

Presentación

Integrantes:

Luis David Marte Vasquez - 2023-1165

Emill Peralta Encarnación - 2023-1151

Frankoris Rodriguez Ortiz - 2023-1346

Caso:

Caribbean Cinemas (Aprendizaje Automático)

Materia:

Inteligencia Artificial

Profesor:

Ramón Alvarez Santana



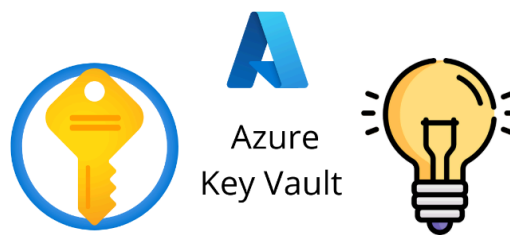
Fecha: 16/7/2024

Recursos en la nube que usarán y módulos de Azure ML que emplearán

Recursos en la nube se están utilizando en lo que a cloud computer se refiere estamos utilizando la máquina **standard_DS3_v2** con 4 núcleos que es la que nos permite entrenar nuestro modelo.

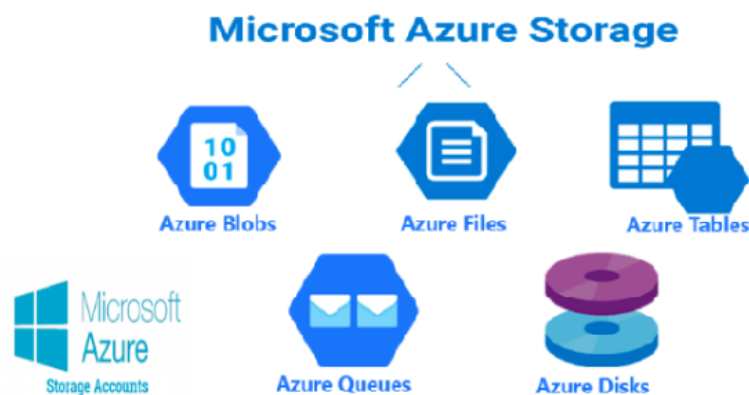
Estas imágenes son los recursos en la nube utilizados para nuestra implementación de un modelo de aprendizaje automático que generará recomendaciones a nuestros clientes como empresa de caribbean cinemas que somos.

Azure Key Vault: Para gestionar y asegurar secretos y claves criptográficas utilizamos estas herramientas ya que son sumamente importantes y beneficiosas en lo que a seguridad de nuestro proyecto se refiere..



Azure Storage Account: Para almacenar grandes volúmenes de datos, como los CSV de entrenamiento que utilizaremos y los resultados del modelo que arroje al ser entrenado son necesarios para nuestro uso posterior guardarlos, por lo que tener una cuenta de estas nos permite tener una serie de opciones para gestionar nuestro almacenamiento.

Azure Blob Storage: Una vez creada nuestra cuenta, nos optamos por el Azure Blob Storage ya que entendemos que de los servicios de almacenamiento es el que mejor se adapta a nuestras necesidades en estos momentos, este nos permite guardar nuestros datasets de grandes volumen para poder entrenar nuestro modelo de implementación de un recomendador, los que no hará estar en constante mejora.



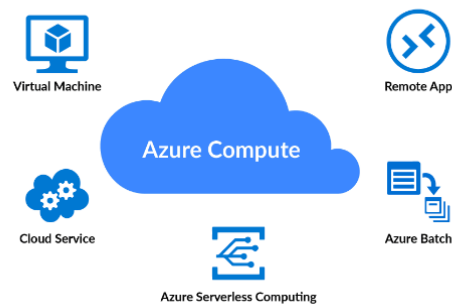
Dentro de la computación y virtualización perced para el entrenamiento de los modelos hemos utilizado:

Azure Compute Instances: Máquinas virtuales configuradas para el desarrollo y pruebas de código, entrenar modelos ademas de necesitar una serie de recursos en cuanto a hardware se refiere, este tambien necesita un ambiente propio para ser configurado por lo que utilizamos **sk-learn1.0:40** ya que satisface con creces las necesidades actuales que requerimos para comenzar el entrenamiento del model, este entorno cuenta con las siguientes dependencias:

```
name: sklearn-1.0
channels:
- anaconda
- conda-forge
dependencies:
- python=3.9
- pip=21.3.1
- pandas~=1.3.0
- scipy~=1.10.0
- numpy~=1.22.0
- pip:
  - wheel~=0.38.1
  - azureml-core==1.56.0
  - azureml-defaults==1.56.0.post1
  - azureml-mlflow==1.56.0
  - azureml-telemetry==1.56.0
  - scikit-learn~=1.0.0
  - joblib~=1.2.0
  # vulnerabilities
  - cryptography>=42.0.4
  - certifi >= 2023.07.22
  - requests>=2.32.3
  - idna>=3.7
  # azureml-automl-common-tools packages
  - py-spy==0.3.12
  - debugpy~=1.6.3
  - ipykernel~=6.0
  - tensorboard
  - psutil~=5.8.0
  - matplotlib~=3.5.0
  - tqdm~=4.66.3
  - py-cpuinfo==5.0.0
  - torch-tb-profiler~=0.4.0
```

