Andres Felipe Garcia Ravazze / Sara Gomez

Reporte Proyecto 2 Analitica Computacional

**1) Preguntas de negocio que quiere resolver para su cliente seleccionado:**

* ¿Cuáles son las características más relevantes que afectan la suscripción a un depósito?

La idea inicial para evaluar esta pregunta, es generar regresiones lineales para evaluar todas las variables posibles y clasificarlas dependiendo de su Beta.

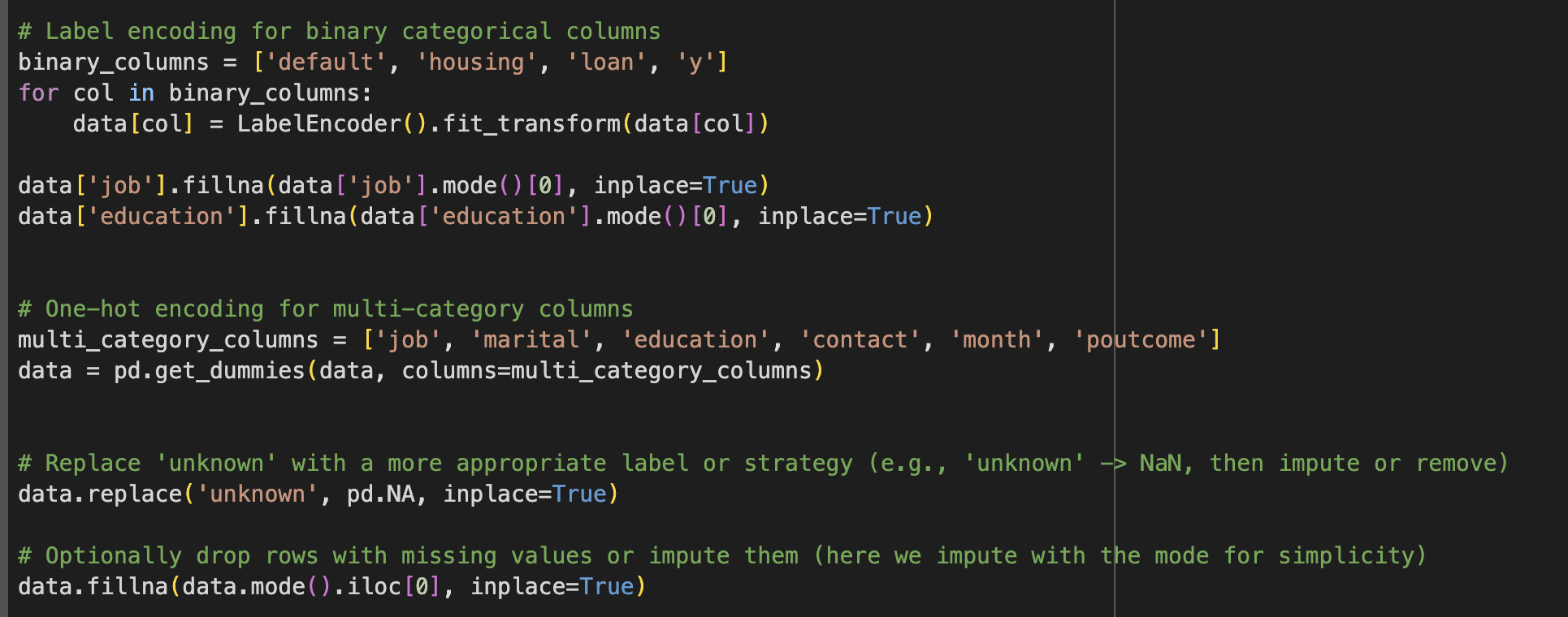
* ¿Cómo podemos utilizar el historial de pagos de los clientes en meses anteriores y sus préstamos y contactos pasados para predecir qué clientes tienen mayor probabilidad para realizar un depósito?

Realizar un modelo predictivo de clasificación para poder predecir esta probabilidad.

