

PROYECTO FINAL

MODELADO PREDICTIVO EN ÓRDENES DE VENTA POR INTERNET

Integrantes Equipo

- Moscardo Federico
- Walker Agustín
- Yacob Santiago



Tabla de Contenido

1. Contexto comercial
2. Problema comercial
3. Definición de objetivos del análisis
4. Descripción de los datos
5. Análisis exploratorio de datos (EDA)
 - 5.1. ¿Cuál es la región con mayor participación en ventas online?
 - 5.2. ¿La región/estado con mayor incremento en ventas online?
 - 5.3. ¿Cuáles son las categorías con tendencia a incrementar o disminuir su cantidad de ordenes online?
 - 5.4. ¿Cómo es la situación de las órdenes de compra canceladas?
 - 5.5. ¿Se podría identificar algún patrón al respecto?
 - 5.6. ¿Existen factores externos que podrían incidir en la cancelación de órdenes de venta?
6. Algoritmo Elegido
7. Cross Validation
8. Insights y Next Steps

Contexto comercial

El **incremento** exponencial de personas con **acceso a internet** en los últimos años, la **globalización** y la reciente **pandemia**, han conformado el escenario ideal para situar a las ventas en línea como reales protagonistas de la economía mundial.

La tendencia indica que **las personas están cambiando sus hábitos de consumo**, disminuyendo su presencia en shoppings y centros comerciales e incrementando las compras desde la comodidad de un dispositivo con conexión a internet.

Dado lo descrito, una importante **empresa de comercio electrónico de Estados Unidos** ha recopilado información sobre las órdenes de compra generada por clientes en forma online en un determinado período de tiempo, de tal manera de disponibilizarla al equipo de Data Science para que se pueda analizar e identificar patrones de consumo, tendencias de venta, agrupaciones en clústeres o toda información que permita la toma de decisiones y el ajuste interno, de tal manera de **reducir costos operativos y maximizar márgenes de utilidad**.

Problema comercial

La tarea del equipo de Data Science consiste en trabajar los datos proporcionados, realizar una limpieza que permita manipularlos correctamente con el mínimo error y proporcionar visualizaciones que respondan las preguntas específicas que tiene el cliente, que se mencionan a continuación.

Definición de objetivos del análisis

Cómo se mencionó, el cliente desea que se analice la información de órdenes de compra de los últimos años a fin de identificar patrones, tendencias de consumo y potenciales segmentaciones.

El cliente tiene un conjunto específico de preguntas a las que le gustaría obtener respuesta:

- ¿Cuál es la región con mayor participación en ventas online? ¿La región/estado con mayor incremento en ventas online?
- ¿Cuáles son las categorías con tendencia a incrementar o disminuir su cantidad de órdenes online?
- ¿Cómo es la situación de las órdenes de compra canceladas?
- ¿Se podría identificar algún patrón al respecto?

- ¿Existen factores externos que podrían incidir en la cancelación de órdenes de venta?

Descripción de los datos

Dataset Original:

- **order_id:** id de la orden de compra
- **order_date:** fecha de la orden de compra
- **status:** estado de la orden
- **item_id:** id producto comprado
- **qty_ordered:** cantidad ordenada del producto
- **price:** precio del producto
- **value:** cantidad por precio
- **discount_amount:** monto de descuento
- **total:** precio final de compra
- **category:** categoría del producto
- **payment_method:** método de pago
- **cust_id:** id del cliente
- **Gender:** género del cliente
- **Age:** edad del cliente
- **Customer_since:** fecha de alta del cliente
- **County:** condado
- **City:** localidad
- **State:** estado
- **Zip:** código postal
- **Region:** región
- **Discount_Percent:** % descuento asociado

