Causal Inference Final Report

TEAM 17:Yuting Xin, Wenzhe Ou, Chian Hsieh, Sharang Jindal, Yali Li

3/1/2020

Business Context

A retail company with an online website has multiple marketing channels, including email, TV, and social media platform. The spending on advertising accounted for 50% in the overall budget. As a result, optimizing the Return On Investment (ROI) of investing advertisements in each campaign would be the priority. The company started an email campaign, which sends two kinds of emails, one featuring Men's merchandise, the other promoting Womens merchandise. They tracked the customers' behaviors during the campaign such as visiting the website through the link in email, making a purchase and their spendings. The dataset contains information of customers who have purchased in the last 12 months. With the dataset, we are aiming at evaluating the effectiveness of the email campaign.

Key Question

How does the email campaign affect visits, conversion and spending?

Key Findings

From our analyses, we discovered that the email campaigns have effects on website visits and conversions but not on spending. We also found that historical spending, whether customers have purchased products in the past 12 months and recency are important factors that affect the effectiveness of the email campaigns. We also recommend the company implement Causal Forest model when applying marketing campaigns.

Data Cleaning

Load packages

```r
library(dplyr)
# install.packages('grf')
library(grf)
library(tidyverse)
library(ggplot2)
```

Load dataset

```r
data = read.csv("email.csv")
```

Separate dataset into three groups

```r
men<-data%>%filter(segment=='Mens E-Mail')
women<-data%>%filter(segment=='Womens E-Mail')
no<-data%>%filter(segment=='No E-Mail')
```

Data Description

Basic information about the dataset

The dataset can be downloaded from https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.htm.

It contains information about an email campaign of a company.

Data Exploration

Look at what the dataset looks like

```
head(data)
```

```
##   recency history_segment history mens womens  zip_code newbie channel
## 1      10  2) $100 - $200  142.44    1      0 Surburban      0   Phone
## 2       6  3) $200 - $350  329.08    1      1     Rural      1     Web
## 3       7  2) $100 - $200  180.65    0      1 Surburban      1     Web
## 4       9  5) $500 - $750  675.83    1      0     Rural      1     Web
## 5       2    1) $0 - $100   45.34    1      0     Urban      0     Web
## 6       6  2) $100 - $200  134.83    0      1 Surburban      0   Phone
##         segment visit conversion spend
## 1 Womens E-Mail     0          0     0
## 2     No E-Mail     0          0     0
## 3 Womens E-Mail     0          0     0
## 4   Mens E-Mail     0          0     0
## 5 Womens E-Mail     0          0     0
## 6 Womens E-Mail     1          0     0
```

Each unit of observation is a customer. The treatment variable is segment while the target variables are visit, conversion and spend. The dataset also contains some other features of customers:

Recency: Months since last purchase.

History_Segment: Categorization of dollars spent in the past year.

History: Actual dollar value spent in the past year.

Mens: 1/0 indicator, 1 = customer purchased Mens merchandise in the past year.

Womens: 1/0 indicator, 1 = customer purchased Womens merchandise in the past year.

Zip_Code: Classifies zip code as Urban, Suburban, or Rural.

Newbie: 1/0 indicator, 1 = New customer in the past twelve months.

Channel: Describes the channels the customer purchased from in the past year.

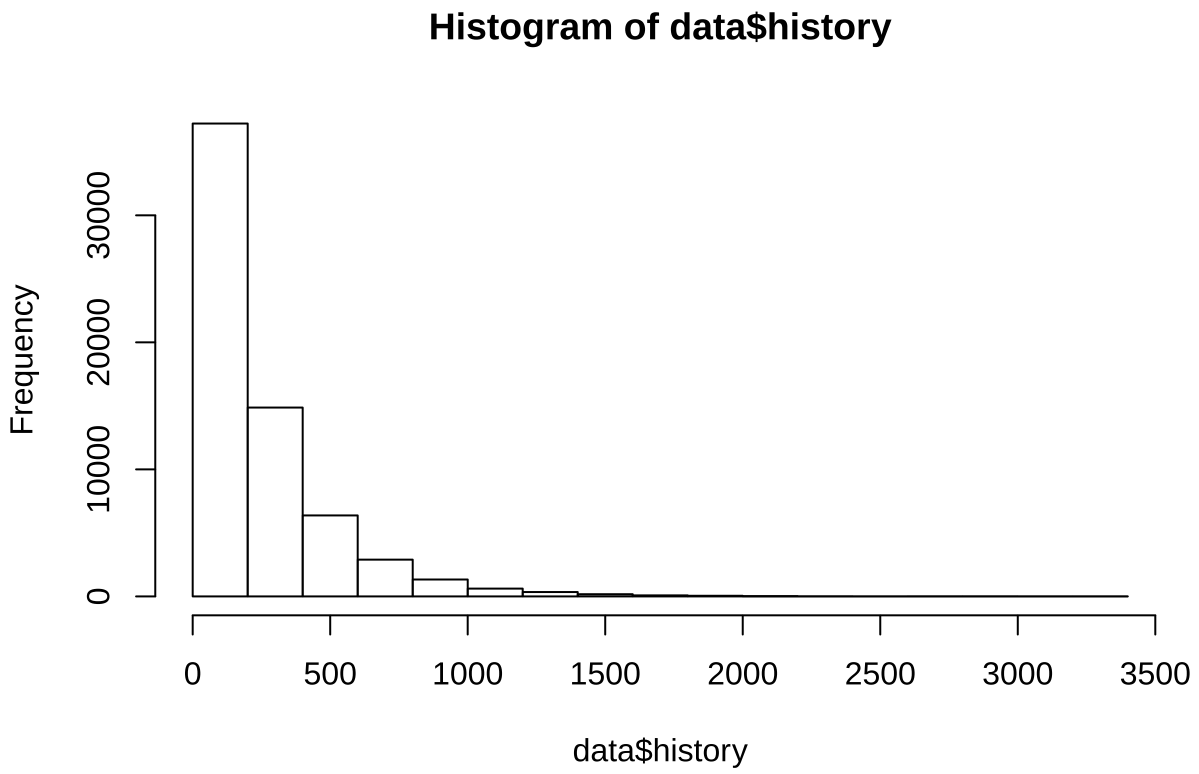
```
nrow(data)
```

```
## [1] 64000
```

There are 64000 observations in the dataset.

