Modelo Predicivo - Proyecto Individual

1. Exploratory Data Analyst

1.1. Reconocimiento

La primera instancia consiste en reconocer los datos recibidos, identificando cuantos datos de entrada tenemos y con qué variables. En este caso para el dataset de entrenamiento tenemos 8999 casos y para el test tenemos 2000. En cuanto a las variables, en total son 10, divididos de la siguiente forma.

Variables categóricas.

- Warehouse
- Mode of shipment
- Costumer rating
- Product importance
- Gender

Variables numéricas.

- Costumer Calls
- Cost of the product
- Prior purchases
- Discount
- Weight

La variable objetivo de estudio es si el pedido llego a tiempo (1) o no (0), que esta consignada en la columna de Reached on Time.

Se hace una revisión, donde se ve que no hay valores nulos en los datos.

1.2. Outliers

Se hace un análisis de outliers de las variables númericas, de cost of the product, discount y weight. Para esto se usa la regla de las 3 sigmas

$$rango = \overline{X} \pm 3 * \sigma^2 \tag{1}$$

Se identifica que todos están dentro del rango, sin embargo al hacer el histograma sobre la variable weight se identifican valores fuera de la distribución y que adicional también están fuera del rango de los valores del test, por tanto se descartan.

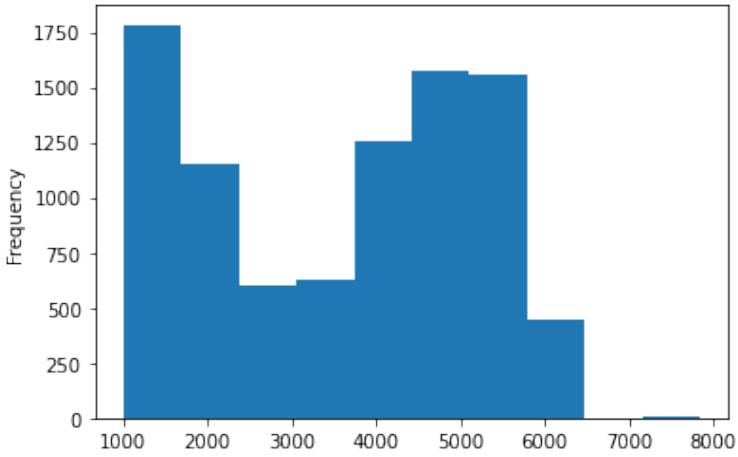

Figura 1: Histograma para la variable Weight

1.3. Relación entre variables

La mejor forma de revisar la relación entre variables es por medio de gráficas donde se contraste las distribución entre ellas. Para eso se usa la herramienta de pair plot, sin embargo para utilizar la herramienta las variables deben ser númericas, por tanto se usa la herramienta de label encoder para hacer una asignación a cada categoria de las variables.

A partir del pairplot, no se ve relación en ninguna de las variables, ni tampoco que alguna tenga una influencia significativa en un resultado favorable o no. Por ejemplo si se revisa el modo de envío, la mayor concentración está en, ya que este fue el medio más utilizado, sin embargo tiene una tendencia igual entre pedidos entregados y no.

Sin embargo se puede destacar que los pedidos que llegaron tarde están concentrados en el menor descuento, indicando que este puede ser uno de los parámetros importante para el modelo. Por otra parte los pedidos de mayor peso y mayor precio tienen una mayor concentración de pedidos entregados tarde.

