Preparación de los datos

Tabla de contenido

Introducción

El presente documento tiene como finalidad la documentación del proceso y el código generados para la preparación de los datos en el marco de la prueba técnica solicitada.

A continuación se listan y se explican todos los procedimientos llevados a cabo sobre los datos con el fin de prepararlos para la construcción del modelo.

El presente proyecto se construye en su totalidad utilizando herramientas de libre distribución del software R.

Configuración

La configuración del entorno de trabajo adecuado es necesaria para el correcto desarrollo del código que realizará al tarea. En este primer capítulo se requieren herramientas de procesamiento de datos y de programación básica.

Paquetes

La configuración del entorno en cuanto a software se realiza por medio del aprovisionamiento de paquetes. Todos los paquetes utilizados se encuentran disponibles bajo licencias de software libre, la mayoría de estos paquetes ha sido desarrollado por la empresa RStudio, cuya labor a nivel mundial es reconocida como pionera en este campo.

```
library("readr")
library("dplyr")
library("tidyr", exclude = "extract")
library("magrittr")
```

Carga de datos

Los datos son cargados y examinados para su posterior transformación.

```
read_csv(
  "01_datos/pcac_encuesta.csv"
  ) -> pcac_encuesta
read_csv(
  "01_datos/pcac_oportunidades_comer.csv"
  ) -> pcac_oportunidades_comer

read_csv(
  "01_datos/pcac_mac_gpi_clientes.csv"
  ) -> pcac_mac_gpi_clientes

read_csv(
  "01_datos/pcac_capacidad_gerentes.csv"
  ) -> pcac_capacidad_gerentes

read_csv(
  "01_datos/pcac_mac_gpi_ecas.csv"
  ) -> pcac_mac_gpi_ecas

read_csv(
  "01_datos/pcac_mac_gpi_tenencia_prod.csv"
  ) -> pcac_mac_gpi_tenencia_prod

read_csv(
```

```
  "01_datos/pcac_planta_comercial2.csv"
  ) -> pcac_planta_comercial2
```

Los conjuntos de datos son los siguientes:

pcac_mac_gpi_clientes: Contiene la información de los clientes del segmento específico que estamos trabajando, la información de su zona y así mismo, el gerente y ejecutivo asignados, así mismo cuenta con la categorización comercial de inversión (A, B o C) y un score de cliente (mientras más alto mejor) que indica la deseabilidad comercial de este. Este score es generado en un proceso independiente a este modelo.

```
pcac_mac_gpi_clientes %>% glimpse

## Rows: 34,145
## Columns: 38
## $ num_doc_cli             <dbl> 7.929490e+17, 3.972598e+18, 2.753978e+18,
2.89…
## $ cod_tipo_doc_cli        <dbl> 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1,
1…
## $ num_doc_cli_dv          <dbl> 7.929490e+17, 3.972598e+18, 2.753978e+18,
2.89…
## $ cli_val                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0…
## $ cli_pan                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0…
## $ cli_per                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0…
## $ num_doc_gte_inv         <dbl> 1.872057e+18, 1.872057e+18, 1.872057e+18,
1.87…
## $ nombre_gte_inv          <chr> "3684e207099dc9903324",
"3684e207099dc9903324"…
## $ cod_gte_inv             <dbl> 8.648915e+18, 8.648915e+18, 8.648915e+18,
8.64…
## $ ciudad_gte_inv          <chr> "SANTA MARTA", "SANTA MARTA", "SANTA MARTA",
"…
## $ region_gte_inv          <chr> "CARIBE", "CARIBE", "CARIBE", "CARIBE",
"CARIB…
## $ cod_region_gte_inv      <dbl> 3, 3, 3, 3, 3, 3, 3, 4, 3, 3, 3, 5, 3, 3, 3,
4…
## $ nombre_gte_regional_inv <chr> "9e68ac4e7387674d5115",
"9e68ac4e7387674d5115"…
## $ estado_gte_inv          <chr> "ACTIVO", "ACTIVO", "ACTIVO", "ACTIVO",
"ACTIV…
## $ cod_ejec_bco            <dbl> 5.011586e+18, 5.011586e+18, 5.011586e+18,
5.01…
## $ nombre_ejec_bco         <chr> "613eaa48144480e766d8",
"613eaa48144480e766d8"…
## $ cod_suc_ejec_bco        <dbl> 0, 0, 0, 0, 0, 0, 0, 68, 0, 0, 0, 74, 0, 0,
0,…
## $ cod_dane                <dbl> NA, NA, NA, NA, NA, NA, NA, 73001, NA, NA,
NA,…
## $ ciudad                  <chr> NA, NA, NA, NA, NA, "Medellin", NA, "IBAGUE",
…
```

```
## $ region_ejec_bco_1       <chr> "CARIBE", "CARIBE", "CARIBE", "CARIBE",
"CARIB…
## $ cod_region_ejec_bco_1   <dbl> 3, 3, 3, 3, 3, 1, 3, 4, 3, 3, 3, 5, 3, 3, 3,
1…
## $ region_ejec_bco         <chr> "CARIBE", "CARIBE", "CARIBE", "CARIBE",
"CARIB…
## $ cod_region_ejec_bco     <dbl> 3, 3, 3, 3, 3, 3, 3, 4, 3, 3, 3, 5, 3, 3, 3,
3…
## $ gerenciamiento_bco      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
1…
## $ cod_ciiu                <dbl> 8512, 10, 10, 8512, 10, 90, 90, 2012, 10,
8512…
## $ ciiu                    <chr> "EDUCACIÓN PREESCOLAR", "ASALARIADOS",
"ASALAR…
## $ segm_cli                <chr> "PREFERENCIAL", "PREFERENCIAL",
"PREFERENCIAL"…
## $ subsegm_cli             <chr> "PREFERENCIAL PLUS", "PREFERENCIAL PLUS",
"PRE…
## $ marca_lista_ctrl        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, NA, 1,
…
## $ cod_calif_riesgo_bco    <chr> "G1", "G1", "G1", "G1", "G1", "G1", "G2", NA,
…
## $ perfil_riesgo_val       <chr> "MODERADO", "MODERADO", "CONSERVADOR",
"AGRESI…
## $ riesgo_sarlaft_bco      <chr> "RIESGO MUY BAJO", "RIESGO BAJO", "RIESGO MUY
…
## $ marca_mac_inv2          <chr> "B", "B", "B", "B", "C", "C", "B", "B", "A",
"…
## $ score_modelo2           <dbl> 0.19, 0.06, 0.12, 0.09, 0.03, 0.03, 0.06,
0.00…
## $ banca_gte_cons          <chr> "PREFERENCIAL", "PREFERENCIAL",
"PREFERENCIAL"…
## $ cliente_nuevo           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0…
## $ marca_mac_inv           <chr> "B", "B", "B", "B", "C", "C", "B", "B", "A",
"…
## $ score_modelo            <dbl> 0.19, 0.06, 0.12, 0.09, 0.03, 0.03, 0.06,
0.00…
```

