Влияние типа прихода клиента на LTV

Ihar Alishkevich

2022-06-13

Table of Contents

## Введение

Проводится исследование на основе реальных данных образовательной школы **Itgenio**. Используются реальные данные *LTV*, полученные путем когортного анализа[[1]](#footnote-20) и разбитые на 3 группы по типу прихода:

1. таргетинговая кампания[[2]](#footnote-22);
2. реферал[[3]](#footnote-24);
3. остальные.

В связи с когортным отображением, из чего следует большое число колонок, данные удобнее просмотреть в браузере[[4]](#footnote-25).

ltv\_target\_df <- read\_sheet(  
 "https://docs.google.com/spreadsheets/d/1LTSw4nZUKFMoh0HJ88Gs36aJQaaxq4GkPCu7drYTj4E/edit?usp=sharing",   
 sheet = "LTV target"  
)

## v Reading from "Влияние типа прихода клиента на LTV".

## v Range ''LTV target''.

ltv\_referral\_df <- read\_sheet(  
 "https://docs.google.com/spreadsheets/d/1LTSw4nZUKFMoh0HJ88Gs36aJQaaxq4GkPCu7drYTj4E/edit?usp=sharing",   
 sheet = "LTV referral"  
)

## v Reading from "Влияние типа прихода клиента на LTV".

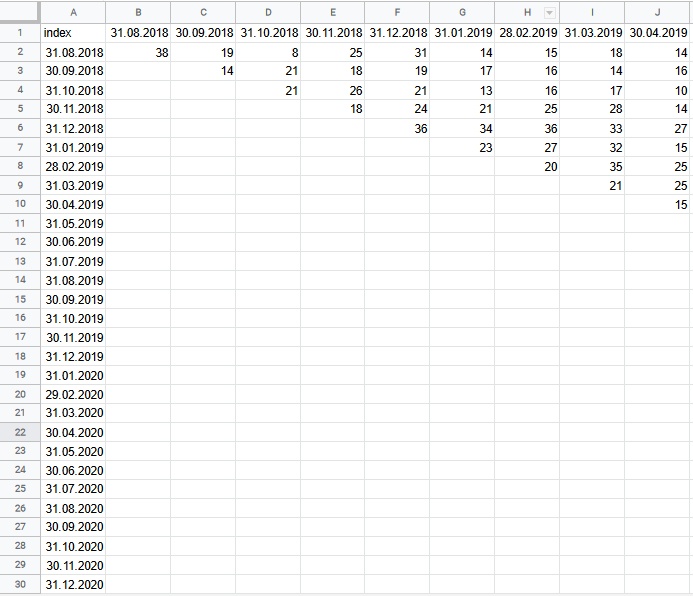
## v Range ''LTV referral''.

ltv\_unknown\_df <- read\_sheet(  
 "https://docs.google.com/spreadsheets/d/1LTSw4nZUKFMoh0HJ88Gs36aJQaaxq4GkPCu7drYTj4E/edit?usp=sharing",   
 sheet = "LTV unknown"  
)

## v Reading from "Влияние типа прихода клиента на LTV".

## v Range ''LTV unknown''.

# Если не удается считать данные через google, можно воспользоваться привычной загрузкой csv из github. Например:  
# ltv\_unknown\_df <- read.csv("https://raw.githubusercontent.com/Delictum/Rmarkdown/main/%D0%92%D0%BB%D0%B8%D1%8F%D0%BD%D0%B8%D0%B5%20%D1%82%D0%B8%D0%BF%D0%B0%20%D0%BF%D1%80%D0%B8%D1%85%D0%BE%D0%B4%D0%B0%20%D0%BA%D0%BB%D0%B8%D0%B5%D0%BD%D1%82%D0%B0%20%D0%BD%D0%B0%20LTV%20-%20LTV%20referral.csv")

Данные содержат наблюдения за 29 месяцев. Первое значение каждой строки (столбец index) отображают, когда клиент впервые завершил оплаченное занятие. На основе этого попадает в соответствующую когорту. 

Остальные столбцы отображают LTV когорт в данном месяце. Для примера когорта клиентов от 2018-10-31 на дату 31.10.2018 принесли прибыли в 21 условных единиц. Используемая версия R: 4.1.2.

