Predictive Analytics for E-Commerce Customer Behavior and Demand Forecasting

Team Members: Shravan Surineni, Shuyu Mao {shravans, surimao}@stanford.edu

Motivation

- Predictive Analytics to predict user intentions towards a certain product, or category on an e-commerce site, based on historical interactions with a website is very useful for advertising, recommendation engines and for demand forecasting.
- Clickstream data can be used to quantify search and purchase behaviors
- The goal of the project is to identify activity patterns of the users that lead to purchase decisions and develop a model to mimic user behavior
- We will use logistic regression and deep learning/neural networks to predict user activity on the website based on clickstream data

Data Preprocessing

Dataset: InfiniteAnalytics a startup by MIT- alum provided user browsing data including clickstream data and the order information correspondingly for recent 10 days in November for a ecommerce site that sells nutrition products.

We split 80% of original data to be training set, and rest 20% into test set. i.e., first 7 days of data for training, and the remaining 3 days for test set.

Data Preprocessing:

- Tune categorical value and numerical value to Indicator value
- Assign weightage for some feature inputs based on intuition
- Link user browsing data with order data by find the matching sample based on same user_id and product_id combination.

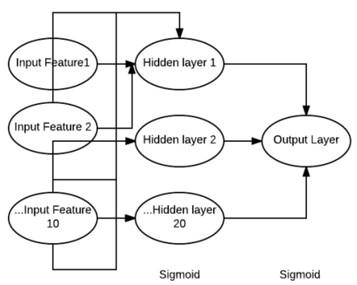
Features

- Client Type
- User need recommendation or not
- Device(Mobile/PC)
- Did user search for any product
- Cookie info(Device, user, IP address)
- How user go to current product page(thru search, click etc.)
- Product ID
- recommendation tag followed by customer
- Registered User or not
- Time stamp

Models

Neural Network

- The input layer reads 10 features of samples linked user browsing & order data.
- The hidden layer we used sigmoid function which has 20 hidden units
- The output layer we used sigmoid activation function to return the probability that user would place an order.
- We added an l1-regulization term of 0.001 to avoid outfitting.



Logistic Regression

- We use logistic regression to train our inked user browsing &order data from training set
- predict whether the user would make a purchase for certain product given user browsing data.

Objective function:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}.$$

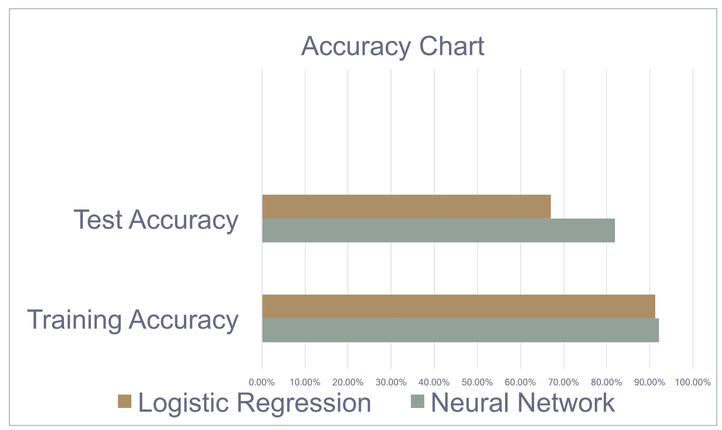
- If the output is greater than 0.5,then predict that user will make order.
- Otherwise, predict false that user won't make order based on the browsing info for certain products.

Result

Model	Training set accuracy	Test Set Accuracy	Sample Size (Train/Test)
Logistic regression	91.3%	67%	80k/20k
Neural Network	92.1%	82%	80k/20k

- We have achieved 91.3% accuracy on training set(80k samples) using logistic regression, and 67%(20k samples)accuracy on test set.
- For our neural network model, we achieved 92.1%(80k samples) accuracy on training set and 82%(20k samples) accuracy on test set.

Analysis



- Logistic regression shows high training set accuracy, but a lower test set accuracy pointing to overfitting
- Intuition is that Logistic regression will be limited in its ability to predict user behavior, due to limited feature set
- Neural Network behaves better, as the advantage is that the hidden layers can replicate the process of consumer's purchase behavior and explain the data
- Using neural network, the test set accuracy has been improved a lot from 67% to 82%, even if training accuracy between two models are similar, indicating the advantage of NN

Next Step & Future Work

- More pre-processing to increase the data corresponding to the actual orders as it provides better insights and improve accuracy
- Use longer time-series data(e.g 3 months or longer) for training to avoid the effect of certain month over customer behavior.
- Expand our model to predict top 10 selling products next week instead of a logistic model.
- Gather data from other e-commerce website selling multiple types of products to generalize our model for more types of products

Reference

- [1] Ricardo Filipe Fernandes e Costa Magalh‹es Teixeira(2015),Using Clickstream Data to Analyze Online Purchase Intentions, FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO