Nuño Sosa Victor Adrian

Plataformas de Hosting

En los últimos años, la industria turística ha vivido un cambio importante con el ingreso de las plataformas online (Booking, Trivago, Airbnb, Expedia, etc), que ayudan a comparar productos como rentas de lofts, departamentos, habitaciones de hotel, casas de fin de semana, etc. Para el consumidor, operador y para el inversionista de este tipo de productos, es conveniente conocer más a detalle el comportamiento de dicho mercado, esto se puede llevar a cabo con los datos que recopilan dichas plataformas.

En este proyecto, se presenta un análisis de este tipo de producto, tomando como base los datos generados por estas plataformas.

Datos

Los insumos son 3 tablas de con peso de 1.8Gb aprox. cada una, en la que se muestran, datos generales de los “rooms”, datos particulares y el historial de reviews de las mismas (comentarios en texto).

Realizando transformaciones y cruces de estas tablas, obtenemos una tabla analítica de datos final con las siguientes variables:

- Total de recamaras
- Total de baños
- Total de camas
- Tarifa de limpieza
- Número de reviews
- Número de reviews por mes
- Precio
- Review Scores Rating
- Review Scores Communication
- Review Scores Location
- Guests Included
- Política de cancelación
- Ciudad
- País
- Tipo de propiedad
- Tipo de habitación

Exploración Inicial

Dentro del conjunto de variables solamente se clasifican como discretas: Ciudad, País, Política de cancelación, Tipo de propiedad, Tipo de habitación, Número de camas, Número de habitaciones, Número de baños.

El restante se clasifica como continuas.

Muestreo

Se tiene un total de aprox. 40K productos con 37 variables, por lo que no es necesario realizar muestreo para carga o para modelado de datos.

Pre-Ingeniería

Se realizan operaciones con las fechas de: primer review, último review, última actualización y antigüedad del host, con la finalidad de determinar las siguientes variables:

- Tiempo del producto en plataforma
- Tiempo del host en plataforma
- Tiempo desde la última actualización
- Tiempo desde la última visita.

Análisis Exploratorio Discreto

Normalización

Todas las variables discretas requirieron de normalización salvo el número de camas.

Unarias

No existieron variables unarias.

Análisis Exploratorio Continuo

Missings

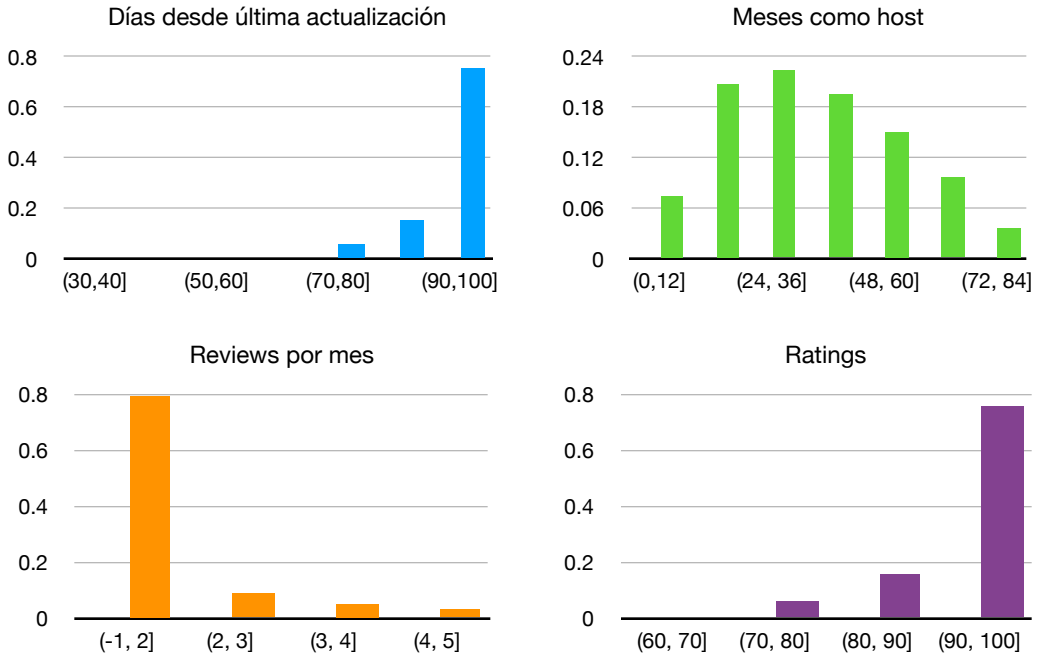
Se eliminó el 2% de las variables por contener valores nulos o ausentes.

