****

**Trabajo Final**

**Deep Learning**

Profesor:

Mauricio A. Valle PhD,

Programa:

Magister en Data Science UAI.

Presentado por los alumnos:

Efraín Aguilera

Felipe Leyton

Pedro Morales

Paulina Ávila

Diciembre 2021

**INTRODUCCIÓN**

El trabajo final del ramo Deep Learning tiene por objeto la aplicación de la materia vista en clases, es decir, alguno de los modelos de redes neuronales aplicada sobre una base de datos real. El objetivo es modelar un problema de clasificación, que permita ofrecer una aproximación a la solución de la problemática.

Se debe proponer el modelo, reportar la problemática, contexto, dataset, arquitectura del modelo, variaciones y desempeños logrados.

En este marco, el trabajo desarrollado por este grupo toma una aplicación móvil utilizada por conductores de vehículos que entrega descuentos en restaurantes u otros servicios similares. A continuación, detalles al respecto:

* Hay una aplicación móvil utilizada por conductores de vehículos, que entrega a éstos, ofertas, descuentos o similares (“cupones”).
* Estos cupones son ofrecidos, a través de esta plataforma digital, por distintas empresas (restaurantes, cafeterías, bares u otros).
* El conductor puede aceptar o no el cupón; en caso que lo acepte, en principio debería cambiar su recorrido para llegar al restaurante en el que aceptó el cupón.
* Este aplicativo puede asimilarse a la aplicación Waze, en Chile, que sigue el recorrido del vehículo, pero en este caso se agrega la publicidad de ofertas de restaurantes u otros servicios. Así, si el conductor acepta el cupón, debiera cambiar su recorrido hacia el lugar donde se encuentra el restaurante, por ejemplo.

**PROBLEMA**

* El aplicativo envía publicidad masiva sobre todos los servicios con descuentos a todos/as conductores/as, indistintamente si éstos/as pudieran ser parte del mercado objetivo de dichos cupones.
* Esto vuelve ineficiente y poco efectiva la publicidad enviada, pudiendo focalizarla exclusivamente a aquellos grupos de clientes que les interesaría hacer uso de dichos cupones.

**OBJETIVO**

* Predecir si un/a conductor/a aceptará o no un cupón para acercarse a alguno de los distintos rubros que se ofrece a través de la aplicación móvil.

**FUENTE DE INFORMACIÓN**

* Los datos fueron extraídos del sitio web:

[https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation#](https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation)

* Esta base de datos fue generada a partir de una encuesta de Amazon Mechanical Turk.
* Ésta recolectó información sobre un conjunto de atributos que se espera permitan predecir si el/la conductor/a aceptará o rechazará el cupón.

**ENTENDIMIENTO DE LOS DATOS**

Análisis de datos

* La base de datos contiene un total de 12.684 registros.
* Cuenta con 26 columnas, donde 18 atributos son categóricos y 8 numéricos.
* La variable “Y” corresponde a si acepta o no el cupón. Es un tipo de dato “int” con valores de 1 y 0, donde las personas que aceptan el cupón (1) representan el 57% del total de registros y las que no lo aceptan (0) representan el 43% del total. Lo cual dada su distribución porcentual permite apreciar que los datos no están desbalanceados.



