Udacity Machine Learning Nanodegree Capstone Proposal

Xian Wee

Domain Background:

The Starbucks Capstone Challenge simulates an empirical business problem. Starbucks seeks to send offers to customers to encourage them to make purchases. The company aims to select customers with the highest probability of making a purchase, without wasting offers on customers who would not likely buy or who would buy even without an offer. Ideally the company would also select the “right” offers to give the customers - the offers most likely to induce the customer to make a purchase. This project is appealing to me because it solves a practical, real-world problem faced by well-known company.

Problem Statement:

Should Starbucks send an offer to a customer and which offer should the company select?

Solution Statement:

Given demographic data of the customers and properties of the offer, a trained machine learning model will predict the likelihood of a customer making a purchase. To decide which offer to give the customer, one approach could be to hold all the features constant for a customer, vary the type of offer and make separate predictions for each offer to determine which offer to send the customer.

Benchmark model:

The primary model I'll be using will be a deep feedforward neural network with 3 or more hidden layers or perhaps a convolutional neural network which can counter overfitting through regularization. A benchmark model I could use will likely be a model outside the category of neural networks, relatively simple, and one that has been widely used in academia and industry in the past. k-nearest neighbors seem to be a model often used as a benchmark. Another good candidate may be a linear model such as a support vector machine with a linear kernel.

Sources:
https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/
https://datascience.stackexchange.com/questions/8785/what-is-a-benchmark-model
https://www.quora.com/What-is-the-linear-model-in-machine-learning

Evaluation metrics:

To evaluate the performance of the model I may look at accuracy, precision and recall. The company may not be too concerned about accuracy, after all it's playing a numbers game and perhaps expects to experience a fairly large number of incorrect predictions.

The initial intuition is that Starbucks would want to minimize lost sales - it's worse for the company to lose out on a potential sale by not providing an offer than to hand out an offer that is not used. Perhaps the revenues from a few sales more than make up for the costs of many offers given to customers. With this primary motivation, Starbucks may place more weight on recall, to reduce the number of false negatives, meaning if the model predicts the customer will not make a purchase after being shown an offer, the customer actually does not make a purchase. Another way to frame it is that the

company would like to correctly identify as many of the people who would buy as possible and is willing to overshoot on identifying customers as potential buyers. So Starbucks may be willing to tolerate the model scoring lower on precision, meaning it would allow for the model to produce more false positives.

Project design:

- Process the data into a useable format
    - Create a dataset with features and labels
        - The features will be the demographic data and the properties of the offers
        - The labels will be the outcome of the customer making a purchase or not
    - Split the data into training, validation and test data
    - The relevant data are the customers who both received and viewed the data, so filter out the customers who did not receive or did not view the offers
- Select the machine learning model to train and make predictions
- Build the model, train using training data, validate and test the model
- Deploy the model to a web API endpoint so that a user can enter demographic data, receive an offer and indicate if he/she likes the offer
- To determine which offer to provide the customer, I may try to hold all test customers' features constant, vary the type of offer and make separate predictions to see which offer has the highest probability of encouraging the user to make a purchase.