Enrique Mejía Fontanot

Examen para el puesto de Sr. Data Scientist en OPI Analytics

Objetivo

Diseñar una solución al problema de optimizar el llenado diario de paletas en máquinas expendedoras dados los costos de transporte y almacenamiento y cumpliendo con mantener por debajo del 2% las máquinas que se quedan sin paletas al final del día.

Variables

Se definen las siguientes variables para conceptualizar el problema, comenzando por las variables que nos permiten ubicarnos en cuanto al tiempo y las máquinas expendedoras de paletas:

- id_maquina = número entero consecutivo identificador de máquina expendedora, el número de máquina va de 1 a 4000
- t = número entero consecutivo identificador del período temporal. Cada número equivale a un día y se sabe que existen al menos 5 años de datos en la base.

Después siguen las variables que se definen siempre para una máquina y tiempo dado. Es decir, que nos dan información del estado de la máquina en ese momento:

- PDI = Paletas disponibles iniciales del período
- PR = paletas retiradas durante el período
- PDF = paletas disponibles al final del período
- C = capacidad de almacenamiento de paletas, donde C es unitario por paleta
- S = cantidad unitaria surtida de paletas al principio del día
- D = variable indicadora del surtido, que toma el valor de 1 si S > 0 y el de 0 si S = 0

Relaciones

Notamos que se deben de cumplir varias condiciones dada nuestra definición del problema. Dada una máquina i y un tiempo t se cumplen:

- Las paletas disponibles al inicio del período t son iguales a las paletas finales en t -1 más el surtido de paletas en t:

$$PDI_{i,t} = PDF_{i,t-1} + S_{i,t} \qquad \text{(ecuación 1)}$$

- Las paletas disponibles al final de t son iguales a las paletas disponibles al inicio más las paletas retiradas durante t:

$$PDF_{i,t} = PDI_{i,t} + PR_{i,t} \qquad \text{(ecuación 2)}$$

Notar que la ecuación 2 también define PDI en función de PDF y PR.

Para estimar la cantidad óptima de paletas a surtir, podemos despejar esta variable de la primera ecuación para observar que se trata de la diferencia entre las paletas disponibles al inicio de t y las disponibles al final de t-1:

$$S_{i,t} = PDI_{i,t} - PDF_{i,t-1} \qquad \text{(ecuación 3)}$$

Notamos que la cantidad a surtir depende de la diferencia entre las paletas que quedaban en t-1 y las que estarán disponibles al inicio de t. Es decir que la cantidad ideal de surtido está en función de la demanda de paletas que serán retiradas de cada máquina y de las paletas que quedan en la máquina al final del periodo anterior.

Necesitamos una tercera ecuación, que incorpora la restricción de máquinas sin paletas disponibles dado nuestro modelo de negocio. Para un periodo t dado, la suma de las variables indicadoras D entre el total de máquinas n debe ser menor a 0.02.

$$\delta = \frac{1}{n}\sum_{i=1}^{n} D_i < 0.02 \qquad \text{(ecuación 4)}$$

Tabla de datos históricos de consumo de paletas

Sabemos que contamos con 5 años de datos históricos para cada máquina. Estos se pueden poner en una tabla donde cada renglón corresponde a una máquina y a un período dado.

id_maquina	t	PDI	PR	C	S	PDF	D
1	1	300	100	300	300	200	1
1	2	200	50	300	0	150	0
1	3	150	50	300	0	100	0
1	4	100	50	300	0	50	0
1	5	200	50	300	200	150	1

Notamos que se trata de una serie de tiempo de 5 años para cada máquina, lo que nos permitirá explotar esta información para predecir las paletas retiradas del siguiente periodo dado el anterior.

Definición de la función objetivo

Definimos el problema de la siguiente manera. Dado que conocemos el estado final de t, para t+1, minimizar la función de costos (suma de costos de almacenamiento más suma de costos de transporte) cumpliendo la restricción de paletas disponibles. Es decir:

Minimizar

$$1\sum_{i=1}^{n} PDI_{i,t+1} + 100\sum_{i=1}^{n} D_{i,t+1} \qquad \text{(función de costos operativos)}$$

Sujeto a

$$\delta = \frac{1}{n}\sum_{i=1}^{n} D_{i,t+1} < 0.02$$

Esbozo de la solución

Ya que ambas partes de la función objetivo del problema dependen de una variable que desconocemos (la cantidad de paletas retiradas de t+1), no podemos encontrar una solución analítica de manera directa. Por ende, analizamos el problema para encontrar una heurística que funcione dados nuestros datos históricos tan vastos. La solución se explica a continuación:

Para t+1, predecir $PR_{i,t+1}$ para determinar $S_{i,t+1}$ y $PDI_{i,t+1}$ de manera óptima.

Ya que contamos con una serie de tiempo para cada máquina, podemos ajustar un modelo predictivo (que puede ser un modelo autorregresivo como un ARIMA) para estimar la demanda del día siguiente. Una vez teniendo la estimación, podemos minimizar los costos de almacenamiento y transporte dadas las condiciones que tenemos que cumplir.

Dado $\hat{PR}_{i,t+1}$ (valor predicho de las paletas retiradas), usamos la ecuación 3 para determinar el valor de $\hat{S}_{i,t+1}$ y la ecuación 1 para determinar el valor de $\hat{PDI}_{i,t+1}$. Al final del período t+1, se debe comprobar que se cumplió la condición determinada por la ecuación 4. De lo contrario, se toma la solución como inválida. Y se debe ajustar el modelo para predecir mejor las paletas retiradas y ofrecer una nueva solución hasta que esta cumpla con la restricción de δ.

Es importante notar que, la información acumulada de cinco años de operación de surtido de nos permite probar nuestra solución sin tener que cambiar la operación diaria. El primer paso sería desarrollar el modelo predictivo y calcular el vector de la cantidad predicha que se debe surtir a cada máquina. Se puede seguir con la operación normal, y comparar los resultados de la solución normal versus la solución predictiva. Se deben obtener métricas históricas de ambas soluciones, que registren el porcentaje de máquinas que se quedaron sin paletas y el costo operativo de cada día. Se debe seguir ajustando el modelo predictivo, hasta que este supere a la operación normal y genere la confianza de que se está respetando la restricción de mantener a los clientes siempre contentos y sin encontrarse con máquinas sin paletas disponibles.

Para evaluar si la solución tiene impacto positivo, se puede calcular la diferencia en costos operativos entre ella y la solución anterior. Se puede además comparar contra otros modelos predictivos para hacer un proceso de mejora continua. Ya que no sabemos a priori cuál será el mejor modelo, debemos seguir monitoreando estas métricas de manera permanente para poder seguir mejorando la solución.