Azure Compute Clusters: Clusters escalables para entrenar nuestro modelo de machine learning a gran escala, ya que entrenar este tipo de modelos requiere un nivel de consumo de recursos bastante elevado como se explicó anteriormente, por lo que incidimos en elegir esta máquina por su precio y adaptabilidad para el entrenamiento de nuestro modelo.

```
cluster-training
Standard_DS3_v2
4 vCPU (núcleos), 14 GB, 28 GB (almacenamiento), 0.23 $/h
Nodos: 0 inactivos, 0 ocupados, 1 sin aprovisionar
```



Para todo lo que ha sido el desarrollo, testeo y entrenamiento de modelos utilizamos:

Azure Machine Learning Workspace: Es el entorno principal donde gestionamos nuestros experimentos, modelos, recursos y es el que nos permitió generar una serie de soluciones de inteligencia artificial, facilitándonos la implementación y creación de nuestros modelos.



Azure designer: Este módulo nos permitio crear una solución por medio de un diseñador sencillo y sumamente eficaz, que nos permite crear una serie de pasos a ser ejecutados siguiendo un flujo de trabajo que puede ser repetido y cambiado las veces que sea necesario:

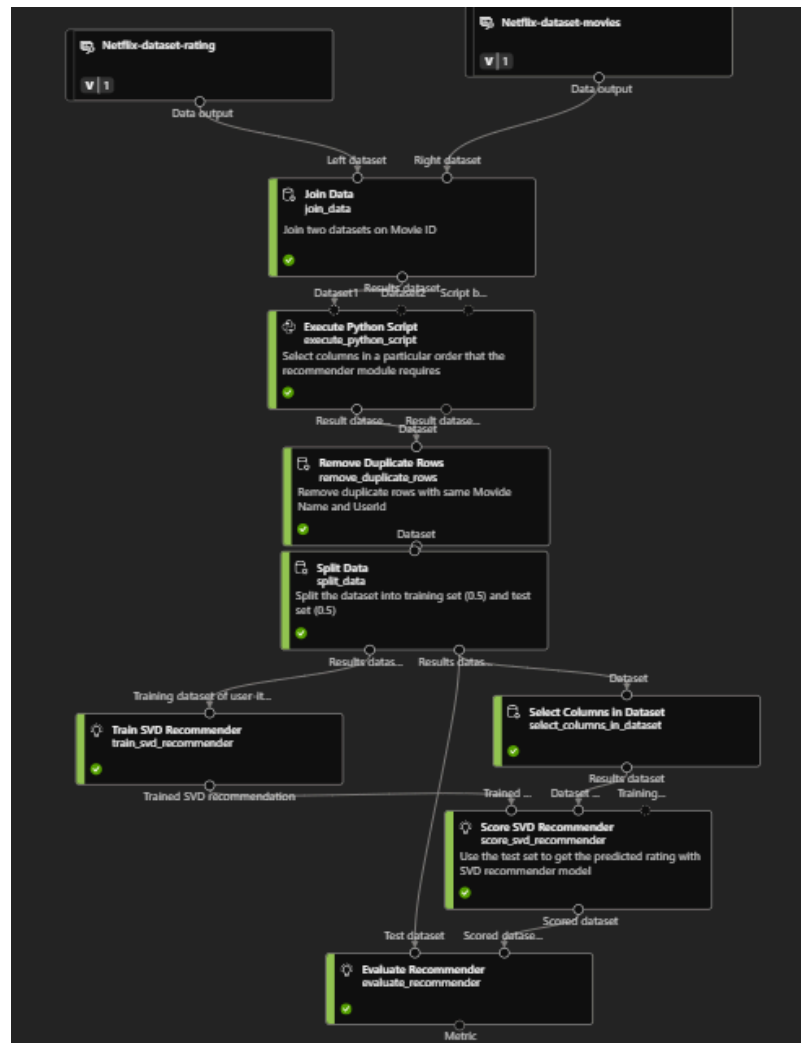
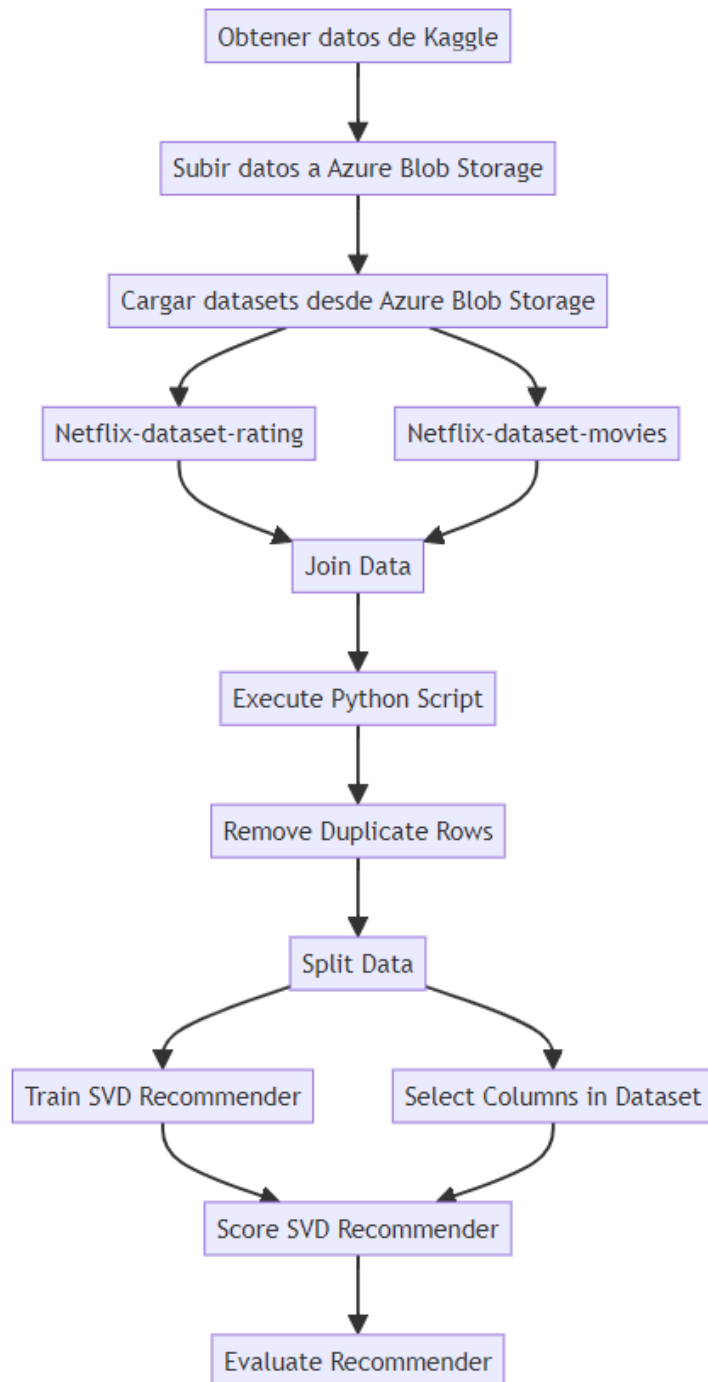


Diagrama de flujo de los datos y proceso



Flujo explicado al detalle:

Obtener y Cargar Datos:

- **Obtenemos los datos de Kaggle:**
 - Descargamos los datasets necesarios desde la plataforma Kaggle.
- **Subimos los datos a Azure Blob Storage:**
 - Subimos los datasets descargados a Azure Blob Storage para facilitar el acceso y la integración con Azure Machine Learning.

Canalización en Azure Machine Learning Designer:

- **Cargamos los datasets desde Azure Blob Storage:**
 - Cargamos los datasets `Netflix-dataset-rating` y `Netflix-dataset-movies` desde Azure Blob Storage en Azure Machine Learning Designer.
- **Unimos los datasets:**
 - Usamos el nodo `Join Data` para unir los datasets en base al ID de la película (`Movie ID`).
- **Ejecutamos un script de Python:**
 - Utilizamos el nodo `Execute Python Script` para ejecutar un script de preprocesamiento que realiza las transformaciones necesarias en los datos unidos.
- **Removemos duplicados:**
 - Empleamos el nodo `Remove Duplicate Rows` para eliminar cualquier fila duplicada que tenga el mismo `Movie ID` y otros campos relevantes.
- **Dividimos los datos:**
 - Usamos el nodo `Split Data` para dividir los datos en conjuntos de entrenamiento y prueba (por ejemplo, 80% para entrenamiento y 20% para prueba).
- **Entrenamos el modelo de recomendación SVD:**
 - Usamos el nodo `Train SVD Recommender` para entrenar el modelo de recomendación SVD utilizando el conjunto de datos de entrenamiento.
- **Seleccionamos columnas en el conjunto de datos:**
 - Utilizamos el nodo `Select Columns in Dataset` para seleccionar las columnas relevantes necesarias para el modelo de recomendación.
- **Calificamos con el modelo SVD:**
 - Usamos el nodo `Score SVD Recommender` para obtener las predicciones de calificaciones utilizando el conjunto de datos de prueba.
- **Evaluamos el modelo de recomendación:**
 - Utilizamos el nodo `Evaluate Recommender` para evaluar el rendimiento del modelo de recomendación utilizando métricas como la precisión, la cobertura y el error cuadrático medio (RMSE).

Ventajas del Algoritmo SVD y la Supervisión No Automática en Sistemas de Recomendación de Películas

El uso del algoritmo SVD (Descomposición en Valores Singulares) en sistemas de recomendación de películas tiene varias ventajas, especialmente en el contexto de sistemas de recomendación colaborativos. Aquí te explico por qué SVD puede ser una buena elección y la razón detrás del uso de supervisión no automática:

Ventajas del Algoritmo SVD en Recomendadores de Películas

1. **Reducción de Dimensionalidad:** SVD es una técnica de reducción de dimensionalidad que descompone la matriz de usuarios y películas en tres matrices más pequeñas. Esto ayuda a reducir el ruido y encontrar patrones latentes que son útiles para hacer recomendaciones precisas.
2. **Manejo de Datos Escasos:** En sistemas de recomendación, las matrices de calificaciones son generalmente muy dispersas, es decir, la mayoría de las entradas son nulas porque no todos los usuarios califican todas las películas. SVD maneja bien estas matrices dispersas, ayudando a predecir las calificaciones faltantes de manera efectiva.
3. **Descubrimiento de Factores Latentes:** SVD ayuda a descubrir factores latentes que influyen en las preferencias de los usuarios. Por ejemplo, puede identificar que a un usuario le gustan las películas de un género específico, incluso si no ha calificado muchas películas de ese género.
4. **Personalización:** Debido a su capacidad para identificar patrones latentes, SVD puede proporcionar recomendaciones altamente personalizadas, mejorando la experiencia del usuario.
5. **Eficiencia Computacional:** Aunque el cálculo inicial de SVD puede ser intensivo, una vez que las matrices están descompuestas, la predicción de calificaciones es muy eficiente.

Supervisión No Automática

La supervisión no automática, o supervisión manual, en el contexto de sistemas de recomendación, se refiere a la intervención humana en el proceso de entrenamiento del modelo. Aquí hay algunas razones por las cuales se puede preferir:

1. **Interpretabilidad:** Los sistemas supervisados manualmente tienden a ser más interpretables porque los expertos pueden ajustar y verificar el modelo a medida que se entrena. Esto es crucial en sistemas de recomendación para asegurarse de que las recomendaciones sean razonables y explicables.
2. **Mejora de la Calidad:** La intervención humana puede ayudar a mejorar la calidad del modelo, especialmente en las primeras etapas del desarrollo, donde se pueden

identificar y corregir errores o sesgos que un algoritmo automático podría pasar por alto.

3. **Control y Ajuste Fino:** Los expertos pueden ajustar manualmente los parámetros y el enfoque del modelo según el contexto específico de la aplicación, lo que puede resultar en un rendimiento mejorado y más acorde con los objetivos del negocio.
4. **Validación y Verificación:** La supervisión manual permite una validación y verificación más exhaustiva del modelo, asegurando que las recomendaciones generadas sean relevantes y útiles para los usuarios.

En resumen, el algoritmo SVD es una excelente opción para los sistemas de recomendación de películas debido a su capacidad para manejar matrices dispersas y descubrir patrones latentes. La supervisión no automática proporciona un control adicional que puede mejorar la calidad y la interpretabilidad del modelo.

Formato de salida propuesto de la información

Resultados del modelo:

1. **¿Qué tipo de resultados o predicciones generará el modelo?**
 - El modelo generará predicciones de calificación para películas específicas para cada usuario en el conjunto de prueba.
2. **Estimará un porcentaje según los datos específicos o verificará para que tire items?**
 - El modelo proporcionará predicciones de calificación numérica para películas, no porcentajes directos.
3. **¿En qué formato se entregarán estos resultados?**
 - Los resultados se entregarán en formato JSON.

