Proyecto final Data Science - Comisión 28330

ANÁLISIS DE LOGÍSTICA DE PRODUCTOS EN UN ECOMMERCE

Delgado, Jose Carlos - Rivero, Julián - Silvetti, Mariano

Tabla de contenidos

Presentación

1. Presentación de la temática

En el presente proyecto, abordaremos la información respectiva a la logística de productos en una empresa genérica de e-commerce. Disponemos de un dataset en el que se detallan, entre otros, datos inherentes a las características de los productos enviados, compradores y condiciones de entrega. Intentaremos, a través del análisis exploratorio del dataset, desarrollar un modelo que permita predecir con cierta precisión si, a partir de sus condiciones, un envío podrá entregarse o no en el tiempo y forma pactados.

2. Desafíos e interrogantes

Entre las interrogantes que surgieron a partir de una lectura superficial del dataset y su perspectiva de negocio, podemos mencionar las siguientes:

¿De qué manera se distribuyen los productos entre los distintos almacenes?

¿Cuál es el costo de los productos? ¿Influye en su disposición en los almacenes?

¿Qué tan diversas son las características físicas (en este caso, el peso) de los productos y qué dificultades puede representar?

¿Influye de alguna manera el precio de los productos en las calificaciones de los compradores?

¿Qué papel tienen las características demográficas del comprador en las características del producto?

¿Los productos en promedio fueron entregados en tiempo y forma?

Si la importancia del producto es elevada.¿Por este motivo se entrega con mayor frecuencia a tiempo o no varía si el producto es de mayor relevancia?

¿Se observan distintas tendencias en la correcta entrega de los productos según sus características?

3. Objetivos

Planteadas las preguntas que servirán de disparadores en nuestro análisis, nos propusimos los siguientes objetivos:

Describir las características de los productos, los compradores y las condiciones de entrega, interiorizando al lector en la información analizada.

Detectar la existencia (o no) de **patrones y correlaciones** entre las características previamente expuestas.

Desarrollar un modelo de ML para predecir exitosamente si un envío podrá o no realizarse correctamente.

Probar diferentes algoritmos de clasificación, establecer métricas comunes y comparar su efectividad, para poder seleccionar el que mejor se adapte al caso presente.

Optimizar el modelo, experimentando con los parámetros que ofrecen los algoritmos usados.
Proponer futuras mejoras, minimizando su margen de error.

4 . Tabla de versionado

I. Análisis univariado	20/04/2022	Primera versión. Manipulación inicial del dataset y análisis univariado con gráficos de matplotlib.
II. Primer entrega	02/05/2022	Planteo de interrogantes en el notebook y análisis bivariado y multivariado con gráficos y heatmaps de matplotlib.
III. Algoritmo de clasificación	09/05/2022	Creación de algoritmos de clasificación: Kneighbors y Random Forest
IV. Segunda entrega	25/05/2022	Responder a las interrogantes de la primera entrega con los algoritmos de clasificación y su respectivo entrenamiento del modelo.
V. Tercer entrega	15/06/2022	Confección de la documentación ejecutiva del modelo. Adaptación de la presentación a las rúbricas y a los comentarios de entregas previas.

Adquisición y análisis del dataset

1. Obtención y fuentes

El dataset para este proyecto fue recolectado de la comunidad kaggle. A continuación se agrega la fuente.

https://www.kaggle.com/datasets/prachi13/customer-analytics

2. Variables del dataset

A continuación, detallamos las características de cada variable y sus unidades de medida.

Variable	Tipo de variable	Detalle
ID	int64	ID único de cada cliente
Warehouse_block (Bloque de almacén)	object	Subdivisión del almacén donde se encuentra el producto (A, B, C, D, E)
Mode_of_Shipment (Modo de envío)	object	Modo de transporte del envío, sea marítimo, aéreo o terrestre. (flight, ship, road)
Customer_care_calls (Llamadas de atención al cliente)	int64	Número de llamados realizados a atención al cliente relacionados a la operación
Customer_rating (Calificación del cliente)	int64	Calificación del cliente en una escala del 1 al 5, siendo 1 el valor más bajo y 5 el más alto.

