PROGRAMACIÓN PARA CIENCIA DE DATOS II

Jorge Andres Melo Mayorga
Facultad de Ingeniería, Ciencia de Datos, Universidad Compensar
Actividad Practica Aplicada, Etapa de Transferencia

Sebastián Rodriguez Muñoz
Docente

27 de septiembre de 2024

Contenidos temáticos:

- Métodos Estadísticos Computacionales.
- Diseño de experimentos computacional.
- Manuales de manejo de software específico.
- Protocolo de presentación final de trabajo.

Descripción de la actividad:

El propósito de esta actividad es desarrollar sus habilidades para presentar los resultados que se han obtenido a lo largo del curso. Con esto en mente el entregable que se estará evaluando es un dashboard con los principales descubrimientos sobre la problemática que fue planeada al inicio. Teniendo esto en cuenta, se realizan las siguientes recomendaciones para mejorar el modelo que se ha estado trabajando.

1. Evalué los resultados que ha obtenido en relación con la problemática que se definió durante la primera actividad. Asegúrese, que cada uno de los análisis que haya planteado sea con el objetivo de mejorar la compresión que se tiene del problema.
2. Realice una evaluación del rendimiento que presenta el modelo actual, la habilidad que tiene de ofrecer información sobre la problemática. Define si debe mejorar el modelo iterativamente, o se debe replantear el enfoque de la pregunta inicial para obtener mejores resultados.
3. Considere la recopilación de datos o un procesamiento adicionales de acuerdo a las técnicas que se vieron en el apartado teórico de esta etapa. Puede realizar diferentes y comprobar como mejora el rendimiento dependiendo de la calidad de los datos utilizados.
4. Experimente con los hiperparámetros que definen el modelo. Es importante que tenga métricas que le permitan evaluar el desempeño de los mismos con el objetivo de optimizar los resultados obtenidos. experimentación controlada con los hiperparámetros es una práctica recomendable.

5. Considere si alguno de las métricas del modelo mejoraría información del algún aspecto del problema. La mejora de la capacidad predictiva se puede lograr mediante la creación de nuevas características o la adaptación de las existentes.
6. Considere la Regularización: En caso de que se detecte un sobreajuste en el modelo, se debe considerar la aplicación de técnicas de regularización. Estas implican la incorporación de términos de penalización en la función de pérdida.

Mejore el modelo iterativamente hasta que las evidencias indiquen que el modelo es fiable y se tenga una comprensión avanzada del problema que permitan realizar planteamientos para solucionar el problema.

Una vez este satisfecho con las conclusiones que ha obtenido del modelo, es momento de presentarlo ante una audiencia. Como se mencionó anteriormente, se debe realizar un tablero dashboard con las conclusiones que se hayan obtenido del modelo. A continuación, se presenta un guía de las actividades que debe realizar.

1. Comienza importando las bibliotecas necesarias, como Dash, Plotly y Pandas. Luego, carga los datos de tu modelo de ciencia de datos en un DataFrame de Pandas para que estén listos para su uso.
2. Define las visualizaciones que deseas mostrar y utiliza la biblioteca Dash para crear componentes como gráficos, tablas y filtros interactivos. Organiza estos componentes de manera lógica en la interfaz del dashboard.
3. Da vida a tu dashboard mediante el uso de elementos interactivos. Utiliza callbacks de Dash para permitir a los usuarios seleccionar datos específicos, cambiar escalas o aplicar filtros dinámicos. Asegúrate de que las visualizaciones respondan en tiempo real a las acciones de los usuarios.
4. Agrega contexto y narrativa a tus datos. Utiliza texto descriptivo y explicativo en el dashboard para contar una historia coherente. Acompaña tus puntos clave con gráficos y visualizaciones relevantes para respaldar tu narrativa.
5. Personaliza la apariencia del dashboard para que sea atractivo y coherente con tu narrativa. Utiliza colores, fuentes y diseños que mejoren la legibilidad y la

comprensión. Presta atención a la organización y al espacio en blanco para una presentación ordenada.

6. Realiza pruebas exhaustivas del dashboard. Asegúrate de que todas las interacciones funcionen correctamente y de que la narrativa sea clara. Obtén retroalimentación de otros usuarios o colegas para mejorar la calidad de la presentación y, finalmente, optimiza el rendimiento del dashboard para una carga eficiente.
7. Organiza toda la documentación, archivos y scripts requeridos para ejecutar el dashboard en una carpeta. Realiza pruebas en otros PC o en máquinas virtuales para asegurarte que todo siga funcionando como se espera.

