Modelo Predictivo para la Gestión Eficiente del Arribo de Embarcaciones - Waypoint LLC

1. Introducción

Antecedentes

Waypoint LLC, fundada en 2010 por expatriados que brindaron apoyo al Ejército de EE. UU. en las guerras de Afganistán e Irak, posee una destacada trayectoria en logística y apoyo a misiones, habiendo participado en más de 8000 proyectos a nivel mundial. La compañía se especializa en soluciones logísticas en puertos y lugares de difícil acceso para el gobierno de EE. UU. en entornos desafiantes. Con una licencia global obtenida en 2020, Waypoint está habilitada para atender navíos de la Armada de Estados Unidos en 30 regiones del mundo hasta 2030.

Desde entonces, la compañía ha estado expandiendo sus operaciones internacionales, logrando para 2024 un incremento del 13,2% en su captación del mercado y posicionándose como una de las tres empresas más importantes del sector.

2. Descripción general del problema

Contexto del problema

El funcionario encargado de solicitar suministros o servicios debe primero completar un formato con el código del servicio requerido. Este formato es enviado a la oficina central, donde se revisa y distribuye entre todos los licitantes autorizados en un modelo de subasta invertida. En este modelo, gana el licitante que pueda entregar todos los servicios requeridos en el tiempo requerido al menor costo. Estas ofertas pueden tener un tiempo de preparación que varía desde el mismo día de la licitación.

Waypoint LLC busca ganar más licitaciones en su unidad de negocio de atención a barcos (GMAC). Para lograrlo, necesita hacer inversiones estratégicas que le permitan tener una capacidad de reacción rápida y disponer de los insumos suficientes para cubrir las demandas de sus clientes en el tiempo y con la calidad esperados. Con este fin, Waypoint desea predecir dónde será la próxima oferta. Para ello, comparte un dataset de más de 8000 ofertas de distintos barcos y lugares alrededor del mundo en los periodos de 2021 - 2024, lo que permite evaluar diversos escenarios y planificar con más tiempo.

Alcance del proyecto

El equipo de tomadores de decisión de Waypoint tiene como objetivo resolver la siguiente cuestión:

"¿Cómo puede Waypoint predecir a qué puerto llegará un barco seleccionado por el usuario?"

Para lograr esto, se plantea abordar las siguientes necesidades:

1. Análisis Descriptivo del Barco: Basado en información histórica que incluye:
   - Tiempo en puerto
   - Tiempo en altamar
   - Tiempo interarribo
   - Puertos más transitados
2. Predicción de Arribo utilizando ML:
   - Implementar herramientas de Machine Learning para predecir el puerto de arribo del barco seleccionado.
   - Proporcionar un porcentaje de certidumbre en la predicción realizada.

3. Descripción de los datos

El dataset se encuentra en Github, estará dispuesto para ser versionado, este consta de un conjunto de 8837 muestras con 16 dimensiones listadas a continuación:

Contenido del dataset

Dimensión	Tipo de Valor	Descripción
RTOP	String	Request Task Order Proposal o identificador único de oferta
Project #	String	Identificador único de proyectos ganados usados al interior de la compañía
Combo	String	Campo que agrupa cuando varias ofertas se licitan en un único conjunto
Status	String	Status de la oferta manejado internamente por la compañía, este puede ser: - New Bid=Oferta nueva sin trabajar - Build=Oferta en proceso de diseño - Submitted=Oferta sometida y en espera de recibir respuesta - Not Submitted=Oferta que no fue licitada por parte de la empresa

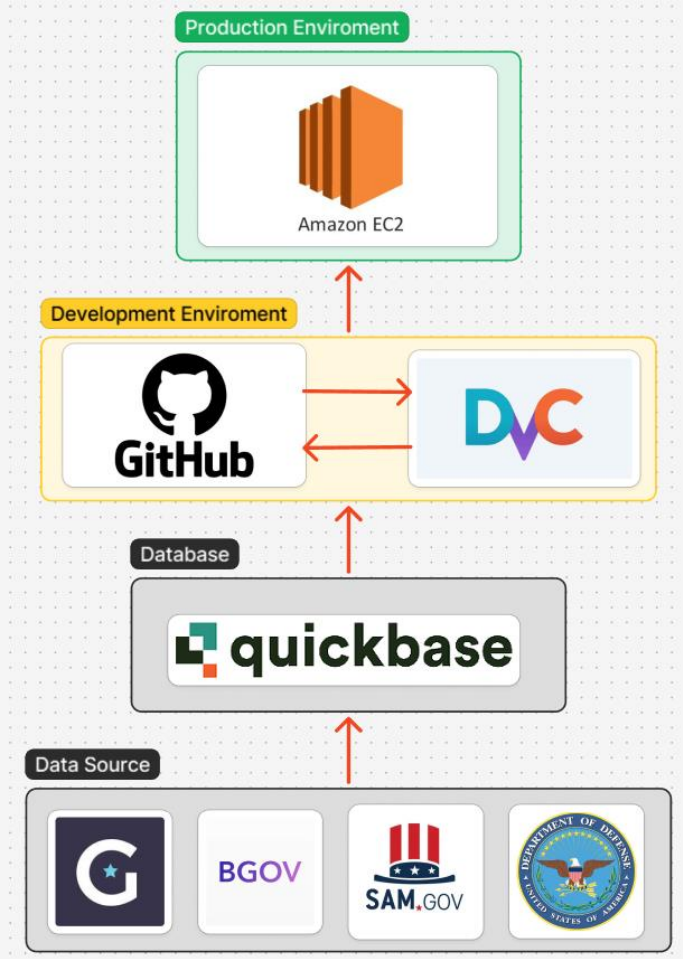
		- Won=Oferta ganada por la compañía - Lost=Oferta pérdida por la compañía - Cancelled=Oferta de la que el gobierno se retracta de licitar
Ship - Name	String	Nombre del barco
Ship ID - Ship Classification	String	Tipo de estructura de barco
Responsible	String	Personas responsables en ejecutar la oferta a nivel interno
Region Name	String	Region a la cual pertenece la oferta
Country	String	País a la cual pertenece la oferta
Business Unit	String	Unidad de negocio a la cual pertenece la oferta (GMAC, WEXMAC, PROVISIONS)
Awarded Company	String	Compañía que gana la licitación
Arrival	Date	Llegada del barco a puerto
Departure	Date	Salida del barco a puerto
Deadline(mmdd yyyy)	Date	Fecha máxima para poder licitar
Time to Execute	String	Tiempo entre Arrival y Departure

