Implementación de un modelo para crear clasificaciones según los perfiles de cliente usando Clustering

Presenta: Haile Jacobo Meneses Moreno

Introducción

Contamos con datos provenientes de una campaña de marketing directo de una institución bancaria de Portugal. Esta campaña está basada en llamadas telefónicas que tienen como objetivo que los clientes contraten un depósito a plazo fijo.

Estas llamadas con diversa duración y frecuencia obtenían como resultado en estos datos que el depósito es contratado o no (si/no), los resultados basados en frecuencia y duración obtienen diversos resultados según el momento de la temporada.

Objetivo del Proyecto

Aumentar la cantidad de clientes que contraten un depósito, estableciendo un instrumento que permita clasificar a los clientes y poder ayudar a crear una oferta de productos orientada en el perfil del cliente; buscando evitar un sesgo que limite la cantidad de clientes que puedan contratar servicios con la entidad.

Decisiones o procesos específicos que se quieren mejorar o automatizar con ML.

Con los datos que se tienen actualmente se pueden identificar diversos perfiles a los que pueden ofrecerse instrumentos que estén orientados a ellos, buscando no excluir a la mayor cantidad de clientes que puedan tener la oportunidad de contratar algún tipo de producto que la entidad pueda crear con los datos clasificados con este instrumento.

¿Se podría resolver el problema de manera no automatizada?

Si, aunque implica realizar una tarea compleja, con una cantidad de variables que puedan hacer difícil la identificación de los perfiles y que además puede ocupar un periodo de trabajo muy extenso.

Metodología Propuesta

Segmentar a los clientes para encontrar grupos homogéneos en los datos me hace decantarme por un algoritmo como K-means Clustering; el objetivo es agrupar a los clientes según su perfil para identificar a qué grupo pertenece a cada cliente, permitiendo a la entidad tomar la decisión de crear productos y estrategias orientados a cada perfil y así aumentar las contrataciones.

La idea básica es que el algoritmo intentará agrupar a los clientes en K grupos (clusters), donde cada cliente dentro de un grupo es similar entre sí y diferente de los otros grupos.

Después de aplicar K-Means, cada cliente pertenece a uno de los clusters, y podemos interpretar estos grupos y saber a qué tipo de perfil de cliente representa cada cluster.

Para evaluar el rendimiento de este algoritmo se tiene que utilizar una métrica distinta a los modelos supervisados, ya que en el clustering no se tienen etiquetas de verdad. Entre los que considero utilizar se encuentran los siguientes:

Inercia o Distorsión Intra-clúster (Inertia)

- Esta métrica mide la suma de las distancias cuadradas de los puntos al centroide más cercano en cada clúster.

- Un valor más bajo de inercia indica que los puntos están más cerca de sus centroides, lo que sugiere una mayor cohesión en los clústeres.

- Esta métrica es útil para identificar el número óptimo de clústeres usando la técnica del codo, es decir, buscando el punto en el gráfico donde una reducción adicional de clústeres no genera una disminución significativa en la inercia.

Silhouette Score o Índice de Silueta, que es una de las métricas más utilizadas para evaluar la calidad de un clustering. Mide qué tan cerca están los puntos dentro del mismo cluster y qué tan alejados están de los otros clusters. Toma un valor entre -1 y 1, donde:

- Valor cercano a 1: Los puntos están bien agrupados y claramente separados de los otros clusters.
- Valor cercano a 0: Los puntos están en el borde de un cluster o mal agrupados.
- Valor negativo: Los puntos probablemente están asignados al cluster incorrecto.

Adicionalmente, implementar múltiples métricas dará una visión más robusta de la calidad de la segmentación y garantizará que se puedan tomar mejores decisiones a la hora de aplicar las agrupaciones.

Una de las que pretendo utilizar es la Calinski-Harabasz Index, la cual mide la relación entre la dispersión dentro de los clusters y la dispersión entre los clusters.

Esta métrica es muy eficiente desde el punto de vista computacional y funciona bien para evaluar la calidad general del clustering. Permite medir qué tan "bueno" es el agrupamiento en términos de la dispersión dentro de los clusters y la separación entre ellos.

Como recurso adicional para evaluación, considero además la posibilidad de sumar el GAP Statistic, una métrica muy útil para determinar el número óptimo de clusters en un modelo de clustering como K-Means.

Dado que el objetivo es segmentar a los clientes bancarios, usar el Gap Statistic puede ayudar a determinar el número óptimo de clusters. Esto es crucial para asegurar que los grupos que se encuentren no son solo un producto del azar, sino que representan perfiles de clientes bien diferenciados.

Por ejemplo, podría usar el Gap Statistic para decidir si los clientes deben agruparse en 3, 4 o 5 segmentos, lo que influirá en las estrategias de personalización de productos bancarios.

Datos Disponibles

Dentro del conjunto de datos relacionados con el perfil financiero disponibles para trabajar con este algoritmo usaré previo al Análisis de Componentes Principales (PCA) los siguientes:

- Edad
- Estado Civil
- Trabajo
- Educación
- Balance de cuenta bancaria
- Crédito impagado (Sí/No)
- Hipoteca (Sí/No)
- Préstamos personales (Sí/No)

Métrica de Éxito del Proyecto

Independientemente de tener en consideración la métrica técnica del algoritmo, en este caso el Índice de Silueta, el verdadero éxito de la implementación de este algoritmo se mide por el impacto que los clusters tienen en el negocio.

En este caso, se busca aumentar la cantidad de clientes que contraten un depósito y ofrecer productos basados en perfiles adecuados. Una de las métricas de negocio que podría utilizar sería la Tasa de Conversión por Cluster:

Una vez que se segmente a los clientes en diferentes clusters, se puede medir la tasa de conversión dentro de cada grupo:

Tasa de conversión = Número de clientes que contratan un producto / Total de clientes en el cluster.

Un logro sería observar una tasa de conversión más alta después de implementar las recomendaciones basadas en los clusters, en comparación con las campañas anteriores.

Por ejemplo, si antes se tenía una tasa de conversión del 10% y después del clustering se aumenta al 15-20% en ciertos clusters, eso sería un claro indicativo de éxito.

Responsabilidades Eticas y Sociales

La principal meta de implementar un proyecto como el que se presenta, radica en evitar el sesgo y que la entidad se plantee ofrecer nuevos productos a más clientes que pueden ser diferenciados o excluidos si se implementa un modelo que los excluya por razones que tradicionalmente les puede hacer inviables a contratar un producto financiero (edad, estado civil, antecedentes bancarios, etc.).

Pero adicionalmente se tendrían en cuenta las siguientes consideraciones:

- Evitar el sesgo y la discriminación mediante la revisión de los datos y el monitoreo de los clusters generados.
- Ser transparentes sobre cómo se toman las decisiones y por qué se asignan productos a determinados grupos.
- Respetar la privacidad y los derechos de los clientes, manejando los datos de manera responsable y con su consentimiento.
- Fomentar la inclusión financiera en lugar de la exclusión, ofreciendo productos a un espectro más amplio de clientes.
- Garantizar la supervisión humana y la posibilidad de que los clientes apelen decisiones automatizadas.