pcac_mac_gpi_ecas: Especifica la relación del equipo comercial, es decir, información de los gerentes y sus ejecutivos asociados.

```
pcac_mac_gpi_ecas %>% glimpse

## Rows: 392
## Columns: 23
## $ num_doc_gte_inv          <dbl> 3.145651e+18, 5.317036e+18, 6.267087e+18,
6.…
## $ cod_gte_inv              <dbl> 2.503505e+18, 2.733425e+18, 3.068611e+18,
3.…
## $ nombre_gte_inv           <chr> "7488028053aeb8b71806",
"b7a18c8d2d7dcb31efb…
## $ num_doc_ejec_bco         <dbl> 2.982694e+18, 5.925214e+18, 5.015768e+18,
9.…
## $ cod_ejec_bco             <dbl> 2.913003e+18, 3.658608e+18, 7.943928e+18,
```

```
4.…
## $ nombre_ejec_bco           <chr> "faea4c01880c19284bf8",
"0b336be9d541a25d7e0…
## $ cod_sucursal              <dbl> 728, 0, NA, NA, NA, NA, NA, NA, 197, 0, 0,
1…
## $ cod_dane                  <dbl> 76147, NA, NA, NA, NA, NA, NA, NA, 20001,
NA…
## $ ciudad                    <chr> "CARTAGO", NA, NA, NA, NA, NA, NA, NA,
"VALL…
## $ num_doc_gte_regional      <dbl> 6.942089e+18, 7.190493e+18, 6.942089e+18,
6.…
## $ cod_gte_regional          <dbl> 1.889575e+18, 6.556604e+18, 1.889575e+18,
1.…
## $ nombre_gte_regional       <chr> "55822639df68182d9945",
"407943d753237da8b52…
## $ desc_tipo_oficial_gte_bco <chr> "EJECUTIVO SENIOR CENTRALIZADO", "EJECUTIVO
…
## $ region_gte_inv            <chr> "SUR", "ANTIOQUIA", "SUR", "SUR", "SUR",
"SU…
## $ cod_region_gte_inv        <dbl> 5, 1, 5, 5, 5, 5, 5, 4, 3, 3, 3, 1, 1, 1,
2,…
## $ ciudad_gte_inv_1          <chr> "PEREIRA", "MEDELLIN", "CALI", "CALI",
"CALI…
## $ region_ejec_bco           <chr> "SUR", "ANTIOQUIA", NA, NA, NA, NA, NA, NA,
…
## $ cod_region_ejec_bco       <dbl> 5, 1, NA, NA, NA, NA, NA, NA, 3, 3, 3, 1,
1,…
## $ ciudad_ejec_bco_1         <chr> "CARTAGO", "CALDAS", NA, NA, NA, NA, NA,
NA,…
## $ ciudad_gte_inv            <chr> "pereira", "medellin", "cali", "cali",
"cali…
## $ ciudad_ejec_bco           <chr> "cartago", "caldas", NA, NA, NA, NA, NA,
NA,…
## $ cod_dane_gte              <dbl> 66001, 5001, 76001, 76001, 76001, 76001,
760…
## $ cod_dane_ejec_bco         <dbl> 76147, 5129, NA, NA, NA, NA, NA, NA, 20001,
…
```

pcac_oportunidades_comer: Contiene las oportunidades comerciales que cada cliente posee. Estas son generadas por un proceso independiente a este modelo.

```
pcac_oportunidades_comer %>% glimpse

## Rows: 247,863
## Columns: 7
## $ num_doc_cli         <dbl> 1.213306e+18, 5.494563e+18, 6.157656e+18,
2.357349…
## $ cod_tipo_doc_cli    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,…
## $ id_oport            <chr> "T-RF-C-5", "F-RV-C-8", "F-RF-C-8", "F-RV-C-8",
"F…
## $ cat_oport           <chr> "Renta Fija", "Renta Variable", "Renta Fija",
"Ren…
## $ nombre_activo_oport <chr> "Compra TES TF 27", "Compra FIC Renta Alta
Convicc…
```

```
## $ cod_producto        <dbl> 23, 16, 23, 16, 23, 23, 12, 9, 23, 23, 14, 12,
14,…
## $ producto            <chr> "tiene_rta_fija_valores",
"tiene_rta_alta_convicci…
```

pcac_mac_gpi_tenencia_prod: Indica para cada cliente si ya posee un producto y de igual manera si lo utiliza.

```
pcac_mac_gpi_tenencia_prod %>% glimpse

## Rows: 66,802
## Columns: 7
## $ num_doc_cli      <dbl> 4.295825e+18, 4.783457e+18, 7.233170e+18,
4.561384e+1…
## $ cod_tipo_doc_cli <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
3,…
## $ num_doc_cli_dv   <dbl> 4.295825e+18, 4.783457e+18, 7.233170e+18,
4.561384e+1…
## $ cod_prod         <dbl> 33, 33, 33, 33, 23, 12, 4, 4, 6, 27, 27, 5, 6, 27,
16…
## $ nombre_prod      <chr> "tiene_cuenta_ahorros_fe",
"tiene_cuenta_ahorros_fe",…
## $ tenencia         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,…
## $ usa_producto     <dbl> NA, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
0…
```

