**INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA**
**ESCUELA DE POSTGRADO**

# Aplicación de Técnicas de Minería en el proceso de Cobranza

**AUTOR: Comunello de Sá, Fellippe (Leg. Nº 104018)**

**TUTOR: Dra. Leticia Gómez**

**TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIAS DE DATOS.**

**BUENOS AIRES**

**SEGUNDO CUATRIMESTRE, 2019**

Índice

1. Introducción ..... 1
2. Marco Teórico ..... 2
   - Fechas de Gestión, ..... 3
   - Personal Disponible, ..... 3
3. Definición del Problema ..... 5
4. Hipótesis ..... 6
5. Objetivo ..... 6
   - Objetivo General ..... 6
   - Objetivos Específicos ..... 6
6. Metodología ..... 7
   - ETL ..... 7
   - Minería de Datos ..... 7
   - Modelo de decisión ..... 7
   - SAS ..... 8
7. Datos ..... 9
   - Fuentes principales de información ..... 9
   - Tratamiento de datos ..... 10
8. Modelo ..... 16
   - Modelo Analítico ..... 16
   - Resultados Obtenidos ..... 17
9. Prueba inicial del Modelo ..... 20
   - Muestra de Prueba ..... 20
   - Conceptos de Prueba ..... 20
   - Resultados Obtenidos ..... 21
10. Prueba del Modelo en ambiente productivo ..... 22
   - Muestra de Prueba ..... 22
   - Conceptos de Prueba ..... 23
   - Resultados Obtenidos ..... 23
11. Conclusión ..... 24
12. Referencias ..... 25

# 1. Introducción

La cobranza es un importante servicio prestado por las empresas que maneja a los clientes morosos. Es un proceso estratégico y clave para generar valor a un rango de clientes y el camino inicial para alguna posible recuperación judicial. Cobranzas es un área dentro de una organización cuyo objetivo es convertir posibles pérdidas en posibles ingresos, utilizando el contacto como herramienta para avisar o revisar la “necesidad” de cumplimento de su obligación o deuda.

En el proceso de gestión existen varias formas y tácticas para alcanzar el contacto con el cliente, tales como: cartas, llamadas telefónicas, mensajes al celular o presencial. El método más difundido y donde se presenta una mejor respuesta es vía telefónica, donde un cobrador, pudiendo ser un empleado/a de la empresa o un tercer agente, habla con el cliente intentando dar soporte y medios para la cancelación de la deuda. Junto con esa interacción se toman notas del contacto para posibles interacciones futuras.

Las grandes empresas, usando bancos como base principal de referencia, necesitan de grandes áreas de cobranza para atender un variado público de clientes. Cuentan con un proceso bastante interactivo para llegar al cliente, siendo soportados por sistemas de llamadas automáticas para una mayor performance. Esos sistemas son esenciales, ya que el volumen de llamadas necesarias para intentar entrar en contacto con todos los clientes de la cartera es muy alto y sería imposible hacerlo manualmente.

Cobranzas es un módulo esencial para mantener la integridad del ciclo del negocio/Crédito, siendo el puente para el mantenimiento de clientes existentes y futuros.

## 2. Marco Teórico

El estudio a ser realizado tiene como referencia un área de cobranza de clientes morosos de una institución bancaria donde la misma tiene como objetivo comunicar, vía teléfono o carta, la necesidad de regularización de su situación financiera morosa. Esta área cuenta con 110 operadores trabajando de lunes a sábado, desde las 9 hasta las 20, distribuidos en 3 turnos de 5 horas.

El objetivo del área es asegurar que el flujo de clientes morosos se mantenga estable, para ello utiliza la comunicación preventiva y de mantenimiento para recolectar promesas de pago.

Existe un alto volumen de clientes entrando en mora, debido a la cantidad de clientes con que cuenta la institución. Por eso, es necesario operar con sistemas de apoyo para asegurar la ejecución de un gran volumen de llamadas.

Las sub-áreas vitales del sector para alcanzar ese número de llamadas están constituidas por 3 frentes principales:

Planeamiento ⇨ Discador ⇨ Operador

Planeamiento – Analiza y establece objetivos en base al riesgo tomado por la institución. Utiliza las informaciones generadas en el área, del sistema de registro y productos para mejorar las estrategias y definir, alineados con los intereses de la institución, cuál será la gama de clientes a trabajar en el día.

Discador – Realiza y controla las llamadas. Cuenta con un sistema automático para realizar los llamados donde detectando una voz humana, la direcciona a un operador, entrega un mensaje de voz virtual y realiza registro de todas las interacciones del operador con el sistema.

Operador – Realiza el contacto directo con el cliente cuando es establecida una conexión telefónica. Intenta obtener promesas de pago, vender herramientas de solución y realiza anotaciones de todos los detalles de la llamada.

Existen algunas otras áreas que forman parte del ciclo de cobranza como: Calidad, Pre Legal, Legal y Operaciones. Como no están en el ciclo del estudio, no serán detalladas.

Utilizando los datos de cobranza, el área de riesgo genera el Score (puntuación). Un score bajo, es considerado un riesgo, es decir, un mal pagador. Diariamente, el sistema de gestión de clientes selecciona esos clientes malos y que estén entre 3 y 180 días de mora. Junto con la puntuación del Score se crea un plan: a qué cliente hay que llamar primero, que estarán en un listado que llamaremos de "File Diario" o FD, y a cuales hay que postergar a una cierta cantidad de días, entrando así en la fila de espera.