Reportes y métricas:

1. **¿Qué reportes y métricas necesitas generar?**
 - Se necesitan métricas como precisión, cobertura y RMSE (Error Cuadrático Medio) para evaluar el rendimiento de nuestro modelo de recomendación entrenado, por lo que la calidad de datos influirá en estas métricas que nos dicen la calidad de nuestro modelo que se encuentra en base a un 85 de 100.
2. **¿En qué formato se entregaremos estos reportes?**
 - Los reportes se entregarán en un formato estructurado, preferiblemente JSON para facilitar su procesamiento y análisis posterior ya que es fácil de manejar.
3. **Predicciones de Calificación:**
 - Para cada usuario en el conjunto de prueba, el modelo genera predicciones de calificación para diferentes películas.
4. **Formato de Salida:**
 - Cada entrada en el resultado contiene la siguiente información:

- **Usuario:** ID del usuario para quien se hace la recomendación.
- **Película Recomendada:** Nombre o ID de la película recomendada.
- **Predicción de Calificación:** Valor estimado de la calificación que el usuario podría dar a la película.

Ejemplo:

Este sería un ejemplo de inferencia de datos a nuestro modelo y de la respuesta que este generará a nuestros usuarios, en este ej seria el caso de que el modelo estuviese desplegado:

```
{ "data":
  {
    "User_ID": 1888204,
    "Name": "Something's Gotta Give",
    "Rating": 4
  }
}
```

Y esta seria las recomendaciones que les daria, evidentemente seria en un formato json que sería entregada para asi manejar esta información dentro de nuestra plataforma de caribbean.

Recomendaciones:

- Película: "The Shawshank Redemption"

Predicción de Calificación: 4.8

- Película: "Inception"

Predicción de Calificación: 4.5

- Película: "Pulp Fiction"

Predicción de Calificación: 4.7

5. Interpretación:

- Las predicciones de calificación más altas indican las películas que el modelo sugiere que el usuario probablemente disfrutará más, basado en sus patrones de calificación históricos y similitudes con otros usuarios.

6. Utilidad:

- Estos resultados pueden ser utilizados para personalizar recomendaciones de contenido para usuarios en plataformas de streaming, mejorando así la experiencia del usuario y aumentando la satisfacción.

Costos de la propuesta

Costo de los Recursos de Computación: Estimamos utilizar los clústeres de computación durante aproximadamente 40 minutos para el entrenamiento del modelo. Utilizaremos una máquina standard DS3_v2 con un costo de \$0.23 USD por hora.

Cálculo del Costo de la Computación: Convertimos el tiempo a horas: 40 minutos = $40/60 = 0.6667$ horas.

Costo total de computación = Tiempo en horas * Costo por hora = $0.6667 * \$0.23 \approx \0.1533

Costo de Azure Blob Storage: Azure Blob Storage tiene un costo de \$0.15 USD por GB. Considerando que nuestro modelo no pesa menos de 1GB y los conjuntos de datos suman aproximadamente 300MB, contando con miles de filas:

Costo total de almacenamiento = $1\text{GB} * \$0.15/\text{GB} = \0.15

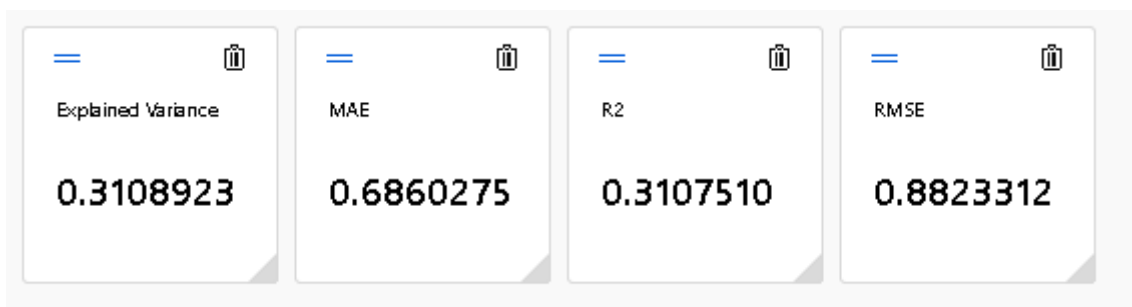
Total de Costos Aproximados:

- **Computación:** \$0.1533
- **Almacenamiento:** \$0.15

Beneficios:

- Utilización eficiente de recursos de Azure para un costo bajo.
- Escalabilidad y flexibilidad en el manejo de grandes volúmenes de datos.
- Integración sin costos adicionales con otros servicios externos.