Multicolinealidad

Aplicando el método de VarClus se obtienen como variables más representativas: Días desde la última actualización, Antigüedad del host en plataforma, Rooms rentadas por el host y Número de reviews por mes.

Univariado

Se muestran los histogramas de las variables obtenidas por varclus:



Outliers

La tasa de outliers es baja, se eliminan el 4.5% de registros por ser valores extremos.

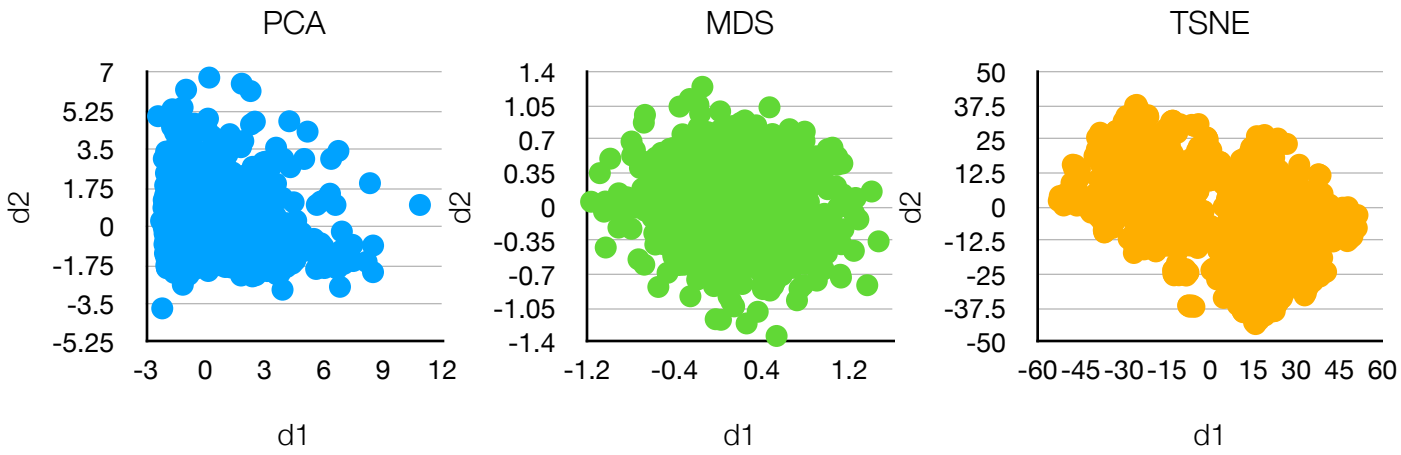
Modelación No Supervisada

Dada la tabla analítica de datos que se ha construido, se toma una muestra del 3% para lograr representatividad y las transformaciones no consuman complejidad computacional en demasía.

Definición de Espacios

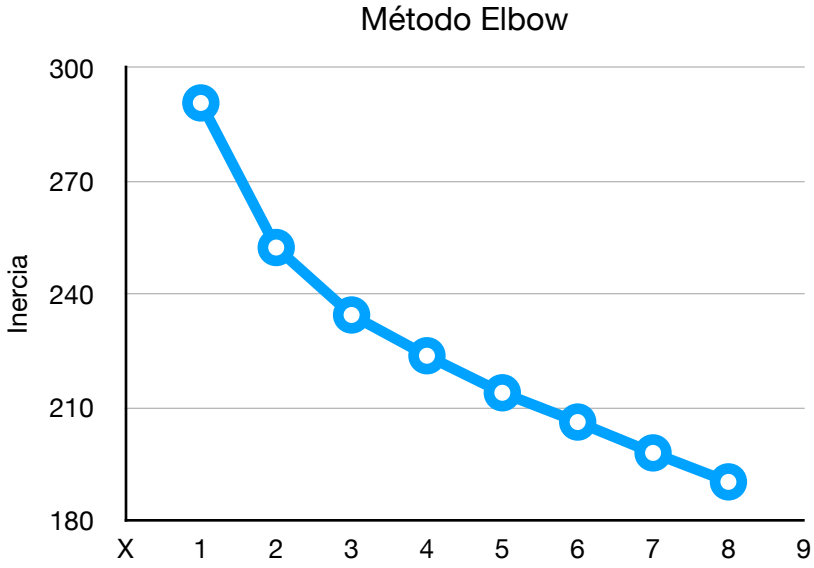
Se realizan las 3 transformaciones de datos correspondientes para visualización en dos dimensiones: Componentes principales (PCA), Escalamiento multidimensional (MDS) e Incrustación de vecinos cercanos (TSNE), ademas de la hipercaja o datos crudos. Dentro de estas transformaciones, se realiza imputación de valores ausentes con la mediana y procesos de escalación Standard y MinMaxScaler, según corresponde.