Entregable:

Debe entregar una carpeta comprimida que contenga los siguientes archivos que evidencien su trabajo en el proyecto.

- Informe sobre el reportando todo el proceso de desarrollo del proyecto, se deben presentar una justificación de las decisiones tomas a lo largo del proceso.
- Carpeta con todos los archivos requeridos para visualizar el dashboard.
- Un archivo de text con dos links: el primero con un link al repositorio en GitHub del proyecto. Y otro link para visualizar el dashboard in Binder.
- Diligencie únicamente para actividades tipo foro.

DASHBOARD DE E-COMMERCE - ANÁLISIS DE DATOS Y PUBLICACIÓN RESULTADOS

El proyecto consiste en realizar el análisis y visualización de resultados de los datos de e-commerce, basándose en el dataset público de comercio electrónico brasileño proporcionado por Olist. Fuente: Brazilian E-Commerce Public Dataset by Olist URL: *https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce*.
Conjunto de datos públicos de comercio electrónico brasileño de pedidos realizados en Olist Store. El conjunto de datos tiene información de 100 000 pedidos de 2016 a 2018 realizados en varios mercados de Brasil. Sus características permiten ver un pedido desde múltiples dimensiones: desde el estado del pedido, el precio, el pago y el rendimiento del flete hasta la ubicación del cliente, los atributos del producto y, finalmente, las reseñas escritas por los clientes. También publicamos un conjunto de datos de geolocalización que relaciona los códigos postales brasileños con las coordenadas de latitud y longitud. Se trata de datos comerciales reales, que han sido anonimizados y las referencias a las empresas y socios en el texto de la reseña han sido sustituidas por los nombres de las grandes casas de Juego de Tronos.

OBJETIVOS

Comprender cómo los factores geoespaciales, los tiempos de entrega y los costos de envío se relacionan con los retrasos, con el fin de optimizar la logística y mejorar la eficiencia en las entregas. Para lograrlo, me enfocaré en los siguientes aspectos:

A. **Identificar los tiempos de entrega**: relación con los costos asociados en diferentes regiones para detectar áreas geográficas específicas donde los retrasos son más frecuentes.
B. **Utilizar modelos de regresión**: Analizar cómo variables como la distancia, el costo de envío y los métodos de transporte afectan los tiempos de entrega. Esto me permitirá cuantificar la influencia de cada factor y priorizar las acciones de mejora.

C. **Examinar la variabilidad en los tiempos de entrega**: En función de factores como los días de la semana, las horas pico y las condiciones meteorológicas. Esta información será crucial para mejorar la planificación logística.

D. **Evaluar el impacto de los costos de envío**: en los tiempos de entrega y en la satisfacción del cliente. Esto me ayudará a determinar si es necesario ajustar las tarifas o renegociar acuerdos con proveedores logísticos.

E. **Dashboard**: para visualizar el resultado del análisis obtenido.

CONTENIDO

El proyecto está organizado de manera estructurada y cargado en mi GitHub:[1]
https://github.com/jandresmelo/EDUCATIVO/tree/a8ccd4c6574db01e0cefcbdad155ddccc95f4a4d/OLIST_ECOMMERCE

1. **Datos descargados**:
   Contiene los datasets originales utilizados en el proyecto. Esta sección almacena los datos sin procesar tal como fueron descargados desde la fuente. Carpeta: OLIST_ECOMMERCE/00_DatosDescargados
2. **Base de datos PostgreSQL**:
   Almacena los datos estructuradamente para su posterior análisis. Incluye la configuración de la base de datos PostgreSQL con extensiones GIS para manejar datos geoespaciales y realizar consultas complejas. Carpeta: OLIST_ECOMMERCE/01_BaseDatosSQL.
3. **Análisis descriptivo de los datos**:
   Contiene los scripts y notebooks que realizan el análisis exploratorio y descriptivo de los datos, utilizando tanto PostgreSQL GIS como Python. Esta sección incluye visualizaciones, estadísticas descriptivas y mapas geoespaciales.
   Carpeta: OLIST_ECOMMERCE/02_AnalisisPython

---

[1] *https://github.com/jandresmelo/EDUCATIVO/tree/a8ccd4c6574db01e0cefcbdad155ddccc95f4a4d/OLIST_ECOMMERCE*

4. **Regresión lineal y logística**:

   Desarrolla los modelos de regresión lineal y logística para predecir diferentes variables de interés. Incluye la implementación de los modelos, la evaluación del rendimiento y la interpretación de los resultados.