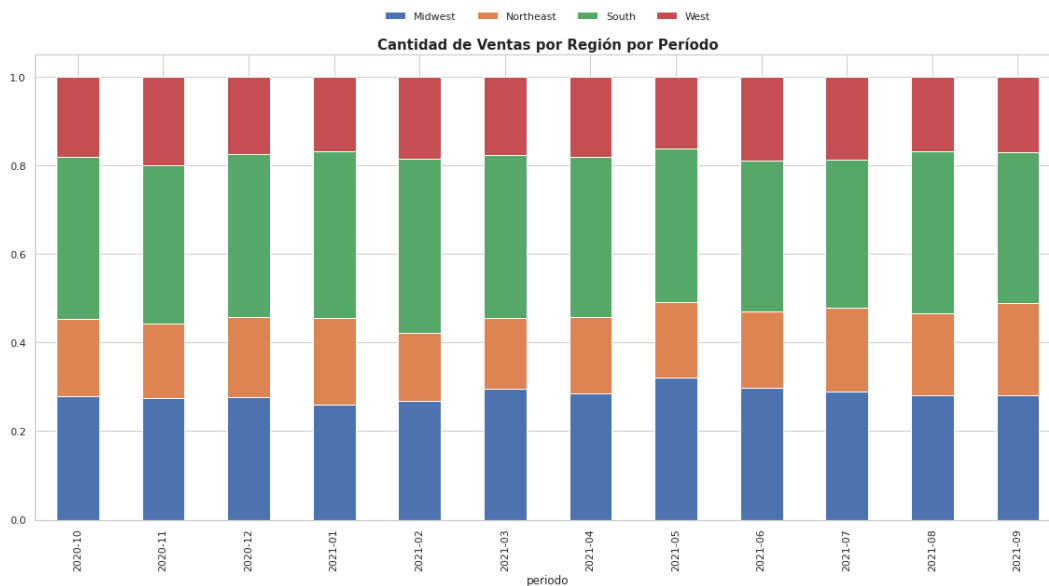
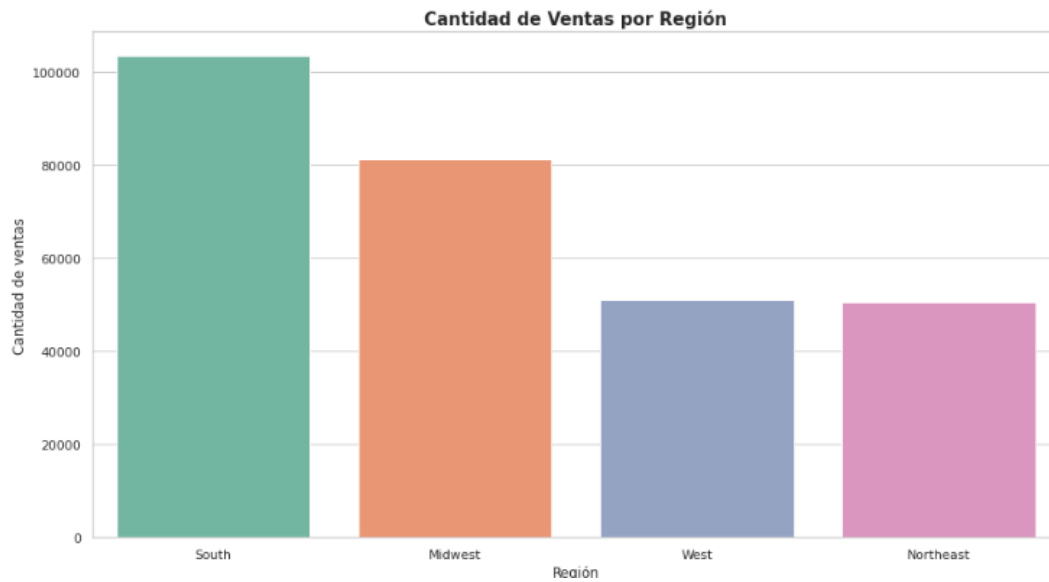
Dataset API:

- **date:** fecha del valor de las acciones
- **open:** valor apertura acción
- **high:** valor máximo acción
- **low:** valor mínimo acción
- **close:** valor cierre acción
- **volume:** volumen de acciones

Análisis exploratorio de datos (EDA)

¿Cuál es la región con mayor participación en ventas online?

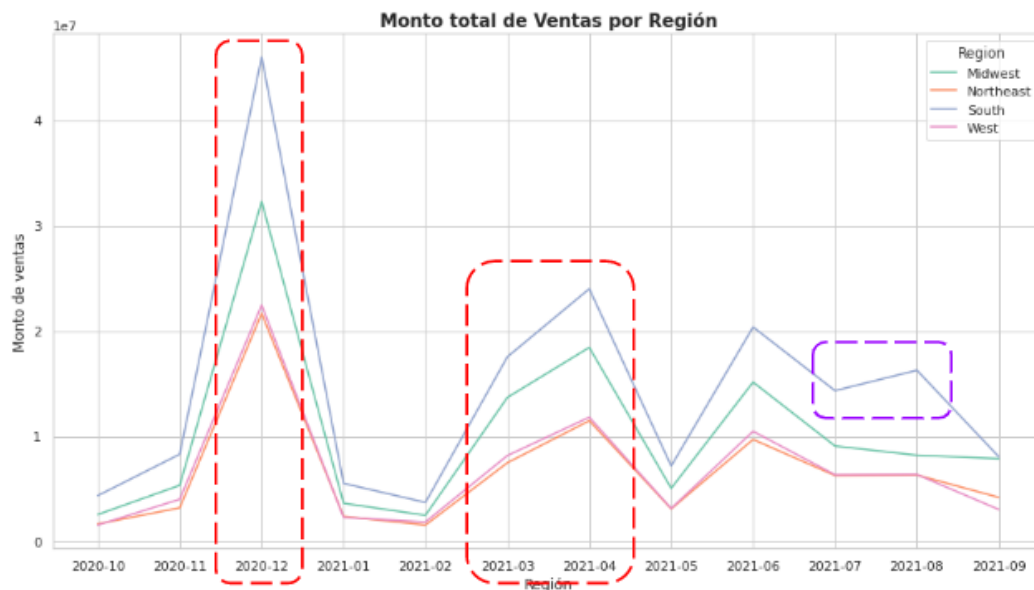
South USA es la región con mayor participación en órdenes on-line. Además **tiene un liderazgo sostenido** a lo largo de los períodos en análisis.



