Inteligencia de negocios

Proyecto 1

Luccas Rojas - 201923052 28/03/2023

Tony Santiago Montes - 202014562

Brian Manuel Rivera - 202015320

Tabla de contenido

Entendimiento del negocio

Objetivos:

Nuestro principal objetivo va será poder predecir si una reseña de una película es positiva o negativa. Esto con el fin de que la empresa interesada pueda usar esta información para por ejemplo decidir que películas meter a una plataforma de streaming o que películas bajar de dicha plataforma por tener malas reseñas de los clientes.

Criterios de éxito:

Nuestros criterios de éxito estarán dados por cuántas reseñas podemos clasificar de manera correcta entre reseñas positivas y reseñas negativas. Esto será cuantificado a través de un porcentaje que presenta el número de películas que son clasificadas correctamente sobre el total de películas a clasificar.

Técnicas y algoritmos:

Para poder lograr casificar las reseñas entre positivas y negativas a través del aprendizaje automático vamos a utilizar distintos algorítmos de clasificación, como lo son Regresión Logística, Naive Bayes y SVM. Evaluaremos el comportamiento de los 3 algorítmos para así poder mejorar la predicción entregada. Esto nos permitirá a través de un proceso de entrenamiento generar un modelo que nos permitirá predecir si una reseña nueva es positiva o negativa.

Oportunidad:

Como empresa observamos que este proyecto tiene una gran oportunidad de negocio, ya que con el modelo de aprendizaje automático que deseamos desarrollar la empresa podrá aumentar su número de clientes y así sus ganancias. Esto ya que va a ser capaz de tener mejor contenido en su sitio de streaming y así atraer más clientes.

Beneficiados:

El beneficiario es la organización que va a poder utilizar la información entregada por el modelo para tomar desiciones que aporten al crecimiento del negocio.

Entendimiento de los datos

Unicidad:

Pudimos notar que no había reseñas duplicadas, por lo que los datos contaban con total unicidad.

Completitud:

Notamos tambíen que no había datos nulos dentro de los datos entregados, por lo que contamos ocn datos completos que por el momento podríamos utilizar.

Consistencia:

Para este caso confiamos en la fuente que nos proporcionó los y así en su consistencia.

Validez:

La mayoría de los datos parecen estar entregados de forma correcta y cumplen con los criterios del negocio, no obstante, hay unas excepciones, como reseñas en inglés que son datos que van a ser removidos para no afectar el modelo generado.

Preparación de los datos

En nuestro proceso para preparar los datos, lo primero que hicimos fue revisar la cantidad de reseñas escritas en inglés, más que todo para saber si era o no un procentaje significativo. Luego hicimos varios pasos para perfeccionar los datos antes de entrenar el modelo que fueron:

Convertir nuestra variable resultado de string a numérica, donde 0 es negativa y 1 positiva para así poder entrenar los diferentes modelos

Quitar todos los caracteres que no fueran ASCII de las reseñas

Convertir todas las palabras a minúsculas para unificar palabras que contengan minusculas y mayúsculas bajo la misma semántica.

Eliminar la puntuación de las palabras, de igual forma que lo anterior para unificar a la semántica de palabras que tienen puntuación.

Cambiar los números a texto para mantener la unicidad.

Remover las "stop words" que hacen referencia a todas aquellas palabras que tienen una correlación muy baja o nula con la variable de salida que es el sentimiento. Esto se hace eliminando conectores y artículos de los textos.

El proceso de vectorizar la reseña lo dejamos dentro del pipeline

Modelamiento

Para poder generar un modelo óptimo que le permita a la empresa sacar de manera precisa la categoría de una reseña, vamos a implementar 3 algorítmos de clasificación que son planteados posteriormente, además vamos a realizar un gridsearch sobre cada algoritmo para asegurarnos de que los hiperparámetros escogidos para cada uno de los algoritmos sean los mejores.