Visualización



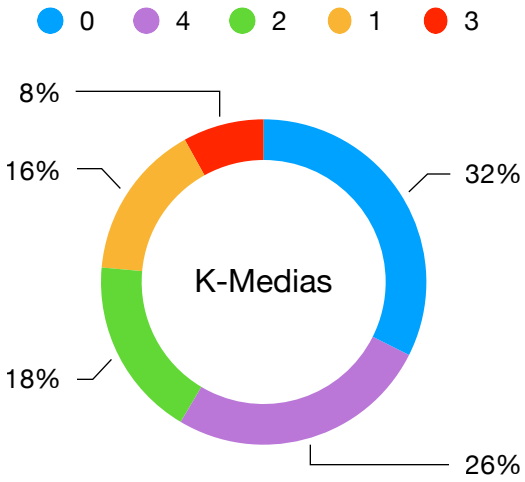
Selección del número de clústers

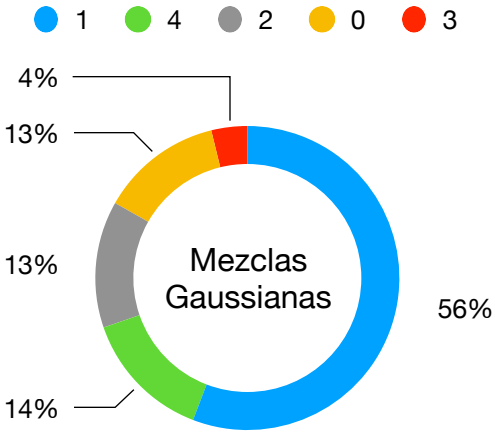
A simple vista, las transformaciones de los datos no dan una idea tan clara de cuántos ilustres debemos elegir, por lo que mostramos la gráfica con el método de Elbow (Momento de Inercia - Técnica del Codo)