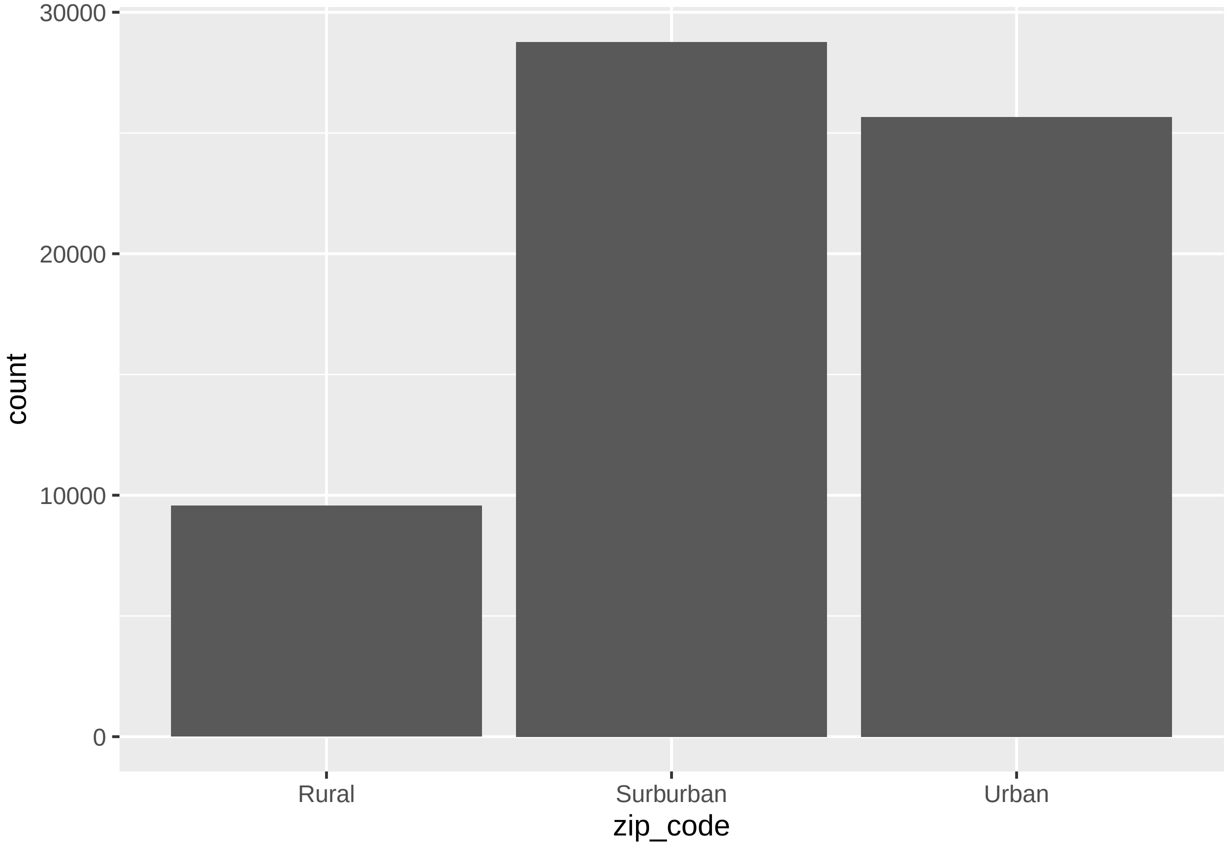
Look at the summary statistics and distributions of variables

```
hist(data$history)
```



The distribution of historical spending of customers is right-skewed.

```
ggplot(data, aes(x=zip_code)) + geom_bar()
```



We can see that most of the customers are in urban or suburban areas.

```
data %>% dplyr::group_by(visit) %>% dplyr::summarise(count=dplyr::n())
```

```
## # A tibble: 2 x 2
##   visit count
##   <int> <int>
## 1     0 54606
## 2     1  9394
```

```
data %>% dplyr::group_by(conversion) %>% dplyr::summarise(count=dplyr::n())
```

```
## # A tibble: 2 x 2
##   conversion count
##        <int> <int>
## 1          0 63422
## 2          1   578
```

```
data_spend <- data %>% filter(spend!=0)
summary(data_spend$spend)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.99   32.27   80.80  116.36  153.35  499.00
```

We can see that there are significantly more people who didn't visit and convert. As for spending, for customers who purchased, the average spending is $116.36.

Experimental Design

Randomization Check

```
# As the space is limited, I only show randomization check for one variable.
summary(aov(mens ~ segment, data = data))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## segment         2      0 0.09846   0.398  0.672
## Residuals   63997  15833 0.24740
```

As there is no significant difference between any groups of any variable, we can conclude that randomization was properly conducted.

```
data %>% dplyr::group_by(segment) %>% dplyr::summarise(count=dplyr::n())
```

```
## # A tibble: 3 x 2
##   segment       count
##   <fct>         <int>
## 1 Mens E-Mail   21307
## 2 No E-Mail     21306
## 3 Womens E-Mail 21387
```

There are three groups of customers: a group that didn't receive email, a group that received emails advertising on men products and a group that received emails advertising on women products. The smallest group is 21306. Let's look at the power of the experiment.