Modelo de regresión logística

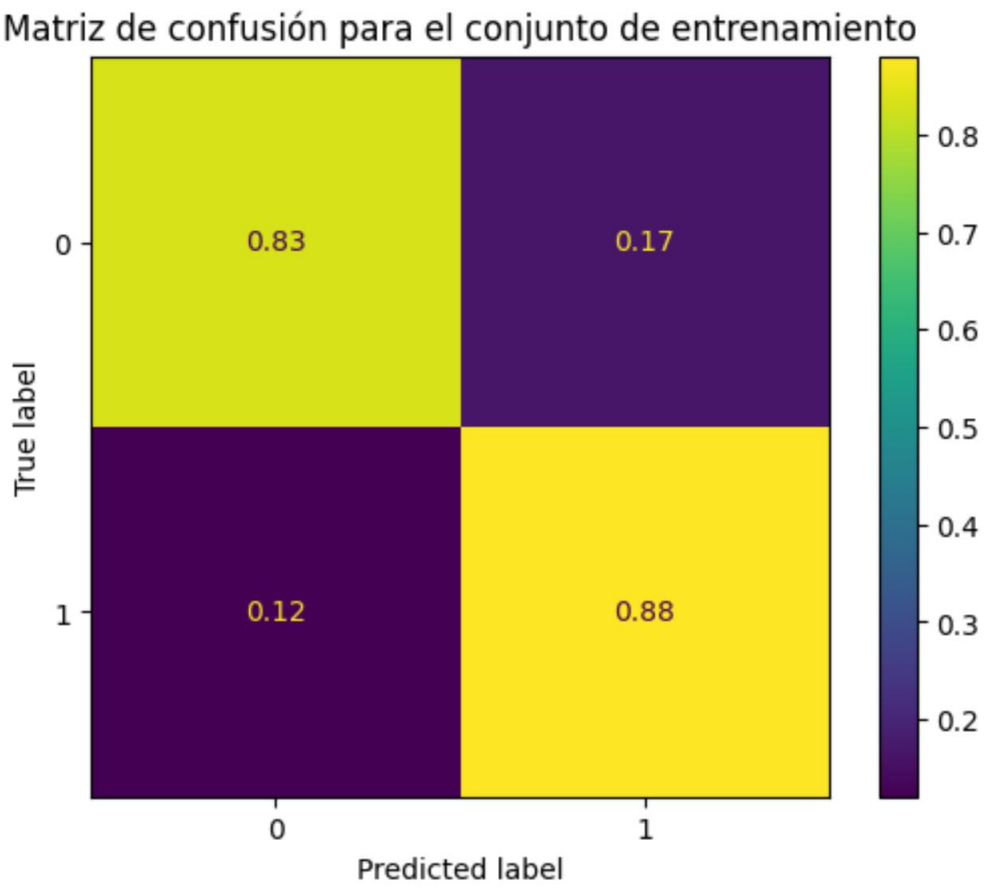
El algoritmo de regresión logística es una buena opción para clasificar opiniones binarias de películas, ya que puede predecir si una opinión es positiva o negativa. Por ejemplo, si queremos clasificar comentarios de películas como positivos o negativos, podemos utilizar un modelo de regresión logística para entrenar un clasificador basado en características de los comentarios. El modelo podría utilizar características como el número de palabras positivas y negativas en el

comentario, la presencia de palabras clave relacionadas con la trama o el género de la película, entre otros. El modelo puede luego predecir la probabilidad de que un comentario sea positivo o negativo. Con esta información, podemos determinar si una opinión es mayoritariamente positiva o negativa y usar esta información para tomar decisiones.

Modelado

Esta implementación utiliza la regresión logística como algoritmo de clasificación, el cual es adecuado para problemas de clasificación binaria como la clasificación de opiniones de películas. Se utiliza la técnica de vectorización TfidfVectorizer para convertir el texto en características numéricas y se integra en una tubería junto con la regresión logística. Además, se utiliza la búsqueda en cuadrícula para encontrar los mejores valores de los parámetros del modelo. La regresión logística es fácil de implementar y puede proporcionar probabilidades de clasificación en lugar de simples predicciones binarias, lo que es útil para la evaluación del modelo. La combinación de la regresión logística, la vectorización TfidfVectorizer y la búsqueda en cuadrícula permite obtener una alta precisión en la clasificación de opiniones de películas.

```
F1: 0.8628628628628628
```



Conclusión del algoritmo

Los resultados obtenidos a través de las matrices de evaluación indican que el proyecto ha logrado una predicción efectiva de aproximadamente el 86% ± 1%, evidenciado por el alto valor en las

métricas de recall y precisión cercanos a 1. Estos resultados respaldan el desempeño general del proyecto y demuestran que el modelo es capaz de clasificar adecuadamente las opiniones de las películas.

