**Dell Shopping Cart Analysis**

**CSCI 6838 – Capstone**

**March 21, 2018**

Dawood Moazzem

Dell

Gary D. Boetticher

University of Houston – Clear Lake

Gorantla, Varun Kumar Chowdary

GorantlaV4942@UHCL.edu

Singh, Nikhil

SinghN7727@UHCL.edu

Tandel, Sonia

TandelS4845@UHCL.edu

Table of Contents

[I. Introduction 1](#_Toc510025352)

[II. Requirements 2](#_Toc510025353)

[*Data Definition* 2](#_Toc510025354)

[III. Modeling Process 4](#_Toc510025355)

[IV. System Design 5](#_Toc510025356)

[*Flowchart*: 5](#_Toc510025357)

[V. Implementation 6](#_Toc510025358)

[VI. Milestones 8](#_Toc510025359)

[VII. Results 9](#_Toc510025360)

[VIII. Conclusion 10](#_Toc510025361)

[IX. References 11](#_Toc510025362)

1. Introduction

Today, the websites are facilitating users to buy products of their choice. There has been an exponential growth in the number of Internet retailers and at the same time the amount of data generated has increased. Along with the customer purchase data, the websites are gaining a tremendous amount of user-specific information in the form of clicks made, frequency of purchases, etc. Analyzing customer data can help organizations make decisions for achieving better sales, improving revenue and increasing customer retention.

The challenges organization face includes extracting insights from the overwhelming amount of customer data using predictive analysis while maintaining performance. The customer data can provide insight into traffic variables (column ‘prop’) pertaining to the page visits, page views, and unique visitors.

This project performs predictive analysis for Dell customer’s website. It seeks to predict whether a customer is going to abandon a shopping cart before it happens. The data contains factors like user-specific information, user’s geo-location information, the website visited, the source of the clicks and social media information which affect the customer hit rate in varying scores. PySpark, Pandas, NumPy and other python libraries are used to perform data pre-processing and cleaning. The project aims to build a model based on the machine learning techniques.

1. Requirements
2. *Functional and Non-Functional Requirements*

The following table contains the requirements along with the functionality type and priority.

*Table 1 - Requirements*

|  |  |
| --- | --- |
| **Requirement** | **Date** |
| Data Acquisition and Reading | 21/02/2018 |
| Data Cleaning/Reducing | 19/03/2018 |
| Data Transformation and Grouping | 19/03/2018 |
| Data Massaging | 13/04/2018 |
| Model Building-Testing and Training | 26/04/2018 |
| Deployment | 26/04/2018 |

1. *System Requirements*

The following table contains the system requirements.

*Table 2 - System Requirements*

|  |  |
| --- | --- |
| RAM | 6GB or more |
| Processor | 2 or more |
| Storage | Cloud |
| Operating System | Linux, Windows |

1. *Software Requirements*

Tool/Platform which allows to store, load and iterate through massive amounts of data. Preferred: Anaconda, Jupyter notebook with Python 3.

Must support python libraries like Pandas, PySpark, NumPy, etc.

### *Data Definition*

The datasets used in this project have been obtained from the Dell proprietary data resource. The data is in the form of clickstream. The datasets used is: ‘hit\_data.tsv’ which contains the user-specific information, user’s geo-location information, the website visited, the source of the clicks and social media information. ‘column\_header.tsv’ contains all the header information.

The dataset consists of 1103464 rows and 971 columns.

1. *Dataset Specifications:*

|  |  |
| --- | --- |
| hit\_data.tsv: | 3.7 GB |
| column\_header.tsv | 11.5 KB |
| No. of rows | 1103464 |
| No. of columns | 971 |

1. *List of Attributes:*

The following are list of some of the attributes of dataset.

|  |  |
| --- | --- |
| **Attribute** | **Value** |
| accept\_language | en-US,en;q=0.8 |
| browser\_height | 640 |
| browser\_width | 1024 |
| post\_referrer | http://www.dell.com/support/home/us/en/19/products/?app=drivers |
| post\_prop14 | http://www.dell.com/support/home/us/en/19/product-support/servicetag/1gkfq61/drivers |
| post\_event\_list | 221,236,110,131,132,000,000,000,000,000,000,000,000,000,000,000,000,000 |
| post\_visid\_high | 1658310662 |
| post\_visid\_low | 925910793 |
| post\_evar11 | 0|0|0|0|0|2410|2514|12089|16468|17547|17547|n|n|n|n|n|y|y|y|y|y|n |
| post\_evar53 | us|en|dhs|19|esupport-home|productsupport|servicetag |
| post\_prop35 | usa:Consumer Online |
| post\_t\_time\_info | 29/0/2018 14:6:48 1 360 |

1. Modeling Process

The Data-Requirement gathering (Data cleaning and Pre-processing / Initial Visualization) involves cleaning and preprocessing the data. Removing duplicate values, columns with low variance, high correlation and null values results in a reduction of 3.6 Gigabytes. Initial visualization is performed on the data to identify inconsistencies and eliminate outliers in the data.

