# Machine Learning Engineer Nanodegree

**Capstone Proposal**
Raj Nikhil Choul
September 30th, 2018

# Proposal

*Domain Background*

E-commerce is one of those sectors which can benefit immensely from machine learning and various advanced neural network algorithms. This is because not only the data required to build these model easily accessible but it is also relatively easy to deploy them in production to help improve the customer experience. It is no surprise then that some of the biggest e-commerce companies around the world were the pioneers when it comes to building and deploying machine learning models to improve various aspects of shopping on their website.

One of the most impactful application of machine learning on an e-commerce website would be to help find customers what they may need or want without having them to explicitly search for it. This features becomes even more useful for customers of a website, which have a large assortment of products and for website whose customers frequently come back to buy similar products. This not only improves customers' shopping experience but it also builds customer loyalty. With that said I'm interested in working on this project because I work in e-commerce industry too and this experience would be very helpful for my career.

*Problem Statement*

An year ago, Instacart put out its data related to prior customer purchase in the public domain. It was also a competition on Kaggle. As part of this project I will need to build a model that predicts which previously purchased products will be in a customer's current order. Since the problem requires to predict whether a product will be bought in the current order it can be considered as a binary classification problem where all the products previously bought by a user have to classified in one of two bins i.e. reorder or no-reorder.

*Datasets and Inputs*

The data for this project was provided by Instacart[1] an year ago. The data was initially part of a Kaggle competition but Instacart has since made the data publicly accessible. The data is divided into 6 datasets. Below are the details of each of the datasets.

1. Aisles: This dataset contains data about category of products placed in this aisle such as Kitchen Supplies, Packaged meats, etc. This dataset has 134 distinct aisle names.
2. Departments: contains data about categories of products such as Beverages, Alcohol, Pets, etc. There are 21 such departments in the data.
3. Products: Products dataset contains names of the products sold on Instacart and also has details of which aisle the product is in and which department the product belongs.
4. Order Products (Prior & Train): These dataset are divided into two, one for prior data and train data. Prior dataset has details of orders prior to the most recent/current order. Train dataset has the same details as that in prior but is provided to train the model on the current order.

   These datasets contain details of order id, IDs of products included in that order, order number in which the product was added to the cart and whether the product was reordered or not represented by 1 or 0 respectively.
5. Orders: This dataset contains details such as order id, user id, eval set i.e. Prior, Train and Test, order number for the user, day of the week the order was placed, hour of the day the order was placed, and number of days it has been since most recent prior order (capped at 30).

All these datasets together provide a history about a customer's prior purchases and the trends in those purchases such as which products a user buys, at what interval are they bought, are there any combinations that are frequently bought together and so on. These characteristics will allow to predict which products a customer might reorder in the current order.

*Solution Statement*

In order to build a model to predict which products will be bought by a user in the current order, I will first have to engineer a few features such as duration between two orders for a product, how long has it been since a product was last ordered, etc. Once I have collated all the features into a dataset I will experiment with few classification algorithms using grid search to determine which algorithm returns the best results. In the Kaggle competition, F1 score was the metric used to evaluate an algorithm. So, I will measure the model I develop with the same metric. The F1 score for the first placed model was 0.4091449. So my aim would be to get close that score.

*Benchmark Model*

One of the models available from the competition to benchmark my algorithm is the model that was placed second[2]. The F1 score of this model was 0.4082039. Since the model uses the same metric for evaluation it should be easy to compare the model I develop.

In this model, the author created several engineered features such as grouping customers, products, what kind of products do these customers prefer, how often a product is bought,

how long has it been seen being bought previously, etc. The benchmark model uses XGBoost to develop the model for this problem. Since the model was measured with F1 score the reorder probabilities needed to be converted to binary i.e. 1 for reorder and 0 for no-reorder. In order to convert these probabilities to binary a threshold must be set on the probability beyond which a product can be deemed reorder or 1 in a given order.

*Evaluation Metrics*

For this model, I will use F1 score to evaluate the model. This is the same metric used in the Kaggle competition as well. The model has to suggest products to customers they may find useful and since the products are mostly necessity based the model has to be accurate with not only what product it suggests but also be accurate with products it should not suggest. This means the model should maintain a balance between precision and recall. Precision and Recall tend to trend in opposite direction therefore to maintain a balance between the two without having to measure each individually we use F1 score. Since F1 is the harmonic mean of precision and recall it a good metric to evaluate this model.

$$precision = True\ Positive/(True\ Positive + False\ Positive)$$
$$recall = True\ Positive/(True\ Positive + False\ Negative)$$
$$f1\ score = 2.(precision.recall)/(precision + recall)$$

*Project Design*

I will work on preparing one model for predicting reorder. But before embarking on model building the first step I will perform is data exploration steps such as

1. Rank products by frequency of their purchase and quantity
2. Grouping similar products which may be replaceable with each other such as Cola and Fridge Pack Cola
3. Build the correlation matrix for the feature in the data

The next step will be to build engineered features such as duration between two orders for a product, how long has it been since a product was last ordered, etc.

In the following step, I can begin building the model. Since this is a classification problem as mentioned previously, I will be testing out algorithms such as Logistic Regression, various decision trees such as Random Forest, ensemble methods and SVM. I will use the grid search method to determine the best algorithm with optimum hyperparameters. As mentioned earlier, these algorithm will be evaluated with F1 score.

*Citations*

[1]:*"The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 09/30/2018*

[2]:http://blog.kaggle.com/2017/09/21/instacart-market-basket-analysis-winners-interview-2nd-place-kazuki-onodera/