**GRUPO 05**

# MODELO DE PREDICCIÓN

## Web Scraping para extracción de feriados y días no laborables

Riabykh et al. (2022) indica que los días festivos y los días contiguos tienen una gran influencia en los retiros realizados en ATMs. Por lo que se decidió explorar detalladamente el impacto de las fechas festivas en los patrones de retiro en ATMs, abordando cuestiones como la cantidad de días festivos, tipos de celebraciones, duración y popularidad de estas. Por lo que, **se desarrolló con anterioridad** una herramienta de web scraping y así, se extrajo, antes de la competencia, los días festivos, y así calcular el día festivo, el día anterior y siguiente al festivo, así mismo del Holiday Sequence que será explicado luego. De esta manera, se busca mejorar la precisión y confiabilidad de los resultados obtenidos.

## Análisis y limpieza de datos

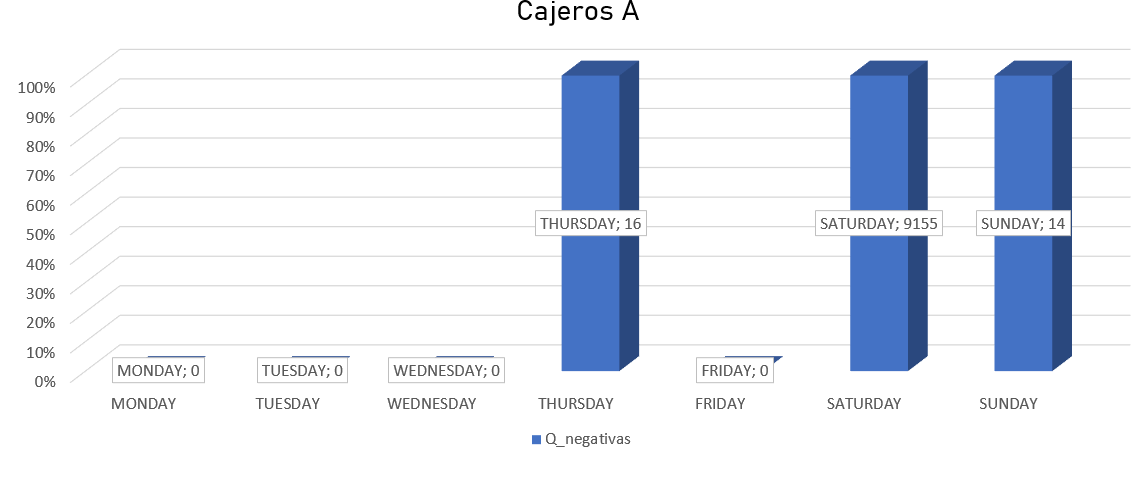
### Manejo de datos atípicos

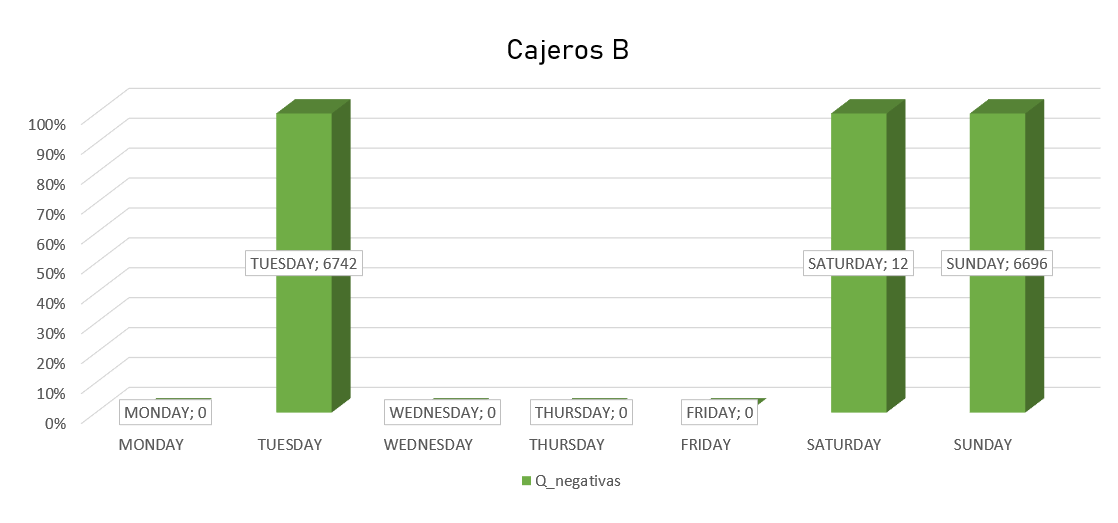
Antes de empezar con el entrenamiento del modelo se tomó acciones frente a los casos de datos atípicos.