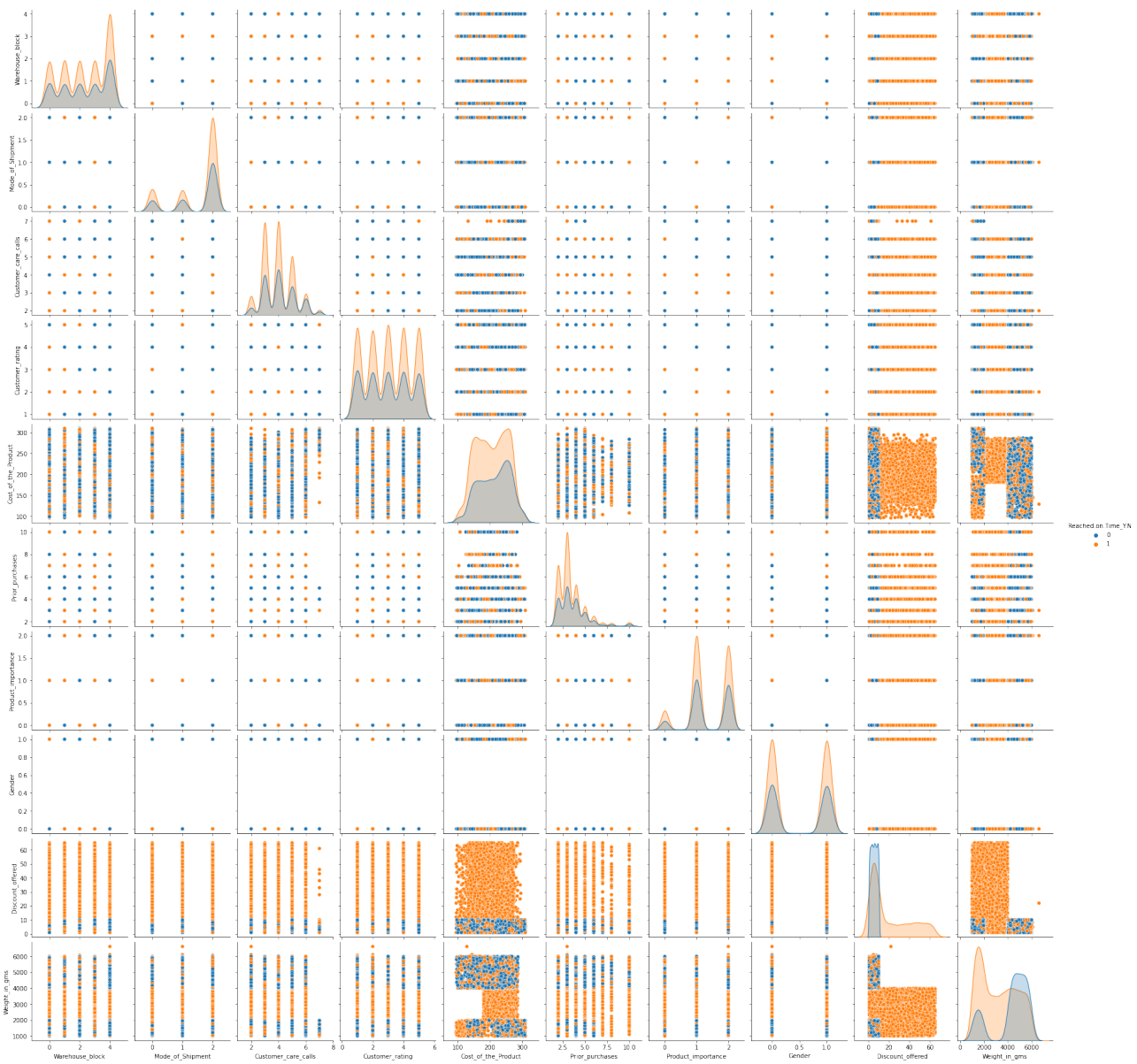

Figura 2: Pairplot del Dataframe

De igual forma se hace una revisión de la matriz de correlación entre variables numéricas, sin embargo no hay ninguna altamente relacionada, lo cual es consecuente con el diagrama de dispersión.

