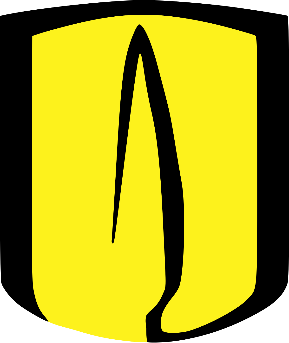
**INFORME PROYECTO ANALITICA DE TEXTOS**

**ANALISIS DE OPINION SOBRE HOTELES**

**Diego Alejandro González Vargas**

**Leonidas Villamil Pachón**

**David Burgos Méndez**

****

**Universidad de los Andes**

**2023-10**

**Introducción:**

El presente trabajo tiene como objetivo la puesta en práctica de conceptos de aprendizaje automático al servicio de problemáticas reales que se puedan ver beneficiados de los mismos. En este sentido, se considera de vital importancia la esquematización de las tareas realizadas:

* Identificación del problema & oportunidades de negocio
* Preparación de los datos
* Uso de modelos de aprendizaje automático, su corrección y mejoras.
* Exposición de resultados en términos de negocio

**Identificación del problema & oportunidades de negocio:**

|  |  |
| --- | --- |
| ELEMENTO | RESPUESTA |
| Oportunidad/problema Negocio | El problema de negocio identificado consiste en que los hoteles reciben cientos de comentarios y debido a la falta de tiempo o a la falta de personal no tienen el tiempo de atender y leer cada uno de ellos. No obstante, cada uno de los hoteles de la plataforma se encuentran fuertemente interesados en la mejora de cada uno de los elementos identificados como negativos dentro de sus servicios. En este sentido, se ve una oportunidad de negocio en un sistema que permita identificar aquellos comentarios que sean considerados como negativos, de tal modo que los hoteles puedan concentrarse solo en el análisis de estos y mejorar sus instalaciones |
| Enfoque analítico: Descripción del requerimiento desde el punto de vista de aprendizaje automático | Desde un punto de vista más técnico, se considera importante ver que esta es una labor predictiva, ya que lo que se busca es poder determinar la categoría de un dato. Adicionalmente, se debe mencionar que se cuenta con un conjunto de datos etiquetados(como positivos, negativos o neutros) que permitieron realizar el modelo. A partir de lo anterior, se puede inferir que es los modelos requieren de tareas de aprendizaje automático de tipo supervisado, ya que los datos tienen etiquetas. Así mismo, como cada una de estas etiquetas NO son numéricas y se trata por tanto de unas categorías discretas, se identifica que lo que se puede buscar es una tarea de aprendizaje de clasificación. |
| Organización y rol dentro de ella que se beneficia con la oportunidad definida | A partir de la labor realizada se pueden ver beneficiados los siguientes roles dentro de la organización:   * Hoteles: A partir de la identificación de los comentarios negativos que se realizan de cada uno de ellos, se pueden iniciar tareas correctivas y que busquen mejorar la calidad de los servicios de cada uno de los hoteles. * Administrador del sistema: Mediante la identificación de los hoteles con un mayor numero de comentarios negativos, se pueden refinar un poco más la organización de los hoteles en la exposición de los mismos en el landing page o en el momento de realizar búsquedas y aplicar filtros. |
| Técnicas y algoritmos a utilizar | Como se explico anteriormente, dado que se trata de un modelo predictivo que lo que busca es la predicción de la categoría de un texto a partir de las palabras utilizadas en el mismo, se proponen las siguientes estrategias:   * Árbol de Decisión: Esta técnica fue ampliamente utilizada en el desarrollo del curso y cumple a cabalidad cada una de las restricciones mencionadas anteriormente a partir del calculo de la influencia de cada uno de los datos en la variable objetivo, se ve afectado por factores (hiperparámetros) como la el método de organización: entropía o Gini, la profundidad del árbol y el mínimo de elementos necesarios en un nodo para poder realizar la división. * Regresión Logística: Esta es una técnica de análisis de datos predictiva que permite evaluar la pertenencia a una categoría de una variable a partir de una serie de factores definidos. Este es un modelo de regresión que está hecho para ser usado para clasificación binaria. El resultado de aplicación de un modelo de regresión logística es un número entre 0 y 1 que representa la probabilidad de pertenencia del elemento a la categoría evaluada. * SVM: SVM, o Support Vector Machine, es un algoritmo de aprendizaje automático que se utiliza comúnmente para la clasificación y la regresión. SVM funciona encontrando el hiperplano que mejor separa los datos en dos o más clases diferentes. El hiperplano es el límite de decisión que divide el espacio de características en dos regiones: una para cada clase. SVM busca el hiperplano óptimo que maximiza la separación entre las dos clases. La mayoría de los clasificadores de texto se basan en la representación de texto como vectores numéricos, que es la forma en que SVM maneja los datos. SVM es particularmente efectivo para lidiar con datos dispersos y ruidosos, lo que es común en los análisis de texto debido a la presencia de términos irrelevantes y ruidosos que pueden afectar el rendimiento de la clasificación. SVM también puede manejar múltiples clases, lo que lo hace adecuado para la clasificación de texto que involucra más de dos clases. Además, SVM tiene la capacidad de hacer frente a problemas de sobreajuste, que a menudo son un problema en la clasificación de texto debido a la alta dimensionalidad de los datos y la cantidad limitada de datos de entrenamiento. SVM puede lograr esto a través del uso de técnicas de regularización, como la restricción de la complejidad del modelo, lo que ayuda a evitar que el modelo se ajuste demasiado a los datos de entrenamiento. |

