HW2_6238077121

How does a supermarket use data analytics to predict pregnancies? What data and model will you use? (Explain)

For a supermarket to effectively utilize data analytics to make predictions about pregnant women, it should encompass the cross-industry standing process for data mining. Data mining is a field of data science that is a combination of big data and data analytics for discovering patterns that exist in a wide array of data sets.

Business understanding

The ultimate decision of the supermarket is to identify pregnant supermarket customers using data mining. However, to be able to effectively carry out data mining, the supermarket needs various data sources that will be helpful in identifying pregnancy. Next, it needs to clean that data into a usable format. Then, the data needs to be explored initially to develop an understanding of the data. The questions become “What are the data that will be helpful for identifying pregnant supermarket customers?” and “What data models are suitable for such kinds of data?”. Simply put, the supermarket wants to target marketing toward pregnant women effectively. To develop a better understanding of this business requirement, the supermarket should conduct market research to find out what are kinds of products pregnant women usually purchase, at what time of the week or month they make purchases, and how much budget they have. This information will help identify what kind of data to collect and where to collect them. Then, the project plan can be produced where team members are assigned the task of research and data collection.

Data understanding

On-hand data

The following process is the assess the supermarket’s current situation. As a supermarket, it is necessary to keep records of each customer’s purchases, in other words, their shopping history. The shopping history will include the time of the transaction, the products, their count, and the amount paid. This data may or may not be anonymous. Transaction history can be recorded as time series data

Thus, the supermarket can turn to the membership system to identify and match the shopping history of a certain customer. The membership will provide further demographic information including names, ages, gender, and income level. Together, demographic data can help identify people who are more likely to be pregnant. For example, a pregnant woman who is bearing a child is more likely to have a higher income level. These data can be collected as structured data that includes demographics, purchase history, and membership information.

With the customer purchases data on hand combined with the membership data, the supermarket can identify patterns (note: clustering?) in terms of the category of the products that pregnant women buy. For instance, supplements, vitamins, diapers, baby clothing, or baby formulas.

External data

Other types of data that may not be readily available in the supermarket database that is useful include social media data, search engine data, health data, and other third-party data. The supermarket can

combine the on-hand data with these external data to develop a clearer image of their customers and develop insights that can help them identify pregnant customers.

Social media data like Facebook and Twitter posts can provide information about user interests, user activities, and user demographics that are related to babies and parenting.

Search engine data can be used to make targeted marketing efforts for baby-related products for pregnant women. Through the process of web scraping and using the search engine APIs, developers may access search data in a structured format. Such data includes click-through rates and search terms.

Data from search engines and social media maybe be incomplete or in unfavorable structures. Thus, if the supermarket has the human resources available, they may conduct surveys or focus groups with pregnant women directly to identify the types of products they purchase, as well as the previously mentioned information types. This method could be time-consuming and expensive, so the supermarket can also decide to buy the data from third-party aggregators.

Summary

The supermarket can use on-hand data including the customer's purchase history together with their record of memberships, combined with the knowledge of what products pregnant women tend to purchase to identify pregnant individuals from their database. Once identified, the information belonging to those pregnant women can be further used to develop the models. Age, gender, and income are the numerical values that should be recorded. Income level, total purchase of baby products over time, and total purchase per time can be identified.

For data that aren't readily available on hand such as search engine data and social media data, they can be achieved through web scraping or search engine APIs. These data may be incomplete, so the supermarket can conduct further interviews or purchase data from third-party instead.

Data preparation

Once the understanding of what the data will be used for and the data are obtained, the supermarket moves on to the data preparation stage.

Once the necessary data are obtained, the supermarket should clean the data to fill in the missing values, eliminate the noise, and remove duplicates, outliers, or inconsistencies.

Since the data comes from multiple sources, the supermarket needs to combine it into a structured format through the process of data integration.

Then, the validity of data should be checked to make sure that the data obtained are accurate, consistent, and complete. This can be done by cross-checking data from different sources.

Finally, the data should be transformed into a suitable format for modeling.

By the end of this phase, the supermarket should end up with complete, structured data that are normalized and optimized for the use of modeling.

Modeling

Regression

Demographic information such as age and gender can be combined with purchase history, the items purchased, and the search behavior and perform regression. A regression model like linear or logistic regression can be utilized here to identify the relationship between different input and outcome variables and can give insights about the probability of pregnancy detected.

Clustering

In clustering, the data are unlabeled but are divided into groups. Data like the customer purchase history that includes the list of items can be input into a clustering algorithm such as k-means to identify patterns in the products that customers purchase. This algorithm helps identify the hidden structure of the data by identifying the shared features of the data, which in this case are the kind of products that people purchase.

Classification

Classification models identify the customer into categories, which in this case are 1. pregnant and 2. not pregnant. Labeled data such as demographic data, membership data, search engine data, and purchase history that are from the customers who are labeled as pregnant and not pregnant achieved through surveys/interview/purchased data can be used to train a classification algorithm that will predict whether another set of unlabeled data belongs to a pregnant customer or not. Some classification models that can be used include decision trees and random forests.

Similarity matching

Customers identified as pregnant can have their purchase history be similarity-matched to the purchase history of the other customers. In addition, the demographic can also be included to find similar demographic individuals with similar purchase histories of identified products related to babies and pregnant woman care. Using algorithms such as k-nearest neighbors, the other potentially pregnant customers can be predicted.

Evaluation

Once the different kinds of data are fed into their respective suitable machine learning algorithms, the results can be summarized through visualization tools to be presented to the business and help answer the “what customers are pregnant”, the business objective.

The performance of each model shall be compared and chosen for deployment.

Then, the models themselves shall be evaluated whether they are built on the correct principles and correct business/data understandings.

Deployment

Finally, the model is then deployed to either an existing application or a new one. Once deployed, the model should be monitored to ensure that it performs its task well (identifying pregnant women). The model shall be continuously monitored and maintained to ensure that it continues to work well. The model shall be retrained as necessary in response to the change to the business problem and the new data over time.