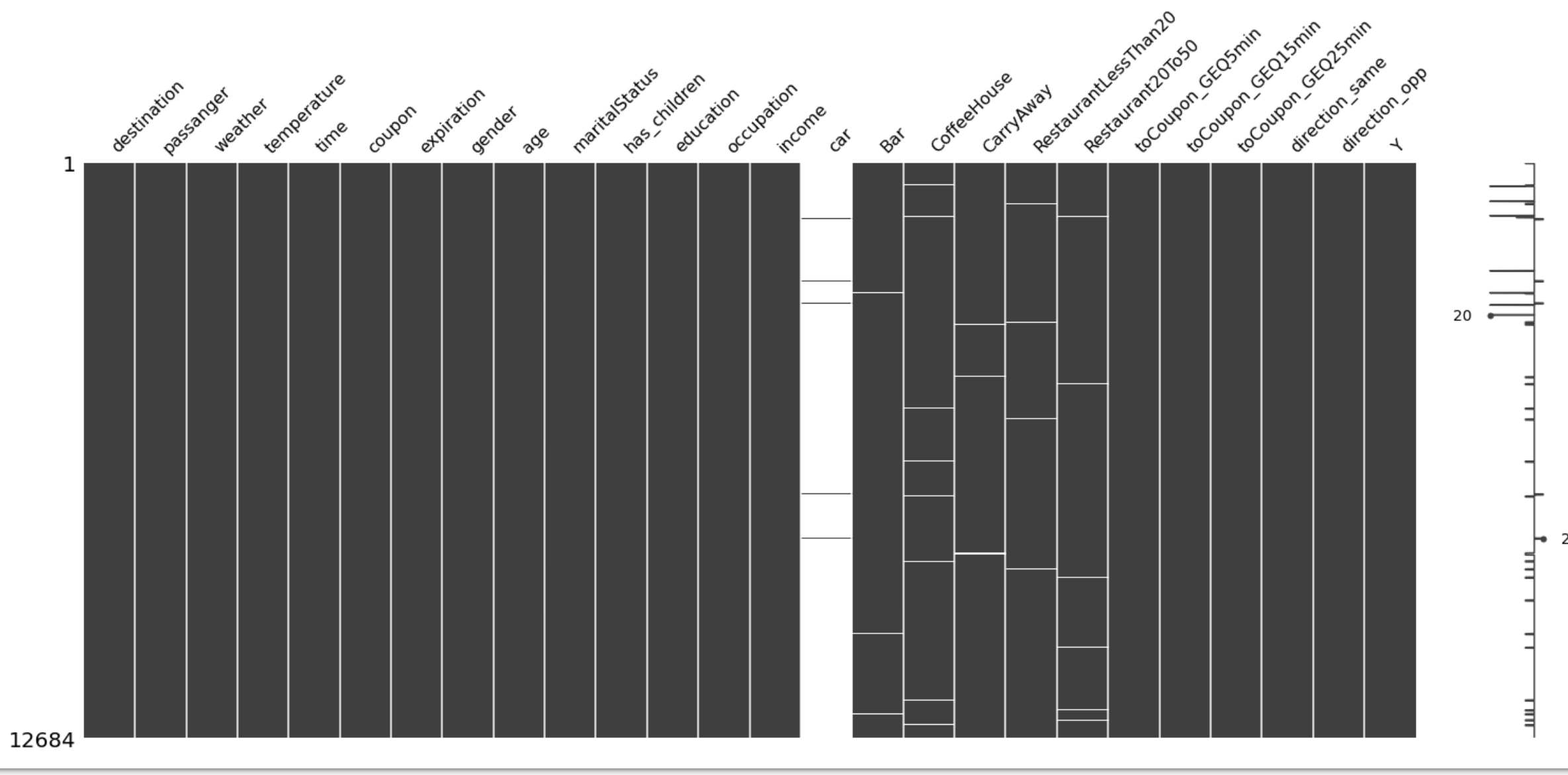
**Tabla N°1:** Variables de la base de datos.

Las variables corresponden a:

* Características del/a conductor/a: Edad, Género, Educación, con/sin hijos, Estado civil, Ocupación (área de la industria donde trabaja), etc.
* Condiciones ambientales: clima, temperatura.
* Hora del viaje.
* Tipo de cupón: % descuento en restaurante, en cafetería o en bar.
* Frecuencia mensual de consumo en restaurantes, cafeterías o bares.

