Análisis Exploratorio de Datos y Machine Learning para Paxer

LA-CoNGA physics International Network School 2022

Victor Guzmán

05/12/2022

Latin American alliance for Capacity buildiNG in Advanced physics

LA-CoNGA physics

Cofinanciado por el programa Erasmus+ de la Unión Europea

1. Data Science for Business
2. Objetivos
3. Metodología
4. Resultados obtenidos e interpretación
5. Conclusión y perspectiva

Data Science for Business

- ¿Por qué analizar datos?
- Data Science para las empresas
- Industria hotelera

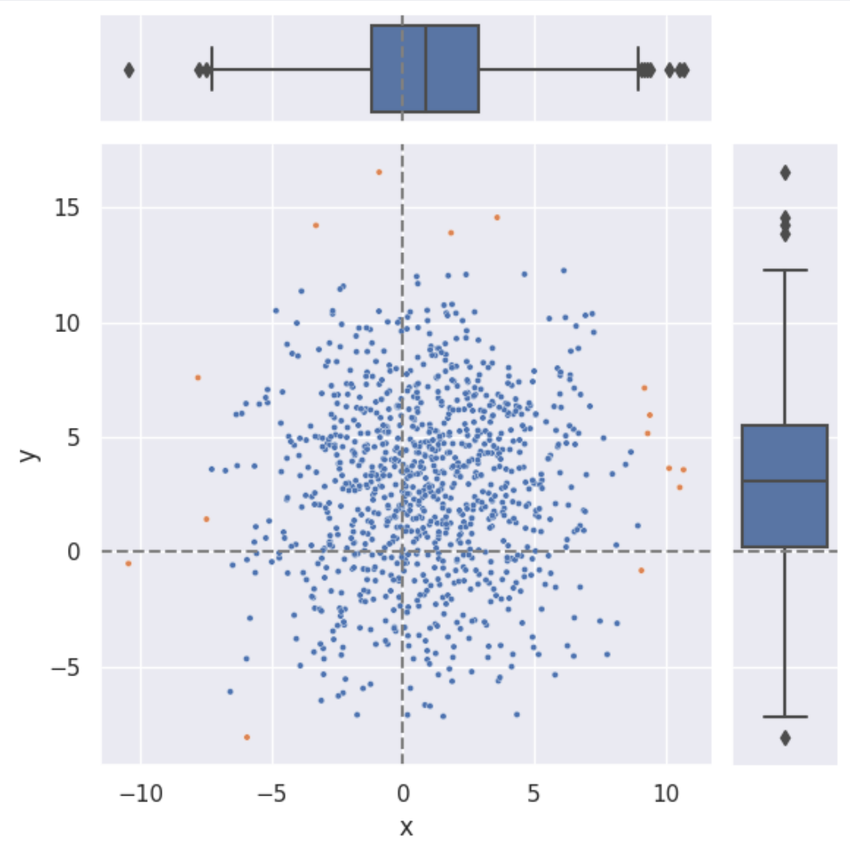
¿Por qué analizar datos?

Descripción del estado actual de la organización
(o de procesos específicos)

Detección de situaciones anómalas

Estudiar la causa de eventos particulares

Extraer insights/perspectivas



Data Science para las empresas

DataCamp (2020) - Blog

Relevancia: El 89% de las empresas están priorizando la relevancia de los datos.

Advancing data literacy in the post-pandemic world - Paris21 (2022) - Meeting

Necesidad: Urgencia de ir más allá del entendimiento ad-hoc.

IBM Global AI Adoption Index (2022) - Report

Adopción: Creció la tasa de adopción de IA hasta el 35 %

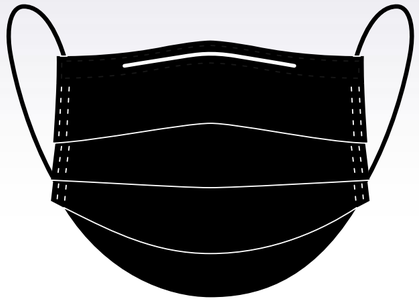
Sostenibilidad: dos tercios (66 %) de las empresas están ejecutando o planeando aplicar AI para abordar sus objetivos de sostenibilidad.

Usos: Automatización, Toma de decisiones informadas,...

Pandemia y oportunidades con IA

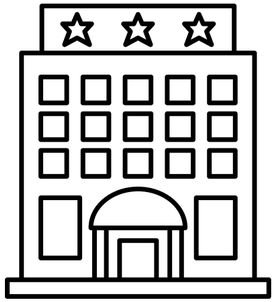
Quantifying the economic impact of COVID-19 on the U.S. hotel industry (2021) - Article

Pérdidas: La ganancia por habitación cayó en un 86%, muchos hoteles cerraron.