Modelo de multinomial Naive Bayes

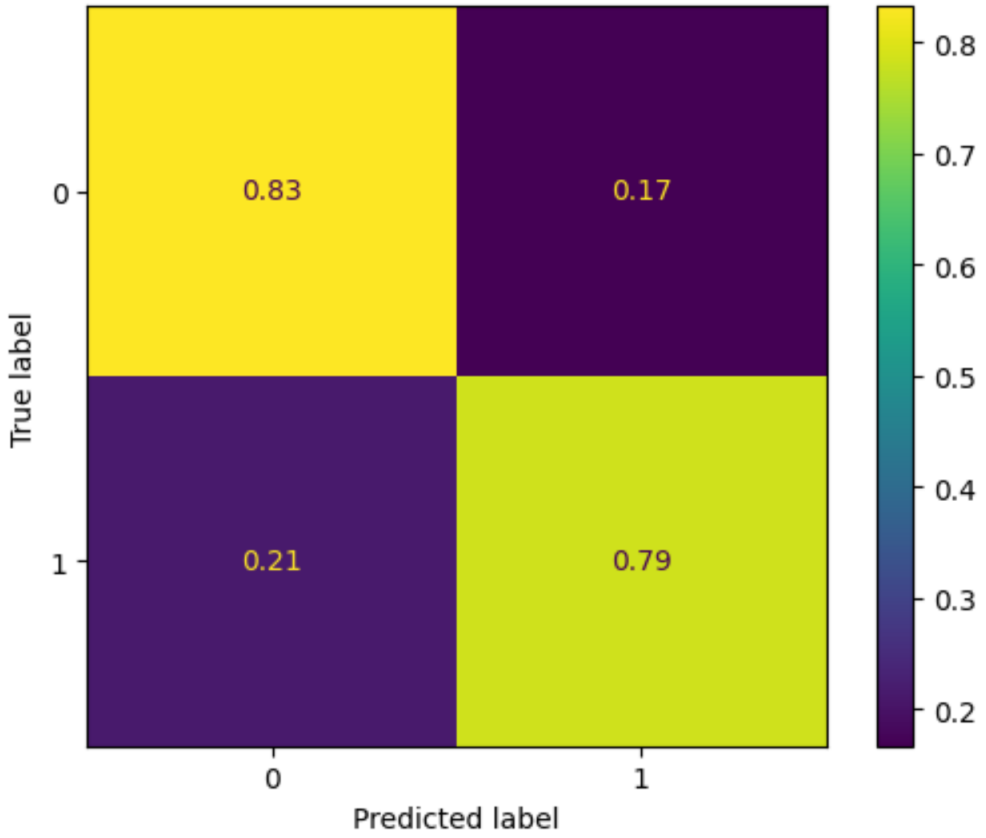
El algoritmo de Bayes, más específicamente el de Naive Bayes se suele utilizar el los algoritmos de clasificación de texto, como lo es el la clasificación de satisfacción de un usuario basado en la reseña de una película, que se trata de una tarea de clasificación binaria. Multionomial Naive Bayes es una variante de Naive Bayes que tiene un mejor desempeño y es más comúnmente utilizado para datos discretos. En el caso de aplicación, el algoritmo de Multinomial Naive Bayes tiene diversas ventajas sobre otros algoritmos, dado que al realizar una implementación de Bag of Words mediante CountVectorizer, se tiene un conjunto de datos discretos que contemplan la frecuencia de ocurrencia de ciertas palabras o frases; adicionalmente MNB provée un factor de probabilidad para la clasificación de sentimientos que permite obtener mejores resultados en el análisis del sentimiento de una persona de acuerdo a las palabras de su reseña.