¿Cuál es la región/estado con mayor incremento en ventas online?

Existen **2 picos** bien marcados en los meses de **Diciembre 2020** y **Marzo/Abril 2022**. Tendremos que ahondar más adelante con detalles de estos comportamientos.

En la **región South** tenemos un **aumento de las ventas en monto total**, en el mes de **Agosto 2021** que no se observa en otras regiones. También buscaremos más detalles al respecto.



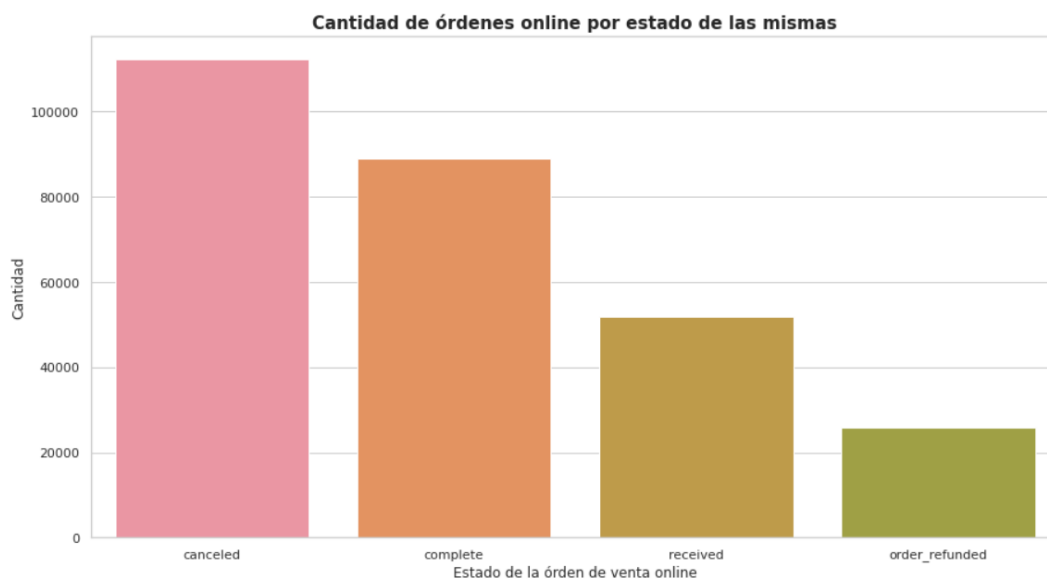
¿Cuáles son las categorías con tendencia a incrementar o disminuir su cantidad de órdenes online?

Sin dudas, observando el gráfico, se deberá poner foco en la categoría **Mobiles & Tablets** la cuál es la dominante en sumatoria de ventas en comercio online. También habría que prestar especial atención a categorías **Books, Entertainment y Men's Fashion**, principalmente por los decrecimientos observados en las órdenes efectuadas.



¿Cómo es la situación de las órdenes de compra canceladas?

En la búsqueda de insights o patrones, se realizó un gráfico de barras en el cuál se agrupan la cantidad de pedidos online por estado de los mismos. Realmente llamó la atención que la categoría con mayor repeticiones sea "cancelado", casi por un 30% más que la categoría siguiente "completado" (que suponíamos por lógica sería la primera categoría).



Esta situación le resultó de particular interés al equipo de Data Science, por lo que se decidió compartir esta visualización al cliente, para conocer si se encontraban al tanto de tan alta tasa de cancelaciones. El cliente se encontró sumamente sorprendido con esta información y pidió ahondar específicamente en este rubro, dado que una orden de venta online genera una cantidad muy importante de información interna y procesos de validación, lo cuál resulta en erogaciones de dinero que se terminan perdiendo si la orden se cancela.