## Гипотеза

Выдвигается предположение, что клиенты, пришедшие благодаря усилиям таргетинговых кампаний, приносят наибольший LTV в сфере образования.

## Вычисления

### Предварительная обработка данных

Для примера возьмем данные за дату 2018-12-31.

num\_row = 5  
  
ltv\_u\_df <- t(ltv\_unknown\_df[num\_row, -1])  
ltv\_r\_df <- t(ltv\_referral\_df[num\_row, -1])  
ltv\_t\_df <- t(ltv\_target\_df[num\_row, -1])  
head(ltv\_u\_df)

## [,1]  
## 31.08.2018 NA  
## 30.09.2018 NA  
## 31.10.2018 NA  
## 30.11.2018 NA  
## 31.12.2018 21  
## 31.01.2019 24

И объеденим в единый data.frame, предварительно избавившись от пустых значений. В том числе добавим столбец с типом factor для идентификации полученного LTV по типу прихода клиента.

u\_df <- data.frame(  
 ltv = c(ltv\_u\_df[!is.na(ltv\_u\_df[,1]),], use.names = F),  
 type = "unknown",  
 date = names(c(ltv\_u\_df[!is.na(ltv\_u\_df[,1]),], use.names = T))  
)  
t\_df <- data.frame(  
 ltv = c(ltv\_t\_df[!is.na(ltv\_t\_df[,1]),], use.names = F),  
 type = "target",  
 date = names(c(ltv\_t\_df[!is.na(ltv\_t\_df[,1]),], use.names = T))  
)  
r\_df <- data.frame(  
 ltv = c(ltv\_r\_df[!is.na(ltv\_r\_df[,1]),], use.names = F),  
 type = "referral",  
 date = names(c(ltv\_r\_df[!is.na(ltv\_r\_df[,1]),], use.names = T))  
)  
res\_df <- rbind(u\_df, t\_df, r\_df)  
res\_df$type <- as.factor(res\_df$type)  
str(res\_df)

## 'data.frame': 75 obs. of 3 variables:  
## $ ltv : num 21 24 19 20 18 13 10 7 6 9 ...  
## $ type: Factor w/ 3 levels "referral","target",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ date: chr "31.12.2018" "31.01.2019" "28.02.2019" "31.03.2019" ...

### Осмотр данных

Взглянем на несколько покзателей из каждой группы:

res\_df %>%  
 group\_by(type) %>%  
 summarise(  
 count\_type = n(),  
 mean\_ltv = mean(ltv, na.rm = TRUE),  
 median\_ltv = median(ltv, na.rm = TRUE),  
 sd\_ltv = sd(ltv, na.rm = TRUE)  
 )

## # A tibble: 3 x 5  
## type count\_type mean\_ltv median\_ltv sd\_ltv  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 referral 25 11.2 12 9.94  
## 2 target 25 11.1 6 12.1   
## 3 unknown 25 9.64 7 5.94

Для полноты картины провизуализируем данные:

res\_df %>%   
 ggvis(~date, ~ltv, stroke = ~type) %>%  
 group\_by(type) %>%  
 layer\_paths() %>%  
 layer\_points(fill := "white") %>%  
 add\_axis("x", title = "", properties = axis\_props(  
 axis = list(stroke = "red", strokeWidth = 3),  
 grid = list(stroke = "lightblue"),  
 ticks = list(stroke = "blue", strokeWidth = 1),  
 labels = list(angle = 45, align = "left", fontSize = 12)  
 )) %>%  
 add\_axis("y", title = "LTV") %>%  
 add\_legend("stroke", title = "Тип прихода",  
 properties = legend\_props(  
 title = list(fontSize = 16),  
 labels = list(fontSize = 12, fill = "#00F"),  
 gradient = list(stroke = "red", strokeWidth = 2),  
 legend = list(x = 500, y = 50)  
 )  
 )

А также через ящик с усами:

res\_df %>%  
 ggvis(~type, ~ltv) %>%  
 layer\_boxplots() %>%  
 add\_axis("x", title = "Тип прихода клиента") %>%  
 add\_axis("y", title = "LTV")