Metodos de recoleccion y almacenamiento

Los datos fueron recolectados a través del tiempo por la compañía en su propio data warehouse y serán extraídos de allí. Estos pueden ser conectados en un futuro a la herramienta de elección por medio de API's para tener información en tiempo real. También se usarán herramientas externas como govtribe o Bloomberg Government para hallar y complementar ofertas que tengan datos incompletos.

Herramientas y su uso

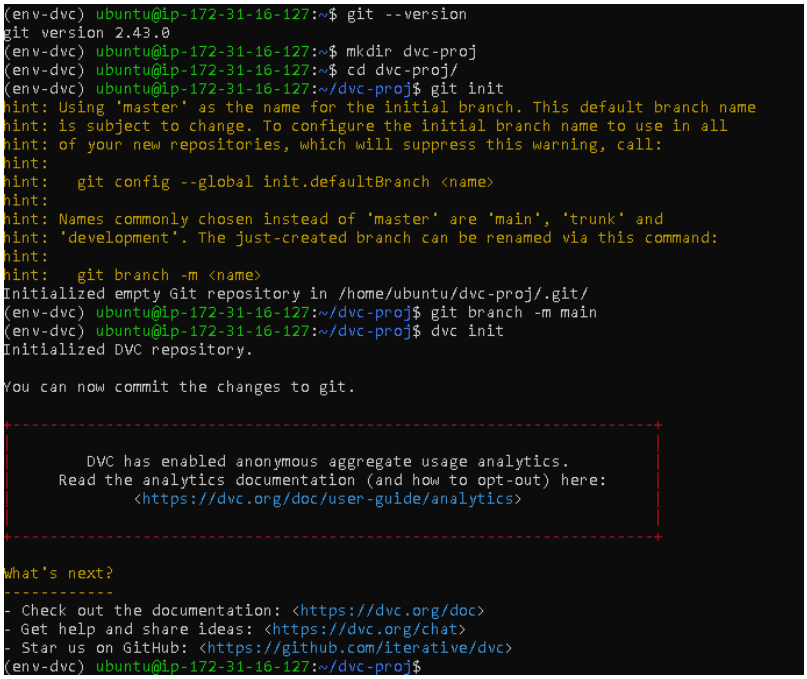
Para este despliegue se considera el data warehouse de la compañía, que usa tecnología de RESTful API de QuickBase (QB). También se tendrá en cuenta el versionado de los datos que se puede hacer tanto en QB como en Github por medio de la librería DVC de Python. El código de la app se desarrollará en un documento de Jupyter como base técnica y la app será versionada en GitHub, la cual se desplegará en EC2 de AWS.



Repositorio

GitHub: https://github.com/anfisbena/MIAD-DSA/tree/main

Este incluye el repositorio de DVC. A continuación se presenta una evidencia de su creación:



