Revisión de sentimientos en comentarios de libros

Natalia Sanabria, Jorge Esguerra y Cristhian Forigua

Tabla de contenido

01

Contexto del problema

Contexto del problema

Se tienen datos sobre los comentarios de libros vendidos en la tienda Amazon Kindle Store, entre 1996 y 2014. Estos datos son relevantes para editoriales, tiendas de libros e incluso potenciales clientes.

Objetivos de Negocio

Creación de un sistema que permita ayudar a las editoriales y a Amazon a identificar el sentimiento del público con respecto a un libro, basado en reseñas que le han dado.

Relevancia

- Punto de vista de Amazon: Un sistema que sea capaz de determinar si la percepción pública sobre una obra es positiva o negativa podría ayudar al negocio a recomendar libros a ciertos tipos de audiencias, y también a buscar contratos de exclusividad de venta de diversas obras a través de Amazon Kindle.

- Punto de vista de Editoriales: Un sistema que sea capaz de determinar si la percepción pública sobre una obra facilitaría la labor de buscar explorar nuevas obras con contenidos similares, buscando aumentar el margen de venta .

Tarea desde el punto de vista de ML

- Generación de un clasificador binario que, dado un texto plano que representa la reseña de algún lector, determine si el sentimiento con respecto a la obra es positivo o negativo.

- Definición de la apreciación de una obra. Los datos que se utilizan para la tarea en cuestión incluyen un valor de calificación de 1-5 otorgado por la persona quien hizo la reseña, para este proyecto se propone que un sentimiento negativo equivale a una calificación menor o igual a 3, mientras que una calificación superior a 3 indica una afinidad o sentimiento positivo del lector hacia el libro.

- El proyecto se considera exitoso si se logra generar un clasificador que sea capaz de acertar, con una exactitud mayor al azar (50%), el sentimiento (positivo o negativo) de una reseña en texto plano.