### Проверка гипотезы

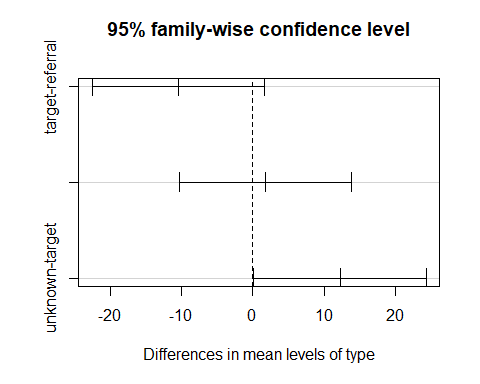
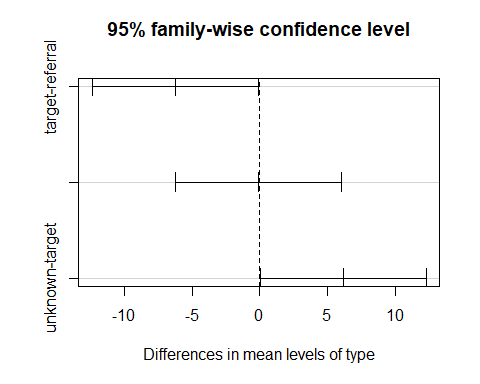
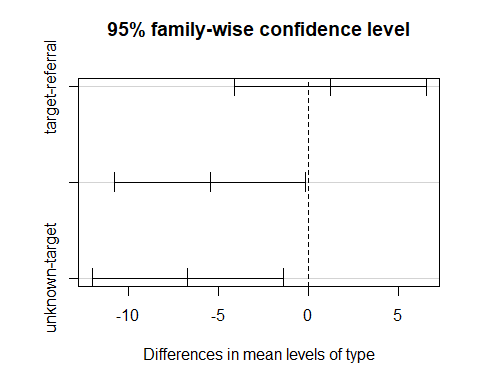
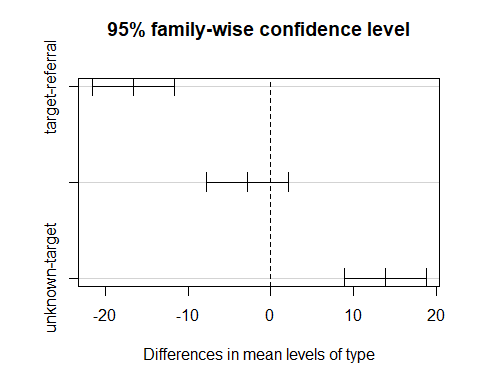
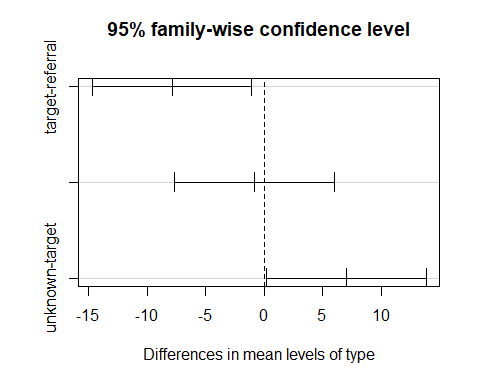
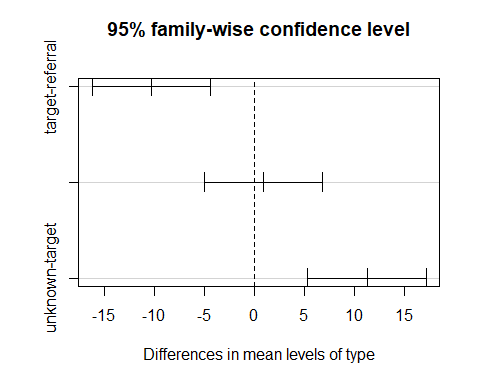
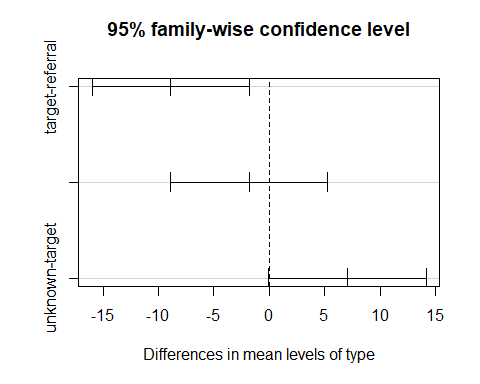
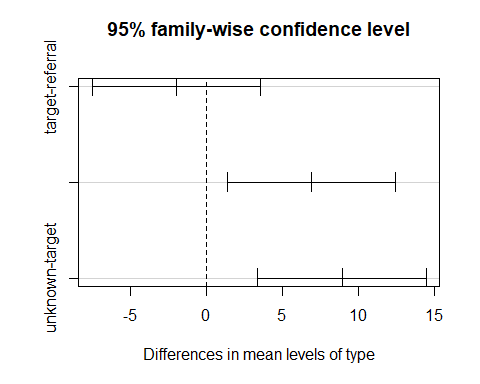
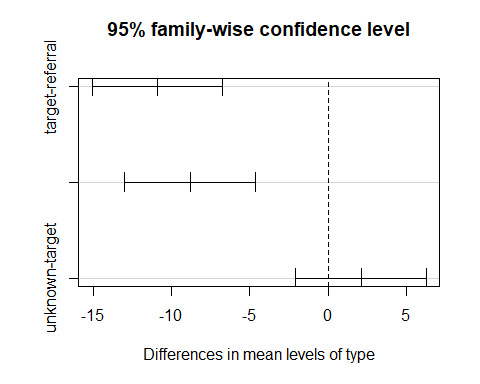
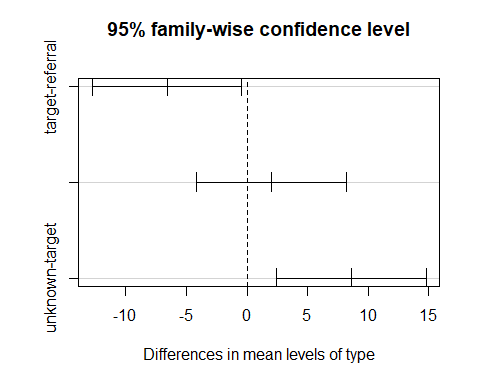
Проверим есть ли статистически значимая разница при помощи дисперсионного анализа:

anova\_one\_way <- aov(ltv ~ type, data = res\_df)  
summary(anova\_one\_way)

## Df Sum Sq Mean Sq F value Pr(>F)  
## type 2 40 19.85 0.213 0.809  
## Residuals 72 6711 93.21

Как видно из результата, p-value значительно выше 0.05. Исследуем оставшиеся когорты. Если значений для одной из когорт не окажется, пропустим их, выведя соотвествующее сообщение. Применив дисперсионный анализ и если найдена статистическая разница, найдем где она находится при помощи критерия Тьюки [[5]](#footnote-34) (и выведем графики для наглядности):

num\_row = 5  
  
get\_ltv\_df <- function(num\_row) {  
 current\_cohort <- names(ltv\_unknown\_df[, num\_row])  
 ltv\_u\_df <- t(ltv\_unknown\_df[num\_row, -1])  
 ltv\_r\_df <- t(ltv\_referral\_df[num\_row, -1])  
 ltv\_t\_df <- t(ltv\_target\_df[num\_row, -1])  
   
 u\_df <- data.frame(  
 ltv = c(ltv\_u\_df[!is.na(ltv\_u\_df[,1]),], use.names = F),  
 type = "unknown",  
 date = names(c(ltv\_u\_df[!is.na(ltv\_u\_df[,1]),], use.names = T))  
 )  
 t\_df <- data.frame(  
 ltv = c(ltv\_t\_df[!is.na(ltv\_t\_df[,1]),], use.names = F),  
 type = "target",  
 date = names(c(ltv\_t\_df[!is.na(ltv\_t\_df[,1]),], use.names = T))  
 )  
 r\_df <- data.frame(  
 ltv = c(ltv\_r\_df[!is.na(ltv\_r\_df[,1]),], use.names = F),  
 type = "referral",  
 date = names(c(ltv\_r\_df[!is.na(ltv\_r\_df[,1]),], use.names = T))  
 )  
   
 if ((u\_df$ltv == rep(0, nrow(u\_df))) | (t\_df$ltv == rep(0, nrow(t\_df))) | (r\_df$ltv == rep(0, nrow(r\_df)))) {  
 return(c(current\_cohort, "Недостаточно данных"))  
 }  
   
 res\_df <- rbind(u\_df, t\_df, r\_df)  
 res\_df$type <- as.factor(res\_df$type)  
   
 anova\_one\_way <- aov(ltv ~ type, data = res\_df)  
 p <- summary(anova\_one\_way)[[1]][["Pr(>F)"]][1]  
 if (p > 0.05) {  
 return(c(current\_cohort, p, "aov > 0.05"))  
 }  
   
 tukey\_result <- TukeyHSD(anova\_one\_way)  
   
