MCDA5580 Assignment 3

Team Member

- Hemalatha Srinivasan A00452621
- Ajay Jain A00455849
- Kin Wa, Chan A00467755

Table of Content

Executive Summary

This report gives a clear picture of the association of website functions on Simplycast.com. It will give various action usage frequencies which can associate in user level and session level. The report details the user behavior of the portal, and highlights areas for improvement to enhance customer satisfaction and drive revenue growth. The report also discusses the association between various actions and their results obtained for various processes.

Based on the report, we can be able to see there is a frequent usage of ManageTab and ProjPreview tab and less usage of InsertLinkAnchor. Users find challenges in using InsertLinkAnchor both at user level and session level.

For future goals, to overcome these challenges in few actions, we can follow below strategies:

1. Provide easier access (e.g. sort on the top of the page) to frequent used functions like ManageTab & ProjPreview
2. Provide shortcut for frequent used action sets, like { ManageTab ,ProjPreview, SendNow} and {ProjPreview, SendNow}
3. By Providing effective help documentation for understanding the functionality using screenshots
4. Additional support by giving Chatbot assistance for any topic related to portal usage which can make their experience easy
5. Offering trial periods for using a few paid functionalities which can make user to get attracted to the functionality. Through which we can increase the probability of buying chance by the customers

Objectives

SimplyCast is a company that provides software solutions for marketing communication automations (e.g., email automation, Twitter automation). The company website "https://www.simplycast.com/" Played an important role in the business, it serves many functions including the presentation of production information, show use cases/ demos, providing customer service and support. It is important for the company to promote its products and maintain its services to customers.

In this report a study will perform to analyze the associations between functions from the browsing history of the website to achieve these goals:

- Reveal user behavior, adjust marketing strategy. E.g., Track associations of the products browsed, develop packages to attract customers.
- Improve user experience in UI. E.g., Discover frequent used functions and make them conspicuous on UI,

- Discover potential UI design problems E.g., Users have difficulties getting access to a function due to wrong descriptions or over complexity on UI.

Data Analysis

The browsing history data (from 20, July, 2015 to 17, December, 2015) is gathered from https://www.simplycast.com/ for analysis. There are two sets of data: browsing history in customer level and browsing history in session level:

Customer Level Browsing History (39096 rows)

Data Field	Data Type	Description
id	Integer	Unique ID to identify a user
milestone	String	Function accessed in website

Session Level Browsing History (173082 rows)

Data Field	Data Type	Description
user_id	Integer	Unique ID to identify a user
date	Date	Date of access
milestone_name	String	Function accessed in website

The Customer Level Browsing History recorded all functions accessed by each user and in Session Level Browsing History, the “date” field is added to specify which date the user accessed the function. All functions accessed by a user on the same date is supposed within a session.

There is no null value or invalid value found in the browsing history extracted.

Design/ Methodology/ Approach

Overview

In order to analyze the associations between functions in the websites, Apriori algorithm is used. It is because the data is not labeled so an unsupervised learning method is needed. In addition, this method could be able to list out all association rules with specific support and confidence which are intuitive and easy to communicate with end users.

There are two types of browsing history extracted from the website: Customer level browsing history and Session level browsing history. Analysis by Apriori algorithm will be carried out separately on these two sets of data and final summary will be made based on the results.

Customer Level Analysis

Data Columns:

ID – User id

Milestone – Various actions on the website performed by the user. Only unique values per user in extracted data. There are 112 unique milestones, count plot of top 20 is shown below.

