Laboratorio 2 - Clustering

Objetivos

- Comprender el proceso asociado con una tarea de semgentación, en especial a nivel de la preparación de datos e interpretación del resultado.
- Aplicar tres algoritmos de clústering como k-means y dos de libre elección para resolver el objetivo de la organización.
- Analizar e identificar los hiperparametros adecuados para los modelos de clústering.
- Obtener conclusiones a partir de los segmentos identificados que sean útiles para una organización.
- Comparar los 3 algoritmos seleccionados y explicar cuál recomiendan a la organización.

Herramientas

- Librerías principales de Python para procesamiento y visualización de datos como: pandas, sklearn, seaborn, numpy y matplotlib.
  Se recomienda usar la última distribución disponible de Anaconda Individual Edition, pueden encontrar el instalador en este enlace.
- Ambiente de desarrollo: JupyterLabs en distribución de Anaconda.

Enunciado

Descripción de negocio

BancAlpes es una entidad bancaria que está realizando una campaña de fidelización para aumentar la retención de clientes. En concreto, con su estrategía busca fidelizar especialmente a las personas que poseen tarjetas de crédito. Esta campaña busca ofrecer los mejores productos, servicios y recomendaciones a sus clientes de acuerdo con sus características. Por esta razón, ha recurrido a ustedes como consultores para que le entreguen al equipo de marketing información que pueda ayudarlos a orientar mejor sus campañas. En particular, el equipo de marketing espera que le provean grupos a los cuales pueden dirigir sus campañas y las caracteristicas de los clientes en dichos grupos. BancAlpes considera que esta nueva estrategia de marketing le permitirá aumentar el consumo de productos y servicios y por supuesto, aumentar la retención de clientes.

Los datos proporcionados al equipo contienen el resumen de los clientes del último año que poseen tarjetas de crédito, en particular, su información y movimientos financieros. Pueden acceder a los datos y al diccionario de datos a través de los siguientes enlaces: BancAlpes y diccionario.

Instrucciones

La empresa quiere seguir apropiando la metodología Crisp-DM, en específico en su nueva versión ASUM-DM, para el desarrollo de estas iniciativas en analítica de datos, por lo cual le sugiere realizar los siguientes pasos:

1. **Perfilamiento de los datos:** en esta etapa es importante saber cuántos datos se tienen (filas y columnas), el tipo de datos de las columnas, cual es la integridad de los datos, cuál es su distribución (discreta o continua). Para esto, es útil aplicar estadística descriptiva sobre los datos, señalando sus principales estadísticos: media, varianza, desviación estándar, etc., para el caso de las columnas numéricas. Enc aso de datos categóricos recuerde que es importante conocer las categorías, los números de registro por categoría, en especial para las categorías con mayor representación en los datos. Recuerde incluir en esta etapa el análisis a nivel de calidad de datos.

2. **Preparación de datos:** es el procedimiento llevado a cabo para transformar los valores actuales de acuerdo con los algoritmos a utilizar y el objetivo de negocio a resolver. Por ejemplo, manejar los datos nulos (missing values) o los valores atípicos (outliers) .

3. **Modelamiento:** en este paso se lleva a cabo la elección del modelo con el que queremos cumplir nuestra tarea y su refinamiento. En este caso usando el algoritmo de K-Means como base y deberán, realizar y compararlo con otros dos algoritmos como clustering jerárquico, DBScan, HDBScan, Gaussian Mixture. No obstante, en ambientes profesionales la elección del algoritmo, hará parte de su tarea de consultoría. Adicionalmente, se espera que sepan interpretar las funciones implementadas (JupyterLab), seleccionar los hiper-parámetros del modelo y justificar su elección. La sugerencia es explorar la generación de clústeres con distintos atributos que puedan llevar a mejores valores de coeficiente de silueta al igual que a mejor comprensión de los grupos por parte de la organización.