 # Согласно документации:  
 # There are print and plot methods for class "TukeyHSD". The plot method does not accept xlab, ylab or main arguments and creates its own values for each plot.  
 base::plot(tukey\_result)   
 return(c(current\_cohort, tukey\_result))  
}  
  
lapply(c(2:(ncol(ltv\_unknown\_df)-2)), FUN = function(x) get\_ltv\_df(x))



## [[1]]  
## [1] "31.08.2018" "Недостаточно данных"  
##   
## [[2]]  
## [1] "30.09.2018" "Недостаточно данных"  
##   
## [[3]]  
## [1] "31.10.2018" "Недостаточно данных"  
##   
## [[4]]  
## [1] "30.11.2018" "0.808662874783857" "aov > 0.05"   
##   
## [[5]]  
## [1] "31.12.2018" "0.0688831097983599" "aov > 0.05"   
##   
## [[6]]  
## [[6]][[1]]  
## [1] "31.01.2019"  
##   
## [[6]]$type  
## diff lwr upr p adj  
## target-referral -6.608696 -12.792778 -0.4246134 0.033508519  
## unknown-referral 2.000000 -4.184082 8.1840822 0.719323151  
## unknown-target 8.608696 2.424613 14.7927779 0.003925894  
##   
##   
## [[7]]  
## [[7]][[1]]  
## [1] "28.02.2019"  
##   
## [[7]]$type  
## diff lwr upr p adj  
## target-referral -10.909091 -15.067577 -6.750605 9.853431e-08  
## unknown-referral -8.818182 -12.976668 -4.659696 1.029589e-05  
## unknown-target 2.090909 -2.067577 6.249395 4.535636e-01  
##   
##   
## [[8]]  
## [1] "31.03.2019" "0.838012215785422" "aov > 0.05"   
##   
## [[9]]  
## [[9]][[1]]  
## [1] "30.04.2019"  
##   
## [[9]]$type  
## diff lwr upr p adj  
## target-referral -2.0 -7.532283 3.532283 0.6612864574  
## unknown-referral 6.9 1.367717 12.432283 0.0109494078  
## unknown-target 8.9 3.367717 14.432283 0.0008106627  
##   
##   
## [[10]]  
## [1] "31.05.2019" "0.124346722495526" "aov > 0.05"   
##   
## [[11]]  
## [[11]][[1]]  
## [1] "30.06.2019"  
##   
## [[11]]$type  
## diff lwr upr p adj  
## target-referral -8.888889 -15.99093492 -1.786843 0.01078170  
## unknown-referral -1.833333 -8.93537936 5.268713 0.80814288  
## unknown-target 7.055556 -0.04649047 14.157602 0.05186335  
##   
##   
## [[12]]  
## [[12]][[1]]  
## [1] "31.07.2019"  
##   
## [[12]]$type  
## diff lwr upr p adj  
## target-referral -10.3529412 -16.277286 -4.428596 3.067396e-04  
## unknown-referral 0.8823529 -5.041992 6.806698 9.310841e-01  
## unknown-target 11.2352941 5.310949 17.159639 9.489599e-05  
##   
##   
## [[13]]  
## [[13]][[1]]  
## [1] "31.08.2019"  
##   
## [[13]]$type  
## diff lwr upr p adj  
## target-referral -7.875 -14.696012 -1.053988 0.02022151  
## unknown-referral -0.875 -7.696012 5.946012 0.94817429  
## unknown-target 7.000 0.178988 13.821012 0.04315120  
##   
##   
## [[14]]  
## [1] "30.09.2019" "0.188511876795768" "aov > 0.05"   
##   
## [[15]]  
## [[15]][[1]]  
## [1] "31.10.2019"  
##   
## [[15]]$type  
## diff lwr upr p adj  
## target-referral -16.642857 -21.606103 -11.679611 1.680414e-09  
## unknown-referral -2.785714 -7.748960 2.177531 3.675837e-01  
## unknown-target 13.857143 8.893897 18.820389 1.192404e-07  
##   
##   
## [[16]]  
## [[16]][[1]]  
## [1] "30.11.2019"  
##   
## [[16]]$type  
## diff lwr upr p adj  
## target-referral 1.230769 -4.092857 6.5543954 0.83938940  
## unknown-referral -5.461538 -10.785165 -0.1379123 0.04333687  
## unknown-target -6.692308 -12.015934 -1.3686815 0.01096715  
##   
##   
## [[17]]  
## [[17]][[1]]  
## [1] "31.12.2019"  
##   
## [[17]]$type  
## diff lwr upr p adj  
## target-referral -6.25000000 -12.37228782 -0.1277122 0.04460751  
## unknown-referral -0.08333333 -6.20562115 6.0389545 0.99938517  
## unknown-target 6.16666667 0.04437885 12.2889545 0.04806423  
##   
##   
## [[18]]  
## [1] "31.01.2020" "0.128693423062993" "aov > 0.05"   
##   
## [[19]]  
## [1] "29.02.2020" "0.0535272782214378" "aov > 0.05"   
##   
## [[20]]  
## [1] "31.03.2020" "0.349473576295044" "aov > 0.05"   
##   
## [[21]]  
## [[21]][[1]]  
## [1] "30.04.2020"  
##   
## [[21]]$type  
## diff lwr upr p adj  
## target-referral -10.50 -22.6183039 1.618304 0.09737306  
## unknown-referral 1.75 -10.3683039 13.868304 0.92981302  
## unknown-target 12.25 0.1316961 24.368304 0.04725866  
##   
##   
## [[22]]  
## [1] "31.05.2020" "0.963130582152476" "aov > 0.05"   
##   
## [[23]]  
## [1] "30.06.2020" "0.191562049512868" "aov > 0.05"   
##   
## [[24]]  
## [1] "31.07.2020" "0.956892862186442" "aov > 0.05"   
##   
## [[25]]  
## [1] "31.08.2020" "0.991031447539367" "aov > 0.05"   
##   
## [[26]]  
## [1] "30.09.2020" "0.484728486989564" "aov > 0.05"   
##   
## [[27]]  
## [1] "31.10.2020" "0.939097085608934" "aov > 0.05"