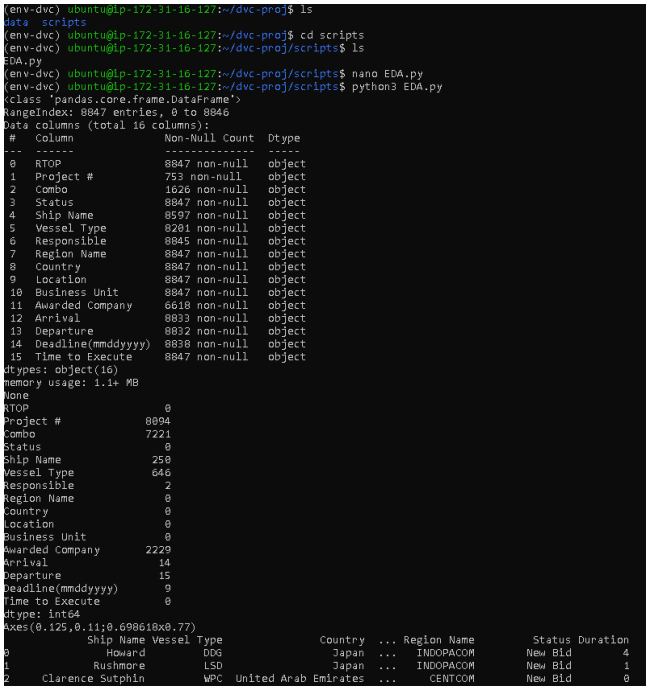
Se realizó la configuración inicial del proyecto de control de versiones usando Git y DVC en el entorno Ubuntu. Mediante el comando mkdir dvc-proj creamos el directorio para el proyecto donde inicializamos y configuramos Git para el control de las versiones en el directorio creado. Cambiamos la rama predeterminada a main utilizando git config --global init.defaultBranch main y confirmamos la rama principal con git branch -m main. Se inicializa DVC con el comando dvc init para gestionar el versionado de datos dentro del proyecto, completando así la configuración necesaria para el manejo de versiones tanto de código como de datos.

Estándares de metadatos

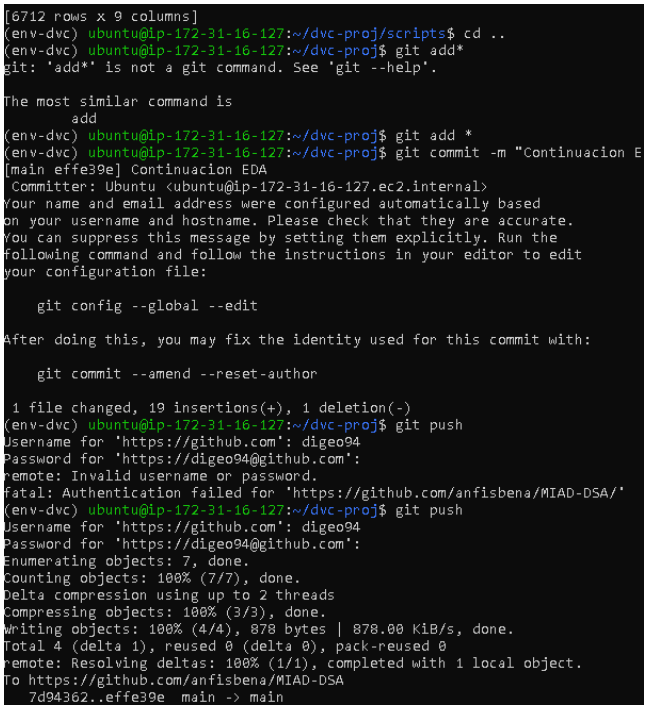
Todo metadato de la información recolectada debe ser etiquetado en su correspondiente dimensión y almacenado en centros de datos seguros que cumplan con los requisitos del contrato de licitación. La información debe ser almacenada por no menos de 7 años, siguiendo los protocolos de acceso y seguridad anteriormente mencionados.

Estrategia de respaldo

Los datos serán respaldados regularmente por el proveedor warehouse QuickBase (QB), el cual tiene mecanismos automáticos para garantizar la seguridad de la información. Dado que la aplicación actúa como una interfaz temporal, los datos utilizados se procesarán sin necesidad de respaldo adicional por parte de la aplicación. Además, el equipo responsable gestionará la seguridad y los respaldos, cumpliendo con los estándares de protección y asegurando que la información sea accesible solo a usuarios autorizados para minimizar riesgos durante la ejecución del proyecto.



```
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ ls
data  scripts
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ cd scripts
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj/scripts$ ls
EDA.py
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj/scripts$ nano EDA.py
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj/scripts$ python3 EDA.py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8847 entries, 0 to 8846
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   RTOP                8847 non-null   object
 1   Project #           753 non-null    object
 2   Combo               1626 non-null   object
 3   Status              8847 non-null   object
 4   Ship Name           8597 non-null   object
 5   Vessel Type         8201 non-null   object
 6   Responsible         8845 non-null   object
 7   Region Name         8847 non-null   object
 8   Country             8847 non-null   object
 9   Location            8847 non-null   object
 10  Business Unit       8847 non-null   object
 11  Awarded Company     6618 non-null   object
 12  Arrival             8833 non-null   object
 13  Departure           8832 non-null   object
 14  Deadline(mmddyyyy)  8838 non-null   object
 15  Time to Execute     8847 non-null   object
dtypes: object(16)
memory usage: 1.1+ MB
None
RTOP                   0
Project #           8094
Combo               7221
Status                 0
Ship Name            250
Vessel Type          646
Responsible            2
Region Name            0
Country                0
Location               0
Business Unit          0
Awarded Company     2229
Arrival               14
Departure             15
Deadline(mmddyyyy)     9
Time to Execute        0
dtype: int64
Axes(0.125,0.11;0.698618x0.77)
          Ship Name Vessel Type              Country  ... Region Name         Status Duration
0             Howard         DDG                Japan  ...    INDOPACOM        New Bld        4
1           Rushmore         LSD                Japan  ...    INDOPACOM        New Bld        1
2     Clarence Sutphin         WPC  United Arab Emirates  ...     CENTCOM        New Bld        0
```



```
[6712 rows x 9 columns]
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj/scripts$ cd ..
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ git add*
git: 'add*' is not a git command. See 'git --help'.

The most similar command is
        add
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ git add *
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ git commit -m "Continuacion E
[main effe39e] Continuacion EDA
 Committer: Ubuntu <ubuntu@ip-172-31-16-127.ec2.internal>
Your name and email address were configured automatically based
on your username and hostname. Please check that they are accurate.
You can suppress this message by setting them explicitly. Run the
following command and follow the instructions in your editor to edit
your configuration file:

    git config --global --edit

After doing this, you may fix the identity used for this commit with:

    git commit --amend --reset-author

 1 file changed, 19 insertions(+), 1 deletion(-)
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ git push
Username for 'https://github.com': digeo94
Password for 'https://digeo94@github.com':
remote: Invalid username or password.
fatal: Authentication failed for 'https://github.com/anfisbena/MIAD-DSA/'
(env-dvc) ubuntu@ip-172-31-16-127:~/dvc-proj$ git push
Username for 'https://github.com': digeo94
Password for 'https://digeo94@github.com':
Enumerating objects: 7, done.
Counting objects: 100% (7/7), done.
Delta compression using up to 2 threads
Compressing objects: 100% (3/3), done.
Writing objects: 100% (4/4), 878 bytes | 878.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/anfisbena/MIAD-DSA
   7d94362..effe39e  main -> main
```