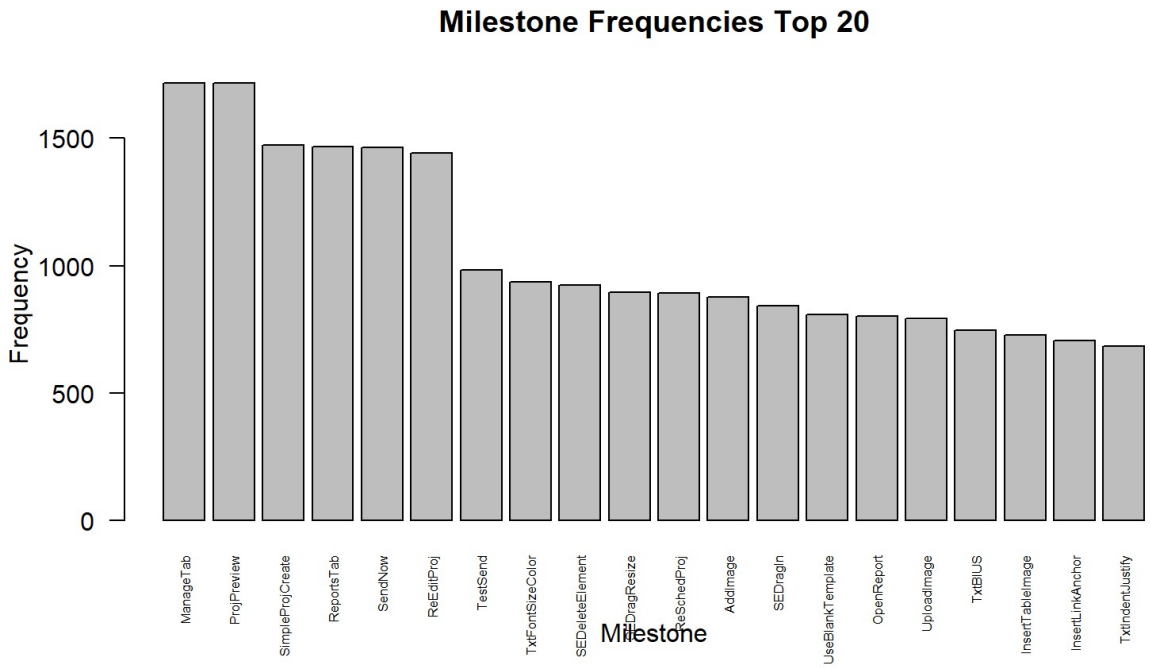

Figure 1 Customer Level Milestone Frequency Plot

Apriori algorithm is used to analyze the user level data. Here is the list of top 20 rules (support =0.2, confidence =0.5) ordered by lift.

	lhs		rhs	support	confidence	coverage	lift
[1]	{AddImage, SEDragResize}	=>	{UploadImage}	0.2044951	0.9124294	0.2241216	3.630182
[2]	{AddImage, SimpleProjCreate}	=>	{UploadImage}	0.2238050	0.9075738	0.2465970	3.610864
[3]	{AddImage, SEDeleteElement}	=>	{UploadImage}	0.2013295	0.8995757	0.2238050	3.579042
[4]	{SEDragResize, UploadImage}	=>	{AddImage}	0.2044951	0.9907975	0.2063944	3.568905
[5]	{SimpleProjCreate, UploadImage}	=>	{AddImage}	0.2238050	0.9901961	0.2260209	3.566738
[6]	{UploadImage}	=>	{AddImage}	0.2484964	0.9886650	0.2513454	3.561223
[7]	{AddImage}	=>	{UploadImage}	0.2484964	0.8950969	0.2776195	3.561223
[8]	{ProjPreview, UploadImage}	=>	{AddImage}	0.2067110	0.9878971	0.2092434	3.558457
[9]	{UploadImage, SEDeleteElement}	=>	{AddImage}	0.2013295	0.9875776	0.2038620	3.557306
[10]	{AddImage, ProjPreview}	=>	{UploadImage}	0.2067110	0.8908595	0.2320355	3.544364
[11]	{ProjPreview, TxtFontSizeColor}	=>	{TxtBIUS}	0.2044951	0.7600000	0.2690725	3.222604
[12]	{ProjPreview, TxtBIUS}	=>	{TxtFontSizeColor}	0.2044951	0.9500000	0.2152580	3.206250
[13]	{TxtBIUS}	=>	{TxtFontSizeColor}	0.2206394	0.9355705	0.2358341	3.157550
[14]	{TxtFontSizeColor}	=>	{TxtBIUS}	0.2206394	0.7446581	0.2962963	3.157550
[15]	{TxtFontSizeColor}	=>	{TxtIndentJustify}	0.2019626	0.6816239	0.2962963	3.152635
[16]	{TxtIndentJustify}	=>	{TxtFontSizeColor}	0.2019626	0.9341142	0.2162077	3.152635
[17]	{SEDragResize, SimpleProjCreate}	=>	{AddImage}	0.2003799	0.8588874	0.2333017	3.093757
[18]	{SEDragIn}	=>	{AddImage}	0.2200063	0.8234597	0.2671732	2.966145
[19]	{AddImage}	=>	{SEDragIn}	0.2200063	0.7924743	0.2776195	2.966145
[20]	{ProjPreview, SEDragResize}	=>	{SEDeleteElement}	0.2000633	0.8598639	0.2326686	2.933380