En primer lugar, los registros que contienen como dato de demanda de este día un valor nulo o directamente 0 fueron eliminados directamente de los datos al representar el 0.17% del total de datos.

Por otro lado, para los casos de registros que contiene como dato de demanda de ese día un valor negativo se optó por simplemente convertir el valor a positivo. Esta decisión se debe a que, en primer lugar, se realizó un análisis por tipo de cajero sobre qué días de semana son los más propensos a tener un valor negativo como demanda. El resultado de esto reveló que para el caso de los cajeros te tipo ‘A’, los días que alguna vez han tenido demanda negativa solo son los jueves, sábados y domingos, siendo el sábado el día mayoría de estos registros; mientras que los días en donde se han registrado demandas negativas para los cajeros de tipo ‘B’ son los martes, sábados y domingos, siendo en este caso los martes y domingos los que contienen la mayoría de los registros.

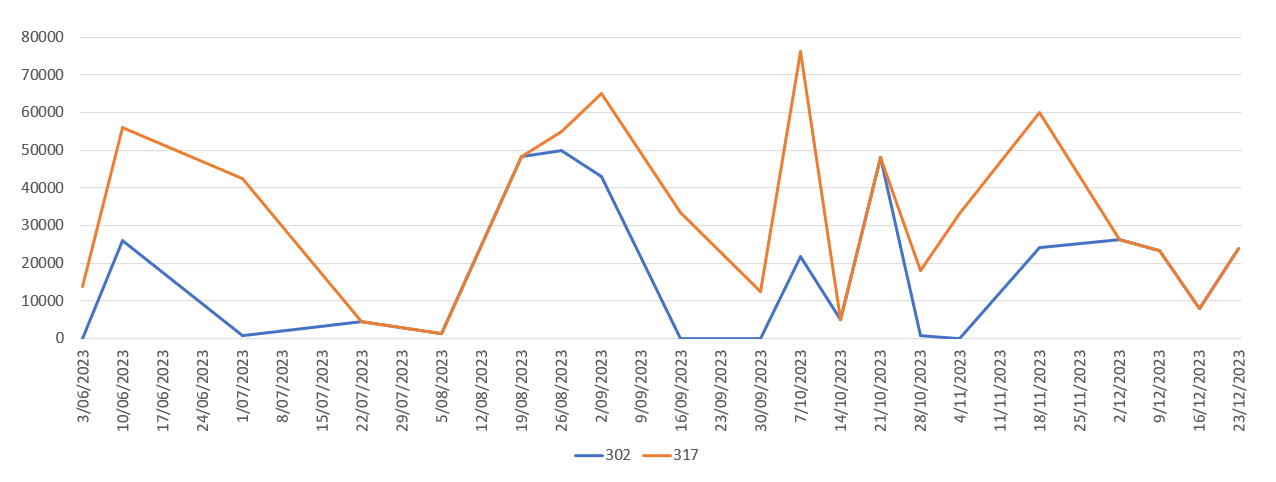
Los resultados de esto se pueden ver en los gráficos a continuación.





A raíz de este análisis se optó por usar una función Rolling\_mean entre los registros con datos positivos para poder simular la demanda que debió ser estos días de demanda negativa. Sin embargo, al revisar estos datos a usar se pudo observar la irregularidad de estos, provocando que los datos a generar por la función no tengan la mejor credibilidad y posiblemente dañen el rendimiento del modelo de predicción.

A continuación, se muestra un gráfico de ejemplo de los registros de demanda positivos de los días sábado de 2 cajeros de tipo ‘A’ para mostrar la irregularidad hallada en los datos.