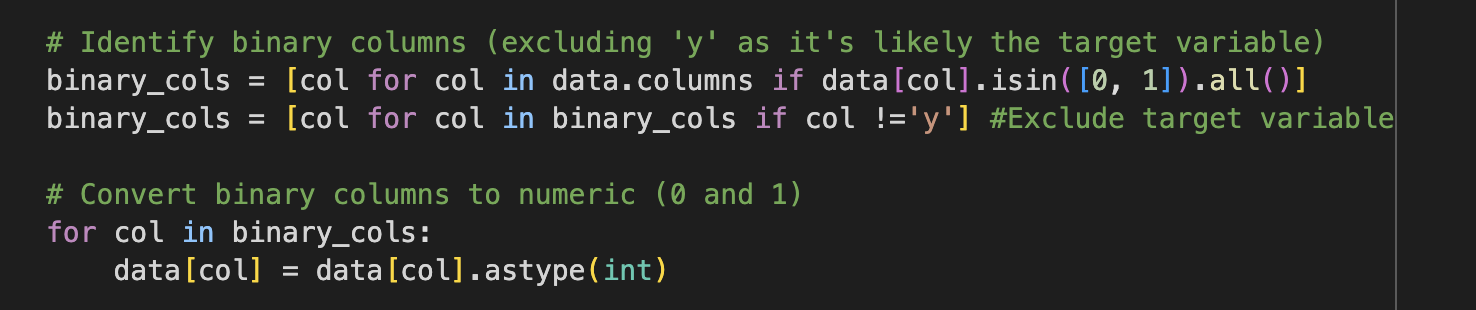
**2) Alistamiento de datos:**

En primer lugar se identificaron ciertas variables en específico (job, marital, education, contact, month, poutcome), lo especial de estas variables es que están divididas en subcategorías, es decir que dentro del data frame inicial estas no tienen un valor numérico si no dicen el estado. Por ejemplo en la variable jobs te dicen admin. haciendo relación al tipo de trabajo y así con todas. Para poder manejar este tipo de información de mejor manera, se decidió dividir estas variables por categorías, es decir crear una nueva columna (variable) que tenga el tipo de variable en específico, para job estará dividida en job \_”nombre del trabajo” osea job\_admin.

Esto se hizo mediante un label encoding, si inicialmente la fila tendría el trabajo relacionado a la nueva columna se pondría true mientras si no false y de esta manera se llenaría todo el dataframe con todas las variables.



Ahora bien, como establecimos anteriormente las variables divididas ahora se encuentran en formato dummies (True False), para poder realizar un modelo o un análisis preliminar hay que transformar estas variables a un formato binario que se pueda leer en un modelo, esto se hizo mediante una codificación binaria para 0 (False) y 1 (True).

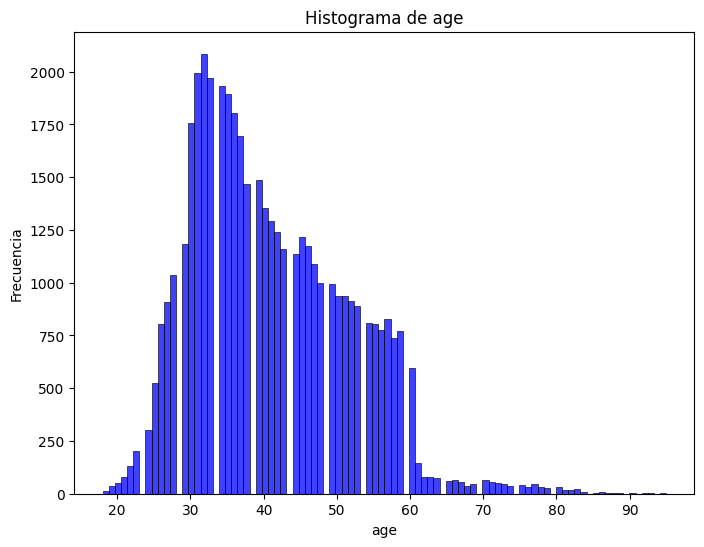


**3)Visualizaciones de datos:**

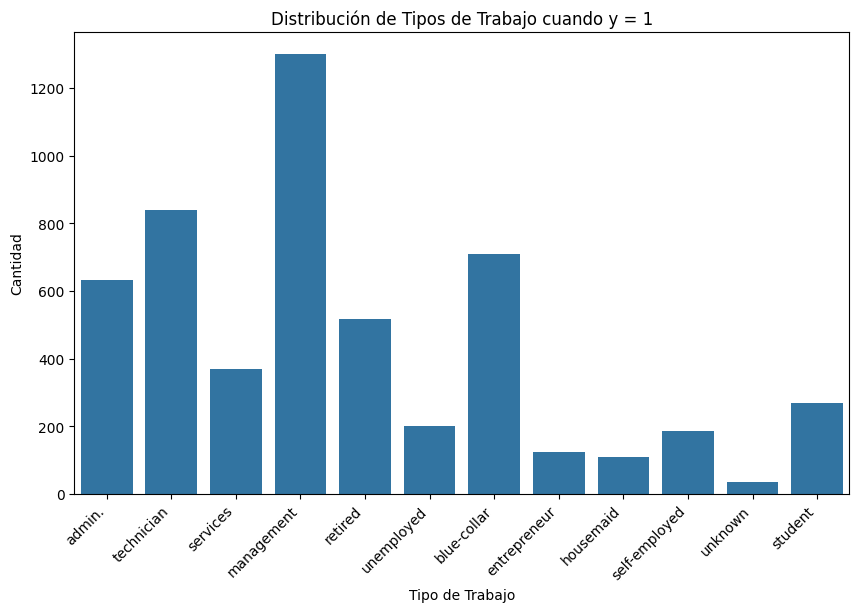
Una vez se alistaron los datos ahora si podemos realizar un análisis preliminar de variables para ver su comportamiento.

