Informe Reto Machine Learning - Análisis de datos

Solución Reto Machine Learning - Análisis de datos

Parte 1.1: Análisis Exploratorio de Datos (EDA)

Para abordar el reto de realizar un análisis exploratorio de datos exhaustivo, se implementaron varias etapas clave en el análisis del dataset de restaurantes. El objetivo fue explorar el dataset y sus correlaciones, identificando conclusiones o correlaciones clave que permitan entender en profundidad la información reflejada por los datos. El proceso no se limitó a un análisis superficial, sino que se realizó con la profundidad adecuada, utilizando diversas visualizaciones y técnicas de exploración y manipulación de datos. Todo el proceso se documentó en un notebook de Jupyter, empleando el stack de herramientas de EDA del ecosistema Python, incluyendo pandas, matplotlib, seaborn y scipy.

Etapas del Análisis Exploratorio de Datos:

1. **Descripción General del Dataset**: Se realizó una primera exploración para comprender la estructura y las características básicas del dataset, incluyendo el análisis de las variables y la identificación de valores nulos o datos faltantes.
2. **Análisis Univariado:** Se examinaron las distribuciones de las variables individuales para entender mejor sus características y comportamientos.
3. **Análisis Bivariado:** Se exploraron las relaciones entre pares de variables, lo que permitió identificar patrones y posibles correlaciones significativas.
4. **Análisis de Correlación**: Se calcularon y analizaron las correlaciones entre todas las variables relevantes para identificar aquellas con una relación significativa con la facturación anual y otros factores clave.
5. **Visualización y Conclusiones Clave**: Se utilizaron diversas visualizaciones para representar los hallazgos de manera clara y comprensible, apoyando las conclusiones con estadísticas clave y visualizaciones efectivas. Para ver las imágenes ir al notebook EDA.ipynb adjunto.

Con este enfoque metodológico, se lograron identificar conclusiones y correlaciones significativas que proporcionan una comprensión profunda de los datos del dataset de restaurantes, sustentadas por visualizaciones y estadísticas clave.

Descripción General del Dataset.

El dataset consta de 3493 registros y 34 columnas. con las siguientes columnas:

#	Variable	Descripción
1	Registration Number	Número de registro
2	Annual Turnover	Volumen de negocios anual
3	Cuisine	Tipo de cocina
4	City	Ciudad
5	Restaurant Location	Ubicación del restaurante
6	Opening Day of Restaurant	Día de apertura del restaurante
7	Facebook Popularity Quotient	Índice de popularidad en Facebook
8	Endorsed By	Respaldado por
9	Instagram Popularity Quotient	Índice de popularidad en Instagram
10	Fire Audit	Auditoría contra incendios
11	Liquor License Obtained	Licencia de alcohol obtenida
12	Situated in a Multi Complex	Situado en un complejo múltiple
13	Dedicated Parking	Estacionamiento dedicado
14	Open Sitting Available	Asientos al aire libre disponibles
15	Restaurant Tier	Nivel del restaurante
16	Restaurant Type	Tipo de restaurante
17	Restaurant Theme	Tema del restaurante
18	Restaurant Zomato Rating	Calificación del restaurante en Zomato
19	Restaurant City Tier	Nivel de la ciudad del restaurante
20	Order Wait Time	Tiempo de espera del pedido
21	Staff Responsiveness	Capacidad de respuesta del personal
22	Value for Money	Relación calidad-precio
23	Hygiene Rating	Calificación de higiene
24	Food Rating	Calificación de la comida
25	Overall Restaurant Rating	Calificación general del restaurante
26	Live Music Rating	Calificación de música en vivo
27	Comedy Gigs Rating	Calificación de actuaciones cómicas
28	Value Deals Rating	Calificación de ofertas de valor
29	Live Sports Rating	Calificación de deportes en vivo
30	Ambience	Ambiente
31	Lively	Animado
32	Service	Servicio
33	Comfortability	Comodidad
34	Privacy	Privacidad

Al

Observar el resumen estadístico y la información del dataset, notamos que hay valores nulos en varias columnas, siendo particularmente notable en:

Variable	Cantidad de Nulos
Facebook Popularity Quotient	99
Instagram Popularity Quotient	56
Restaurant Tier	49
Overall Restaurant Rating	212
Live Music Rating	765
Comedy Gigs Rating	2483
Value Deals Rating	2707
Live Sports Rating	3288
Ambience	25

Análisis de la distribución de la facturación anual para las variables categóricas:

El análisis de la tabla muestra que existen diferencias significativas en la facturación anual según el tipo de cocina (F-statistic: 5.138, p-value: 2.28E-12), la ciudad (F-statistic: 1.706, p-value: 9.35E-12), el día de apertura del restaurante (F-statistic: 1.565, p-value: 7.21E-21), y el tipo de endoso (F-statistic: 5.775, p-value: 0.0031). Por otro lado, no se encontraron diferencias significativas en la facturación anual según la ubicación del restaurante (F-statistic: 3.437, p-value: 0.0638) ni según el tipo de restaurante (F-statistic: 2.56, p-value: 0.0533), aunque ambas variables están cerca del umbral de significancia del 5%. Finalmente, no hay diferencias significativas en la facturación anual según el tema del restaurante (F-statistic: 0.766, p-value: 0.8286).