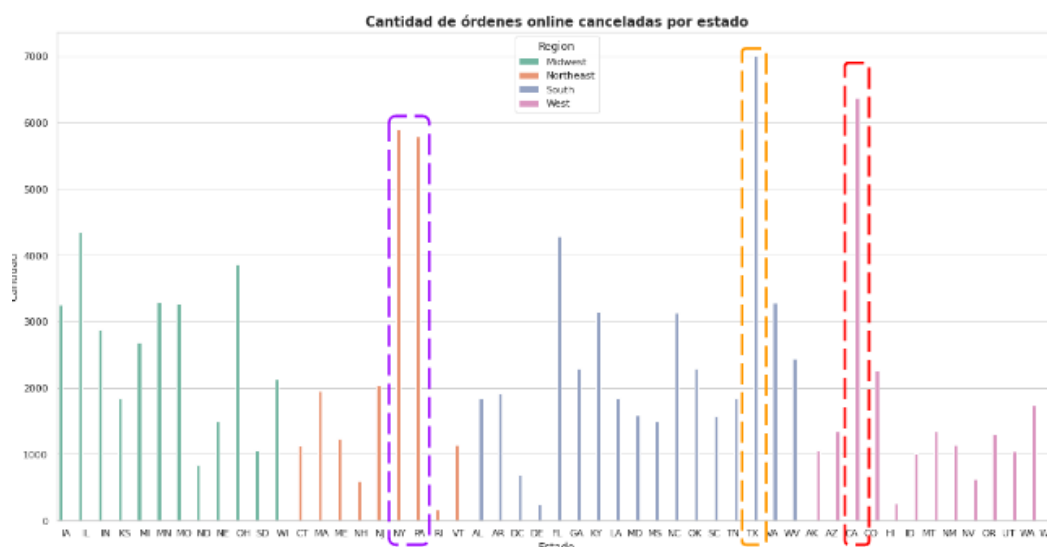
Por lo tanto, el equipo de científicos de datos dedicará una importante cantidad de recursos en ahondar sobre las órdenes de compra canceladas, de tal forma de identificar patrones que permitan clusterizar las órdenes de venta que se cancelan, a la vez que evaluar la posibilidad de crear un algoritmo que prediga cuando una orden será cancelada.

¿Se podría identificar algún patrón al respecto?

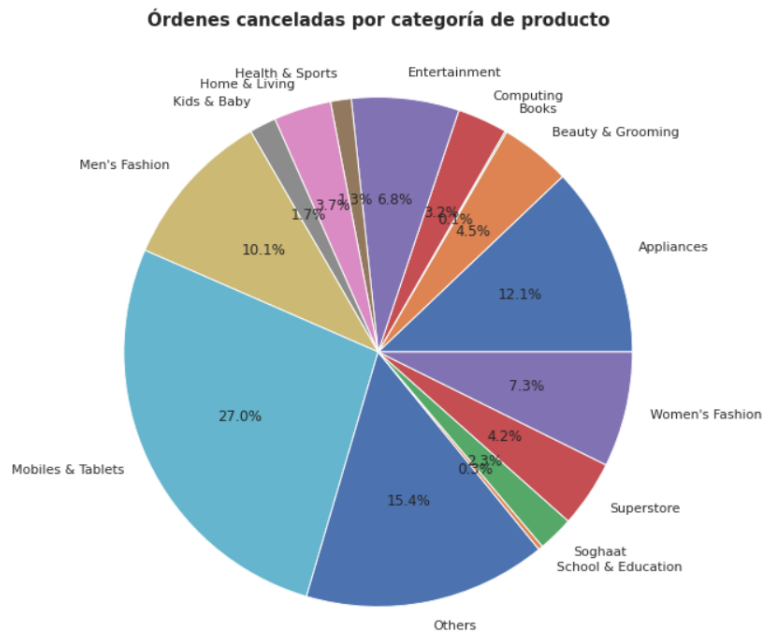
Si hacemos un corte por región, **no hay concentración de cancelaciones por región**, por lo que se descarta esta hipótesis

Sí tenemos **concentración de cancelaciones en algunos estados**, tal es el caso de:

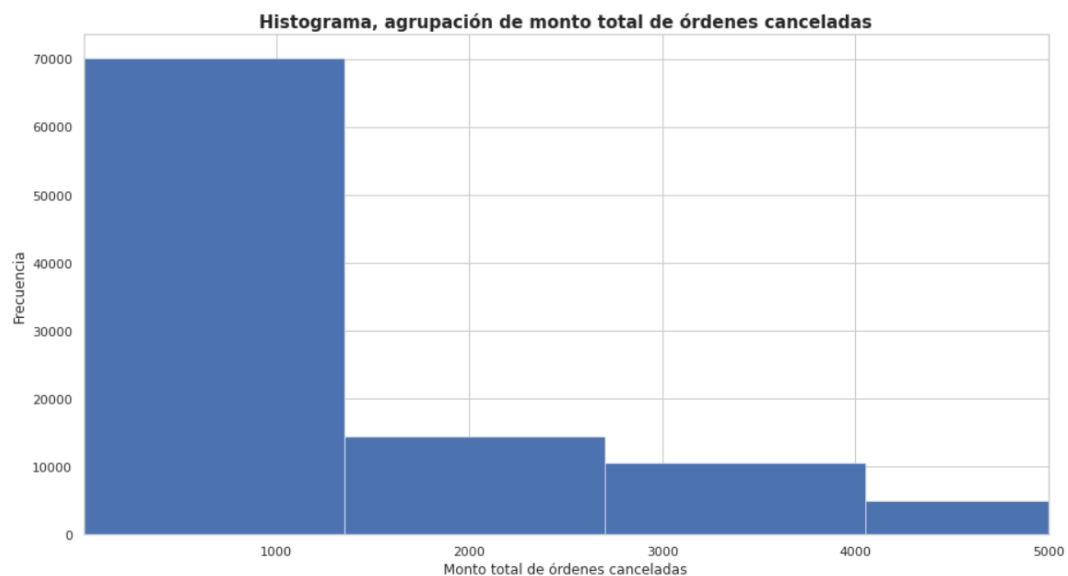
- CA (California - West)
- NY/PA (New York y Pensilvania - Northeast)
- TX (Texas - South)



Mobiles & Tablets llega casi a 30% de sus órdenes canceladas. Tiene coherencia, ya que como veíamos antes es la categoría con mayor cantidad de ventas, es un dato importante que tenemos que considerar en el análisis por el tipo de producto

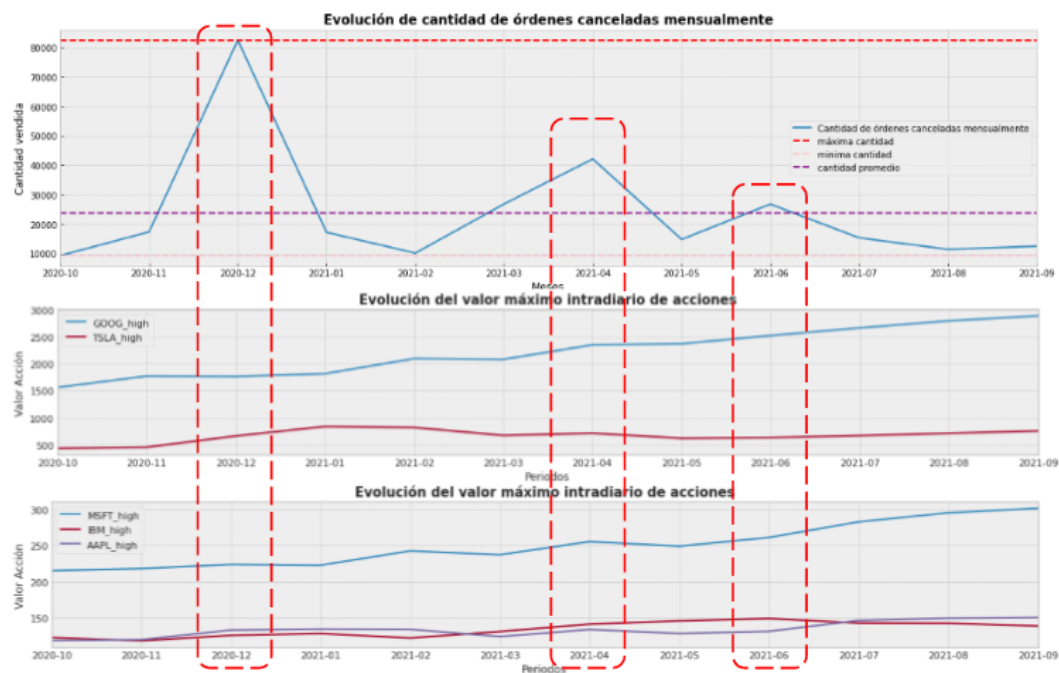


Casi el **90% de las órdenes canceladas** fueron órdenes de un **valor inferior a los 2500 USD**



