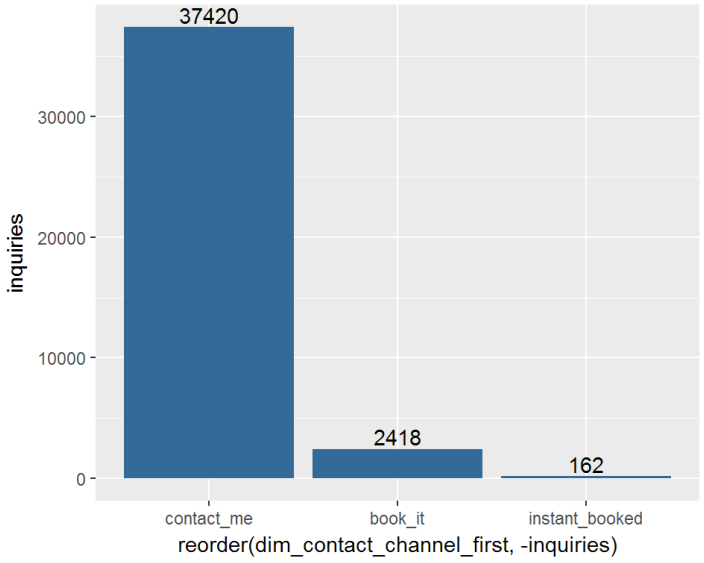
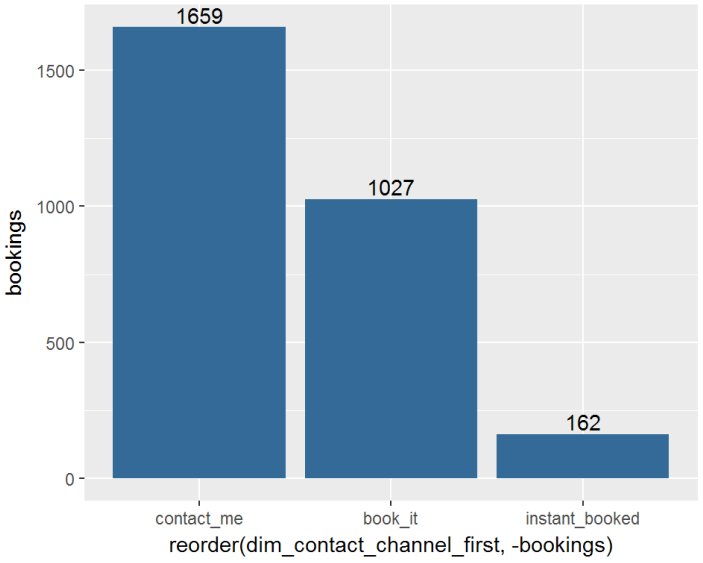
Airbnb Challenge - Bruce Hao

Executive Summary

Data summary:

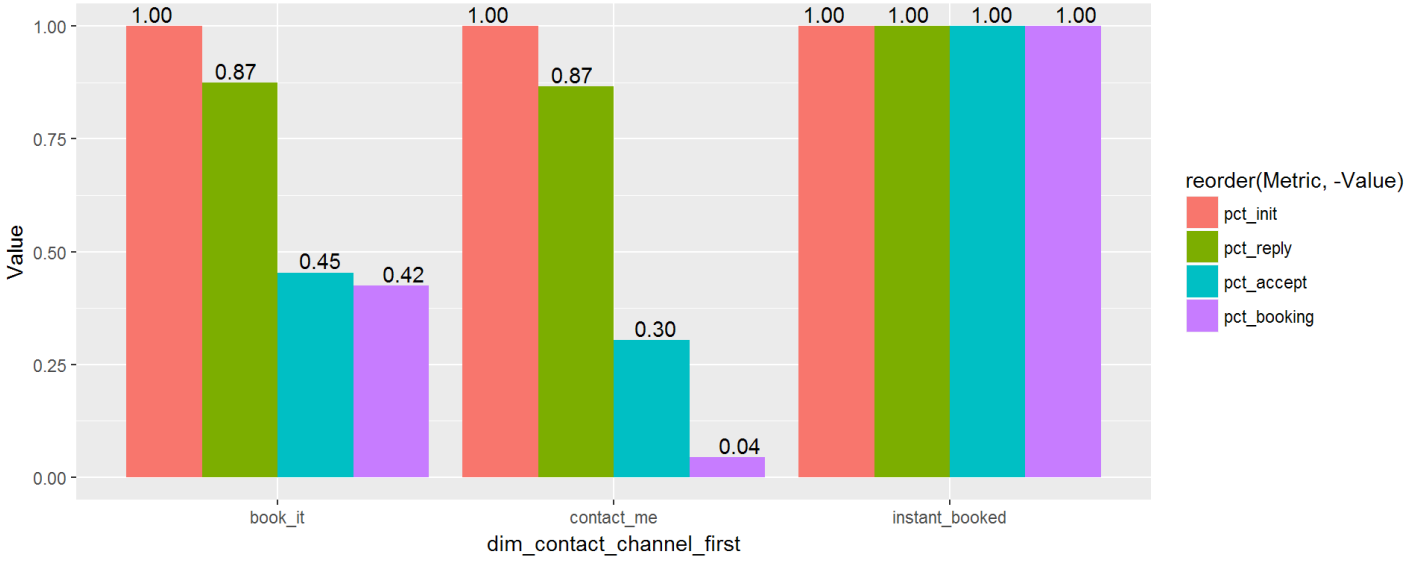
- Out of 40,000 total inquiries, there were 2,848 (7.1%) bookings
- 93.6%, 6.0% and 0.4% of inquiries came from contact_me, book_it and instant_booked, respectively
- 58.3%, 36.1% and 5.7% of bookings came from contact_me, book_it and instant_booked, respectively





Opportunities to increase the number of successful bookings in Rio de Janeiro:

- Contact Me is the largest channel but only 14.4% of acceptances result in bookings - *This is a large opportunity. A 1% improvement would result in ~7% more bookings through this channel, so we need to understand how to improve the ratio of bookings to acceptances*
- Book It is also a major source of bookings but only 45.3% of inquiries are accepted - *This is also an important opportunity, although a 1% improvement would only result in ~2% more bookings through in this channel*
- Instant Booked unsurprisingly has a 100% conversion rate but only represents 5.7% of total bookings - *Understanding how many listings offer instant booking and why may shed light on how to increase host participation in this channel*
- Host initial reply rates to guest inquiries averages 87.1% - *This could also be a large opportunity as any improvement should increase conversions across relevant channels, so we need to understand why hosts are not always replying to guest inquiries*



