Predicting Customer Purchase Intent in E-commerce using Clickstream Data

SI 671 Data Mining Final Project | Michelle Cheng (michengz@umich.edu)

Background

With the prevalence of web tracking tools such as Google Analytics, e-commerce platforms are able to utilize click log data to inform about user behaviors and predict future actions. These analyses and predictions on behavioral patterns can help companies enhance user experience of their websites

Motivation

This project aims to explore and compare different methods on processing clickstream data to predict user purchase intent early on a user session.

Research Questions

Can Sequence Modeling on clickstream data outperform common feature-based classification models?

Are we able to predict purchase intents with limited volume of click-stream data? What is the minimum length of clickstream sequence required to make acceptable predictions?

How does the level of interaction (e.g. page-level, event-level) affect prediction performances? Does categorization of pageviews help with improving performance?

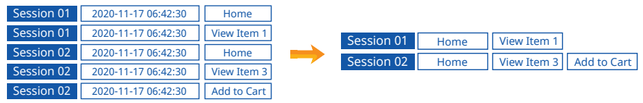
Dataset & Data Preprocessing

Google Online Merchandise Store Dataset

2,946,000 rows of click logs with 87840 unique user sessions (95.8% sessions with no purchases, 4.2% sessions with purchase)

Data Preprocessing

- **Labeling:** Session labeled as Purchase (1) if a purchase event is in sequence, otherwise, labeled as Non-Purchase (0).
- **Sessionize:** Transform click logs to event sequences that are grouped by sessions in preparation for modeling sequential data

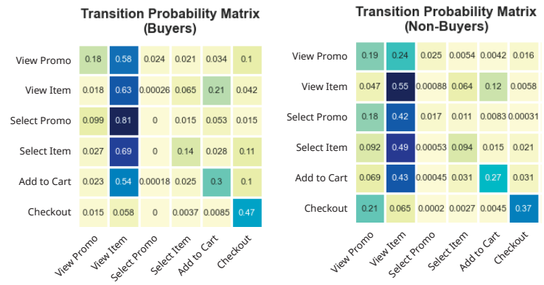


- **Undersampling:** Since the dataset is highly imbalanced with only 4.2% of purchase data, we downsampled our dataset to a 2:1 ratio of non-purchase data and purchase data.
- Purchase-related events in clickstreams are removed from training data to prevent data leakage

Methods

Sequence Discrimination with Markov Chains

To perform sequence discrimination, two separate transition probability matrices are first generated for each class (buyers and non-buyers):



Based on the transition probability matrices, the log odds ratio (also known as sequence discrimination value) of an observing sequence can then be calculated to determine its odds of stemming from either the buyers class or non-buyers class.

$$S(x) = log\frac{P(x \mid buyer)}{P(x \mid non{-}buyer)}$$

Sequence Classification with LSTM

- A common method used in the field of NLP to classify text sequences is the Long Short Term Memory (LSTM) model, a type of RNN that can learn longer patterns.
- In this project, we treat each clickstream sequence as sentences, in which each sequence can be represented as embeddings inside the neural network. Since our event streams are in different lengths, each sequence are transformed into padded sequences before passing in as training inputs.

session_id	padded_sequence	purchase
1000009091	[P3, P4, P1, P1, 0, 0, 0, 0, 0, 0]	0
1000075392	[P2, P3, P3, P4, P3, P3, P4, P2, P3, P3]	0
1000172345	[P2, P5, P7, P5, P5, P4, 0, 0, 0, 0]	0
1000175006	[P2, P3, P2, P2, P1, P2, P1, P4, P9, 0]	1
1000178538	[P2, P3, P3, P5, 0, 0, 0, 0, 0, 0]	0

Feature-based Standard Classification

- The non-sequential features extracted for training standard classification are the amount of clicks of each page in a sequence. Similar to storing term frequencies using Bag of Words (BoW) in NLP, we transform our data into feature vectors where each event represents a single feature.

	P1	P2	P3	P4	P5	P7	P9
[P3, P4, P1, P1] →	1	0	1	1	0	0	0
[P2, P3, P3, P4, P3, P3, P4, P2, P3, P3] →	0	2	6	2	0	0	0
[P2, P5, P7, P5, P5, P4] →	0	0	0	1	2	1	0
[P2, P3, P2, P2, P1, P2, P1, P4, P9] →	2	4	1	1	1	0	1
[P2, P3, P3, P5] →	0	1	2	0	1	0	0

- The Support Vector Classifier (SVC) was tested among other classifiers such as Logistic Regression and Random Forest, and was selected as our standard classification model for comparing with other methods.

Experiments & Results

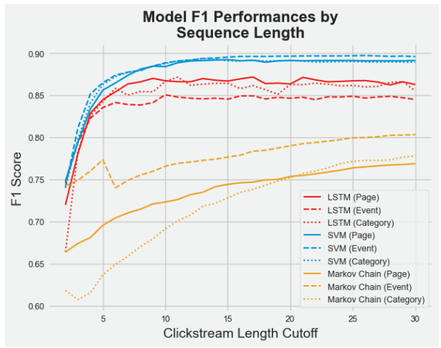
Order Estimation for Markov Chain Model: The order (k) of the Markov Chain Model is determined based on the AIC and BIC criterias (a likelihood ratio for model comparison) As a result, our Seuqence Dicrimination Method was implemented using a First-Order Markov Chain Model, in which every state only depends its preceding state.

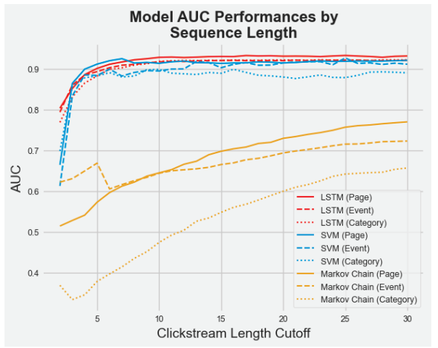
Cluster	Order	AIC	BIC
Purchase	0	358740.99630	358740.99630
Purchase	1	173331.35302	173694.08849
Purchase	2	173411.35302	174136.82396
Non-Purchase	0	1145267.89748	1145267.89748
Non-Purchase	1	505611.58550	506135.84022
Non-Purchase	2	505713.58550	506762.09494

Model Comparison:

Each model is trained and evaluated on different **clickstream lengths** (from 2 to 30 clicks long), and is also trained at different **interaction levels** (Page level, Event level, and Category Level). Models are evaluated using **F1 score** and **Area Under Curve (AUC).**

Results





Conclusion

- Overall, the feature-based standard classification model outperformed the sequential models in terms of f1 score. Results from the feature-based model are also more stable (less fluctuation) compared to models trained with sequential data. The LSTM sequential model slightly outperforms the SVM model in terms of AUC score when longer lengths of clickstreams are available for training.
- Given that most models reach f1 scores of 85% at the clickstream length of 4-5, we should have at least 4-5 click events in our observing sequence to ensure its credibility of making the right predictions.
- No significant patterns of whether a specific interaction level of clickstream performs better on models. However, we can tell from the results that the LSTM model almost always obtained higher scores when training on page-level sequences.

Future Steps

- Time-based features such as the amount of time spent on each page to be futher explored and incorporated to the feature-based classification model.
- Can consider performing customer segmentation using clustering methods to capture detailed user intentions (e.g. casually browsing, abandoned cart).