Proyecto de ML para incrementar el gasto anual de clientes en la tienda

Introducción

Se trata de una tienda exclusiva especializada en la confección y venta de ropa a medida. Se destaca por ofrecer consultorías altamente personalizadas. Los clientes visitan la tienda para recibir asesoramiento directo de estilistas expertos que ayudan a crear piezas únicas que se ajusten perfectamente a sus preferencias y medidas. Tras la sesión de consultoría, los clientes pueden hacer pedidos de ropa mediante una aplicación móvil o del sitio web de la empresa.

Objetivos del Proyecto

1. ¿Cuáles son los objetivos del negocio?

Aumentar las ventas evaluando dónde concentrar los esfuerzos: mejorando la experiencia de los clientes en el sitio web o en la aplicación móvil

2. ¿Qué decisiones o procesos específicos desea mejorar o automatizar con ML?

Se busca optimizar las decisiones empresariales relacionadas con la experiencia del cliente y automatizar la predicción del gasto anual de cada cliente.

3. ¿Se podría resolver el problema de manera no automatizada?

Si bien la predicción del gasto anual de cada cliente se puede hacer utilizando hojas de cálculo o software estadísticos, el uso de machine learning permitirá automatizar este proceso mediante entrenamientos programados que incorporen rápidamente las transacciones diarias sin necesidad de intervención manual.

Metodología propuesta

4. ¿Cuál es el algoritmo de Machine Learning más adecuado para resolver este problema? ¿Cómo justifica la elección de este algoritmo?

La tienda cuenta con un conjunto de datos actualizado que incluye información identificativa de cada cliente,

Variables independientes

1.la suma anual que ha gastado en la tienda

2. el tiempo dedicado a interactuar en el sitio web

3. tiempo dedicado a interactuar en la aplicación móvil,

4. el estado de miembros: entiendo si el cliente se ha dado de baja o no

Según el apartado del pdf en cuestión:

¿Se podría resolver el problema de manera no automatizada? ***Si bien la predicción del gasto anual de cada cliente*** *se puede hacer utilizando hojas de cálculo o software estadísticos, el uso de machine learning permitirá automatizar este proceso mediante entrenamientos programados que incorporen rápidamente las transacciones diarias sin necesidad de intervención manual.*

Se puede deducir que la variable dependiente u objetivo a predecir es numérica y es el gasto anual de cada cliente.

Al disponerse de muchas variables históricas, y tenemos la variable suma anual que ha gastado en la tienda del pasado, se trata de un problema de ML supervisado, ya que tenemos etiquetas de salida que conocemos (variable dependiente), suma anual que ha gastado cada cliente. Lo primero que probaría es **un modelo de regresión lineal múltiple** con **variables numéricas independientes, 2, 3** de las comentadas anteriormente por ser el más simple de todos.

Para validar si el modelo se ajustara correctamente, miraría la métrica R2, **el coeficiente de determinación**. Si fuera muy cercano a 1 entonces el modelo estaría muy bien ajustado, aunque deberíamos vigilar el tema del sobreajuste. Obviamente, además miraríamos mes a mes, si con las variables 2 y 3, si la predicción de la compra mensual encaja con lo realmente gastado por el usuario. Esto nos indicaría si harían faltas más variables o no a incorporar al modelo.

En caso de que no obtuviéramos buenos resultados con el modelo de regresión, se evaluarían otros modelos como el modelo de regresión polinomial. Antes deberíamos transformar las dos variables, la 2 y la 3 elevándolas al cuadrado, a la tercera, etc. Y probar muchas combinaciones con un modelo de regresión lineal múltiple hasta encontrar un R2 bastante bueno y cercano a 1 para garantizar un buen ajuste. En realidad, lo bien ajustado que fuera el modelo, debería pactarse con la política de la empresa, para saber cuál sería el error aceptado.

El algoritmo principal de ML hará una predicción del gasto anual y mensual del cliente. Para mejorar la experiencia del cliente, que es la otra vertiente que nos quedaba, la empresa utilizaría otros algoritmos de ML para proponerle más o menos productos recomendados según el gasto mensual previsto. Estos algoritmos tendrían como objetivo incentivar al cliente a superar el gasto mensual previsto por el algoritmo principal.

¿Qué métricas de evaluación se utilizarán para medir el rendimiento del modelo?

Como el modelo que utilizaremos será un modelo de regresión lineal, la bibliografía nos indica cuáles son las métricas más adecuadas para medir el rendimiento del modelo en estos casos.

Error Absolut medio (MAE)

Esta podría ser la métrica más obvia de evaluación del modelo de regresión. Sería calcular el promedio del error entre el valor y predicho (y_pred) con el modelo de regresión versus el valor (y_actual) observado en valor absoluto. Este cálculo con valor absoluto evita que errores de signo contrario se compensaran entre sí y nos pudieran dar un error promedio bajo.

```
MAE = Σ|y_pred - y_actual| / n
```

La ventaja del MAE, es que tiene unidades de la variable de salida (objetivo), y por tanto podemos comparar más fácilmente si un error es grande o no. Por ejemplo, imaginemos que de la cartera de clientes que vamos a estudiar vemos que los valores observados de la variable objetivo, en nuestro caso el gasto anual en la tienda, se encuentran en el rango de 300 € a 700 € para diferentes años. Si con un modelo de regresión obtuviéramos un MAE de 100 €, pues claramente sería un error relativo muy elevado comparado con los valores y observados.

