II. Construcción de un modelo de machine learning

Información del DataSet

Hay 18 179 guests, 2118 hosts y 3172 listings.

Primero, cambio los tipos de las fechas a objetos *datetime* para poder operar con ellos, y cambio los NaN de las columnas “m_guest_first” y “m_first_message_length_in_characters” a 0 para convertir la columna a *int*.

La columna “dim_guest_language” está vacía, por lo que la boto. Por ende, no nos interesa tampoco “dim_host_language”.

Creo 3 columnas target, booleanas, “guest_accepted” y “guest_booked”. Posteriormente elimino “guest_booked” por estar directamente correlacionada con “guest_accepted”, y tendríamos información repetida. El 99% de las veces que un guest es aceptado, hace el booking.

Luego creo las siguientes columnas:

- diff_reply_interaction: Tiempo que toma el host en responder al primer mensaje enviado. (en minutos)
- stay: duración de la estadía en la inquiry. (en días)
- diff_ckeckin_interaction: delta de tiempo entre el día que el guest se planea hospedar y el momento de la primera interacción. (en días)

Separo features existentes para poder comparar sus valores.

- interaction_hour: hora de la primera interacción
- interaction_weekday: día de la semana de la primera interacción (0-lunes, 6-domingo)
- interacion_month: mes de la primera interacción

- checkin_weekday: día de la semana del checkin de la inquiry.
- checkin_month: mes del checkin

Factorizo,

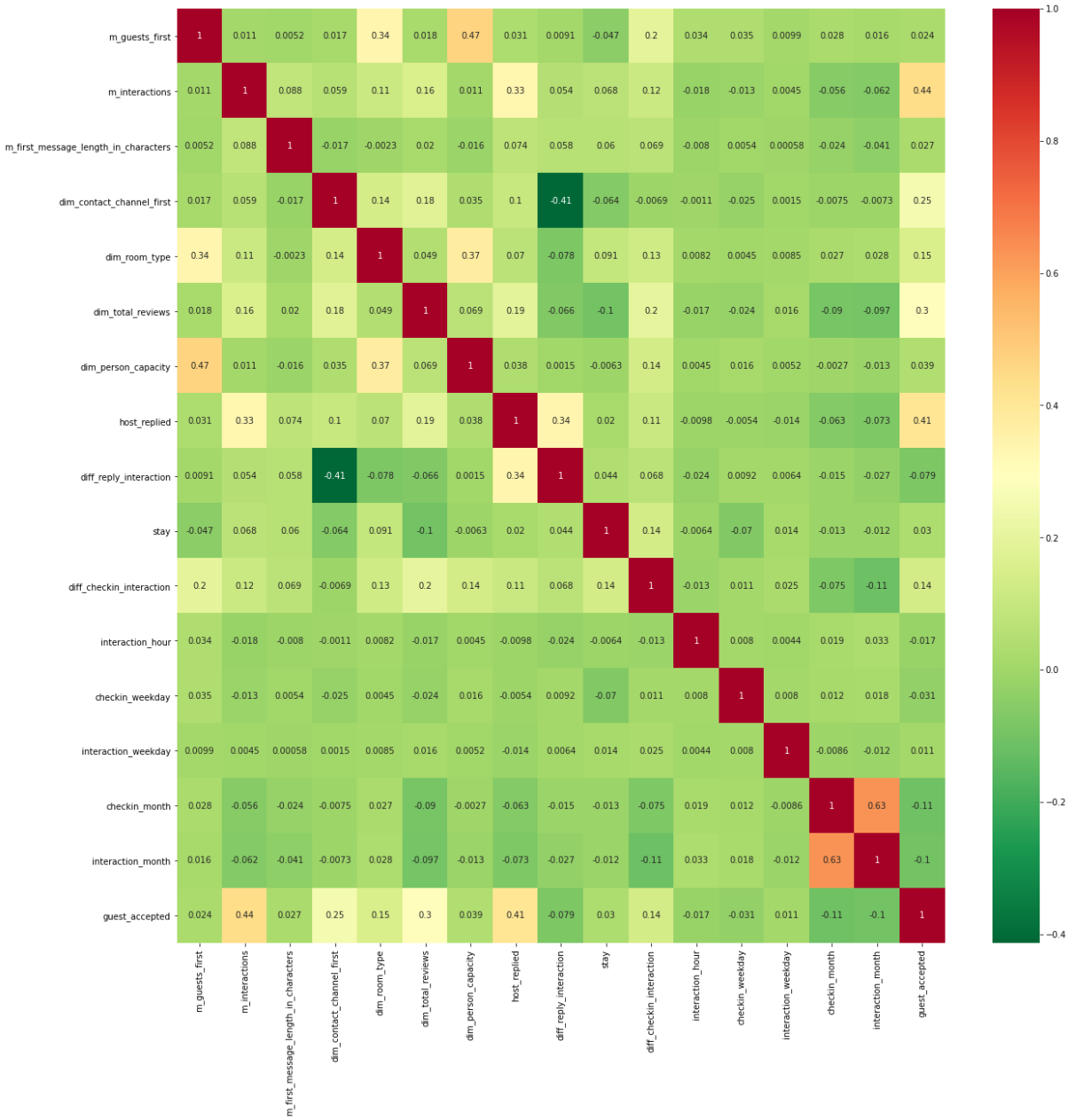
- dim_contact_channel_first: 0 bookit, 1 instant booking. Posteriormente, descarto esta feature porque el 100% de los que usan instant booking son aceptados. No aporta valor al estudio.
- dim_room_type: 0 private room, 1 entire home, 2 shares room

Por último, elimino las columnas con las que no se puede tratar, son información repetida y columnas que ya traducimos, o que no tienen causalidad.