Continuando con el manejo de versiones, realizamos la carga inicial de los datos y efectuamos el primer commit al repositorio Git. Este proceso se llevó a cabo en el archivo EDA.py, ubicado en el directorio de scripts creado para el proyecto. A medida que avancemos con el análisis exploratorio de datos, seguiremos utilizando este mismo archivo EDA.py para registrar los cambios en el procesamiento de datos, haciendo un commit al repositorio cada vez que se genere una actualización. Este enfoque garantiza que siempre contemos con una versión actualizada y segura del proyecto, y que el script de análisis esté respaldado de forma consistente en la plataforma de control de versiones.

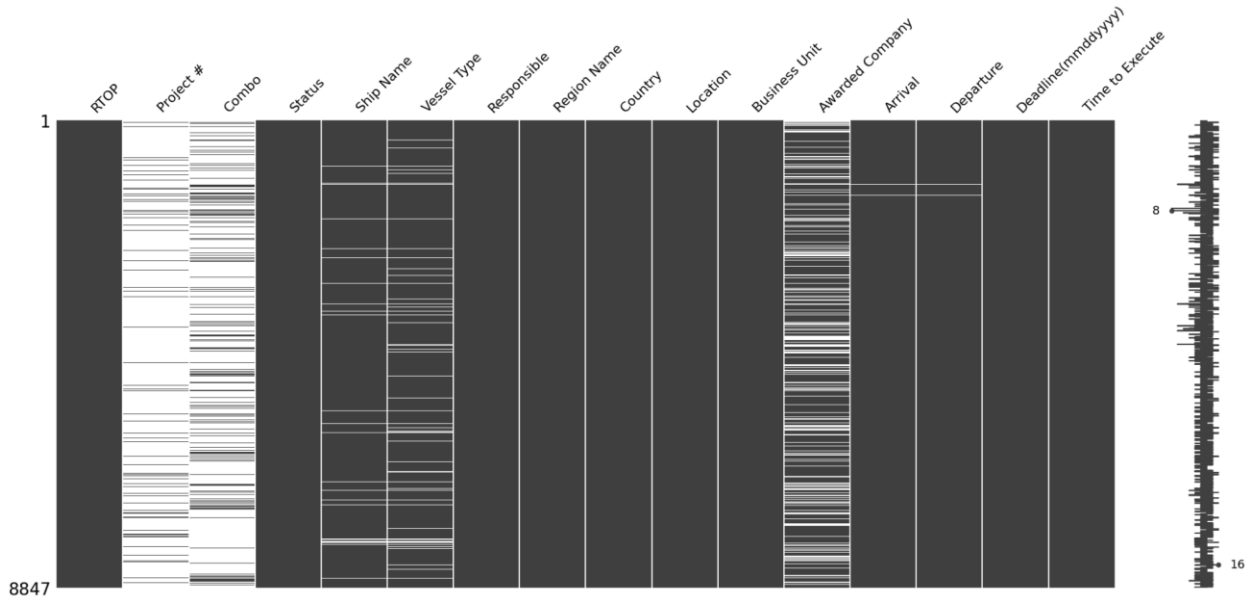
4. Exploración de los datos

Se cambian los nombres de las dimensiones para simplificar su entendimiento

- Ship ID - Ship Classification == Vessel type
- Ship - Name == Ship Name

Completitud

De la totalidad del dataset se identifican grandes cantidades de muestras con información incompleta, luego de indagar con el equipo de expertos de la compañía se llega a la conclusión que hay dimensiones innecesarias que afectan la completitud del Dataset



