Northeastern University

Toronto

Initial Analysis Report

Soni Manan ***NUID**: 002982645*

Savaliya Parth ***NUID**: 002982302*

Dave Dhairya ***NUID**: 002110382*

College of Professional Studies, Master of Professional Studies in Analytics.

***Subject**: ALY6015*

Under the guidance of

Dr. Prof. Alex Maizlish

Introduction

Data has been released by H&M for their benefits and put as competition on Kaggle. All of the dataset contains data regarding their products and customers like clothing's, accessories and customers membership status and their activeness. These data can be used for recommendation of their product based on sales. Initially it looked like we need to feature variables and subset them to get proper predictions.

Descriptive Analysis

- Articles dataset has 105502 samples and 25 variables.
- Customers dataset has 1371980 samples and 7 variables in which postal and id are hashed so useless as well as many rows has NA which cant be replaced with mean and median so replaced with mode.
- Most of the values are character and others are assigned id.

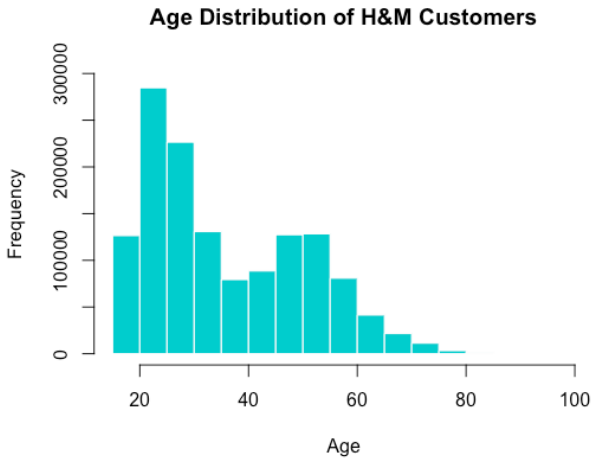
```
'data.frame':   105542 obs. of  25 variables:
 $ article_id                  : int  108775015 108775044 108775051 110065001 110065002 110065011 111565001 111565003 111586001 1115
93001 ...
 $ product_code                : int  108775 108775 108775 110065 110065 110065 111565 111565 111586 111593 ...
 $ prod_name                   : chr  "Strap top" "Strap top" "Strap top (1)" "OP T-shirt (Idro)" ...
 $ product_type_no             : int  253 253 253 306 306 306 304 302 273 304 ...
 $ product_type_name           : chr  "Vest top" "Vest top" "Vest top" "Bra" ...
 $ product_group_name          : chr  "Garment Upper body" "Garment Upper body" "Garment Upper body" "Underwear" ...
 $ graphical_appearance_no     : int  1010016 1010016 1010017 1010016 1010016 1010016 1010016 1010016 1010016 1010016 ...
 $ graphical_appearance_name   : chr  "Solid" "Solid" "Stripe" "Solid" ...
 $ colour_group_code           : int  9 10 11 9 10 12 9 13 9 9 ...
 $ colour_group_name           : chr  "Black" "White" "Off White" "Black" ...
 $ perceived_colour_value_id   : int  4 3 1 4 3 1 4 2 4 4 ...
 $ perceived_colour_value_name : chr  "Dark" "Light" "Dusty Light" "Dark" ...
 $ perceived_colour_master_id  : int  5 9 9 5 9 11 5 11 5 5 ...
 $ perceived_colour_master_name: chr  "Black" "White" "White" "Black" ...
 $ department_no               : int  1676 1676 1676 1339 1339 1339 3608 3608 3608 3608 ...
 $ department_name             : chr  "Jersey Basic" "Jersey Basic" "Jersey Basic" "Clean Lingerie" ...
 $ index_code                  : chr  "A" "A" "A" "B" ...
 $ index_name                  : chr  "Ladieswear" "Ladieswear" "Ladieswear" "Lingeries/Tights" ...
 $ index_group_no              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ index_group_name            : chr  "Ladieswear" "Ladieswear" "Ladieswear" "Ladieswear" ...
 $ section_no                  : int  16 16 16 61 61 61 62 62 62 62 ...
 $ section_name                : chr  "Womens Everyday Basics" "Womens Everyday Basics" "Womens Everyday Basics" "Womens Lingerie"
...
 $ garment_group_no            : int  1002 1002 1002 1017 1017 1017 1021 1021 1021 1021 ...
 $ garment_group_name          : chr  "Jersey Basic" "Jersey Basic" "Jersey Basic" "Under-, Nightwear" ...
 $ detail_desc                 : chr  "Jersey top with narrow shoulder straps." "Jersey top with narrow shoulder straps." "Jersey to
p with narrow shoulder straps." "Microfibre T-shirt bra with underwired, moulded, lightly padded cups that shape the bust and provid
```