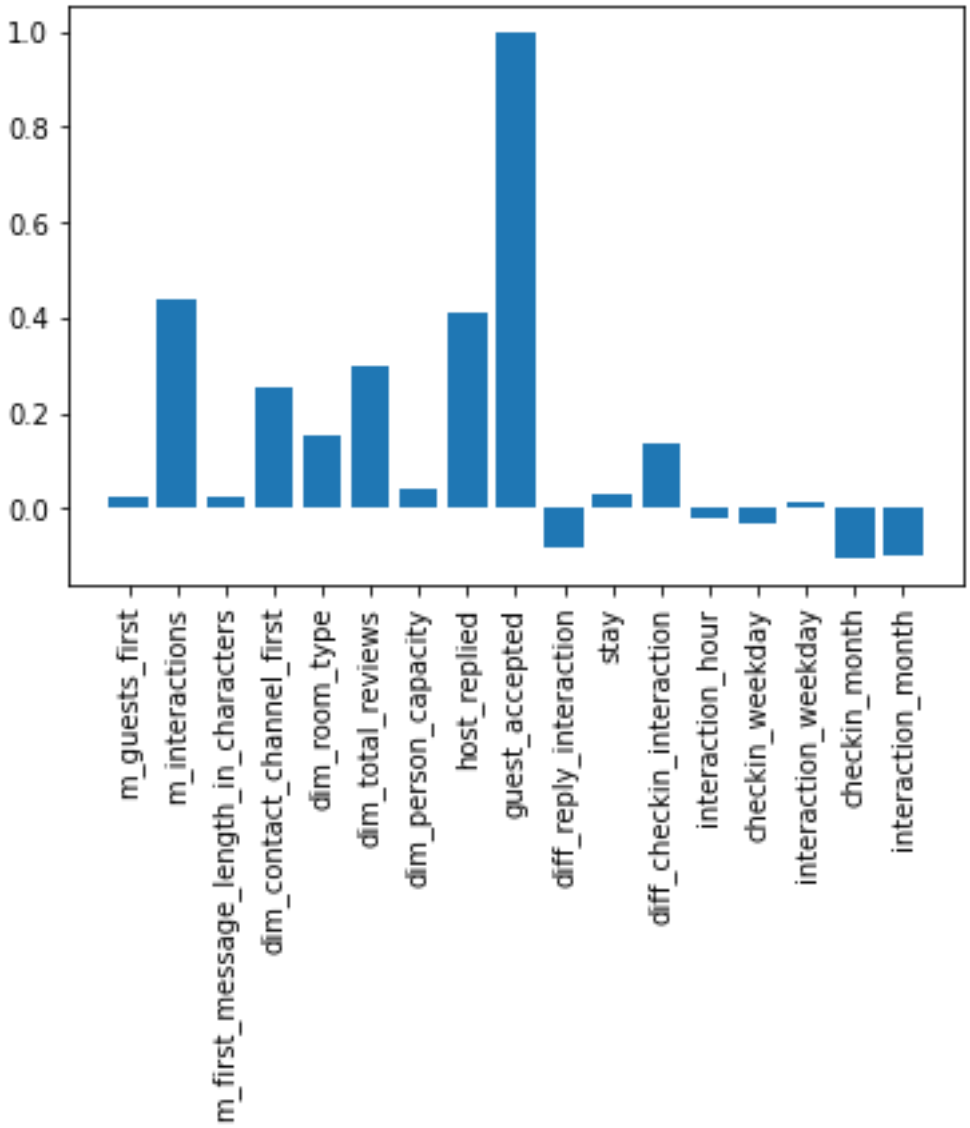
Correlación entre variables

Primero, usamos el coeficiente de correlación de Kendall para buscar una correlación entre todas las variables del DataSet. Decidí usar este coeficiente porque tratamos con variables ordinales.

Creo un heatmap para visualizar estos coeficientes, y dar una idea de las features más relevantes:



Los coeficientes de correlación que más nos interesan son para la variable guest_accepted, que es nuestra variable target. A continuación, dichos coeficientes.



Para seguridad, también calculo el coeficiente de correlación point-biseral con respecto a guest_accepted, ya que se trata de una variable binaria. Tenemos coeficientes similares.

Análisis estadístico por feature.

Por el momento, los factores que más parecen impactar en la aceptación del huésped son la cantidad de mensajes intercambiados, si el host responde (si el host responde, hay 70% de probabilidad que el guest sea aceptado), el tipo de cuarto, las total reviews, el mes de la interacción y del checkin, y cuanto antes se hace la inquiry.

En el notebook examino cada una de las features y doy más información al respecto.

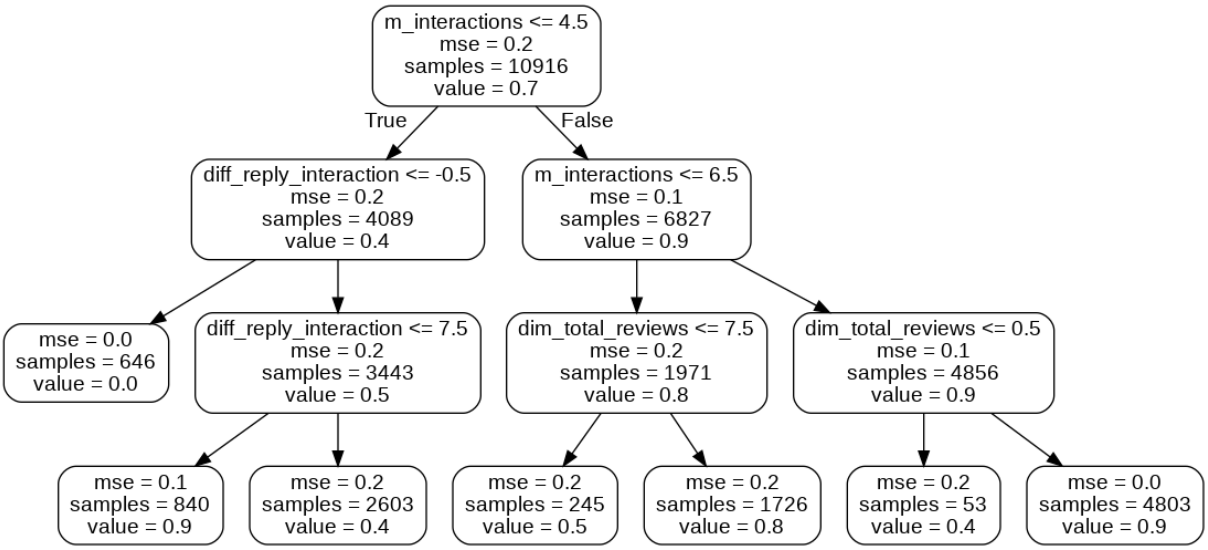
Random Decision Forest

Utilizo el algoritmo de random decision forest, con 1000 n-estimators, y un random-state de 42 para obtener las features más importantes.

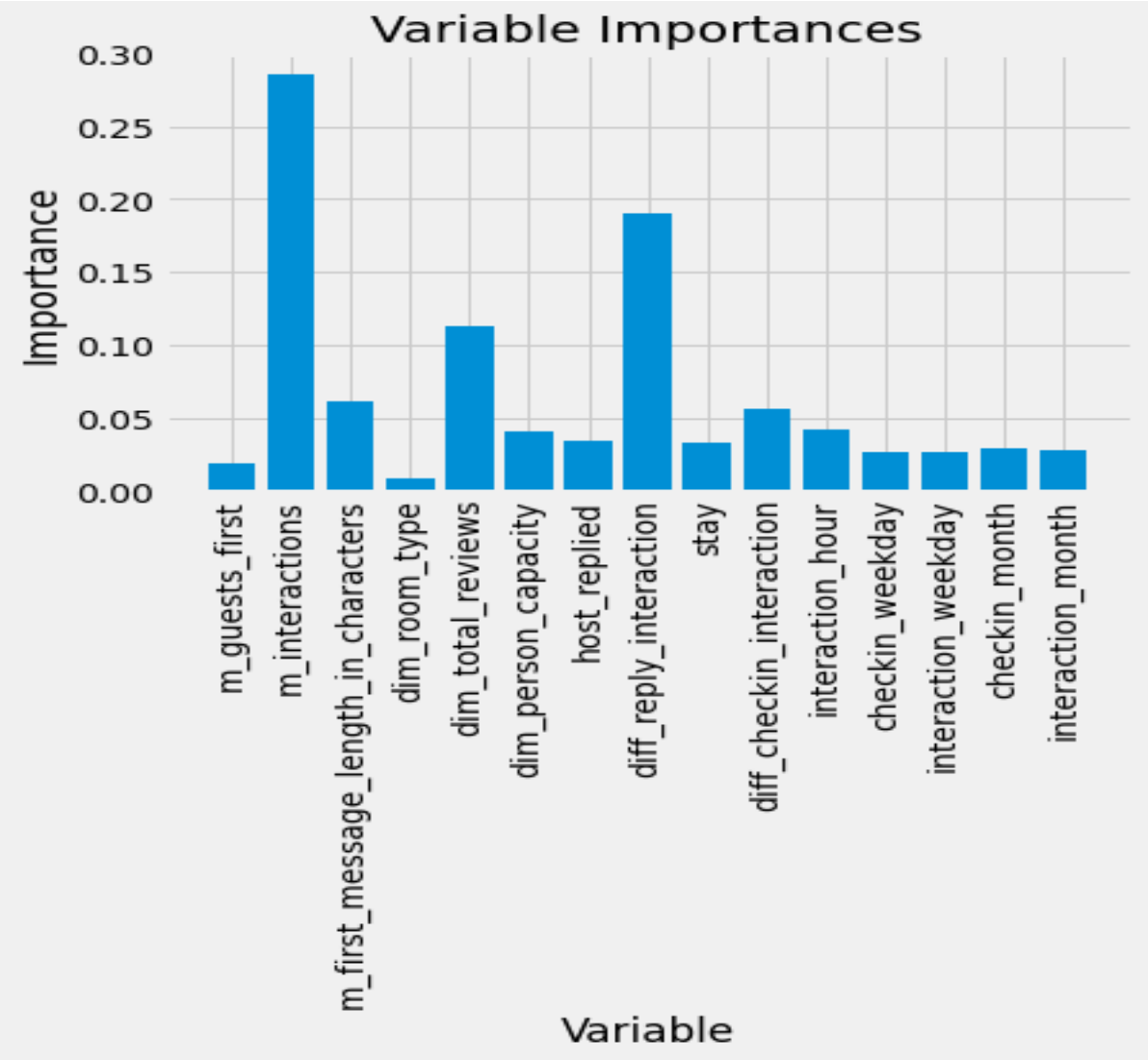
Obtengo:

- Mean absolute error: 0.2
- Mean Squared error: 0.1
- Root Mean Squared Error: 0.32

Limito el largo del árbol a 3 niveles, para visibilidad:



Y las features con mayor importancia según este modelo.