Limpieza de datos

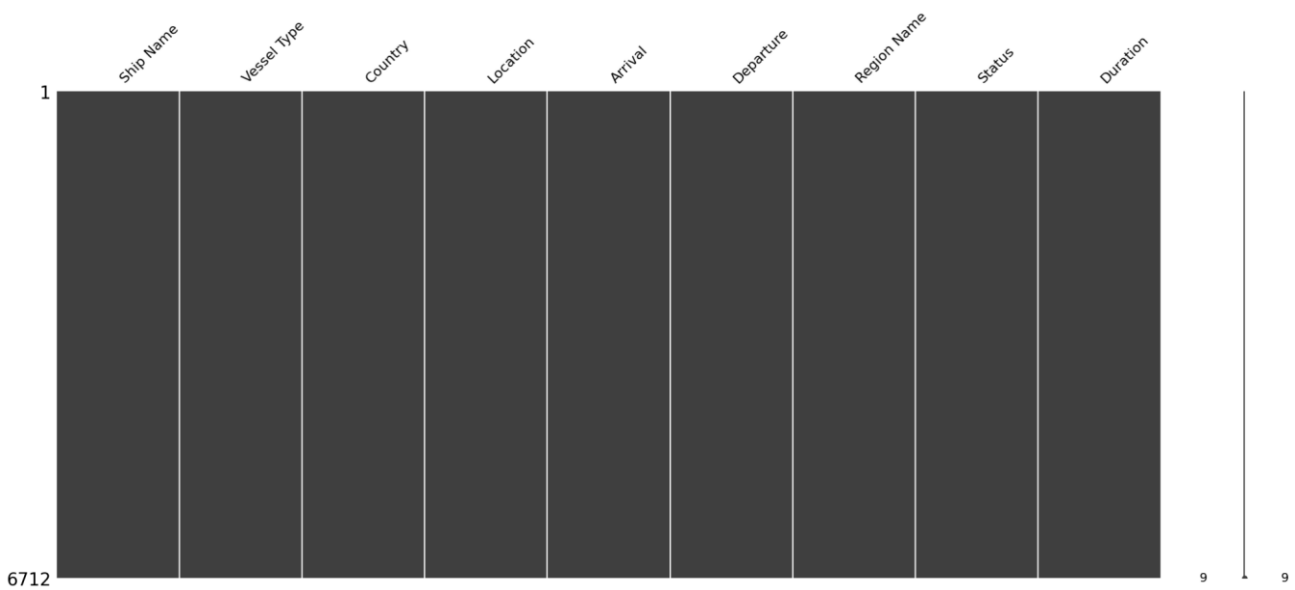
Ajustándose al alcance del proyecto, se reduce la dimensionalidad del dataset a sus valores primordiales para el caso, por ende se van a conservar:

- Ship Name
- Vessel type
- Region Name
- Country
- Location
- Arrival
- Departure
- Time to Execute

Información a considerar

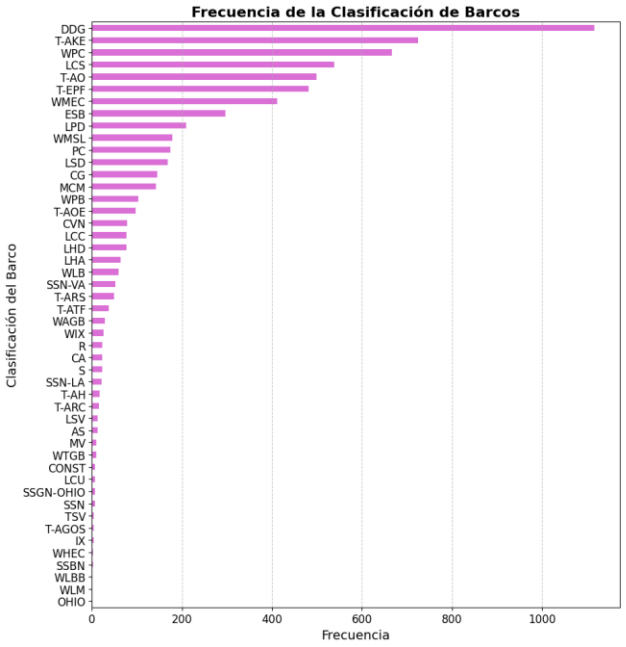
- Aquellas muestras que contengan valores nulos de Arrival, Departure, Ship Name y location inicialmente serán limpiados del modelo
- Los Champions del proyecto también informan que ofertas con días en ejecución menores a 0 y mayores a 60 serán desechadas debido a un posible error humano
- Toda las ofertas cuyo Status sea Cancelled, no serán tomadas en cuenta