Error Cuadrático medio (MSE)

Es muy parecido al anterior, pero en este caso, el error de cada una de las predicciones (y_predi) respecto al valor observado (y_actual), lo elevemos al cuadrado en vez de aplicar el absoluto, para evitar que errores de signo contrario se compensen entre sí, y para amplificar los errores de los valores que se consideran *outliers* (los que están muy por encima de dos desviaciones estándar, por ejemplo).

Es decir, si justamente queremos que el modelo sea muy sensible a los outliers o valores anómalos, valores que están fuera del patrón habitual de los valores, entonces esta métrica de evaluación será ideal para que se mida este efecto.

```
MSE = Σ(y_pred - y_actual)^2 / n
```

Aun así, la gran desventaja, es que el MSE no tiene unidades de salida comparables con la variable y observada (la objetivo), y por tanto su interpretación puede ser menos intuitiva.

Raíz cuadrada del error cuadrático medio (RSME)

Tal como dice su nombre se calcula como la raíz cuadra del MSE, la métrica anterior. Justamente se hace así, para que las unidades del error, sean comparables con las unidades de la variable objetivo y.

Coeficiente de determinación (R^2)

Como el modelo que aplicaríamos a la tienda en principio sería un modelo de regresión lineal, otra métrica de evaluación sería el R^2, **el coeficiente de determinación**. Esta métrica indica cuánta varianza en los valores de la variable dependiente puede ser explicada por el modelo. Donde la varianza total es la suma de las diferencias cuadradas entre los valores reales y su media Este parámetro tiene la ventaja, que siempre nos dará entre 0 y 1 independiente de los datos que tengamos, y eso es útil para evitar interpretaciones distintas. Un valor muy cercano a 1, nos indicará que el modelo de regresión lineal estaría muy bien ajustado a nuestros datos y, al contrario, valores cercanos a 0 indicaría un muy mal ajuste.

Es importante señalar, que en caso que nuestro modelo de regresión fuera no lineal y el modelo estuviera muy bien ajustado a los datos, este coeficiente de determinación R^2 no sería el adecuado, pues nos daría valores de R^2 bajos por ser el modelo no lineal y en cambio tendríamos un buen ajuste. Eso es importante recalcarlo, aunque como hemos señalado hacemos la hipótesis que nuestro modelo se ajustará a un modelo de regresión lineal.

```
R² = 1 - (MSE / varianza_total)
```

Como conclusión, utilizaremos **el coeficiente de determinación R^2 y** el **RSME como métricas de evaluación** del modelo de regresión lineal para la tienda de ropa**.**

Datos Disponibles

5. ¿Qué datos están disponibles para abordar este problema?

La tienda cuenta con un conjunto de datos actualizado que incluye información identificativa de cada cliente, la suma anual que ha gastado en la tienda, el tiempo dedicado a interactuar tanto en el sitio web como en la aplicación móvil, y el estado de miembros.

Métrica de Éxito

6. ¿Cuál es la métrica de éxito para este proyecto?

Aumento en el Gasto Anual Medio por Cliente. Esta métrica reflejaría directamente la efectividad del modelo al mejorar las decisiones de la empresa.

Responsabilidades Éticas y Sociales

7. ¿Qué responsabilidades éticas y sociales es importante tener en cuenta?

Completar por los estudiantes

Como hemos visto la ética en la implementación de algoritmos de Machine learning de forma simplificada puede tener diferentes pilares donde los más destacables serían: **Transparencia, Justicia, Privacidad, Responsabilidad e Impacto social.**

Transparencia

La empresa comunicará, que se utilizarán todos los datos de interacción por la web, así como otros de las compras, para mejorar la experiencia del cliente, y así poderle ofrecer los productos más adecuados y de un coste que se ajusten a algo más del presupuesto mensual del cliente.

Justicia

En el caso del género, el algoritmo no tratará los datos de manera diferente por ser un hombre o una mujer, sino que aplicará los algoritmos de la misma manera, sin incidir en cómo serían los resultados para favorecer más las recomendaciones de compras mensuales a un tipo de género pensando que históricamente, el género femenino es más propenso a comprar ropa.

Privacidad

El cliente debe dar siempre su consentimiento para que la empresa recoja y almacene sus datos privados, como el tiempo de interacción en la web o móvil, y las

compras realizadas. Además, se debe proporcionar una manera clara y fácil para que el cliente pueda darse de baja y eliminar sus datos cuando lo desee.

Responsabilidad

Si los datos del cliente se filtran sin autorización, la empresa debe garantizar, en el momento de la autorización de la recopilación, que será responsable de indemnizar al usuario por cualquier perjuicio causado.

Impacto social

Se debería hacer difusión de esta mejora de experiencia a nivel de marketing a todo grupo cultural de la misma manera, para que todo el mundo tuviera las mismas oportunidades de acceso a la aplicación.