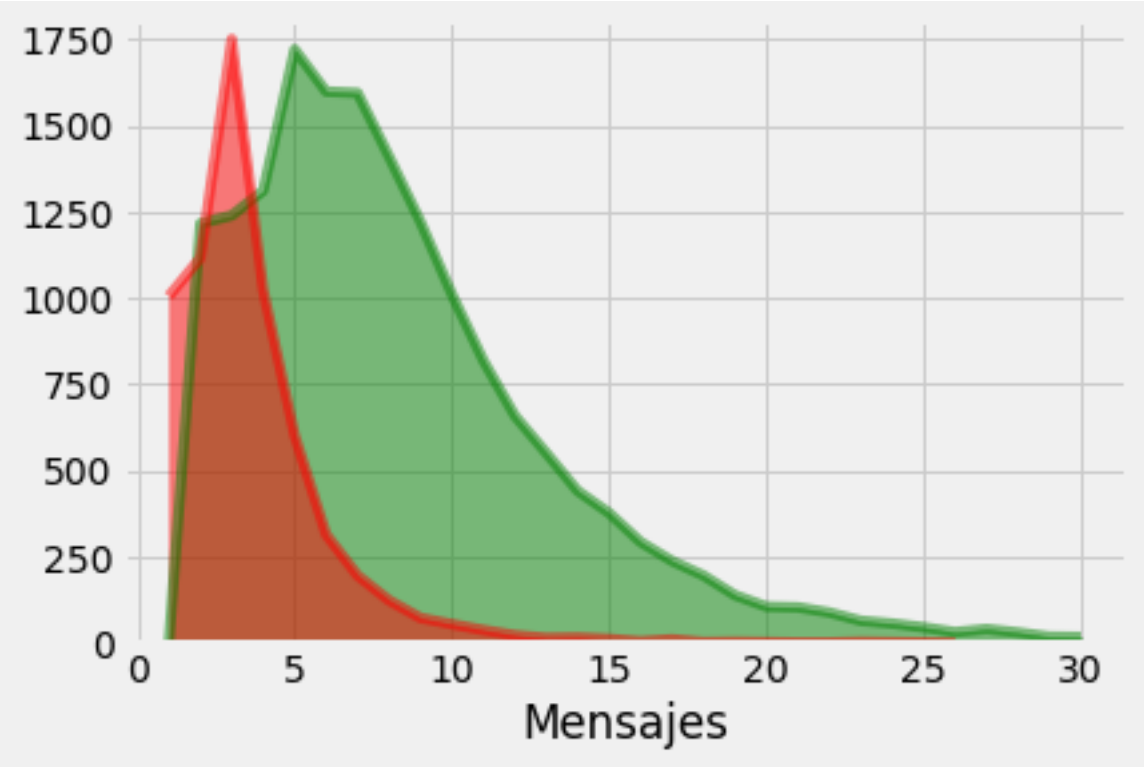
Con los valores:

Variable	Importancia
m_interactions	0.29
diff_reply_interaction	0.19
dim_total_reviews	0.11
m_first_message_length_in_characters	0.06
diff_checkin_interaction	0.06
dim_person_capacity	0.04
host_replied	0.04
interaction_hour	0.04
stay	0.03
checkin_weekday	0.03
interaction_weekday	0.03
checkin_month	0.03
interaction_month	0.03
m_guests_first	0.02
dim_room_type	0.01

Las variables más importantes según el modelo son, la cantidad de mensajes que se envían entre guest y host, el tiempo que tarda el host en responder, las reviews del listing, el largo de los mensajes y el delta entre el checkin y el momento de la primera interacción.

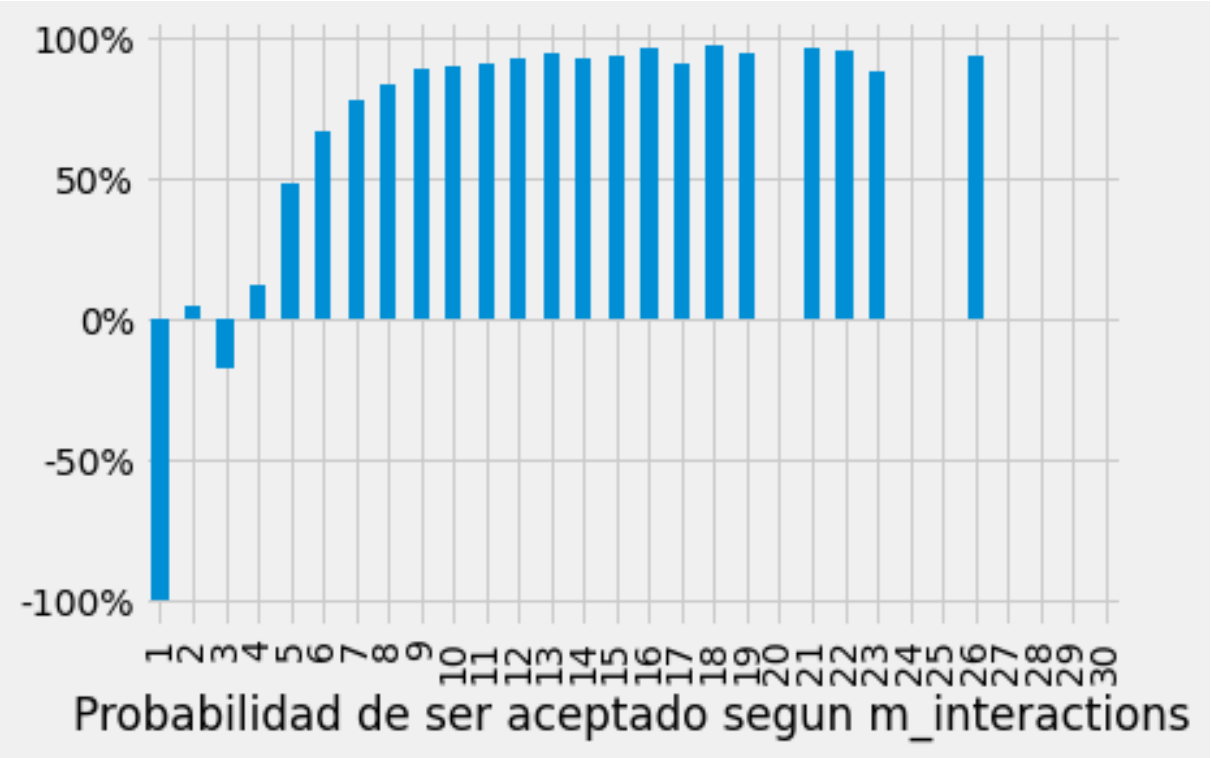
Análisis de las features importantes

1. m_inteactions



Aceptación con respecto a la cantidad de mensajes enviados

Un huésped que es aceptado intercambia en promedio 8.3 mensajes, mientras que un usuario denegado intercambia en promedio 3.59 mensajes.
La moda de la curva verde es de 5, y la de la curva roja de 3.
75% de los usuarios aceptados interactúa menos de 11 mensajes en total.
En cuanto a los huéspedes no aceptados, 75% intercambia menos de 4 mensajes.