Métricas de efectividad del sistema propuesto



Varianza Explicada (Explained Variance): 0.3108923

El modelo explica alrededor del 31% de la variabilidad en los datos. Hay margen significativo para mejorar.

Error Absoluto Medio (MAE): 0.6860275

En promedio, las predicciones del modelo se desvían en 0.686 unidades de los valores reales. Un valor más bajo sería mejor.

Coeficiente de Determinación (R2): 0.3107510

El modelo explica aproximadamente el 31% de la varianza en los datos, indicando una necesidad de mejora.

Raíz del Error Cuadrático Medio (RMSE): 0.8823312

Las predicciones del modelo se desvían en promedio en 0.882 unidades de los valores reales. Un valor más bajo es deseable.

Dificultades al realizar el proyecto

Problema 1: Datos Insuficientes para Entrenar el Modelo

Los datos no son suficientes ni adecuados para entrenar el modelo de manera efectiva.

Causas

- **Cantidad insuficiente:** No hay suficientes datos.
- **Calidad Baja:** Datos con errores o valores faltantes.

Solución

- **Recolección de Más Datos:** Obtener datos adicionales.
- **Mejorar la Calidad de los Datos:** Corregir errores y valores faltantes.
- **Optimización del Modelo:** Optimizar el modelo.

Problema 2: Dificultades en la Integración de Servicios en Azure

Otro problema común al trabajar con este proyecto de Azure fue la dificultad para integrar algunas cosas de Azure de manera efectiva. Esto afectó mucho a la funcionalidad y el rendimiento del sistema.

Causas

- **Complejidad de la Configuración:** Los servicios requieren configuraciones, lo que fue complejo y nos daba errores.
- **Falta de Conocimiento Técnico:** Nosotros al no tener conocimiento necesario para configurar y gestionar los servicios se nos complicó.

Soluciones

- **Documentación y Recursos de Azure:** Utilizamos la documentación de Azure y videos para entender mejor cómo configurar e integrar los servicios.
- **Uso de Plantillas y Automatización:** Utilizar plantillas de Azure Resource Manager (ARM) y herramientas de automatización.

Posibles mejoras de la propuesta

Crear nuestro data para mejorar el Entrenamiento

Crear un conjunto de datos personalizado puede mejorar la precisión y el rendimiento del modelo.

Beneficios

- **Datos Relevantes:** Más adecuados para el problema específico.
- **Mejor Calidad:** Control y eliminación de errores.
- **Más datos:** Datos suficientes para mejorar el entrenamiento.

Estrategias

- **Generación de Datos Sintéticos:** Aumentar la cantidad de datos.

- **Colaboración:** Trabajar con otros departamentos para obtener datos útiles.
- **Preprocesamiento y Limpieza:** Limpiar los datos.

Riesgos y limitaciones

Riesgos:

- **Calidad de los datos:** Asegurar la precisión y completitud de los datos es fundamental para evitar errores y decisiones basadas en el modelo.
- **Complejidad del Problema:** Problemas complejos pueden requerir más recursos o técnicas avanzadas, lo cual podría aumentar los costos y el tiempo de desarrollo del proyecto.

Limitaciones:

- **Características:** Elegir adecuadamente las variables es importante, ya que esto influye en la precisión y lo bien que puede salir el modelo.
- **Costos y Recursos:** Mantener modelos de aprendizaje automático son costosos en tema de recursos y por eso estamos limitados.

Procedimiento realizado para resolver el problema

Descripción clara del problema o desafío:

- **Problema:** Caribbean Cinemas desea seguir la tendencia de empresas líderes en la distribución de multimedia, como Netflix y YouTube, utilizando datos históricos de sus clientes para personalizar su oferta y mejorar sus estrategias de marketing.
- **Desafío:** Crear un sistema de clustering para agrupar a los clientes en categorías específicas basadas en sus preferencias de contenido. Esto permitirá a Caribbean Cinemas hacer recomendaciones más precisas y personalizadas, y desarrollar estrategias de marketing más efectivas dirigidas a distintos clústeres de clientes.