**Entendimiento y preparación de los datos:**

Para el análisis de cumplimiento de los objetivos de esta etapa se propone la consideración diferenciada de los resultados de las 2 etapas que la componen:

1. **Perfilamiento de datos**: Para poder llevar a cabo esta actividad, se considera el análisis de las características de los datos que tiene el negocio y sus características. En primer lugar, se debe considerar que los datos vienen en formato CSV(separado por comas), que una vez leídos se puede determinar que tienen dimensiones de 5000filasx6columnas, donde cada una de las filas corresponde a un comentario realizado por un cliente, y las columnas describen información del comentario como
   1. Título: encabezado del comentario
   2. Rating: calificación otorgada por el usuario de 0 a 5
   3. Texto: el contenido escrito por el usuario con sus percepciones del hotel.
   4. Ubicación: lugar donde se encuentra el hotel
   5. Nombre del hotel
   6. Etiqueta: Hace referencia a si el comentario es considerado positivo, negativo o neutro.

Adicionalmente, se consideró pertinente la evaluación de la existencia de registros duplicados, pero no se encontraron. Posteriormente, se evaluó la existencia de nulos en las variables de interés principal (texto y etiqueta) y se encontró que tampoco se tiene esta problemática para el conjunto de datos evaluados.

Así mismo se consideró sacar algunas estadísticas descriptivas de los comentarios, como sigue:

* + Conteo: Esta variable permite visualizar el número total de palabras en el comentario.
  + Conteo por palabra: Esta nueva columna permite conocer el número de repeticiones de cada palabra dentro del comentario.
  + Palabra más repetida: A partir de la información presentada en la columna de conteo por palabra, se puede hallar la palabra más repetida dentro de cada comentario.
  + Mínimo: hace referencia al mínimo de caracteres de las palabras en el comentario
  + Máximo: hace referencia al máximo de caracteres de las palabras en el comentario

Finalmente, se realizó una descripción de los datos de la columna de interés (“Etiqueta”) y se encontró de más del 50% de los datos se encuentran etiquetados como Buenos. Así mismo, se observó que la desviación estándar de los datos se encontraba sobre los 0,79 puntos.