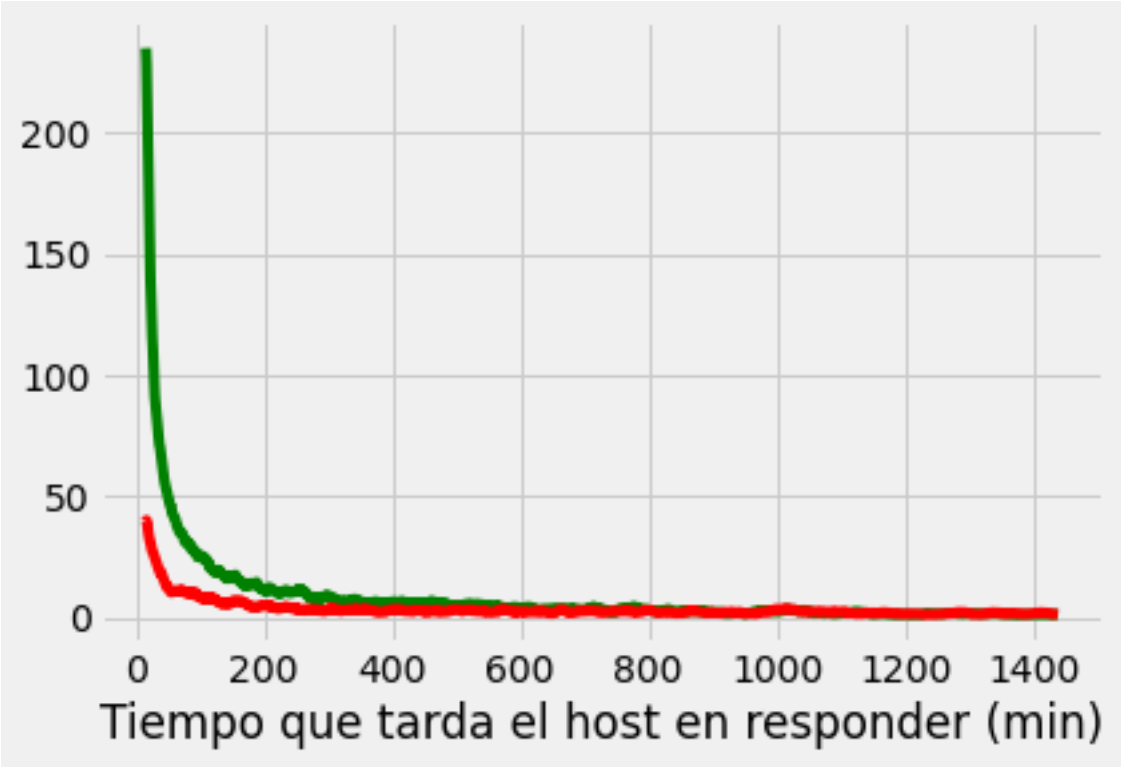


Es lógico que si únicamente un mensaje es enviado en la conversación, el guest sea rechazado.

Un usuario que interactúe al menos 7 veces cos el host, tiene más de 75% de probabilidades de ser aceptado.

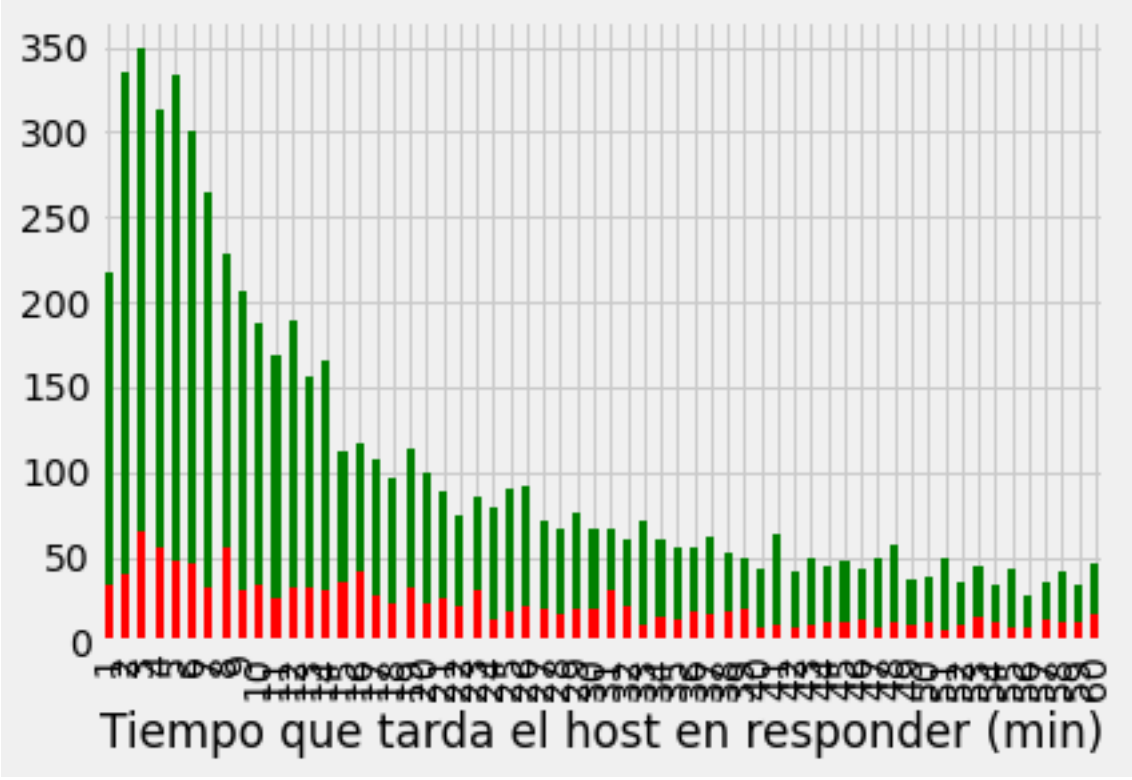
Si se interactúa exactamente 3 veces con el host, hay un 20% de probabilidad de no ser aceptado.

2. diff_reply_interaction

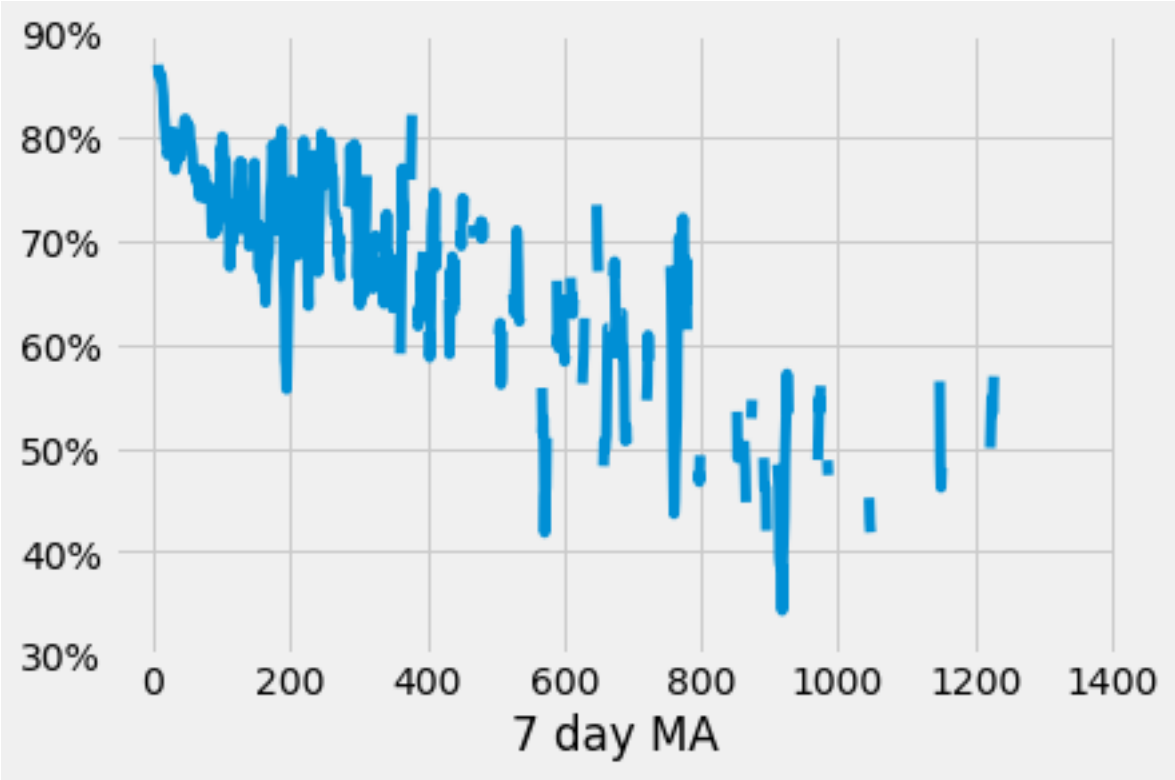


Usuarios aceptados o rechazados según el tiempo que tardó el host en responder.