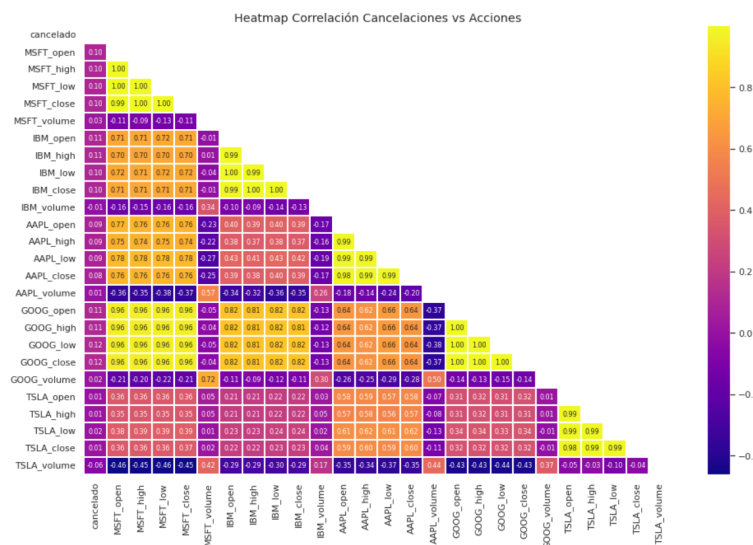
¿Existen factores externos que podrían incidir en la cancelación de órdenes?

No se observa una **correlación directa** entre la cantidad de **órdenes canceladas** mensualmente con las variables de referencia promedio de las **acciones de las principales compañías tecnológicas**.



En el **análisis OLS** que se realizó con todas las variables de las acciones (valor de apertura y cierre, máximo y mínimo) **también arrojó resultados negativos**.

En el **heatmap** vemos gráficamente mismo resultado: **no hay correlación**.



Algoritmo Elegido

Decision Tree

El árbol de decisión modelado tiene en cuenta variables numéricas como ser la cantidad ordenada, precio de venta, descuentos así como también variables originalmente categóricas transformadas a numérico, como ser categoría y métodos de pago.


Se transformó la columna de datos "status" (originalmente variable categórica) en numérica tomando valor 1 si la orden fue cancelada y valor 0 si la orden no fue cancelada.

Luego de generado y entrenado el método de clasificación propuesto, se utilizó el 30% del dataset para testear el mismo, observándose la capacidad de predicción del mismo en la matriz de confusión.

Se calcularon métricas de accuracy, precision, sensibilidad, especificidad y f1 para evaluar el modelo; arrojando valores aceptables de predicción de órdenes canceladas y no canceladas. Por último, se tomaron ejemplos puntuales para testear la predicción del modelo.



La exactitud del modelado, es decir la proporción entre lo predicho y la totalidad de los casos arroja un coeficiente de 0.7918, el cuál refleja que el modelo se adapta relativamente bien a nuevos valores (los proporcionados por el dataset de test); aunque se podría mejorar dado que no predice correctamente el 20.82% de los nuevos datos ingresados.



Contemplando las métricas de precisión, la precisión positiva y negativa arrojan porcentuales de 73% y 83%; por lo que hay un aceptable rango de predicción teniendo en cuenta los falsos positivos y negativos. El modelo clasifica en forma correcta los nuevos datos ingresados.

A la hora de hablar del recall de sensibilidad y especificidad, los coeficientes obtenidos son 0.74 y 0.83, lo que consideramos como valores aceptables, aunque se podría mejorar el recall de sensibilidad, o la capacidad de identificar los verdaderos positivos que en nuestro caso serían las órdenes predichas como canceladas y efectivamente canceladas en el test set.

Por último, el F1 score arroja coeficientes de 0.83 para los casos no cancelados y 0.74 para los cancelados, lo que sugiere poca presencia de falsos positivos y falsos negativos. No obstante, entendemos que el recall podría mejorarse, especialmente en el aspecto de predecir correctamente las órdenes que por regla del modelo se cancelarán, lo que termina siendo la finalidad última del modelado predictivo.