Variable	Tipo de variable	Detalle
Cost_of_the_Product (Costo del producto)	int64	Costo del producto enviado, expresado en USD.
Prior_purchases (Compras anteriores)	int64	Número de compras anteriores efectuadas por el cliente.
Product_importance (Importancia del producto)	object	Importancia asignada por la compañía al producto (low, medium, high)
Gender (Género)	object	Género del cliente; en este caso, M (male) o F (female).
Discount_offered (Descuento ofrecido)	int64	Descuento otorgado en el producto, expresado en %.
Weight_in_gms (Peso en gms)	int64	Peso del paquete expresado en gramos.
Reached.on.Time_Y.N (Llegó a tiempo)	int64	Variable binaria donde 1 implica que el producto no llegó a tiempo, y 0 que sí lo hizo.

3. Análisis exploratorio

El propósito principal del análisis exploratorio de datos consiste a grandes rasgos en estudiar los datos antes de hacer cualquier supuesto. Puede ayudar a identificar errores obvios, así como comprender mejor los patrones dentro de los datos, detectar valores anómalos atípicos o eventos anómalos, y encontrar relaciones interesantes entre las variables.

Nos sirve como una introducción general para poder trabajar con un mejor entendimiento de los datos, así establecer métricas para la implementación de los modelos de Machine Learning.

Data Acquisition:

Se extrae de la fuente kaggle los datos para trabajar. Se observaron 10999 datos con 12 variables categóricas y numéricas. Importamos las librerías pertinentes para el análisis.

Data Wrangling:

La cantidad de variables del conjunto de datos es de 12 y no se encuentran valores nulos. Sin embargo se hace hincapié en algunas variables más que en otras debido a su mayor relación con los objetivos del negocio, a fin de responder los interrogantes que se plantean inicialmente.

Principalmente haremos enfoque en la variable “Llego a tiempo” (**Reached.on.Time_Y.N**) y su relación con las demás. La variable categórica ID es irrelevante para nuestro análisis.

Análisis Exploratorio:

Estadísticas

A continuación se muestra las estadísticas descriptivas de las variables numéricas:

	count	mean	std	min	25%	50%	75%	max
ID	10999.0	5500.000000	3175.282140	1.0	2750.5	5500.0	8249.5	10999.0
Customer_care_calls	10999.0	4.054459	1.141490	2.0	3.0	4.0	5.0	7.0
Customer_rating	10999.0	2.990545	1.413603	1.0	2.0	3.0	4.0	5.0
Cost_of_the_Product	10999.0	210.196836	48.063272	96.0	169.0	214.0	251.0	310.0
Prior_purchases	10999.0	3.567597	1.522860	2.0	3.0	3.0	4.0	10.0
Discount_offered	10999.0	13.373216	16.205527	1.0	4.0	7.0	10.0	65.0
Weight_in_gms	10999.0	3634.016729	1635.377251	1001.0	1839.5	4149.0	5050.0	7846.0
Reached.on.Time_Y.N	10999.0	0.596691	0.490584	0.0	0.0	1.0	1.0	1.0

Implementación de Machine Learning, métricas y optimización del modelo

1. Variable seleccionada para el modelo

Como teníamos como objetivo principal poder predecir si un producto iba a llegar o no a tiempo según las características del pedido, decidimos escoger **Reached.on.Time_Y.N** como la variable a predecir.

Después de hacer nuestros respectivos análisis bivariado y multivariado, decidimos tomar como principales variables para el modelo,:

1. **Cost_of_the_Product**
2. **Weight_in_gms.**

Esto debido a que ambas variables tenían una buena correlación entre si un producto llegaba a tiempo o no y que las variables enfocadas al cliente no tenían correlación a la variable a predecir.