Contexto del Problema

Identificación del Problema:

Caribbean Cinemas, al igual que muchas otras empresas, ha reconocido el valor de los datos de sus clientes y ha estado almacenando información histórica sobre ellos. Al observar el éxito de empresas líderes en distribución de multimedia, como Netflix y YouTube, que utilizan datos para personalizar recomendaciones y estrategias de marketing, Caribbean Cinemas ha identificado una oportunidad para mejorar su propia oferta. La compañía se dio cuenta de que podía utilizar los datos de sus clientes para homologar a sus clientes con etiquetas específicas y hacer recomendaciones más personalizadas, lo cual podría mejorar significativamente la experiencia del cliente y las estrategias de marketing.

Causas Raíz del Problema:

Caribbean Cinemas enfrenta varios desafíos debido a la falta de segmentación de clientes, subutilización de datos históricos, ausencia de tecnologías avanzadas de machine learning, falta de personalización en la comunicación y competencia en el mercado. Esto resulta en una experiencia genérica para los clientes, limitaciones en la personalización y optimización de estrategias de marketing, y una menor efectividad en la retención y fidelidad de los clientes frente a competidores que utilizan estrategias avanzadas de personalización.

Identificar y abordar estas causas raíz es crucial para que Caribbean Cinemas pueda desarrollar un sistema de clustering efectivo y mejorar sus estrategias de marketing, aprovechando al máximo los datos de sus clientes y ofreciendo una experiencia más personalizada.

Objetivos Específicos

Segmentación de Clientes:

- **Descripción:** Desarrollar un sistema de clustering para segmentar a los clientes en grupos basados en sus preferencias y comportamientos.
- **Resultado Esperado:** Identificación de clústeres específicos de clientes, como aquellos que prefieren películas de acción y romance, contenido político, etc.

Personalización de Recomendaciones:

- **Descripción:** Implementar un sistema que ofrezca recomendaciones de películas y promociones personalizadas según los clústeres de clientes identificados.
- **Resultado Esperado:** Aumento en la satisfacción y retención de los clientes debido a recomendaciones más relevantes y atractivas.

Optimización de Estrategias de Marketing:

- **Descripción:** Desarrollar estrategias de marketing dirigidas a los distintos clústeres de clientes para mejorar la efectividad de las campañas.
- **Resultado Esperado:** Mayor retorno de inversión (ROI) en campañas de marketing y promociones, así como una mejor utilización del presupuesto de marketing.

Incremento en Ventas y Asistencia:

- **Descripción:** Aumentar la frecuencia de visitas al cine y el gasto de los clientes en entradas y concesiones.
- **Resultado Esperado:** Incremento en las ventas de boletos y productos relacionados, así como un aumento en la asistencia general al cine.

Mejora en la Experiencia del Cliente:

- **Descripción:** Ofrecer una experiencia de cine más personalizada y satisfactoria para los clientes.
- **Resultado Esperado:** Aumento en la fidelidad y satisfacción del cliente, lo que lleva a una mayor lealtad y recomendación del cine.

Análisis Continuo y Ajuste de Estrategias:

- **Descripción:** Monitorear y analizar continuamente el comportamiento de los clientes y la efectividad de las recomendaciones y estrategias de marketing.

- **Resultado Esperado:** Capacidad para ajustar y mejorar continuamente las estrategias basadas en datos y retroalimentación, asegurando la relevancia y efectividad a largo plazo.

Desarrollo de una Ventaja Competitiva:

- **Descripción:** Posicionar a Caribbean Cinemas como un líder innovador en el uso de datos para la personalización de la experiencia del cliente.
- **Resultado Esperado:** Mejora en la posición competitiva de la empresa en el mercado, atrayendo a más clientes y manteniendo a los actuales frente a la competencia.

Los criterios de éxito para Caribbean Cinemas incluyen:

- **Precisión en la Segmentación de Clientes:** Segmentar a los clientes en clústeres significativos, medido por métricas de cohesión interna y separación externa.
- **Relevancia de las Recomendaciones Personalizadas:** Asegurar que las recomendaciones sean relevantes, evaluado por la tasa de clics, tasa de conversión y encuestas de satisfacción.
- **Mejora en la Eficiencia de las Estrategias de Marketing:** Aumentar la efectividad de las campañas, medido por el ROI, tasa de apertura de correos y ventas incrementales.
- **Incremento en Ventas y Asistencia:** Aumentar ventas y asistencia al cine, medido por el incremento porcentual en ventas y número de visitas por cliente.
- **Satisfacción y Fidelidad del Cliente:** Mejorar la satisfacción y lealtad, medido por encuestas, Net Promoter Score y tasa de retención.
- **Eficiencia Operativa y Continuidad del Análisis:** Mantener un sistema de análisis continuo, medido por la calidad y frecuencia de los informes y ajustes realizados.
- **Desarrollo de una Ventaja Competitiva:** Posicionar a Caribbean Cinemas como líder en personalización, medido por comparaciones con competidores, análisis de mercado y reconocimiento en la industria.

