Logística de envíos: ¿Cuándo llega?

Mentoría DiploDatos 2019

Práctico: Aprendizaje supervisado y no supervisado

Motivación

En la actualidad, cada vez más productos se comercializan a través de una plataforma online. Una de las principales ventajas de este sistema es que el usuario puede recibir el producto en su domicilio en una fecha determinada. Pero, ¿cómo sabemos qué día va a llegar? ¿A partir de qué datos podemos predecir la demora del envío? En este práctico se trabajará con datos de envíos de MercadoLibre, el e-commerce más grande de Latinoamérica, analizando y modelando el problema de logística de envíos para poder responder ¿cuándo llega?

Descripción del dataset

Datos: El conjunto de datos seleccionado para realizar el práctico corresponde a un muestreo aleatorio no uniforme de 500.000 envíos de MercadoLibre. Estos envíos fueron realizados en Brasil en el período comprendido entre Octubre de 2018 y Abril de 2019 (las fechas han sido modificadas y adaptada a un período de tiempo diferente, conservando el día de la semana y considerando los feriados correspondientes). Los datos comprenden variables tanto categóricas como numéricas.

El dataset presenta la siguiente información:

- Sender_state: Estado de Brasil de donde sale el envío.
- Sender_zipcode: Código postal (de 5 dígitos) de donde sale el envío.
- Receiver_state: Estado de Brasil a donde llega el envío.
- Receiver_zipcode: Código postal (de 5 dígitos) a donde llega el envío.
- Shipment_type: Método de envío (normal, express, super).
- Quantity: Cantidad de productos en un envío.
- Service: Servicio del correo con el cual se realizó un envío.
- Status: Estado del envío (set: listo para ser enviado, sent: enviado, done: entregado, failed: no entregado, cancelled: cancelado).
- Date_created: Fecha de creación del envío.
- Date_sent: Fecha y hora en que se realizó el envío (salió del correo).
- Date_visit: Fecha y hora en que se entregó el envío al destinatario.
- Shipment_days: Días hábiles entre que el envío fue enviado (salió del correo) y que fue entregado.

Objetivo

El objetivo de este práctico es realizar de manera completa el proceso de desarrollo de un modelo de aprendizaje automático supervisado y no supervisado para determinar cuándo llega un envío. Se busca desarrollar el conocimiento práctico sobre dicho proceso, desde la definición de los datasets, la elección y análisis del modelo y las métricas propias para la problemática. También se busca ganar conocimiento sobre la diferencia entre modelos de aprendizaje supervisado y no supervisado, y sus usos. Un último objetivo es desarrollar habilidades de comunicación de la información obtenida a partir de los datos de manera clara y sencilla.

Método

A partir de lo estudiado en las clases teóricas y prácticas de las materias “Aprendizaje supervisado” y “Aprendizaje no supervisado”, y considerando el práctico realizado para la mentoría de la asignatura “Introducción al aprendizaje automático”, realizar un informe en formato de notebook o interactivo, en el cual se respondan, y justifiquen, las siguientes preguntas (además de cualquiera otra información extra que se considere de relevancia sobre la problemática):

- Emplear y calcular la performance de un modelo de aprendizaje supervisado de tipo support vector machine, separando para ello en dos clases: envíos rápidos (demoran menos de 3 días) y lentos (demoran 3 o más días). ¿Existen outliers? ¿Son los datos linealmente separables? Si no lo son, ¿qué kernel creen se podría emplear para realizar la proyección?
- En la problemática propuesta, podemos dividir la experiencia del usuario en tres situaciones diferentes: el envío llegó antes de lo prometido, el día prometido o después del día prometido. Definir una o más métricas que nos permitan determinar la performance de los modelos desarrollados considerando las tres posibles experiencias de usuario de interés.
- Determinar, desarrollar y calcular la performance (según las métricas definidas en el punto anterior) de un modelo de aprendizaje supervisado de tipo ensemble de modelos (considerando todas las clases posibles). ¿Cuál es la diferencia con los modelos más sencillos del práctico anterior? ¿Cómo es la performance en este caso? ¿Cómo creen que puede incrementarse aún más la performance utilizando este tipo de modelos?

- Elegir las dos features que considere más relevante y desarrollar un modelo de clustering (sin utilizar la información del target, i.e., *shipment_days*).
- Realizar una visualización de los diferentes clusters.
  ¿Cómo son los tamaños relativos?
  ¿Parecen adecuados para el problema de interés?
  ¿Representan a todas las clases?
  ¿Qué medida podemos utilizar para inspeccionar el contenido de los clusters? Aplicarla al problema y comentar el resultado obtenido.
  ¿Qué modificaciones deberíamos llevar a cabo para que los clusters sean más representativos del problema de interés?
- Repetir el análisis anterior empleando en el modelo de clustering un número mayor de clusters (que difiera del anterior en al menos 3). ¿Qué diferencias observa en ambos casos?
- Utilizando todos los features y el número de clusters que considere mejor para el problema de interés, desarrollar un modelo de clustering. Calcular accuracy y la performance según las métricas definidas en el segundo punto. ¿Cómo es la performance comparada a la del modelo de aprendizaje supervisado? Comente sobre cuál modelo es mejor y por qué.
- Emplear un embedding sobre las features seleccionadas (por ejemplo, PCA) y emplear el mismo modelo de clustering desarrollado en el punto anterior. ¿Cómo se modifican las métricas? ¿Y el tiempo de entrenamiento? Comente sobre las ventajas y/o desventajas de aplicar el embedding.

Esta comunicación debe estar dirigida para un público técnico pero que desconoce los aspectos propios del problema a resolver (por ejemplo, sus compañeros de clase). Se evaluará, principalmente, la claridad del mensaje presentado, el uso de las herramientas, los conceptos y los modelos desarrollados en las clases teóricas.

Además se debe realizar una breve comunicación en pdf (3 páginas máximo) dirigida a un stakeholder del proyecto (por ejemplo, manager), comentando los hallazgos y problemas encontrados, y las posibles acciones a tomar.

Estructura del informe

El informe debe contar con la estructura propia de un reporte de un experimento científico. Esto implica que debe tener un objetivo claro, una introducción a la problemática a resolver en dicho informe (no únicamente al problema general), una descripción de los datos a emplear, el desarrollo propiamente dicho del experimento y las conclusiones que se obtuvieron.

En el informe se deberá brindar una descripción del dataset suministrado (columnas, tipo de variables, valores extremos, etc.), las visualizaciones realizadas que sean pertinentes para la resolución del práctico, un análisis del modelo seleccionado, el análisis y las respuestas a las preguntas indicadas anteriormente, y las conclusiones.

Entrega

La entrega del informe final será antes del día 10 de Octubre. El notebook donde se realicen los cálculos y gráficos debe encontrarse en un repositorio al cual se pueda acceder.