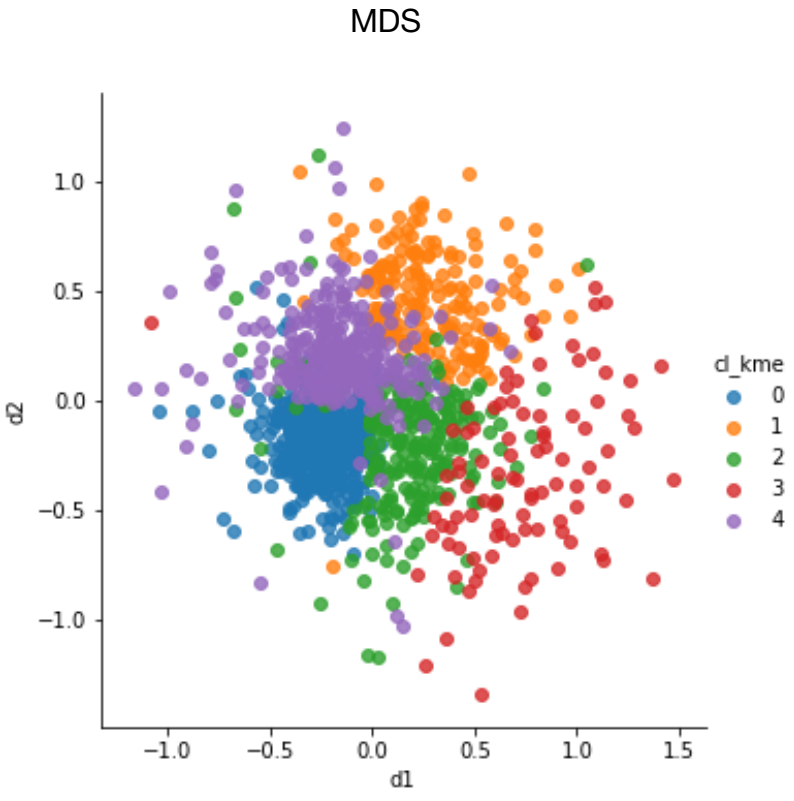
Modelación

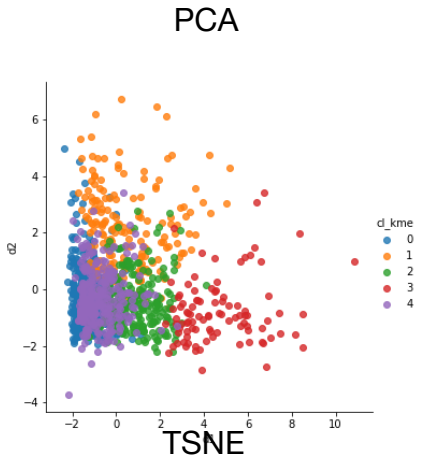
Se realizan los modelos por optimización (K-Medias) y por el método difuso (Mezclas Gaussianas).

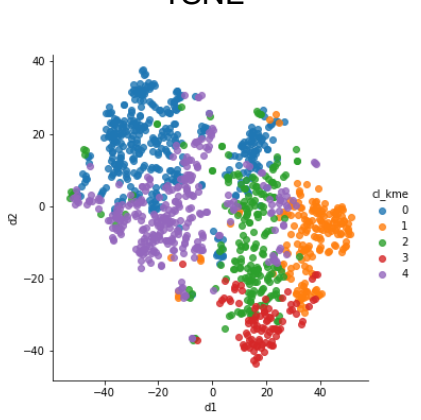




Visualización de Clusters







Significancia estadística para perfilar (Kruskal)

Todas las variables mantienen buena significancia estadística para perfilar.

Pruebas Post-Hoc (Tukey)

Los clusters mantienen buen nivel de diferenciación entre sí para todas las variables del perfilamiento.