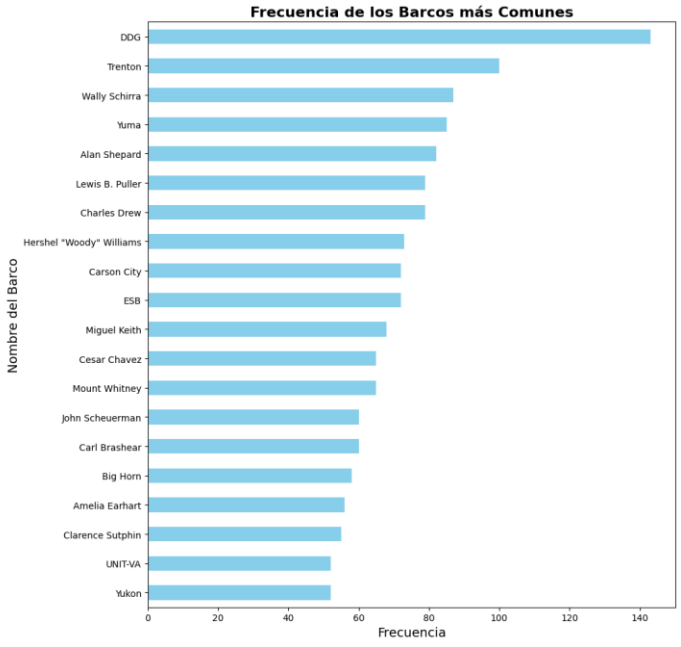
Debido a esto se limpiaron 2135 ofertas que representan 24% de los datos



Frecuencia de ofertas por barcos

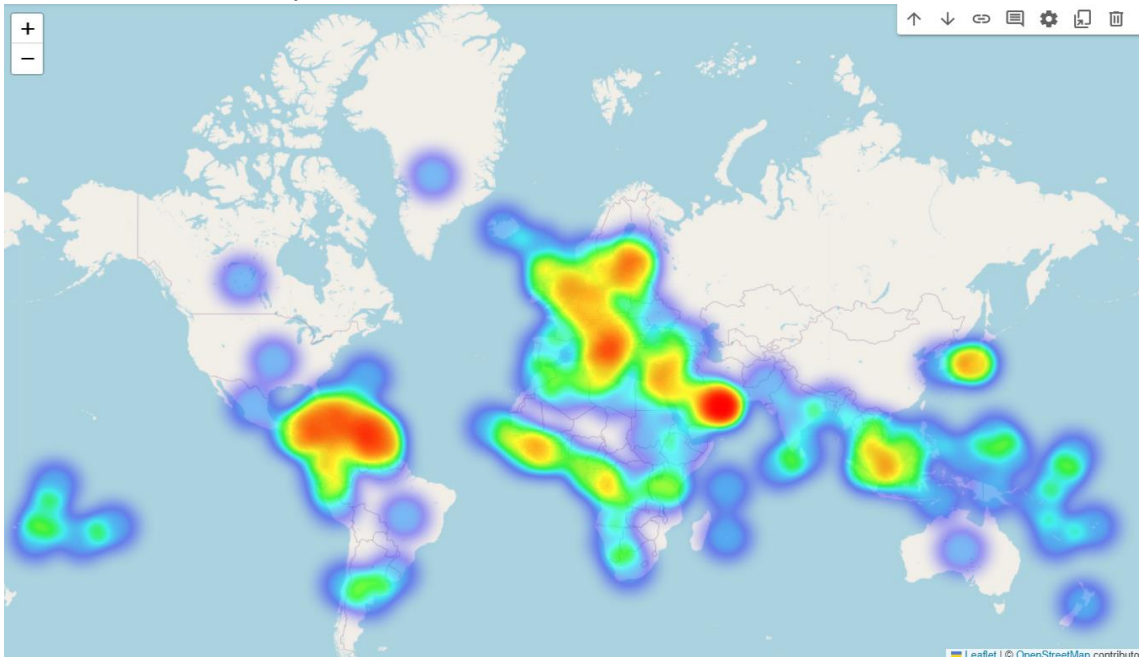
Este valor nos permite identificar que barcos tienen más información en el sistema, por ende, nos podría dar mejores resultados