The Model Building - Training phase focuses on building and training a predictive analysis model to identify the users who have abandoned the cart.

The Validation phase and evaluation phase validates the outcome (prediction) of the resultant model of the previous phase against the known set of results.

The Testing phase and Deployment phase evaluates the resultant model derived from the validation phase by applying the model to the sample data.

1. System Design

### *Flowchart*

Data Cleaning and Pre-processing phase

Model Testing phase

Model Training phase

Model Validation phase

1. Implementation

Some steps recommended by Adobe Clickstream Analytics (ACA) [1] are used in the Data cleaning phase.

Data cleaning and Preprocessing steps followed:

1. Remove rows where value of column exclude\_hit is less than 0: the hits for this value are to be excluded as the visit number is not incremented for this value. (ACA)
2. Filter rows with source 5,7,8,9 (ACA)

|  |  |
| --- | --- |
| Source | Explanation |
| 5 | Contains generic ecommerce data source information. |
| 7 | Transaction id/ customer id data source file |
| 8 | Search center data sources |
| 9 | Adobe social summary metrics |

1. Merge the columns post\_visid\_low and post\_visid\_high to obtain a unique user\_id column. (ACA)
2. Filter users according to the cart data i.e. who has accessed cart.
3. Remove columns with the null values exceeding some threshold (50%, 20%).
4. Remove duplicate columns.
5. Remove columns with high correlation and low variance.
6. Obtain the list of unique values of user\_id column, determined by users who had made a purchase and who abandoned the cart by checking for value “1” in post\_event\_list column for each row.
7. Obtain the rows of users who abandoned cart and who made a purchase using these lists of unique user\_id values.
8. User grouping: extract users who abandoned the cart and who made a purchase and group them accordingly by using groupby functions.

*Column information*:

Some columns include:

*Table 3 - Column Information*

|  |  |  |
| --- | --- | --- |
| Column name | Column description | Data type |
| Post\_Event\_list | Comma separated list of numeric IDs representing events passed in from the customer. | String (64k characters) |
| Post\_UniqueId | Combination of post\_visid\_high and post\_visid\_low. | 64-bit unsigned integer |
| Post\_Referrer | Page prior to the current page | String (244 characters) |
| Post\_Prop14 | Reflecting the path information |  |
| Post\_Evar53 | Conversion variable | String (100 characters) |
| Post\_Evar71 | Conversion variable | String (100 characters) |

1. Milestones

**Key code:**

V =Varun Kumar Chowdary Gorantla

N = Nikhil Singh

S = Sonia Tandel

|  |  |  |  |
| --- | --- | --- | --- |
| **Due Date** | **Activity** | **Percent Contribution** | **Status** |
| Feb. 8 | Construct a project website | V (33.33), N (33.33), S (33.33) | Done: Jan. 27 |
| Feb. 25 | Data Cleaning | V (33.33), N (33.33), S (33.33) | Done: Feb. 26 |
| Mar. 19 | Data Reducing and Grouping | V (33.33), N (33.33), S (33.33) | Done: Mar. 18 |

1. Results

The initial dataset has been preprocessed, which resulted in dataset with reduced row and column count. This dataset contains data of users who has accessed their cart at least once. This dataset is used to derive information about the users who made a purchase or abandoned cart and group them accordingly. This is a cleaned dataset and it can be transformed into required format for building a predictive model by using machine learning techniques.

1. Conclusion

The main concern for online retailers these days is persuading the customer to buy their product. Few customers buy the product and few of them add the product to the cart ultimately not making a purchase. The data cleaning and data preprocessing has been achieved at present.

Using machine learning techniques, the future goal is to build a model to capture the patterns exhibited by customers who add a product to the cart but abandon it without making a purchase. The model can also be used to predict if a user will abandon a shopping cart or not. This model intends to help the organizations retain their customers thus, increasing the sales and improving the revenue.

1. References

[1] Adobe. (n.d.). Clickstream (Legacy) Data Feeds Help [Blog post]. Retrieved from <https://marketing.adobe.com/resources/help/en_US/sc/clickstream/>

[2] Adobe. (n.d). Traffic Variables sProps Inside Omniture SiteCatalyst [Blog post]. Retrieved from <https://theblog.adobe.com/traffic-variables-sprops/>

[3] Adobe. (n.d). Conversion Variables – Part I Inside Omniture SiteCatalyst [Blog post]. Retrieved from <https://theblog.adobe.com/conversion-variables-part-i/>