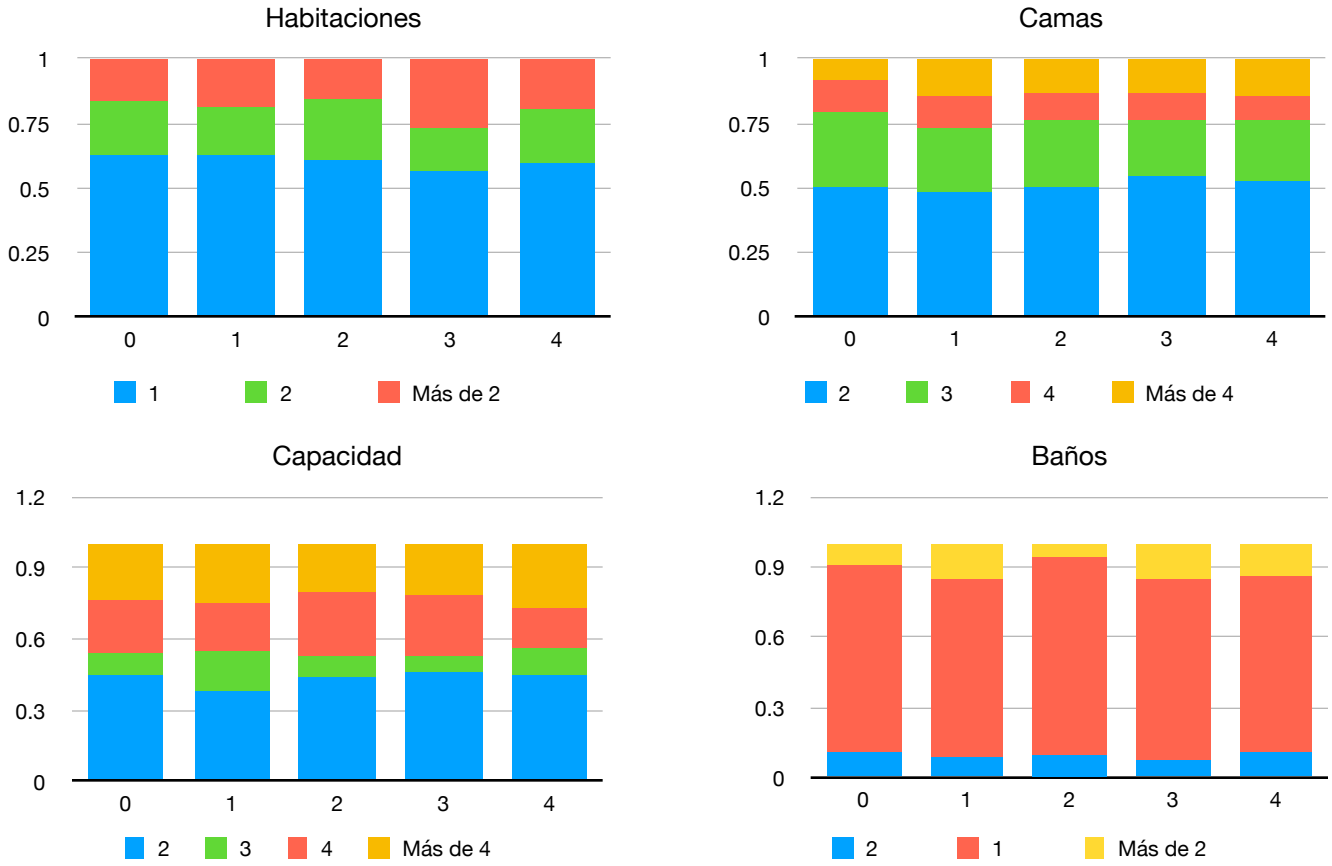
Perfilamiento

Cluster	Reviews	Score General	Score Descripción	Score Limpieza	Score Comunicación	Valor vs Pago	Meses de producto en plataforma	Meses de host en plataforma
0	18.9	97.3	9.9	9.8	10	9.8	11.7	19.9
1	76.4	93.8	9.6	9.5	9.8	9.4	45.7	59.2
2	16.9	91	9.4	9.2	9.8	9	19.3	31
3	9.7	81.4	8.4	8.2	8.9	8.3	22.3	38.8
4	12.5	97.8	9.9	9.8	10	9.8	20.4	48.1

Con las medias de cada variable, se pueden definir los cluster:

- **Cluster 0:** Va comenzando en la plataforma, ha mostrado ser profesional, su host no es tan experimentado pero tiene potencial.
- **Cluster 1:** Mucha experiencia dentro de la plataforma, ha logrado mantener un buen nivel de calificaciones - Altamente recomendado.
- **Cluster 2:** Tiene pocas visitas para el tiempo que lleva en plataforma, calificaciones medianas, también muestra muchas áreas de mejora.
- **Cluster 3:** Es el peor calificado de la plataforma, tiene pocas visitas por lo mismo a pesar de tener un buen host, el producto es malo.
- **Cluster 4:** Bueno en general, buenas calificaciones, sin embargo tiene un bajo ratio de visita, a pesar de tener un host muy experimentado. Es posiblemente un producto no atractivo para las masas.

Las variables discretas, muestran una distribución muy uniforme por cluster:



Lo cual no permite del todo extraer información extra para enriquecer el perfilamiento continuo, sin embargo, los siguientes ejercicios podrán mostrarnos mayor información.

Modelación Supervisada

Para realizar el análisis supervisado, se toma en cuenta una variable objetivo relacionada al número de visitas por mes, un número alto de visitas por mes indica un buen negocio en general para un producto de este tipo. Por ello, se busca encontrar el tipo de producto que esta dentro del 85% mas visitado. Para ello, utilizaremos un algoritmo de Scoring para permitir explicar con más detalle.

Procesamiento de Datos

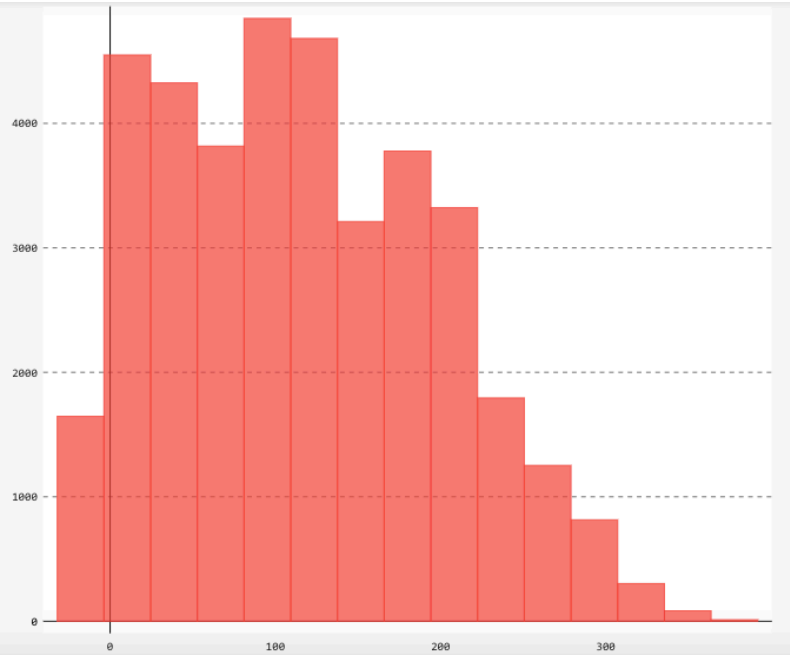
Inicialmente se discretean las variables continuas para hallar el Information Value (IV) para que con las mejores se realice la Transformación Entropica (WOE). Estas son las variables con mejor IV:

Variable	Information Value
Días desde la última visita	1.09876
Número de visitas	0.852451
Rating general	0.518596
Mínimo de noches	0.373854
Ciudad	0.275178
País	0.259982
Rooms rentadas por el host	0.155763
Anuncios del host	0.150407
Rating de limpieza	0.141083
Precio	0.138012

Modelación

Una vez teniendo lo anterior, se entrena una regresión logística, obteniendo un 85% de score.

Distribución del Score



Scorecard

Días desde última visita	Puntaje
De 0 a 2 meses	-30
De 4 a 6 meses	41
De 6 meses a 1 año	51
Más de 1 año	56
De 2 a 3 meses	-22
De 3 a 4 meses	2

Número de Visitas	Puntaje
1.0	64
2.0	58
3.0	54
4.0	47
Más de 4	-3

Rating General	Puntaje
De 1 a 85	26
De 86 a 90	11
De 90 a 95	4
De 96 a 97	0
De 98 a 99	-6
100	25

Anuncios del host	Puntaje
2.0	5
Al menos 1	13
Más de 2	3

Rooms rentadas por el host	Puntaje
1	16
2	1
3	-4
Más de 3	-2

Ciudad	Puntaje
Copenhagen	27
London	14
Los-angeles	5
Otros	8
Paris	15
Rome	3

País	Puntaje
Australia	14
Canada	9
France	19
Italy	-1
Others	17
Spain	0
United states	2

Mínimo de Noches	Puntaje
1.0	-2
2.0	0
3.0	13
4.0	47
Más de 4	80

Precio	Puntaje
De 0 a 50	9
De 100 a 200	6
De 201 a 500	23
De 51 a 99	2
Más de 500	57

Rating Limpieza	Puntaje
10.0	6
8 o menos	36
9.0	8

Negocio

Se define como corte los 50 puntos, es decir, las combinaciones que obtengan menos de 50 puntos son exitosas y las que obtengan mas de 50 puntos son no exitosas.

Se eligen tres combinaciones exitosas (están dentro del 85% más visitado de la plataforma), tomando en cuenta las variables que pueden ser decididas por el host-inversionista:

- Departamento en Roma, Italia con precio menor a 200 dls por noche, y uno o dos días como mínimo de renta.
- Departamento en Los Angeles, Estados Unidos con precio menor a 100 dls por noche, con un mínimo de una noche de renta.
- Departamento en España o Canadá, con precio de 50 a 100 dls y de preferencia mínimo una noche en renta.

Estas fueron selecciones que minimizan el puntaje basado en el producto y sus características, las variables que no son determinadas por el host-inversionista o que no son inherentes al producto en si y que ayudan a obtener éxito son:

- Cuidar que no pasen más de 3 meses sin visita.
- Cuidar obtener calificaciones en rating general mayores a 95
- De ser posible, que el host sea experimentado (que tenga varios anuncios en la plataforma y haya realizado un buen nivel de visitas a otros host)
- Muy importante cuidar la limpieza para obtener calificaciones mayores a 9.

Los siguientes modelos, nos permitirán enriquecer estos resultados.

Aprendizaje Profundo

Text Mining

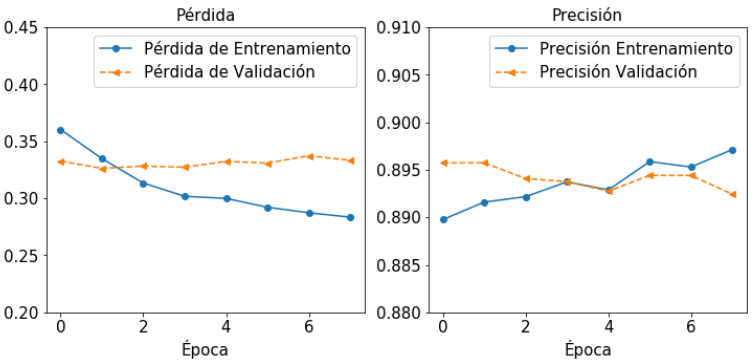
Para realizar un análisis más profundo de los productos, se puede analizar el texto de las reviews. Por lo que se propone un modelo en tensorflow utilizando la

arquitectura de memorias de largo plazo (LSTM) con un modelo de clasificación para los productos que tienen una calificación mayor al 90%.

