2023 - 2024

ACTIVIDAD 2

Máster en Big Data y Data Science

05MBID – Minería de datos

Nombre: Sergio German Quispe

Fecha: 27/12/2023

viu Universidad Internacional de Valencia

TABLA DE CONTENIDO

1. PLANTEAMIENTO DE PROBLEMA Y OBJETIVO A ALCANZAR

1.1 CONTEXTO:

Una empresa de automóviles tiene planes de ingresar a nuevos mercados con sus productos existentes. Después de una intensa investigación de mercado, dedujeron que el comportamiento del nuevo mercado es similar al del mercado existente. En su mercado actual, el equipo de ventas ha clasificado a todos los clientes en 4 segmentos (A, B, C, D). La categoría A es quien tiene la mejor categoría económica y D la no tan mejor.

1.2 SITUACIÓN PROBLEMÁTICA

La empresa enfrenta el desafío de requerimiento de un algoritmo que permita clasificar o segmentar los clientes de acuerdo con la categoría (A, B, C, D) al nuevo mercado. A pesar de las similitudes generales en el comportamiento del mercado, podría haber sutiles diferencias que podrían afectar la precisión del modelo de segmentación al aplicarlo al nuevo conjunto de clientes.

Adicional, necesitan la categorización de los nuevos clientes para poder enviarles ofertas promocionales de acuerdo con su categoría económica (A, B, C, D), así poder ofrecer productos premium a los que tienen una mejor categoría y generar mayores ingresos.

1.3 OBJETIVOS

Objetivo General:

A) Implementar un algoritmo de clasificación que prediga con precisión el segmento al que pertenece un nuevo cliente en el mercado objetivo, usando la metodología KDD Process

Objetivos Específicos:

B) Identificar las características claves que son determinantes en la clasificación de clientes

C) Identificar y evaluar posibles riesgos asociados con la implementación del modelo

D) Determinar si la cantidad de datos a usar es suficiente o es requerido más datos para lograr una buena precisión

E) Apoyar al área de ventas desarrollando el modelo para que puedan identificar oportunidades de ventas adicionales, como upselling (ofrecer productos premium) y cross-selling (ofrecer productos complementarios), para maximizar los ingresos por cliente de acuerdo con su categoría.

2. ELECCION DE LA BASE DE DATOS Y SELECCIÓN DE DATOS UTILES

2.1 ORIGEN DE LOS DATOS

Los datos son obtenidos de la plataforma Kaggle[1], el dataset es de una empresa de automóviles anónima donde los datos son atributos o características de sus clientes, no hay información sensible ya que se maneja el ID. El dataset se llama **Customer Segmentation Classification**, se cuenta con una data para train y una data para test, ambos en formato csv.

2.2 TAMAÑO Y CARACTERÍSTICAS DE LOS DATOS DE ORIGEN

Los datos de train constan con un total de 8068 datos y los datos de test con un total de 2627 datos. No es una data pesada por lo cual no se requiere una buena máquina con buen procesador o Ram, es data promedio.

El dataset cuenta con los siguientes atributos:

A) Datos categóricos:

- **Gender:** Género del cliente
- **Ever_Married:** Estatus mariscal (si es casado o no)
- **Graduated:** Si el cliente es graduado o no
- **Profession:** Profesión del cliente
- **Spending_Score:** Puntaje de gasto del cliente
- **Var_1:** Categoría anonimizada del cliente (categoría 7 es más alto hasta categoría 1)

B) Datos numéricos:

- **ID:** Identificador del cliente
- **Age:** Edad del cliente
- **Work_Experience:** Años de experiencia del cliente
- **Family_size:** Tamaño de familia del cliente (1 indica vive solo)

Segmentation: Etiqueta o variable objetivo de tipo categórico (Clase A tiene mejor nivel económico)

2.3 SELECCIÓN DE DATOS

Teniendo en cuenta las variables se plantea a usar todas al inicio, el ID se elimina después de analizar datos duplicados, ya que el ID es un campo que garantiza los datos únicos del cliente. Por lo cual se comenzaría a analizar los datos teniendo en cuenta 6 datos categóricos, 4 datos numéricos y 1 campo más que sería la etiqueta **Segmentation**.