02

Comprensión y procesamiento

982.619

Registros

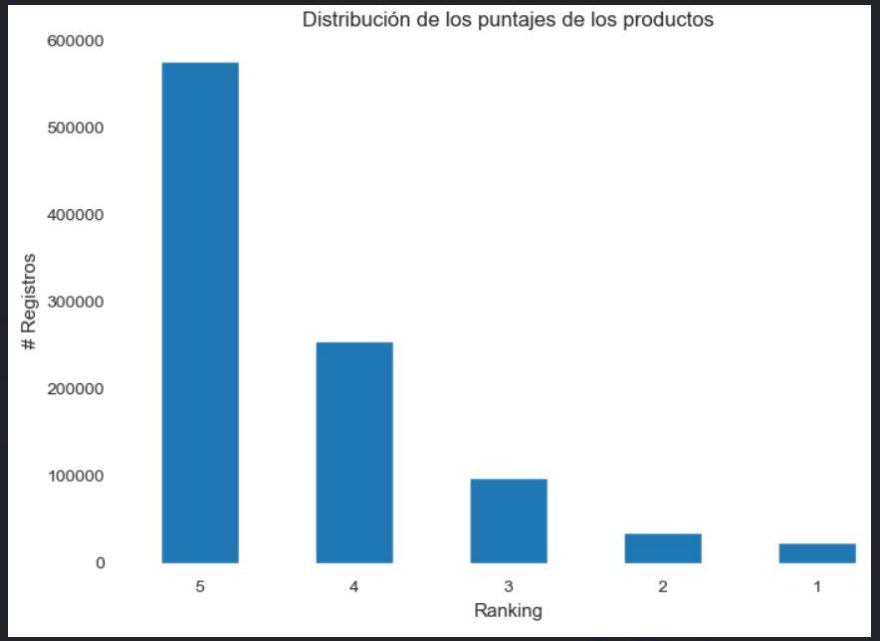
10

Variables

0

Valores atípicos

Distribución del puntaje asociado



Datos

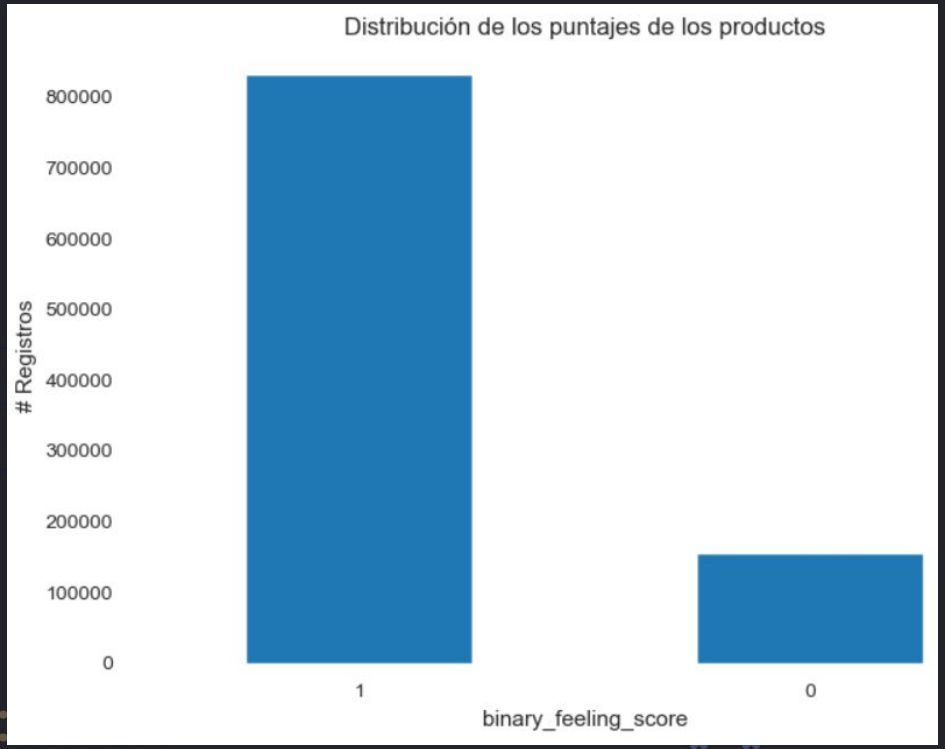
Se tienen 983000 reseñas distintas, cuya distribución de calificaciones se ve en el gráfico.

El set de datos es imbalanceado con respecto al Ranking otorgado, que es en últimas la clase que queremos utilizar como variable dependiente de nuestros modelos de clasificación.

Completitud

23 registros con valores nulos fueron eliminados

Binarización del puntaje asociado



Transformación

Se transforma esta variable Ranking (overall) a una columna binaria donde 1 indica un sentimiento positivo y 0 lo contrario, mediante la siguiente regla:

Si Ranking <= 3, entonces es un sentimiento negativo hacia la obra.

De lo contrario, la reseña indica un sentimiento positivo hacia la obra.

Dado que hay bastantes más reseñas de puntajes mayores a 3, se aplica técnica de sub-sampleo para datasets **desbalanceados**.

Preprocesamiento de las reseñas

01

Limpieza

Eliminación de palabras no ASCII, signos de puntuación, *stopwords*, contracciones y traducción de los números

02

Tokenización

Transformación de texto completo de una reseña a un arreglo

03

Normalización

Proceso de *Stem* y *lemmatize*

Desbalance de los datos

Se utilizó la técnica de under-sampling ya que como son datos de lenguaje natural es difícil generar nuevos datos.

Se muestrearon los datos para mantener una relación de 2:1 con la clase mayoritaria.