Product recommendations that could address the opportunities identified:

- Contact Me bookings-to-acceptance ratio:
  - The number of interactions is by far the most important variable when it comes to the bookings-to-acceptance ratio, with sweet spot between 10-16 interactions. *Before recommending product changes to encourage more interactions, a deeper understanding of interactions (successul and unsuccessful) is needed*, e.g. isolating interactions to those prior to booking, understanding time lapse between interactions, understanding how interactions typically end, etc.
  - Reply time is a distant second in terms of importance but may prove more actionable. The faster the host replies to the initial inquiry, the higher the ratio. We cannot draw causal conclusions without experimentation, but *encouraging hosts to reply more quickly may increase host engagement and result in higher conversions*
- Book It acceptance-to-inquiry ratio:
  - Here again, the number of interactions is the most important variable, with a sweet spot between 8.5-12 interactions. Similar to the case above, a deeper understanding of interactions is needed before suggesting any changes to the product
  - The number of total reviews is the second most important variable, but again a causal relationship cannot be established without experimentation
  - Although not quite as explanatory, reply time and lead time do have explanatory effect and are perhaps more actionable. Similar to above, faster reply times are associated with higher acceptance rates, and longer lead times are associated with lower acceptance rates. *To the extent that guests systematically struggle to find bookings that are further in the future, we may incentivize hosts to accept bookings with longer lead times (or at least understand why they might be more adverse to such inquiries)*
- Instant Booked host participation:
  - Only 1.2% of listings in the area have made instant booked bookings. Since conversion rates are unsurprisingly high for this channel, *we should encourage hosts to allow instant booked through promotions and/or other incentives (or understand why they might be adverse to instant booked and how we might mitigate their concerns)*
- Host initial reply rates:
  - Again, the number of reviews is an important factor in reply rates, but of course a guest cannot book and review if the host never replies
  - Lead time is the second most important factor, with longer lead times associated with lower reply rates. Again, this suggests that hosts may be somewhat adverse to bookings too far into the future
  - As with all ratios, there exists some natural limit, so *we should compare reply rates in this region to others to determine how much room for improvement there might be and perhaps gain some insights on how to effect improvement*

Other data and next steps that may be useful to deepen our analysis and understanding of the market:

Incorporating the following qualitative and quantitative data will likely prove helpful in informing actual product changes. By having multiple data points and analyses supporting the same conclusions, we can be more confident that the observed relationships are real. Experimentation like A/B tests would need to be run to determine any causal relationships, however.

- Nature and quality of interactions - as interactions are the major factor in conversions, we should conduct further research. For example, it appears that conversions are maximized between 8-16 interactions, but is this simply related to the quantity of interactions? Are there systematic differences between successful and unsuccessful interactions? How many interactions take

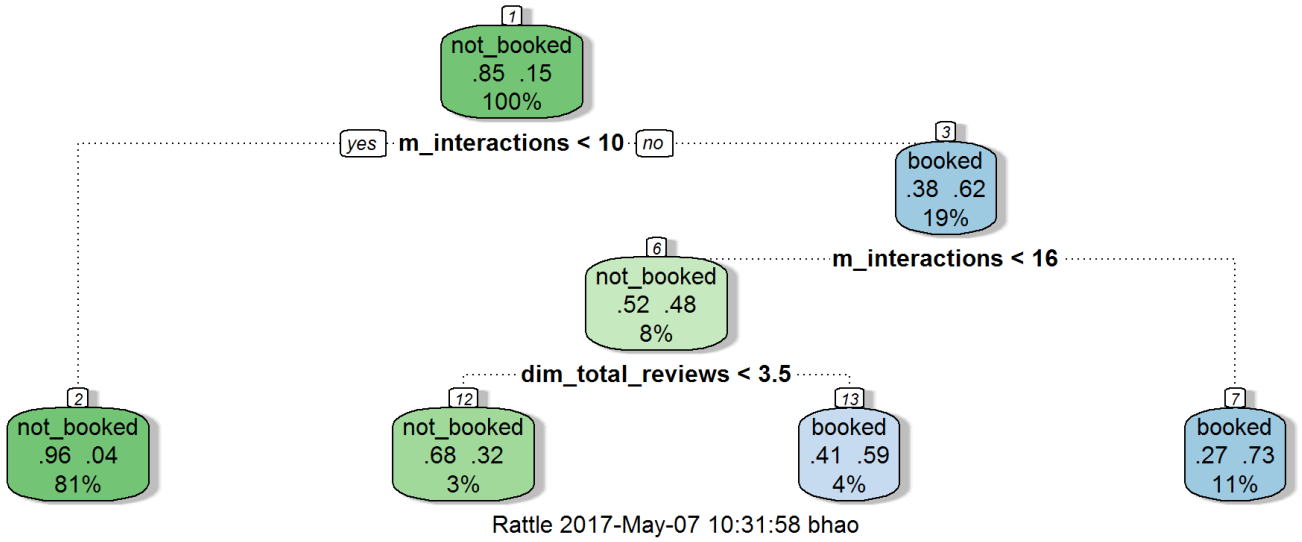
place before bookings and after? Are there any inquiries that convert despite fewer interactions? What drives interactions? What curtails them?

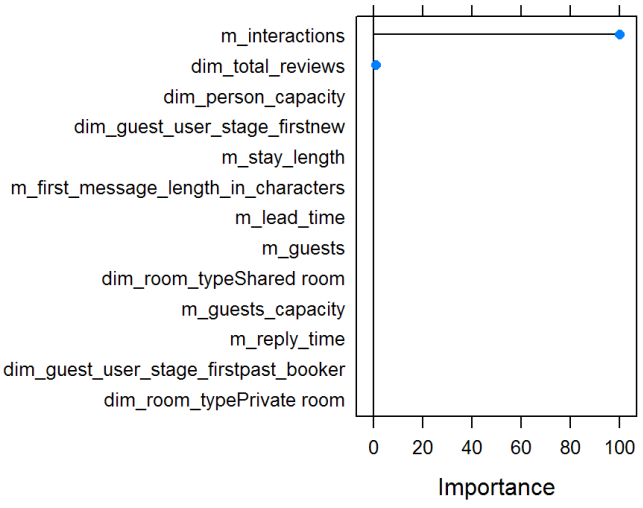
- Percent of users who write reviews - the number of listing reviews is also an important variable; however, further research would be necessary to determine causal relationships with other variables
- Quality of listing reviews - the quality of listing reviews may be a better indicator than the quantity since the relationship may be less nebulous
- Quality of guests - data on photo, profile (how much is completed), reviews, verification, etc. - these data may help provide more actionable insights into how guests may improve acceptance rates
- Quality of host - data on photo, profile (how much is completed), reviews, etc. - these data may help provide more actionable insights into how hosts may improve booking rates, especially in the absence of listing reviews