1. Mercado objetivo:

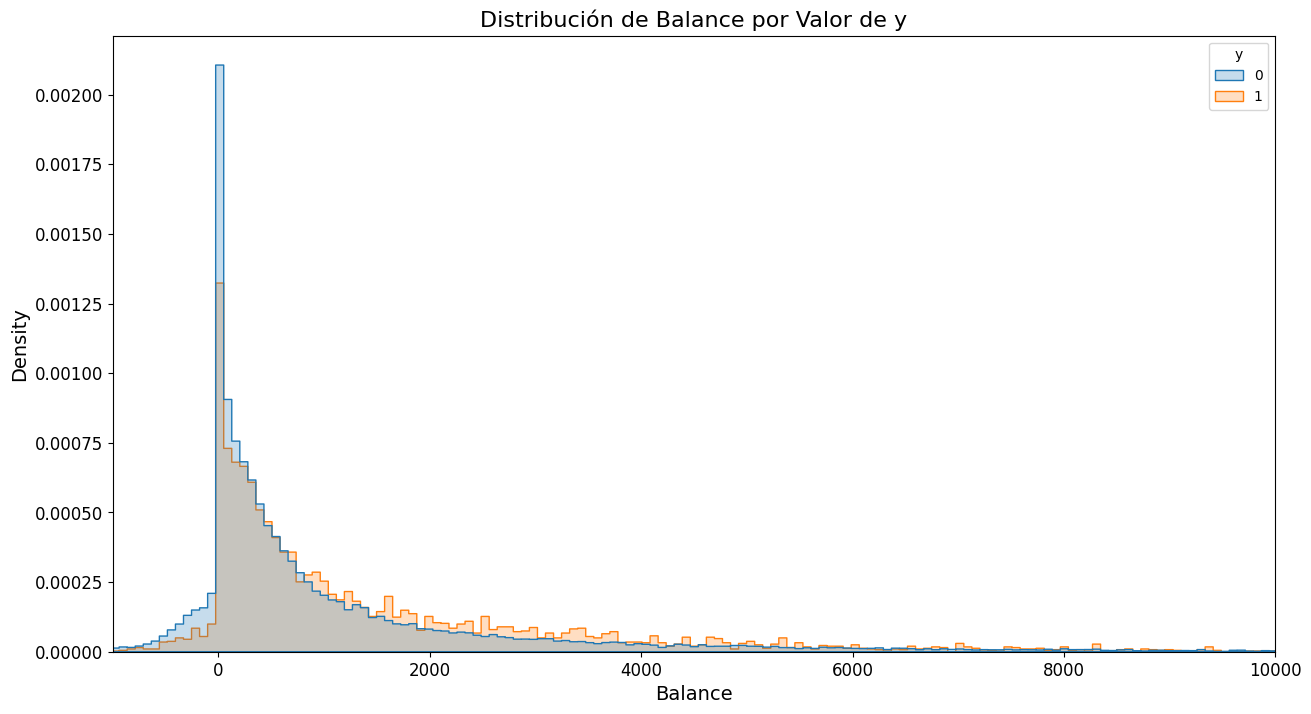
* Visualización: Diagramas de barras o gráficos de dispersión que muestran la relación entre características como edad, nivel educativo, salario y la probabilidad de realizar un depósito.
* En primer lugar se decidió hacer un historiografía de la edad para determinar si existía una edad en particular predominante, y en efecto el rango más visto es entre 30 a 40. Con esto nos hacemos claramente una idea de cuál es nuestro mercado objetivo, así podemos realizar después más análisis con otras variables para determinar que variables afectan mas el resultado.

