**Científico de Datos**

**COLFONDOS**

*Tiempo estimado de la prueba: 4 horas (puede disponer de más o menos horas si lo requiere)*

Para esta prueba utilizará como fuente la base de datos de ventas de una tienda de retail, la cual contiene información de ventas, compradores, devoluciones y gerentes de región. Los datos están disponibles en el Excel adjunto. También deberá utilizar para el punto del modelo predictivo las bases de datos de Afiliados Activos.txt y Afiliados Fugados.txt.

**PREPARACIÓN DE DATOS**

Utilizando SQL o Bigquery de GCP, usted debe:

**Realizar limpieza, validación y preparación de datos**

1. Tabla de Ventas (Orders):
   1. Cuando algún campo crítico es nulo, no debe ser incluido en reportes ni en las bases finales, pues la compañía lo considera como un error. Validar que los siguientes campos no estén nulos: Order Date, Ship Date, Customer ID.
   2. Si los campos Product ID, Category, Sub-Category o Product Name están vacíos, se debe cambiar el valor del campo a “Otros”.
   3. El valor en ‘Sales’ y en ‘Quantity’ nunca pueden ser negativos. Si fueran negativos, esas ventas se deben excluir.
   4. Los valores en los campos ‘Category’ y ‘SubCategory’ deben estar escritos siempre en mayúsculas (todas las palabras) para que coincidan con las registradas en la base de datos.
   5. Las ventas se cuentan siempre por el mes de la compra (order date), independientemente del día en que se realicen. Todas las fechas en la columna ‘order date’ deben ajustarse para que muestren el primer día del mes de la compra. (Ej: si la compra fue el 23/07/2010, la fecha debe quedar ajustada a 2010/07/01. Siempre en el primer día del mes). 🡪 El formato en el que se requiere la fecha es YYYY-mm-dd. No puede tener hora.
2. Tabla de clientes (Customers):
   1. Esta tabla no puede haber clientes repetidos. Se deben eliminar los duplicados.
   2. En el caso que haya un cliente repetido con segmentos/ciudad/estado/código postal/región diferentes; se debe tomar para estos campos aquella ciudad que aparezca más veces en la base de datos (moda).
   3. Al final sólo puede haber 1 línea por cliente (Customer ID, Customer Name).

**Entregables punto de preparación de datos:**

1. El output del punto anterior debe ser 1 tabla de SQL o Bigquery de GCP (**Tabla de SQL o Bigquery de GCP consolidada**) en donde estén unidas las tablas, para que en esa tabla quede la información de las ventas (Orders), toda la información del comprador (nombre, zona, ciudad, etc) de cada producto, el nombre del gerente de la zona a la que pertenece ese comprador y claramente identificados los artículos que fueron regresados a las tiendas, y adicionalmente se debe generar en la tabla anterior un campo de ingresos totales *(Sales x quantity).*

**MODELO PREDICTIVO**

La Dirección de Inteligencia de Negocios de Colfondos tiene a su cargo liderar estrategias analíticas que le permitan a la compañía mejorar sus indicadores estratégicos. Una de las variables con mayor relevancia para nosotros es la fuga de los clientes, la cual ha ido creciendo sostenidamente durante los últimos años.

Teniendo en cuenta lo anterior, la Dirección considera pertinente desarrollar modelos de propensión de fuga de cada uno de sus clientes, a fin de enfocar los esfuerzos de retención, por lo cual, usted deberá desarrollar un modelo de propensión de fuga que permita enfocar los esfuerzos de retención de la compañía.

Para llevar a cabo esta tarea usted dispone de la siguiente información:

**Afiliados Activos:** Contiene información histórica (2017-2020) de afiliados activos. Los afiliados activos son aquellos que sostienen una relación contractual con la compañía.

**Afiliados Fugados:** Contiene información histórica (2017-2020) de afiliados fugados. Los afiliados fugados son aquellos que han sido trasladados a otra administradora de fondos de pensiones (AFP), es decir, dichas personas no tienen una relación contractual con la compañía.

Glosario de Campos relacionados con las bases de datos a utilizar:

**ID:** código de identificación único del afiliado o cliente

**Periodo:** Periodo mensual para identificar si el cliente y/o afiliados se encuentra activo o fugado.

**Año:** Periodo anual para identificar si el cliente y/o afiliados se encuentra activo o fugado.

**Estado:** Denota el estado del afiliado con la compañía. Para este caso solo existirán 2 estados, activos y fugados. Un cliente solo puede tener un estado.

**Género:** Clase o tipo a que pertenece una persona.

**Edad:** Tiempo que ha vivido una persona desde su nacimiento

**IBC:** El IBC significa Ingreso Base de Cotización y es el valor que se debe tomar como base para calcular la seguridad social y los aportes parafiscales de una persona, sea dependiente o independiente. Comúnmente se identifica como el salario de una persona.

**Ciudad:** Corresponde a la ciudad de residencia del cliente o afiliado.

**Origen Afiliado:** Corresponde a la procedencia del afiliado, es decir, si un afiliado tiene la marca “Vinculación Inicial” indica que el cliente fue ingresado por primera vez a una administradora de fondo de pensiones (AFP) por efecto de su primer trabajo. Si un cliente tiene la marca “Traslado de AFP” indica que el cliente fue trasladado de una AFP a otra AFP.

**Origen AFP:** Denota a que AFP pertenecía el afiliado luego de haberse traslado de una AFP a otra AFP.

**Nivel de Satisfacción:** Calificación sobre la percepción de servicios recibidos por parte del afiliado o cliente. La calificación varia de 1 a 10, siendo 10 la máxima calificación y 1 la mínima calificación.

**PQR:** Denota si el afiliado ha realizado alguna Petición, Queja o Reclamo.

**Entregables punto de modelo predictivo:**

1. Una presentación en la que se muestre lo siguiente:
   1. Análisis previo de variables.
   2. Modelo desarrollado, validación de supuestos y una breve discusión del porqué eligió este modelo.
2. Código en Bigquery de GCP, SQL, R o Python con el procesamiento previo de la información.
3. Código en Bigquery de GCP, SQL, R o Python con los resultados del modelo, este código debe correr y los resultados deben coincidir con los expuestos en la presentación.

**ENTREVISTA**

Una vez finalizadas las dos secciones anteriores se programará una entrevista para realizar algunas preguntas con relación al proceso realizado. La entrevista podrá durar aproximadamente 1 hora en la cual **15 minutos** serán para preguntas relacionadas con esta prueba.

**Los entregables de esta prueba son:**

1. **1 Tabla de SQL o Bigquery de GCP Consolidada** (Después de Limpieza, Validación, Preparación y unión con las demás tablas) y con las **tablas adicionales solicitadas** en la sección de “**Entregables punto de preparación de datos**”. 🡪 Adjuntar igualmente las **querys realizadas en SQL o Bigquery** para llevar a cada una de las tablas solicitadas.
2. **1 Código Bigquery de GCP, SQL, R o Python** con el procesamiento previo de la información del modelo predictivo y el modelo predictivo.
3. **1 Presentación** con el análisis previo de variables, el modelo desarrollado, la validación de supuestos y una breve discusión del porqué eligió este modelo.
4. **Resolución de dudas o inquietudes** en la entrevista en el espacio de 15 minutos.