PRIMERA ENTREGA DE PROYECTO DE APRENDIZAJE NO SUPERVISADO

Título: Arquetipos de clientes de banco Tibamoa

Presentado por:

Wendy Barreda
Harold Rodríguez
José Rivera
Carlos Andrés Rodríguez

1. Resumen:

Con este trabajo se busca responder a la pregunta con datos de un banco real, cuyo nombre se mantiene confidencial, de cómo se pueden agrupar con base en sus comportamientos y características, de manera que se les puedan diseñar productos o servicios acordes a ellos.

Hasta el momento se han desarrollado los siguientes pasos: primero se hizo un análisis de los datos disponibles en crudo que nos permitió entender que la mayoría de los datos son de variables categóricas, también nos pudimos dar cuenta que, aunque no son muchos registros, la calidad de datos es muy buena dado que no se tienen registros faltantes. Para esta entrega buscamos, en primer lugar, hacer una revisión más detallada de la bibliografía en cuanto a las situaciones en las que se han resuelto problemas similares por medio de las herramientas de aprendizaje no supervisado, en segundo lugar, una revisión más detallada de los datos y proponer una metodología de las que hemos visto en el curso para resolver este problema.

2. Introducción:

Uno de los participantes en el grupo trabaja en una empresa de asesoría de Customer Experience y, una de las actividades que desarrolla actualmente la empresa es la definición de arquetipos de clientes. Los arquetipos de clientes son agrupaciones de clientes de acuerdo con sus comportamientos y características que, permiten desarrollar a partir de esas características o comportamientos comunes, estrategias de desarrollo productos o servicios, estrategias de comunicación y canales más acordes con sus características.

Actualmente esta actividad se hace con información de entrada que es en su mayoría cualitativa y dependiendo la experiencia del consultor. Se piensa que al usar las técnicas de Aprendizaje No Supervisado, se podría hacer de una manera más precisa y con mayor soporte. Sería excelente poder replicar el método que se use en este caso para poderlo desarrollar en otros clientes bancarios o no bancarios, pero dado que se cuenta con datos del sector bancario, se haría en este caso con ellos.

El proceso actual de determinación de arquetipos es casi artesanal, se busca establecer cuáles variables cuantitativas sirven para agrupar los datos de manera que los polarice. Se buscan dos variables que agrupen los datos de la mejor manera, pero dada la naturaleza manual del proceso, nunca se usan más de 2 variables por la complejidad que podría traer agrupar de manera manual más conjuntos de clientes. Esperamos en este trabajo poder crear una metodología que pueda ser fácilmente replicable en otro tipo de clientes y, que nos permita con técnicas más precisas y

automáticas, hacer la determinación de arquetipos incluso en casos en los que las variables que expliquen el comportamiento de los datos sean más que dos, que es el límite de variables que se manejan actualmente.

3. Revisión preliminar de antecedentes en la literatura

Como ya se mencionó, el objetivo es agrupar a los clientes de un banco de acuerdo con sus características, pero esto no es algo nuevo. Actualmente existen diversos proyectos de carácter nacional e internacional que buscan alcanzar objetivos similares.

Un ejemplo de ello es el proyecto de "Segmentación de clientes con afectación en sus servicios Área Analytics TIGO" [1]. Un proyecto realizado por alumnos de la Universidad de Antioquia donde se busca segmentar a los clientes con afectaciones por los servicios de la compañía TIGO con la finalidad de atender las necesidades y aumentar la satisfacción de estos. En este proyecto se implementó un algoritmo no supervisado (K-Means), lo cual reafirma el hecho de que el problema de la segmentación de clientes de banco puede ser solucionado con algoritmos no supervisados.

Otro caso similar lo podemos encontrar en el articulo publicado en la IEEE "Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning" [2]. Este articulo se publicó como parte de la 3rd International Conference on Advances in Computing, donde los autores se enfocan en la segmentación de clientes para la venta de productos, buscan conocer al cliente para descubrir tendencias que les ayuden a dirigir los productos adecuados a los clientes adecuados para el aumento de ventas. Para poder clasificar a los clientes se utilizó información demográfica tal como sexo, edad, ingresos, etc., información geográfica y datos de comportamiento.

