TALLER N°2 MACHINE LEARNING – Predicciones con Algoritmos ML

A. Objetivo General:

- Desarrollar un modelo de Machine Learning que permita a la empresa de producción cinematográfica "CineMagix" prever el éxito financiero de futuras películas, aplicando los conocimientos adquiridos durante el curso en las áreas de análisis exploratorio de datos (EDA), preprocesamiento de datos, modelado de ML y evaluación de modelos, y siguiendo la metodología CRISP-DM.

B. Objetivos Específicos:

1. **Análisis Exploratorio de Datos (EDA)**:
   - Identificar tipos de variables y, en consecuencia, sus tendencias, patrones, posibles anomalías en los datos de las películas.
   - Analizar las relaciones/correlaciones/asociaciones entre las distintas características de las películas y su éxito financiero.
2. **Preprocesamiento de Datos**:
   - **Manejo de Valores Nulos**:
     - Implementar y justificar técnicas de imputación y estrategias para valores nulos y atípicos, considerando las implicancias en los modelos y la calidad de los datos.
   - **Ingeniería y Codificación de Características**:
     - Desarrollar y justificar estrategias de ingeniería de características, incluyendo la creación de nuevas variables y la elección de técnicas de codificación para variables categóricas, evaluando su impacto y aplicabilidad en el modelado predictivo.
   - **Evaluación de Estrategias de Preprocesamiento**:
     - Evaluar la efectividad de las estrategias de preprocesamiento y su influencia en la calidad y utilidad de los datos para el modelado, comunicando sus impactos y justificando las decisiones tomadas.
   - **Adherencia a Mejores Prácticas**:
     - Asegurar que todas las decisiones y estrategias aplicadas estén alineadas con las mejores prácticas de ML, siendo fundamentadas y detalladas en su presentación y justificación.
3. **Modelado de Machine Learning**:
   - Implementar y comparar al menos tres modelos de ML (Regresión Lineal, Árboles de Decisión, Regresión Logística y SVM) para predecir el éxito financiero de las películas.
   - Asegurar que los modelos se validen y se optimicen adecuadamente utilizando técnicas como la validación cruzada y el ajuste de hiperparámetros.
4. **Evaluación y Interpretación del Modelo**:
   - Evaluar los modelos desarrollados utilizando métricas adecuadas para proporcionar una interpretación clara de su rendimiento.
   - Interpretar los resultados del modelo para obtener insights que puedan ser útiles para "CineMagix" en la toma de decisiones.
5. **Comunicación de Resultados**:
   - Desarrollar una comunicación clara y comprensible de los hallazgos, insights y recomendaciones derivadas del modelo de ML.
   - Presentar la solución de manera estructurada, siguiendo la metodología CRISP-DM, y asegurando que las recomendaciones estén alineadas con los objetivos de negocio de "CineMagix".

6. **Nota**: Es importante mencionar que los Consultores deberán adherirse estrictamente a los algoritmos de Machine Learning especificados (Regresión Lineal, Árboles de Decisión, Regresión Logística y SVM) y justificar claramente sus elecciones y estrategias en cada etapa del proyecto.

C. CONTEXTO DE NEGOCIO: CINEMAGIX - DESCIFRANDO EL ENIGMA DEL ÉXITO CINEMATOGRÁFICO

En el vibrante y despiadado mundo del cine, la productora CineMagix ha sido una constante, tejiendo historias que han resonado tanto en taquillas como en corazones durante más de dos décadas. Sin embargo, la industria cinematográfica ha evolucionado, y el encanto del cine ahora está fuertemente influenciado por patrones que parecen esquivos y enigmáticos. La competencia es feroz, las apuestas son altas, y la magia del cine ahora requiere tanto de arte como de ciencia.

El Dilema:

El CEO de CineMagix, Alex Sterling, ha observado un patrón peculiar: algunas películas que eran consideradas seguras, con actores estelares y directores renombrados, han fracasado, mientras que algunas bajo el radar han sido éxitos sorprendentes. ¿Hay una fórmula oculta para el éxito que están pasando por alto? Alex se pregunta si los datos que han acumulado meticulosamente a lo largo de los años pueden revelar los secretos de la taquilla.

La Misión:

Su equipo, como consultores expertos son convocados por Alex, quien les presenta un desafío: Descifrar el enigma del éxito cinematográfico utilizando el arsenal de datos históricos de películas que CineMagix ha acumulado. Alex, siendo un ferviente creyente en el poder de los datos, quiere que los asesores exploren profundamente en los números, los géneros, los directores, las estrellas y todas las variables disponibles, y extraigan patrones, secretos y estrategias que puedan iluminar el camino de CineMagix hacia futuros éxitos de taquilla.