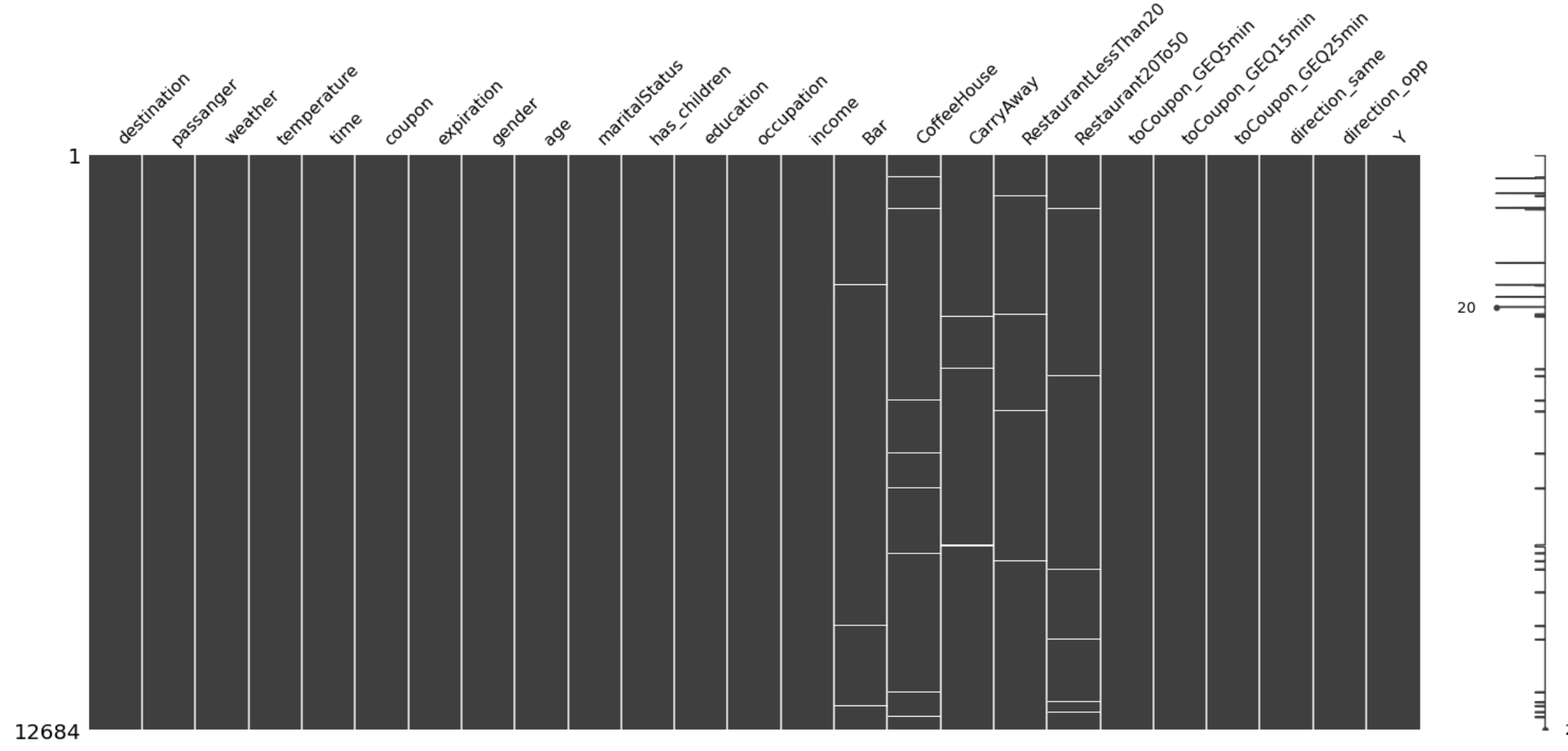
Nota: La base de datos utiliza el término “Passenger” para referirse al conductor/a del vehículo.

La siguiente gráfica representa una visualización de la cantidad de registros por columna. Se aprecia que el campo “car” de tipo categórico, contiene muchos datos vacíos, por lo que se propone eliminarlo del dataset de cara al análisis, creación y posterior evaluación del modelo

****

**Gráfico N°1:** Variables y datos nulos en las mismas

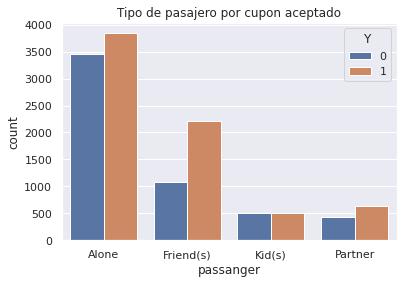
Una vez eliminado el atributo “car” del dataset, se visualiza una base más limpia. Además, se eliminaron todos los valores nulos de la base de datos, los cuales no eran representativos.