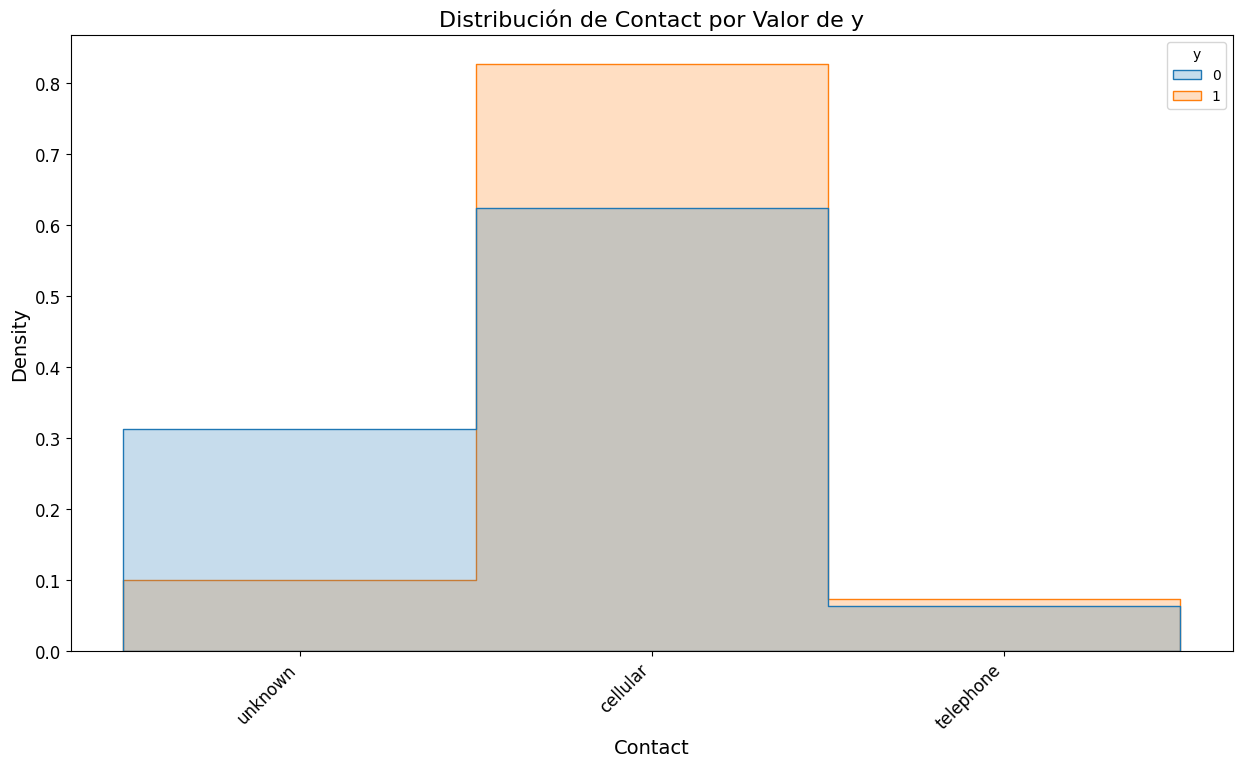


Ahora identificamos el tipo de trabajos cuando y (nuestra variable de aceptamos) es positiva, con esto podemos entender que nuestro mercado objetivo no ronda entre los 30-40 años si no también en personas con cierto tipo de trabajos como administrador, management, blue-collar y tecnico.



Para finalizar este análisis se decidió establecer una relación entre el balance (saldo) y la tasa de éxito para determinar nuestros esfuerzos hacia dónde deberían ir dirigidos y por qué medio.



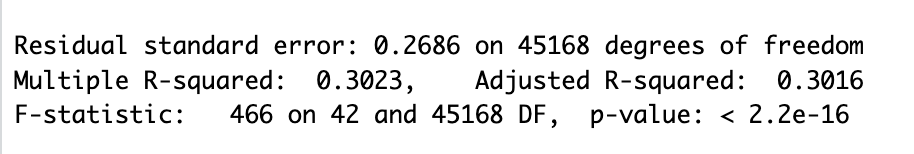
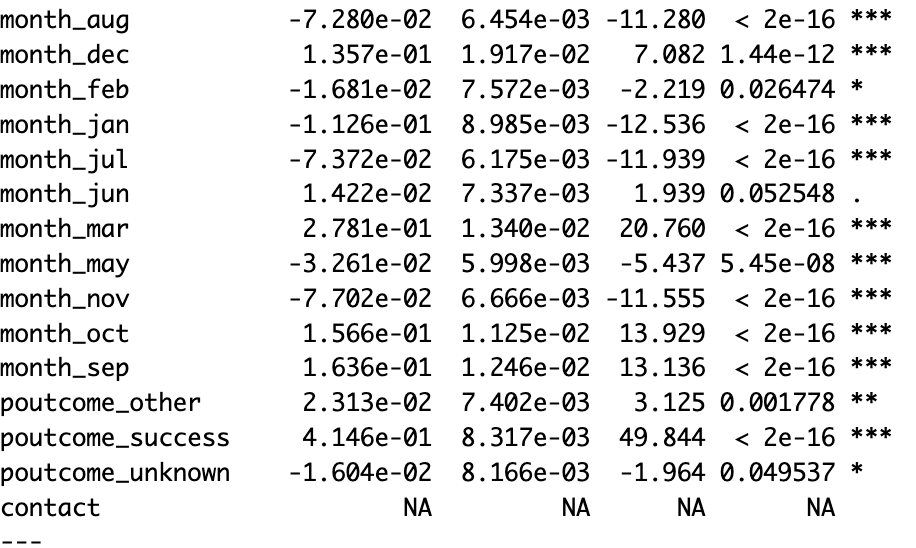