   Carpeta: OLIST_ECOMMERCE/03_Regresión

5. **Dashboard interactivo**:

   Un dashboard interactivo desarrollado en Python utilizando Streamlit. Este dashboard permite a los usuarios explorar los datos de manera dinámica y visualizar los principales resultados del análisis.

   Carpeta: : OLIST_ECOMMERCE/04_Dashboard, publicado en: *https://educativo-qiqzvyawdt4jmyzjtafrth.streamlit.app/#resultados-analisis-de-datos-bd-olist-e-commerce*

6. **Documento del proyecto**:

   Proporciona una descripción detallada del proyecto, incluyendo el objetivo, metodología, resultados y conclusiones. Este documento sirve como una guía completa para entender el desarrollo y los hallazgos del proyecto. Ubicación: OLIST_ECOMMERCE/05_Documento

CARACTERÍSTICAS DEL PROYECTO

- **Análisis de clientes y pedidos**: Número de clientes por estado, mapa de calor de ubicación de clientes, distribución de órdenes por estado del pedido.
- **Análisis de retrasos**: Distribución de retrasos en la entrega, boxplot de retrasos por método de pago, Q-Q plots de residuos de regresión lineal y Random Forest.
- **Información de productos y vendedores**: Radar chart para la cantidad de vendedores en las principales ciudades, visualización de categorías de productos mediante caras personalizadas.
- **Métodos de pago y evaluaciones**: Distribución de métodos de pago, valor de pagos por número de cuotas, análisis de reseñas.
- **Matriz de correlaciones**: Análisis de correlación entre variables clave.

CONCLUCIONES

La actividad propuesta por el docente señala la importancia fundamental de la integración de conocimientos estadísticos y herramientas tecnológicas avanzadas para llevar a cabo un análisis de datos efectivo y significativo. A través de la construcción de un dashboard interactivo utilizando Python y tecnologías como Streamlit, PostgreSQL, y diversas bibliotecas de visualización, se logra no solo procesar y analizar grandes volúmenes de datos, presentando así los resultados de una manera clara y comprensible.

El conocimiento estadístico es esencial para comprender los datos en profundidad, identificar patrones, tendencias y posibles anomalías. Durante el ejercicio, se aplicaron técnicas estadísticas como la regresión lineal y la evaluación de modelos (MSE, R^2) para predecir y entender la relación entre diversas variables, como el retraso en la entrega y el precio de los productos. Sin una sólida comprensión de estos conceptos, el análisis podría haber sido superficial, perdiendo la oportunidad de extraer información valiosa para la toma de decisiones.

El Uso de Herramientas Tecnológicas para el Análisis de Datos, como Python, PostgreSQL, y bibliotecas de visualización como Matplotlib, Seaborn, y Folium, fue crucial para el éxito del análisis. Estas herramientas permitieron manejar grandes conjuntos de datos, realizar análisis complejos, y generar visualizaciones claras y persuasivas. Además, la implementación en Streamlit permitió que el análisis sea accesible de manera interactiva a través de un dashboard, facilitando la interpretación de los resultados y mejorando la experiencia del usuario final.

La visualización de datos geográficos mediante mapas proporcionó una dimensión adicional crucial al análisis. Herramientas como Folium permitieron representar espacialmente la distribución de clientes y los retrasos en las entregas, revelando patrones que no habrían sido evidentes en gráficos convencionales. Esta capacidad de visualizar datos geográficos es

especialmente útil en escenarios de comercio electrónico, donde la ubicación geográfica puede influir significativamente en la logística y la experiencia del cliente. Los mapas no solo embellecen los informes, sino que también permiten una comprensión más profunda y accesible de los datos analizados.

En conjunto, este ejercicio resalta cómo la sinergia entre conocimiento estadístico y herramientas tecnológicas avanzadas puede transformar datos en decisiones estratégicas, permitiendo a las organizaciones optimizar operaciones y mejorar sus servicios en un entorno competitivo.

Anexo 1 - Ajuste de Análisis Estadístico.
Anexo 2 – Ajuste de Regresión.
Anexo 3 – Resultados Dashboard.