Como podemos ver, la segmentación de clientes es algo que puede sumar valor en diversas áreas por lo cual se ha convertido en algo común entre las industrias, y, por ende, han surgido diversas empresas que ofertan este tipo de servicios como lo es el caso de Brands by Rappi y su producto Behavioral Audiences [3], el cual ofrece la segmentación de los usuarios en 26 categorías generales y más de 280 subcategorías dependiendo de los datos. Otra compañía que ofrece soluciones de segmentación es Zendesk, una empresa de CRM dedicada especialmente a la atención al cliente. Zendesk en 2020 público un articulo "Segmentación de clientes: cómo llegar efectivamente a nuestro público objetivo" [4], donde habla de la importancia de la segmentación de clientes puntualizando que esto ayuda a mejorar el proceso de marketing, atraer y convertir leads, diferenciarse en el mercado, mejorar la experiencia del cliente, entre otras.

Por todo lo anterior podemos concluir que encontrar los arquetipos de los clientes de un banco es un proyecto viable y con sustento para poder realizar con algoritmos no supervisados.

4. Descripción detallada de los datos

Los datos son proporcionados en un archivo de texto plano extensión xlsx que lleva por nombre baseTibamoa. Las dimensiones del set de datos son:

- Número de registros: 215
- Número de variables: 46

Para realizar un análisis descriptivo se utilizó la librería pandas en Python para poder convertir el set en un dataframe; una vez creado el dataframe podemos observar las 46 variables y su tipo de dato

Descripción de variables recibidas:

Nombre de la variable	Tipo de dato
ID_Consutoriamigo	object
APERTURA_DT	object
EDAD	int64
GENERO	object
ESTRATO	int64
ACTIVIDAD_ECONOMICA	object
ESTADO_CIVIL	object
RANGO EDAD	object
BASE	object
FECHA_ENCUESTA	object
Probabilidad_Recomendacion	int64
Productos_Tibamoa	object
Frecuencia_uso_CA	object
Uso_ppal_CA	object
Principal_uso_Cred Consum	object
Principal_uso_TC_Tibapuntos	object
Identificacion_afirm_comunicacion	object
Identificacion_afirm_conocimiento	object
Identificacion_afirm_disp_compra	object
2_aspectos_mas_importantes_escoger_ entidad_financiera	object
Que_impotante_Satisfaccion_EF	object
Busqueda_info_producto	object
Solicitud_productos	object
Consulta_info_y_caracteristicas_prod	object
Transacciones (compras, pagos, abonos, etc.)	object
modo_Realizar_consulta_o_requerimiento	object
Radicar_queja_reclamo	object
Primer_canal_alternativo	object
Web B Tibamoa	object
App B Tibamoa	object
Banca en línea B Tibamoa	object
Chat WhatsApp B Tibamoa	object
Teléfono B Tibamoa	object
Cajeros electrónicos	object
Oficinas B Tibamoa	object
Correo electrónico B Tibamoa	object
Disposicion_recibir_comunic_Tibamoa	object
Información de beneficios	object
Información de ofertas comerciales	object
Notificaciones de estados de trámites	object

Sí Tibamoa excelente, estaría dispuesto a:	object
Si igual tiempo seguiría cliente	object
Otras entidades que lo atienden	object
Banco Principal	object
Billetera digital preferida	object
Que_valora de bco ppal	object

De lo anterior podemos observar que la mayoría de las variables son categóricas por lo cual se tendrá que aplicar algún tipo de transformación para incluir los datos en el proceso del entrenamiento del modelo

Análisis de datos faltantes o nulos:

También se analizan el número de registros faltantes (nulos o sin información).

Nombre de la variable	# de registros faltantes	% de registros faltantes
Información de beneficios	21	9.767442
Información de ofertas comerciales	21	9.767442
Notificaciones de estados de trámites	21	9.767442
Billetera digital preferida	211	98.139535

Al analizar las características de la base de datos, se concluye que los datos faltantes en realidad no son faltantes, sino que, la respuesta no aplica por lo que se pueden considerar vacíos y con el fin de conservar la completitud de la base, debido a que no se debe a ausencia de información, y por ende no causan sesgo, se pueden obviar las observaciones.

