Modelo de predicción de gastos mensuales

Daniela Márquez Campos, Karla Andrea Palma Villanueva, Adrian Pineda Sánchez, Kevin Antonio González Díaz

Resumen
El objetivo principal de esta propuesta es lograr mejorar la salud financiera de los clientes a través de la predicción de sus gastos por mes. La predicción de gastos futuros permite a los clientes planificar sus finanzas de manera más efectiva, al tomar decisiones financieras informadas y ajustar los presupuestos en consecuencia. Para realizar la predicción de gastos utilizamos el modelo RandomForestRegressor de la librería de Python sklearn, en el que obtuvimos una precisión de 0.96, lo que nos indica que las predicciones son muy cercanas a los valores reales. El objetivo es mostrar al cliente el monto total de gastos que se preveé que el cliente tendrá el próxmo mes, además del despliegue de una gráfica donde muestre las predicciones de sus mayores gastos de acuerdo a la naturaleza de la compra.

Palabras clave
Predicción — Seguridad financiera — Regresión Lineal

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Índice

1. Introducción

1.1 Introducción

El estrés financiero es un problema fuerte en los clientes puesto que se producen como consecuencia de los problemas económicos que ocasionan una sensabilidad de vulnerabilidad en la que se teme la escasez; por lo cual resulta inprescindible crear una alternativa cuya propuesta genere la prevención de dicha problemática a la vez que se fomenta el crecimiento económico del cliente.

Según os expertos se tiene claro que la salud financiera de los clientes debería ser un objetivo prioritario (y no sólo del sector financiero) en el futuro. Ayudar a los clientes a mejorarla reduciendo su estrés y contribuir a alcanzar su bienestar financiero, se presenta como una oportunidad para fortalecer la confianza de los clientes en las entidades como asesoras.[1]

Al fomentar la salud financiera en los clientes se fomenta una relación beneficiosa entre las empresas que proveen servicios financieros, puesto que procuras a los usuarios al tener como prioridad estratégica su bienestar económico, a la vez que se hace uso eficiente de las herramientas tecnológicas.

En este camino, es necesario también contar con un buen nivel de educación financiera que permita hacer frente a los retos económicos que se presenten y que se convierta en una herramienta indispensable para tomar las decisiones más adecuadas. Una mayor inclusión financiera aumentará el bienestar y el crecimiento económico sostenido y sostenible de los países. [1]

1.2 Definición del problema

Como se mencionó anteriormente, es un gran problema el que experimentan las personas al no regular sus finanzas por no contar con un amplio conocimiento en su educación financiera. Muchas veces nosotros mismos no sabemos a profundidad en qué sector llegamos a invertir más, o el impacto que nuestros gastos llegan a tener en nuestra economía; impidiendo as+i que se tenga un autocuidado de nuestra salud financiera. La solución a esta problemática es presentar un modelo de predicción en el cual se toma en cuenta el historial económico de los clientes por mes, para poder realizar una estimación de los gastos que pudiera llegar a tener el usuario el mes posterior.

Primeramente, se deben considerar muchos factores que puedan ayudar a la predicción del modelo, tomando en cuenta el uso de la tarjeta del usuario para poder modelar las inversiones. Por otro lado, un reto que se presenta es tener que limpiar el dataset asignado de modo que se logre erradicar cualquier celda vacía que perjudique la precisión de nuestro modelo.

Asimismo, debemos elegir qué modelo y método es el que mejor se adapta para resolver el problema. Inclusive, será útil probar con varios modelos para comparar los resultados y

finalmente elegir el que obtenga los mejores.

2. Modelación del problema

2.0.1 Identificación de las variables y posibles parámetros a tomar en cuenta para la solución del problema

En el dataset recabado existían algunas celdas sin información alguna, por lo cual la precisión de nuestro modelo resultaría menor. Por lo cual, optamos por realizar una nueva base de datos en la que se llenara esta información con ayuda de la media existía entre los gastos que sí fueron realizados.