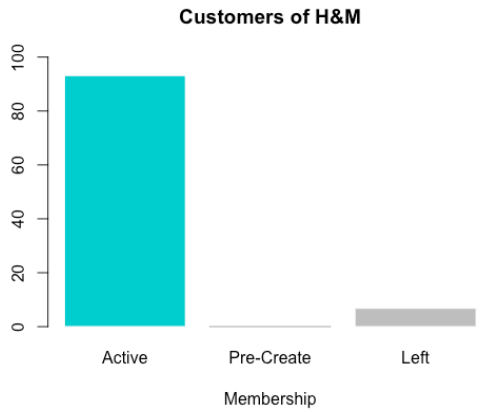
```
   Elders Youngsters
   37.24158   62.75842
```

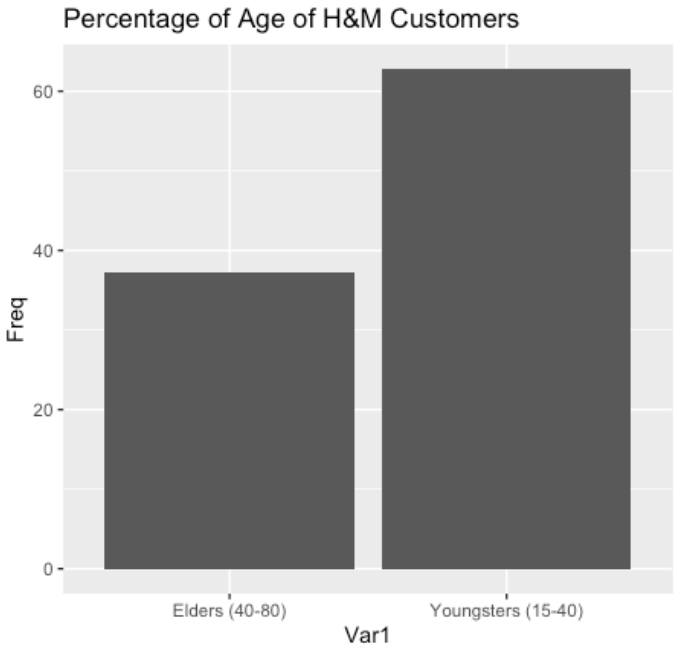
```
summary(clubMemberStatus$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  16.00   24.00   32.00   36.38   49.00   99.00   13575
```

- Mean age is 36.38 and we consider age 15 to 40 are youngsters who are into good fashion where 37% elders of H&M customers.
- Age distribution of H&M is not normal as we can judge like H&M is brand for upcoming fashion and provide clothing's according to west. So most of are in the range of 0-40 and little increment between 50-70 age group.
- 90% are active who follows news for whatever reason, some have pre-created membership and few left like 7% has left membership.

