**Target Big Data – Aleksandr Panteleev – 64200406**

**Case description**

The case is dedicated to Target collecting data on purchases of customers and conducting analysis to identify women that are potentially going through pregnancy. The process starts when a new customer shops for the first time and gets a discount card, providing company with his/her address and other personal information. After that, every time a person makes a purchase, information about it is stored in the company`s information system. In addition to that, Target also purchases data from third party providers to gain better insights into consumer preferences. That allows the company to create personalized promotions for each customer and send them via email/by post.

In this particular case, it was stated, that Target`s data science team discovered that pregnant women tend to purchase certain combinations of products, they normally do not buy. Based on what products were bought, statisticians also managed to estimate time until delivery. For example, pregnant women tend to buy larger quantities of lotion, supplements and cotton balls during specific periods of pregnancy. The model allowed Target to identify pregnant women with 87% accuracy. There was a case, when father of a girl discovered his daughter was pregnant after receiving personalized promotion from Target.

**Advantages and disadvantages of data source**

It cannot be denied, that gaining data for business intelligence the way, Target is doing has numerous advantages. First of all, information on past purchases of clients in Target stores is a primary source of information for the company. As a result of that, the company does not have to go through the process of purchasing data from third party providers, overbidding other interested companies. That obviously gives company an exclusive access to information on its customers.

Secondly, by finding correlations of purchases between certain groups of products, the company can get some information prior to other competitors and be ahead of them. For example, by identifying potentially pregnant women prior to them disclosing this information to the public, Target could be the first company to contact a person, thus being ahead of Walmart, Costco and other competitors.

Last but not least, identifying consumers, who are currently experiencing some life-changing events (marriage, divorce or pregnancy) is of high importance to companies. During these periods, individuals become the most prone to changing their habits. As a result of that, if identified on time and addressed properly, consumers could be converted into life-long customers.

On the other hand, there are some serious drawbacks to using this source of data. First of all, Target is not the only company that has loyalty programs. As a result of that, if proved being valuable, the approach could be easily replicated by other competitors. Obviously, competitive edge will be quickly lost, once every company starts exploiting it.

Secondly, there is always a question of ethics. Customers may get really upset about company “spying” on them, especially when it comes to such private aspect of life as pregnancy. In order to prevent such controversies, Target decided to mix pregnancy products with other ones, so that women would think that these products appeared in promotion “by chance”.

Thirdly, Target can only obtain information once an actual purchase is made. Amazon or Walmart.com on the other hand can get information even prior to the purchase, while customer is still considering, which product/brand to choose. However, that is an advantage online shops have over physical ones and not Target-specific one.

**Accuracy of data**

When it comes to the issue of data accuracy, there are two sides of the coin. Due to the fact, that Target is conducting analysis of customer preferences based on past purchases in its stores, the company can get really accurate information, as it is basically an administrative data, no estimates are needed. When compared with information from public surveys, administrative data obviously appears to be a more trust-worthy source. As a result of that, there likely is no issue of “[Garbage in, garbage out”.](https://en.wikipedia.org/wiki/Garbage_in,_garbage_out" \t "_blank)

On the other hand, inferences drawn from the data collected are not necessarily accurate. First of all, for model to be accurate, there has to be actual relationship between variables/processes (analyzing purchases of male customers to predict their pregnancy is probably not the way to go). Secondly, even if some relation exists, one has to build a model correctly, as otherwise there may be too many false positive or false negative results, which would definitely imply additional costs to company or lost opportunities. Having accurate data does not necessarily imply obtaining meaningful and reliable results.

**Big data issue**

When it comes to question, whether the term “big data” is applicable in the case of Target, three Vs requirements should be considered. There is absolutely no doubt, that the volume requirement is satisfied. Target Corporation has 1900+ stores and generates 90+ billion dollars of annual revenue.

When it comes to the variety, the issue is somewhat complicated. Of course, I am not fully aware of the way Target collects and utilizes data. However, based on the description of the case and the idea behind the algorithm that would allow company to identify women that are potentially pregnant, it appears that data is in fact quite homogeneous. I would suspect, that the algorithm could be based on something as trivial (when taking the size of Target into consideration) as logistic regression, where quantities of a products purchased serves as regressors of a model. Due to the fact that information, that is generated by customer making a purchase is quite standard, I believe that condition of variety of sources is not fulfilled in case of Target.

Coming to velocity, it appears that this condition is hardly fulfilled as well. Of course, data is transferred at high speeds. However, technically speaking, delay of one hour (or even day) would not be that big of a deal. There is probably no difference, if the company “discovers” some woman is pregnant one hour later than it could. It would still need to print promotion and send it to customer`s address, contracting USPS or private posting service. Compare it to operations of NYSE or NASDAQ, where fraction of a second can bring company a competitive advantage or put it out of business.

Based on the combination of these factors, it appears to me, that this case does not really represent Big Data in its essence.

**Key findings**

I believe, the first key findings in Target case is that there are certain moments of our live, when we are more prone to changing habits, such events being divorce, marriage or pregnancy. I understand, that moving to another city or moving out of parents’ house can result in change of purchasing habits, as person adapts to new conditions. However, I thought that person is as likely to change purchasing behavior in any other period of life, provided there are reasons for that: new store is constructed in the neighborhood, person discovers that one shop offers better products at the same price point.

Secondly, retailers not only send personalized promotions with products, that you have purchased in the past or products that belong to the same cathegory but are simply of different brand, but also those that could be correlated to the products that you have previously purchased. Target identifying pregnant women is a good example of that.