[1] Kaggle: Plataforma online que ofrece datasets, competencias de ciencia de datos y cursos afines

Se realiza gráficos usando la librería matplotlib y seaborn para conocer un poco más acerca de los datos que se cuentan, también se analiza la variable objetivo 'Segmentation' y se evidencia que no está desbalanceada.

3. PREPROCESAMIENTO Y TRANSFORMACION DE DATOS

3.1 DETECCION Y TRATAMIENTO DE VALORES NULOS

En este apartado se busca detectar si hay datos nulos en el dataset. Al visualizar ello tengo este escenario:

Escenario: Hay datos nulos en 6 columnas, pero no tan moderado, optamos por imputar:

En el caso de datos categóricos: Imputamos por la moda las variables categóricas 'Ever_Married', 'Graduated', 'Profession' y 'Var_1', teniendo en cuenta que no hay demasiados datos nulos, ya que si es considerable y reemplazamos por el valor más repetitivo no sería una buena variable.

En el caso de los datos numéricos 'Work_Experience' y 'Family_Size', se opta imputar por la mediana ya que es menos sensible a outliers y proporciona una estimación más robusta del centro de la distribución.

Instrumentos: Función isna().sum() de la librería Pandas para sumar datos nulos

Instrumentos: Librería sckikit-learn que facilita la imputación de datos faltantes con diferentes estrategias, incluyendo 'most_frequent' (para la moda) y 'median' (para la mediana)

3.2 DETECCION Y TRATAMIENTO DE DATOS DUPLICADOS

En este apartado analizamos los valores duplicados, para ello contamos con el ID de cliente que nos garantiza la unicidad de los datos. En este caso al analizar los datos, no se encuentra datos duplicados.

Instrumentos: duplicated() de Pandas para identificar datos duplicados

Una vez hemos pasado esta etapa, se elimina el campo ID ya que se considera que no aporta más al modelo.

3.3 ENCODEAR VARIABLES CATEGORICAS

En el dataset se cuenta con las siguientes variables catégoricas: Gender, Ever_Married, Graduated, Profession, Spending_Score. Se encodea debido a que hay algoritmos de clasificación como la regresión logística que es un algoritmo que generalmente requiere que las variables categóricas sean codificadas para su entrada.

Instrumentos: LabelEncoder de sklearn.preprocessing

3.4 ANALISIS UNIVARIANTE

Se realiza el análisis univariante para conocer la distribución de cada variable que es mediante un histograma. En este caso las variables 'Age', 'Work_Experience' y 'Family_Size' no siguen una distribución normal, por lo cual, la regresión logística puede ser una opción sólida para clasificar la categoría del cliente, incluso si los datos no siguen una distribución normal.

Instrumentos: Librería matplotlib para crear histogramas por variable

3.5 DETECCION Y TRATAMIENTO DE OUTLIERS

Para la detección de outliers se puede optar por hacer un diagrama de caja (boxplot), y adicional se complementa con el análisis univariante para analizar si en las variables se cuenta con outliers o no.

En este caso, no se ha detectado outliers en el dataset, solo en la variable 'Age' se realiza un análisis a detalle para analizar si se considera outliers a los clientes alejados del punto de concentración. En este contexto, tratar a los clientes con edad avanzada como outliers podría no ser apropiado, ya que representan una parte sustancial del conjunto de datos, ya que se visualiza que clientes de 78 años han comprado 29 automóviles, de 80 años han comprado 25 y 85 años han comprado 22.

Instrumentos: Seaborn para hacer el boxplot

3.6 ANALISIS DE CORRELACION

En este punto, debemos detectar la correlación entre variables ya que debemos evitar a toda costa variables con correlación alta, porque significa que están introduciendo problemas al modelo que luego entrenaremos para resolver el problema.

Una correlación débil indica que las variables no están linealmente relacionadas entre sí. Esto puede ser beneficioso ya que se está construyendo un modelo de regresión logística, la independencia de las variables predictoras puede ayudar a evitar problemas de multicolinealidad. En este caso se encontró correlación débil entre las variables, a excepción de 2 variables que tuvieron correlación moderada.