Posteriormente se regeneró una nueva base de datos en la cual se dividieron los gastos de cada cliente al mes y ya procurando que no existiera alguna celda vacía. Ahora bien, los gastos de igual forma se categorizaron de acuerdo al giro, en los cuales haciendo los *clusterings* correspondientes nos dio un total de 28 variables que se usarán para nuestro modelo de predicción.

En los datos correpondientes, se puede notar que en cada cliente se determina su id, el tipo de pago ya sea débito o crédito, lugar de origen, tipo de comercio, giro de gasto, edad, fecha. Esta información se optó por clasificarla en meses dependiendo del uso de la tarjeta que le daba el cliente.

Uno de nuestros objetivos principales de nuestro modelo es ayudar al cliente a determinar en qué giro o tipo de gasto a invertido más dinero, por consiguiente, optamos por realizar una conversión de las variables categóricas encontradas en el apartado de giro a numéricas, de tal forma que se realizara una estimación más exacta del porcentaje desembolsado por parte del usuario.

El modelo que utilizamos RandomForestRegressor es muy útil en estos casos, en donde de forma automática el algoritmo decide qué variables son relevantes y les da mayor peso, por lo que no fue un aspecto que requiriera un análisis exhaustivo. Pero definitivamente las variables que debían estar de forma inmediata eran los montos totales de por meses de acuerdo al tipo de la compra. También queríamos darle un peso a la categoría de la transacción, la del giro nombre, por ejemplo, pero para utilizar el modelo RandomForestRegressor, necesitábamos reorganizar nuestros datos para evitar posibles errores en el aprendizaje del algoritmo.

La columna giro nombre, la tuvimos que convertir a un tipo numérico, utilizando labelEncoder que es una técnica usada en el procesamiento de datos y en el aprendizaje automático para convertir etiquetas de texto o categorías en valores numéricos para que puedan ser utilizados como entrada en modelos de aprendizaje automático. Pero el problema fue que teníamos datos desde 1,2, 3... 28, los cuales no representaban una cantidad, sólo representaban una clase, por lo que decidimos pasarlos como columnas de clases binarias usando one hot encoding, para poder aplicar el algoritmo. La edad la claificamos en 4 categorías porque consideramos que la edad es un factor clave para determinar la predicción de un gasto futuro, y los agrupamos de acuerdo a la cantidad y tipos de datos que se tenían, desde 1 a 123. Para el país consideramos tres categorías: México, Estados Unidos y Otros países, en donde de igual forma usando one hot encoding creamos las 3 columnas. Y en total nos quedaron las siguientes variables predictorias:

- $edad_0$ Rango de edad de 0-20 años.
- $edad_1$: Rango de edad de 20-10.
- $edad_2$: Rango de edad de 30-40 años.
- $edad_3$: Rango de edad de 40-60 años
- $edad_4$: Rango de edad de 50-60 años.
- $edad_5$: Rango de edad de 60 años o más.
- $pais_{mx}$ País de origen: México.
- $pais_{us}$ País de origen: Estados Unidos.
- $pais_{others}$ País de origen: Otros.
- **Giro** $Nombre_n$ 28 columnas giro nombre
- **noviembre**
- **diciembre**
- **enero**
- **febrero**

3. Creación del modelo

Despues de la transformacion de la base de datos a una en donde se modela en base a las variables categoricas a traves de su conversion en forma binaria, anexando las columnas anteriormente nombradas dentro del Dataset, se procede a la busqueda de la implementacion de un modelo de aprendizaje automatico que nos prediga el 'Gasto total' del proximo mes en base a la suma total de las categorias iguales entre si.

Esto para obtener una estimacion con alto grado de fiabilidad del monto total que cada cliente espera percibir dentro del proximo mes. Por lo tanto, a traves de realizar un analisis exploratorio de las posibles alternativas que podemos esperar, se procede a analizar los modelos y seleccionar uno en cuestion de las especificaciones de nuestro problema.

Es necesario tener en cuenta la naturaleza de los datos, es decir, variables categóricas y numéricas que se utilizan para predecir una variable numérica.