Resumen de métricas de la performance del modelo:

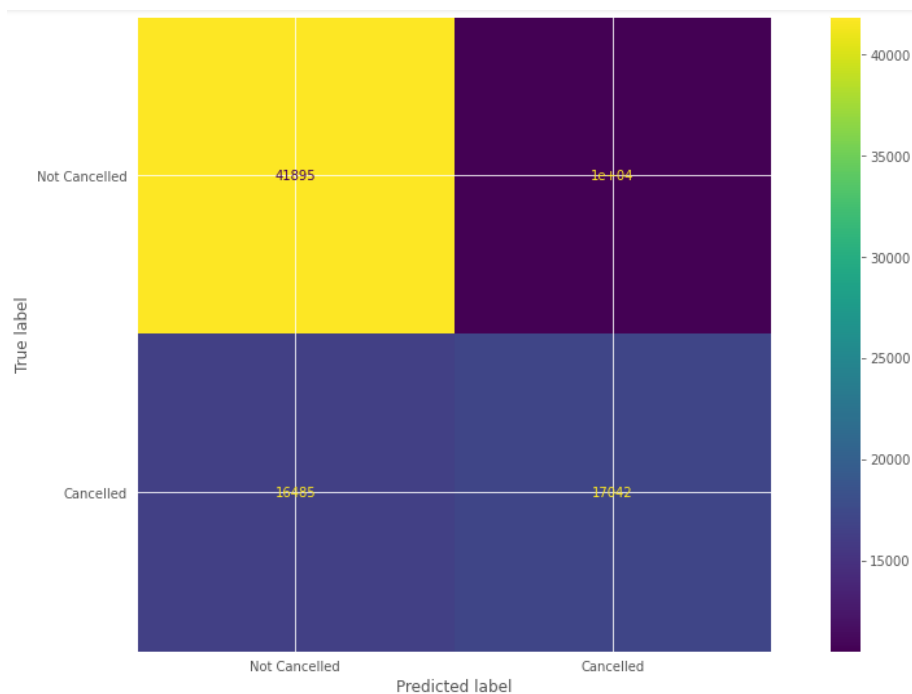
- Accuracy: 79,18%
- Precision_positiva: 73%
- Precision_negativa: 83%
- Recall_sensibilidad: 74%
- Recall_especificidad: 83%

Bayes Naive Classifier

Se propuso una alternativa al modelo de clasificación a partir de árbol de decisión; con lo cuál realizamos un modelado de clasificación a partir del Teorema de Bayes, el cuál relaciona probabilidades condicionales de eventos. El modelado generativo utilizado se denomina Gaussian Naive Bayes, el cuál asume distribución gaussiana a la vez que independencia de las características seleccionadas.

A la hora de iniciar el modelado, se realizó una feature selection utilizando Select K Best, el cuál analiza y selecciona características con mejores índices "k". El procesamiento del feature selection nos arrojó 5 características que resultaron las más adecuadas a utilizar, con lo cuál se analizó la correlación entre las mismas para ratificar o refutar el feature selection seleccionado.

Luego de generado y entrenado el método de clasificación propuesto, se utilizó el 30% del dataset para testear, observándose la capacidad de predicción del modelado a partir de las métricas construidas y la matriz de confusión.



Resumen de métricas de la performance del modelo:

- Accuracy: 68,6%
- Precision_positiva: 62%
- Precision_negativa: 72%
- Recall_sensibilidad: 51%
- Recall_especificidad: 80%

Al analizar el modelado de clasificación a partir de Gaussian Naive Bayes, tanto la matriz de confusión como las métricas nos brindan información contundente: El modelado no presenta una exactitud que consideramos aceptable; todas las métricas presentaron valores inferiores al modelado de clasificación a partir de árbol de decisión.

Claramente nos encontramos en un caso de underfitting dado que el modelo encuentra dificultad para encontrar patrones y predecir órdenes que a futuro se cancelarán. Teniendo en cuenta las propuestas de modelado de clasificación realizadas, sugerimos avanzar en el moldeado y corrección de modelo de clasificación a partir de árbol de decisión.

Cross Validation

La validación cruzada (cross-validation en inglés) es una técnica de evaluación de modelos de aprendizaje automático que se utiliza para estimar la precisión de un modelo en una muestra de datos. La idea detrás de la validación cruzada es simple: en lugar de utilizar una sola partición de los datos para entrenamiento y otra para pruebas, se dividen los datos en varias particiones y se realizan varios entrenamientos y pruebas, de tal manera que cada partición de los datos se utilice una sola vez para pruebas.

La validación cruzada es útil porque permite obtener una estimación más precisa de la precisión del modelo en datos nuevos. Si se utiliza una sola partición de los datos para entrenamiento y pruebas, es posible que el modelo se ajuste demasiado bien a esos datos, lo que puede llevar a una sobreestimación de la precisión. Al utilizar varias particiones de los datos para entrenamiento y pruebas, se minimiza este riesgo y se obtiene una estimación más precisa de la precisión del modelo en datos nuevos.