Power of the experiment

```
power.t.test(n  =21306, sig.level = 0.1, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 21306
##           delta = 0.02408528
##              sd = 1
##       sig.level = 0.1
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

We can reasonably detect a change no smaller than 0.02.

Threats to Causal Inference

After examining the empirical context and the sample data we used, we are faced one of the endogeneity factors, *Sampling Error*. The 64,000 customers may not be representative of the whole customer population, so our conclusions may not be applicable to the whole customer base.

Causal Inference methods:

Problem Description

In order to evaluate whether women's and men's campaigns are successful, we perform statistical tests to confirm whether customer visits, conversions or spendings on the website are affected by men's campaign or women's campaign.

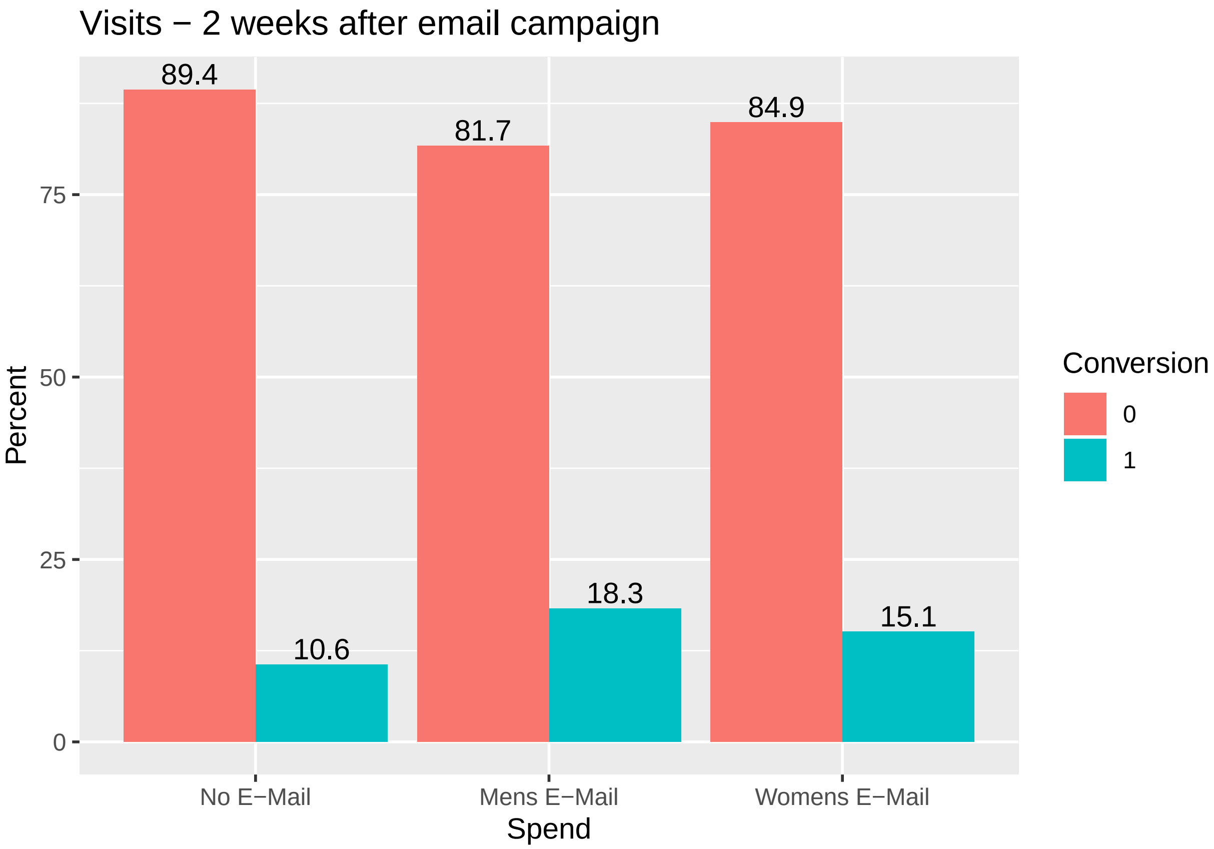
Approach

Overall impact of Campaigns

As visits and conversions are all binary variables, the team used logistic regression to identify whether Men's or Women's campaign affects the probability of the customers' visits or conversions significantly compared to customers getting no email. Similarly, linear regression has been used to identify effects of Men's or Women's campaign on spending of customers on the website. These techniques quantify the impacts of men's and women's campaigns, which also help in evaluating relative performance of each campaign.

Overall impact of campaigns on visits to website:

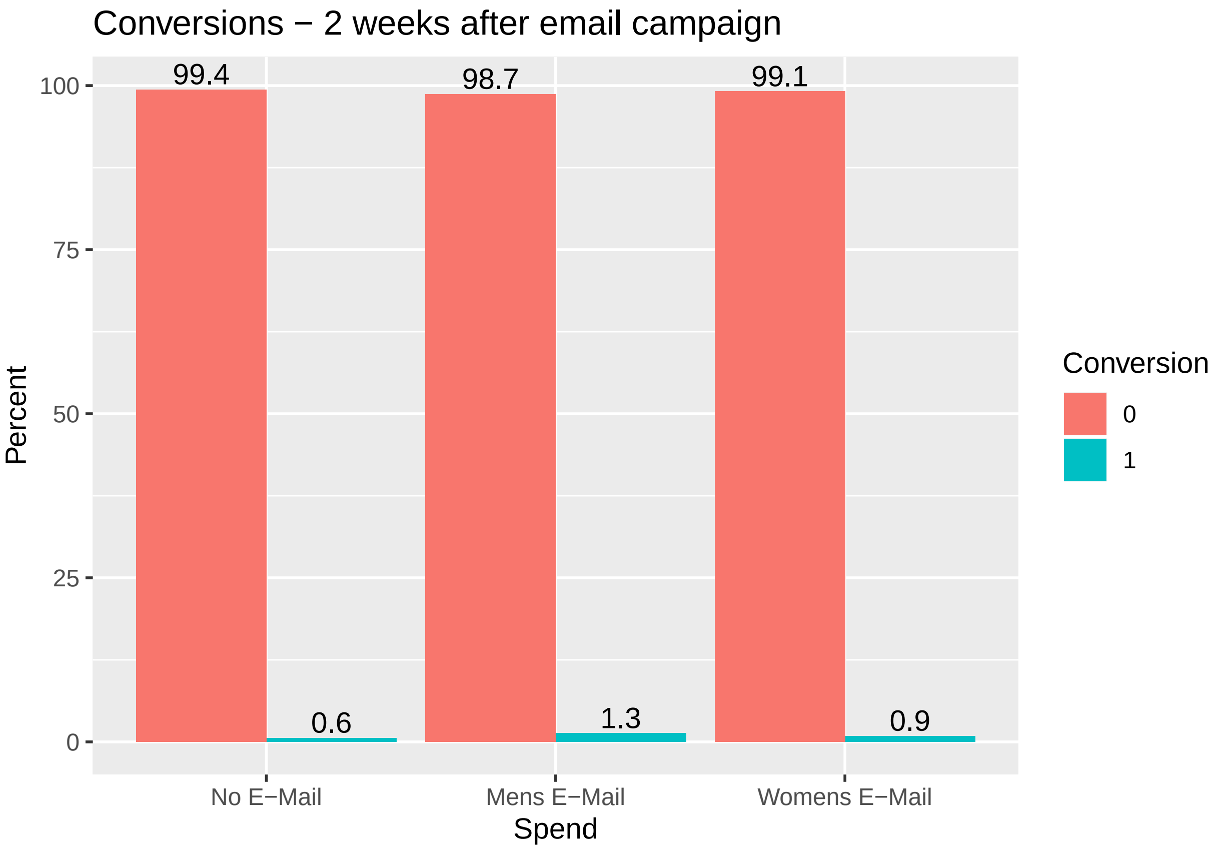
The following plot shows that 10.6% customers who did not get any email visited the website, compared to 18.3% customers who received men's emails and 15.1% customers who received women's emails. Higher proportion of customers who received Men's or Women's emails visited the website, however, we need to evaluate whether this change is significantly different from customers who did not receive any emails.



Results from the logistic regression in section M1.1 (Appendix) confirm that there is a difference in visit behavior of customers who did not receive any emails and customers who received either Men's or Women's emails. Further, we see that log(odds) of visiting the website improve by 0.63 for customers receiving men's emails and 0.40 for customers receiving women's emails. So, overall Men's campaign is more successful in increasing visits on the website.

Overall impact of campaigns on conversions:

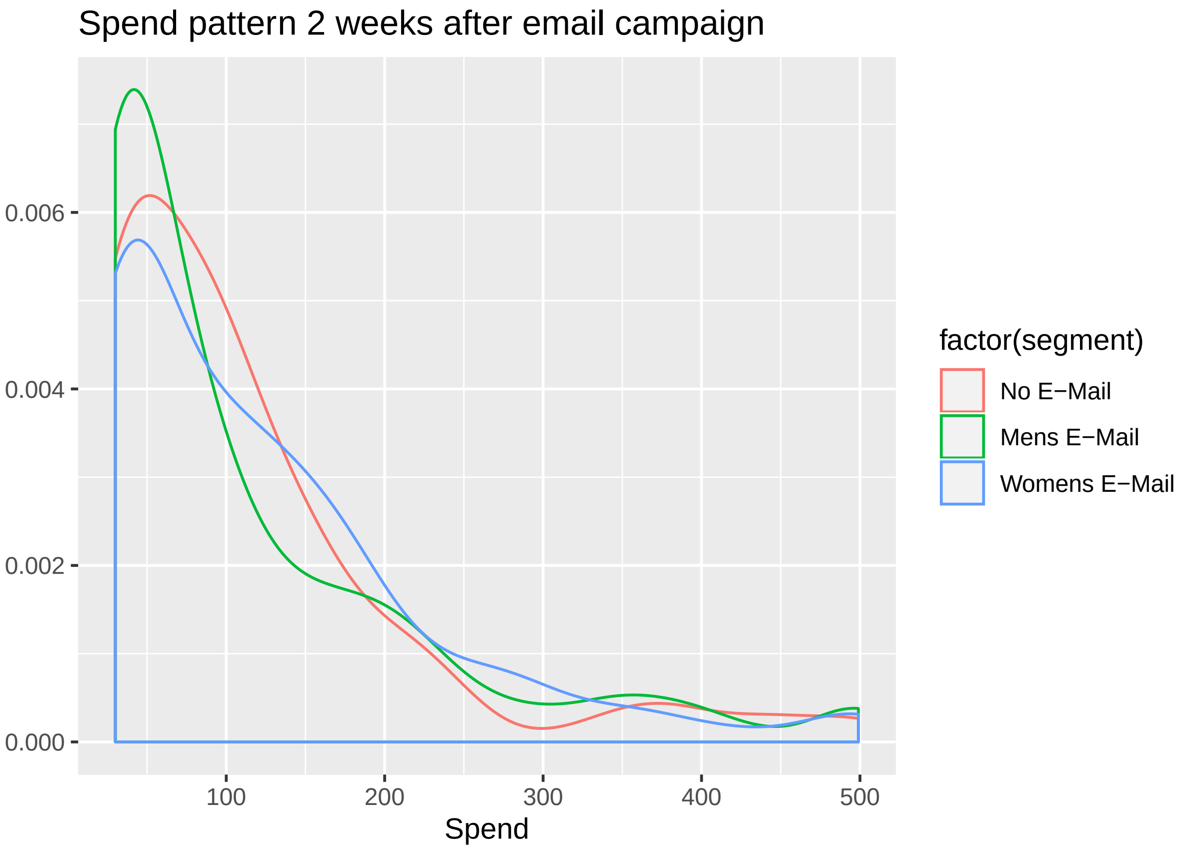
The following plot shows that 0.6% customers who did not get any email converted on the website, compared to 1.3% customers who received men's emails and 0.9% customers who received women's emails. Higher proportion of customers who received Men's or Women's emails purchased products from the website, however we need to evaluate whether this change is significantly different from customers who did not receive any emails.



Results from the logistic regression in section M1.2 (Appendix) confirm that there is a difference in purchase behavior of customers who did not receive any emails vs. customers who received either Men's or Women's emails. Further, we see that log(odds) of visiting the website improve by 0.79 for customers receiving men's emails and 0.43 for customers receiving women's emails. So, overall Men's campaign is more successful in increasing purchase on the website.

Overall impact of campaigns on spending:

The following plot shows that spending on the website is right skewed. Most of the customers (in all categories) who purchase, spend less than $150 and very few customers spend more $300. Customers who received men's emails spent less, customer's who received women's emails had a more balanced distribution and customers who did not receive any emails were somewhere in between. However, the plot does not provide any conclusive evidence whether campaigns lead to more or less spending on the website. Next, we use statistical models to help confirm this. This plot also indicates that while running the statistical tests, we should use logarithmic transformation of the outcome variable to counter the right-skew.



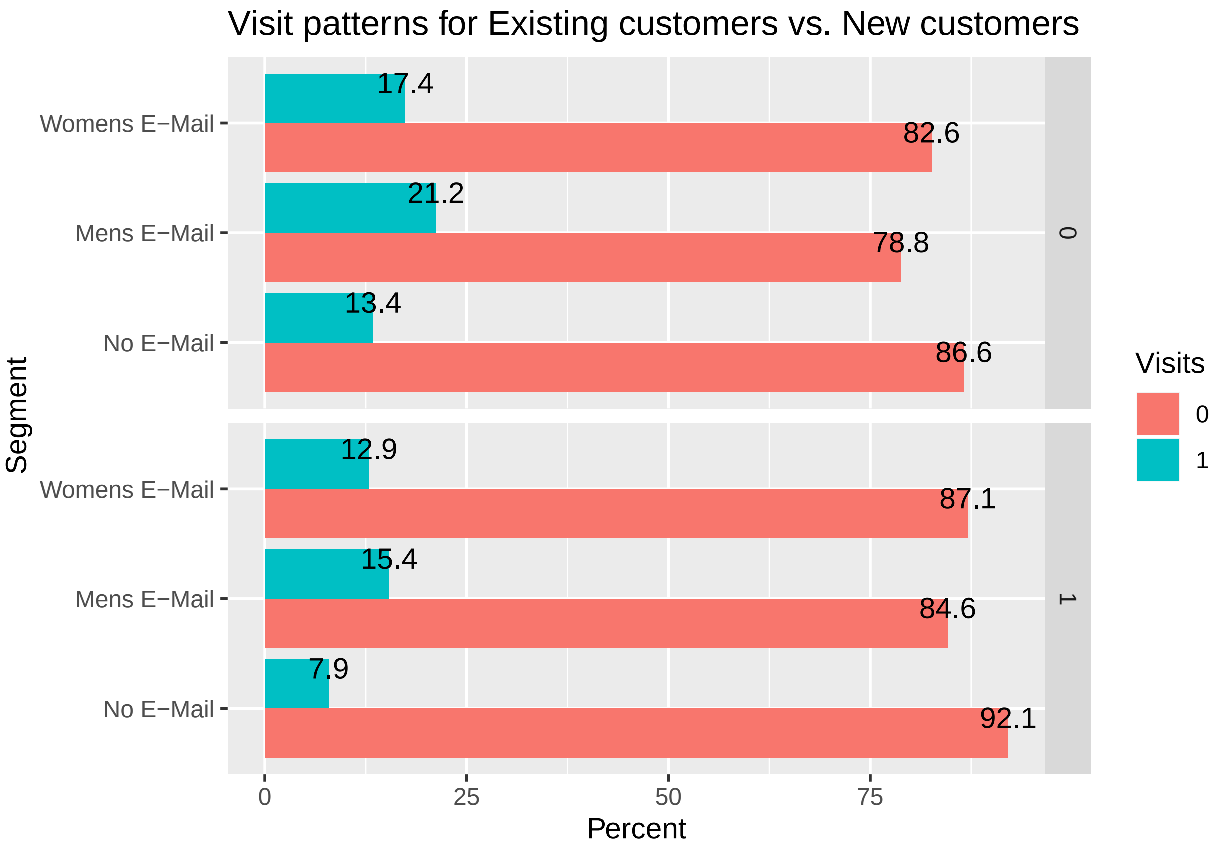
Results from the following statistical test in M1.3 (Appendix) show that there is no significant difference in spending of customers on the website. So, it is safe to say that campaigns do not impact the amount spent on the website by a particular customer.

Impact of Campaigns on categories of customers

The team used logistic regression to identify whether Men's or Women's campaign affects one category of customers more than another category, for example, does the campaign increase likelihood of visits equally for new and existing customers or is the effect different.

Impact of campaigns on visits for existing and new customers:

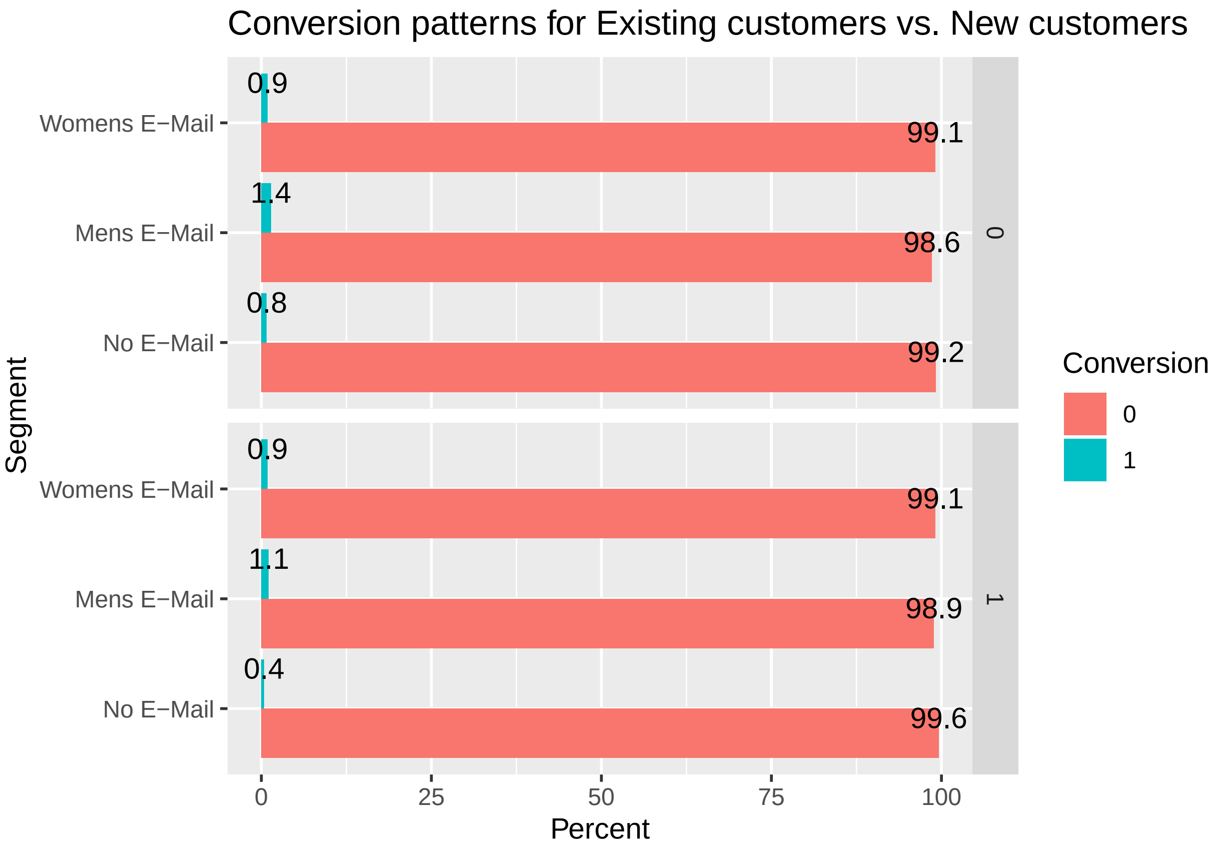
From the following plot, we can infer that existing customers who do not receive any emails are more likely to visit the website compared to new customers who do not receive any emails. Customers who receive men's and women's emails in both the groups are more likely to visit the website. In order to see if these changes are significant we use statistical tests.



Results of the test in section M2.1 (Appendix) confirms the following items:
– New customers are significantly less likely to visit the website compared to existing customers
– Sending men's and women's email to existing customers has a significant positive impact on visits
– Effect of men's and women's email on new customers is significantly higher than existing customers
– For existing as well as new customers men's campaign increases likelihood of visits more compared to women's campaign

Impact of campaigns on conversions for existing and new customers:

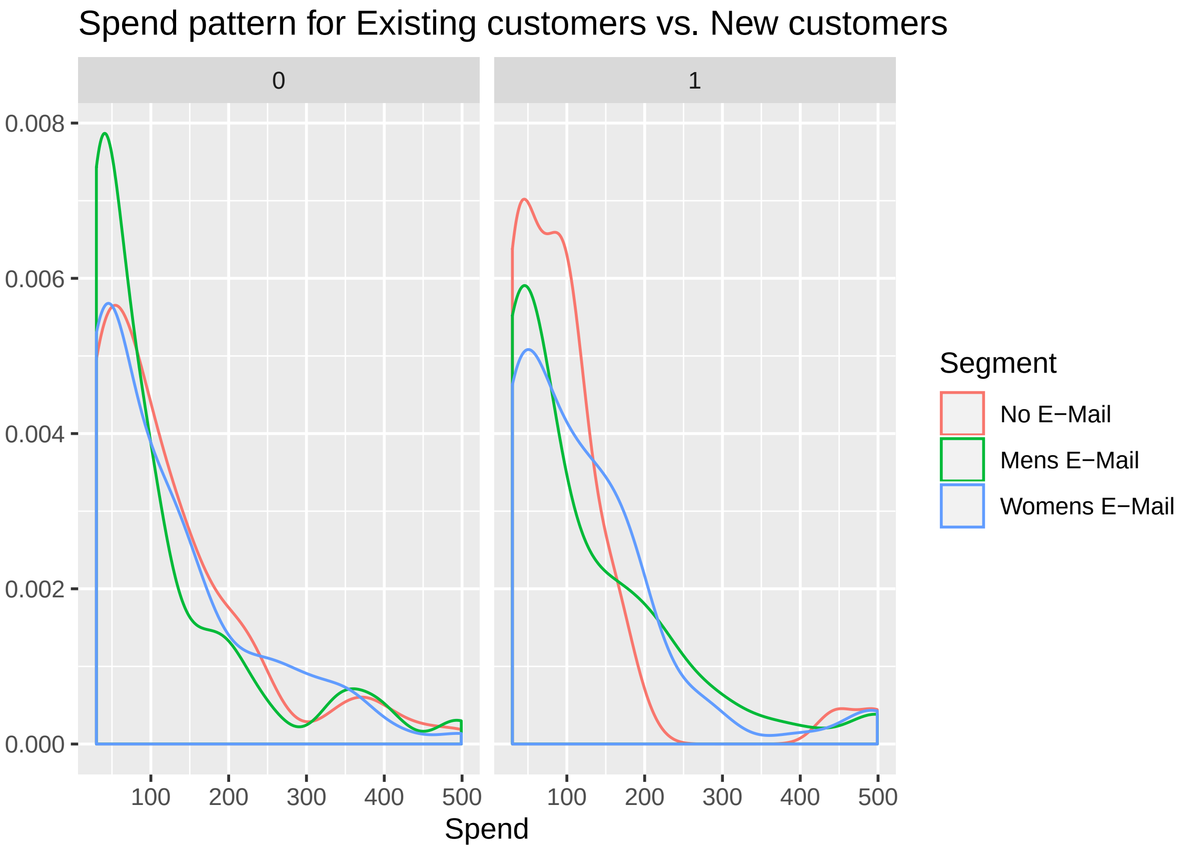
From the following plot, we can infer that existing customers who do not receive any emails are more likely to make purchases on the website compared to new customers who do not receive any emails. Customers who receive men's and women's emails in both groups are more likely to visit the website. In order to see if these changes are significant we use statistical tests.



Results of the test in section M2.2 (Appendix) confirm the following items:
– New customers are significantly less likely to make purchase compared to existing customers
– Sending men's to existing customers has a significant positive impact on conversions
– Effect of men's and women's email on new customers is significantly higher than existing customers
– For existing as well as new customers men's campaign increases likelihood of conversion more compared to women's campaign

Impact of campaigns on spending for existing and new customers:

For the plot, we see that new customers who make purchases and do not receive any email spend less compared to existing customers who do not receive any emails. Moreover, existing customers who receive men's emails spend less compared to existing customers who do not receive any emails. To check whether there is a significant difference in spending, we run some statistical tests.



Results from statistical tests in section M2.3 (Appendix):
– There is no significant difference in spending patterns of new and existing customers
– There is no significant change in customer spending behavior for men's and women's email campaign