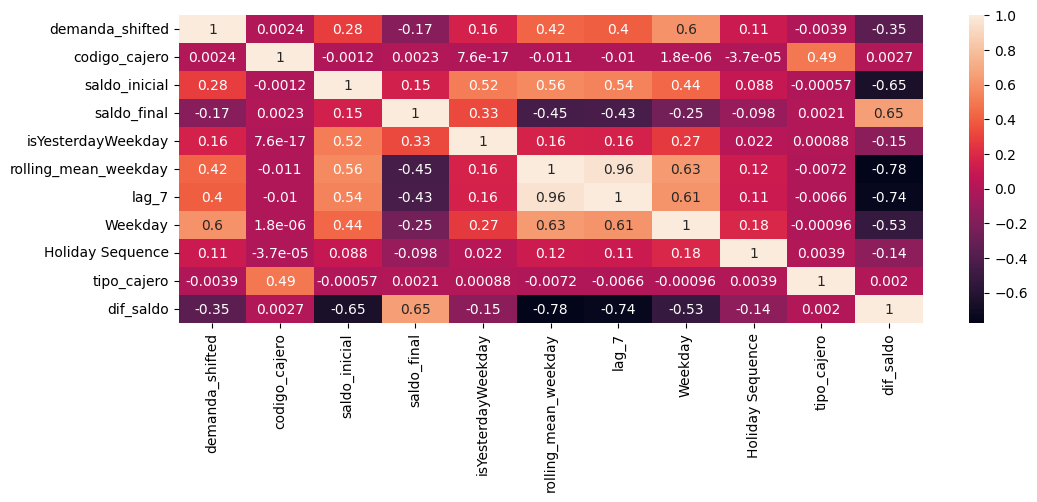


### Adición de variables adicionales

Según las investigaciones de Asad et al. (2020), Gorodetskaya et al. (2021) y Riabykh et al. (2022) los anticipos y fechas de pago son variables que influyen en la demanda de los ATM. Por lo que en este archivo se calcularán y agregarán estas variables. Además, En Perú, generalmente, la fecha de pago se realiza el último día del mes, siempre y cuando, sea un día laborable (día de semana desde lunes hasta viernes) y no sea feriado. En caso se incumpla estos criterios, se calcula la fecha anterior más cercana que sí cumpla. Po otro lado, Asad et al. (2020) indica que la semana de pago podría ser más precisa que el día específico, como fue calculado anteriormente, por lo que para calcular la semana de pago, en base al último día del mes y quincena, se capturará los 3 días anteriores y los 3 días en adelante, todos ellos serán considerados como la semana de pago. Asimismo, se crearon variables que dependen de los días anteriores como la demanda promedio del mismo dos días del pasado (es decir, el promedio de dos Jueves pasados o dos miércoles). La demanda máxima de la semana, la demanda de hace 7 días, entre otros. La inclusión de todas estas variables está sustentada en la literatura revisada y se encuentras a continuación:

|  |  |
| --- | --- |
| **Variable** | **Descripción** |
| Type | Tipo de dia (Laborable, feriado, feriado sector público, etc.) |
| Weekday | Dia de la semana (Lunes, martes, miércoles …, domingo) |
| Holiday Sequence | Tipo de dia anterior, tipo de día hoy, tipo de día mañana |
| isYesterdayHoliday | ¿Ayer fue feriado? (True / False) |
| isHoliday | ¿Hoy es feriado? (True / False) |
| isTomorrowHoliday | ¿Mañana es feriado? (True / False) |
| isYesterdayWeekday | ¿Ayer fue dia laborable? (True / False) |
| isTomorrowWeekday | ¿Mañana es dia laborable? (True / False) |
| isWeekday | ¿Hoy es dia laborable? (True / False) |
| isPaymentDay | ¿Es el día de pago (quincena o fin de mes)? (True/False) |
| isPayweek | ¿Es la semana de pago (quincena o fin de mes)? (True/False) |
| rolling\_mean\_weekday | Demanda promedio de los últimos 2 días del mismo día actual (Por ejemplo: Si hoy es jueves, esta variable será el promedio de demanda de los últimos 2 jueves |
| rolling\_max | Valor máximo hasta ahora de esta semana |
| lag\_7 | La cantidad de demanda 7 días antes. |
| rolling\_max\_weekday | La demanda máxima de los últimos 2 días del mismo día actual. |
| rolling\_std | Desviación estandar semanal |
| rolling\_mean | Demanda promedio mensual. |

Posteriormente a esto, se realizó un análisis de correlación de las nuevas columnas frente a la columna objetivo, esto para poder garantizar que las nuevas adiciones sean verdaderamente de ayuda. El resultado de este proceso se puede observar en la figura a continuación.