La moda de las dos curvas es de 3 minutos
75% de los usuarios aceptados tardaron menos de 309 minutos en recibir una respuesta.
Mientras que 75% de los rechazados tardaron más de 53 minutos en recibir una.



Usuarios aceptados o rechazados a los que les respondieron en una hora o menos

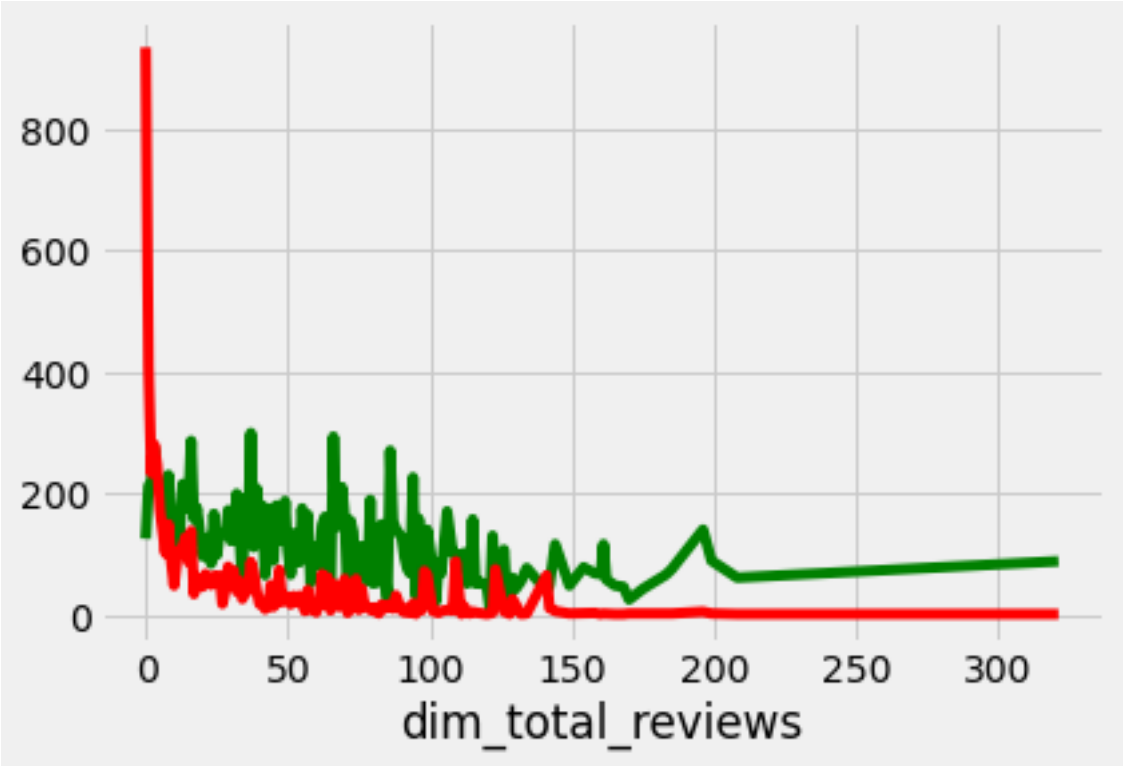


Probabilidad de ser aceptado según cuando tarda el host en responder (minutos, media móvil simple de 7 ***minutos (no días)***

Usuarios aceptados/rechazados si el host tarda exactamente 1 minuto en respondes.

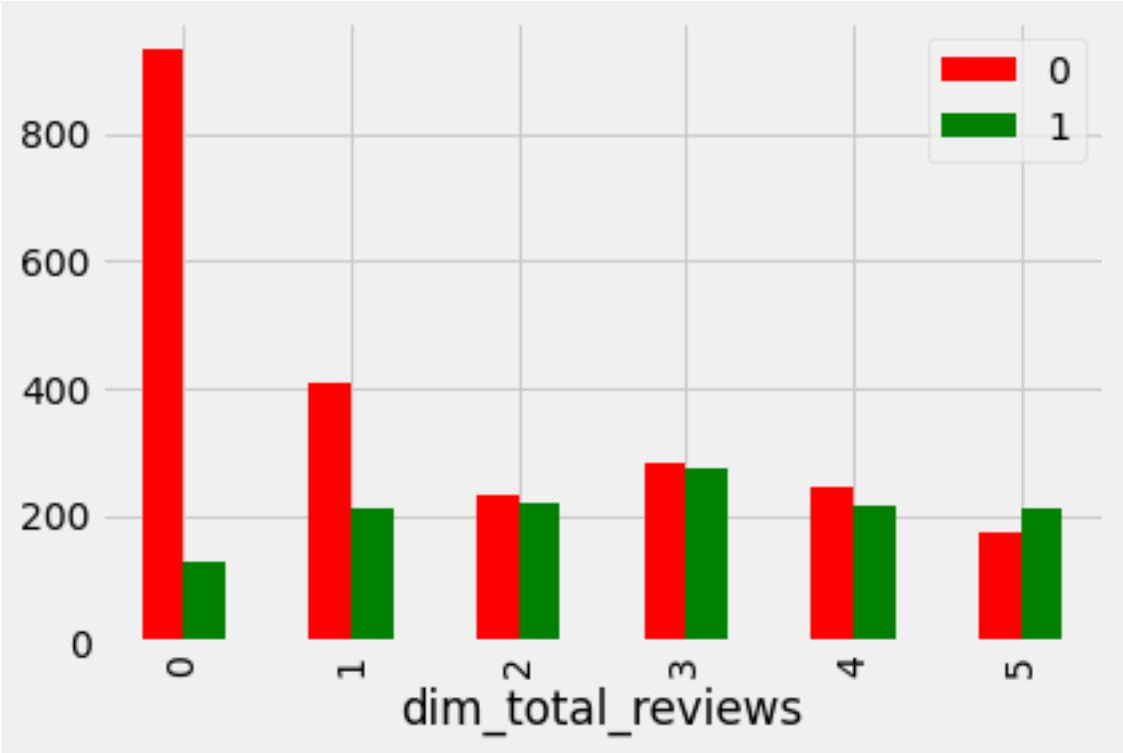
88.8% de los guests a los que le respondieron en menos de 1 minuto fueron aceptados.
Y 99.6% de los hosts que respondieron casi inmediatamente aceptaron al huésped.