4. **Validación:** En modelos de aprendizaje no supervisado la validación de los modelos es un reto importante que deben asumir los consultores. En este caso, con el uso de clustering y en particular del algoritmo de K-Means, dado que se trata de un algoritmo particional, la calidad desde el punto de vista cuantitativo, relacionada con los datos utilizados y con el valor de “k” seleccionado, puede validarse utilizando el coeficiente de silueta y ser ajustado siguiendo guías como el método del codo. Información adicional para complementar su interpretación puede encontrarla en este enlace. Una vez realizada la validación cuantitativa, se tiene una segunda parte de la validación que corresponde a la descripción de los resultados obtenidos (clústers y conclusiones del proceso), para la comprensión por parte de la empresa. Esta parte la denominados validación cualitativa.

5. **Visualización:** El estudiante debe proponer una visualización de los resultados obtenidos en un tablero de control, de tal manera que el cliente tenga la capacidad de entender los resultados obtenidos en su labor de consultoría.

Entregables

- Código desarrollado para consultarlo en JupyterLab o una herramienta equivalente.
- Informe del laboratorio con el desarrollo y la evidencia de cada una de las etapas, es importante que incluyan imágenes de gráficas y su interpretación. Se espera que el informe no supere las 8 páginas. Además, una presentación corta para el negocio con los resultados, las conclusiones y algunas recomendaciones del proceso realizado.

Nota: Recuerde que la presentación debe estar orientada al área de marketing, por lo que se recomienda evitar el uso de términos muy técnicos y utilizar un lenguaje con el que el área esté familiarizada. Se les sugiere centrarse en la interpretación de los resultados.

- Realice su análisis siguiendo los pasos estipulados previamente en la sección de instrucciones, interprete los resultados obtenidos explícitamente en el informe del laboratorio, así mismo registre sus

conclusiones. Al final de su informe, haga una comparación de los algoritmos implementados para identificar sus ventajas y desventajas. Recuerde indicar el miembro del grupo que trabajó cada algoritmo.

Algunas preguntas que pueden guiar su desarrollo son:

1. ¿Qué criterios son importantes para la selección del modelo?
2. ¿Cómo medir la calidad del modelo construido? ¿Cómo saber que el modelo construido tiene una buena calidad?
3. ¿Qué retos tienen estos modelos no supervisados, si se quisieran aplicar a nivel profesional?
4. ¿Cómo varía la calidad del modelo obtenido si aplico diferentes algoritmos?
5. ¿Qué pasa en la ejecución de los algortimos si se incluyen variables categóricas nominales. ¿Qué tratamiento se les debe aplicar?
6. ¿Qué variables afectan negativamente el score de la silueta? ¿Són realmente necesarias esas variables para determinar los segmentos en las campañas de fidelización?

Instrucciones de Entrega

- El laboratorio se entrega en grupos de máximo 3 estudiantes. Estos estudiantes pueden ser de diferentes secciones.
- Recuerde hacer la entrega por la sección unificada en Bloque Neón, antes del domingo 6 de marzo a las 22:00.
  Ese será el único medio por el cual se recibirán entregas.

Rúbrica de Calificación

A continuación se encuentra la rúbrica de calificación.

Nota: Los siguientes porcentajes hacen referencia a la nota grupal, que corresponde a un 80% de la nota inidiviudal.
El 20% restante se calcula según el puntaje obtenido en la implementación del algoritmo del cual el estudiante estuvo a cargo dentro del grupo.

Concepto	Porcentaje
Descripción y análisis del perfilamiento de los datos y de las tareas sugeridas de transformación	15%
Descripción del preprocesamiento realizado, según el algoritmo utilizado	10%
Implementación de K-means, descripción de las decisiones más importantes asociadas a la implementación del algoritmo y los hiperparámetros configurados	6%
Implementación de un segundo algoritmo, descripción de las decisiones más importantes asociadas a la implementación del algoritmo y los hiperparámetros configurados	10%

Concepto	Porcentaje
Implementación de un tercer algoritmo de libre elección, descripción de las decisiones más importantes asociadas a la implementación del algoritmo y los hiperparametros configurados	10%
Análisis de los resultados obtenidos y justificación del modelo recomendado para el caso propuesto	24%
Presentación para BancAlpes con resultados, recomendaciones dadas a la empresa y visualización	20%
Notebook asociado	5%

Sugerencias

Los siguientes enlaces pueden serle de utilidad para la implementación en Python.

- Ejemplo de K-Means usando Sklearn
- Artículo de Towards Data Science: K-means with Sklearn
- Artículo de Towards Data Science: Introducción al análisis de datos con Python