Sintetizando los resultados del análisis de correlación, las columnas nuevas que presentaron mejor correlación con la columna objetivo son ‘rolling\_mean\_weekday’ y ‘lag\_7’, seguidas por ‘weekday’.

Sin embargo, el hecho de que las variables ‘rolling\_mean\_weekday’ y ‘lag\_7’ presenten una correlación muy similar a la columna objetivo puede causar multicolinealidad en estas columnas. Este problema es causado gracias a que ambas variables son de naturaleza muy parecida, siendo una el promedio móvil de 2 días sobre el día actual y la otra el valor del día de la semana pasada sobre el día actual, por lo tanto, se optó por usar solamente la columna ‘rolling\_mean\_weekday’ debido a que tiene más sustento al ser mencionada en estudios similares al caso actual.

Además, para las variables categóricas, se aplicó la técnica de OneHotEncoding y TargetEncoding.

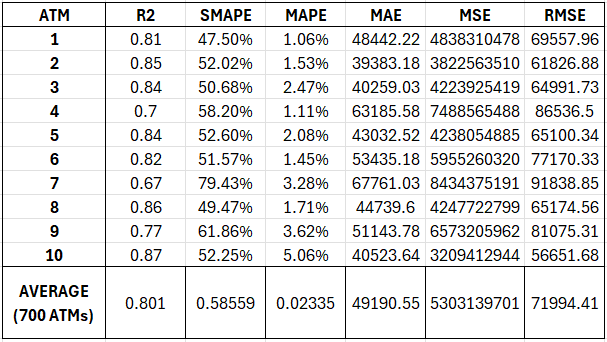
## Modelo de Predicción

Entre los modelos probados para la predicción de la demanda de ATM’s, Random Forest Regressor fue el que tuvo el menor índice de error. Los datos no fueron escalados ya que por la naturaleza del modelo, realizar esta tarea podría mantener o empeorar ligeramente los resultados, por lo que no fue realizado. Además, se separó los datos en entrenamiento, validación y test, en la proporción de 80%, 10% y 10% respectivamente. Asimismo, se utilizó GridSearchCV para la selección de los mejores hiperparámetros, los parámetros son mostrados a continuación.

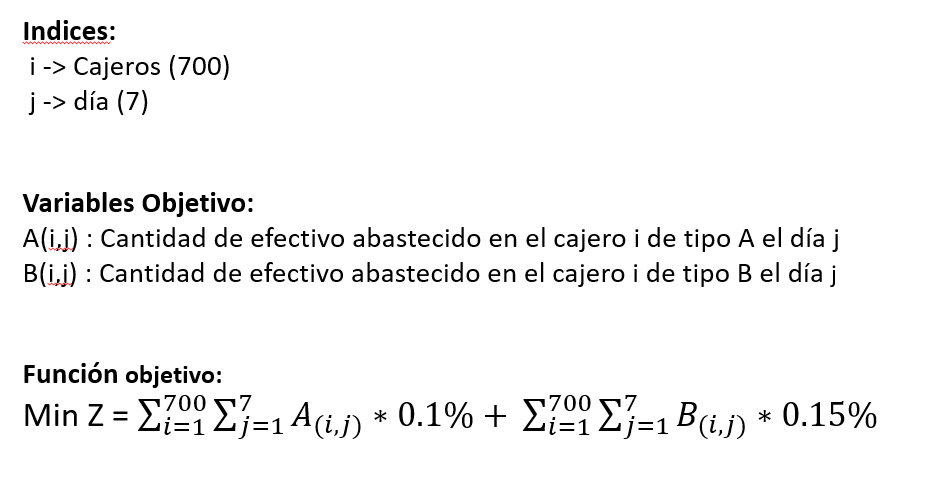
|  |
| --- |
| rfr = RandomForestRegressor(random\_state=2024, verbose=0, n\_jobs=5)          param\_grid = {              'n\_estimators': [100, 200, 300],              'max\_depth': [None, 10, 20],              'min\_samples\_split': [2, 5, 10],              'min\_samples\_leaf': [1, 5, 10]          }            grid\_search = GridSearchCV(rfr, param\_grid, cv=5, scoring='neg\_mean\_squared\_error', n\_jobs=5) |

## Resultados

Se calcularon las métricas R2, SMAPE, MAPE, MAE, MSE, y RMSE para las predicciones realizadas por el modelo Random Forest Regressor. En la siguiente tabla, se muestra el desempeño de los primeros 10 ATMs y promedio general de los 700 ATMs.



