Logística de envíos: ¿Cuándo llega?

Mentoría DiploDatos 2019

Práctico: Introducción al aprendizaje automático

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 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. Un tercer 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 la materia “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):

- En el práctico anterior se respondió al siguiente enunciado: “A la hora de determinar la promesa de entrega de un envío (fecha estimada de llegada), ¿cuáles son los features que consideran pueden tener mayor relevancia? ¿Cuál es el valor a predecir?”. Recupere esa respuesta y presente un breve resumen de los features que consideraron de mayor relevancia y el target seleccionado para predecir.
- El primer paso para desarrollar un modelo de aprendizaje automático es contar con datos limpios. ¿Qué pasos harían para limpiar el dataset?
- Es necesario poder separar el dataset en un conjunto de entrenamiento y en uno de test. ¿Cómo realizaría esta separación? ¿Qué tamaño emplearía para cada uno considerando que partimos de 500.000 datos?
- Dados los datos que disponemos y el target antes seleccionado, ¿qué tipo de modelo emplearían (regresión o clasificación)?
- Definir el modelo a utilizar, entrenar y evaluar el mismo utilizando los valores por defecto propios de la librería *scikit-learn*. Analizar los resultados obtenidos en el contexto de la problemática (por ejemplo, ¿por qué creen que para ciertos valores del target tiene mejor performance que para otros?).
- Modificar los hiperparámetros propios del modelo, y volver a entrenar y evaluar. ¿Por qué se eligió dicho valor para modificar? ¿Qué consecuencias tuvo? ¿Mejoró la performance del modelo? Analice los resultados obtenidos en el contexto de la problemática.
- En los puntos anteriores se seleccionó un modelo de regresión o bien uno de clasificación. Realice una prueba con un modelo del otro tipo y comente sobre las métricas y los resultados obtenidos. ¿Por qué tuvo mejor o peor perfomance?

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.

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 26 de Julio, con una muestra previa de avance el día 19 de Julio. El notebook donde se realicen los cálculos y gráficos debe encontrarse en un repositorio al cual se pueda acceder.