3. dim_total_reviews



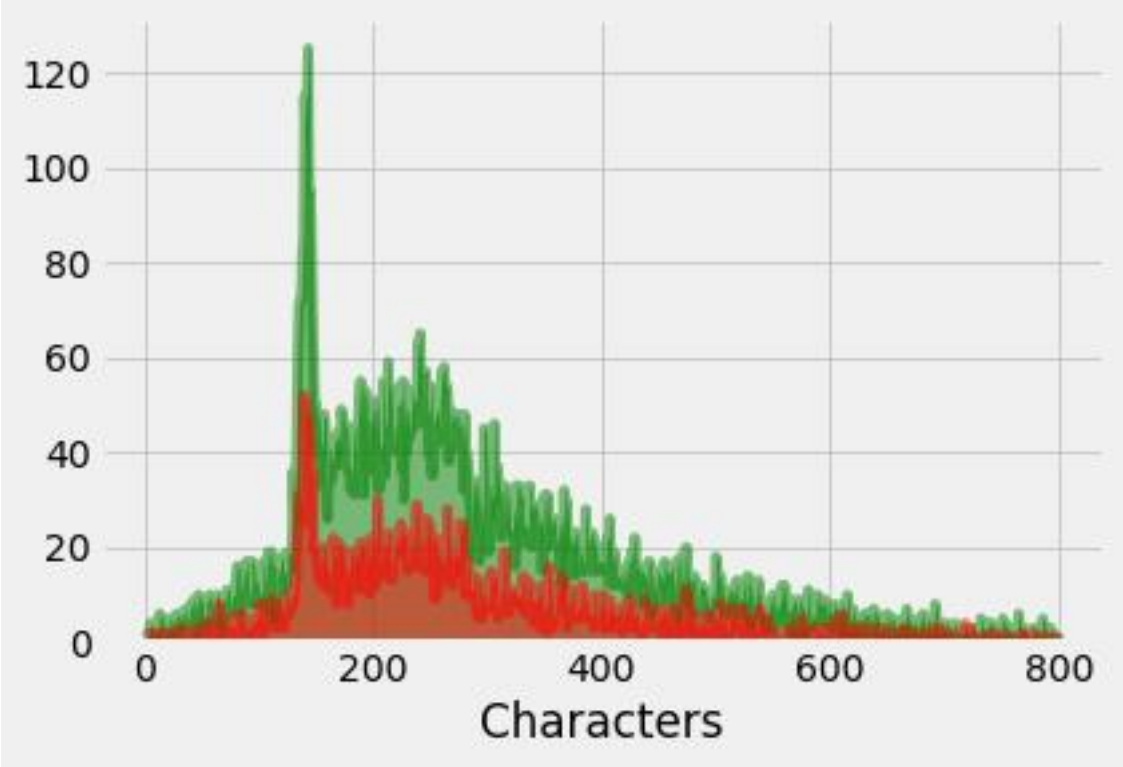
Usuarios aceptados o rechazado según la cantidad de reviews del listing.

75% de los usuarios son rechazados en listings de menos de 46 reviews, mientras que 75% de los aceptados, lo son en listings de más de 24 reviews.



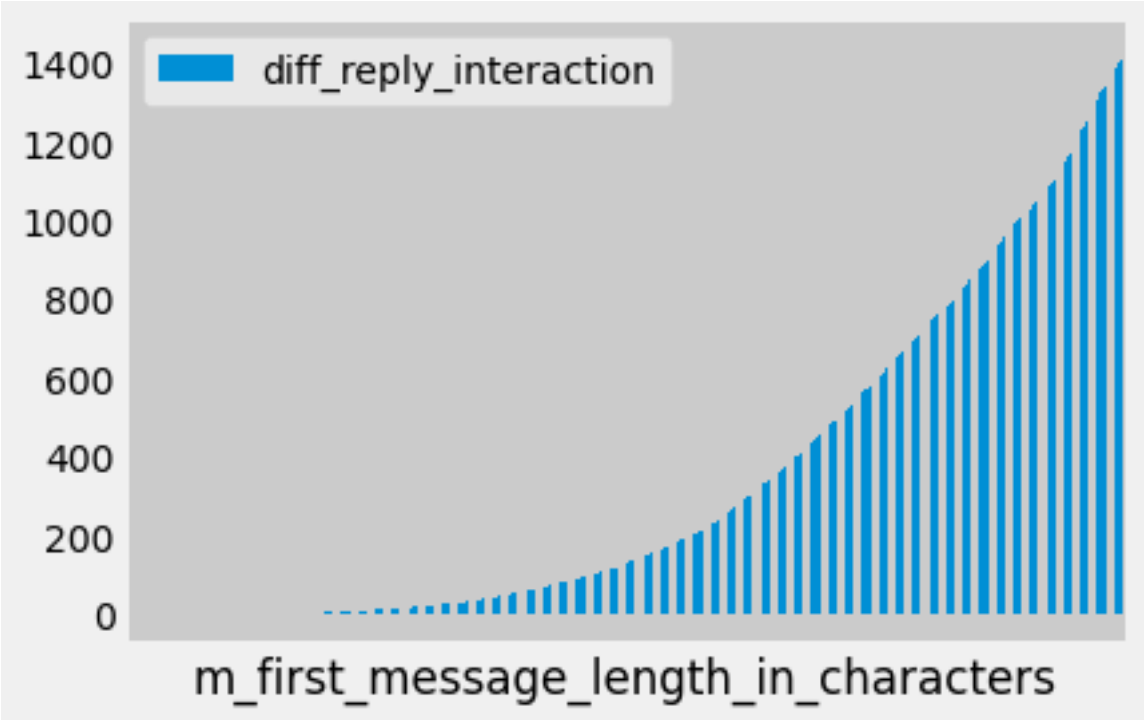
Usuarios aceptados o rechazado según listings con menos de 5 reviews.

4. m_first_message_length_in_characters



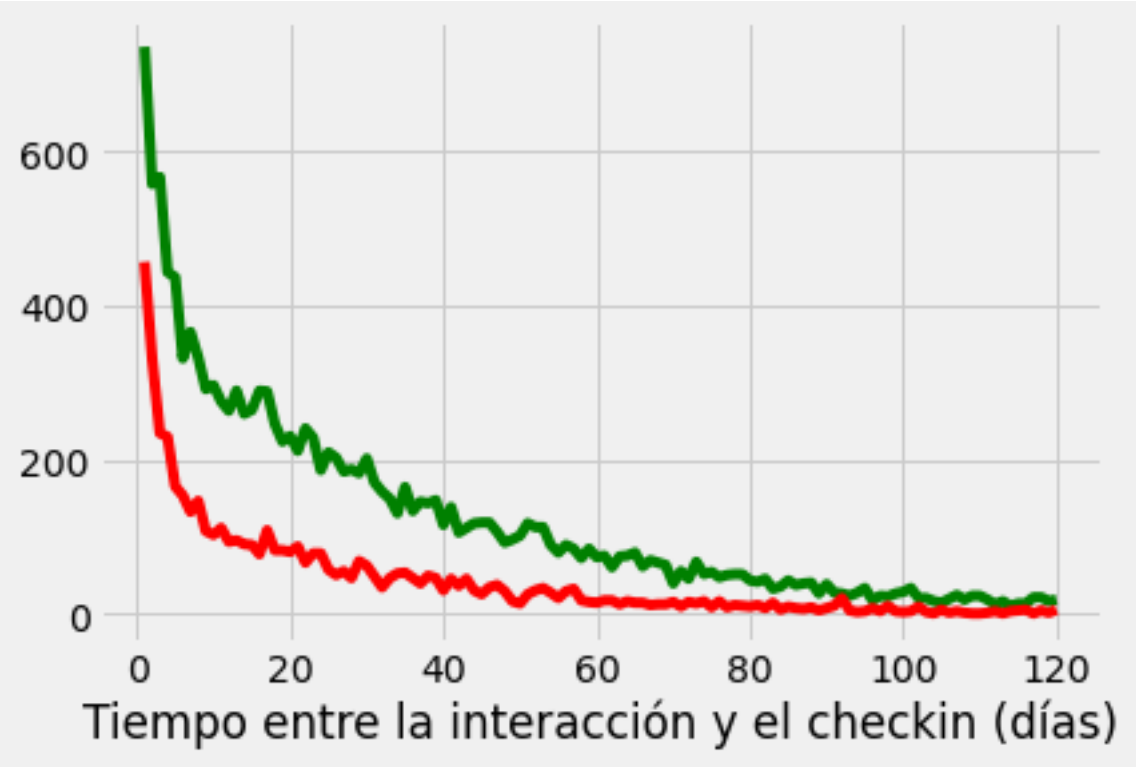
Cantidad de huéspedes aceptados o rechazados según número de caracteres del primer mensaje enviado.