Rol de los miembros del equipo y trabajo realizado.

Búsqueda de Datos (Data Gathering) - Emill Peralta Enc

Responsabilidades:

- Identificación de fuentes de datos: Localizar todas las posibles fuentes de datos que puedan ser útiles para el proyecto. Esto puede incluir bases de datos internas, archivos de registro, API externas, datos públicos, etc.

- Extracción de datos: Descargar, recopilar y consolidar los datos desde las diversas fuentes identificadas.
- Limpieza de datos: Asegurarse de que los datos sean consistentes, eliminando duplicados, gestionando valores faltantes y corrigiendo errores.
- Transformación de datos: Formatear y estructurar los datos de manera que sean adecuados para el análisis y el entrenamiento del modelo.

Selección del Algoritmo (Algorithm Selection) - Frankoris Rodriguez Ortiz

Responsabilidades:

- Análisis del problema: Comprender en profundidad el problema que se intenta resolver con el modelo de recomendación. Esto puede incluir análisis exploratorio de datos y discusión con partes interesadas.
- Revisión de algoritmos: Investigar los distintos algoritmos de aprendizaje automático que podrían ser aplicables, como sistemas de filtrado colaborativo, sistemas basados en contenido, o una combinación híbrida.
- Evaluación de algoritmos: Comparar los algoritmos seleccionados en términos de precisión, rendimiento y otros parámetros relevantes utilizando conjuntos de datos de prueba.
- Justificación de la selección: Documentar y justificar la elección del algoritmo basado en el análisis y las pruebas realizadas.

Entrenamiento del Modelo (Model Training) - Luis David Marte Vasquez

Responsabilidades:

- Preparación del entorno: Configurar el entorno de entrenamiento en Azure, asegurando que se cuente con los recursos computacionales necesarios.
- Entrenamiento del modelo: Ejecutar el proceso de entrenamiento del modelo seleccionado utilizando los datos preprocesados.
- Ajuste de hiperparámetros: Optimizar los hiperparámetros del modelo para mejorar su rendimiento.
- Validación del modelo: Evaluar el modelo entrenado utilizando conjuntos de datos de validación y ajustar según sea necesario.
- Documentación: Registrar el proceso de entrenamiento, los ajustes realizados y los resultados obtenidos.

Infraestructura del sistema que proponen como solución

Infraestructura Propuesta:

Ingesta y Almacenamiento de Datos:

- **Azure Data Factory:** Para orquestar y automatizar la ingesta de datos desde múltiples fuentes, como bases de datos transaccionales, archivos CSV, y otros repositorios de datos.
- **Azure Blob Storage:** Almacenamiento escalable para datos históricos y nuevos datos de clientes en formato no estructurado y semiestructurado.

Procesamiento y Análisis de Datos:

- **Azure Databricks:** Plataforma de análisis unificada basada en Apache Spark para procesamiento masivo de datos y creación de modelos de clustering.
- **Azure Synapse Analytics:** Para análisis de grandes volúmenes de datos y ejecución de consultas SQL sobre los datos almacenados.

Modelado y Aprendizaje Automático:

- **Azure Machine Learning:** Creación, entrenamiento y despliegue de modelos de clustering (por ejemplo, K-means clustering) para segmentar a los clientes en grupos homogéneos.

Visualización y Reportes:

- **Power BI:** Herramienta de visualización para crear informes interactivos y paneles de control que muestran los resultados del clustering y las recomendaciones personalizadas.

Integración y Automatización:

- **Azure Logic Apps:** Automatización de flujos de trabajo para integrar el sistema de recomendaciones con otras aplicaciones y servicios de Caribbean Cinemas.
- **Azure Functions:** Funciones serverless para ejecutar código en respuesta a eventos específicos, como la actualización de datos de clientes o la generación de recomendaciones.

Seguridad y Cumplimiento:

- **Azure Active Directory:** Gestión de identidades y control de acceso para asegurar que solo personal autorizado tenga acceso a los datos y modelos.
- **Azure Key Vault:** Almacenamiento seguro de claves, secretos y certificados utilizados en el sistema.

Monitoreo y Mantenimiento:

- **Azure Monitor:** Supervisión continua del rendimiento del sistema y los modelos de aprendizaje automático.
- **Azure Log Analytics:** Recolección y análisis de datos de registro para identificar y resolver problemas rápidamente.