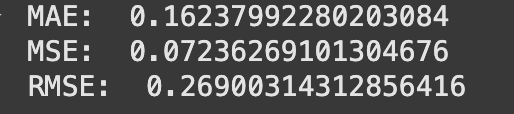
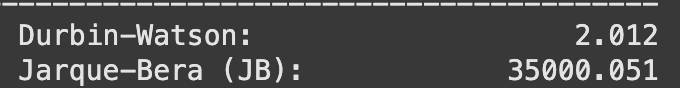
Finalmente podemos establecer con estas gráficas que el salario o balance de los éxitos ronda entre los 50 a 2000 euros. Ya habiendo establecido nuestro mercado objetivo que serían personas entre 30-40 años con trabajos de Admin., Management, Blue-Collar y Technician. Decidimos averiguar por qué medio se tendría la mayor tasa de éxitos y encontramos que sería por medio de llamadas al celular.

**4) Modelamiento**

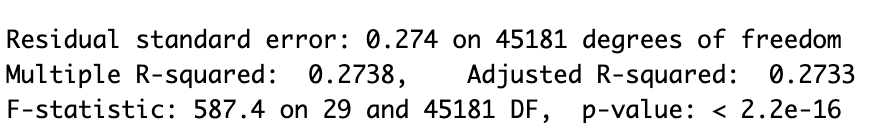
En primer lugar se entendió que la variable de respuesta que denominamos “y” se basaba si la persona hacía un depósito o no, estos valores registrados como 0 (fallido) y 1 (exitoso). En este orden de ideas, se decidió hacer un modelo de clasificación a base de redes neuronales para poder predecir dependiendo de diferentes factores si esa persona realizara un depósito. De esta manera posteriormente se podrían calcular los ingresos dependiendo de la matriz de confusión.

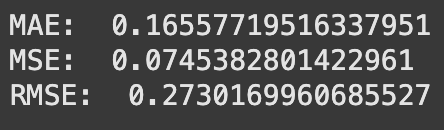
Para poder realizar este modelo inicialmente hay que determinar si todas las variables son relevantes a tener en cuenta y si considerar un alto número de variables resulta en un modelo significativo. Para ello, se decidió hacer una Regresión Lineal evaluando primeramente todas las variables para determinar si existe un buen modelo o hay variables a cambiar. Este análisis se hizo mediante R:



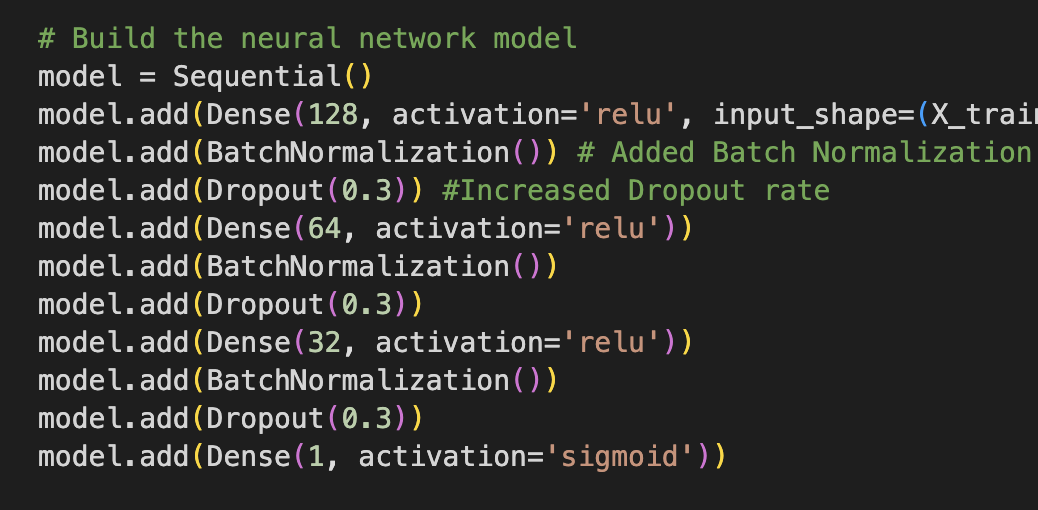


El resultado es un modelo significativo con un error bastante pequeño, al mismo tiempo se revisaron las variables significativas del modelo y al mismo tiempo se hizo la regresión en python y no dieron problemas ni de autocorrelación ni de normalidad en los residuos. Sin embargo, si encontramos un problema de multicolinealidad debido a la categoría de los meses, como todos pertenecen a una temporalidad diferente y tienen un efecto diferente sus varianzas son demasiado distintas. En este orden de ideas, llegamos a la siguiente idea, como se evaluaron por diferentes columnas los contactos por meses y todas mostraron ser significativas por excepción de Junio se decidió unirlas las columnas con una particularidad, esta columna se llamará contacted y describir si la persona fue contactada no necesariamente en un mes en especifico pero si serlo, por ende si la columna es 1 es que fue contactada en los últimos meses y 0 si no lo fue. Adicionalmente, se decidió eliminar variables que no eran significativas y que consideramos irrelevantes sobretodo para un personal en una compañía financiera ('previous', 'education\_unknown', 'education\_secondary', 'marital\_single'). Volvimos a evaluar el modelo tanto en R como en Python y obtuvimos unas estadísticas muy parecidas: Un modelo significativo, unos errores bastante bajos y

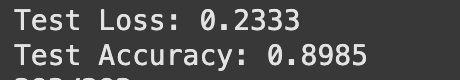


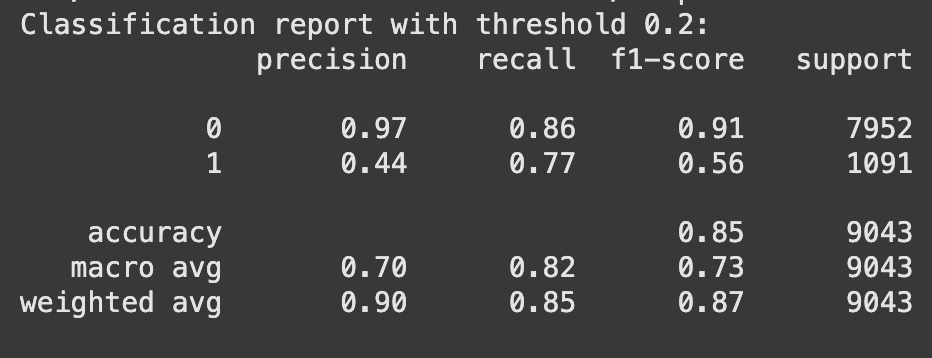


Pasado este punto y sabiendo qué modelo vamos a usar, se deicidio hacer un modelo predictivo de clasificación con redes neuronales con función de activación Sigmoide, de esta manera el modelo devolverá la probabilidad que tendría el usuario de realizar un depósito de acuerdo a varias variables.

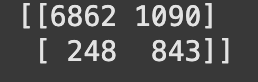


En esta instancia se decidió evaluar diferentes cantidades de umbrales para encontrar la más óptima, la forma en que decidimos la manera mas optima seria mediante una matriz de confusión e intentar maximizar en True Positive, por ello se hizo una búsqueda binaria y se determinó que el umbral que máxima el accuracy de nuestra variable de interés. Cabe recalcar que al evaluar el modelo el error esperado fue bastante pequeña y el nivel de accuracy fue del 90% es decir que nuestro modelo es bastante robusto.





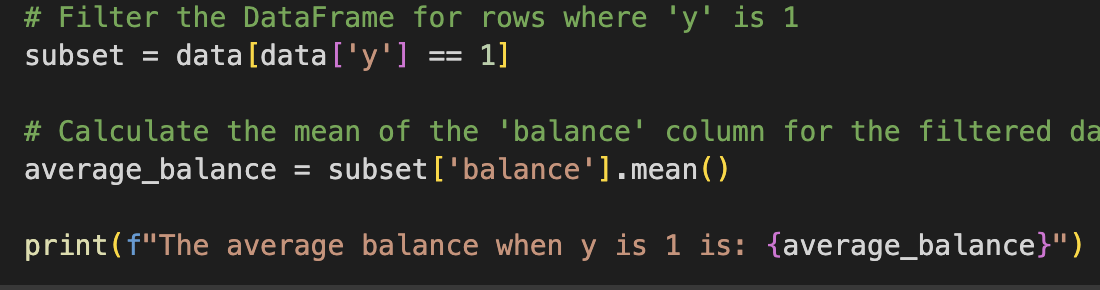
En esta imagen podemos ver que con un 77% evaluamos correctamente los valores de True Positive