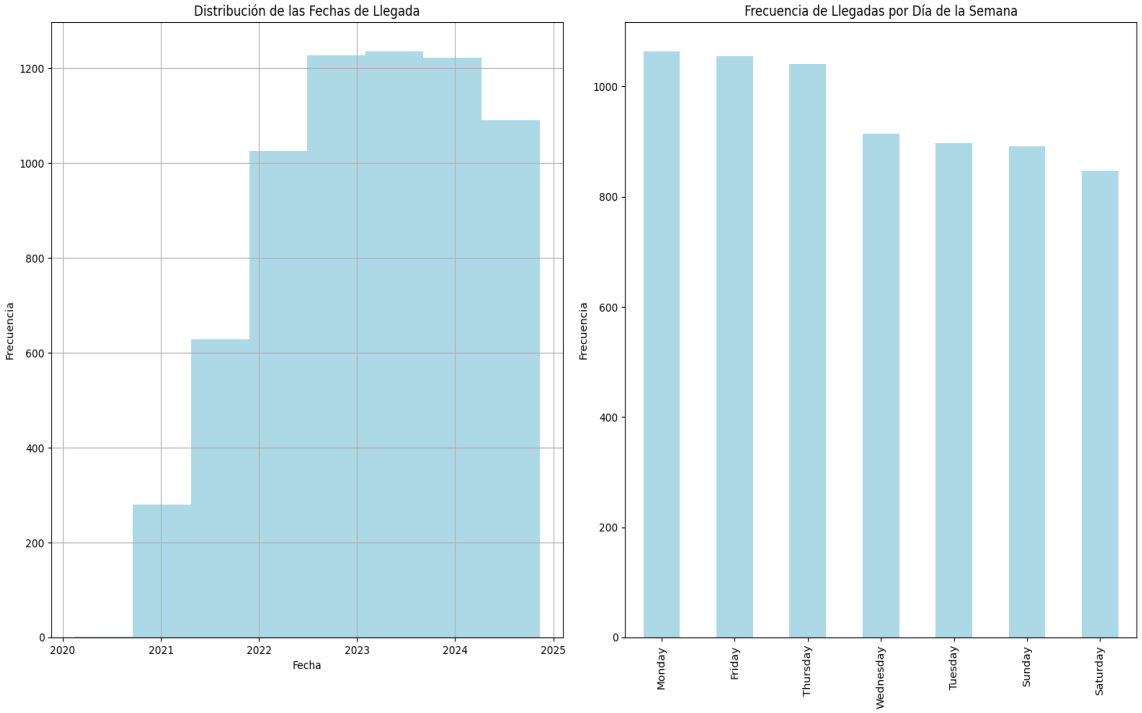
Mapa de calor de ofertas por país

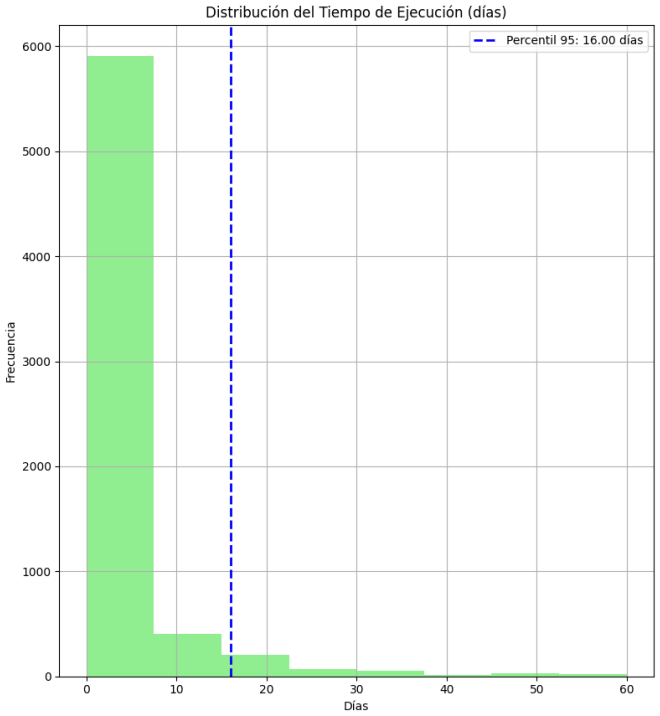
Este insumo nos permite identificar qué regiones tienen más concentración de ofertas y por ende de información para desarrollar los modelos

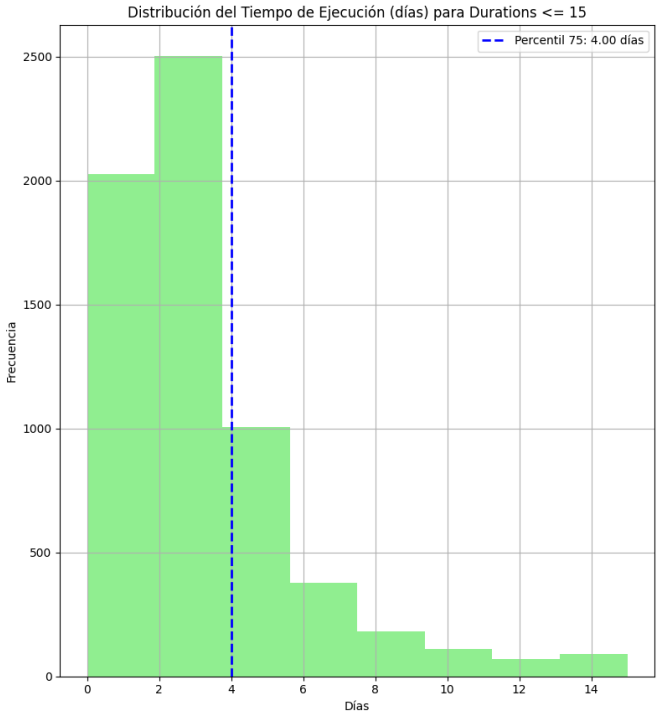


Distribución de ofertas de forma temporal

Está rubrica nos permite identificar, según el barco, las frecuencias de ofertas por día de la semana y la tendencia por barco de tiempo que se encuentra en puerto.