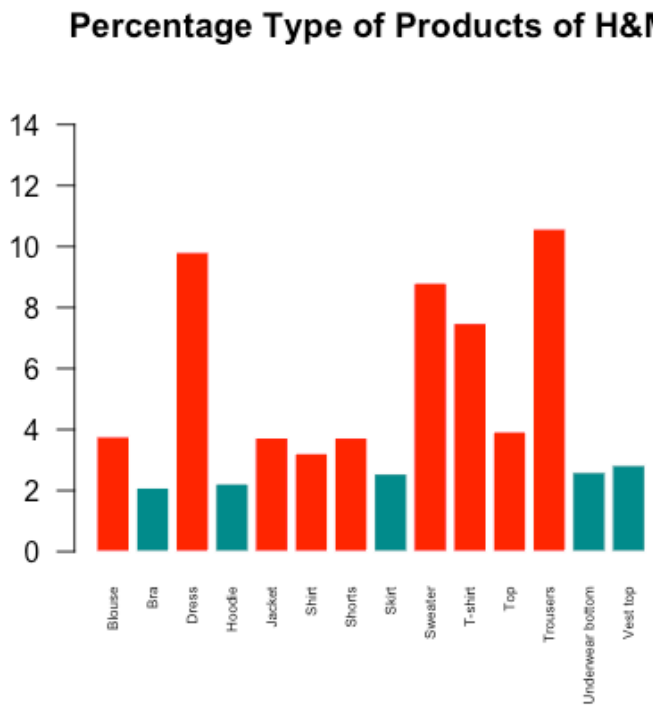
- Cleaning such as removing hasn't take place while making this analysis as it can affect.



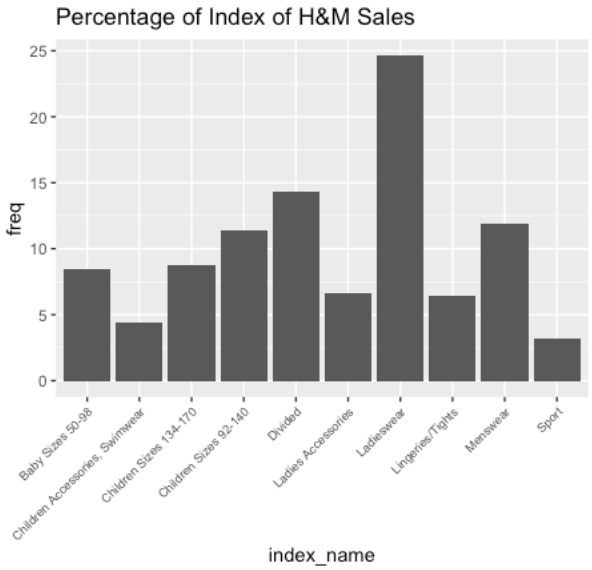




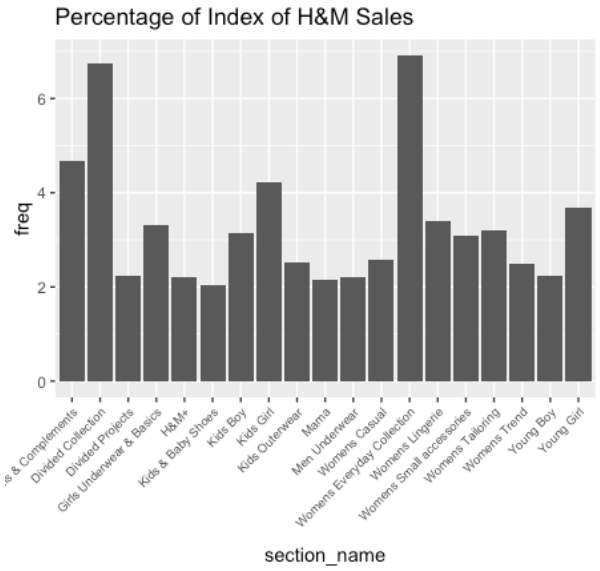
- 90% are active who follows news for whatever reason, some have pre-created membership and few left like 7% has left membership.
- As we can see, Trousers, Dresses, Sweaters, Tees and Top are top 5 whose cumulative relative frequency is almost 45%. Among that Trousers and Dresses tops with 11 and 10% respectively.



- Index name are just like category of articles.
- Ladieswear, tops with 25% in this where Divided, and Menswear are almost 15%. Following that, Childrenswear are at 10%.