Figure 2 Customer Level Rules Listing

Network Plot of the rules:

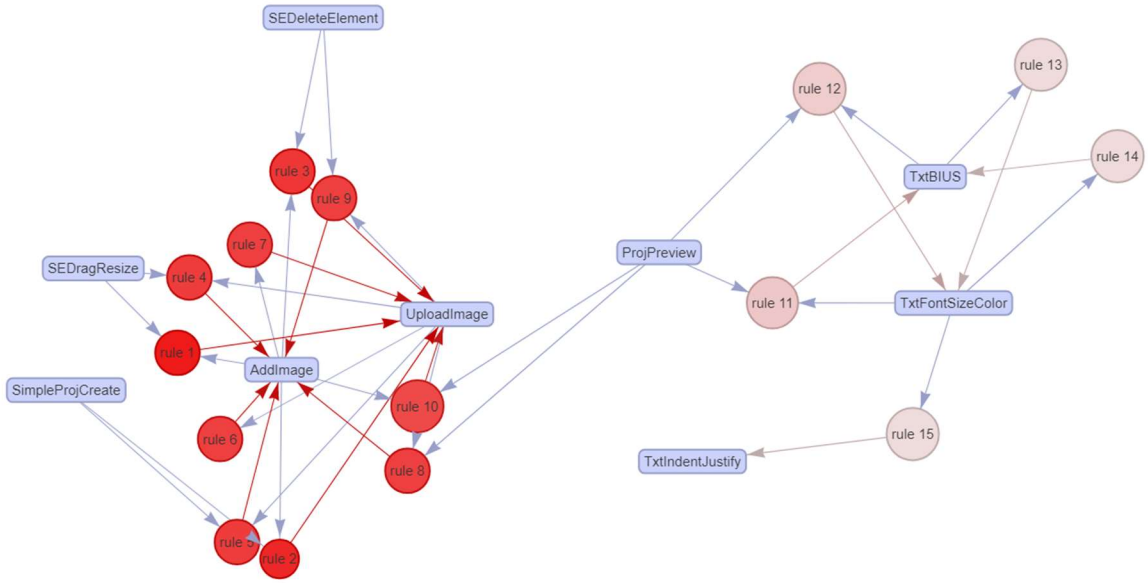

Figure 3 Customer Level Network Plot

Most frequent item set by support

	items	support
[1]	{ManageTab}	0.5432099
[2]	{ProjPreview}	0.5428933
[3]	{ManageTab, ReportsTab}	0.4283001
[4]	{ManageTab, SendNow}	0.4169041
[5]	{ManageTab, ProjPreview}	0.4159544
[6]	{ProjPreview, ReEditProj}	0.3979107
[7]	{ProjPreview, SendNow}	0.3972776
[8]	{ManageTab, ReEditProj}	0.3817664
[9]	{ProjPreview, ReportsTab}	0.3710035
[10]	{ReEditProj, SendNow}	0.3697373
[11]	{ManageTab, ProjPreview, SendNow}	0.3665717
[12]	{ReportsTab, SendNow}	0.3577081
[13]	{ManageTab, ProjPreview, ReportsTab}	0.3551757
[14]	{ManageTab, ProjPreview, ReEditProj}	0.3485280
[15]	{ManageTab, ReportsTab, SendNow}	0.3472618
[16]	{ManageTab, ReEditProj, SendNow}	0.3469452
[17]	{ReEditProj, ReportsTab}	0.3437797
[18]	{ProjPreview, ReEditProj, SendNow}	0.3374486
[19]	{ManageTab, SimpleProjCreate}	0.3301678
[20]	{ManageTab, ReEditProj, ReportsTab}	0.3301678

Figure 4 Customer Level Frequent Item Set

Analysis

Top rule is “{AddImage, SEDragResize} => {UploadImage}”. This rule has lift of 3.63 and confidence of 0.91 which is very high. Support for this rule is 0.20 which indicates almost 20% of transactions contain this itemset. This rule indicates that when user accesses AddImage and SEDragResize then there is high probability that user will also access UploadImage.

Next two rules are similar to the first one. SEDragResize is replaced by SimpleProjCreate and SEDeleteElement. These rules also have life in (~3.6) range and confidence in (~0.90) range.

