Proyecto Final

Juan David Rojas Gacha

1. Introducción

En vista de la situación que se presentó en el mundo a raíz de la pandemia, las ventas por internet han tenido un auge, en particular, en Colombia. Caso contrario a Brasil, donde las ventas por internet son normales y no son un gran tabú para su población. A pesar de esto se busca en este proyecto responder a preguntas relacionadas a la logística en la entrega de los productos comercializados por la red.

Queriendo indagar y trabajar sobre este tipo de situaciones en la plataforma de datos abiertos kaggle se encontró la base de datos: `https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_customers_dataset.csv`. Allí se encuentran 9 datasets diferentes que albergan información detallada de compra, más **NO** de información personal de los usuarios. Dada lo amplio que es el territorio brasileño el estudio se restringió a compradores en el estado de Rio de Janeiro y las 5 ciudades más grandes: Rio de Janeiro, Niteroi, Nova Iguaçu, São Gonçalo y Duque de Caxias.

Se decidió usar está información para responder a preguntas relativas a disposición logística de los envíos. Una de clasificación para los clientes: si se hace un pedido, *¿llegará antes de la fecha propuesta por el vendedor o llegará después?* esto teniendo en cuenta, que por motivos de fuerza mayor como fechas de compras excesivas (como fin de año) o por estar el vendedor en una ciudad con poco transito a Rio de Janeiro pueda tardar. Por otro lado, pero también pensando en los clientes o una posible caída de la plataforma de correos se sugiere responder: *¿cuál es el valor del flete (envío) basado en información de la compra y el producto?*

Por último, se hace una segmentación de los vendedores, basados en geolocalización, e historial de venta de productos regidos por las características de peso y volumen. Esto con el fin de encontrar posibles grupos de ciudades en donde la circulación de productos pesados sea mayor, o por el contrario de productos de tamaño y peso pequeño, todo esto pensando en mejorar la disposición de vehículos de transporte, maquinaria en los centros de acopio y disposiciones de espacio para albergar los mismos.

2. Métodos

Con el fin de responder a las preguntas de clasificación y regresión se hizo uso de las variables: año, mes y día de la semana de la compra, distancia entre comprador y vendedor, geolocalización del comprador y del vendedor, volumen, peso y precio del producto. Los modelos usados para la clasificación fueron: *regresión logística, LDA, QDA, Naive Bayes, vecinos más cercanos (kNN), Máquinas de Vectores de Soporte con kernel lineal, sigmoide y rbf, Clasificador Bagging con árboles de decisión, Clasificador Random Forest y Clasificador XGBoost*. Para los modelos que no son basados en

árboles se usaron componentes principales y para los demás la base de datos normal, sin estandarizar. Un detalle importante en este proceso de clasificación fue el hecho de que se debió hacer un UnderSampling de los datos, pues la proporción de entregas atrasadas era mucho menor que las a tiempo, un 12 %. Para la evaluación de cada modelo se hizo uso de `GridSearchCV` y se usó como medida de error el área bajo la curva ROC, esto último al percibir que el accuracy no era suficiente para diferenciar un buen desempeño de los modelos.

Modelando el valor del flete (problema de regresión), se estandarizaron los valores de entrenamiento y se usaron los modelos: *regresión lineal, Lasso, Ridge, Máquinas de Vectores de Soporte con kernel sigmoide y rbf, Regresor Bagging con árboles de decisión, Regresor Random Forest y Regresor XGBoost*. A diferencia de la clasificación en ningún modelo se uso PCA y en la regresión lineal se hizo un estudio estadístico para eliminar variables. La medida de error fue el valor R^2 y para evaluar la desviación promedio en el valor del flete, se uso el error medio cuadrático.

Y en la segmentación se usaron los modelos: *k-means, k-medoids y clustering jerárquico*, como medida de comparación y evaluación se usaron el score y la silueta.

3. Resultados

Tanto para la clasificación como la regresión el mejor modelo fue el *Random Forest*. Los intervalos de confianza con un 95 % fueron:

- Área bajo la curva ROC: $(0.7077, 0.7981)$ con valor promedio de 0.7515. En el caso de la matriz de confusión, tiene una efectividad de 87.5 % para acertar que llegará a tiempo, mientras que sólo 56.7 % para decir que no lo hará en el test dataset.
- Error medio cuadrático (valor del flete): $(5.1370, 8.0777)$ con valor promedio de 6.5333 reales, que son aproximadamente 4600 pesos colombianos, lo cuál no es un valor alto para tener de excedente en un flete. El R^2 en este caso fue de 0.76.

Para la segmentación, el mejor modelo fue *k-means* con valor de silueta más alto y usando 8-clusters. Se encontró que la segmentación categorizo regiones por su posición en el país y adicionalmente por las dimensiones de los productos comercializados, de hecho fue interesante encontrar que las categorías disponibles en cada segmentación correspondía con las variables de volumen y peso.

4. Conclusiones

Los modelos de machine learning, tanto supervisados como no supervisados generan respuestas alineadas con la realidad, siempre y cuando se escojan las variables adecuadas y se entienda el contexto del problema. Lo ocurrido en el problema de clasificación es una buena muestra de esto, pues allí fue necesario entender la desventaja al tener menos de una clase que de otra.

Sorprende la efectividad del modelo *Random Forest* quien terminó siendo el mejor para los dos problemas, seguido en ambos del *XGBoost*, estos resultan ser ensambles de varios modelos sencillos (árboles) y abre las puertas a explorar ensambles entre diferentes modelos, como por ejemplo las potentes *máquinas de vectores de soporte* con estos creados a partir de árboles y por qué no con una red neuronal.