Las variables "Tipo de Cocina", "Ciudad", "Día de Apertura del Restaurante" y "Respaldado Por" presentan diferencias significativas en la facturación anual, sugiriendo que son factores determinantes para el éxito financiero de un restaurante. En contraste, las variables "Ubicación del Restaurante" y "Tipo de Restaurante", aunque no alcanzan significancia al nivel del 5%, se encuentran cerca del umbral, indicando una posible relevancia. Por último, la variable "Tema del Restaurante" no muestra diferencias significativas, lo que implica que no es un factor crucial en la facturación anual.

Análisis de la distribución de la facturación anual para las variables numéricas:

Los análisis indican que la variable "Número de Registro" no tiene una correlación significativa con la facturación anual. En contraste, el "Índice de Popularidad en Facebook" y el "Índice de Popularidad en Instagram" muestran una correlación positiva moderada, lo que sugiere que una mayor popularidad en estas plataformas está asociada con una mayor facturación. Las variables "Auditoría de Incendios", "Licencia de Licor Obtenida", "Ubicado en un Complejo Multifuncional",

"Estacionamiento Dedicado" y "Disponibilidad de Asientos Abiertos" no presentan correlaciones significativas.

La variable "Categoría del Restaurante" tiene una correlación negativa moderada, indicando que los restaurantes de menor categoría tienden a tener una mayor facturación. La "Calificación del restaurante en Zomato" muestra una correlación positiva débil. Otras variables, como "Tiempo de Espera del Pedido", "Categoría de la Ciudad del Restaurante" y "Calificación del Personal", no presentan correlaciones significativas.

Las variables "Valor por Dinero", "Calificación de Higiene", "Calificación de la Música en Vivo" y "Calificación de Deportes en Vivo" muestran correlaciones positivas moderadas, sugiriendo que mejores calificaciones en estos aspectos están asociadas con una mayor facturación. La "Calificación de la Comida" y la "Calificación General del Restaurante" tienen correlaciones positivas débiles. Las variables "Calificación de Shows de Comedia", "Calificación de Ofertas de Valor" y "Ambiente" no presentan correlaciones significativas. Asimismo, variables como "Servicio", "Comodidad" y "Privacidad" tampoco muestran correlaciones significativas con la facturación anual.

El análisis de las variables numéricas revela que la popularidad en redes sociales, tanto en Facebook como en Instagram, presenta correlaciones positivas moderadas, lo que indica que una mayor popularidad en estas plataformas está asociada con mayores ingresos. De manera similar, las calificaciones de higiene, la respuesta del personal, la música en vivo y la transmisión de deportes en vivo también muestran correlaciones positivas moderadas, sugiriendo que estos factores están relacionados con una mayor facturación.

En contraste, la categoría del restaurante presenta una correlación negativa moderada, lo cual indica que los restaurantes de menor categoría tienden a tener una mayor facturación. Este resultado, aunque contraintuitivo, merece un análisis más profundo para comprender sus implicaciones más adelante.

Las demás variables analizadas no muestran correlaciones significativas con la facturación anual, lo que sugiere que pueden no ser factores críticos para el éxito financiero de los restaurantes.

Conclusiones clave:

La popularidad en Facebook e Instagram presenta una correlación positiva moderada con la facturación anual, lo que sugiere que una mayor presencia en estas plataformas puede contribuir a aumentar los ingresos. En cuanto a las calificaciones de higiene y la responsabilidad del personal están moderadamente correlacionadas con la facturación anual, indicando que estos aspectos son importantes para el éxito financiero. La presencia de música en vivo y la transmisión de deportes en vivo también muestran una correlación moderada con una mayor facturación, sugiriendo que estos eventos pueden atraer a más clientes, la variedad de cocina y el tipo de endoso muestran diferencias significativas en la facturación anual, destacando su relevancia en la estrategia de marketing, la ciudad y la ubicación del restaurante son factores importantes que afectan significativamente la facturación anual, subrayando la importancia de una buena ubicación geográfica. Existen diferencias significativas en la facturación anual según el día de apertura, lo que podría estar relacionado con factores estacionales o estrategias específicas en el lanzamiento de nuevos restaurantes.