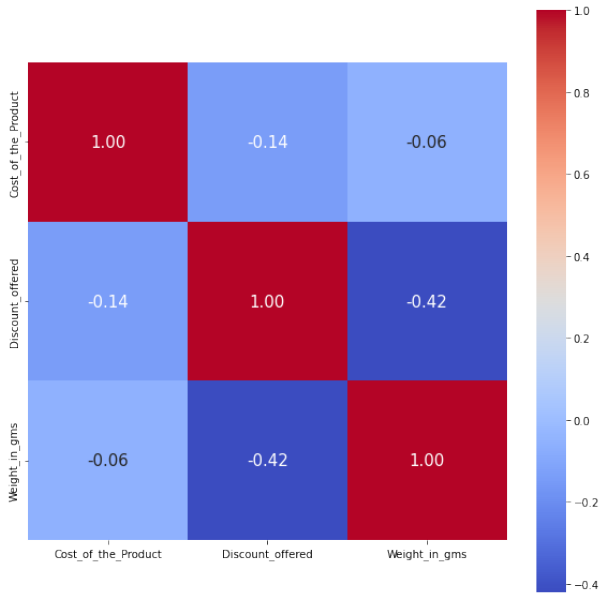

Figura 3: Matriz de Correlación

Del la gráfica del pairplot, se ve un comportamiento de distribución similar entre el Warehouse y el modo de envío. Por ello se hace una prueba chi cuadrado para verificar si son varibles independientes o no. Para ello el primer paso es generar la tabla de frecuencias.

Una vez tenemos la tabla de frecuencias aplicamos la prueba de independencia chi-2, que maneja las siguientes pruebas de hipótesis:

- H_0: Los resultados obtenidos son independientes
- H_1: Los resultados obtenidos son dependientes

Para este caso el valor del p-valor es 1, que para un $\alpha = 0{,}05$ implica que no se rechaza la hipótesis nula, es decir que las variables son independientes.

2. Modelo Predictivo

Lo primero que se debe reconocer para escoger el modelo predictivo, es identificar que tipo de problema estamos enfrentando, que para este caso es de clasificación binaria. Siendo así, para este problema se plantea trabajar con dos tipos de modelo; el de regresión logística y el de support vector machine (SVM). En el primero tenemos una distribución de probabilidad ajustada a una función sigmoide nos da un valor de 0 o 1, mientras que en el otro tenemos un plano que nos separa los valores de 0 y 1.

Para poder ejecutar cualquiera de los algoritmos primero debemos escoger con que variables de va a entrenar el modelo. De acuerdo al análisis hecho previamente, inicialmente se va a trabajar con las variables de: llamadas, costo, descuento y peso. De igual forma se va a hacer predicciones con diferentes variables para identificar con cual tiene mejor rendimiento.

Debido a que se quiere evitar un overfitting en el modelo, de los datos de Train se toma un 70 % para entrenamiento y el restante como testeo. Se evalúa el modelo con las métricas de recall y accuracy, tanto para los datos que se tomaron para el entrenamiento como con los restantes de la evaluación. Es por esto que en las tablas de resultados mostradas a continuación para el test se muestran dos resultados.

Finalmente se hace el entrenamiento de los diferentes modelos. Para el SVM se hizo una prueba de rendimiento para evaluar cuales son los mejores hiperparámetros a usar, que son:

Los resultados obtenidos con los diferentes modelos y parámetros escogidos son:

- C: 1
- Coef0: 0
- Gamma: 0.1
- Kernel: rbf

Intento	**Variables**	**Recall (Train)**	**Accuracy (Train)**	**Accuracy (Test)**	**Recall (Test)**
1	Llamadas Precio Descuento Peso	0.74 0.71	0.67 0.66	0.47	0.66
2	Descuento Peso	0.75 0.72	0.66 0.66	0.47	0.65
3	Precio Peso	0.82 0.81	0.67 0.67	0.45	0.80
4	Llamadas Precio Importancia Peso	0.82 0.80	0.66 0.66	0.45	0.76
5	Llamadas Precio Importancia Descuento Peso	0.74 0.71	0.67 0.66	0.47	0.66

Cuadro 1: Resultados obtenidos con Regresión Logística

Intento	**Variables**	**Recall (Train)**	**Accuracy (Train)**	**Accuracy (Test)**	**Recall (Test)**
1	Llamadas Precio Descuento Peso	0.79 0.77	0.68 0.68	0.48	0.53
2	Descuento Peso	0.65 0.62	0.70 0.69	0.48	0.66
3	Precio Peso	0.71 0.68	0.70 0.69	0.45	0.51
4	Llamadas Precio Importancia Peso	0.70 0.68	0.69 0.69	0.45	0.50
5	Llamadas Precio Importancia Descuento Peso	0.61 0.60	0.71 0.70	0.48	0.52

Cuadro 2: Resultados obtenidos con SVM