Results and Conclusions

Impact of campaigns on other segments of customers: The team conducted similar statistical tests for other categories of customers. Here are the condensed results from the analysis:

Location:
– Rural customers are more likely to visit the website and make purchase on the site compared to suburban customers
– Men's and Women's campaigns increase likelihood of customers visiting the website significantly
– Men's and Women's campaigns do not affect the likelihood of purchase or amount spent on the website

Preferred channel:
– Customers with preferred channel as Phone have significantly less visits to website compared to Web and Multichannel customers
– Women's and Men's campaigns increase likelihood of purchase and visits to the website significantly. Men's campaign is more successful than women's campaign.
– There is no difference in effect of campaigns across customer segments

Past purchase - Men's products:
– Customers who purchased men's products in the past are more likely to visit the website
– Women's email has a more positive effect on visits and conversions customers who did not purchase men's products compared to customers who purchased men's products
– Men's email has a positive effect on visits for both the groups. The effect is higher for customers who purchased men's product in past
– Campaigns do not have significant effect on spending patterns in these segments

Past purchase - Women's products:
– Customers who purchased women's products in the past are more likely to visit the website
– Both campaigns have significantly positive impact on visits on the website – Men's campaign has a significant impact on the conversions but women's campaign does not
– Campaigns do not have significant effect on spending patterns in these segments

Recent purchase:
– Customer who have made recent purchases are more likely to visit the website
– Both campaigns have a positive impact on visits and conversions for all segments
– Effect of campaigns does not differ by segment
– Campaigns do not significantly impact on spending

Historical spend:
– Customers with high historical spend are more likely to visit the website compared to customers with low spend
– There is no significant difference in conversions and spend of customers in different historical spend segments
– Both campaigns have a positive impact on visits and conversion, however men's campaign is more effective for all segments
– No significant change in spending behavior for customers across history segments

Causal Forest

In the previous section, we have shown that the treatment effect of email campaigns varies for different sub groups. In the approach we take in the last section, customers in the same group have the same treatment effect, which is not accurate enough.

In this section, we are trying to predict the treatment effect of email campaigns for each individual customer. We apply a recently introduced technique, which is causal forest, to help us achieve it.

Train Model and Make Predictions

Description

First, we are going to build four different causal forest models for visits of men's email, conversion of men's email, visits of women's email, and conversion for women's email separately because conversion and visits are affected by campaign but spending is not.

We will split the dataset into training and testing dataset by a ratio of 80/20. After training the model, we will use our model to predict the testing dataset.

For the purpose of demonstration, we will illustrate the process of one model in this section, which is visits for men's email campaign .The detail of other three models will be in the appendix.

Visits for Men's Email

Create dataset