1. **Preparación de los datos:** Para poder realizar una correcta preparación de los datos, teniendo en cuenta que se trataba de un proyecto de analítica de textos, se plantearon los siguientes pasos:
   1. No consideración de las demás columnas explicativas: al tener el objeto de estudio en el análisis de los comentarios, no se hace necesaria la consideración de las columnas de título, rating, locación, etc.
   2. Reemplazo de caracteres numéricos por sus palabras: Dado que los algoritmos de análisis de texto, específicamente las herramientas de stemming y lematización (explicados más adelante) no pueden interpretar número planos, se hace necesario el cambio de cada uno de ellos por las palabras que los representan.
   3. Convertir a minúsculas: Dado que el significado de una palabra no cambia de acuerdo a si cada uno de sus caracteres está en mayúscula o en minúscula, para facilitar las labores de comprensión del mensaje y reducir el número de términos diferentes a analizar, se consideró importante la homogenización de las palabras a minúscula.
   4. Remover puntuación: El objetivo de los signos de puntuación es dar sentido a cada una de las oraciones del texto de las contiene. Sin embargo, los temas tratados en el mismo no cambian. Así las cosas, para los objetivos planteados en este proyecto, y ayudando una vez más a la reducción del lenguaje a su base más fundamental, se consideró importante la eliminación de todos los signos de puntuación de los comentarios.
   5. Remover tildes & ñ’s: dado que algunas de las herramientas utilizadas no pueden comprender caracteres ascii, y en aras de no perder estos caracteres en su totalidad, se decidió convertir cada uno de ellos a sus caracteres más similares en ASCII imprimible: las letras vocales en minúscula y la n.
   6. Remover caracteres especiales: Como se explicó anteriormente, cada una de las herramientas utilizadas para el desarrollo de los modelos no comprenden caracteres fuera de los ASCII. Así las cosas, se hace una función que me permita normalizar las palabras limpiándolos de todos estos caracteres.
   7. Eliminación de stopwords: dentro del contexto de analítica de textos, las stopwords hacen referencia a las palabras comunes dentro de los textos que no aportan al significado de las mismas, pero si ensucian el análisis computacional de los mismos. Ejemplos de estas palabras son los artículos que acompañan los sustantivos. Para resolver este problema, se decidió la eliminación de estas palabras.
   8. Tokenización: Para poder entrenar un sistema que funciona con listas de palabras, se necesita cambiar cada uno de los comentarios por las listas de las palabras que lo componen. Esta transformación se le conoce como tokenización y fue contemplada dentro del proyecto
   9. Stemming: Cada una de las palabras tienen un significado específico dentro del español. Así mismo, cada palabra tiene muchas variaciones en las cuales mantiene su significado(como las conjugaciones de los verbos). En este sentido, se realizó una reducción del vocabulario mediante la transformación de cada una de las palabras a aun base fundamental.
   10. Lematización: El proceso de Stemming da como resultado bases fundamentales de cada una de las palabras del vocabulario. Sin embargo, cada una de estas palabras pueden NO corresponder a palabras reales del español. En este sentido, los resultados arrojados posteriormente por los modelos no podrán ser interpretados con facilidad. En este sentido, se hizo necesaria la traducción de cada una de estas bases a palabras reales mediante el proceso de lematización.
   11. Armado de la matriz: Una vez se tiene cada una de las palabras en sus bases fundamentales se arma un nuevo dataframe de pandas que tiene las siguientes características: Las columnas corresponden a cada uno de los términos resultantes de los 10 procesos anteriores. Y las filas corresponden a una matriz de repetición que indica en cada una de las casillas el número de existencias de la palabra en el comentario. En este sentido, cada una de las filas corresponde un comentario.
   12. Conversión de las variables explicativas del modelo: Dado que ni la representación binaria de las palabras ni la representación de las repeticiones de la misma en los comentarios son formas significativas de analizar la importancia de una palabras dentro de un comentario, y menos dentro del contexto de clasificación que se aborda, se propone la utilización de un modelo tfidf que calcule el peso de las palabras con respecto a su protagonismo dentro del comentario.