## Выводы

Согласно полученным данным можно сделать вывод, что присутствует некая *третья* переменная, которая определяет результат. Тем не менее из полученных данных часто проскакивает зависимость между неизвестным типом прихода к остальным как самая незначимая. Что это может быть за третья переменная? Вероятно она может быть связана непосредственно с проведенной таргетинговой кампанией в тот период времени. Но точного ответа нет, поэтому нельзя заявить, что **клиенты, пришедшие благодаря усилиям таргетинговых кампаний, приносят наибольший LTV**.

1. Статья про [когортный анализ и немного об LTV](https://vc.ru/s/productstar/134090-chto-takoe-kogortnyy-analiz). [↑](#footnote-ref-20)
2. Статья из Википедии про [таргетинговую рекламу](https://ru.wikipedia.org/wiki/%D0%A2%D0%B0%D1%80%D0%B3%D0%B5%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D0%B0%D1%8F_%D1%80%D0%B5%D0%BA%D0%BB%D0%B0%D0%BC%D0%B0). [↑](#footnote-ref-22)
3. Пришли по приглашению другой семьи. [↑](#footnote-ref-24)
4. Данные из google-drive “[Влияние типа прихода клиента на LTV](https://docs.google.com/spreadsheets/d/1LTSw4nZUKFMoh0HJ88Gs36aJQaaxq4GkPCu7drYTj4E/edit?usp=sharing)”. [↑](#footnote-ref-25)
5. Имеется k выборок равного объёма n из нормально распределённой совокупности. Проверяется гипотеза о статистической неразличимости средних:

   $$\\n\_0 = \bar\mu\_1 = \bar\mu\_2 = ... = \bar\mu\_k$$

   . [↑](#footnote-ref-34)