Modelado

En específico en esta implementación se aplicó primero un CountVectorizer, como técnica de vectorización, que se encarga de tomar todas las reseñas preprocesadas, y convertirlas en una bolsa de palabras. Posteriormente se ejecuta el algoritmo Grid Search para la búsqueda del hiperparámetro 'alpha' del modelo Multinomial Naive Bayes, que representa un porcentaje de probabilidad adicional "ficticio" de la frecuencia de ciertas palabras o conjuntos de palabras, y que permite evitar que exista una probabilidad nula y así mejorar la precisión del modelo.

```
F1: 0.8034371643394199
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7
```



Conclusión del algoritmo

Según lo que se puede observar en las matrices de confusión presentadas anteriormente, y según el score obtenido anteriormente; se tiene una precisión promedio del modelo de alrededor del 82%, que es bastante buena. Así mismo, dentro del modelo el mejor valor para aplha seleccionado, se encuentra entre 2.5 y 4.5, con el fin de maximizar la precisión del algoritmo.

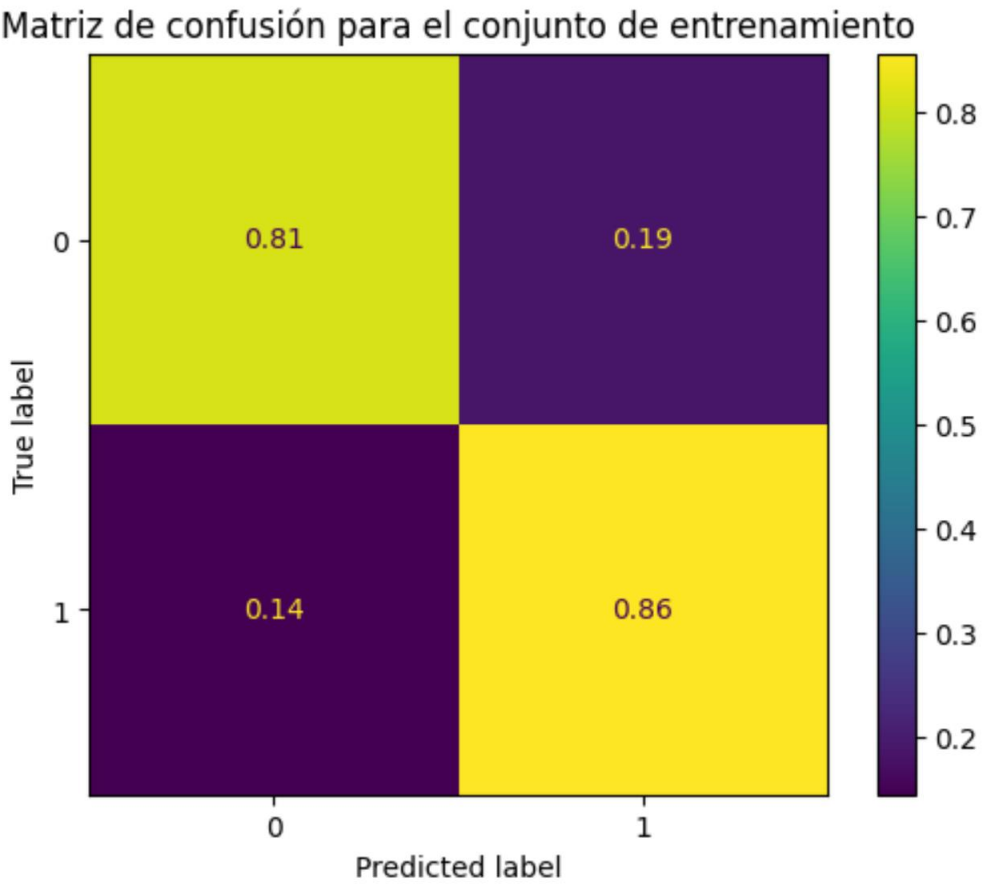
Modelo con SVC

El algoritmo de Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) es un algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión de datos. La idea principal detrás de SVM es encontrar un hiperplano en un espacio de alta dimensión que maximice la distancia entre las clases. El algoritmo SVM es útil para problemas de clasificación binaria y multiclase y puede manejar conjuntos de datos no lineales mediante el uso de un truco de kernel para transformar los datos a un espacio de mayor dimensión. La elección del kernel es un aspecto importante en el rendimiento del modelo SVM y se puede elegir entre diferentes tipos de kernels, como lineal, polinomial y radial. Además, el algoritmo SVM es eficiente en la clasificación de conjuntos de datos con muchas características. Sin embargo, puede ser sensible a la selección de hiperparámetros y requiere más recursos computacionales que otros algoritmos de aprendizaje supervisado.