A continuación, se presentan algunos modelos de regresión que podrían ser útiles para predecir el gasto futuro de los clientes:

- **Regresión lineal**: Es un modelo que utiliza una función lineal para predecir una variable numérica a partir de una o más variables numéricas. En este caso, se podría utilizar una regresión lineal múltiple para predecir el gasto futuro de los clientes en función de las variables categóricas y numéricas.

- **Regresión Logistica**: Es un modelo que se utiliza para predecir la probabilidad de un evento binario (sí/no) a partir de una o más variables numéricas o categóricas. En este caso, se podría utilizar una regresión logística para predecir la probabilidad de que un cliente gaste más o menos en el próximo mes en función de las variables categóricas y numéricas.

- **Árboles de decisión**: Es un modelo que utiliza una estructura de árbol para dividir los datos en subconjuntos más pequeños y predecir una variable numérica o categórica a partir de las variables categóricas y numéricas. En este caso, se podría utilizar un árbol de decisión para predecir el gasto futuro de los clientes en función de las variables categóricas y numéricas.

- **Random Forest**: Es un modelo de ensamble de árboles de decisión que utiliza una combinación de varios árboles de decisión para predecir una variable numérica o categórica. En este caso, se podría utilizar un Random Forest para predecir el gasto futuro de los clientes en función de las variables categóricas y numéricas.

Por lo tanto, de acuerdo a las necesidades de nuestro proyecto, se decidio ahondar en el modelo de tipo Random Forest, siendo mas especificos en el tipo de **Random Forest Regression** debido a las siguientes razones:

El algoritmo de **Random Forest Regression** es una buena opción para este tipo de problema, ya que es capaz de manejar tanto variables numéricas como categóricas, y es robusto frente a valores atípicos y errores en los datos. Además, este algoritmo tiene la capacidad de manejar múltiples entradas y producir resultados precisos con alta eficiencia computacional.

En particular, el Random Forest Regression es un tipo de algoritmo de aprendizaje automático basado en ensamblaje que combina múltiples árboles de decisión para realizar predicciones precisas. En cada árbol individual, se toma una muestra aleatoria de los datos y se construye un árbol de decisión. Luego, se combinan las predicciones de todos los árboles individuales para producir una predicción final.

Algunas de las ventajas del algoritmo Random Forest Regression son:

Puede manejar datos faltantes y valores atípicos de manera efectiva. Es altamente preciso en comparación con otros algoritmos de regresión. Es resistente al sobreajuste debido a la técnica de ensamblaje utilizada. Es capaz de manejar tanto variables numéricas como categóricas sin la necesidad de transformaciones adicionales.

En resumen, el Random Forest Regression es una buena opción para este problema debido a su capacidad para manejar múltiples entradas, datos faltantes, valores atípicos y variables categóricas, así como su alta precisión y eficiencia computacional.

Ya elegido nuestro modelo de prediccion, se procede al entrenamiento del modelo en torno a subdividir los datos en entrenamiento y prueba, con una distribucion 80/20, donde primeramente se implementara con un valor conocido que es el del mes de Marzo para entrenar a nuestro modelo y compararlo con su valor real, donde posteriormente se utilizara para predecir el 'Gasto total' del mes de abril.

Por lo tanto, dividimos nuestras variables predictoras en base a las demás características, que después de la transformación de datos, se extrapolarián a ser las siguientes: edad del cliente, categoría de la compra, país de origen, sexo y cuatro meses de gastos totales; mientras que nuestra variable objetivo seleccionada fue 'marzo', ya que el algoritmo necesita entrenarse antes de ser usado, teníamos que darle el resultado de la predicción.

Por lo que despues de probarlo inicialmente con un random state = 42 y una profundidad = None (que significa que el árbol crecerá hasta que todas las hojas contengan un número mínimo de muestras, que suele ser 1 por defecto), asi como un n estimators = 100, pudimos observar con un tiempo moderado de procesamiento, el rendimiento a traves del coeficiente de determinacion fue equivalente a un x^2.