**Gráfico N°2:** Variables y datos nulos en las mismas, sin incluir la variable “car” (eliminada).

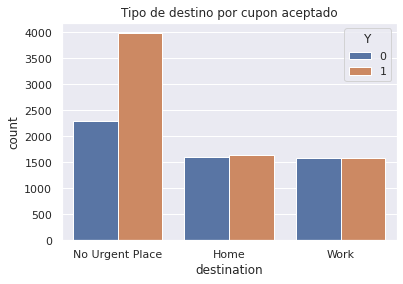
A continuación, presentamos algunas visualizaciones que permiten dar a conocer la comparación de los datos en base a sus categorías con gráficos de tipo barra:

Se puede observar que la mayor cantidad de registros se encuentran en la categoría de conductor sin acompañante “Alone”, pero que la única que marca una diferencia significativa entre aceptar y no el cupón es cuando el conductor va acompañado de amigo/s “Friends”.



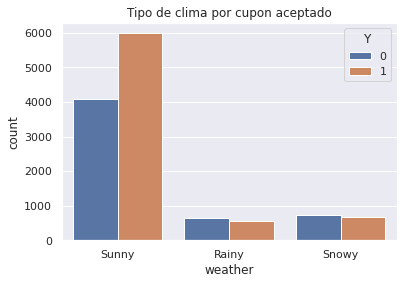
**Gráfico N°3:** Número de cupones no aceptados y aceptados según tipo de conductor (pasajero).

También se puede observar que existe una alta aceptación del cupón cuando la persona no tiene urgencia de llegar a destino.



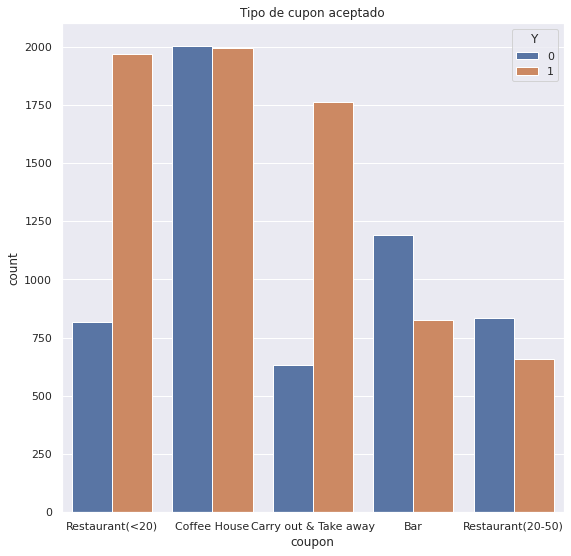
**Gráfico N°4:** Número de cupones no aceptados y aceptados según urgencia del viaje.

Se aprecia una alta aceptación del cupón cuando el día está soleado. En este caso, no se cuenta con la variable fecha para determinar si los datos son de una época particular del año o en distintas épocas, tampoco se conoce la fecha en la que se realizó por ende es difícil determinar la influencia de esta variable para el análisis.



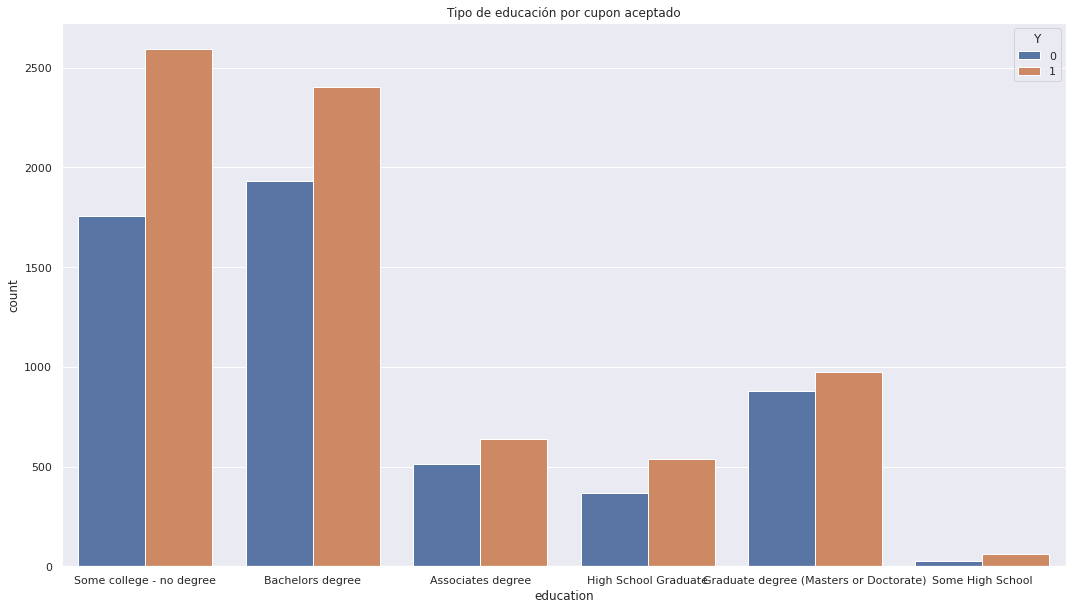
**Gráfico N°5:** Número de cupones aceptados y no aceptados según tipo de clima.