Parte 1.2: Modelos para predecir la facturación anual de los restaurantes.

El objetivo del análisis fue predecir la facturación anual de los restaurantes utilizando diversas técnicas de modelado. A continuación, se presenta el procedimiento estructurado seguido para la elaboración y evaluación de los modelos predictivos, documentado en un notebook de Jupyter utilizando el stack de herramientas de Python.

Preparación del Dataset

Para llevar a cabo el análisis y la creación de modelos predictivos, se siguieron varios pasos en la preparación del dataset. Primero, se importaron los datos utilizando la librería pandas y se realizó una exploración inicial para comprender la estructura del dataset, identificar el tipo de variables presentes y detectar valores faltantes. Posteriormente, se abordó la limpieza de datos mediante la imputación o eliminación de valores nulos según la variable en cuestión y la conversión de variables categóricas a variables dummy para facilitar su uso en los modelos.

En la etapa de ingeniería de características, se crearon nuevas variables a partir de las existentes para capturar información adicional y se normalizaron y estandarizaron las variables para asegurar que todas tuvieran la misma escala. Este proceso garantizó que los datos estuvieran listos para el modelado predictivo.

Creación de los modelos

En la creación de modelos predictivos, el primer paso fue dividir el dataset en conjuntos de entrenamiento y prueba utilizando la función train_test_split de scikit-learn. Esto permitió garantizar que los modelos se entrenaran con una parte de los datos y se evaluaran con otra parte no vista previamente, asegurando una evaluación objetiva de su desempeño.

Posteriormente, se seleccionaron varios algoritmos de aprendizaje automático para su comparación. Entre los modelos considerados se incluyeron la Regresión Lineal, Bosques Aleatorios, Gradient Boosting, Support Vector Machines (SVM) y Redes Neuronales. Esta diversidad de modelos permitió evaluar distintas aproximaciones y seleccionar la más adecuada para el problema en cuestión.

El siguiente paso involucró el entrenamiento de cada modelo en el conjunto de entrenamiento. Para asegurar la consistencia y robustez de los resultados, se utilizó la validación cruzada. Esta técnica permitió evaluar el rendimiento de los modelos en diferentes subconjuntos del conjunto de entrenamiento, proporcionando una estimación más precisa de su desempeño general y ayudando a identificar el modelo más confiable.

Evaluación de Modelos

Las métricas de evaluación utilizadas para analizar el rendimiento de los modelos incluyeron el RMSE (Root Mean Square Error), MAE (Mean Absolute Error) y R^2 (Coeficiente de Determinación). Estas métricas proporcionaron una visión integral sobre la precisión y capacidad predictiva de cada modelo. La comparación del rendimiento de los modelos se realizó en el conjunto de prueba, permitiendo identificar el modelo con el mejor desempeño basado en las métricas de evaluación. Este proceso garantizó la selección del modelo más adecuado para predecir la facturación anual de los restaurantes. Además, se llevaron a cabo visualizaciones de los resultados para comparar las predicciones del modelo con los valores reales. Esto incluyó gráficos de residuos que facilitaron el análisis de los errores de predicción, proporcionando información valiosa sobre la precisión y confiabilidad del modelo seleccionado.

Ajuste de Hiperparámetros

Para la optimización de los hiperparámetros, se emplearon técnicas como Grid Search y Random Search con el objetivo de identificar los parámetros óptimos para los modelos. Posteriormente, el modelo seleccionado se reentrenó utilizando estos hiperparámetros óptimos. Finalmente, se llevó a cabo una evaluación en el conjunto de prueba para confirmar la mejora en el rendimiento del modelo.

Presentación de Resultados

El proceso de documentación se realizó meticulosamente, registrando cada paso del análisis y modelado en el notebook de Jupyter. Se incluyeron visualizaciones y explicaciones detalladas para cada etapa del proceso, asegurando una comprensión clara y completa del análisis. Ver notebook adjunto Modelos.ipynb.

Conclusiones clave:

En el análisis de modelos predictivos para la facturación anual de restaurantes, se realizaron varias iteraciones de preprocesamiento y optimización, obteniendo los siguientes resultados:

Resultados Iniciales (Iteración 1): Se aplicó imputación y codificación básica de datos. Los modelos de regresión lineal, árboles de decisión y redes neuronales presentaron un desempeño deficiente, con MSE elevados y valores negativos de R^2, indicando que los modelos no capturaban adecuadamente la variabilidad de los datos.