```
men_no<-rbind(men,no)
```

Split dataset

```
set.seed(666)
men_no$treat<-as.numeric(men_no$segment)*-1+2
cases <- sample(seq_len(nrow(men_no)), round(nrow(men_no) * .8))
train_men<-men_no[cases,]
test_men<-men_no[-cases,]
```

Train dataset

```
X = model.matrix(~ ., data = train_men[, 1:8])
Y = train_men$visit
W = train_men$treat
Y.hat <- predict(regression_forest(X, Y))$predictions
W.hat <- rep(0:1, length.out=length(Y))
params <- tune_causal_forest(X, Y, W, Y.hat, W.hat)$params
# Use these parameters to train a regression forest.
tuned.forest_visit_men <- causal_forest(X, Y, W,
  Y.hat = Y.hat, W.hat = W.hat, num.trees = 5000,
  min.node.size = as.numeric(params["min.node.size"]),
  sample.fraction = as.numeric(params["sample.fraction"]),
  mtry = as.numeric(params["mtry"]),
  alpha = as.numeric(params["alpha"]),
  imbalance.penalty = as.numeric(params["imbalance.penalty"]))
```

Make predictions on testing dataset

```
preds_men_visit_tune<-predict(
  object = tuned.forest_visit_men,
  newdata = model.matrix(~ ., data = test_men[, 1:8]),
  estimate.variance = TRUE
)
```

Merge predicted value into testing dataset

```
test_men$preds_men_visit_tune<-preds_men_visit_tune$predictions
```

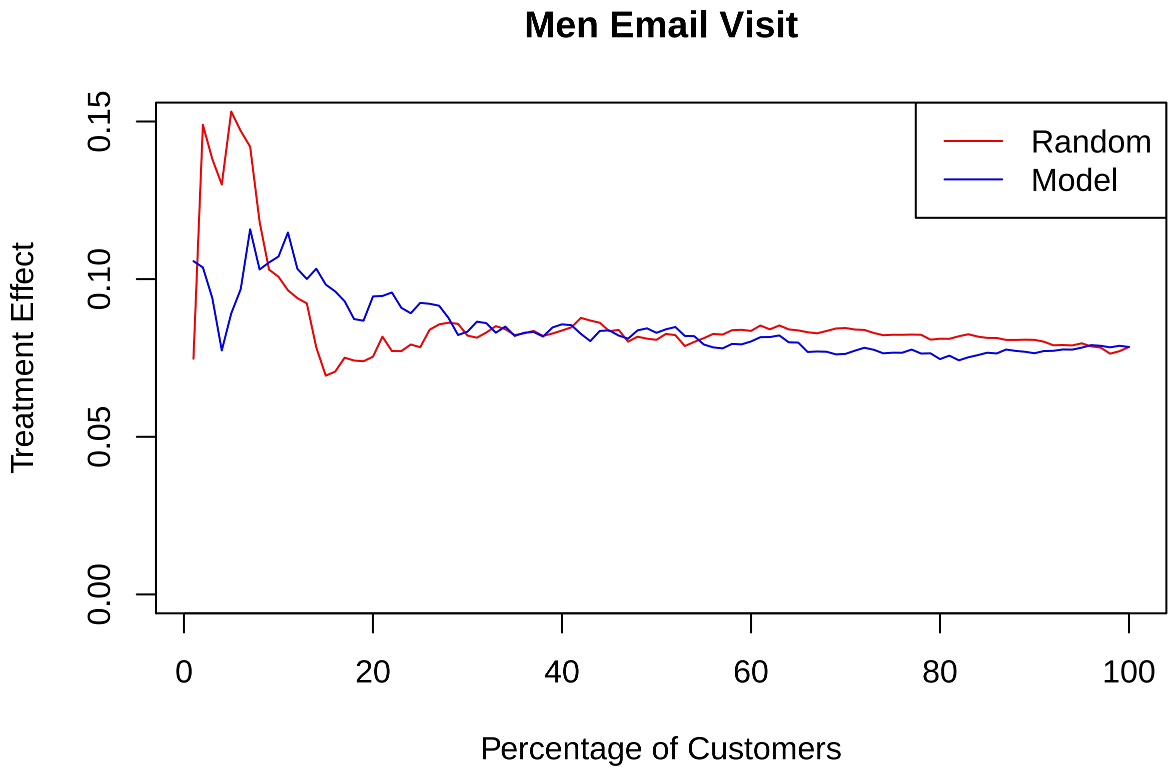
Evaluate Performance

In this section, we will use plots to visualize the performance of causal forest models we build. Each graph has two lines, one representing the model performance and the other representing that of random selection. X-axis is the percentage of customers and y-axis is the treatment effect. Hence, if a point is at (20,0.04), it means if we choose 20% of customers from our testing dataset, the treatment effect will be 0.04 for them. If the model line is above the random line, it means we will achieve a better treatment effect if we select customers by random forest model instead of random selection. The gap between two lines is the improvement of our model from random selection

Visits for Men's Email