# MODELO DE OPTIMIZACIÓN DE ABASTESIMIENTO

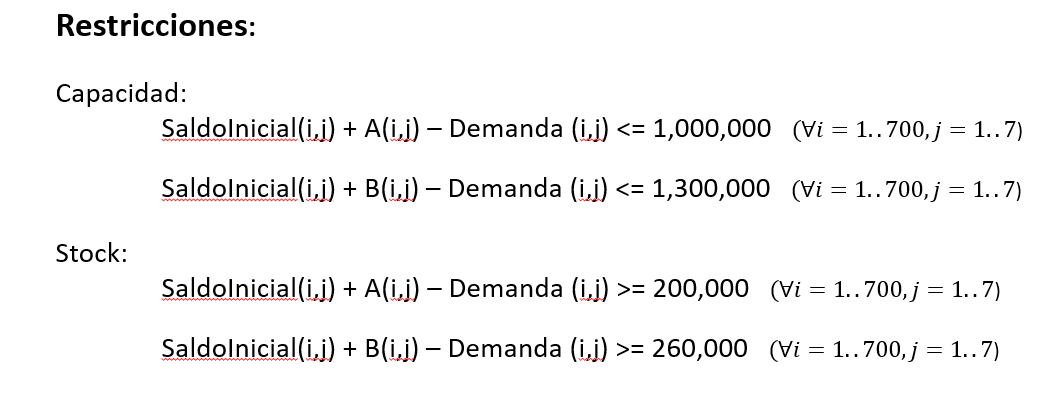


En primer lugar, se establecerán los índices, como son 700 cajeros, se establecerá un índice i desde el 1 hasta el 700 para hacer referencia a los cajeros. Así también se hará un índice j, referente al día de la semana donde se encuentra.

Nota: Los indices aquí van a tomar un valor de 1 a 301 para B y un valor de 302 a 700 para A, esto es porque los primeros 301 cajeros son de tipo A y los siguientes son de tipo B.

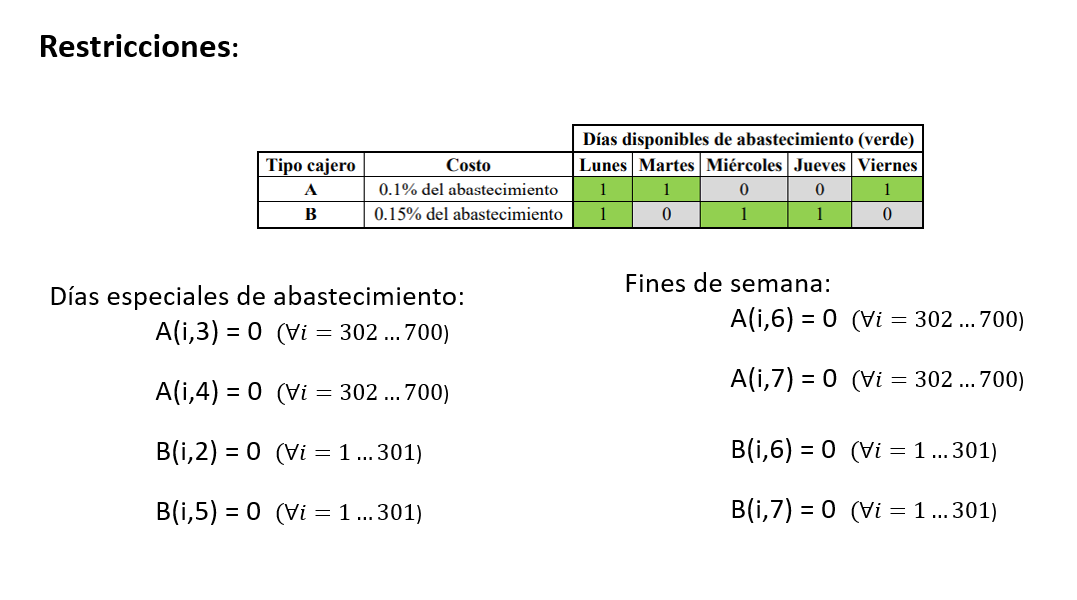
Los índices son esenciales para establecer las variables objetivo:

* A(i,j) Cantidad de efectivo en el cajero i de tipo A el día j
* B(i,j) Cantidad de efectivo abastecido en el cajero i de tipo B el día j



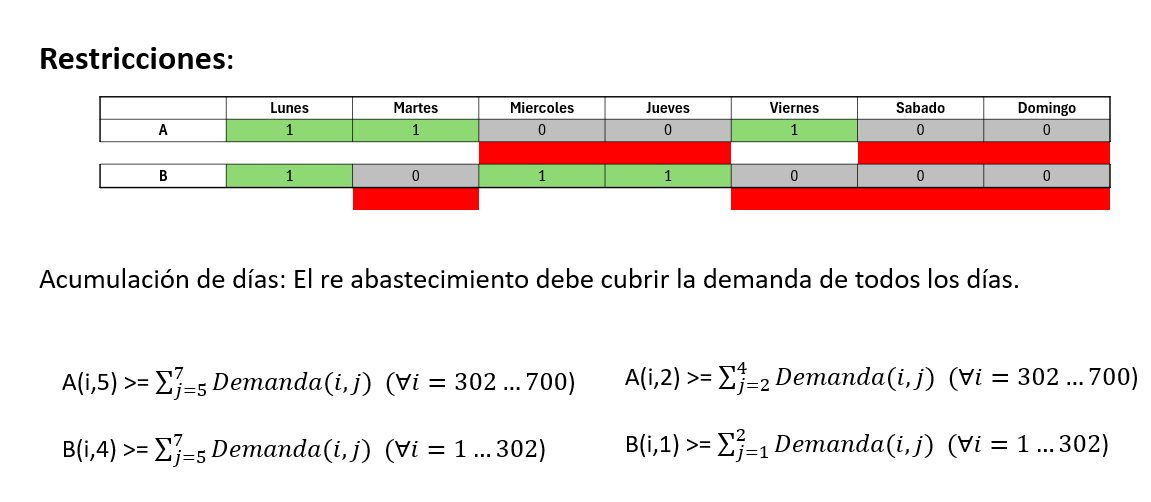
Para ello se plantearán restricciones. En primer lugar, una restricción respecto a su capacidad máxima, que representa la suma entre el saldo inicial en el cajero i el día j más la cantidad a reabastecer en el cajero A con id i en el día j, debe ser menor a igual a 1 millon, debido a que tiene su capacidad máxima de abastecimiento. Asimismo se debe hacer la restricción con el cajero B, con el saldoinicial(i,j) mas la cantidad de efectivo a reabaastecer menos la demanda (i,j) es menor a

Para la siguientes restricciones, se plantea el stock mínimo, para el cajero A, este debe tener una capacidad superior a 200000 que es el monto mínimo y para el B debe tener un valor mayor a 260000.

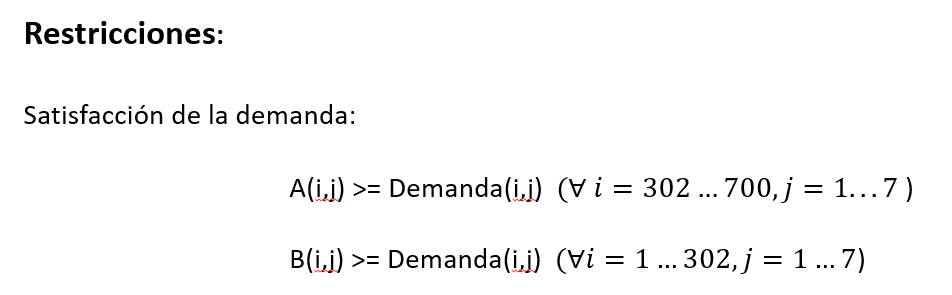


Las siguientes restricciones son respecto al re abastecimiento. Solo se puede re abastecer ciertos días según el tipo de cajero. Los cajeros A se pueden re abastecer lunes, miércoles, y viernes. Por lo que el resto de días, no se re abastecerán y tomarán un valor de 0.

Y respecto a los días fines de semana, también se asignan 0 porque no se pueden re abastecer esos días.



