**Click Path Analysis Updates**

(WIP document)

Data sources (BigD data sources)

Omniture

MDS

Clickstream sessionizer

Customer resolution

So far:

\*At the session level

Aim:

* First cluster the sessions
* Predict **purchase/no-purchase** for each cluster above

Data prep:

* Brand = BR
* Time frame = March 2015
* ~370K (374,706) sessions in March 2015
* Session level data for the above brand and time frame
* Pre Processing
  + Compute Time on Page from the raw data
  + Use Prop33 and Pagename to generate the clickstream
  + Filter the values from the clickstream which are directly indicative of a purchase e.g “order confirm, shipping, billing”
  + Filter the sessions without any shopping cart interaction
* Data Sample Statistics
  + No of sessions in the sample ~ 178K (178,529)
  + No of sessions with a cart interaction ~70K (70,968
  + No of sessions with a cart + purchase ~ 13K (13,296)
* Distribution of length of sessions for purchase vs no purchase

**Clustering results**

* **Sessions with cart + purchase**
* Clustering methodology:
  + K-means with k ranging from 2 to 10
  + Picked k=5 (chosen for interpretability of clusters)
* Classification methodology
  + Features used = n-grams with n <= 3
  + Feature weights = Tried counts, TFIDF scores, time on page. Picked counts since all 3 fair equally. (+- 1%)
  + Feature space size = 23K (23,000)
  + Feature selection = Chi2 & L1, L2 Regularization. 5K (5000) features post feature selection
  + Linear SVC with L2 Regularization
  + Link to Top Features

RESULTS: Accuracy = 0.83, F1 score = 0.81

1. Short sequences of type *<category, product, zoom>*

Results: Accuracy - 54.00%, Mean Session Length – 48, Size - 3708

EXAMPLE:

1. Long & Repeating sequences of type *n0\*<category, product, zoom>*

Results: Accuracy – 96.77%, Mean Session Length – 261, Size - 62

EXAMPLE:

1. *<category, product, zoom>* pattern followed by

*n1\*<inline\_bag\_add>, <alt\_view\_product\_detail & view larger>*

& *n2\* <shopping\_cart>*.

Results: Accuracy – 69.22%, Mean Session Length – 81, Size - 1780

EXAMPLE:

1. *n3\*<alt\_view\_product\_detail & view larger>*

Results: Accuracy – 0.80%, Mean Session Length – 130, Size - 590

EXAMPLE:

1. Sessions that start with *<shopping bag>*. Mostly, a continuation of a previous session.

Results: Accuracy – 24.84%, Mean Session Length – 20, Size - 7156

EXAMPLE:

**Clustering results**

* **Sessions with cart + no-purchase**
* Clustering methodology:
  + K-means with k ranging from 2 to 10
  + Picked k=5 (chosen for interpretability of clusters)

1. Short sessions

Results: Percentage of Positive Predictions – 2.34%, Mean Length – 9, Size - 23906

EXAMPLE:

1. *<n0 \* category, product, zoom >* but excluding the pattern

*<category, product, zoom>*

Results: Percentage of Positive Predictions – 29.15%, Mean Length -150, Size - 320

EXAMPLE:

1. *Not a clean cluster*

Results: Percentage of Positive Predictions – 6.40%; Mean Length–32, Size - 8199

EXAMPLE:

1. *<category, product, zoom> & <n1\*category, product, zoom>*

Results: Percentage of Positive Predictions – 12.94%; Mean Length-69, Size - 724

EXAMPLE:

1. More number of *<quicklooks>* and *<inline bag adds>* than

*<category, product, zoom>*

Results: Accuracy – 24.84%, Mean Session Length – 20, Size - 2335

EXAMPLE:

\*Aim: predict **purchase/abandoned cart/neither** at the click level

* Classification methodology (as before)
  + Features used = n-grams with n <= 3
  + Feature weights = counts
  + Feature space size = 23K (23,000)
  + Feature selection = Chi2 & L1,L2 Regularization. 5K (5000) features post feature selection
* Classifiers trained on sessions with cart interactions only.
  + Subset 1 – Cart interaction but checkout not initiated, No Purchase
  + Subset 2 – Checkout initiated, No Purchase
  + Subset 3 - Purchasers
* Classifier 1 (96%) - Distinguishers between Non Purchasers and Potential Purchasers
* Classifier 2 (62%) – Refines the Potential Purchaser decision to identify cart abandoners vs purchasers.

\*Classification process

\*Results

\*Click by click predictions

\*Session length distribution for the test data set

\*Other approaches tried to improve classifier 2 (62% classifier)

1. Use an ensemble of classifiers based on
   1. Session Length
   2. Session Cluster
2. Defined a Count Vectorizer specific to this subset of data.