```
test_order1<-test_men[order(-test_men$preds_men_visit_tune),]
```

```
set.seed(123)
rows<-sample(nrow(test_men))
test_men<-test_men[rows,]
x<-seq(0.01,1,0.01)
model=c()
random=c()
for (i in x){
  t1<-lm(visit~treat, head(test_order1,round(i*length(test_men$recency))))
  a<-summary(t1)$coef[2]
  t3<-lm(visit~treat, head(test_men,round(i*length(test_men$recency))))
  b<-summary(t3)$coef[2]
  model<-append(model,a)
  random<-append(random,b)}
plot(random,type='l',ylim=c(0,0.15),xlab = 'Percentage of Customers',
     ylab='Treatment Effect',col='red',main='Men Email Visit')
lines(model,col='blue')
legend("topright", legend=c("Random", "Model"),
       col=c("red", "blue"), lty=1, cex=1)
```



```
men_visit_matrix<-data.frame(model=model,random=random)
write.csv(men_visit_matrix,'men_visit_matrix.csv')
```

Conclusion

As we can see from the above graphs, most of the time, for both visits and conversions, our model has higher treatment effects than random selection.

Identify Important Factors

Because causal forest is an expanded version of the random selection model, we can also study factors that are important when we build the model. In this section we will show the top three most important factors for each model.

Visits for Men's Email

```
importance_visit_men<-tuned.forest_visit_men %>%
  variable_importance() %>%
  as.data.frame() %>%
  mutate(variable = colnames(tuned.forest_visit_men$X.orig)) %>%
  arrange(desc(V1))
head(importance_visit_men,3)
```

```
##            V1 variable
## 1 0.44655052  history
## 2 0.09198606  recency
## 3 0.06857143   newbie
```

Conclusion

As we can see from above, history and recency are the two factors that are important. They are also very important in all other three models. "History" represents the spending history of a customer and "recency" represents the past purchase of a customer in the past 12 months. This means that a customer's past history is highly related to the treatment effects of the email campaigns. Causal forest model cannot identify a clear relationship there. Hence, it would be a great idea to invest in the Customer Relationship Management(CRM) department to figure out the reason behind it. Also, it is important that we can develop and build loyal customer programs in the future.

Final Recommendation

a. Implement causal forest model for future email campaign.As the causal forest model performs better than random selection, in the future, if we want to apply email campaigns to customers, we can use the causal forest model to help us choose the most valuable customers, and hence maximize the effects of email campaigns.

b. Invest on CRM and develop loyal customers. As mentioned above, history and recency are the most important factors for visits and conversions for both email campaigns. Investing in CRM can help attract more customers to visit and purchase on the website.

c. Both campaigns are successful in increasing visits on the website and increasing overall conversions on the website. These campaigns should be rolled out to the entire customer base. However, these campaigns do not affect customer spending behaviors on the website, we might need to tweak the campaigns or resort to some other methods to increase spending on the website. Among the two campaigns, Men's emails are more successful in increasing visits and conversions.

Limitations

1. The lack of demographic data of customers. If there is more demographic data of customers, we are able to do more heterogeneity analysis and understand the effects of campaigns across different segments of customers better. The company can collect customer demographic data by offering customers incentives to fill in some of their personal information when they sign up.
2. Sample selection bias. In the experiment, we are not clear how samples were selected. It is likely that customers are not representative of the whole customers base. The company can consider using stratified sampling to ensure the sample is representative.
3. Interference. Communication between each treatment group may affect the original action of customers. The company can try sampling customers from different regions to decrease the interference between treatment and control groups.

Appendix

Randomization Check:

```r
summary(aov(womens ~ segment, data = data))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## segment         2      0 0.07828   0.316  0.729
## Residuals   63997  15842 0.24754
```

```
summary(aov(recency ~ segment, data = data))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## segment         2      7   3.327    0.27  0.763
## Residuals   63997 787386  12.303
```

```
summary(aov(history ~ segment, data = data))
```

```
##                Df    Sum Sq Mean Sq F value Pr(>F)
## segment         2 4.718e+04   23589   0.359  0.698
## Residuals   63997 4.199e+09   65619
```

```
rand5 <- aov(newbie ~ segment, data = data)

tbl <- table(data$channel, data$segment)

chisq.test(tbl)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tbl
## X-squared = 3.6324, df = 4, p-value = 0.458
```

Statistical testing: Evaluate success of women's and men's campaigns

M1.1 Visits - Overall

```
model_visit = glm(visit ~ factor(segment), data = data, family = "binomial")
summary(model_visit)
```

```
## 
## Call:
## glm(formula = visit ~ factor(segment), family = "binomial", data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6353  -0.6353  -0.5730  -0.4738   2.1179  
## 
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                  -2.13050    0.02224  -95.80   <2e-16 ***
## factor(segment)Mens E-Mail    0.63272    0.02844   22.25   <2e-16 ***
## factor(segment)Womens E-Mail  0.40684    0.02930   13.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 53387  on 63999  degrees of freedom
## Residual deviance: 52871  on 63997  degrees of freedom
## AIC: 52877
## 
## Number of Fisher Scoring iterations: 4
```

M1.2 Conversion - Overall

```
## 
## Call:
## glm(formula = conversion ~ factor(segment), family = "binomial", 
##     data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.1588  -0.1588  -0.1332  -0.1072   3.2133  
## 
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                  -5.1570     0.0908 -56.797  < 2e-16 ***
## factor(segment)Mens E-Mail     0.7901     0.1097   7.201 5.97e-13 ***
## factor(segment)Womens E-Mail   0.4371     0.1165   3.750 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 6592.1  on 63999  degrees of freedom
## Residual deviance: 6536.1  on 63997  degrees of freedom
## AIC: 6542.1
## 
## Number of Fisher Scoring iterations: 8
```

M1.3 Spend - Overall

```
## 
## Call:
## lm(formula = log(spend + 1) ~ factor(segment), data = data %>% 
##     filter(spend > 0))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.04390 -0.91949 -0.01909  0.62334  1.85735 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  4.42881    0.07490  59.133   <2e-16 ***
## factor(segment)Mens E-Mail   -0.07155    0.09040  -0.791    0.429    
## factor(segment)Womens E-Mail  0.04875    0.09607   0.507    0.612    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8273 on 575 degrees of freedom
## Multiple R-squared:  0.004168,   Adjusted R-squared:  0.0007042 
## F-statistic: 1.203 on 2 and 575 DF,  p-value: 0.301
```

M2.1 Visits - Newbie