pcac_encuesta: Define los valores de una encuesta realizada a los distintos equipos comerciales en todo el país donde se definen los tiempos que dedican a las distintas actividades en cada producto y la cantidad de veces que lo realizan en el año.

```
pcac_encuesta %>% glimpse

## Rows: 1,804
## Columns: 10
## $ nombre_comercial                    <chr> "30de3fa56a2819285bb8",
"fcfc374…
## $ cedula_comercial                    <dbl> 4.035832e+18, 8.848587e+18,
2.06…
## $ codigo_de_vendedor                  <dbl> 2.727520e+18, 7.191292e+18,
1.02…
## $ cod_producto                        <dbl> 100, 100, 100, 100, 100, 100,
10…
## $ producto                            <chr> "Vinculaciones Bancolombia
Panam…
## $ tipo_de_solicitud                   <chr> "Vinculación a Bancolombia
Panam…
## $ categoria_cliente                   <chr> NA, NA, NA, NA, NA, NA, NA, NA,
…
## $ etapa_del_producto                  <chr> "Venta", "Venta", "Venta",
"Vent…
## $ total_promedio_volumen_por_semana   <dbl> NA, NA, NA, NA, NA, NA, NA, NA,
…
```

```
## $ total_promedio_tiempo_min_x_actividad <dbl> 50, 90, 90, 90, 90, 15, 20,
120,…
```

pcac_capacidad_gerentes: establece el tiempo de atención disponible que tienen los gerentes para realizar sus actividades comerciales.

```
pcac_capacidad_gerentes %>% glimpse

## Rows: 50
## Columns: 5
## $ cod_gte_inv          <dbl> 6.348457e+18, 5.892352e+18, 3.418692e+18,
2.72752…
## $ cod_region_gte_inv   <dbl> 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2…
## $ sistematica_anual    <dbl> 85050, 85050, 85050, 85050, 85050, 85050, 85050,
…
## $ tiempo_instrum_resta <dbl> 9120, 9120, NA, 9120, 9120, 9120, 9120, 9120,
912…
## $ tiempo_restante      <dbl> 75930, 75930, 75930, 75930, 75930, 75930, 75930,
…
```

pcac_planta_comercial2: Especifica detalles de la planta comercial (Gerentes de inversión).

```
pcac_planta_comercial2 %>% glimpse

## Rows: 50
## Columns: 11
## $ num_doc_gte_inv         <dbl> 1.891524e+18, 5.317036e+18, 3.322303e+18,
5.43…
## $ nombre_gte_inv          <chr> "ccd630b17e7511bb06a9",
"b7a18c8d2d7dcb31efb6"…
## $ receptor_valores        <chr> "c7d98e8c39b09959d75f",
"1d9c62f429857f3d0ca4"…
## $ cod_gte_inv             <dbl> 8.760414e+18, 2.733425e+18, 6.462951e+17,
5.25…
## $ ciudad_gte_inv          <chr> "BOGOTÁ", "MEDELLIN", "MEDELLIN", "BOGOTÁ",
"M…
## $ region_gte_inv          <chr> "BOGOTÁ", "ANTIOQUIA", "ANTIOQUIA", "BOGOTÁ",
…
## $ cod_region_gte_inv      <dbl> 2, 1, 1, 2, 1, 4, 4, 1, 1, 4, 3, 3, 1, 1, 2,
1…
## $ banca_gte_inv           <chr> "PREFERENCIAL", "PREFERENCIAL",
"PREFERENCIAL"…
## $ cargo_gte_inv           <chr> "GERENTE COMERCIAL PREFERENCIAL", "GERENTE
COM…
## $ cod_sap_gte_inv         <chr> "31897763384400dd67a6",
"3774031a39034848e1e6"…
## $ nombre_gte_regional_inv <chr> "ac78bcc3f95093b6b918",
"407943d753237da8b527"…
```

Procesamiento

El procesamiento de los datos se documenta a continuación. Este procesamiento se lleva a cabo con el objeto de conocer los requerimientos de tiempo de los ejecutivos y el tiempo disponible de los gerentes. Estos tiempos son el insumo principal para el modelo.

Tiempo por producto

En primera instancia es necesario conocer el tiempo requerido para las labores de cada producto. estos tiempos se encuentran dentro de los datos de la encuesta realizada a vendedores. Es posible obtener los tiempos mencionados utilizando el siguiente código.

```
pcac_encuesta %>%
  group_by(cod_producto) %>%
  summarise(
     tiempo_x_producto =
       mean(total_promedio_tiempo_min_x_actividad, na.rm = TRUE)
  ) %>%
  filter(
    tiempo_x_producto != 0
    ) -> tabla_producto_tiempo
```

El código anterior realiza las siguientes acciones:

- Agrupa los datos por producto.
- Obtiene el tiempo promedio por cada producto.
- Filtra los productos cuyos tiempos sean iguales a cero.