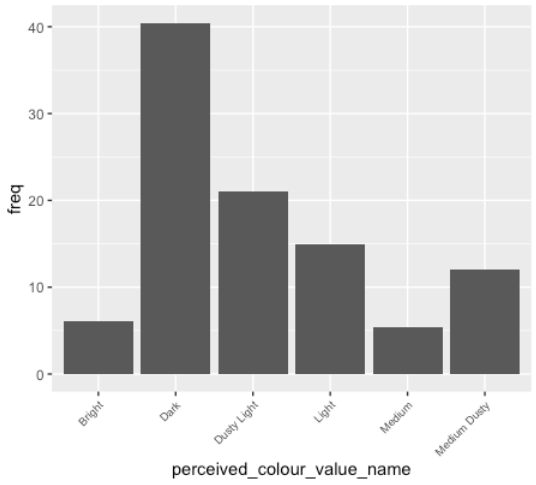


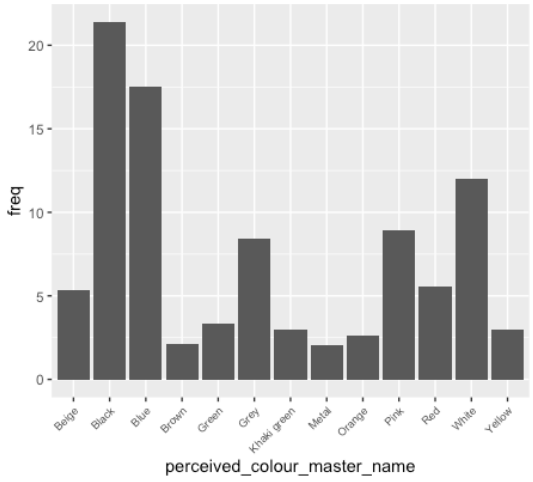
- This is uneven distribution of indexes at H&M. Following upper diagram, this follows same as that but in descriptive way. For ladieswear this has different ladieswear section like Everyday collection, Casual, etc. We can answer questions by subsetting or labelling this.



- This is uneven distribution of indexes at H&M. Following upper diagram, this follows same as that but in descriptive way. For ladieswear this has different ladieswear section like Everyday collection, Casual, etc. We can answer questions by subsetting or labelling this.

- This is overall colour preferences of customers. We can see Dark has preferred by 40% people. Dusty light and light has preferred by 20 and 15% customers respectively.

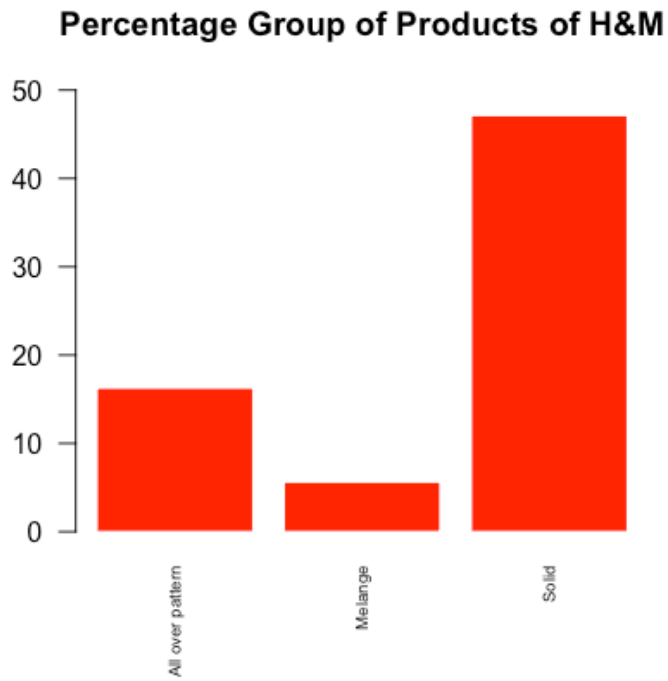




- To be precise, in Dark, Black and Blue are highly preferred by 37%. White and Pink are second highest preference of customers.