4. Resultados

En la Figura 1 se muestra la primera gráfica que se despliega al cliente que es la predicción del gasto total que se obtuvo para el próximo mes del cliente 4, la barra de predicción aparece en color amarillo para diferenciar los montos reales representados por las barras negras. El algoritmo tomó en cuenta las características específicas del cliente y los gastos realizados durante los meses anteriores y desplegó el resultado que se observa.

En la Figura 2, se muestra un ejemplo de los diversos tipos de gastos clasificados que fueron realizados por un cliente en un mes, cabe destacar que este información es de suma importancia que se conocida por el mismo, puesto que le ayudará a determinar y valorar sus finanzas y en el área en la cual está invirtiendo más.

En la Figura 3 se presenta el desgloce de los gastos del cliente por mes y en total ordenados del monto mayor al

Figura 1. Ejemplo de predicción de datos un cliente por mes

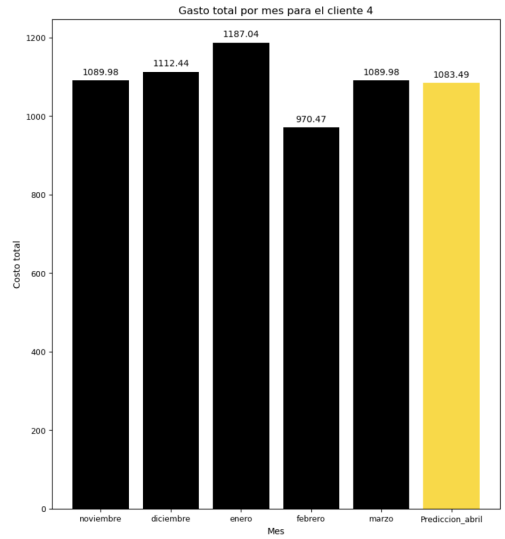
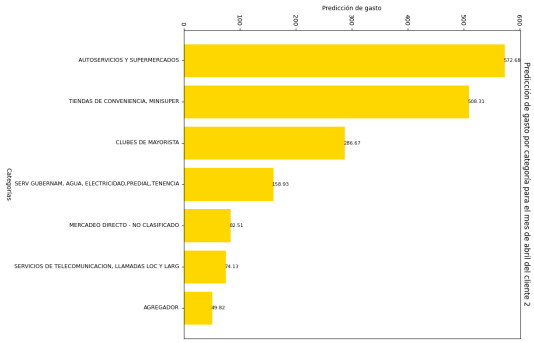


Figura 2. Clasificación de gastos de un cliente al mes



menor, agrupados por la categoría de la compra. Esto permite al usuario visualizar el nombre el comercio donde hizo las compras y ver detalladamente la proporción de sus gastos.

Figura 3. Desgloce de gastos del cliente 4

id_cliente	nombre_comercio	giro_nombre	mcc_nombre	December	February	January	March	November	Total
4	OPENPAY*TRANSPAIS UNICVICTORIA TAM	AGREGADOR	AGREGADOR	616.72	322.33	765.92	0.0	0.00	1704.97
4	MERCADO PAGO 3 CIUDAD DE MEX001	OTROS	MERCADO DIRECTO - VENTAS POR CATALOGO	475.98	0.00	0.00	0.0	0.00	475.98
4	ZAP ITALICA CD VICTORIA T028	RETAIL	ZAPATERIAS	0.00	0.00	199.07	0.0	0.00	199.07
4	CFE CONTIGO MU MEXICO DF DF	GOBIERNO	SERV GUBERNAM, AGUA, ELECTRICIDAD,PREDIAL,TENE...	0.00	0.00	117.24	0.0	0.00	117.24
4	CONEKTA*OXXOSPIN CIUDAD DE MEX001	OTROS	SERVICIOS RELACIONES PUBLICAS (SERVICIOS CONSU...	0.00	0.00	0.00	0.0	11.79	11.79
4	OPENPAY*TRANSPAIS SAN VICTORIA TAM	AGREGADOR	AGREGADOR	0.00	10.45	0.00	0.0	0.00	10.45

Referencias

[1] BBVA. Salud financiera: una clave para fortalecer la confianza de los clientes, 2023.