Las dos curvas son muy similares. Existe un pico de usuarios que envían 140 caracteres. Es probable que sea ocasionado por algún cambio en la plataforma.

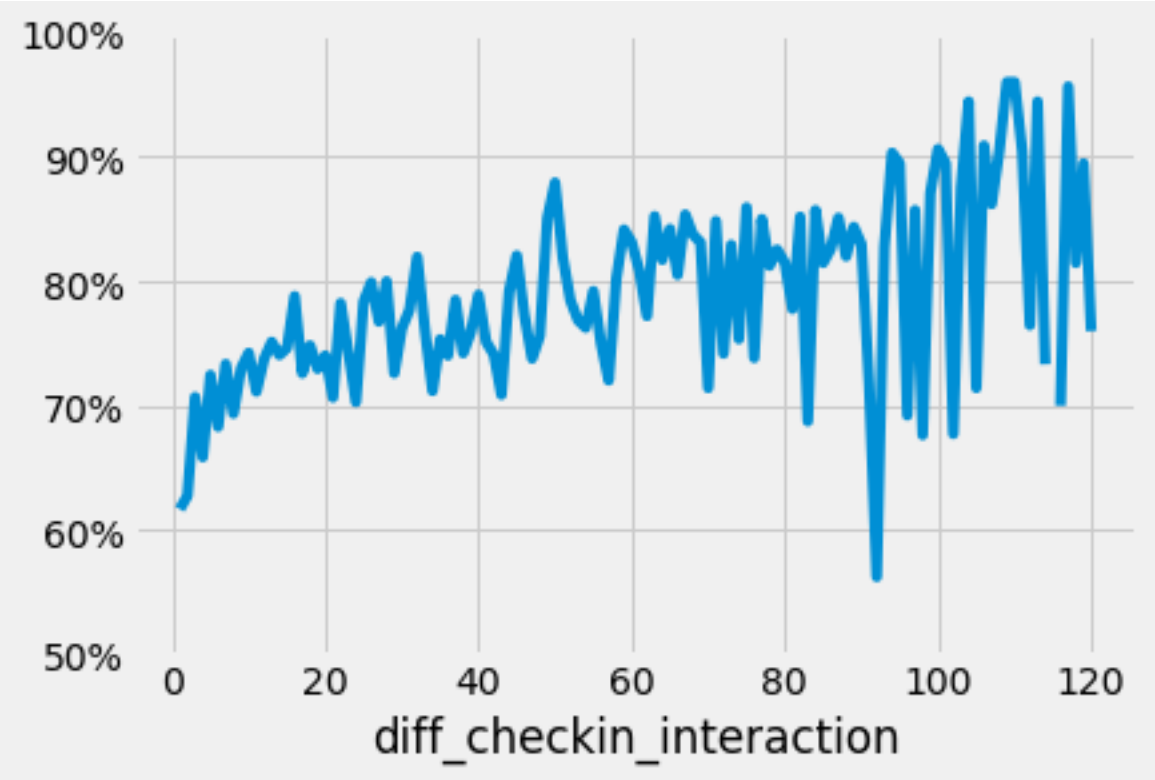


Sin embargo, hay una clara correlación entre cuanto tarda en responder un host y la cantidad de caracteres que hay en el primer mensaje, de aquí la importancia de esta variable.

5. diff_checkin_interaction

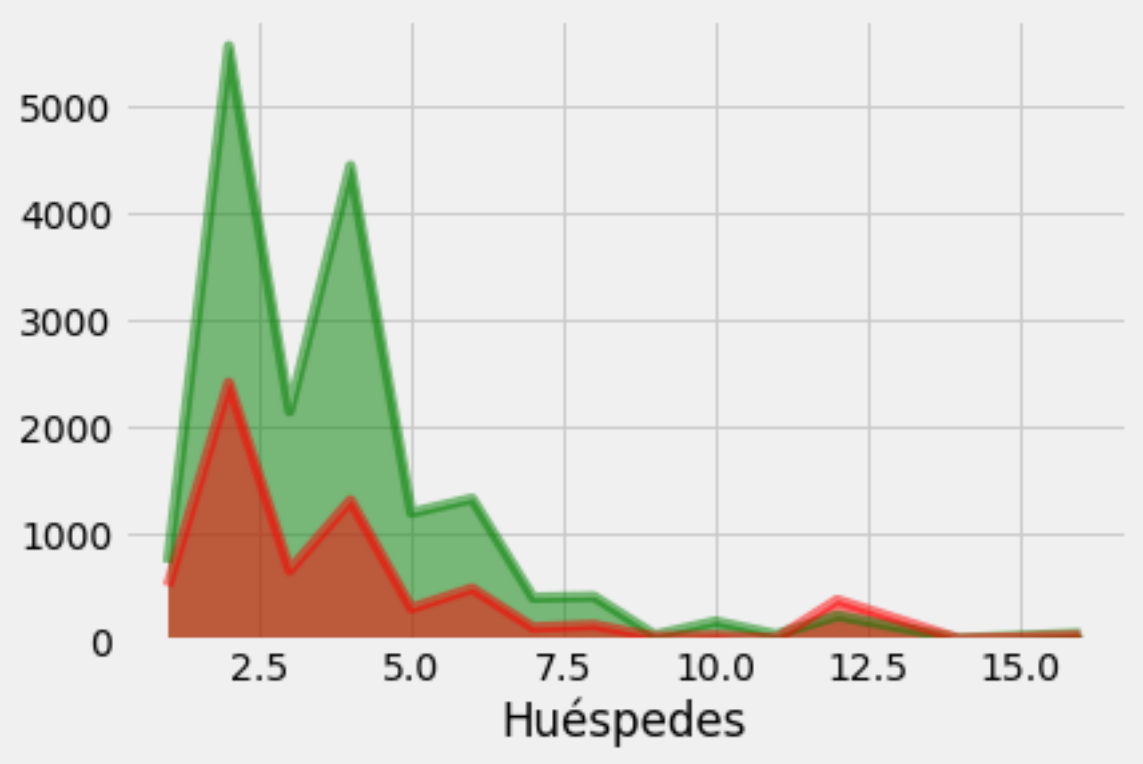


Pocos usuarios solicitan un checkin con más de 2 meses de anticipo, 75% de los usuarios lo hacen con menos de 45 días. Sin embargo, es más probable que el usuario sea aceptado si esta diferencia de tiempo es mas grande (ver gráfico siguiente).