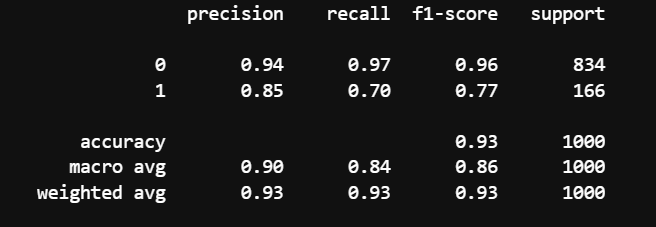
**RESUMEN DE RESULTADOS DE CADA UNO DE LOS MODELOS:**

**REGRESION LOGISTICA:**

TOP 60 PALABRAS MAS INFLUYENTES:

|  |  |
| --- | --- |
| LOS MAS NEGATIVOS | LOS MENOS NEGATIVOS |
| Atributos de mayor peso:  coeficientes caracteristicas  15347 8.941404 pesim  19079 8.301812 suci  14777 7.193221 pared  17889 6.664400 saban  258 6.618358 360o  5667 6.618358 cristobal  19086 6.401692 sucio  15353 5.788610 pesimo  10390 5.713847 horror  15143 5.538509 pequena  14516 5.353812 othels  6311 5.342946 des  6121 5.085786 deficient  2256 5.056166 asin  12581 5.051376 malisim  12583 5.039001 malisimo  3754 5.014911 car  6505 4.976696 desear  20624 4.943521 vending  7402 4.935846 educacion  11433 4.925861 jabon  17890 4.915990 sabana  11819 4.893131 lament  19566 4.873960 teni  6858 4.827941 dieciocho  6857 4.752698 diecioch  10389 4.734474 horrible  10388 4.734446 horribl  14856 4.659604 pasable  5742 4.635676 cuatro | Atributos de menor peso:  coeficientes caracteristicas  12124 -8.305515 limpio  4798 -8.099142 comod  1614 -7.264969 amplio  1606 -7.190908 ampli  15205 -6.735808 perfect  15209 -6.511150 perfecto  8625 -6.483618 excelent  8627 -6.457231 excelente  11442 -6.455201 jaen  12104 -6.440171 limpi  7653 -6.395587 encant  7257 -5.905734 dud  20036 -5.747600 tranquil  21122 -5.673265 wifi  12745 -5.653362 maravill  1506 -5.495771 amabl  9586 -5.372193 genial  20056 -5.327909 tranquilo  13188 -5.215485 mil  20270 -5.127103 ubicacion  977 -4.981255 agrad  8329 -4.977498 estacion  5735 -4.938576 cuarenta  4799 -4.865051 comoda  15300 -4.779587 personal  7263 -4.768605 dudar  4063 -4.716520 cerc  980 -4.636906 agradable  4064 -4.625998 cerca  9573 -4.450995 general |

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

*Imagen 1: Matriz de confusión Imagen 2: Estadísticas de la matriz de confusión*

|  |  |
| --- | --- |
| MAE: | RMSE: |
| Train: 0.3945714285714286  Test: 0.384 | Train: 0.8048069866211933  Test: 0.7916228058025279 |