2. Presentación de algoritmos a utilizar

Para el proyecto tendrémos 3 algoritmos para predecir la variable seleccionada:

1. **Árbol de clasificación**
2. **K Neighbors**
3. **Random Forest**

Consideramos los algoritmos de clasificación como los principales de nuestro proyecto porque podríamos identificar patrones o segmentos para facilitar la predicción de si un producto llegaría a tiempo o no.

2.1. Classification Tree

En este algoritmo lo primero que hicimos fue sacar la variable a predecir **Reached.on.Time_Y.N** y definirla como el target del modelo.

Hicimos la división de la base del 70% para el entrenamiento y un 30% para el test. Entrenamos el modelo y corrimos los cálculos de precisión y fueron los sigueintes:

```
% de aciertos sobre el set de entrenamiento: 0.6699571372905572
% de aciertos sobre el set de evaluación: 0.6803030303030303
```

Observamos que el % de precisión en la evaluación fue mayor y que las dos variables con más importancia fueron:

1. **Cost_of_the_Product**
2. **Prior_purchases**

Validando una de nuestras hipótesis que el costo del producto era una variable con mucha correlación y que nos ayudaría a predecir la variable seleccionada.

2.2. K-Nearest Neighbours

Para este modelo como comentamos anteriormente decidimos usar las siguientes variables:

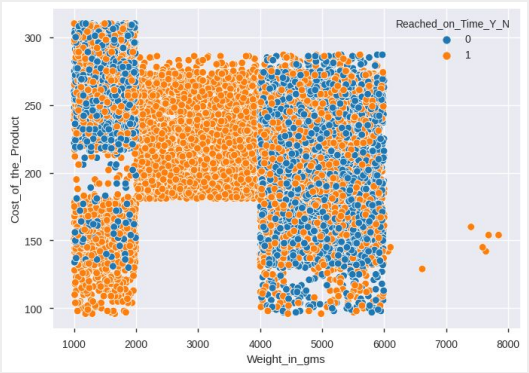
1. **Cost_of_the_Product**
2. **Weight_in_gms**

Ya que al crear un scatterplot pudimos identificar ciertos patrones en los envíos a tiempo que con un algoritmo de KNN nos ayudaría a identificar si un producto llegaría a tiempo o no.

Hicimos un ejercicio con la siguiente información

```
'Weight_in_gms': [3000],
'Cost_of_the_Product': [220]
```

Las cual nos arrojó una predicción de 1 que significa que el producto sí llegaría a tiempo, este es un muy buen algoritmo que podemos seguir usando e implementar en la compañía para identificar productos con un riesgo de no entrega a tiempo y mejorar la experiencia de los clientes.



2.3. Random Forest

Con este algoritmo al igual que con el árbol de clasificación sacamos la variable a predecir **Reached.on.Time_Y.N** y definirla como el target del modelo.

Hicimos una comparación con un nuevo árbol de decisión con uno random, con un random state de 11 en ambos y con 200 estimadores para el random forest y los resultados fueron los siguientes

```
% de aciertos sobre el set de evaluación para DT: 0.9003636363636364
% de aciertos sobre el set de evaluación para RF: 0.6541818181818182
```

Observamos que el % de precisión del árbol de decisión fue mayor y que las dos variables con más importancia del random fueron:

1. **Cost_of_the_Product**
2. **Prior_purchases**
3. **Weight_in_gms**

Quitando la variable de las órdenes pasadas, el costo y el peso siguen siendo las 2 variables con mayor correlación en los modelos.

Conclusiones

Finalmente, una vez revisado los algoritmos elegidos para tener el margen de error más pequeño, decidimos que el algoritmo de árbol de decisión de y random state 11 es el que mejor performance brinda a la hora de trabajar con los datos de esta empresa. Sin embargo, podemos decir que si tomamos los otros modelos y aplicamos un ajuste de parámetros, podremos aumentar la precisión y obtener también buenos resultados.
Con todo el trabajo de análisis exploratorio y la elección de modelos de predicción supervisado, se puede dar una tentativa de respuestas a las preguntas que la empresa se hacía para poder mejorar y ser más eficiente en la logística de productos que comercializa.