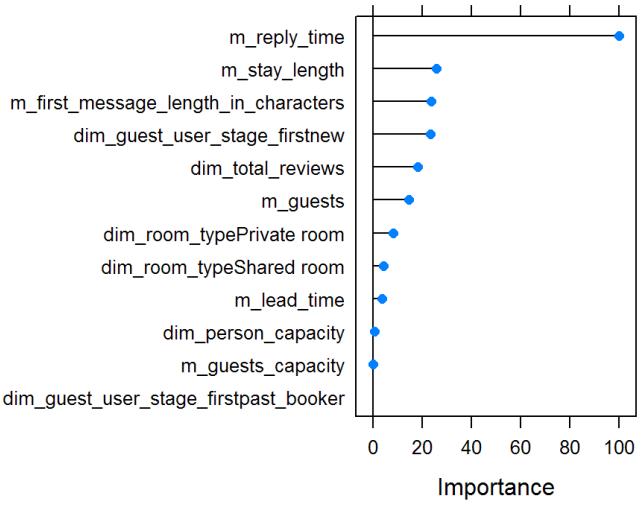
Contact Me - Why are booking rates so low after host acceptance?

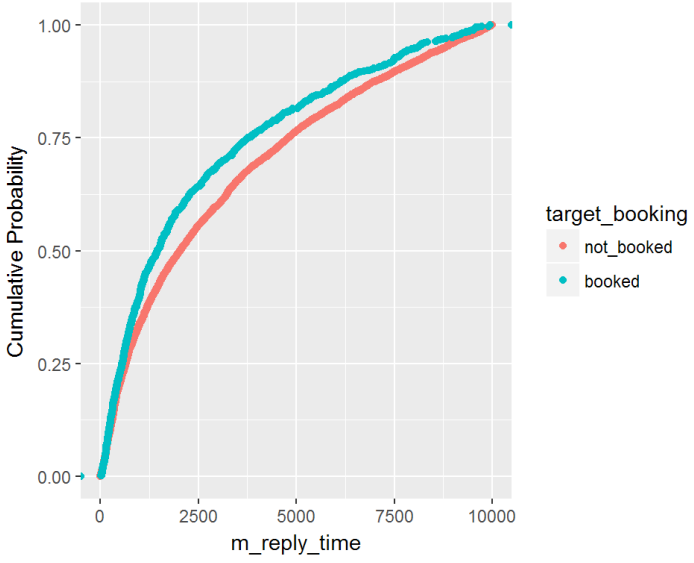
To evaluate this, we first subset the data to include only the contact_me data where the host accepted the booking. We then use decision trees, logistic regression, variable importance measures and empirical cumulative distribution charts to identify which variables are influential to a guest finalizing a booking.

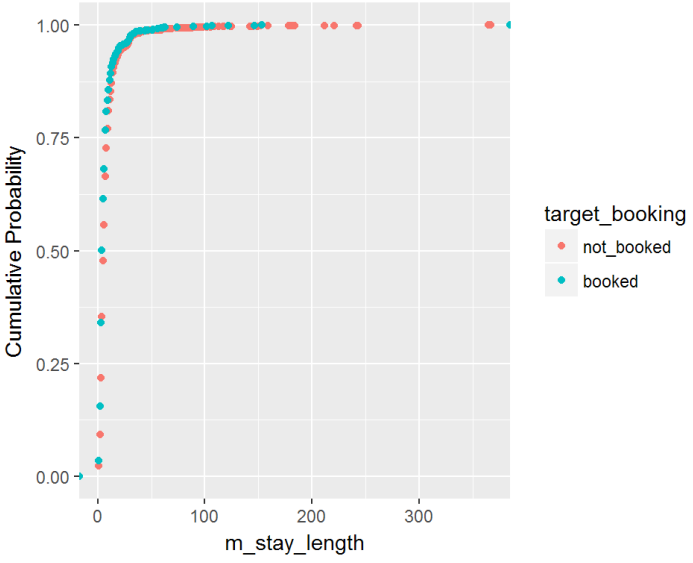
- m_interactions is the most important variable when it comes to explaining bookings. Interestingly, there is a ‘sweet spot’ for m_interactions between ~10 and ~16 with bookings falling off for m_interactions below 10 and somewhat above 16 (perhaps suggesting to a more complex situation requiring more correspondence)
    - Further research into the nature and quality of interactions may be necessary in order understand exactly how actionable this lever may be
- Given that it is unclear exactly how actionable m_interactions is as a lever for bookings conversions, we remove it as a variable to evaluate the impact of the remaining variables
- m_reply_time (custom defined variable) is the next most important variable with faster replies from hosts associated with higher booking rates - this is certainly actionable, and we should encourage and gently remind hosts to reply to guest inquiries promptly
- Other measured effects:
    - Shorter lengths of stay (m_stay_length: custom defined variable) are associated with higher booking rates
    - Booked inquiries exhibit longer initial messages on average - we can gently encourage guests to write lengthier initial messages
    - More listing reviews are associated with higher booking rates - of course, we cannot determine which variable is affecting the other, but we can gently encourage guests to write reviews after stays as more reviews is likely generally beneficial