- To be precise, in Dark, Black and Blue are highly preferred by 37%. White and Pink are second highest preference of customers.

- So people around 50% people prefers Solid over patterns and etc. So if a research question ask like what are the chances of a customer will buy a t-shirt with solid or with some kind of pattern.



Models and Predictions

- A Logistic regression is a type of regression model we use when the answer of variable is like yes or no.
- As we have variables which are categories weather it is menswear or ladies or is it accessory or not.
- So Multiple logistic regression, we will use to answer our research question from which H&M can get benefits.
- There can be multiple because if we need work to improve the model accuracy or there can be multiple independent variables.

Conclusion

From this we can learn how to make a logistic regression model for binary solutions. How to calculate miscalculation (False Negatives and Positives) by model.

Finding the accuracy and Performance by finding confusion matrix, ROC and AUC.

For the dataset, From the Applications, Accept, Fulltime Grads, Outstates and PhD are the variables who leading the model to 98%. As I mentioned, correlation matrix plays important role just like EDA and feature modelling.

Accuracy and AUC is just higher than 90-95% are consider as overfitting model however for qualified data it is possible.

Most of target audience should be youngsters and less elders as number of sales accounted for western styles like trousers, tops, t-shirts and etc.

Next Milestone

1) Next step should be getting confidence using hypothesis testing before going for the model.

Getting confidence is important. Most of data is categorical so we can go with Chi-Square test.

Statically confident is something we need to get as to please the authorities how strongly we are suggesting just instead telling.

2) Splitting the data between train and test and then use GLM to train model and predict using that.

Detect and remove outliers from process. Also identify the outliers using AIC, Cooks, QQ plots.

Getting ROC, AUC, Recall, Precision and Accuracy for the model.

References

H&M Personalized Fashion Recommendations, Provide product recommendations based on previous purchases, H&M Group *Sources*: https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data

Appendix

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

articles = read.csv("./articles.csv")
customers = read.csv("./customers.csv")

psych::describe(customers)
summary(customers)

unique(customers$club_member_status)
library(dplyr)

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

customers[customers$club_member_status == "",] = Mode(customers$club_member_status)
clubMemberStatus = customers %>% filter(club_member_status != "")
clubMemberStatus$club_member_status                                                    =
as.numeric(as.factor(clubMemberStatus$club_member_status))
clubMemberStatus$fashion_news_frequency                                                =
as.numeric(as.factor(clubMemberStatus$fashion_news_frequency))

print(prop.table(table(clubMemberStatus$club_member_status))*100)
barplot(prop.table(table(clubMemberStatus$club_member_status))*100,
        names.arg = c("Active","Pre-Create","Left"),
        xlab="Membership",
        col=ifelse(prop.table(table(clubMemberStatus$club_member_status))*100                >
80,'cyan3','grey'),
        border=ifelse(prop.table(table(clubMemberStatus$club_member_status))*100
<5,'gray','white'),
        main="Customers of H&M",
        ylim=c(0,100))

hist(clubMemberStatus$age,
        xlab="Age",
        main="Age Distribution of H&M Customers",col='cyan3',border='white',
        ylim=c(0,300000))

brackets <- clubMemberStatus %>% mutate(agegroup = case_when(age > 0  & age <= 15 ~ 'Teen',
                                        age > 15  & age <= 40 ~ 'Youngsters (15-40)',
                                        age > 40  & age <= 80 ~ 'Elders (40-80)')) # end
function
```