La Intriga:

Sin embargo, hay un giro. Alex revela que dos películas recientes, que eran consideradas “apuestas seguras”, resultaron ser los mayores fracasos financieros de CineMagix, causando tensiones en el estudio y poniendo en juego futuros proyectos. La presión es alta y los asesores deben navegar a través de los datos, no solo para predecir el éxito, sino para salvar la esencia creativa de CineMagix de las garras de decisiones meramente financieras y seguras que amenazan con apagar la chispa de la creatividad y la innovación en el estudio.

El Desafío Específico:

Los Consultores deben desentrañar y responder (esto es parte de la parte conclusiones-recomendaciones):

- ¿Cómo los géneros han influido en el éxito a lo largo de los años y cómo se puede estratificar eso para futuras producciones?
- ¿Qué impacto tienen las estrellas y directores en el éxito y cómo se puede cuantificar y categorizar ese impacto?
- ¿Cómo los metadatos de una película (duración, clasificación, etc.) se correlacionan con su rendimiento financiero y crítico?

El Viaje:

A medida que los Consultores se embarcan en este viaje, deben explorar, cuestionar, procesar y modelar los datos, descubriendo historias ocultas y tendencias que pueden estar invisibles para un ojo no entrenado. La misión no solo es predecir el futuro, sino también entender el pasado, y asegurarse de que las futuras producciones de CineMagix no solo sean éxitos financieros, sino también cinematográficos, preservando la creatividad y la originalidad en un mundo impulsado por datos y algoritmos.

D. Tareas Específicas (QUE SE REQUIERE EN ESTE TALLER)

La empresa CinaMagix ha reunido un vasto tesoro de datos de diversas películas a lo largo de los años y busca la ayuda de mentes brillantes: Ustedes los Consultores, quienes con su conocimiento en Análisis de Datos y Machine Learning, deberán descubrir los secretos ocultos dentro de los datos:

d.1) Análisis Eploratorio - Exploración del Tesoro de Datos:

Realice un análisis exploratorio y analice las variables en sus dimensiones que le permitan tomar las decisiones para tener un conjunto de datos sólido para predecir como se caracteriza el éxito en Cine y ayudar a CinaMagix en su misterio. **Interprete los resultados.**

d.2) Preprocesamiento - Forjando la Espada del Preprocesamiento:

"*Moldeen los datos, forjando la espada que cortará a través del ruido y las inconsistencias, revelando datos claros y precisos para tu modelado.*"

Con el fin de preparar los datos para el modelado de cómo se conjura el éxito de una película, deberá preprocesar los datos para completar y limpiar, sin hacer supuestos basados en pareceres personales, deben estar basados en alguna práctica o decisión justificada por el análisis de los datos, por ejemplo no se deben eliminar columnas por la falta de datos, no se deben imputar valores con medias o medianas (eso es un último recurso y no es aceptable en nuestro curso). Haga modelos de clasificación para las variables categóricas y haga regresiones para las variables numéricas. Aplique alguna estrategia razonada y justificada por la realidad, que le permita hacer simplificaciones para estos casos.

d.3) Modelamiento – Conjurar Modelos Predictivos:

"*Usen la magia de los modelos de Machine Learning para prever el éxito de una película, empleando tanto técnicas de clasificación como de regresión para descubrir la fórmula del éxito.*"

En alguna parte de su estrategia para ayudar a CineMagix deberá preparar un modelo de regresión para predecir el éxito financiero que CineMagix busca, por lo que deberá modelar con 3 o 4 algoritmos de ML lo que busca CineMagix, los únicos algoritmos permitidos son: Regresión Lineal, Árbol de Decisión Regresión Logística y Support vector Machine. Deberá probar al menos 3 de estos 4 y seleccionar el mejor basándose en las métricas interpretadas para hacer su justificación.

d.4) Optimización de Hiper-parámetros:

"*Refinen su artefacto mágico (modelo) para asegurar que sus predicciones sean tan agudas y precisas como sea posible, validando su poder en diferentes escenarios.*"

Para la búsqueda del mejor modelo que le permita entregar las mejores recomendaciones, los mejores insight y la mejor receta a CineMagix sobre futuras producciones, deberá optimizar el desempeño de su

modelo a través de la optimización de hiper-parámetros. Y posterior a este proceso, decida con que modelo se presentará a recomendar a CineMagix.

d.4) Recomendaciones - Desvelando los Secretos:

"*Revela los secretos descubiertos por tu modelo, proporcionando insights críticos y recomendaciones estratégicas que podrían guiar a CineMagix hacia un futuro de éxitos cinematográficos.*"

Resultado Esperado: CineMagix espera que esta aventura no solo descubra un modelo preciso y robusto que pueda prever el éxito financiero de las películas, sino también que los insights y recomendaciones derivadas de este viaje puedan iluminar su camino hacia decisiones de inversión más sabias y fundamentadas en futuros proyectos cinematográficos.