Tiempo por cliente

En este punto el objetivo es conocer los tiempos requeridos por cliente. Tomando en consideración la base de datos de oportunidades comerciales, es posible establecer el tiempo requerido para prestar atención a las necesidades de cada cliente. Cada oportunidad está vinculada a un producto, de manera que el tiempo por cliente está dado por la suma de los tiempos por producto.

```
pcac_oportunidades_comer %>%
  left_join(tabla_producto_tiempo) %>%
  mutate(
    tiempo_x_producto = replace_na(tiempo_x_producto,
mean(tiempo_x_producto, na.rm = TRUE))
  ) %>%
  group_by(num_doc_cli) %>%
  summarise(
```

```
    tiempo_x_cliente =
       sum(tiempo_x_producto, na.rm = TRUE)
  ) -> tabla_cliente_tiempo
```

El código anterior realiza las siguientes tareas:

- Cruza la tabla de oportunidades comerciales con la tabla de tiempo por producto.
- Para los productos que no tienen datos de tiempo, los imputa usando el valor medio de aquellos que sí tienen.
- Agrupa la base de datos por cliente.
- Obtiene los tiempos de cada cliente sumando los tiempos de sus productos.

Tiempo de cada ejecutivo

El tiempo requerido por cada ejecutivo se obtiene a partir de la suma de los tiempos de atención requeridos por sus clientes. Este cálculo se lleva a cabo de la siguiente manera.

```
pcac_mac_gpi_clientes %>%
  left_join(tabla_cliente_tiempo) %>%
  mutate(
    tiempo_x_cliente = replace_na(tiempo_x_cliente,
mean(tiempo_x_cliente, na.rm = TRUE))
  ) %>%
  group_by(cod_ejec_bco) %>%
  summarise(
    marca_a = sum(as.numeric(marca_mac_inv == "A")),
    marca_b = sum(as.numeric(marca_mac_inv == "B")),
    tiempo_x_ejecutivo =
       sum(tiempo_x_cliente, na.rm = TRUE),
    cod_region_ejec_bco = median(cod_region_ejec_bco),
    clientes = n()
  ) -> tabla_ejecutivo_tiempo_region_marca
```

Por medio de este código se efectúan las siguientes tareas.

- Cruzar la tabla de clientes con la tabla de tiempo por cliente.
- Imputar usando el valor medio del tiempo por cliente para los clientes que no presentan este dato.
- Agrupar los datos por ejecutivo.
- Obtener los tiempos de cada ejecutivo sumando los tiempos de sus clientes.
- Obtener la cantidad de clientes del grupo A por ejecutivo.

- Obtener la cantidad de clientes del grupo B por ejecutivo.
- Obtener la cantidad de clientes por ejecutivo.
- obtener la región de cada ejecutivo.

Tiempo por gerente

El tiempo disponible de cada gerente de inversión está consignado en la tabla correspondiente. Es necesario revisar si el gerente se encuentra activo o retirado. También es pertinente tomar en consideración que la tabla de clientes proporciona información al respecto. En este sentido se toma la decisión de utilizar ambas tablas.

```
pcac_mac_gpi_clientes %>%
  filter(estado_gte_inv == "ACTIVO") %>%
  select(cod_gte_inv, num_doc_gte_inv,  cod_region_gte_inv) %>%
  unique %>%
  left_join(pcac_capacidad_gerentes) %>%
    mutate(
    tiempo_restante = replace_na(tiempo_restante, mean(tiempo_restante,
na.rm = TRUE)),
    sistematica_anual = NULL,
    tiempo_instrum_resta = NULL,
  ) -> tabla_gerente_tiempo_region
```

El código presentado permite:

- Filtrar los genrentes activos en la tabla de clientes.
- Seleccionar las columnas correspondientes al código del gerente y a su región.
- Retirar los registros repetidos, creando así una primera base de gerentes.
- Cruzar la base resultante con la base de datos de gerentes.
- Imputar usando el valor medio del tiempo por gerente para los gerentes que no presentan este dato.
- Retirar los campos de sistemática y tiempo de instrumentación.

Escritura

Finalmente se realiza la escritura de las bases de datos de interés que serán trabajadas en otro archivo.

```
write_rds(tabla_ejecutivo_tiempo_region_marca,
"01_datos/tabla_ejecutivo_tiempo_region_marca.rds")

write_rds(tabla_gerente_tiempo_region,
"01_datos/tabla_gerente_tiempo_region.rds")
```