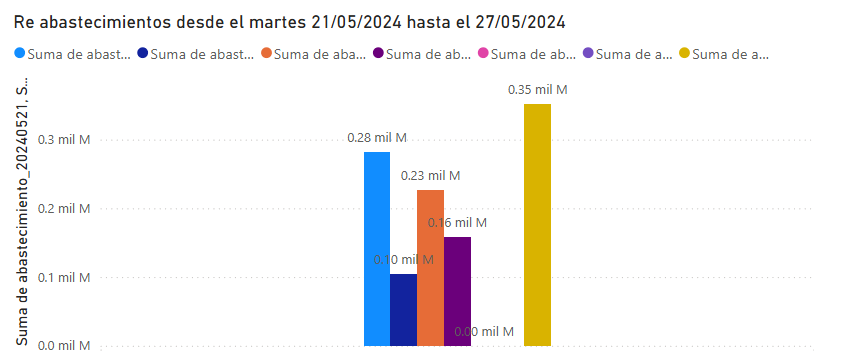
Las siguientes restricciones son respecto a las demandas de los días que no se re abastece, por ejemplo para los de tipo A, son los miércoles, jueves sábado y domingo, por lo que el monto re abastecido un día antes deben ser mayor a la demanda de esos días para poder suplirla. Lo mismo para los cajeros tipo B para los días martes, viernes, sábado y domingo.



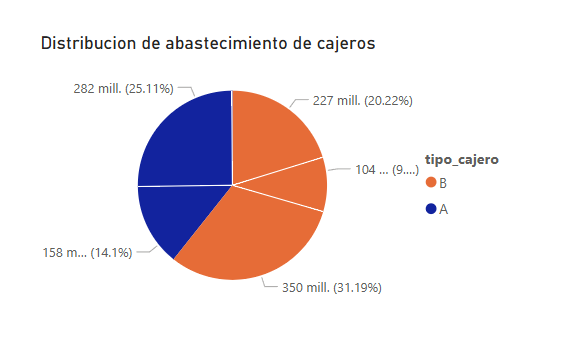
Por último, el monto re abastecido siempre será superior a la demanda establecida a ese día según la predicción para que la pueda re abastecer.



Finalmente al ejecutar el modelo, este es el costo final del re abastecimiento de los cajeros



Se puede visualizar la distribución desde el martes 21/05/2024 hasta el lunes 27/05/2024. De esta gráfica se puede visualizar como los días de mayor entrega en orden son los lunes, seguido de los martes y seguido de los miércoles. Esto puede hablar de como las restricciones se emplean correctamente para asignar cuando es posible distribuir por tipo de cajero.



En el presente gráfico, se puede visualizar la distribución de efectivo entre cajeros, asignadas por días, el tipo A exige 440 millones de efectivo, mientras que el tipo B exige 681 millones de efectivo. La diferencia entre los cajeros es de 224 millones, resultando un costo diferencial entre cajeros al banco de 33.6 millones de soles por el costo mayor de reabastecimiento del cajero tipo B

# CONCLUSIONES Y RECOMENDACIONES

Para el desarrollo y mantenimiento del modelo se considera lo siguiente:

CONCLUSIONES:

* Los valores atípicos generan ruido y su correcto trato es esencial para el modelo y la optimización.
* Considerar los Rolling\_mean (promedios móviles) de la demanda en diferentes periodos de tiempo.
* Los días de mayor distribución son los lunes, martes y jueves para satisfacer la demanda de cajeros en los días de no reabastecimiento
* La mayor demanda de efectivo viene por el cajero tipo B, el cual presenta un mayor costo por reposito pudiendo elevar los costos totales

RECOMENDACIONES

* Investigar el impacto de los días feriados: De acuerdo con la literatura, estos tienen cierto impacto en la predicción de la demanda, de manera que sería ideal poder corroborar al 100% esto con información no simulada.
* Desagregar los ATMs por información geográfica y poblacional: sería interesante poder analizar los cajeros según su ubicación, zonas y lugares (malls, cerca a mall, alejados, etc); además, considerar el impacto del sueldo promedio de zonas podría mejorar las métricas del modelo.
* Realizar clustering ATMs y analizar el impacto de un modelo individual por centroide y posiblemente ensamblar.
* Evaluar una estrategia de reducción de costos de reabastecimiento del cajero tipo B ya que cuenta con una mayor demanda .
* Considerar estrategias para mejorar la oferta de cajeros tipo A y hacerla más atractivas a los clientes ya que cuenta con costos de redistribución mas bajos.