Machine Learning Engineer

Challenge

Cliente: ELECTRIC COMPANY LLC

Contexto: En un marketplace, para vender sus productos, las empresas o personas que desean vender deben crear una **publicación** en la que indiquen con todo el detalle posible el producto que desean vender. Además, hay casos en que el vendedor no puede crear una publicación propia para vender su producto, en cambio, debe ofrecer sus productos a través de **catálogos** predefinidos por el marketplace, un mecanismo que permite agrupar publicaciones que venden exactamente el mismo producto. Finalmente, el marketplace agrupa las publicaciones y los catálogos en **categorías**, según el tipo de producto que se esté vendiendo (Por ej: categoría Iluminación o categoría Ventilación, etc).
Luego, cuando un comprador realiza una compra y después de recibir el producto, el marketplace le realiza una breve encuesta en donde se le solicita al comprador que cuente su experiencia en la compra y haga una devolución calificando el producto que compró.
Estas devoluciones contienen información muy valiosa respecto a la reputación de una marca o de un producto en el mercado y por ello merecen cierta atención. En general cuentan de dos partes:

1. RATE (calificación numérica de satisfacción) y
2. COMENTARIO (texto en el cual el comprador expresa verbalmente su opinión.)

Requerimiento: Nuestro cliente "ELECTRIC COMPANY LLC" es el fabricante de los productos que se venden por distintos vendedores, en distintas publicaciones y desean conocer la reputación de su marca en el marketplace. Para dar una solución, se requiere construir un modelo que permita analizar cada comentario y determinar si el comentario habla bien o mal del producto, del precio, de la calidad, etc. Además se debe confeccionar gráficas o un informe que permita analizar los resultados.

Datos

Dataset: El set de datos disponible consiste en un listado de más de 1400 comentarios de 353 publicaciones en 32 categorías. Algunas de estas publicaciones pertenecen a un mismo catálogo y otras no. Además, algunas categorías poseen un volumen de publicaciones muy diferente a las otras. Entre las columnas se dispone de:

1. id_categoria: identificador único de la categoría.
2. id_catalogo: identificador único del catálogo.
3. id_publicacion: identificador único de la publicación.
4. id_comentario: identificador único del comentario.
5. marca: Marca del producto de la publicación.
6. rate: Puntaje de 1 a 5 otorgado por el comprador.
7. valorization: Apoyo de otros compradores al comentario.
8. title: Descripción breve de la conformidad del comprador.
9. content: Comentario realizado por el comprador.

Para una mejor interpretación, tener en cuenta que:

- Cada registro corresponde a un comentario (id_comentario) de un comprador.
- Una publicación (id_publicacion) puede tener N comentarios.
- Una categoría (id_categoria) contiene N catálogos (id_catalogo).
- Una categoría (id_categoria) contiene N publicaciones (id_publicacion).
- Un catálogo (id_catalogo) contiene N publicaciones (id_publicacion).
- Una publicación (id_publicacion) puede pertenecer a un catálogo (id_catalogo) o no.
- Una publicación no puede pertenecer a más de una categoría.
- Una publicación no puede pertenecer a más de un catálogo.

Cualquier duda o consulta relacionada al contexto, dataset o el requerimiento, escribir a: alvaro@nubimetrics.com

Evaluación

Análisis de los datos: se espera principalmente un análisis de los textos de los comentarios.

Abordaje: se espera una solución escalable, considerandos tiempos y recursos.

Features engineering: el dataset está limpio pero sin etiquetar (en caso que se elija un modelo supervisado).

Tipo de modelo elegido: se espera una solución simple, implementando algún clasificador de 3 clases (Positivo, Negativo, Nulo)

Entrenamiento: se espera que utilice herramientas de validación cruzada y búsqueda de hiper parámetros.

Performance: se espera que evalúe el rendimiento con las herramientas adecuadas a la naturaleza de los datos y del modelo con el uso de curvas ROC, AUC y/o métricas.