Rule 4 is “{SEDragResize, UploadImage} => {AddImage}”. It has life of 3.56 and confidence is very high 0.99. Support is around 0.20. It indicates that it is highly likely that users will access AddImage when they access SEDragResize and UploadImage.

Rule 5 is “{SimpleProjCreate, UploadImage} => {AddImage}”. It is similar to rule 4, only change is that SEDragResize is replaced by SimpleProjCreate.

Rule 6 and 7 are “{UploadImage} => {AddImage}” and “{AddImage} => {UploadImage}”. These have lift of (~3.56) and confidence of (~0.89 to 0.98) and support of (~0.24). Both the rules indicate that AddImage and UploadImage are highly likely to be accessed together.

Rule 8, 9 and 10 are similar to first five rules with ProjPreview added to the itemsets and its different combinations.

Rule 11 is “{ProjPreview, TxtFontSizeColor} => {TxtBIUS}”. It has lift of 3.2, confidence of 0.76 and support of 0.20. Confidence of this rule is lower than other rules but it is still significant. It indicates that it is highly likely that TxtBIUS is accessed when ProjPreview and TxtFontSizeColor are accessed.

Rule 12 is “{ProjPreview, TxtBIUS} => {TxtFontSizeColor}”. It is similar to rule 11 but confidence is 0.95 which is even higher.

Rule 13, 14, and 15 are {TxtBIUS} => {TxtFontSizeColor}, {TxtFontSizeColor} => {TxtBIUS} and {TxtFontSizeColor} => {TxtIndentJustify}. These have lift of (~3.15). These indicates that TxtBIUS, TxtIndentJustify and TxtFontSizeColor are frequently accessed together. It makes sense intuitively as well since they belong to text formatting group of functions.

Session Level Analysis

Session level browsing history is gathered from website to proceed the analysis. There is the summary of the most frequent user accessed functions:

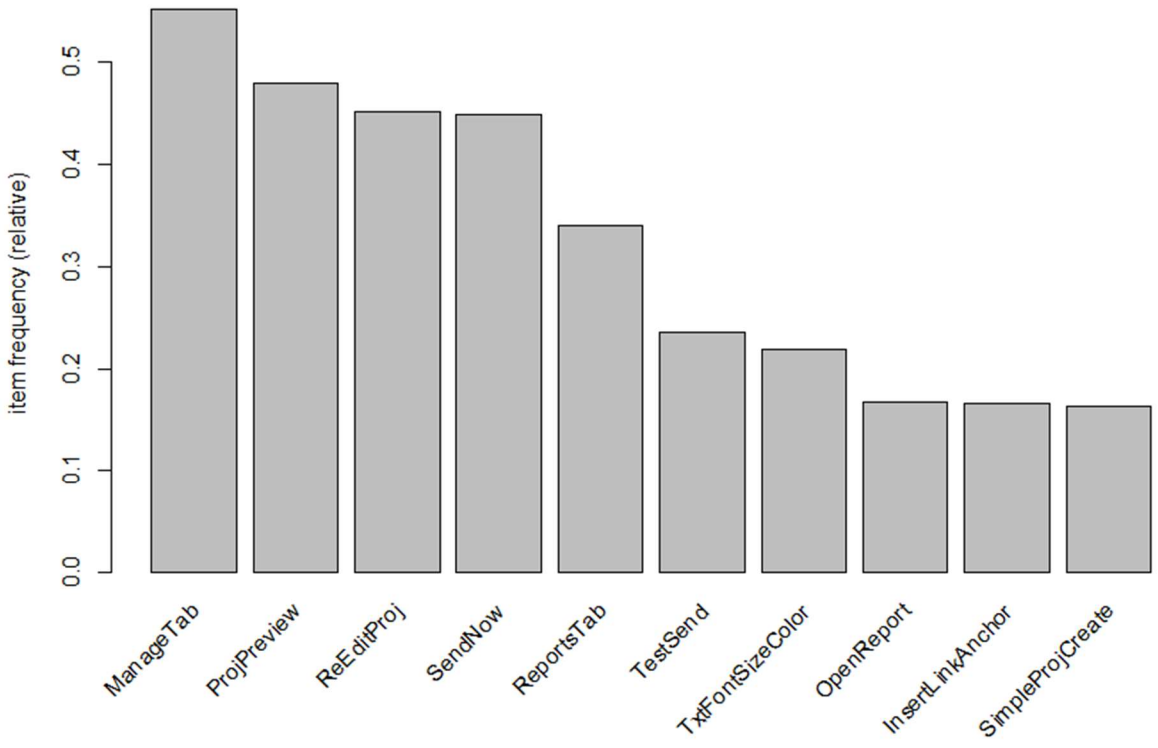

Figure 5 Session Level Milestone Frequency Plot

Apriori algorithm is used to analyze the session level data. Here is the list of top 16 rules (support =0.25, confidence =0.5) ordered by lift.

```
     lhs                          rhs           support   confidence coverage  lift     count
[1]  {ManageTab, ProjPreview} => {SendNow}     0.2541982 0.7968037  0.3190224 1.774001  6282
[2]  {ProjPreview, SendNow}   => {ManageTab}   0.2541982 0.9042752  0.2811071 1.638850  6282
[3]  {SendNow}                => {ManageTab}   0.4047263 0.9010811  0.4491563 1.633061 10002
[4]  {ManageTab}              => {SendNow}     0.4047263 0.7334996  0.5517744 1.633061 10002
[5]  {ReEditProj}             => {SendNow}     0.2681180 0.5930368  0.4521102 1.320335  6626
[6]  {SendNow}                => {ReEditProj}  0.2681180 0.5969369  0.4491563 1.320335  6626
[7]  {ManageTab, SendNow}     => {ProjPreview} 0.2541982 0.6280744  0.4047263 1.310393  6282
[8]  {ProjPreview}            => {SendNow}     0.2811071 0.5864922  0.4793024 1.305764  6947
[9]  {SendNow}                => {ProjPreview} 0.2811071 0.6258559  0.4491563 1.305764  6947
[10] {ProjPreview}            => {ReEditProj}  0.2739854 0.5716336  0.4793024 1.264368  6771
[11] {ReEditProj}             => {ProjPreview} 0.2739854 0.6060145  0.4521102 1.264368  6771
[12] {ProjPreview}            => {ManageTab}   0.3190224 0.6655973  0.4793024 1.206285  7884
[13] {ManageTab}              => {ProjPreview} 0.3190224 0.5781754  0.5517744 1.206285  7884
[14] {ReEditProj}             => {ManageTab}   0.2966455 0.6561353  0.4521102 1.189137  7331
[15] {ManageTab}              => {ReEditProj}  0.2966455 0.5376210  0.5517744 1.189137  7331
[16] {}                       => {ManageTab}   0.5517744 0.5517744  1.0000000 1.000000 13636
```