Se observa que existe preferencia en la aceptación de cupones mayormente asociados a “Restaurant”, “Coffe house” y cupones de “Carry out & Take away”.



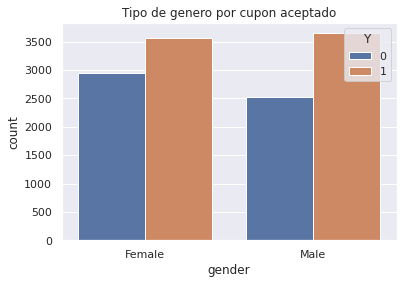
**Gráfico N°6:** Número de cupones aceptados y no aceptados según porcentaje de descuento y lugar que lo ofrece.

En cuanto al gráfico de tipo de “Education” por cupón aceptado podemos decir que sería bueno, debido a la distribución de los datos agrupar algunos campos que tienen pocos registros como “High School Graduate y Some High School, ya que de esta manera podríamos mejorar la performance del modelo.



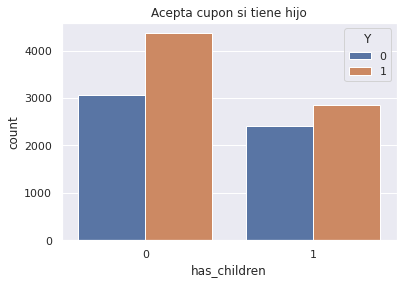
**Gráfico N°7:** Número de cupones aceptados y no aceptados según nivel educacional.

No se aprecia una diferencia importante en relación a la cantidad de cupones aceptados por tipo de género.



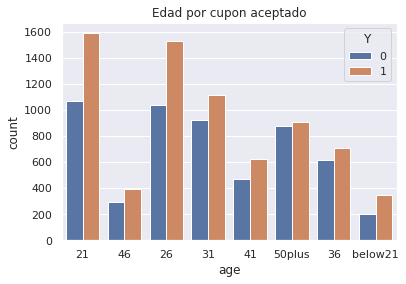
**Gráfico N°8:** Número de cupones aceptados y no aceptados según tipo de género.

Se observa que existe una tendencia a aceptar en mayor cantidad el cupón cuando la persona no tiene hijos.



**Gráfico N°9:** Número de cupones aceptados y no aceptados según si la persona tiene o no hijos.

Se observa que existe una tendencia que denota que mientras menos edad tenga la persona, mayor aceptación tiene del cupón.



**Gráfico N°10:** Número de cupones aceptados y no aceptados según edad.

Una vez hecho el análisis exploratorio de los datos (EDA) podemos determinar que el data set a nuestro criterio cumple con los parámetros necesarios para poder realizar algún tipo de predicción.

**METODOLOGÍA PARA EL PREPROCESAMIENTO, MODELAMIENTO, EVALUACIÓN**

A continuación, el desarrollo de cada uno de estos puntos.

**PREPROCESAMIENTO**

Luego del análisis realizado de los datos, nos pudimos percatar que la variable “Weather” y la variable “Temperature”, cada una tenía 3 valores, y como indicamos en el EDA, dado que no contamos con información adicional para saber en qué período del año y en que fechas se tomaron los datos se decide eliminarlas de la data de estudio.