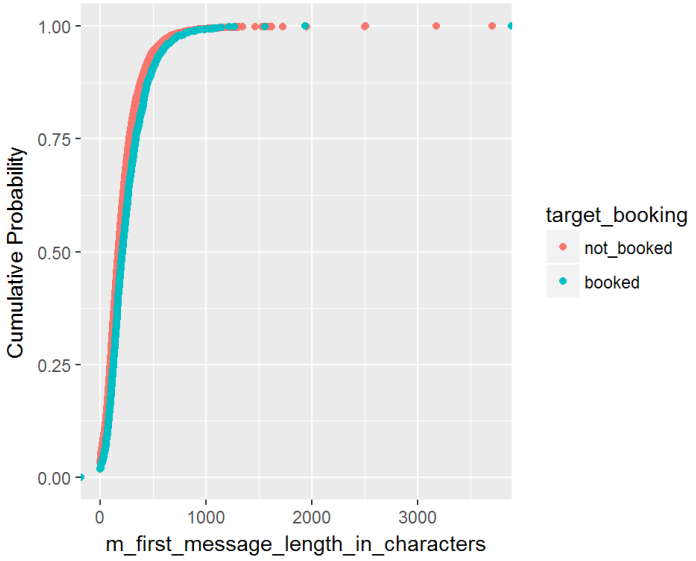

1
not_booked
.85 .15
100%
yes
m_interactions < 10
no
3
booked
.38 .62
19%
6
not_booked
.52 .48
8%
m_interactions < 16
dim_total_reviews < 3.5
2
not_booked
.96 .04
81%
12
not_booked
.68 .32
3%
13
booked
.41 .59
4%
7
booked
.27 .73
11%
Rattle 2017-May-07 10:31:58 bhao

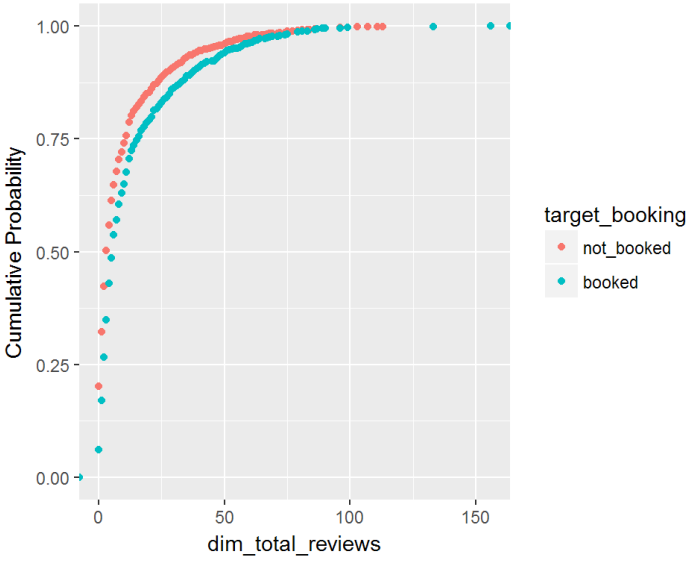

m_interactions
dim_total_reviews
dim_person_capacity
dim_guest_user_stage_firstnew
m_stay_length
m_first_message_length_in_characters
m_lead_time
m_guests
dim_room_typeShared room
m_guests_capacity
m_reply_time
dim_guest_user_stage_firstpast_booker
dim_room_typePrivate room
0 20 40 60 80 100
Importance


m_reply_time
m_stay_length
m_first_message_length_in_characters
dim_guest_user_stage_firstnew
dim_total_reviews
m_guests
dim_room_typePrivate room
dim_room_typeShared room
m_lead_time
dim_person_capacity
m_guests_capacity
dim_guest_user_stage_firstpast_booker
0 20 40 60 80 100
Importance


Cumulative Probability
1.00
0.75
0.50
0.25
0.00
0 2500 5000 7500 10000
m_reply_time
target_booking
not_booked
booked


Cumulative Probability
1.00
0.75
0.50
0.25
0.00
0 100 200 300
m_stay_length
target_booking
not_booked
booked

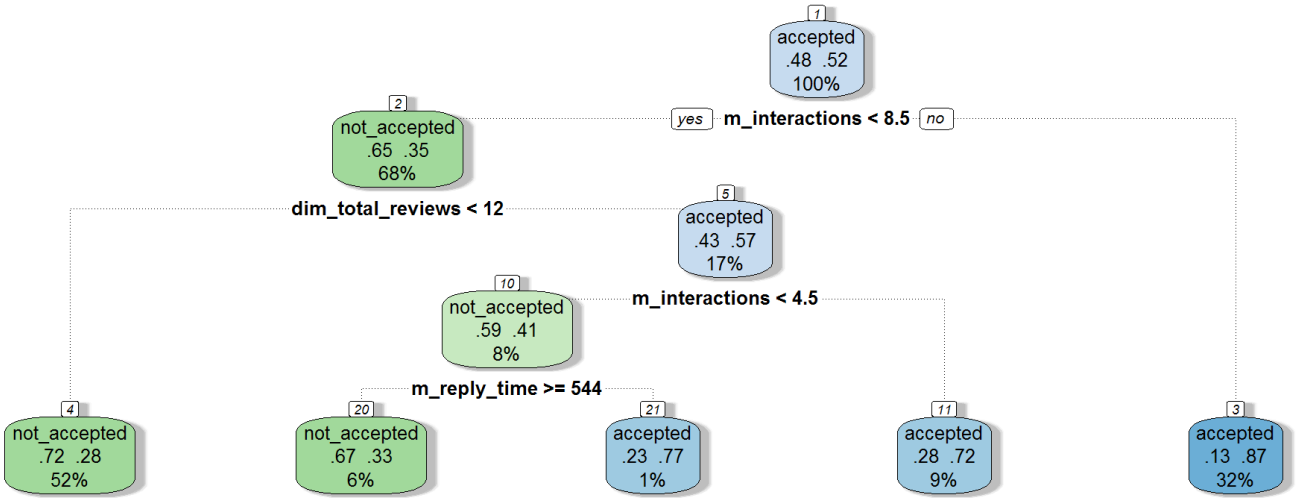




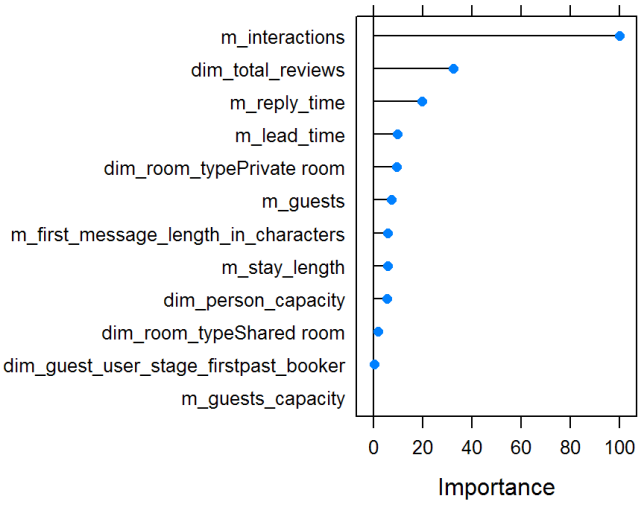
Book It - Why are acceptance rates so low after initial response?

