UNIVERSIDAD TORCUATO DI TELLA

Machine Learning
2023
Trabajo Práctico

Regresión en Retail

Problema

El *forecasting de demanda* es un problema muy recurrente en la industria retail. Poder estimar cuánto volumen se va a operar en los siguientes días constituye una información crucial para poder atacar un montón de otros problemas dentro de la operatoria de una empresa: manejo de compras, organización de depósitos, calendarización de personal y maquinarias, decisiones de marketing, etc.

En base a datos históricos de transacciones realizadas en un retailer, y sumando información contextual como variables sociales, económicas, de calendario, naturales; se puede obtener un gran poder predictivo sobre la demanda de ventas que va a tener un retailer en días venideros.

Datos

En este trabajo vamos a tratar con datos provenientes de un retailer alemán con presencia en diferentes países de europa.

Los datos fueron parte de una competencia de la plataforma Kaggle

(https://www.kaggle.com/competitions/rossmann-store-sales/data).

En dicha competencia el objetivo era predecir un único número correspondiente a las ventas de cada una de las *stores* de la empresa.

En el link correspondiente a los datos nos encontramos con 4 archivos:

- *train.csv*: Es un archivo tabular en donde encontramos filas correspondientes a información sobre a una store en un día particular. En cada una de estas filas tenemos varias columnas con información detallada. Algunas de ellas son:
  - Id, Store, Date: son para identificar de qué día y store es la información de la fila.
  - Sales: la variable dependiente a predecir.
  - Open, Promo, StateHoliday, etc: información particular que indica si una store estuvo abierta ese día, si estaba haciendo alguna promoción, etc.
  - Customers: la cantidad de clientes para dicho par (store, día)

- *test.csv*: Un archivo tabular similar al train.csv pero en donde no se tienen las columnas Sales y Customers.
- *sample_submission.csv*: Un archivo de ejemplo para enviar a la plataforma de competición. Consta simplemente de dos columnas, la primera es un *id* que identifica a un par (store, día). La segunda es un único número correspondiente a la columna Sales.
- *store.csv*: Es otro archivo tabular que contiene información adicional correspondiente a los stores. Es decir, este archivo es independiente de la fecha y cada fila corresponde a información útil sobre cada store en particular. Contiene información determinante de la store que indica el tipo de la misma y proporciona información sobre la competencia directa y las promociones que maneja.

Enunciado

En la competencia mencionada el objetivo final era optimizar la métrica RMSPE (Root Mean Square Percentage Error) sobre el conjunto de test.

En este trabajo práctico nuestro objetivo va a ser más amplio. Por un lado vamos a buscar optimizar la métrica mencionada pero haciendo especial énfasis en experimentar diferentes compenentes de *El camino de Machine Learning* y en poder interpretar nuestros modelos. Por otro lado, vamos a buscar plantear y resolver otro problema de aprendizaje automático diferente al planteado en la competencia.

En particular se pide:

- Explorar diferentes técnicas y modelos para optimizar la métrica de regresión propuesta en la competencia original teniendo en cuenta, como mínimo, los siguientes puntos:
  - Preprocesamiento: Realizar *Feature engineering* tanto por necesidad para los modelos elegidos como por conveniencia para optimizar el RMSPE. Considerar diferentes opciones como la extracción de información a partir de las columnas originales, la codificación de variables categóricas y/o la adición de nuevas columnas en base a datasets externos (por ejemplo de indicadores económicos, sociales o naturales).
  - Considerar al menos 2 modelos diferentes de Machine Learning.
  - Realizar una optimización de hiperparámetros robusta, utilizando técnicas de validación adecuadas (más allá de trabajar con el test.csv provisto por la competencia).
  - Finalizar el problema de regresión enviando las predicciones sobre el conjunto de test a Kaggle y reportando el score obtenido.
- Reportar elementos de interpretación de modelos para explicar diferentes facetas de la regresión. Queda a criterio del grupo elegir qué técnicas se utilizan, pudiendo ser algunas de las vistas en clase como así nuevas técnicas exploradas de manera autodidacta.
- En base al dataset original, definir un segundo problema de Machine Learning diferente al planteado en la competencia. Explicar coloquial y formalmente qué se quiere predecir, si se trata de un problema de clasificación o de regresión, cuál sería su contexto de uso, cómo es la métrica de medición, etc.

Luego del planteo del problema, se debe utilizar algún modelo visto en clase *out-of-the-box* para mostrar algunas primeras aproximaciones a la resolución del problema planteado.

Se debe entregar un informe que desarrolle los puntos propuestos junto con el código implementado. El informe no debe exceder las 10 páginas de extensión. El código debe contener comentarios y presentar una modularización adecuada para que sea entendible por alguien que no realizó la implementación.

Fechas de entrega

- *Formato Electrónico:* Domingo 16 de Julio de 2023, hasta las **23:59 hs**. La entrega se hace mediante el campus virtual de la materia en donde se puede encontrar una carpeta de entrega para subir todos los archivos correspondientes. Deben entregar una única versión por grupo de trabajo.