Aceptación e implementación de la IA en la industria hotelera (2021) - Book

Dinámica: La dinámica en estas empresas y su relación con los clientes ha cambiado.



Objetivos

- General
- Específicos

Objetivos

General

Estudiar el comportamiento de Paxer y de sus clientes por regiones en el transcurso del año 2022

Específicos

- Identificar un conjunto de problemas que permitar entenden el estatus actual en cada Región
- Identificar un conjunto de relaciones que sirvan para dar respuesta a los problemas planteados
- Realizar el proceso de Data Wrangling de acuerdo a los problemas planteados en el contexto de las relaciones establecidas entre las variables
- Realizar el proceso de Análisis Exploratorio de Datos para entender el comportamiento de las variables
- Estudiar la posibilidad de establecer un modelo de machine learning para alguno de los problemas establecidos por medio de la creación de un modelo base

Metodología

- Identificación de un conjunto de problemas
- Data Wrangling
- Análisis Exploratorio de Datos (EDA)
- Ajuste del modelo
- Evaluación del modelo

Identificación de un conjunto de problemas (Python)

Identificación de KPIs en la industria hotelera

Identificación de variables en Paxer

Comportamiento de los clientes de Paxer durante el 2022

Comportamiento de Paxer en el transcurso del año 2022

Data Wrangling y Exploratory Data Analysis

Quality Assurance

- Verificación de datos faltantes
- Modificación del tipo de datos
- Limpieza del dataset
- Validación del dataset modificado
- Interpretación de los objetivos establecidos en términos del nuevo dataframe

Exploratory Data Analysis

- Exploración del comportamiento de variables
- Visualizaciones para identificar la relaciones
- Identificación de eventos anómalos

Modelo base y rendimiento

Modelos de clasificación

Planteamiento del problemas de interés como un problema de clasificación

Modelo base (Aprendizaje Supervisado)

Implementación de modelos usando criterios de información

Evaluación del modelo

- Uso del accuracy, precision, recall y el f1-score
- División en conjuntos de entrenamiento, validación y prueba

Resultados e interpretación

- Exploratory Data Analysis
- Machine Learning Model
- Creación de funciones

Exploratory Data Analysis

Identificación de problemas en el dataset

- Se identificaron datos que habían sido introducidos de manera inapropiada (que podía afectar en ciertos aspectos).

Estado de Paxer y sus clientes (2022)

- El comportamiento de los clientes depende de su ubicación
- Se observan unas temporadas altas y bajas

Creación de funciones

- Cargue, cree y modifique un grupo de archivos csv de forma automática.
- Visualización de histogramas (con diagramas de caja) y conteos
- Exploración numérica de características de las variables

Modelo de Machine Learning para Clasificación y Predicción

Creación de funciones para el modelo:

- Función para preparar el dataset (features y labels)
- Función para aplicar un muestreo con reemplazo
- Identificación del modelo óptimo (exactitud mas alta)
- Implementación del modelo
- Generación de un reporte de clasificación
- Guardar modelo…

Feature engineering

Usando el criterio de impureza Gini:

- Seleccionando 3 variables
- Evaluando su "peso" en la predicción

Model validation

- Validación cruzada
  - ≈90% para todas las métricas
- "Rendimiento" con el conjunto de prueba
  - ≈65% exactitud
  - Muestra no representativa

Conclusiones

- Exploratory Data Analysis
- Machine Learning Model
- Creación de funciones

Paxer y sus clientes en el transcurso del año 2022

Análisis de datos

Volumen de datos: Visualizaciones para datos anómalos.

Países y épocas: Existen dinámicas propias de cada región que dependen en gran manera de la época.

Datos no estructurados: No fueron cubiertos (textos).

Flexibilización de la cuarentena en América Latina: Comportamiento durante el 2022

Resultados e interpretación

Paxer y sus clientes en el transcurso del año 2022

Decenas de variables

Se utilizaron 3 para la predicción

Modelo de clasificación

El desbalanceo de clases debe ser tomado en cuenta (clientes con características diferentes - Segmentación podría ayudar).

Validación y prueba

La diferencia en la validación y la prueba es un indicador de que la muestra no es representativa

Creación de funciones y documentación

- Las buenas prácticas de programación son necesarias considerando la cantidad de variables y los posibles cambios que pueden ocurrir en curso.

- Otras técnicas pudiésen aportar información de interés:
  - Deep Learning (LSTM): Long Short-Term Memory networks
  - Kernel Methods
  - Aprendizaje no supervisado