Por otro lado, considerando la gran dispersión que tenía la variable “Occupation”, que inicialmente tenía 25 valores distintos, se decide agruparlos en 11 grupos, dejando grupos homogéneos en cuanto a la cantidad de registros equitativos, considerando las primeras 6 categorías de forma individual y agrupando el resto de categorías en 5 grupos de similar cantidad de registros.

|  |  |
| --- | --- |
| Grupo 1 | Unemployed |
| Grupo 2 | Student |
| Grupo 3 | Computer y Mathematical |
| Grupo 4 | Sales & Related |
| Grupo 5 | Education & Training & Library |
| Grupo 6 | Management |
| Grupo 7 | Office & Administrative Support, Arts Design Entertainment Sports & Media |
| Grupo 8 | Business & Financial, Retired |
| Grupo 9 | Food Preparation & Serving Related, Healthcare Practitioners & Technical, Healthcare Support, Community & Social Services |
| Grupo 10 | Transportation & Material Moving, Personal Care & Service, Architecture & Engineering |
| Grupo 11 | Protective Service, Life Physical Social Science, Construction & Extraction, Installation Maintenance & Repair, Production Occupations, Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry |

Asimismo, se realizó el mismo tratamiento con la variable “Education”, la cual tenía 6 valores distintos, se convirtieron a 3 grupos, según cantidad de registros.

|  |  |
| --- | --- |
| Grupo 1 | Some college – no degree |
| Grupo 2 | Bachelors degree |
| Grupo 3 | Graduate degree, Associates degree, High School Graduate, Some High School |

Dado que la gran cantidad de los datos eran variables categóricas nominales (11 variables) y variables categóricas ordinales (5 variables) utilizamos la técnica de OneHotEncoder para poder ingresar los datos a nuestro modelo. De esta forma pasamos de tener un data set de 23 variables a uno de 89 variables, el cual se convierte en nuestro data set final para confeccionar la red neuronal.

Luego separamos las variables para entrenar el modelo, por un lado, dejamos la variable objetivo o variable a predecir “Y”, y el resto de las variables en otro conjunto de datos “X”.

Posteriormente dividimos el data set en Train y Test, y finalmente para asegurarnos de que los datos sean más íntegros los escalamos utilizando la función MinMaxScaler().

**MODELAMIENTO**

En relación al modelamiento, se presentan los pasos utilizados, considerando que el modelo seleccionado es de tipo **Clasificación Binaria.**

Identificamos la función de activación:

* Capa inicial: Función de activación Relu con una salida de 100 neuronas y un Rate de 0,5.
* Capa intermedia: Función de activación Relu con una salida de 24 neuronas y un Rate de 0,1.
* Capa de salida: Función de activación Sigmoide con 1 neurona de salida.

Definimos:

* Función de pérdida Loss Binary\_crossentropy.
* Optimizador Adam con un Learning Rate de 0,0001.

Métricas a utilizar:

* Loss
* Accuracy
* AUC
* Inicializador: “Sequencial” (Redes secuenciales, capa una detrás de otra).
* Capa: “Dense” (Todas las neuronas se conectan entre ellas).
* Kernel: “normal” (Todos los pesos de la capa deben tener un número que viene dado de una distribución normal con media 0 y varianza 1).

**EVALUACIÓN**

En este punto, luego de haber hecho el preprocesamiento y posteriormente el Modelamiento, se realiza la Evaluación del modelo.

Aprendizaje:

* Para evaluar nuestro modelo realizamos 32 combinaciones distintas, modificando tamaños de epochs y batch size, partiendo con epochs de 50 y aumentando gradualmente de 50 en 50 hasta llegar a los 300 epochs. Para cada epochs ajustamos el batch size partiendo en 25 y aumentando gradualmente a 50, 100, 150, 200 y finalmente 250.

**TABLAS RESULTANTES AL REALIZAR LA EVALUACIÓN DEL MODELO**

A continuación, los resultados de las pruebas realizadas, según las variaciones mencionadas en el párrafo anterior, en los Epoch (número de iteraciones) y en los Batch (tamaño de los lotes).

