Case Study: Predicting Pregnancy Likelihood of Customers at Target

Bilal Kudaimi

Overview

Just like all publicly traded companies, public retail store chains search for ways to increase their profit margins. This includes marketing campaigns, introducing new products, and even constructing predictive analytic (PA) models to predict customers' buying behavior (arguably the best technique to increase revenue). With the power to predict what customers are likely to buy what, retailers can make better decisions as to what products should be marketed and which customers should be targeted for a marketing campaign. Speaking of targets, this is exactly what the retail giant Target accomplished.

This case study describes a predictive analytic model developed by Target to predict, with high accuracy, which of their female customers was pregnant. Target was already a large corporation by the time Andrew Pole, one of their statisticians, noticed that pregnant women bought more of certain products. He knew which customers were pregnant because Target kept a baby registry, where pregnant women told Target they were pregnant and when they expected to deliver. Pole used his observations to identify 25 key products pregnant women bought more frequently; he then used that information to consolidate a dataset and train/deploy a PA model that could accurately predict which female customers were pregnant. How was this information useful to Target? Since the birth of a child can change customers' store loyalty, Target could send coupons for

infant items/targeted ads to those customers to convince them to shop at Target. In this way, Target can grow their customer base and increase their profit margins.

Data Understanding

The dataset used to train the PA model contained several rows and columns obtained from the baby registry and transaction data. The unit of analysis here was a female customer, so each row represented one such customer, and each column represented information about that customer. By combining the aforementioned baby registry with customer transaction data, the following information became available:

- Customer ID (assigned to each customer based on their information)
- Customer age
- Store location that the customer frequents
- Whether the customer has purchased each of the 25 key products Pole identified
- Whether the customer was pregnant at the time the information was obtained

The PA model was trained using this customer information and the target variable. Here, the target variable was the binary column indicating whether the customer was pregnant at the time of the registry.

Data Preparation

Most of the columns in the dataset are in a binary YES/NO format, so they must first be converted to a 1/0 format, where 1 indicates YES and 0 indicates NO. This is to allow the machine learning model to parse the data. The only columns that were not in a

binary format were customer age and location, but those columns still had to be prepared as well.

Customer ID was not useful to the analysis; it only helped to consolidate the transaction and registry data into the dataset ultimately used for this model. Customer age is an integer, so any ages present as floats were rounded to the nearest whole then converted to integers. Customer store location was converted to a numerical ID. This works because Target has a finite amount of stores. There was no missing data to impute, as Target maintains customer records of those who were on the baby registry.

Modeling

Pole trained and tested a PA model that could assign a “pregnancy score” to each female customer, i.e., a probability that the customer is pregnant at their transaction time. Because the target variable is a float, a form of regression was used to build the model. It is unknown which type of regression was used, however. The data was first split into training and testing data using an 80-20 split, then the model was trained, tested, and evaluated using two metrics, RMSE and R-squared. RMSE is an error indicator, so the lower the better, while R-squared is a decimal between 0-1 showing how well a model fits some data, so the higher the better. RMSE was minimized as much as possible before the model was deployed on never-before-seen customer data. It was found that the model was very accurate, even correctly guessing a teenager was pregnant before she had told anyone!

Deployment

As one can imagine, the deployment of this model caused a large media uproar and raised privacy concerns among the public. The revelation that a teenage girl was secretly pregnant was only the tip of the iceberg. Charles Duhigg published a New York Times article titled “How Companies Learn Your Secrets” which turned the revelation into a debacle. Target cut off contact with Duhigg after this, and people began to wonder if Target always watched customers’ actions when, in fact, it was a just a guess.

Target spoke out, giving assurances to their customers and changed their marketing strategy. They would mix the infant ads with their other ads in case of a false positive (a woman identified as pregnant when she isn’t). After all, they had to keep this PA model in operation as it was only one part of a two-part campaign: Send targeted ads to those identified as pregnant, then send further ads to keep those customers shopping at Target. Eventually, the dust of this debacle settled, and Target was able to identify the best follow-up ads to send to retain those new customers.

Conclusion

Pole’s regression model not only revealed which Target shoppers were pregnant, but it also revealed the products they bought the most of. It was found that pregnant shoppers were more likely to purchase large quantities of soap, cotton balls, unscented lotion, and mineral supplements (calcium, zinc, etc). Interestingly enough, they were also more likely to buy large purses, perhaps to hold more items for the arriving baby.

With the large profits this model has brought to Target, there is the possibility that Target could sell it to other companies. While this isn’t illegal per se, it may alienate

many customers if others decide to use this model. It would benefit Target to weigh the costs of misidentifying pregnancies first. What would have happened if the teenage girl actually wasn't pregnant? It isn't inherently unethical to predict what customer might do either, but great tact must be employed to ensure that companies can benefit from this without harming themselves or their customers.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Technologies for the Professional Data Analyst.* John Wiley and Sons.

Duhigg, C. (2012). *How Companies Learn Your Secrets.* The New York Times. Retrieved on August 6, 2021 from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Hill, K. (2012). *Could Target Sell Its 'Pregnancy Prediction' Score?* Forbes. Retrieved on August 6, 2021 from https://www.forbes.com/sites/kashmirhill/2012/02/16/could-target-sell-its-pregnancy-prediction-score/?sh=65e749cf35be.

Kuhn, G. (2020). *How Target Used Data Analytics to Predict Pregnancies.* Drive Research. Retrieved on August 6, 2021 from https://www.driveresearch.com/market-research-company-blog/how-target-used-data-analytics-to-predict-pregnancies/.

Piatesky, G. (2014). *Did Target Really Predict a Teen's Pregnancy? The Inside Story.* KDNuggets. Retrieved on August 6, 2021 from https://www.kdnuggets.com/2014/05/target-predict-teen-pregnancy-inside-story.html.