Linear Regression: MSE = 5.777425e+14, R^2 = -0.270687

Decision Tree: MSE = 7.330364e+14, R^2 = -0.612241

Neural Network: MSE = 1.398314e+15, R^2 = -2.075452

Resultados Mejorados (Iteración 2): Se implementó la remoción de outliers y la selección de características. Esto mejoró significativamente el desempeño de los modelos, reflejado en menores MSE y valores positivos de R^2

Linear Regression: MSE = 4.077893e+14, R^2 = 0.103108

Decision Tree: MSE = 4.339332e+14, R^2 = 0.045607

Neural Network: MSE = 1.097466e+15, R^2 = -1.413768

Resultados con Validación Cruzada y Nuevos Modelos (Iteración 3): Se añadió validación cruzada y se exploraron nuevos modelos, incluyendo Random Forest. Los resultados mostraron mejoras adicionales.

Random Forest: MSE = 4.269040e+14, R^2 = 0.061067

Optimización de Hiperparámetros (Iteración 4): Se aplicó optimización de hiperparámetros y validación cruzada. Los modelos mostraron mayor estabilidad y mejor desempeño.

Linear Regression: MSE = 4.077893e+14, R^2 = 0.103108, Mean CV MSE = -1.244418e+14

Decision Tree: MSE = 4.339332e+14, R^2 = 0.045607, Mean CV MSE = -1.344919e+14

Neural Network: MSE = 4.285952e+14, R^2 = 0.057347, Mean CV MSE = -1.932010e+14

Random Forest: MSE = 4.167118e+14, R^2 = 0.083484, Mean CV MSE = -1.290426e+14

Inclusión de Gradient Boosting y SVR (Iteración 5): Se añadieron modelos de Gradient Boosting y Support Vector Regression, con optimización de hiperparámetros.

Gradient Boosting: MSE = 4.104605e+14, R^2 = 0.097233, Mean CV MSE = -1.261745e+14

SVR: MSE = 4.548227e+14, R^2 = -0.000337, Mean CV MSE = -1.582058e+14

Mejoras con Codificación y Selección de Características (Iteración 6): Se aplicaron técnicas avanzadas de codificación, imputación y selección de características utilizando Isolation Forest y Lasso Regression. Esto resultó en una mejora significativa en el rendimiento de los modelos.

Linear Regression: MSE = 0.217569,

R^2 = 0.268769

Random Forest: MSE = 0.236267, R^2 = 0.205926

Gradient Boosting: MSE = 0.221447, R^2 = 0.255735

Optimización y Validación Cruzada (Iteración 7): Se refinó la optimización de hiperparámetros y validación cruzada.

Ridge Regression: MSE = 0.217225, R^2 = 0.269926, Mean CV MSE = -0.235154

Random Forest: MSE = 0.236253, R^2 = 0.205976, Mean CV MSE = -0.245763

Gradient Boosting: MSE = 0.220602, R^2 = 0.258577, Mean CV MSE = -0.238739

Creación de Nuevas Características (Iteración 8): Se crearon nuevas características y se aplicó nuevamente la optimización de hiperparámetros y validación cruzada.

Ridge Regression: MSE = 0.216387, R^2 = 0.272743, Mean CV MSE = -0.233622

Random Forest: MSE = 0.232713, R^2 = 0.217872, Mean CV MSE = -0.240553

Gradient Boosting: MSE = 0.223093, R^2 = 0.250206, Mean CV MSE = -0.233701

En resumen, a lo largo de las iteraciones, se observaron mejoras significativas en el rendimiento de los modelos. La optimización de hiperparámetros y la creación de nuevas características resultaron en modelos más robustos y precisos, con Ridge Regression mostrando el mejor desempeño general.

A lo largo de las diversas iteraciones de optimización y preprocesamiento de los modelos predictivos, aunque se observaron mejoras en el rendimiento, los resultados obtenidos no alcanzaron los niveles esperados de precisión y robustez. Los valores de MSE y R^2 indican que aún existe una considerable variabilidad en los datos que no ha sido capturada por los modelos actuales. Como sugerencia para futuras mejoras por tema de tiempos de la prueba, profundizar en el análisis de los datos y aplicar técnicas adicionales de ingeniería de características. Esto incluye la exploración y selección de nuevas variables que puedan tener un impacto significativo en la predicción de la facturación anual de los restaurantes. Además, realizar análisis más detallados y experimentos con diferentes combinaciones de características podría conducir a una mejor comprensión de las relaciones subyacentes en los datos y, por ende, a un mejor desempeño de los modelos predictivos.