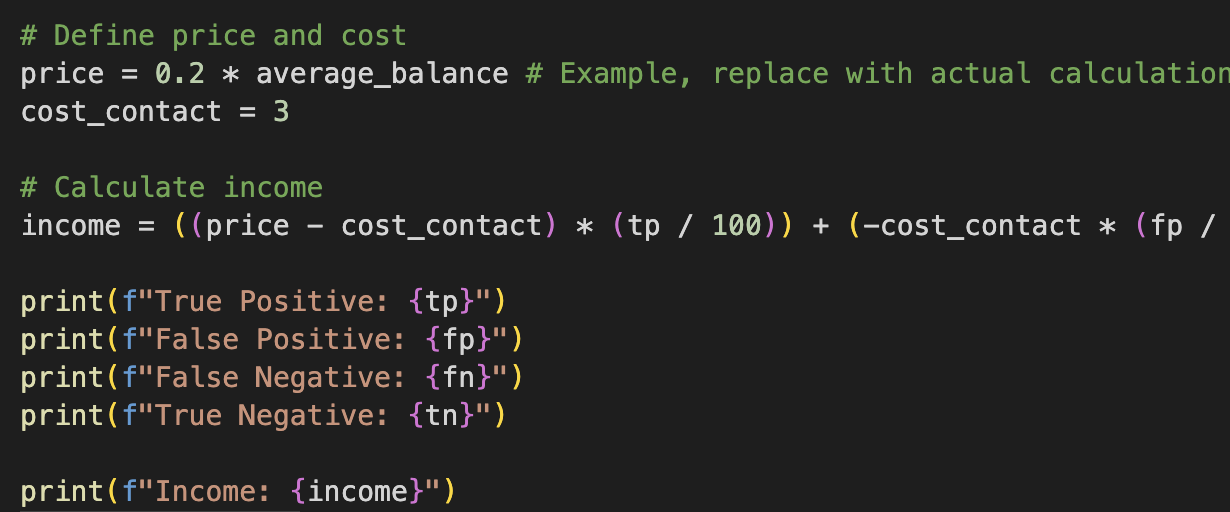


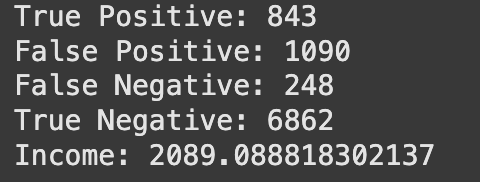
Matriz de confusión

Ahora bien, como nuestro cliente es el área comercial interesada en maximizar las ganancias utilizamos la matriz de confusión para realizar una fórmula para calcular una posible Utilidad. La idea es basar nuestros Utilidad en el valor de *TPx* *(Ingreso-Costo de contacto)- FPx*(*Costo de contacto*)- *FN*x(*Precio*). Para poder dar una conclusión teniendo en cuenta nuestros datos iniciales, necesitamos el valor de Ingresos y el de costos.

Se decidió hacer un supuesto para poder tener un valor numérico, para ingresos se decidió tomar una aproximado desde una fuente confiable como JP Morgan se sabe que un depósito se hará desde el saldo (balance) del usuario sin embargo este es una parte del mismo. Por ello los ingresos son un 20% del balance del usuario mientras que los costos dependen ya de una valor en específico y tomamos un aproximado de 2 euros por costo de contacto. En este orden de ideas, se escogieron los usuarios que realizaron el depósito es decir y == 1 y se promedió su balance para tener una base.