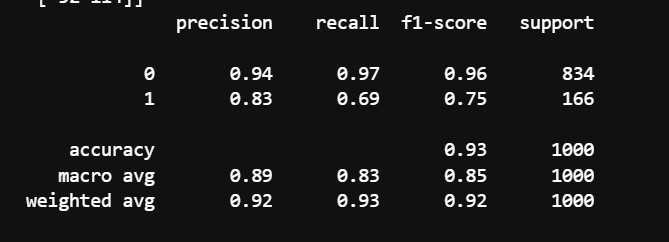
**SVM:**

TOP 60 PALABRAS MAS IMPORTANTES

|  |  |
| --- | --- |
| LOS MAS NEGATIVOS | LOS MENOS NEGATIVOS |
| coeficientes caracteristicas  19079 2.059576 suci  15347 1.769692 pesim  14777 1.725393 pared  19086 1.558407 sucio  15353 1.342498 pesimo  10390 1.301082 horror  17889 1.246254 saban  7402 1.238269 educacion  12227 1.164948 llegar  5764 1.132404 cucaracha  5763 1.132404 cucarach  6311 1.121134 des  12583 1.102125 malisimo  12581 1.101712 malisim  15143 1.094794 pequena  9215 1.087370 fot  9216 1.087370 foto  6067 1.070857 decir  17806 1.066935 rot  7198 1.064190 doscientos  14591 1.059720 pag  3754 1.058942 car  11433 1.056681 jabon  12546 1.053405 mal  3121 1.029723 booking  6121 1.028140 deficient  13598 1.023131 muell  19038 1.009395 sub  10388 1.000376 horribl  6505 0.999822 desear | coeficientes caracteristicas  12124 -2.132645 limpio  12104 -1.678751 limpi  4798 -1.531703 comod  15300 -1.345585 personal  11442 -1.322479 jaen  1614 -1.281776 amplio  15205 -1.275945 perfect  8625 -1.274339 excelent  8627 -1.267876 excelente  1606 -1.253103 ampli  15209 -1.230904 perfecto  7257 -1.208743 dud  9586 -1.208330 genial  1506 -1.204055 amabl  20270 -1.159329 ubicacion  7653 -1.158799 encant  20056 -1.156584 tranquilo  8329 -1.131726 estacion  21122 -1.127341 wifi  20036 -1.092655 tranquil  5735 -1.056165 cuarenta  14802 -1.020268 parking  980 -1.007709 agradable  13423 -1.002270 monitor  977 -0.992397 agrad  7263 -0.988900 dudar  12745 -0.984782 maravill  196 -0.938806 24h  18615 -0.927055 situ  4799 -0.903009 comod |

METRICAS DE CALIDAD:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente 

*Imagen 3: Matriz de confusión // Imagen 4: Estadísticas de la matriz de confusión*

Gráfico, Gráfico de líneas

Descripción generada automáticamente

*Imagen 5: Recall*

**ARBOLES DE DECISIÓN**

TOP 30 PALABRAS MAS INFLUYENTES:

|  |  |
| --- | --- |
| LOS MAS NEGATIVOS | LOS MENOS NEGATIVOS |
| Atributo Importancia  19079 suci 0.088874  12546 mal 0.054483  15347 pesim 0.048254  14591 pag 0.028224  6067 decir 0.018498  14777 pared 0.016993  16819 recepcion 0.015368  12124 limpio 0.013136  12227 llegar 0.012732  10437 hotel 0.011766  6926 diner 0.011340  6121 deficient 0.010861  10390 horror 0.010624  11820 lamentable 0.010390  12581 malisim 0.009353  12189 llam 0.009104  12619 manch 0.009027  9215 fot 0.008789  13794 nefasto 0.008339  6385 desayuno 0.008152  15142 pequen 0.007944  6858 dieciocho 0.007484  980 agradable 0.007478  6311 des 0.007435  11463 jam 0.007408  11170 insoport 0.007381  18892 sorpresa 0.007364  5764 cucaracha 0.007346  3806 carisim 0.007328  19566 teni 0.006921 | coeficientes caracteristicas  12124 -2.132645 limpio  12104 -1.678751 limpi  4798 -1.531703 comod  15300 -1.345585 personal  11442 -1.322479 jaen  1614 -1.281776 amplio  15205 -1.275945 perfect  8625 -1.274339 excelent  8627 -1.267876 excelente  1606 -1.253103 ampli  15209 -1.230904 perfecto  7257 -1.208743 dud  9586 -1.208330 genial  1506 -1.204055 amabl  20270 -1.159329 ubicacion  7653 -1.158799 encant  20056 -1.156584 tranquilo  8329 -1.131726 estacion  21122 -1.127341 wifi  20036 -1.092655 tranquil  5735 -1.056165 cuarenta  14802 -1.020268 parking  980 -1.007709 agradable  13423 -1.002270 monitor  977 -0.992397 agrad  7263 -0.988900 dudar  12745 -0.984782 maravill  196 -0.938806 24h  18615 -0.927055 situ  4799 -0.903009 comoda |

