Hanna Syed
Assignment 3 DAT 412

Introduction and Problem Statement

In the digital age, understanding online customer behavior is crucial for businesses. Inorder to optimize their marketing strategies and improve conversion rates, it is important to recognize trends. The way customers interact with websites provides valuable insights into purchasing decisions. This project develops a logistic regression model to predict whether a customer will make a purchase based on two variables, time spent on the website (in minutes) and the number of pages visited.

Customers who spend more time on a site or go on more pages are generally more interested in making a purchase. However, this relationship can not be assumed to be linear. For example, browsing for a long time without purchasing could reflect the customer's indecisiveness or dissatisfaction with the products offered on these pages. To better capture this model, a logistic regression is used because as it models the probability of binary outcomes. In this case particularly, purchase (1) or no purchase (0).

The goal is to build an accurate prediction model. However things like overfitting, choosing the right learning rate, and regularization strength need to be optimized. By using regularization and testing different parameters can help to make the model perform with great accuracy. Balancing accuracy and simplicity is important to provide reliable purchase predictions that businesses can then use to improve sales.

Conclusion

The logistic regression model developed in this project effectively predicts whether a customer will make a purchase based on the time spent on the website and the number of pages visited. By using the sigmoid function, calculating the cost, applying gradient descent, and adding regularization, the model performed perfectly. With 100% accuracy on the small dataset, it shows that these two variables are strong indicators of purchasing behavior.

While the model's accuracy is seemingly perfect, there may be overfitting since the dataset only contains six data points. Overfitting happens when a model learns specific details and noise from the training data instead of general patterns. To reduce this risk, regularization was used to penalize large weights, helping the model become simpler and more general. Testing different learning rates (alpha) and regularization strengths (lambda) showed that moderate values work best. For example, an alpha of 0.1 and lambda of 0.01 had a low regularized cost.

There are several ways to further improve the model. One way is by increasing the size of the dataset. Using at least 30 data points would make the model statistically significant and more reliable. We can also add more variables to improve the models predictability. By adding more variables to the model, it helps the model better capture the data, making it more accurate predictions. Additional features provide more information about customer behavior, which improves the model. For example, we can add previous purchases, device type or location. We can also test our model on different datasets. In doing so, it can help to conclude how well our model generalizes in new data and tests if it performs consistently across different data points.

Overall, the model performs perfectly with the current data but could benefit from these improvements to ensure its reliability. By using this model, businesses can better predict customer purchasing behavior and improve sales.