Instrumentos: Matplotlib y seaborn para hacer la matriz de correlación

3.7 NORMALIZACION DE DATOS

La regresión logística no es tan sensible a las diferencias de escala como algunos otros algoritmos, como la regresión lineal. Sin embargo, la normalización puede facilitar la interpretación de los coeficientes y ayudar en la convergencia del algoritmo.

Instrumentos: MinMaxScaler

4. MODELO DE MINERIA DE DATOS

Después de haber hecho un análisis y haber PreProcesado los datos se plantea usar la Regresión Logística Multinomial para predecir la categoría económica del cliente (4 clases)

4.1 ADAPTACIONES EN LOS DATOS

Imputación de Datos Nulos: Los datos nulos en las variables categóricas se han imputado utilizando la moda, y en las variables numéricas con la mediana. Esto asegura que no haya valores faltantes que puedan afectar el rendimiento del modelo.

Codificación de Variables Categóricas: Las variables categóricas se han encodeado para que el modelo pueda trabajar con ellas de manera efectiva.

Normalización de la Variable Edad: La variable 'Age' se ha normalizado para asegurar que las características estén en la misma escala, lo cual es importante para la regresión logística.

Adiciona al ello, no se encontraron datos duplicados ni outliers.

4.2 ENTRENAMIENTO DEL MODELO

Se utilizará la implementación de regresión logística multinomial, especificando el parámetro multi_class='multinomial' en la librería de aprendizaje automático, como scikit-learn en Python.

Se pueden ajustar otros hiperparámetros según sea necesario, como la regularización (C), el solver (por ejemplo, 'lbfgs'), y el número máximo de iteraciones (max_iter). El entrenamiento por búsqueda de los mejores hiperparámetros es una buena opción.

Adicional a ello, ya se contaba con una data para train y una para test previamente.

Instrumentos: LogisticRegression

5. DISCUSION E INTERPRETACION DE LOS RESULTADOS

5.1 EVALUACION DEL MODELO

Para evaluar el modelo se plantea usar la matriz de confusión ya que es una clasificación en este caso de 4 clases. Se evaluaría el F1-Score y el accuracy.

Instrumentos: confusion_matrix

5.2 INTERPRETACION DE RESULTADOS

En este apartado se plantea evaluar los resultados y teniendo en cuenta los objetivos planteados.

1. Identificar si los inputs o atributos que se contaban son determinantes para la predicción, solo se eliminó el ID y los demás campos no se quitaron por escasez y había una influencia con el target, aunque no tan destacable la influencia.

2. Evaluar los posibles riesgos asociados con la implementación del modelo:

- ***Riesgo***: El modelo podría estar sesgado hacia ciertas categorías económicas, lo que podría resultar en decisiones discriminatorias. Se plantea realizar un análisis de equidad para evaluar si el modelo muestra sesgo en función de características sensibles

- **Riesgo**: El modelo puede no tener el rendimiento esperado en la práctica.
  Evaluación: Examinar métricas de evaluación como accuracy y F1-score.

- **Riesgo**: El modelo podría no generalizar bien a nuevos datos fuera del conjunto de entrenamiento.
- Evaluación: Evaluar el rendimiento del modelo en un conjunto de prueba separado y utiliza técnicas como la validación cruzada para garantizar que el modelo sea robusto y generalice bien a datos no vistos.

3. Como se ha usado Regresión Logística y funciona bien con la cantidad de datos utilizados, no hay problema con ello, en caso se use redes neuronales si se necesitaría más datos para evitar el overfitting.

4. El objetivo de apoyar al área de ventas con el modelo aún no se ha evaluado, ya que se requiere mejorar la precisión del modelo y tenerlo a un buen nivel antes de llevarlo a producción.

   Posterior a ello, cuando el modelo esté listo sería de buena ayuda para que puedan identificar oportunidades de ventas adicionales, como upselling (ofrecer productos premium) y cross-selling (ofrecer productos complementarios), para maximizar los ingresos por cliente de acuerdo con su categoría.