Arquitectura

Layer	Output	Parametros/Activacion
Embedding	(165,24)	(14539,24)
Spatial Dropout	(165,24)	0.25
LSTM	100	0.5
Dropout	100	0.2
Dense	1	Sigmoidal

Performance



Metrica	Train	Test
ROC	0.774	0.767
ACC	0.895	0.900

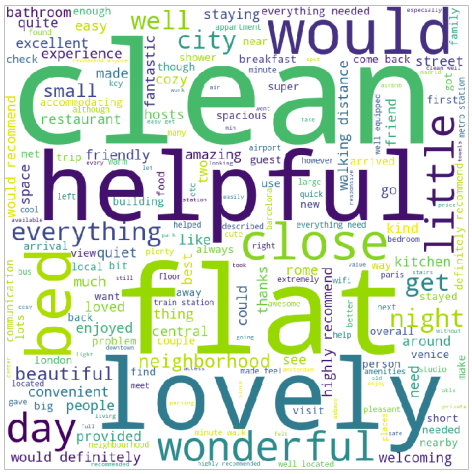
Análisis

En la primer limpieza de las nubes de palabras se obtiene que para las calificaciones mayores a 90, tiene muy marcadas las palabras: beautiful, lovely, wonderful, clean, helpful, neighborhood, amazing, city, would definitely, flat, highly recommend, everything, walking distance, enjoyed, etc.

Mientras que las calificaciones menores a 90 tienen las palabras: clean, helpful, flat, bed, little, would, close.

Score mayor a 90

Score menor a 90



De esta primer entrega, se aprecian las palabras de reconocimiento/confort de las calificaciones mayores a 90 (beautiful, amazing, lovely, wonderful, enjoyed, etc), también se nota con las palabras neighborhood o ciudad, la satisfacción del barrio o del rumbo (ubicación), el compuesto walking distance para hacer referencia a lo céntrico del producto o el compuesto would definitely para hacer afirmaciones futuras con seguridad.

Por su parte, las calificaciones menores a 90, muestran en gran medida la palabra clean, haciendo referencia a temas de limpieza, bed para referenciar la(s) cama(s), little para describir el tamaño de algo y a diferencia de el otro conjunto, el modal would que se asociaría a sugerencias o mejoras del producto.

Score mayor a 90

Score menor a 90

Posterior a una limpieza de las palabras mas notorias en la primer entrega, se obtienen estas dos nuevas nubes de palabras.

Entre las que se destaca para las calificaciones mayores a 90, friendly para describir la hospitalidad, highly recommend como una recomendación de primer nivel, spacious para mencionar el buen tamaño del producto, experience, para referenciar lo agradable del momento.

Por su parte, en las calificaciones menores a 90, se nota convenient, como un adjetivo mesurado, bathroom para hacer referencia al baño.

Esto nos puede dar a notar que la gente que califica bien, no hace referencia a temas inherentes al producto como el baño o las recamaras, que valoran mucho la ubicación o lo céntrico del producto y también la atención del host. Finalmente que es muy posible que te califiquen mal por temas de limpieza.

Clasificación

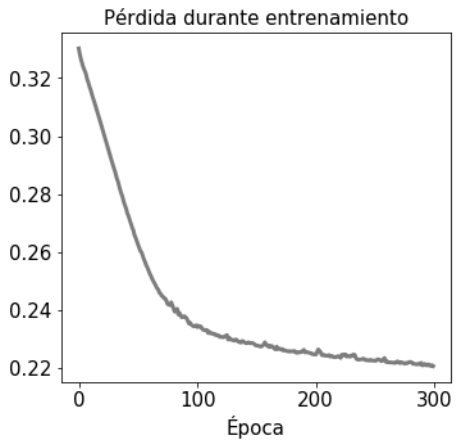
Una vez que se ha explorado el texto de las reviews de cada room, es importante también tratar de replicar este ejercicio pero con nuestra tabla analítica de datos.