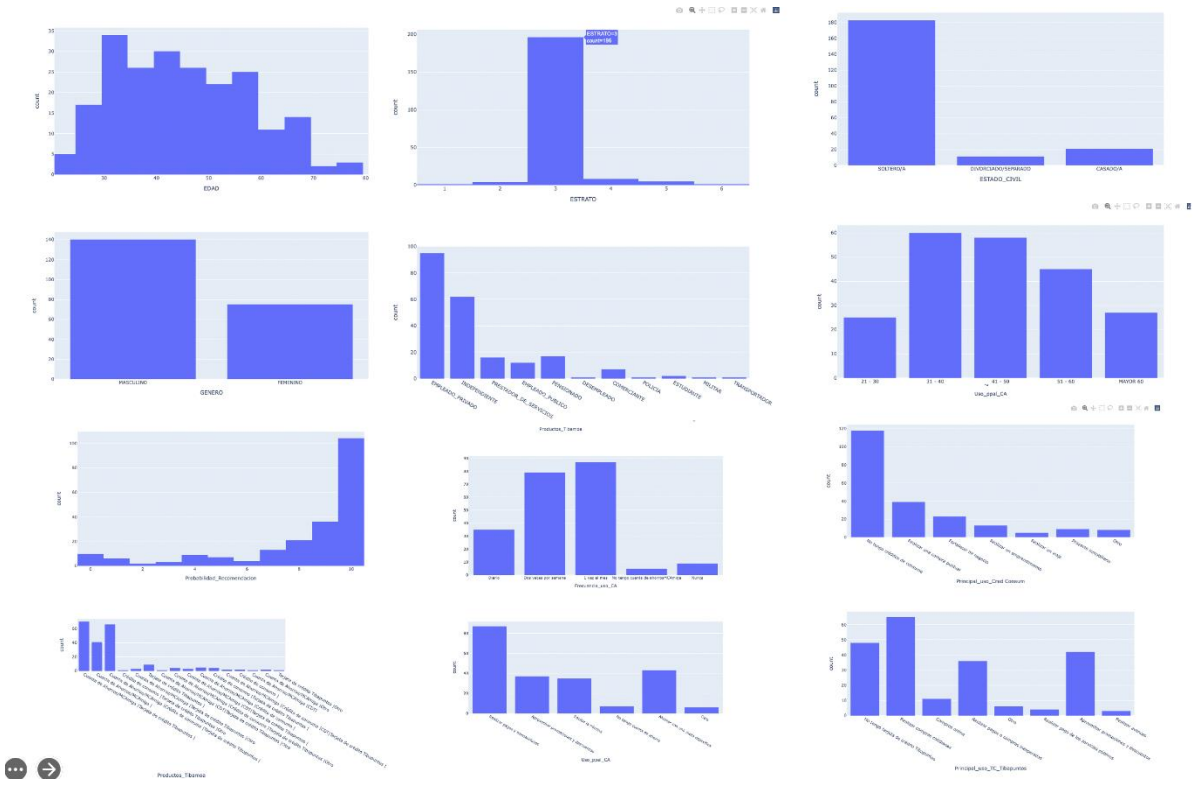
De la estadística descriptiva podemos resaltar lo siguiente:

- El promedio de edad es de 45 años
- Existen más registros del género masculino
- La actividad económica más frecuente es empleado privado
- El estado civil predominante es soltero
- El rango de edad con mayor presencia es de 31 a 40 años
- El promedio de estrato es el 3

Análisis gráfico:

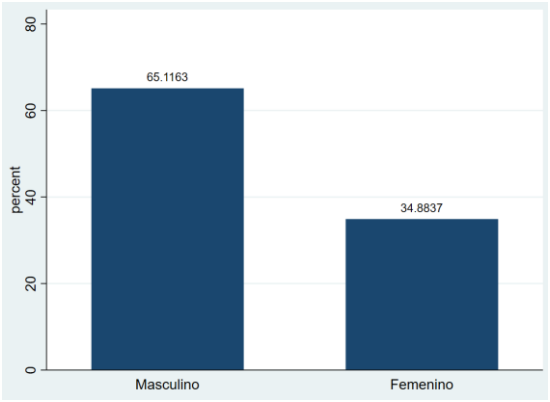
Se realizaron histogramas para cada una de las variables que nos sirvieron para familiarizarnos con las variables. Por razones prácticas, dado que hay una gran cantidad de variables, solo mostramos algunas de ellas:

Figura 1. Histogramas de algunas variables para arquetipo de cliente



Por ejemplo, frente a la variable género podemos observar que el 65% de los registros son masculinos y el 34% pertenece a la categoría femenina.