NOTA: En este punto se podría encontrar con que algunos algoritmos no proveen el mecanismo de la importancia de las variables, por lo que deberá buscar las formas alternativas, como el RFE (Recursive Feature Elimination) remueven iterativamente las características menos importantes. Las últimas en ser eliminadas son las más importantes.

E. Dataset provisto por CineMagix

Los datos que se utilizarán para este proyecto provienen de una extensa colección cinematográfica consolidada a lo largo de los años por CineMagix. Estos datos incluyen información detallada sobre películas, abarcando desde su rendimiento financiero y críticas, hasta detalles específicos de producción como directores y actores principales. El dataset para este Taller-N°2 sobre PREDICCIONES se denomina "cinemagix-movies.csv", el cual representa una muestra de 10 mil películas producidas y estrenadas a lo largo de varias décadas. Esta compilación reúne títulos de diversos géneros, directores y actores, proporcionando una visión panorámica de la industria cinematográfica y sus tendencias a lo largo del tiempo. Las características del dataset son las siguientes:

1. **MovieName**: Nombre de la película.
2. **YearOfRelease**: Año de estreno de la película.
3. **RunTime**: Duración en minutos de la película.
4. **MovieRating**: Rating asignado por los usuarios a la película según IMDb.
5. **Votes**: Número de votos en IMDb para esta película.
6. **MetaScore**: Representa el rating ponderado entregado por los críticos expertos en cine (varía de 0 a 100), y una puntuación de 90 o más se considera excelente. (ver https://www.metacritic.com/)
7. **Gross**: Ingresos brutos de la película en USD $ (dólares estadounidenses).
8. **Genre**: Géneros a los que está asociada la película (pueden ser más de 1)
9. **Certification**: Representa la calificación ya sea cinematográfica de la MPA o de televisión de la FCC. Ambos son sistemas para clasificar contenido y edad recomendada para la audiencia.
10. **Director**: Nombre del o los directores de la película o programa de TV
11. **Stars**: Elenco o actores principales de la película.
12. **Description**: Un breve resumen o trama de la película.

F. Entregables

Para lo anterior se pide entrega, con la estructura del nombre indicado:

1. Archivo **CineMagix-GRUPO-Nnn.ipynb** con todo el código y los análisis. (donde nn=n° del grupo asignado)

2. Documento en pdf (**CinaMagi-GRUPO-Nnn.pdf)** con el siguiente contenido en secciones que el grupo de verá proponer
   i) El análisis descriptivo de los datos, gráficos e **interpretación** de lo todo lo graficado y observado.
   ii) La **explicación de la estrategia** para abordar los aspectos técnicos del preprocesamiento
   iii) **Justificación** de todas las **decisiones** de modelamiento.
   iv) La **propuestas** y **justificaciones** de **grillas** para la optimización de hiper-parámetros.
   v) La est**rategia para responder a CineMagix**.
   vi) Los insights y conclusiones encontradas
   vii) Las recomendaciones y estrategia entregada a CineMagix para mejorar su éxito financiero produciendo películas.
3. Buena documentación y presentación de su código en Jupiter (Python)
   i) Justificación del uso de librerías (la idea es que no haya en el código liberías que no se usan y que resulten de copias de código)
   ii) Nivel de documentación del código. (Orden y claridad del markdown usado para documentar el código)
4. Formalidad del documento pdf (tapa con grupo e integrantes, asignatura, tabla de contenidos, cabecera sobria y pie de página)

Tabla de Evaluación del Taller N°2

Aspecto Evaluado	Puntaje Máximo
1. EDA (profundidad, primeros insights, interpretaciones, uso del analizado, etc.)	21
2. Pre-procesamiento (estrategias, uso de técnicas, como imputan, justificación de decisiones, etc.)	28
3. Modelamiento (uso de 3 algos ML, selección del mejor, uso de métricas de desempeño, etc.)	28
4. Optimización Hiper-parámetros (criterios selección grilla, justificaciones, estrategia,	14
5. Insights, recomendaciones, conclusiones, Propuesta a CineMagix. (estrategia de solución, modelo usado para proponer la estrategia, búsqueda del "éxito financiero", selección de variables para proponer estrategia a CineMagix etc.). También en este punto se resumirá el punto del uso de la metodología CRISP-DM.	28
6. Código (documentado y librerías justificadas)	07
7. Documento (explicaciones y justificaciones de decisiones y estrategias + formalidad)	14
Total Puntos	140

$$Nota\ Final\ Taller\ N°2 = \frac{1}{20}\sum puntaje\_aspecto_i$$

Fecha de Entrega: Sábado 11-Noviembre-2023 20:00 hrs

BUEN DESARROLLO DE PROYECTO !!