METRICAS DE CALIDAD:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamenteGráfico, Calendario

Descripción generada automáticamente

*Imagen 3: Matriz de confusión // Imagen 4: Estadísticas de la matriz de confusión*

**EVALUACION DE LOS MODELOS EN LENGUAJE NATURAL**

A partir del resumen de resultados mostrado, se puede mencionar las siguientes evaluaciones:

* Si se utiliza el modelo de regresión logística, se tendrá una exactitud del 93%, lo que se traduce en que para cada 100 datos, 93 de ellos serán correctamente clasificados como positivos o negativos. Para el caso de SVM, se tiene la misma cantidad de aciertos. Por otra parte, arboles de decisión no destaca en ya que solo tendría 87 aciertos. En esta estadística los mejores modelos son SVM y regresión.
* A partir del modelo de regresión logística se tiene una precisión del 85%, lo que se traduce en que de cada 100 comentarios que el modelo afirme que son positivos, 85 realmente lo son. Para el caso de SVM, se tienen 83 de aciertos. Por otra parte, arboles de decisión vuelve a ser el peor modelo ya que solo tendría 63 aciertos. En esta estadística la regresión logística es el mejor modelo
* Otra característica del modelo de regresión logística, aún con sus parámetros optimizados, es que tiene un recall del 70%, lo que se traduce en que, de todos los comentarios que son negativos, sólo el 30% de los mismos no son bien interpretados por el modelo. Para el caso de SVM, se tiene un 31% de negativos perdidos. Por otra parte, arboles de decisión pierde la increíble cantidad de 42% de los verdaderos negativos. En esta estadística la regresión logística es el mejor modelo
* Adicionalmente, característica que se considera buena del modelo es el F-score: este puntaje va de 0 a 1 y determina que tan bueno es el modelo de acuerdo con el recall y la precisión, ya que ambos son inversamente proporcionales. Este puntaje enseña que el modelo tiene un buen balance entre ambos indicadores, pues el f score se acerca bastante a 1 con respecto a los demás modelos, ya que SVM se encuentra 2 puntos porcentuales por debajo & los árboles de decisión están 17 puntos porcentuales por debajo de la regresión logística
* Aunque inicialmente se habían considerado los parámetros de MSE y MSA, se considera que estos pueden llegar a no ser útiles en la práctica, ya que como mencionaron los expertos en estadísticas, los indicadores de distancia entre el hiperplano calculado por un modelo de regresión logística & cada uno de los datos no son significativos para el análisis.

**EVALUACION DE LOS MODELOS EN TERMINOS DE NEGOCIO:**

Ahora bien, tomando en consideración toda la información que se presentó anteriormente se consideró lo siguiente en términos de negocio:

* La métrica más importante es la de Precisión, ya que para cualquier estrategia que vaya implementar la empresa a partir de los comentarios negativos, es importante que esos comentarios negativos realmente sean negativos, como se enseña a continuación:
  + En caso de que la plataforma quiera enviar los comentarios negativos a cada una de las organizaciones hoteleras para que puedan mejorar, se considera nuevamente necesario que la precisión del modelo sea alta y que realmente esos comentarios enviados sean negativos y permitan mejoras significativas para la empresa.
  + Por otra parte, en caso de que la plataforma quiera utilizar el modelo para filtrar las búsquedas de hoteles de acuerdo con la percepción de los usuarios, tanto la precisión como la exactitud del modelo deben ser optimizados. En este sentido, el modelo cumple con las condiciones establecidas ya que cada uno estos 2 indicadores se encuentran optimizados