Validación de modelado de árbol de decisión

- Accuracy: 0.67 (+/- 0.08)

Validación de modelado Gaussian Naive Bayes

- Accuracy: 0.63 (+/- 0.08)

Insights y Next Steps

INSIGHTS

- Mobiles & Tablets es la categoría dominante en sumatoria de ventas en comercio online
- Books, Entertainment y Men's Fashion tienen caídas de ventas a partir de 21-Q3
- El 37,2% de las órdenes están CANCELADAS
 - no hay concentración de cancelaciones por región
 - Existe una concentración de cancelaciones en los ESTADOS de California, New York, Pensilvania y Texas
 - 90% de las órdenes canceladas fueron órdenes de un valor inferior a los 2500 USD

- Factores externos como los valores de acciones de las principales compañías tecnológicas, no afectan a las cancelaciones de ventas online

NEXT STEPS


- Realizar un deep dive con el equipo comercial, de lo sucedido en los picos de ventas de Dic-2020 y Mar/Abr-2022. Esto podría generar un insight para prepararnos mejor para futuros incrementos de ventas y aprovechar esta demanda estacional
- Realizar un workshop entre equipos comerciales regionales, para analizar la situación de órdenes de ventas de la región South y la concentración de cancelaciones en los ESTADOS de California, New York, Pensilvania y Texas
- Existe una demanda particular en esta región?
- Tenemos diferentes estrategias que nos hacen captar mejor la demanda? Se pueden llevar estas estrategias al resto del país?
- Implementar el modelo de Decision Tree para predecir futuras cancelaciones. Esto permitirá tomar las mejores acciones y reducir el impacto en los costos operativos y maximizar márgenes de utilidad de la compañía

Conclusión del Proyecto

En conclusión, este proyecto de análisis de datos se enfoca en el e-commerce y la tendencia alcista de compras en línea. La pandemia y el aumento exponencial de personas con acceso a internet han dado lugar a un aumento en las compras en línea. Una importante empresa de e-commerce de Estados Unidos ha recopilado información sobre las órdenes de compra generadas por sus clientes en línea para ser analizadas por el equipo de Data Science. La tarea del equipo consistió en trabajar con los datos proporcionados, realizar una limpieza adecuada y proporcionar visualizaciones que respondan a las preguntas específicas del cliente.

El objetivo final es ayudar a la empresa a tomar decisiones informadas y maximizar sus márgenes de utilidad. Para lograr esto, el equipo de Data Science realizó un proceso de data preparation y data wrangling, leer y transformar los datos para su visualización y construir visualizaciones que identifiquen patrones y tendencias en el conjunto de datos.

El equipo de Data Science ha llevado a cabo un análisis exhaustivo para identificar patrones y causas detrás de la elevada tasa de cancelaciones en las órdenes de venta online. Al compartir esta información con el cliente, se logró generar una conciencia sobre el impacto económico que esta situación genera y se decidió profundizar en la investigación.



Inicialmente, se analizó la posible relación entre las cancelaciones y los valores bursátiles de las principales compañías tecnológicas, pero los resultados de los análisis OLS de regresión indicaron que no existía una correlación directa. En este sentido, se descartó este insight.

Posteriormente, el equipo de Data Science construyó un modelo de árbol de decisión para identificar las variables más importantes en la predicción de las cancelaciones. Se transformaron variables categóricas a numéricas y se evaluó el modelo utilizando el 30% del dataset para hacer pruebas. Las métricas de accuracy, precisión, sensibilidad, especificidad y f1 arrojaron valores aceptables, lo que indica que el modelo es capaz de predecir con éxito las cancelaciones y no cancelaciones.

La validación cruzada resultó una técnica fundamental ya que permitió evaluar el rendimiento de un modelo en datos no vistos previamente y proporcionó una estimación más precisa de su desempeño en el mundo real. En definitiva, resultó esencial para garantizar que los modelos de aprendizaje automático sean confiables, generalizables y capaces de realizar bien en tareas específicas.

En general, el modelo propuesto brinda una solución efectiva para identificar las órdenes que están en riesgo de cancelarse y tomar medidas preventivas. Además, los resultados obtenidos son coherentes con la información presentada en la matriz de confusión y las métricas. Por lo tanto, se puede afirmar que el modelo ha sido entrenado adecuadamente y no se encuentra en una situación de underfitting o overfitting.

En resumen, el trabajo realizado por el equipo de Data Science permite conocer mejor los patrones y causas detrás de las cancelaciones de órdenes de venta online, lo que a su vez permitirá al cliente tomar medidas preventivas y mejorar su rentabilidad.