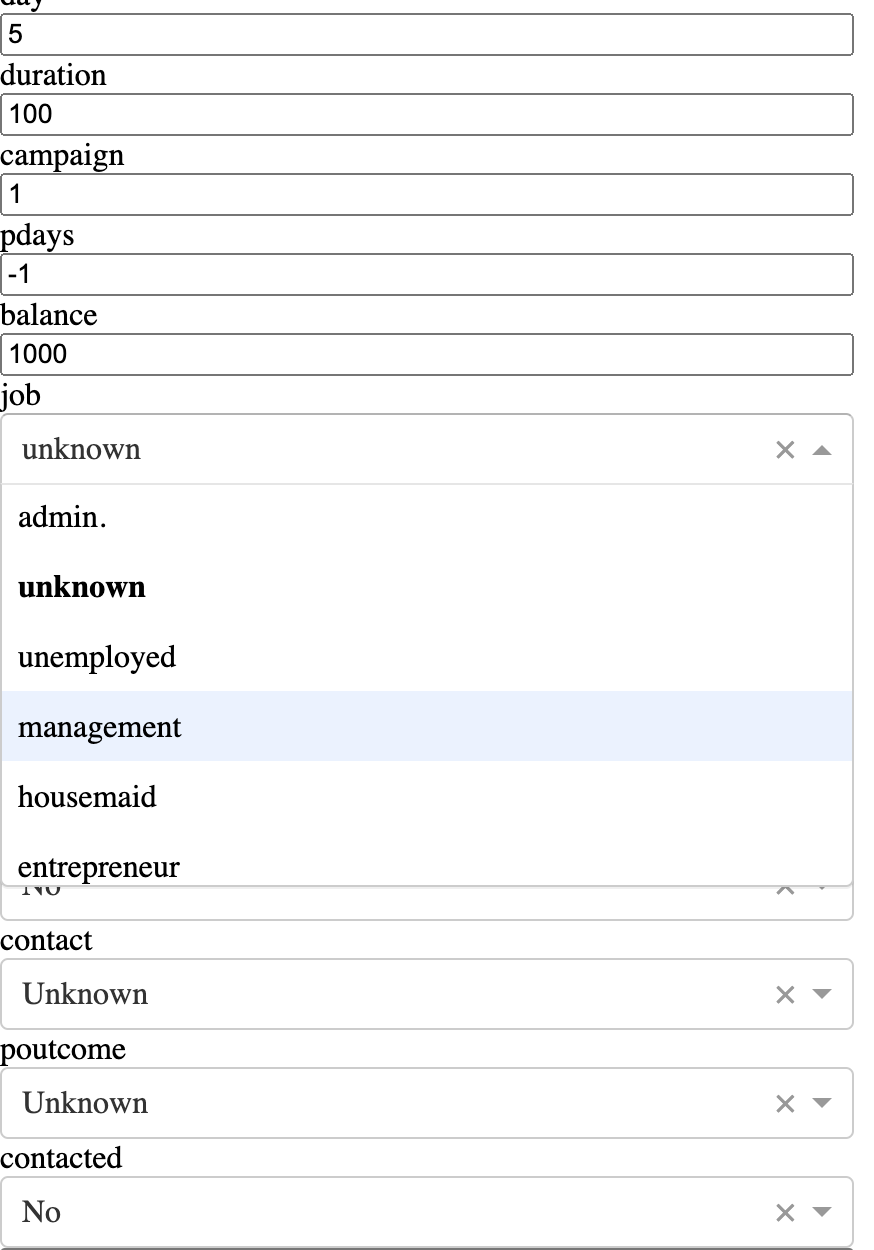






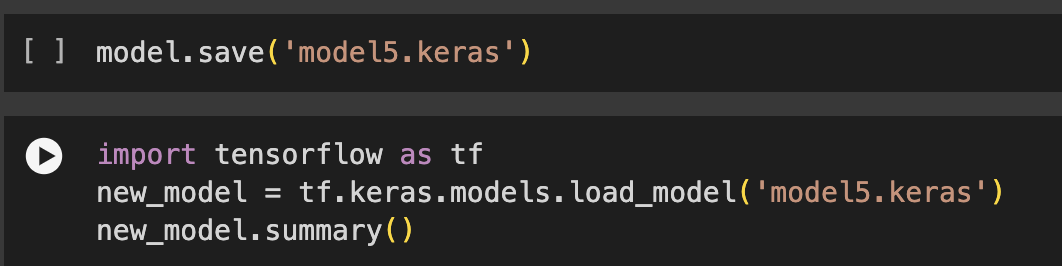
**5) Diseño y desarrollo del tablero**

Se decidió hacer un tablero en Dash que mostrará dos peculiaridades, la idea es que el usuario ingrese los parámetros de entrada según su información personal y que se devolviera la probabilidad y conclusión de que si se va a hacer un depósito. Adicionalmente, se usará el saldo (balance) del usuario para calcular el ingreso de este mismo según la matriz de confusión. Se desplegó por medio de casillas las diferentes posibilidad de las variables cómo trabajo para que apareciera la opción deseada al igual que cómo lo contactaron y demás.



**6)Despliegue**

Para poder desplegar el modelo y no tener que re-entrenar al momento de evaluar se realizó y se subió al código del Dash para que lo abriera directamente a través de la máquina virtual.



Al mismo tiempo para tener la base de datos iniciales se lanzó una instancia en AWS donde se incluirá los datos iniciales

Instancia para tablero Elástica