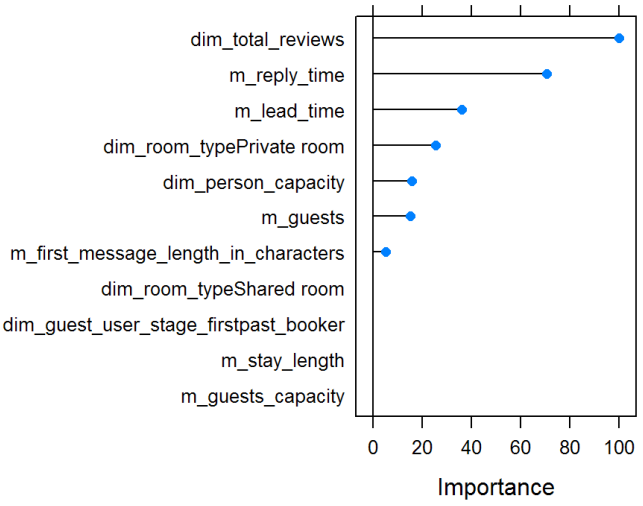
To evaluate this, we first subset the data to include only the book_it data where the host replied to the guest. We then use decision trees, logistic regression, variable importance measures and empirical cumulative distribution charts to identify which variables are influential to a host accepting an inquiry.

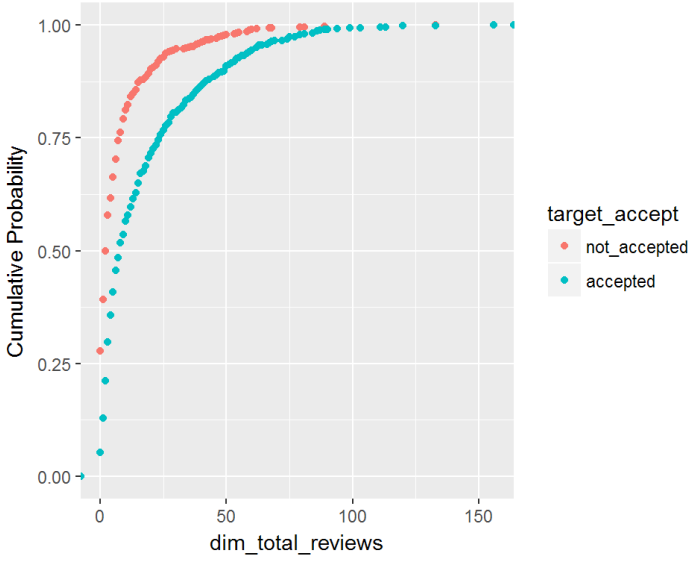
- Again, m_interactions is the most important variables; we proceed as we did above by removing the variable to evaluate the remainder
- dim_total_reviews was the second most important variable - similar logic to above
- m_reply_time is the third most important variable - similar logic to above
- m_lead_time (custom defined variable) interestingly is associated with with shorter lead times associated with higher acceptance rates - we can do more work to understand what is driving this, but to the extent that we find that guests systematically struggle to find bookings that are further in the future, we may encourage hosts to accept bookings with longer lead times

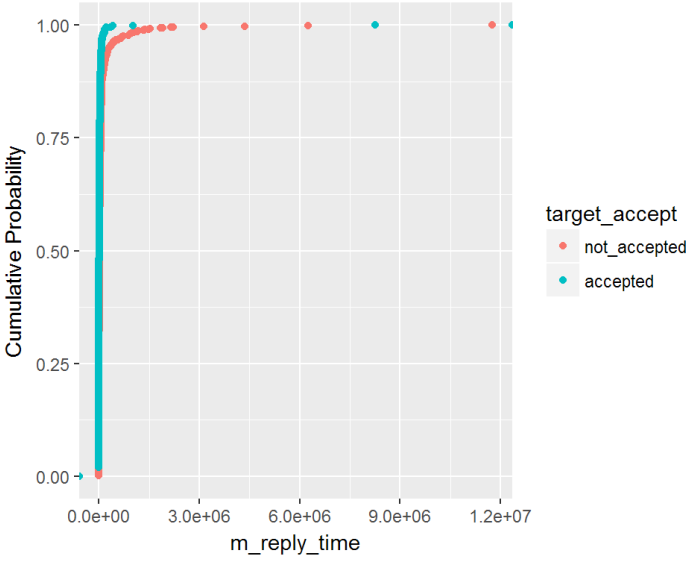


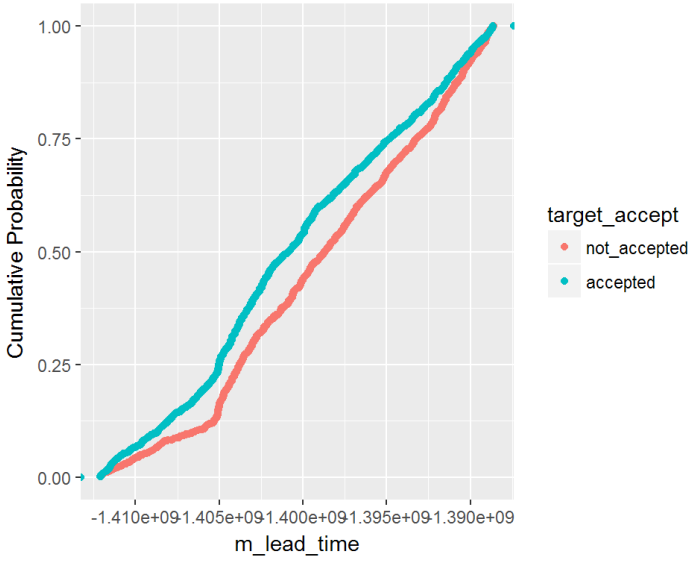
Rattle 2017-May-07 10:33:05 bhao











Instant Booked - How many listings have instant booked as an option?