Figure 6 Session Level Rules Listin

Plot for the rules

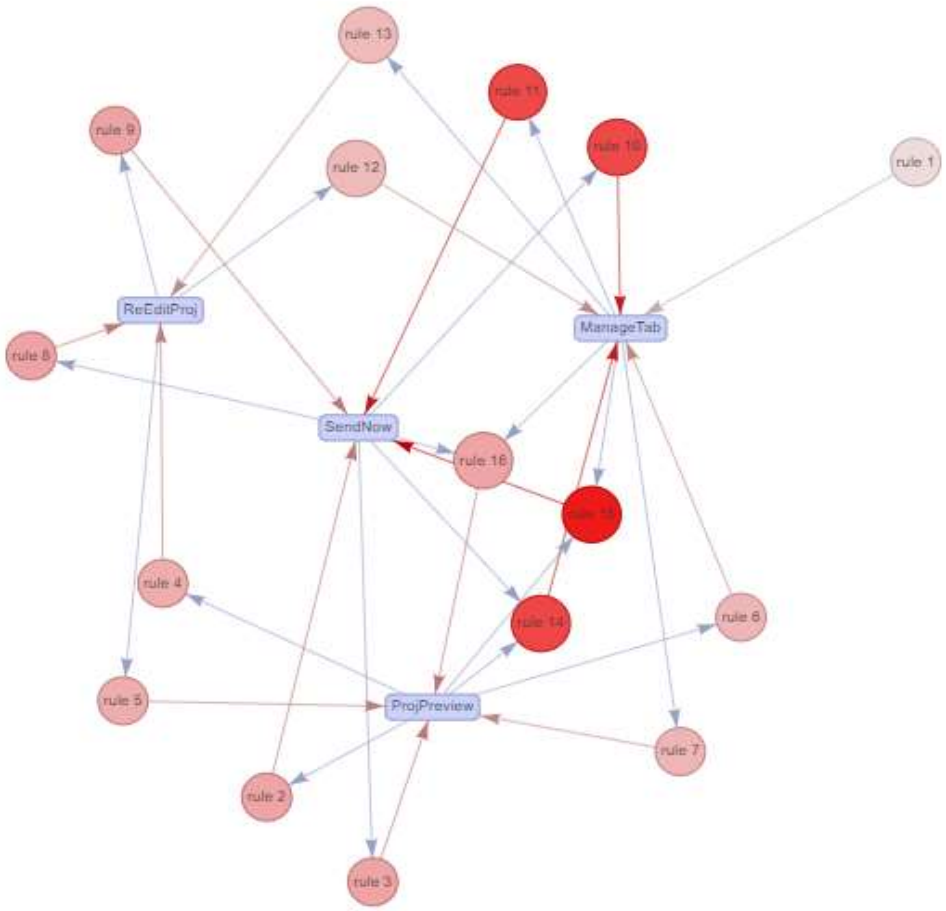

Figure 7 Session Level Network Plot

Most frequent item set by support

	items	support
[1]	{ManageTab}	0.5517744
[2]	{ManageTab, SendNow}	0.4047263
[3]	{ManageTab, ProjPreview}	0.3190224
[4]	{ManageTab, ReEditProj}	0.2966455
[5]	{ProjPreview, SendNow}	0.2811071
[6]	{ProjPreview, ReEditProj}	0.2739854
[7]	{ReEditProj, SendNow}	0.2681180
[8]	{ManageTab, ProjPreview, SendNow}	0.2541982

Figure 8 Session Level Frequent Item Set

Analysis

The top rule is {ManageTab, ProjPreview} => {SendNow} have both high confidence (~0.8) and lift (~1.77) which means user highly likely to access {SendNow} when {ManageTab, ProjPreview} are accessed. Moreover, the support (~0.25) of this rule is high, it implies a quarter of records contains the {ManageTab, ProjPreview} item set. These all make this rule very meaningful.

In the second rule {ProjPreview, SendNow} => {ManageTab}, the items listed are the same as the first rule and also with high value in confidence (~0.9), lift (~1.64) and support (~0.25). Therefore, we can see there is close relationship between the items {ManageTab, ProjPreview}, and {SendNow}.

The rules 3 and 4 are {SendNow} => {ManageTab} and {ManageTab}=> {SendNow} respectively. They have very high support value (~0.4) and similar confidence (~0.73 to 0.9) and lift (~1.63) compared with rules 1 and 2. It shows the associations appear extremely frequent and relationship between {SendNow}, {ManageTab} is strong.

The rules 5 and 6 are {ReEditProj} => {SendNow} and {SendNow} => {ReEditProj}, the patterns are similar to rules 3 & 4 ({ManageTab} is replaced by {ReEditProj}) with high support (~0.27) but lower confidence (~0.59 to 0.6) and lift (~1.32). It shows the association between {ReEditProj} and {SendNow} are strong that should not be ignored.

The remaining rules are somehow the combinations of the most frequent items ({ManageTab}, {ProjPreview}, {ReEditProj}, {SendNow}) . It shows the importance of the items on the website and also verify the consistency of our analysis.

Conclusion

Data Analysis for Association mining is done using Apriori Algorithm in this report. Here we used this algorithm to analyze both customer and session level. Data Analysis involved preparing the data by removing the unwanted data which can diverge the results. Next step involves understanding the data and their behavior based on their usage. With the achieved data, we implement our algorithm to find the data sets which are all used frequently. The end results will reveal the relationship between these data sets.

From the Customer level Analysis, we were able to understand that users use Manage Tab and ProjPreview actions more often than other actions. The users used these actions more than 1500 times and reaches the highest position. They both share an almost equal number of counts from the analytics we have done.

{AddImage, SEDragSize} => { UploadImage } gains highest confidence and Lift value of 0.9124294 and 3.63. This combination of actions on the left for the desired action on the right maintains highest confidence.

When we are analyzing the data based on the support value achieved, it also results in the same action as highest support value namely, ManageTab with support value of 0.5432 and ProjPreview with support value of 0.5428.

