data of customers; age, average income and nationality, I believe there is always a trade-off between too many and too less features that our model does not fall in biasness or variances (underfitting or overfitting). Still to my idea, there can be many good predictors or features that I would ask the manager for;

1. **Season;** it is very important indicator in cases of shopping that a customer has different needs in different seasons. So we can see when a customer normally visits the supermarket.
2. **Number of visits**; This is quite correlated feature to the above one, as in the case of supermarket there must be people coming in almost everytime irrespective of seasonal shopping. This indicator can be a very strong predictor for class 0 (buying many cheap products).
3. **Time of visit**; another important indictor in my point of view could be a time of the customer visit, as normally there must be some peak ranges of time during the day when people would buy expensive products (class 0). For example, not many people of class 0 would be visiting late night to the supermarket.

Approach

There are many model and ML techniques that can be administered in the project. However, I will talk about two models here that by definition of problem seem to perform really well.

Support vector machine was by foundation intended for binary classification problem only. Hence it can play very well on this given problem. To differentiate or separate out 2 classes of customers or datapoints, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some credibility so that future data points can be classified with more confidence. Therefore, I highly prefer this as it produces significant accuracy with less computation power.



K-NN knowing the ground truth of the nearest neighbors or a customer of similar profile, it is highly likely to produce good results. As phsycologically we tend to get inspired by the society a lot and subconsciously we make decision on our shopping as we see people around us of the same profile.

Evaluation of the System

I prefer using the simplest evaluation metric that is Accuracy Score. F1 is usually more useful than accuracy, especially if you have an uneven class distribution. However, according to the data I am given it is a balanced class distribution problem. Hence, we are more concerned about the true positives and true negatives together whereas, F1 is a harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases which is not necessarily needed here.