MGT415 - Data Science in Practice - Problem Set 1

De Becker, Sophie; Di, Yao; Gallese, Mattia; Martiriggiano, Giacomo
College of Management of Technology, EPFL, Switzerland

I. Introduction

This report will present the result of an exploratory data analysis carried out on a customer dataset of Tesco. The raw dataset contained over 7000 unique customers described by over 20 features. The main goal was to investigate the lifetime value of customer that didn't churn.

II. Initial observations

First let's present general observation regarding the distribution of monthly charges. These monthly charges are of interest since they are representative of the value each customer provides the company.

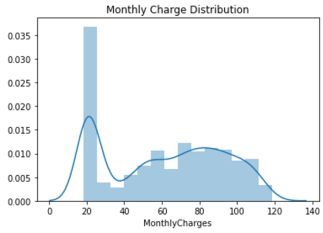

Figure 1. Customer distribution Based on monthly charges

According to Fig 1, we can notice that there is a peak of density at 20 which indicates a significant proportion of client base is paying minimum fees. While the density is low between 30 to 60, there is another peak around 80-90. Therefore, it can be inferred that the average Monthly Charge payed by our customer is not necessarily representative of an "average customer", but rather it is the average of two customer types. Hereafter, an analysis of how the different services impact the monthly charges is presented.

III. Distribution by category

A. Services

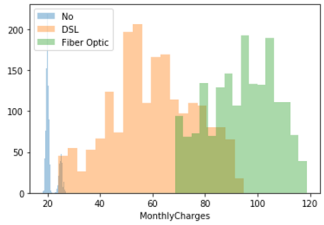

Figure 2. Customer distribution by internet service

The peak for no internet service in Fig. 2 seems to align with the peak seen in the previous graph in figure 1. This suggests that the customers making up the first customer segment around monthly charge of 20 are those without internet services. It is necessary to point out that this may not be a direct causal relation since the price difference could not be driven just by internet services. Indeed, customers not having internet could still have additional offline services and pay a higher monthly charge. However, this graph helps see that the lower charge customers are mostly non-internet users.

The pattern of two peaks is also observed for people who have phone service (Fig. 3). This may be explained

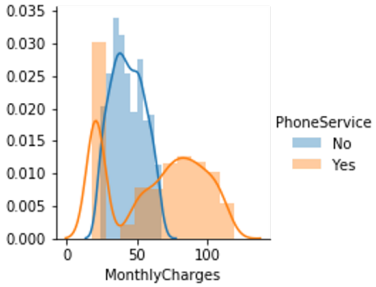

Figure 3. Customer distribution by phone service

by the fact that amongst the group of customers who have phone service the impact of charges associated with other services play a big role in determining monthly charges. However, people that don't have phone services are centered symmetrically around a single peak value. This could be explained by the fact that people without phone service don't have much variation in the other features they subscribe to.

B. Streaming Services

After analysing the basic services provided by Tesco, further analysis is done on streaming service. In Figure 4, it can be observed that among the customers having streaming services, more than 50% subscribes to both Tv and Movies. While, the remaining ones are equally likely to have either the Tv or the Movies service.

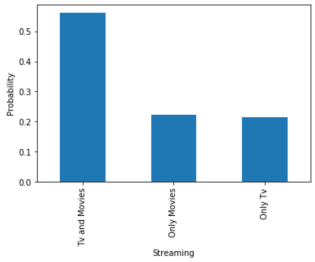

Figure 4. Streaming Services Analysis

C. Tenure

In order to visualize more clearly the relation between tenure and monthly charges data was aggregated by using the average of the monthly charge. The plot shows there is a correlation between tenure and average monthly charge. Although it isn't possible to make any causal inference strictly from this graph, one could make a few hypotheses. For example, it could be that as you stay longer you purchase more features meaning higher charges.

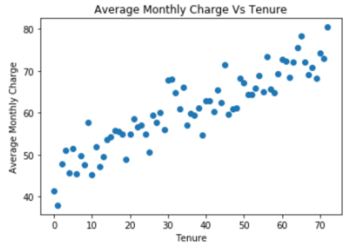

Figure 5. Average Monthly Charges by Tenure