Modelado

El algoritmo SVC ofrece varias ventajas importantes para la clasificación de datos, incluyendo su eficacia en conjuntos de datos complejos y no linealmente separables, su capacidad para generalizar bien para nuevos datos, su flexibilidad en la elección de funciones kernel, su capacidad para manejar datos atípicos o ruido en los datos y su alta precisión en la clasificación binaria. En este caso, la implementación del algoritmo SVC en una tubería junto con la técnica de vectorización TfidfVectorizer y la búsqueda en cuadrícula de los mejores parámetros, permite maximizar las ventajas de este algoritmo y mejorar la precisión de la clasificación de opiniones de películas.

```
F1: 0.8396793587174348
```



Conclusión del algoritmo:

Como se puede observar en las matrices presentadas anteriormente, podemos observar como una precisión con los datos de prueba del 85%, lo que implica un resultado bueno y que permitirá al negocio hacer predicciones con respecto a las reseñas de una película. Este algoritmo tuvo como mejores parámetros un C de 1, un y un kernel linear para maximizar la precisión.

Conclusiones

Matriz de métricas de cada uno de los algoritmos realizados en el proyecto:

	Precisión	Recall	F1 - Score	Soporte
Regresión Logística	85%	85%	85%	961
Multinomial Naive Bayes	82%	82%	82%	961
SVC	85%	84%	84%	961

Los resultados del modelo son importantes para Netflix ya que pueden mejorar su proceso de selección de películas y, por lo tanto, la satisfacción del cliente. El modelo proporciona información valiosa sobre las opiniones de los usuarios en relación a las películas ofrecidas, lo que permite a Netflix seleccionar y promocionar las películas con mejores reseñas y adaptar su biblioteca de contenido a las preferencias de sus usuarios. Sin embargo, se recomienda utilizar el modelo en combinación con otros métodos de selección de películas y actualizarlo periódicamente.

Descripción del trabajo en equipo

Roles

Brian Manuel Rivera:

Rol: líder de datos Brian se encargó en su mayoría de la limpieza de datos, haciendo el tratamiento de las variables, vectorizando las reseñas de entrada, eliminado las palabras que generaran ruido, eliminando las reseñas en inglés y en general dejando los datos listos para poder ser utilizados por el resto del equipo. Brian realizó el algoritmo de Regresión Logística para intentar solucionar el proyecto planteado

Tony Santiago Montes

Rol: líder de negocio y líder de analítica Tony se encargó de revisar los diferentes modelos y escoger el mejor modelo, en definir con qué métricas evaluaríamos los algoritmos realizados y verificar los estándares de calidad necesarios para la entrega del modelo que mejor se comporta. Además, Tony veló por resolver el problema del negocio y dirigir el equipo hacia el fin de completar la tarea que el negocio necesita. También fue el encargado de comunicarse con el experto de estadística y tener listas las preguntas así lograr un trabajo interdisciplinario con esta persona. Tony desarrollo el algoritmo de multinomial Naive Bayes para intentar solucionar el problema.

Luccas Rojas

Rol: líder del proyecto Luccas Rojas fue el encargado de dirigir el proyecto, establecer los días y horas de reunión entre todos los integrantes, en organizar los entregables del grupo y asignar las tareas que cada uno de los integrantes iba a realizar para que fuera lo más equitativo posible. Además, fue el que desarrolló el algoritmo de SVC para intentar solucionar el problema.

Reflexión grupal

Repartición de los 100 puntos

Tony Santiago Montes 33.3

Brian Manuel Rivera 33.3

Luccas Rojas 33.3

Aspectos por mejorar

Reunirnos más seguido

Definir los roles y el trabajo por hacer desde el principio

Mejorar la comunicación entre nosotros para coordinar las tareas y asi no llegar a repetir las tareas que otros ya realizaron

Mejorar la cooperación del equipo para que todos nos ayudemos en las tareas y podamos aprender de lo que el otro hizo.

Tenemos que mejorar también la motivación que tenemos con respecto a las actividades, ya que llegamos a hacer el trabajo sólo por entregarlo.

Debemos aprender a manejar mejor el tiempo para que nos rinda más y tengamos más tiempo para corregir los errores.