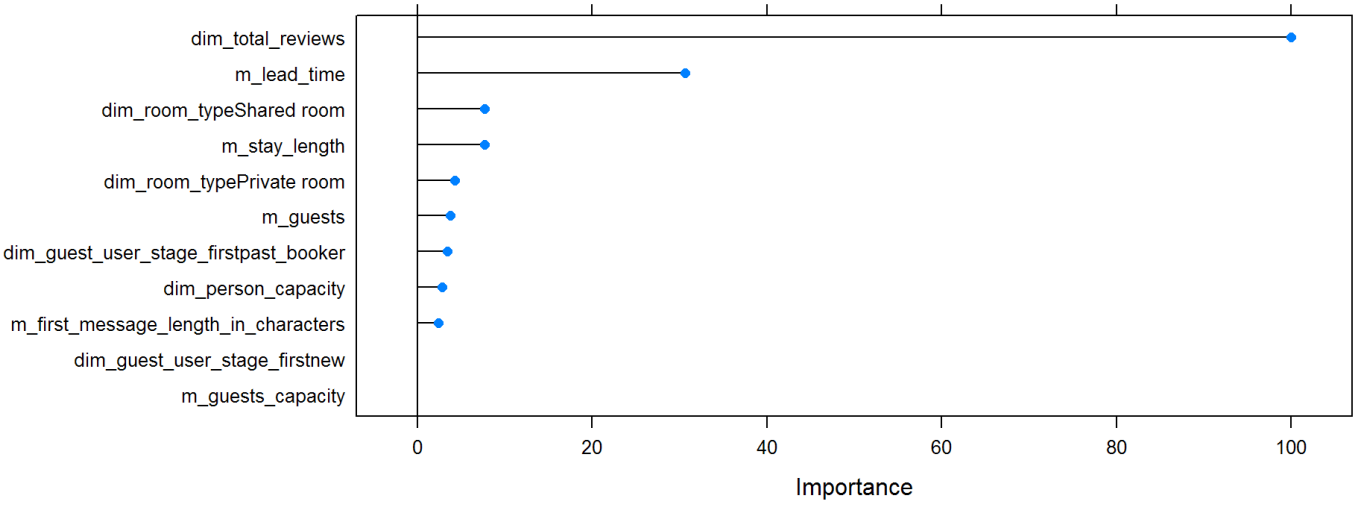
With more data on listings, e.g. which offer instant booked and other listing features, we may be able to discover important levers to drive higher instant booked adoption among the hosts within this region. With the data we have, we can really only estimate the participation level.

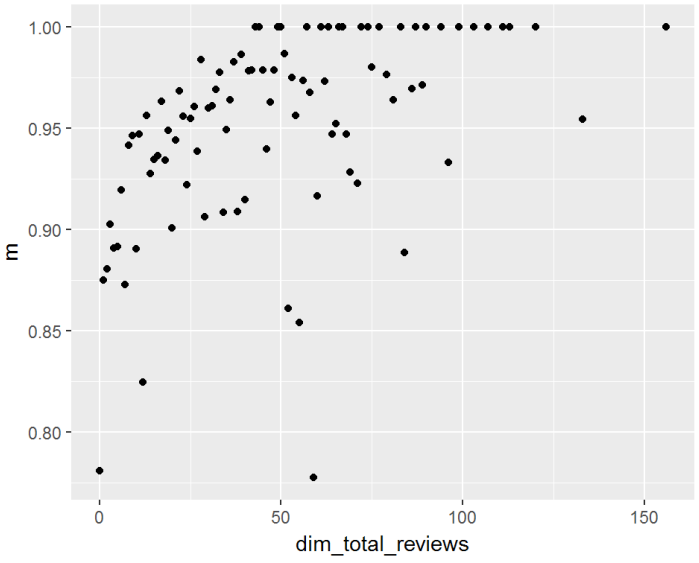
- While based on the data, we cannot tell how many listings offer instant booked, the ratio of unique instant booked listings to all unique listings may provide some color
- Only 1.2% of listings have made instant booked bookings. Since conversion rates are so high for instant booked, we should encourage hosts to allow instant booked through promotions and/or other incentives, perhaps

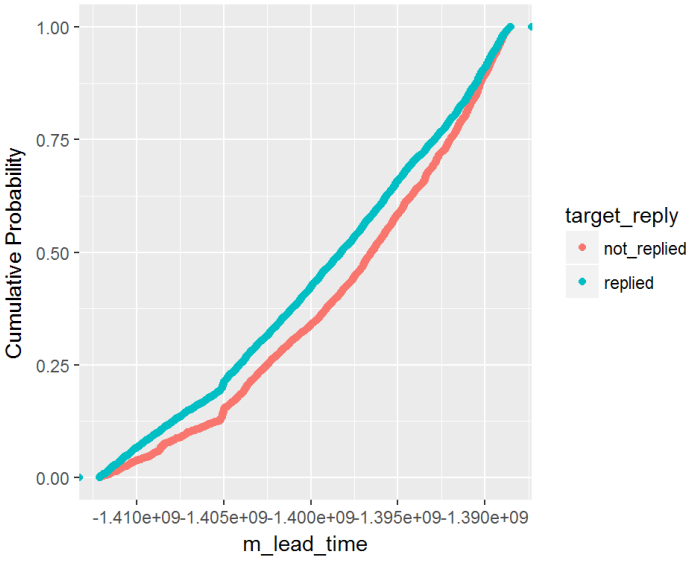
Host Initial Reply Rate - Why are ~13% of inquiries not being replied to by hosts?

With more data on the quality of the guests and hosts profiles, we understand more deeply the drivers behind initial reply rates and potentially find ways to increase reply rates and conversions.

- dim_total_reviews - hosts of listings with more reviews tend to reply much more consistently than other hosts - again, it's impossible to determine a causal relationship here without additional information
- Shorter lead times are associated with higher initial response rates, again suggesting that hosts may be somewhat adverse to bookings for dates further into the future


dim_total_reviews
m_lead_time
dim_room_typeShared room
m_stay_length
dim_room_typePrivate room
m_guests
dim_guest_user_stage_firstpast_booker
dim_person_capacity
m_first_message_length_in_characters
dim_guest_user_stage_firstnew
m_guests_capacity
0
20
40
60
80
100
Importance


1.00
0.95
0.90
0.85
0.80
m
0
50
100
150
dim_total_reviews


1.00
0.75
0.50
0.25
0.00
Cumulative Probability
-1.410e+09
-1.405e+09
-1.400e+09
-1.395e+09
-1.390e+09
m_lead_time
target_reply
not_replied
replied