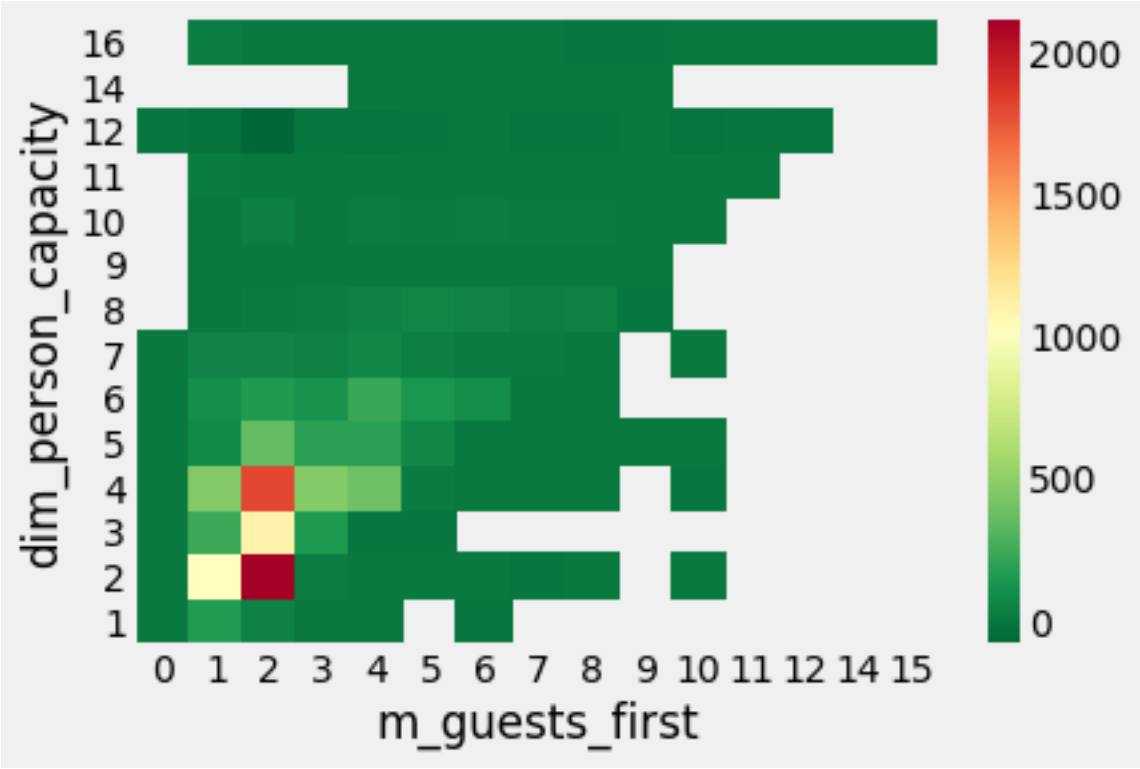


Probabilidad de ser aceptado según cuanto antes se hace la inquiry con respecto al checkin.

6. dim_person_capacity



Cantidad de usuarios aceptados y rechazados según el número de huespedes que desean alojar.



Usuarios aceptados o rechazados según la cantidad de huéspedes de la inquiry y la capacidad del listing.

Se puede observar un hotspot de usuarios aceptados cuando la cantidad de huéspedes y la capacidad del listing están cerca de 2, lo que indica una relación entre estas tres variables.

Recomendaciones para aumentar la probabilidad de aceptación

Basándonos en la primera feature (m_interactions), vemos que entre más mensajes se envíen entre el guest y el host, es más probable que sea aceptado.
Por lo que una opción puede ser el implementar un tipo de respuesta rápida dentro del chat. Por ejemplo, del lado del guest, recomendar mensajes de introducción, que autocompleten rápidamente con información del usuario, para facilitar y agilizar la conversación.
Del lado del host, se pudieran sugerir preguntas que los anfitriones hacen usualmente, e.g. "¿Cuál es el propósito de su viaje?", "¿Es primera vez que viene a esta ciudad?", "¿Es usted fumador?", etc.
De esta manera, se agiliza la conversación entre ambas partes, y se aumenta la probabilidad que el huésped sea aceptado.

A/B testing

No todos los sujetos de prueba han mandado mensajes de no menos de 140 caracteres, por lo que los elimino de la tabla.

Primero, buscamos la tasa de aceptación de cada grupo, la desviación estándar y el *standard error of the proportion* ($\frac{\sigma}{\sqrt{n}}$).

ab_y	acceptance_rate	std_deviation	std_error
control	0.728	0.445	0.004
treatment	0.741	0.438	0.004

Tabla mostrando los valores de la tasa de aceptación, desviación estándar y el standard error of the proportion

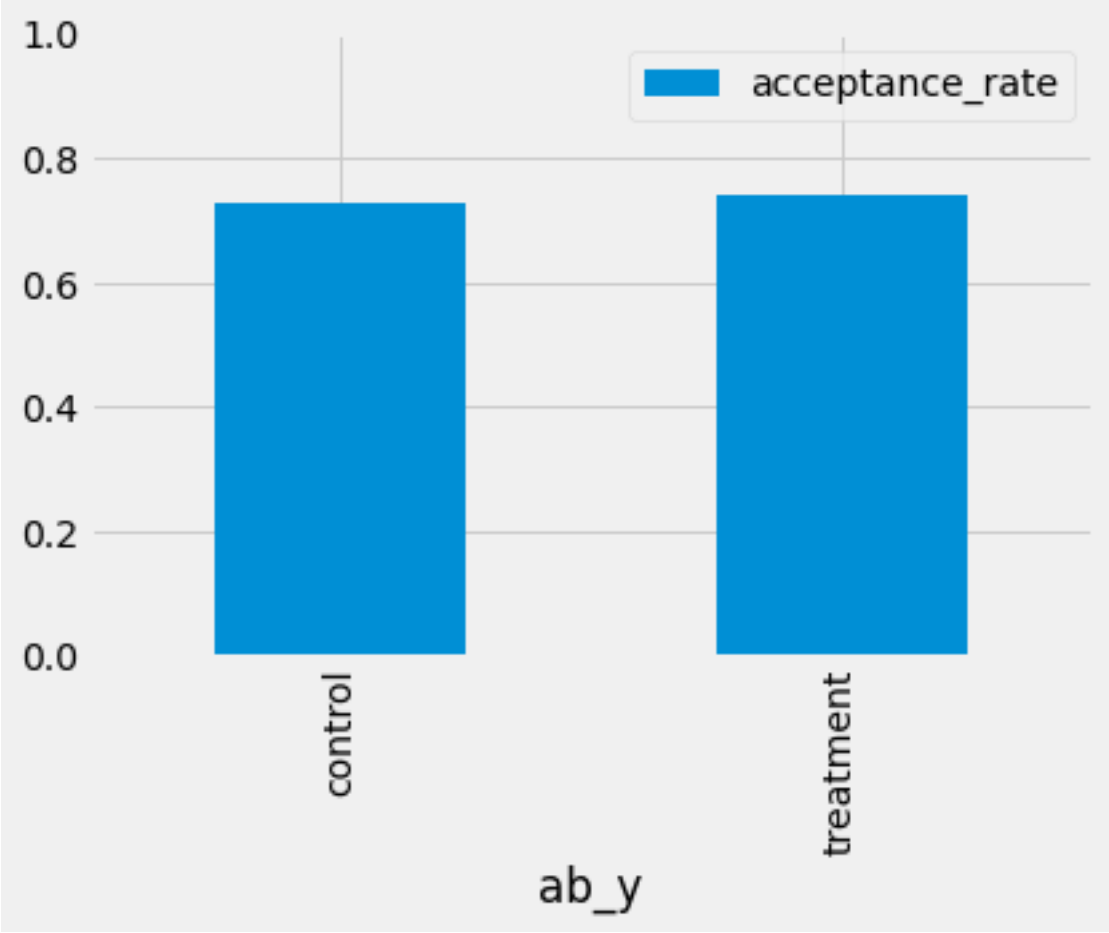


Grafico de tasa de aceptación del grupo de control con respecto a la del grupo de prueba.

Para test la hipótesis, uso las herramientas de la librería statsmodels (proportions_ztest, proportion_confint).
Tomo como nivel de confianza el valos $\alpha = 0.05$

El p-value del test es de 0.026, que es relativamente bajo, lo que nos permite rechazar la hipótesis nula. Hay menos de 5% de probabilidad que la hipótesis nula (que el experimento no tenga ningún efecto) sea correcta.

El intervalo de confianza para el grupo de control es [0.72, 0.735], y para el grupo de treatment es [0.732, 0.749].

Por lo tanto, en el caso de la hipótesis nula ser falsa, tendríamos como máximo (probabilísticamente) un aumento de 2.9% en la tasa de aceptación de los huéspedes.

La respuesta de si debe lanzarse depende del modelo de negocios que utilice la plataforma, y de los costos y el peso de las desventajas que trae implementar esto (por ejemplo, establecer un mínimo de caracteres es generalmente mala User Experience)
Si un aumento de aproximadamente 3% de tasa de aceptación de usuarios es suficiente para ser perceptible a la plataforma, entonces si debería lanzarse.
Además, el experimento no mostró disminución (más que un posible 0.3%) en la tasa de aceptación.