```
age_brackets = as.data.frame(prop.table(table(brackets$agegroup)) * 100)
ggplot(age_brackets,aes(x=Var1,y=Freq))+geom_bar(stat='identity')  +labs(title='Percentage  of Age of H&M Customers')


summary(clubMemberStatus$age)


psych::describe(articles)
str(articles)
# PRODUCT SALES RELATIVE PLOT
salesAsType = articles %>%  count(product_type_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq > 2)
barplot(salesAsType$freq,names.arg=salesAsType$product_type_name,ylim=c(0,15),main='Percentage Type of Products of H&M',cex.names = 0.5,las=2,col=ifelse(sales$freq > 3,'Red','cyan4'),border='white')


salesAsTypeGroup = articles %>%  count(product_group_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq > 1)
barplot(salesAsTypeGroup$freq,names.arg=salesAsTypeGroup$product_group_name,ylim=c(0,50),main='Percentage Group of Products of H&M',cex.names = 0.5,las=2,col='Red',border='white')


salesAsGraphics = articles %>%  count(graphical_appearance_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq > 5)
barplot(salesAsGraphics$freq,names.arg=salesAsGraphics$graphical_appearance_name,ylim=c(0,50),main='Percentage Group of Products of H&M',cex.names = 0.6,las=2,col='Red',border='white')


library('ggplot2')
salesAsIndex = articles %>%  count(index_name) %>% mutate(freq = n / sum(n)*100)
ggplot(salesAsIndex,aes(y=freq,x=index_name))+geom_bar(stat='identity')+theme(axis.text.x = element_text(angle = 45,hjust=1,size=7))+labs(title='Percentage of Index of H&M Sales')


salesAsSection = articles %>%  count(section_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq>2)
ggplot(salesAsSection,aes(y=freq,x=section_name))+geom_bar(stat='identity')+theme(axis.text.x = element_text(angle = 45,hjust=1,size=7))+labs(title='Percentage of Index of H&M Sales')


salesAsColor = articles %>%  count(perceived_colour_value_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq>2)
ggplot(salesAsColor,aes(y=freq,x=perceived_colour_value_name))+geom_bar(stat='identity')+theme(axis.text.x = element_text(angle = 45,hjust=1,size=7))+labs(title='Percentage of Index of H&M Sales')


salesAsColorName = articles %>%  count(perceived_colour_master_name) %>% mutate(freq = n / sum(n)*100) %>% filter(freq>2)
ggplot(salesAsColorName,aes(y=freq,x=perceived_colour_master_name))+geom_bar(stat='identity')+theme(axis.text.x = element_text(angle = 45,hjust=1,size=7))+labs(title='Percentage of Index of H&M Sales')


featureGraphics = as.data.frame(articles$graphical_appearance_name)
featureGraphics$Solid      =      ifelse(featureGraphics$`articles$graphical_appearance_name` =='Solid',1,2)
colnames(featureGraphics) = c("graphical_name",'solid')
featureGraphics$graphical_name = as.numeric(as.factor(featureGraphics$graphical_name))
```

```
library(misclassGLM)

mIndex =  sample(c(1,2), nrow(featureGraphics),
                 replace = T,
                 prob = c(0.7,0.3))
train_x = featureGraphics[mIndex == 1,]
test_x = featureGraphics[mIndex == 2,]
head(train_x)

LR_model <- glm(solid ~ graphical_name,
                data = train_x,
                family = binomial(link = "logit"))
summary(LR_model)

prob.train_x = predict(LR_model,
                       newdata = train_x,
                       type = "response")
cm_data = as.factor(ifelse
                    (prob.train_x >= 0.5,
                      1, 2))

library(caret)
confusionMatrix(cm_data,
                as.factor(ifelse(train_x$solid == 1, 1,2)))

prob.test_x = predict(LR_model,
                      newdata = test_x,
                      type = "response")
prob.test_x

cm_data = as.factor(ifelse
                    (prob.test_x >= 0.5,
                      1, 2))
cm_data
head(cm_data)

confusionMatrix(cm_data,
                as.factor(ifelse(test_x$solid == 1, 1,2)),
                )

library(pROC)
ROC = roc (test_x$solid, prob.test_x)
X =  plot(ROC,
          col = "black",
          ylab = "Sensitivity = TP rate",
          xlab = 'specificity = FP rate')
```

```
#08  Calculate and interpret the AUC.
AUC = auc(ROC)
AUC
```