Airbnb Challenge - CODE

```
# load necessary libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(gridExtra)
library(caret)
library(rpart)
library(rpart.plot)
library(rattle)

# load and format data
contacts = read.csv(paste0(path, 'contacts.csv'), stringsAsFactors = FALSE)
contacts = contacts %>%
  mutate(ts_interaction_first = ts_interaction_first %>% as.POSIXct("%Y-%m-%d %H:%M:%S"),
         ts_reply_at_first = ts_reply_at_first %>% plyr::mapvalues('', NA) %>%
           as.POSIXct("%Y-%m-%d %H:%M:%S"),
         ts_accepted_at_first = ts_accepted_at_first %>% plyr::mapvalues('', NA) %>%
           as.POSIXct("%Y-%m-%d %H:%M:%S"),
         ts_booking_at = ts_booking_at %>% plyr::mapvalues('', NA) %>%
           as.POSIXct("%Y-%m-%d %H:%M:%S"),
         m_reply_time = as.numeric(ts_reply_at_first - ts_interaction_first),
         ds_checkin_first = ds_checkin_first %>% plyr::mapvalues('', NA) %>% as.Date(),
         ds_checkout_first = ds_checkout_first %>% plyr::mapvalues('', NA) %>% as.Date(),
         m_stay_length = as.numeric(ds_checkout_first - ds_checkin_first),
         m_lead_time = as.numeric(ds_checkin_first - ts_interaction_first),
         target_booking = factor(if_else(is.na(ts_booking_at), 0, 1),
                                 labels = c('not_booked', 'booked')),
         target_accept = factor(if_else(is.na(ts_accepted_at_first), 0, 1),
                                labels = c('not_accepted', 'accepted')),
         target_reply = factor(if_else(is.na(ts_reply_at_first), 0, 1),
                               labels = c('not_replied', 'replied'))
         )

listings = read.csv(paste0(path, 'listings.csv'), stringsAsFactors = FALSE)
users = read.csv(paste0(path, 'users.csv'), stringsAsFactors = FALSE)

# combine data sets
# there are 11 user_ids in contacts that are not in users; however, this is not material or worrisome
combined_data = contacts %>% left_join(listings) %>%
  left_join(users, by = c('id_guest_anon' = 'id_user_anon')) %>%
  mutate(m_guests_capacity = m_guests - dim_person_capacity)
```

Executive Summary

```
# total bookings by channel
p1 = contacts %>% group_by(dim_contact_channel_first) %>%
  summarise(inquiries = n()) %>% mutate(inquiries_pct = inquiries / sum(inquiries)) %>%
  ggplot(aes(x = reorder(dim_contact_channel_first, -inquiries), y = inquiries, fill = 0)) +
  geom_bar(stat = 'identity') +
  geom_text(aes(label = inquiries), vjust = -0.25, position = position_dodge(width = 1)) +
  theme(legend.position="none")

p2 = contacts %>% group_by(dim_contact_channel_first) %>%
  summarise(bookings = sum(!is.na(ts_booking_at))) %>% mutate(bookings_pct = bookings / sum(bookings)) %>%
  ggplot(aes(x = reorder(dim_contact_channel_first, -bookings), y = bookings, fill = 0)) +
```

```
  geom_bar(stat = 'identity') +
  geom_text(aes(label = bookings), vjust = -0.25, position = position_dodge(width = 1)) +
  theme(legend.position="none")

grid.arrange(p1, p2, ncol = 2)
```

```
# contact to booking funnel by channel
contacts %>% select(dim_contact_channel_first, ts_interaction_first, ts_reply_at_first,
                    ts_accepted_at_first, ts_booking_at) %>%
  group_by(dim_contact_channel_first) %>%
  summarise(pct_init = sum(!is.na(ts_interaction_first)) / n(),
            pct_reply = sum(!is.na(ts_reply_at_first)) / n(),
            pct_accept = sum(!is.na(ts_accepted_at_first)) / n(),
            pct_booking = sum(!is.na(ts_booking_at)) / n()
            ) %>%
  gather(Metric, Value, -1) %>%
  ggplot(aes(x = dim_contact_channel_first, y = Value, fill = reorder(Metric, -Value))) +
    geom_bar(stat = 'identity', position = position_dodge()) +
    geom_text(aes(label = format(Value, digits = 1)), vjust = -0.25, position = position_dodge(width = 1))
```

Contact Me - Why are booking rates so low after host acceptance?

```
# isolate accepted inquiries within contact_me channel
contact_me = combined_data %>%
  filter(dim_contact_channel_first == 'contact_me' & !is.na(ts_accepted_at_first))
# select potential factors and delete rows with missing data (<12)
contact_me = contact_me %>%
  select(target_booking, m_guests, m_interactions,  m_first_message_length_in_characters,
         dim_guest_user_stage_first, m_reply_time, m_stay_length, m_lead_time,
         dim_room_type, dim_person_capacity, dim_total_reviews, m_guests_capacity)
contact_me = contact_me[complete.cases(contact_me),]

# then isolate which variables impact bookings
# decision tree
set.seed(123)
tree = rpart(target_booking ~ ., data = contact_me)
fancyRpartPlot(tree)

# logistic regression
# logreg_model = glm(as.numeric(target_booking) ~ ., data = contact_me)
# summary(logreg_model)

# review variable importance
myControl = trainControl(verboseIter = FALSE, summaryFunction = twoClassSummary, classProbs = TRUE)  # IMPORTA
glmnet_model = train(target_booking ~ ., data = contact_me, method = 'glmnet',
                     trControl = myControl, preProcess = c('center', 'scale'))
p1 = plot(varImp(glmnet_model))

# remove m_interactions and rerun analysis
contact_me = contact_me %>%
  select(target_booking, m_guests, m_first_message_length_in_characters,
         dim_guest_user_stage_first, m_reply_time, m_stay_length, m_lead_time,
         dim_room_type, dim_person_capacity, dim_total_reviews, m_guests_capacity)
glmnet_model = train(target_booking ~ ., data = contact_me, method = 'glmnet',
                     trControl = myControl, preProcess = c('center', 'scale'))
p2 = plot(varImp(glmnet_model))

# faster replies from hosts are associated with higher booking rates
```

```r
p3 = contact_me %>% ggplot(aes(x = m_reply_time, color = target_booking)) + stat_ecdf(geom = 'point') +
  xlim(c(0, 10000)) + ylab('Cumulative Probability')
# t.test(contact_me[contact_me$target_booking == 'booked', 'm_reply_time'],
#        contact_me[contact_me$target_booking == 'not_booked', 'm_reply_time'])

# shorter lengths of stay are associated with higher booking rates
p4 = contact_me %>% ggplot(aes(x = m_stay_length, color = target_booking)) +
  stat_ecdf(geom = 'point') + ylab('Cumulative Probability')
# t.test(contact_me[contact_me$target_booking == 'booked', 'm_stay_length'],
#        contact_me[contact_me$target_booking == 'not_booked', 'm_stay_length'])

# booked inquiries exhibit longer initial messages on average
p5 = contact_me %>% ggplot(aes(x = m_first_message_length_in_characters, color = target_booking)) +
  stat_ecdf(geom = 'point') + ylab('Cumulative Probability')
# t.test(contact_me[contact_me$target_booking == 'booked', 'm_first_message_length_in_characters'],
#        contact_me[contact_me$target_booking == 'not_booked', 'm_first_message_length_in_characters'])

# More listing reviews are associated with higher booking rates
p6 = contact_me %>% ggplot(aes(x = dim_total_reviews, color = target_booking)) +
  stat_ecdf(geom = 'point') + ylab('Cumulative Probability')
# t.test(contact_me[contact_me$target_booking == 'booked', 'dim_total_reviews'],
#        contact_me[contact_me$target_booking == 'not_booked', 'dim_total_reviews'])

grid.arrange(p1, p2, ncol = 2)
grid.arrange(p3, p4, ncol = 2)
grid.arrange(p5, p6, ncol = 2)
```