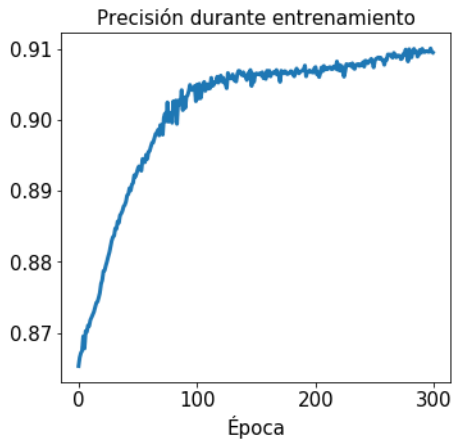
Para ello se propone el mismo target (Calificaciones mayores a 90%).

Arquitectura y Performance

Capa	Activación	Nodos
1	relu	10
2	tanh	40
3	relu	25
4	sigmoid	1

Metrica	Train	Test
ACC	0.91	0.90
ROC	0.94	0.94

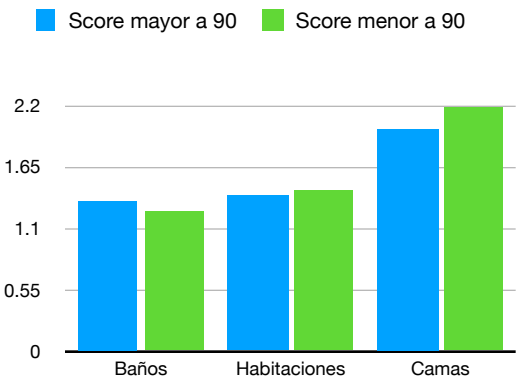


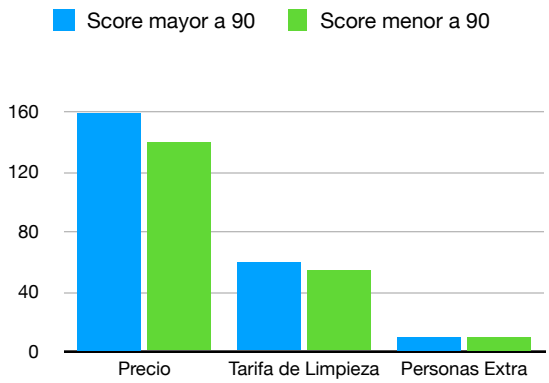


Análisis

Las siguientes gráficas muestran el comportamiento medio de las variables más significativas por clasificación. Lo que permite ver que hay habitaciones con score menor a 90 que tienen mayor número de camas, pero hay casi mismo número de habitaciones para ambos grupos. Lo que quiere decir que no necesariamente mayor número de camas se asocia con mejor calificación. En el caso de los baños se nota una leve mejoría si tienes más baños, un medio baño extra no estaría demás.

Finalmente, el grupo de los mejores calificados tienen mayor precio, lo que implica que al consumidor no le importa pagar más siempre y cuando lo valga, en el caso de la tarifa de limpieza pasa algo similar, hay mayor tarifa de limpieza en los mejores calificados, es decir: no importa si hay que pagar por la limpieza siempre y cuando este limpio (lo cual genera mejor calificación).





Conclusiones

Como se ha podido ver a lo largo del modelamiento de estos datos estructurados y no estructurados, podemos concluir que en definitiva, el paso general para que un room tenga éxito es cuidar el tema de los ratings, un producto bien calificado va a ser muy visitado y eso genera mejor negocio. Los perfilamientos sugieren comenzar de cero con buenas calificaciones y de preferencia con un host experimentado, ser muy enfático en el tema de la limpieza aunque se tenga que cobrar un fee para ello. Para el caso de los demás ratings, como lo son: descripción, comunicación y ubicación, es importante mantener sobre todo los dos últimos, los textos de los reviews revelan la gran importancia de que el room este bien ubicado geográficamente (respecto del entorno y a nivel barrio/colonia), y que el host tenga buena atención en el servicio.

Para el caso de las variables inherentes al room, es sugerible implementar una política de 1 o 2 noches mínimo, no mas, un precio asequible (menor a 200 dls por noche) y en caso de querer invertir en una propiedad a largo plazo como producto de estas plataformas, elegir una combinación ganadora de países y ciudades como las antes mostradas que cuide el tener recamaras suficientes pero no en demasía (un cuarto de tv o studio estaría excelente), hacer un esfuerzo porque tenga un medio baño extra a los que ya posee, así como tratar de que sea lo más espacioso posible.