03

Modelos y resultados

Modelos propuestos

Multinomial NB

Mejor modelo de clasificación

Regresión logística

Random Forest classifier

Comparación de exactitudes de los modelos

83%

Multinomial NB

67%

Regresión Logística

67%

Random Forest

Resultados mejor modelo- Naive Bayes

83%

Precisión

Alta precisión del modelo

83%

Cobertura

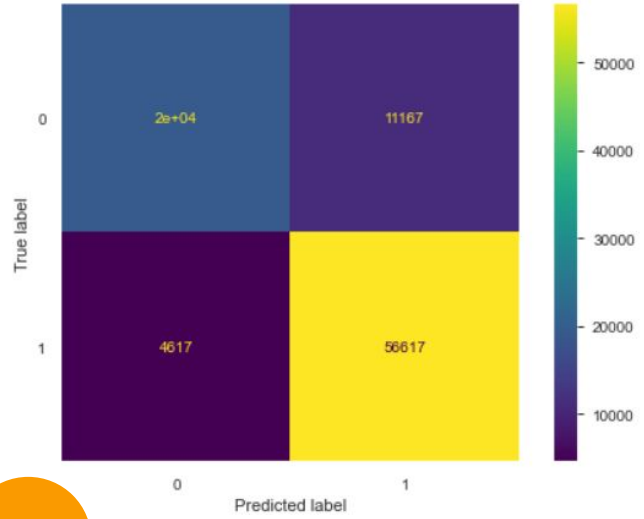
Alta cobertura del modelo

82%

F1_score

Alta relación entre precisión y cobertura

Resultados cuantitativos



Exactitud sobre entrenamiento: 0.86
Exactitud sobre test: 0.83

	precision	recall	f1-score	support
0	0.81	0.64	0.71	30771
1	0.84	0.92	0.88	61234
accuracy			0.83	92005
macro avg	0.82	0.78	0.80	92005
weighted avg	0.83	0.83	0.82	92005

Conclusiones

- El modelo Naive Bayes Classifier es poderoso para las tareas de procesamiento de lenguaje natural

- Para esta tarea, definitivamente el modelo más apropiado, en términos de exactitud y también de tiempo de ejecución fue Naive Bayes. Mientras que RandomForestClassifier demoraba más de una hora en calcularse, con hiperparámetros adecuados, el Naive Bayes Classifier lograba generarse y ajustarse en tan solo cuestión de un minuto, ganándole incluso a la regresión logística que tardaba algunos minutos en ejecutarse. Esto lo hace el modelo más apropiado para hacer parte de un pipeline de machine learning en producción, en donde el negocio probablemente querrá re-entrenar el modelo cuando se dispongan de más datos de reseñas.

Conclusiones

- A futuro, es importante reevaluar la hiperparametrización del modelo de RandomForestClassifier. La limitante de este proyecto, en cuanto a estos modelos, fue el poder de cómputo de la infraestructura donde fue entrenado. Los hiper parámetros tuvieron que limitarse, y no se probaron todas las posibles combinaciones para buscar mejorar el desempeño, puesto que el tiempo de ejecución de estos modelos incrementa radicalmente por cada valor adicional considerado en los métodos de grid search. Esto explica el sub-ajuste del modelo de RandomForestClassifier, que definitivamente no es óptimo para esta tarea de clasificación de reseñas, puesto que el cálculo de múltiples árboles de decisión de profundidades mayores a 15 toma un tiempo muy largo, lo que sería perjudicial si el modelo quisiera re-entrenarse en el futuro o si formara parte de un pipeline ejecutado periódicamente.

- El modelo de LogisticRegression fue el segundo mejor, y también tiene un tiempo de ejecución no muy extenso, lo que lo hace pertinente para ser reevaluado en iteraciones posteriores del proyecto, según lo considere el negocio. No obstante, la exactitud del modelo de Naive Bayes fue notoriamente superior.

Conclusiones

- El modelo de Naive Bayes Classifier es el que recomendamos para continuar con fases posteriores del proyecto. Podría eventualmente hacerse una aplicación que permita, en tiempo real, introducir una reseña y calcular el sentimiento de la misma. De esta manera, podría pedirse a expertos literarios la realización de reseñas sobre libros, buscando evaluar cuáles van a ser los más exitosos, y con ello tomar decisiones de negocio que permitan aumentar las ventas. Esto podría aprovecharlo tanto Amazon Kindle como las editoriales, al buscar hacerse con los derechos de distribución y de publicación de diversas obras prometedoras, respectivamente.