Ad Click Prediction - Classification Problem

Proyecto Final Inteligencia Artificial
Luis Alejandro Ariza Garcia, arizalalejandro@javeriana.edu.co
Inteligencia artificial

abstract, En el siguiente documento encontrara un algoritmo de clasificación en el cual se predice si un cliente concretara o no la compra de un producto mediante el uso de algoritmos de machine learning.

1. **Comprensión empresarial**

Se busca haciendo uso de Inteligencia artificial analizar la base de datos seleccionada, con esta base de datos se busca analizar si un cliente hace click en un anuncio o no, de modo que, al conocer información demográfica del cliente, genero, edad y salario estimado es posible estimar si la persona hará o no una compra.

Es importante conocer si los clientes finalizaran o no una compra es algo que cualquier empresa desea saber, así mismo conocer más información sobre el cliente y de este modo entender de que forma pueden llegar a mas personas segmentando sus clientes.

El objetivo es conocer si un potencial cliente hace click en un anuncio, luego de esto saber que tan probable es concretar una venta.

2. **Comprensión de datos**

Para recopilar los datos se realizo la descarga de la base de datos de kaggle, para realizar el análisis necesario se hace uso de la herramienta Google Colab.

La base de datos esta compuesta por las siguientes características:

1. 'User ID': identificación única para el consumidor
2. 'Edad': edad del customer en años
3. 'Salario estimado': Promedio. Ingresos del consumidor
4. 'Género': si el consumidor era hombre o mujer
5. 'Comprado': 0 o 1 indicado haciendo clic en Anuncio

El user ID identifica el cliente en la base de datos, esto para encontrar un cliente especifico de forma rápida y organizada, es importante conocer la edad de nuestros clientes ya que de este modo es más fácil clasificar que compran las personas según su rango de edades, el salario estimado es un factor muy importante a tener en cuenta, esto nos puede dar una orientación de aproximadamente cuanto dinero esta dispuesto a gastar una persona, el genero nos ayuda a clasificar los potenciales productos que se pueden ofrecer a las personas, y finalmente si la persona finalmente concluyo la compra luego de haber hecho click en el anuncio o no.

3. Preparación de datos

Se decide excluir el User ID ya que este no aporta en el análisis de datos que se desea realizar, ya que es una variable independiente a todos los otros datos y que no necesitamos para realizar la clasificación.

Para corregir los datos se eliminan los datos faltantes, y se rellenan con el promedio de datos existentes, y se rellena con una regresión usando las otras características.

Ya que la base de datos no provee información adicional con la cual se puedan construir datos adicionales se trabajará únicamente con los datos que se tienen.

Se cambian los datos categóricos a datos numéricos, esto hablando del genero de las personas, se debe cambiar estas etiquetas a valores numéricos para poder ser analizados.

4. Modelado

Se decidió utilizar el algoritmo de clasificación para predecir sobre la base de la demografía del cliente como variable independiente, de modo que cuando las personas realicen una búsqueda en base a sus datos se les pueda enseñar información dirigida y productos que probablemente puedan adquirir

Los datos de entrenamiento o «training data» son los datos que usamos para entrenar un modelo. La calidad de nuestro modelo de aprendizaje automático va a ser directamente proporcional a la calidad de los datos. Por ello las labores de limpieza, depuración o «data wrangling» consumen un porcentaje importante del tiempo de los científicos de datos.

5. Evaluación

Se realizo la división de datos haciendo uso del método Sklearn “train_test_split”, de este modo se dividen los datos en tipo train y test, de modo que se utilizan unos para hacer el entrenamiento del algoritmo y otro para probarlo.

Para realizar la normalización de los datos se hizo uso del método “StandardScaler” de Sklearn, normalizando los datos en valores binarios de 0 y 1.

Se decide hacer uso de la PCA para limpiar nuestra data y poder visualizar nuestra data en menos dimensiones, al hacer uso de la PCA el algoritmo verificara en qué dirección tenemos una mayor varianza, evaluamos obteniendo la matriz de covarianza con la cual tenemos la medida de la dispersión de nuestros puntos alrededor del centro de masa, esta matriz nos dice la variación de cada una de las dimensiones con respecto de la demás con lo cual se obtuvo una matriz de covarianza de este modo.

```
Matriz covarianza [0.30129372 0.26352067 0.23388952]

0.7987039105268429
```

Se obtuvo un porcentaje del 79% haciendo uso de PCA

El segundo método utilizado fue la regresión logística, ya que al tener un problema de clasificación binaria en donde queremos saber si realmente el cliente comprara o no es un método bastante útil para predecirlo, se utilizaron diferentes hipótesis para realmente observar con cual de estas se obtiene un mejor resultado, el mejor resultado obtenido fue el de la tercera hipótesis donde se obtuvo que:

```
hipotesis X+X^3+1, el MCC es:  0.8796856499979752
hipotesis X+X^3+1, el F1 es:  0.9500000000000001
```

Como se puede evidenciar el MCC obtenido es de 0.87 haciendo uso de esta hipótesis fue el que mejor se ajusto y el mas cercano a 1 con lo cual hay un buen coeficiente de correlación.

Así mismo evaluando el F1 que obtuvimos es posible evaluar la precisión del modelo, el valor de F1 se utiliza para combinar las medias de precisión y recall en un solo valor, esto resulta practico ya que hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

6. Conclusiones

- El mejor resultado obtenido fue haciendo uso de la regresión logística con la tercera hipótesis, la regresión logística es un buen método para implementar debido al problema que fue abordado.
- Al tener pocas clases no vale la pena disminuir 1 sola característica al hacer uso de la PCA ya que el porcentaje de acierto se disminuye bastante.