I. Epoch: 50 y 100. Batch: De 25 a 250 (25, 50, 100, 150, 200, 250)



II. Epoch: 100 y 150. Batch: 50- 100- 150- 200- 250



III. Epoch: 150- 200-250. Batch: 50- 100- 150- 200- 250



IV. Epoch: 250- 300. Batch: 50- 100- 150- 200- 250



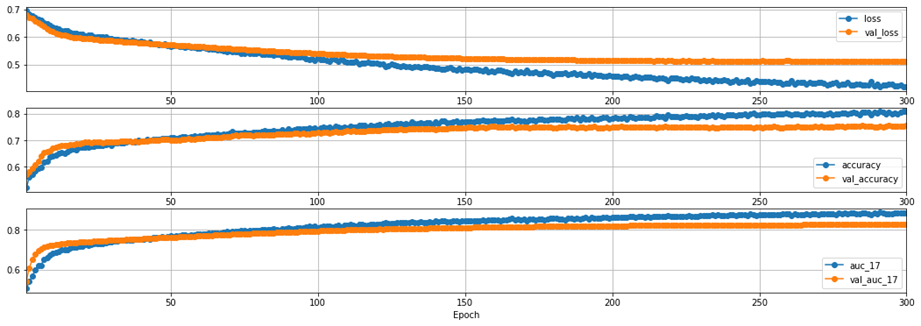
**SELECCIÓN DEL MODELO**

De acuerdo al análisis el modelo que entrega las mejores métricas es el siguiente:

Test 29, correspondiente a la combinación de un batch size de tamaño 100 y un epoch de tamaño 300, con la configuración inicial de la red neuronal.

A continuación, presentamos los resultados de las métricas:

* loss: 0.4218
* accuracy: 0.808
* auc: 0.8843
* val\_loss: 0.5125
* val\_accuracy: 0.7558
* val\_auc: 0.8266

****

**Gráfico N°11:** Resultados obtenidos con el modelo seleccionado.

Imagen de la pantalla de un celular con texto

Descripción generada automáticamente con confianza media

**CONCLUSIONES**

A la luz de los análisis realizados durante el presente documento, se pueden comentar los siguientes puntos relevantes:

* Con el dataset que realizamos la evaluación podemos concluir que para la cantidad de datos que posee, en primer lugar con un batch size de 100 y un epochs de 300, es suficiente para encontrar las métricas con mejor desempeño, ya que al probar con epochs más grandes observamos que los modelos tienden a volverse más inestables y comienzan sus curvas a separarse lo cual podría ser indicio de que el modelo se está sobre entrenando o cayendo en mínimos locales, además la ganancia entre hacer más de 300 epochs no es significativa como para alargar tanto el proceso, considerando también el uso ineficiente de los recursos.
* Dentro de los 32 modelos que analizamos en el escenario descrito anteriormente podemos observar que el mejor Accuracy se encuentra en la evaluación del modelo test 29, que fue el seleccionado por el equipo, no podemos dejar de mencionar que no es el que posee la mejor curva de Loss y AUC, sin embargo, esta dentro de las 3 mejores de los 32 modelos evaluados y nuestra decisión se basó en la métrica de Accuracy y la poco diferencia que tiene con respecto a los otros dos mejores en relación a las otras métricas evaluadas.
* Luego de realizado el análisis, y considerando que en este caso particular la idea es proponer un modelo que permita solucionar una problemática, nos encontramos con dificultades propias del entendimiento del problema, ya que no fue posible obtener un dataset propio, buscamos uno que pudiese ser aplicable al desafío planteado, pero al poco andar realizando el EDA nos encontramos con dificultades para entender el contexto en el que se recolectaron los datos que nos hizo asumir supuestos que no tenemos certeza si efectivamente son así, lo que se resume en que si bien estamos proponiendo un modelo de clasificación binaria las métricas que obtuvimos no nos convencen del todo y creemos que podrían ser mejores si tuviésemos algún contexto más amplio para entendimiento de los datos y un pre procesamiento que apoye ese entendimiento y suponga un mejor desempeño del mismo.