```
## 
## Call:
## glm(formula = visit ~ factor(segment) * factor(newbie), family = "binomial", 
##     data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6895  -0.6180  -0.5358  -0.4052   2.2541  
## 
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  -1.86838    0.02852 -65.506
## factor(segment)Mens E-Mail                    0.55285    0.03712  14.893
## factor(segment)Womens E-Mail                  0.30981    0.03833   8.084
## factor(newbie)1                              -0.59008    0.04584 -12.873
## factor(segment)Mens E-Mail:factor(newbie)1    0.20302    0.05817   3.490
## factor(segment)Womens E-Mail:factor(newbie)1  0.24095    0.05985   4.026
##                                              Pr(>|z|)    
## (Intercept)                                   < 2e-16 ***
## factor(segment)Mens E-Mail                    < 2e-16 ***
## factor(segment)Womens E-Mail                 6.29e-16 ***
## factor(newbie)1                               < 2e-16 ***
## factor(segment)Mens E-Mail:factor(newbie)1   0.000482 ***
## factor(segment)Womens E-Mail:factor(newbie)1 5.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 53387  on 63999  degrees of freedom
## Residual deviance: 52499  on 63994  degrees of freedom
## AIC: 52511
## 
## Number of Fisher Scoring iterations: 5
```

M2.2 Conversion - Newbie

```
## 
## Call:
## glm(formula = conversion ~ factor(segment) * factor(newbie), 
##     family = "binomial", data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.1675  -0.1497  -0.1319  -0.1246   3.3432  
## 
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                   -4.8552     0.1109 -43.795
## factor(segment)Mens E-Mail                     0.5958     0.1384   4.306
## factor(segment)Womens E-Mail                   0.1148     0.1525   0.753
## factor(newbie)1                               -0.7297     0.1933  -3.774
## factor(segment)Mens E-Mail:factor(newbie)1     0.5027     0.2296   2.189
## factor(segment)Womens E-Mail:factor(newbie)1   0.7700     0.2424   3.177
##                                              Pr(>|z|)    
## (Intercept)                                   < 2e-16 ***
## factor(segment)Mens E-Mail                   1.66e-05 ***
## factor(segment)Womens E-Mail                0.451623
## factor(newbie)1                             0.000161 ***
## factor(segment)Mens E-Mail:factor(newbie)1   0.028586 *
## factor(segment)Womens E-Mail:factor(newbie)1 0.001489 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6592.1  on 63999  degrees of freedom
## Residual deviance: 6517.5  on 63994  degrees of freedom
## AIC: 6529.5
##
## Number of Fisher Scoring iterations: 8
```

M2.3 Spend - Newbie

```
##
## Call:
## lm(formula = log(spend + 1) ~ factor(segment) * factor(newbie),
##     data = data %>% filter(spend > 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06343 -0.88067 -0.01494  0.61085  1.90028
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   4.47738    0.09145  48.962
## factor(segment)Mens E-Mail                   -0.16305    0.11400  -1.430
## factor(segment)Womens E-Mail                 -0.02040    0.12576  -0.162
## factor(newbie)1                              -0.14813    0.15970  -0.928
## factor(segment)Mens E-Mail:factor(newbie)1    0.24445    0.18947   1.290
## factor(segment)Womens E-Mail:factor(newbie)1  0.18824    0.20007   0.941
##                                              Pr(>|t|)
## (Intercept)                                    <2e-16 ***
## factor(segment)Mens E-Mail                      0.153
## factor(segment)Womens E-Mail                    0.871
## factor(newbie)1                                 0.354
## factor(segment)Mens E-Mail:factor(newbie)1      0.198
## factor(segment)Womens E-Mail:factor(newbie)1    0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8281 on 572 degrees of freedom
## Multiple R-squared:  0.007402,   Adjusted R-squared:  -0.001275
## F-statistic: 0.8531 on 5 and 572 DF,  p-value: 0.5125
```

Causal Forest model training on conversion of both campaigns and visit of women campaign

Conversion for Men's Email

```
# test_order2<-test_men[order(-test_men$preds_men_conversion_tune),]
```

```
# rows<-sample(nrow(test_men))
# test_men<-test_men[rows,]
# x<-seq(0.01,1,0.01)
# model=c()
# random=c()
# for (i in x){

# t1<-lm(conversion~treat, head(test_order2,round(i*length(test_men$recency))))
# a<-summary(t1)$coef[2]

# t3<-lm(conversion~treat, head(test_men,round(i*length(test_men$recency))))
# b<-summary(t3)$coef[2]

#  model<-append(model,a)
#  random<-append(random,b)
#}

#plot(random,type='l',ylim=c(-0.05,0.05),xlab = 'Percentage of Customers',ylab='Treatment #Effect',col=
#lines(model,col='blue')
#legend("topright", legend=c("Random", "Model"),col=c("red", "blue"), lty=1, cex=1)
```

Visits for Women's Email

```
#test_order3<-test[order(-test$visit_preds),]
```

```
#rows<-sample(nrow(test))
#test<-test[rows,]
#x<-seq(0.01,1,0.01)
#model=c()
#random=c()
#for (i in x){

#  t1<-lm(visit~treat, head(test_order3,round(i*length(test$recency))))
#  a<-summary(t1)$coef[2]

#  t3<-lm(visit~treat, head(test,round(i*length(test$recency))))
#  b<-summary(t3)$coef[2]

#  model<-append(model,a)
#  random<-append(random,b)
#}

#plot(random,type='l',ylim=c(0,0.15),xlab = 'Percentage of Customers',ylab='Treatment #Effect',col='red
#lines(model,col='blue')
#legend("topright", legend=c("Random", "Model"),col=c("red", "blue"), lty=1, cex=1)
```

Conversion for Women's Email

```
#test_order4<-test[order(-test$conv_preds),]
```

```
#rows<-sample(nrow(test))
#test<-test[rows,]
#x<-seq(0.01,1,0.01)
#model=c()
#random=c()
#for (i in x){

#  t1<-lm(conversion~treat, head(test_order4,round(i*length(test$recency))))
#  a<-summary(t1)$coef[2]


#  t3<-lm(conversion~treat, head(test,round(i*length(test$recency))))
#  b<-summary(t3)$coef[2]




#  model<-append(model,a)
#  random<-append(random,b)
# }

# plot(random,type='l',ylim=c(-0.05,0.05),xlab = 'Percentage of Customers',ylab='Treatment
# Effect',col='red',main='Women Email Conversion')
# lines(model,col='blue')
# legend("topright", legend=c("Random", "Model"),col=c("red", "blue"), lty=1, cex=1)
```

```
# women_conv_matrix<-data.frame(model=model,random=random)
# write.csv(women_conv_matrix,'women_conv_matrix.csv')
```

Feature Importance

Conversion for Men's Email

```
# importance_conv_men<-tuned.forest_conv_men %>%
#  variable_importance() %>%
#  as.data.frame() %>%
#  mutate(variable = colnames(tuned.forest_conv_men$X.orig)) %>%
#  arrange(desc(V1))
# head(importance_conv_men,3)
```

Visits for Women's Email

```
# importance_visit_women<-tuned.forest_visit_women %>%
# variable_importance() %>%
#  as.data.frame() %>%
#  mutate(variable = colnames(tuned.forest_visit_women$X.orig)) %>%
#  arrange(desc(V1))
# head(importance_visit_women,3)
```

Conversion for Women's Email

```
# importance_conv_women<-tuned.forest_conv_women %>%
#  variable_importance() %>%
#  as.data.frame() %>%
# mutate(variable = colnames(tuned.forest_conv_women$X.orig)) %>%
#  arrange(desc(V1))
# head(importance_conv_women,3)
```