Book It - Why are acceptance rates so low after initial response?

```r
# isolate replied inquiries within book_it channel
book_it = combined_data %>% filter(dim_contact_channel_first == 'book_it' & !is.na(ts_reply_at_first))
# select potential factors and delete rows with missing data (<12)
book_it = book_it %>%
  select(target_accept, m_guests, m_interactions, m_first_message_length_in_characters,
         dim_guest_user_stage_first, m_reply_time, m_stay_length, m_lead_time, dim_room_type,
         dim_person_capacity, dim_total_reviews, m_guests_capacity)
book_it = book_it[complete.cases(book_it),]

# then isolate which variables impact acceptance
# decision tree
set.seed(123)
tree = rpart(target_accept ~ ., data = book_it)
fancyRpartPlot(tree)

# logistic regression
# logreg_model = lm(as.numeric(target_accept) ~ ., data = book_it)
# summary(logreg_model)

# review variable importance
myControl = trainControl(verboseIter = FALSE, summaryFunction = twoClassSummary, classProbs = TRUE)  # IMPORTA
glmnet_model = train(target_accept ~ ., data = book_it, method = 'glmnet',
                     trControl = myControl, preProcess = c('center', 'scale'))
p1 = plot(varImp(glmnet_model))

# remove m_interactions and rerun analysis
book_it = book_it %>%
  select(target_accept, m_guests, m_first_message_length_in_characters, dim_guest_user_stage_first,
         m_reply_time, m_stay_length, m_lead_time, dim_room_type, dim_person_capacity, dim_total_reviews,
```

```
        m_guests_capacity)
glmnet_model = train(target_accept ~ ., data = book_it, method = 'glmnet',
                     trControl = myControl, preProcess = c('center', 'scale'))
p2 = plot(varImp(glmnet_model))

# More listing reviews are associated with higher acceptance rates
p3 = book_it %>% ggplot(aes(x = dim_total_reviews, color = target_accept)) + stat_ecdf(geom = 'point') +
  ylab('Cumulative Probability')
# t.test(book_it[book_it$target_accept == 'accepted', 'dim_total_reviews'],
#        book_it[book_it$target_accept == 'not_accepted', 'dim_total_reviews'])

# Shorter reply times are associated with higher acceptance rates
p4 = book_it %>% ggplot(aes(x = m_reply_time, color = target_accept)) + stat_ecdf(geom = 'point') +
  ylab('Cumulative Probability')
# t.test(book_it[book_it$target_accept == 'accepted', 'm_reply_time'],
#        book_it[book_it$target_accept == 'not_accepted', 'm_reply_time'])

# Shorter lead times are associated with higher acceptance rates
p5 = book_it %>% ggplot(aes(x = m_lead_time, color = target_accept)) + stat_ecdf(geom = 'point') +
  ylab('Cumulative Probability')
# t.test(book_it[book_it$target_accept == 'accepted', 'm_lead_time'],
#        book_it[book_it$target_accept == 'not_accepted', 'm_lead_time'])

grid.arrange(p1, p2, ncol = 2)
grid.arrange(p3, p4, ncol = 2)
grid.arrange(p5, ncol = 2)
```

Instant Booked - How many listings have instant booked as an option?

```
instant_booked = combined_data %>% filter(dim_contact_channel_first == 'instant_booked')
# nrow(unique(instant_booked %>% filter(!is.na(ts_booking_at)) %>% select(id_listing_anon))) / length(unique(c
```

Host Initial Reply Rate - Why are ~13% of inquiries not being replied to by hosts?

```
# select potential factors and delete rows with missing data (<12)
replied = combined_data %>%
  select(target_reply, m_guests, m_first_message_length_in_characters, dim_guest_user_stage_first,
         m_stay_length, m_lead_time, dim_room_type, dim_person_capacity, dim_total_reviews,
         m_guests_capacity)
replied = replied[complete.cases(replied),]

# decision tree
# set.seed(123)
# tree = rpart(target_reply ~ ., data = replied)
# fancyRpartPlot(tree)

# logistic regression
# logreg_model = lm(as.numeric(target_reply) ~ ., data = replied)
# summary(logreg_model)

# review variable importance
myControl = trainControl(verboseIter = FALSE, summaryFunction = twoClassSummary, classProbs = TRUE)  # IMPORTA
glmnet_model = train(target_reply ~ ., data = replied, method = 'glmnet',
                     trControl = myControl, preProcess = c('center', 'scale'))
p1 = plot(varImp(glmnet_model))
```

```r
# more listing reviews are associated with higher initial response rates
p2 = combined_data %>% group_by(dim_total_reviews) %>% summarise(m = sum(!is.na(ts_reply_at_first)) / n()) %>%
  ggplot(aes(x = dim_total_reviews, y = m)) + geom_point()

# shorter lead times are associated with higher initial response rates
p3 = replied %>% ggplot(aes(x = m_lead_time, color = target_reply)) + stat_ecdf(geom = 'point') +
  ylab('Cumulative Probability')
# t.test(replied[replied$target_reply == 'replied', 'm_lead_time'],
#        replied[replied$target_reply == 'not_replied', 'm_lead_time'])

grid.arrange(p1, ncol = 1)
grid.arrange(p2, p3, ncol = 2)
```