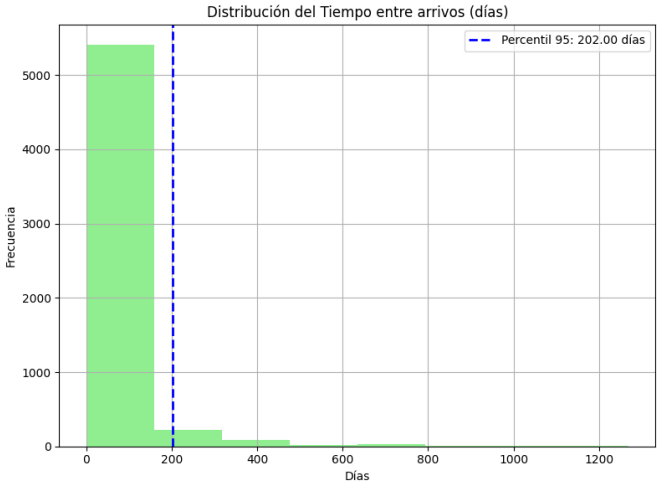


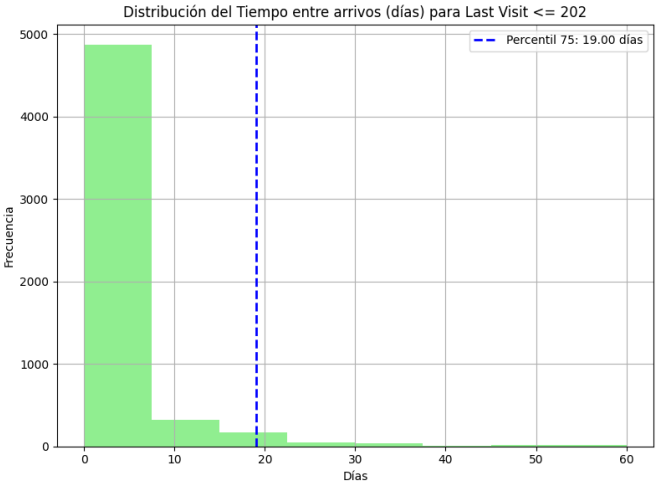




Distribución de tiempo interarribo

Está rubrica nos permite identificar cuanto tiempo tiende el barco a estar en altamar





5. Maqueta del prototipo

En esta maqueta se tuvo en cuenta dos secciones. La primera está relacionada con los resultados del modelo de predicción, en donde se debe seleccionar el barco para predecir su próximo puerto. Y la segunda sección presenta información histórica importante que apoya a Waypoint LLC en su operación.

waypoint

Optimización de llegada de barcos de Waypoint LLC

MODELO DE PREDICCIÓN

A continuación, puede seleccionar el barco en el siguiente menú desplegable para predecir el posible próximo puerto al que este va a llegar.

Trenton

El modelo predice que el próximo puerto es:

Prox. Puerto

Estadísticas del modelo

Precisión de Ubicación: -15.76%
Error Cuadrático Medio de Ubicación: 183.30
R^2 de Ubicación: -0.16

Localización del puerto

INFORMACIÓN HISTÓRICA DEL BARCO SELECCIONADO

Rango de tiempo | Región | Puerto

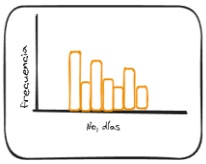
Información:

Tiempo promedio en puerto predicho
3 días

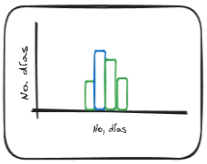
Puerto más frecuentado:
Ponce

Tiempo promedio entre arribos
8 días

Tiempo en puerto predicho

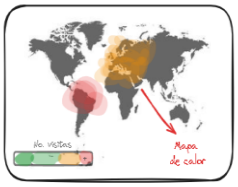


Tiempo en altamar

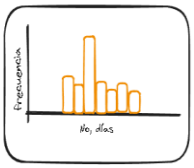


Puertos más visitados por el barco seleccionado

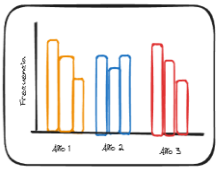
Puertos más visitados por barcos del mismo tipo



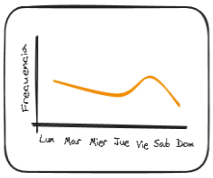
Distribución del tiempo entre arribos (días)



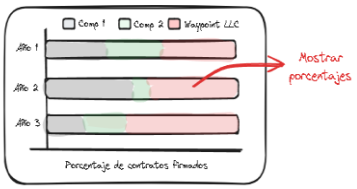
Distribución del tiempo entre arribos (días) del barco por año



Frecuencia de llegada por día de la semana



Contratos firmados del barco seleccionado con Waypoint LLC y otros competidores



6. Reporte de trabajo en equipo

Nombre y Rol	Responsabilidad
Karine Terán: Diseñadora de la Interfaz de Usuario	- Diseño de la interfaz de usuario y creación del Mock-up. - Previsualización del dashboard para asegurar una experiencia de usuario intuitiva y funcional.
Diego Tibaduisa: Ingeniero de Datos e Infraestructura	- Diseño de la infraestructura de datos. - Despliegue de los repositorios y gestión del versionamiento de los datos. - Escalamiento de las aplicaciones que conformarán la APP para asegurar su rendimiento y fiabilidad.
Andrés Arcila: Analista de Datos y Modelado Predictivo	- Creación de los repositorios de DVC y GitHub. - Análisis descriptivo y predictivo de los datos para entender patrones y tendencias.

Entregables

Se encuentran en el repositorio de GitHub