**RESULTADOS**

* Contexto del problema: Se considera un contexto ficticio en el cual los datos corresponden a una plataforma web que oferta diferentes hoteles y sobre la cual se realizaron los comentarios analizados. En este sentido, se considera que el modelo propuesto puede resolver las necesidades del negocio con la predicción de los comentarios
* Impacto en el negocio: Se espera que el modelo afecte positivamente el negocio, ya que se hace una estrategia de optimización de búsquedas, haciendo que cada uno de los usuarios se sienta más satisfecho con los resultados y con los hoteles mostrados. Así mismo, en la medida en que la información de los modelos pueda ser enviada allos hoteles, se espera un aumento de ingresos en la plataforma pues los servicios de los hoteles mejoraran, y por ende se estimula a los clientes a que utilicen la plataforma.
* Limitaciones y riesgos: un riesgo es que el algoritmo tienda a recomendar siempre los mismos hoteles ya que como los comentarios dentro de una plataforma de este estilo no tiene tantos cambios la información del modelo será siempre similar y por ende la clasificación de los hoteles, afectando negativamente a aquellos hoteles con comentarios negativos. Así mismo, el modelo tiene una gran limitación y es su carácter binario. Esto hace que se pierdan las ventajas de las etiquetas con múltiples categorías. Finalmente, el modelo arrastra los sesgos de cada una de las librerías utilizadas
* Recomendaciones: En este espacio se hará un análisis de cada uno de los lotes de palabras que, de la +21.000 tomadas en cuenta, resultaron ser las más significativas para el modelo, bien sea positiva o negativamente:
  + Lista de palabras MÁS negativas: de este lote de palabras se puede inferir que, además de las palabras tradicionalmente conocidas como adjetivos negativos dentro del idioma español, las siguientes palabras(así como todas sus derivadas inmediatas) son consideradas como ítems de primera necesidad para los clientes y pueden determinar que un hotel sea lo suficiente malo como para generar un comentario:
    - Paredes & sabanas: Al ser elementos básicos de la habitación, cualquier novedad de los mismos genera comentarios negativos
    - Insectos: Se encontró a la palabra cucaracha & sus derivados entre las palabras más influyentes, por lo que se infiere que la presencia de insectos es descalificadora para un hotel & sus comentarios
    - Jabón: Las condiciones del mismo parecen ser importantes para los clientes.
    - Ausencia: La presencia de cosas básicas en la habitación es imprescindible para los clientes
    - Pequeño: Los espacios que tenga el hotel son de vital importancia para la incomodidad de los clientes.
    - Caro: El precio hace que el hotel sea más susceptible a recibir comentarios negativos
  + Lista de palabras MENOS negativas: de este lote de palabras se puede inferir que, además de las palabras tradicionalmente conocidas como adjetivos positivos dentro del idioma español, las siguientes palabras (así como todas sus derivadas inmediatas) son consideradas como ítems de plusvalía para los clientes y pueden determinar que un hotel sea lo suficiente malo como para generar un comentario:
    - Limpieza: En contraparte con los adjetivos de suciedad y la referencia a las sábanas y paredes de los hoteles negativos, la limpieza es valorada como elemento de destacamento del hotel
    - Tranquilidad: Este puede ser considerado como en elemento de adjetivo calificativo positivo común
    - Amabilidad: La atención al cliente es importante para la percepción de los clientes de un hotel
    - Ubicación: La locación de un hotel es importante. Así mismo, palabras relacionadas como “estación” también se destacan, por lo que se entiende que las facilidades de transporte de un hotel son de gran valor para el cliente
    - 24h: Los servicios en todo momento son valorados mucho por los clientes como un elemento de plusvalía.
* Plan de acción & Conclusión final: Se considera que aunque la utilización de los modelos es útil, es evidente que el modelo de regresión es el que puede brindar mejores herramientas a la compañía. En este sentido, se invita al negocio a la utilización de los archivos .joblib para la utilización del modelo y el aprovechamiento del mismo.

**ANEXOS EVIDENCIA TRABAJO INTERDISCIPLINARIO:**

|  |  |
| --- | --- |
| Evidencia | Comentarios de Mejora |
|  | * Desestimación del MSA y RMSE para el modelo de regresión * Complementación de la evaluación en términos de negocio (la primera versión no tenía esta sección) * Correcciones de redacción & ortografía |