Figura 2. Histograma de género



Necesidades de transformación de datos:

Conversión de variables dummies: Dado que la mayoría de los datos son categóricos, se debe evaluar el mecanismo para transformarlos en variables dummies según sus características.

Es necesario prestar especial atención a las tres variables mencionadas en la tabla, a continuación, pues son variables generadas a partir de múltiples respuestas de los clientes. Por ejemplo, en la variable "Productos_Tibamoa" los clientes podrían responder que usan Tarjeta de Crédito, Cuenta de Ahorro, Cuenta Corriente, etc. Entonces, el tratamiento de estas variables debe repensarse al momento de trabajar con la observación que contiene varios datos. Una solución pensada es crear una columna por cada producto y generar una dummy que posea o no posea el producto, teniendo en cuenta que las opciones no son mutuamente excluyentes.

Nombre de la variable	Tipo de dato
Productos_Tibamoa	object
2_aspectos_mas_importantes_escoger_ entidad_financiera	object
Otras entidades que lo atienden	object

Recuperación de datos numéricos: las variables a continuación son calificaciones que dan los clientes a los servicios del Banco Tibamoa. Estas variables son categóricas que van de 1 a 10, pero existen dos "valores" que pueden tomar las variables: "No uso este canal" y "No conozco este canal". Se debe decidir el tratamiento para estas categorías, de manera que se mantenga la información numérica y a su vez no implique asignarle un valor al dato, como pasaría si se cambiaran estas categorías por un cero.

Nombre de la variable
Web B Tibamoa
App B Tibamoa
Banca en línea B Tibamoa
Chat WhatsApp B Tibamoa
Teléfono B Tibamoa
Cajeros electrónicos
Oficinas B Tibamoa
Correo electrónico B Tibamoa

5. Cronograma

		Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8
Primera entrega	**Responsables**				X				
Titulo del proyecto	Carlos Rodriguez/ Harold Rodriguez								
Resumen	Carlos Rodriguez/ Harold Rodriguez								
Introducción	Carlos Rodriguez/ Harold Rodriguez								
Revisión preliminar de antecedentes en la literatura	Wendy Barreda/ Jose Rivera								
Descripción detallada de los datos	Wendy Barreda/ Jose Rivera								
Propuesta metodológica	Wendy Barreda/ Jose Rivera								
Repositorio y README	Todos								
Entrega final								X	
Transformación de datos	Carlos Rodriguez/ Harold Rodriguez								
Reducir dimensiones de los datos	Carlos Rodriguez/ Harold Rodriguez								
Aplicar algoritmos de agrupamiento o clusters	Wendy Barreda/ Jose Rivera								
Comparar resultados y sacar conclusiones	Todos								
Documento del proyecto	Todos								
Video	Todos								
Repositorio GitHub	Todos								

6. Propuesta metodológica:

Para la reducción de dimensionalidad se utilizará el método de descomposición en valores singulares SVD dado su reconocimiento y eficiencia. Como la cantidad de registros es pequeña se puede proceder para el análisis de clusters con el método k-medoides ya que su complejidad será baja y no representará problemas computacionales, ofreciendo un mejor resultado ante la presencia

de valores atípicos. Adicional se aplicarán algoritmos de clustering jerárquico para comparar los resultados y poder sacar mejores conclusiones y un análisis más completo del agrupamiento que permita tomar las mejores decisiones en la organización.

7. Bibliografía

[1] S. López Stan, “Segmentación de clientes con indisponibilidad o afectación en sus servicios, Área Data Analytics TIGO”, trabajo de grado profesional, Ingeniería de Sistemas, pregrado, Universidad de Antioquia, Ciudad Universitaria, 2022. Recuperado 1 de septiembre de 2022, de https://bibliotecadigital.udea.edu.co

[2] M. F. Alam, R. Singh and S. Katiyar, "Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning," *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 94-98, Recuperado 1 de septiembre de 2022, de https://ieeexplore.ieee.org

[3] Brands by Rappi. (2022). For Agencies. Recuperado 1 de septiembre de 2022, de https://brands.rappi.com/for-agencies

[4] Douglas Da Silva, D. S. (2020, 25 septiembre). Segmentación de clientes: cómo llegar efectivamente a nuestro público objetivo. Blog de Zendesk. Recuperado 1 de septiembre de 2022, de https://www.zendesk.com.mx/blog/segmentacion-de-clientes/

8. Link a repositorio Github

https://github.com/WendyBarreda/MIAD_ANS_Project