From the Session Level analysis, we can sense that users have more frequency count on ManageTab and ProjPreview actions. ManageTab has a more relative frequency count of 0.5 and above while ProjPreview has the next highest relative frequency of 0.4 and above.

{ManageTab, ProjPreview} => { SendNow } and {ProjPreview, SendNow} => {ManageTab} shares same highest count of 6282 and have confidence of 0.796 and 0.904 respectively. They have the highest lift among other combinations of actions which are 1.774 and 1.638 respectively.

Finally, when we achieved the support value for all the actions based on session value, we received highest support value for ManageTab and {ManageTab, SendNow} with value 0.5517 and 0.4047 respectively.

Appendix

R Script

Customer Level Analysis

```
#rm(list = ls())    # Clear Environment
#cat("\014")       # Clear Console

library(arules)
library(reshape2)
library(arulesViz)
library(Hmisc)
library(tidyr)
library(dplyr)

# read input csv file

user_data <- read.csv("userDistinctMilestoneDec151.csv",
                      header = TRUE)

# Plot frequency count of milestone

freq_table <- sort(table(user_data$milestone), decreasing=TRUE)

print(freq_table)

# calculate the proportion distribution
prop_table <- round(prop.table(freq_table) * 100, 2)
```

```
print(prop_table[1:20])

barplot(freq_table,
        main="Milestone Frequencies",
        xlab="Milestone",
        ylab="Frequency",
        cex.names=0.5,
        las=2
        )

# Plot top 20 milestones

barplot(freq_table[1:20],
        main="Milestone Frequencies Top 20",
        xlab="Milestone",
        ylab="Frequency",
        cex.names=0.5,
        las=2
)

# convert data to basket format
user_basket_data <- pivot_wider(user_data,
                                names_from=milestone,
                                values_from=milestone,
                                values_fn=length,
                                values_fill=0)

# Drop id column
```

```
user_basket_data <- select(user_basket_data, -1)

# loop over each column and convert 1 to TRUE and 0 to FALSE
for (i in 1:ncol(user_basket_data))
 {
 user_basket_data[, i] <-
       ifelse(user_basket_data[, i] == 1, TRUE, FALSE)
 }

# rules mining using apriori
association_rules <- apriori(user_basket_data,
                      parameter= list(supp=0.2,conf=0.5))

# sort the rules by life
rules_sorted <- sort(association_rules,
                by = "lift",
                decreasing = TRUE)

# print top 20 rules
inspect(rules_sorted[1:20],linebreak=FALSE)

# Interactive plot for top 15 rules
plot(rules_sorted[1:15],
    method="graph",
    engine="visNetwork")


# List item set
```

```
itemsets=unique(generatingItemsets(association_rules))

inspect(sort(itemsets,by='support'),linebreak=FALSE)
```

Session Level Analysis

```
library(plyr)

library(arules)

library(arulesViz)


df_session = read.csv("D:/#Spring 2023/5580 - Text
Mining/Assignment3/sessionDistinctMilestoneDec15 (2).csv")


df_session$id = paste(as.character(df_session$user_id) , as.POSIXct(df_session$date, format="%Y-%m-
%d"))



# Transpose source

df_session= ddply(df_session,c("id"),function(dfl)paste(dfl$milestone_name, collapse=","))


# Remove id field

df_session$id = NULL

df_session$date = NULL

df_session$user_id = NULL


# Write to temp file

write.table(df_session,"D:/#Spring 2023/5580 - Text
Mining/Assignment3/session2.csv",quote=FALSE,row.names=FALSE,col.names=FALSE)


# Read temp file
```

```
tr = read.transactions("D:/#Spring 2023/5580 - Text
Mining/Assignment3/session2.csv",format="basket",sep=",")


# Plot Frequent items

#summary(tr)

itemFrequencyPlot(tr, topN=10)


# List the rules

rules<- apriori(tr, parameter= list(supp=0.25, conf=0.5))


inspect(sort(rules,by='lift'))


# Visualize the rules

plot(rules,

    method="graph",

    engine="visNetwork")


# List item set

itemsets=unique(generatingItemsets(rules))

inspect(sort(itemsets,by='support'))
```

Reference/ Citation

Comparison of Association Algorithm

https://www.educba.com/association-rules-in-data-mining/

https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/

Apriori Algorithm

https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

https://www.engati.com/glossary/apriori-algorithm

https://www.javatpoint.com/apriori-algorithm