La cantidad de clientes en el File diario (FD) depende de algunas variables, siendo las más importantes:

- Fechas de Gestión,

Los días a trabajar del mes son divididos en dos categorías, las fechas donde existe un gran volumen de vencimientos y las fechas con un flujo normal. Por ejemplo, la caída de vencimiento de la tarjeta de crédito, donde el volumen crece significativamente contra el volumen de cuentas que están evolucionando en meses posteriores al advenimiento de la mora, corresponde a la primera categoría.

- Personal Disponible,

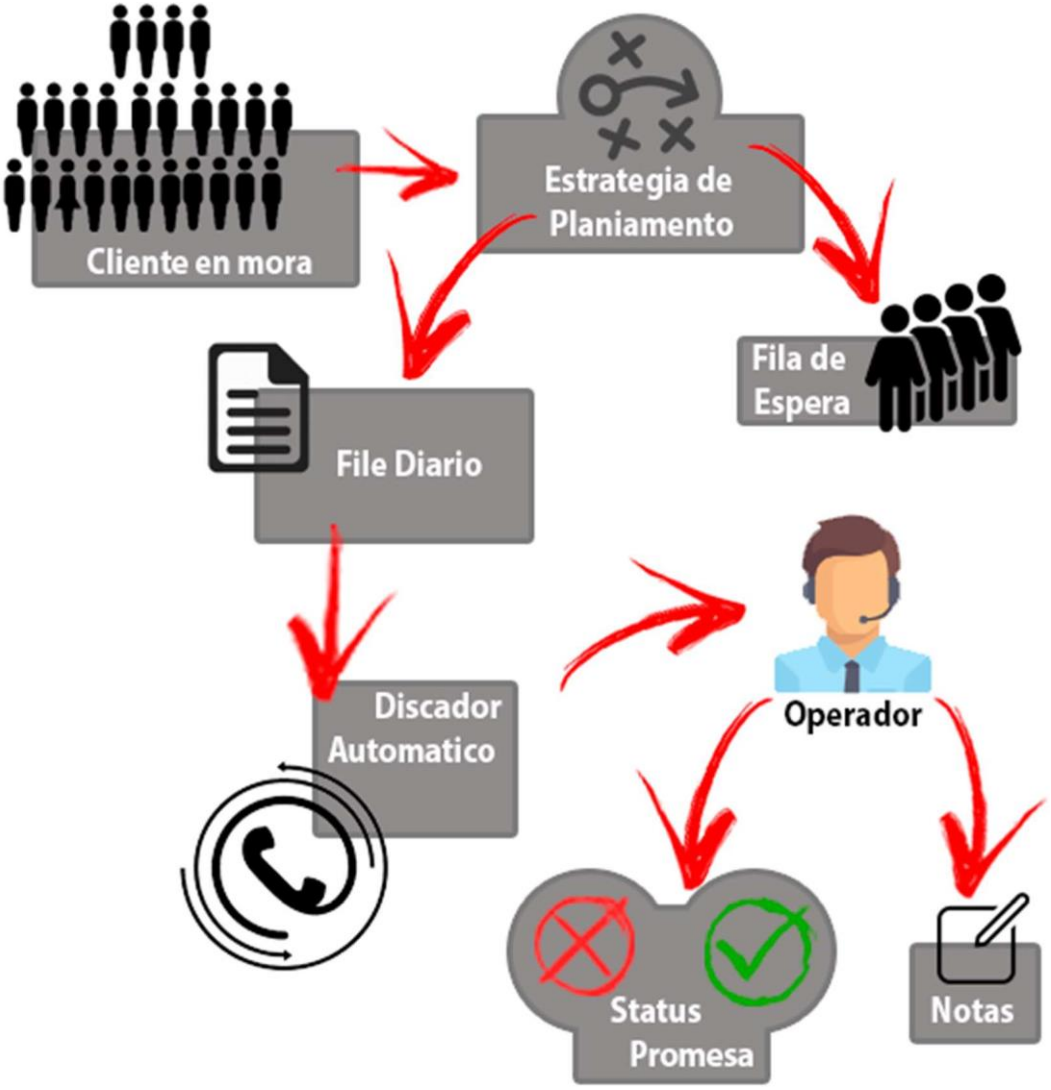
Todos los años es definida la cantidad de personal que estará disponible para realizar llamadas. Ese número define la cantidad de clientes que será posible contactar en el día.

El sistema de llamadas, el Discador, utilizando el FD comienza los llamados buscando establecer una conexión con el número. Si considera un discado completo cuando se establece la conexión física. La demanda de números discados depende directamente de la cantidad de discados completos sobre la cantidad de clientes en el FD, así aumentando o disminuyendo la cantidad de números discado según la cantidad de operadores disponible para contestar las llamadas realizadas.

En el caso de éxito ante la existencia de voz humana, tenemos una conexión y la llamada es direccionada a un Operador disponible, teniendo la posibilidad de hablar con el cliente u otra persona que conoce al cliente del número llamado.

Teniendo en cuenta todo esto, es posible comprender el objetivo diario establecido por Planeamiento: realizar un discado completo para cada uno de los clientes del FD. Como en general no se logra este objetivo recorriendo los clientes del FD, se realiza hasta el máximo de 3 intentos para cada uno de ellos.

El flujo actual de trabajo puede ser visto en la Figura 1.



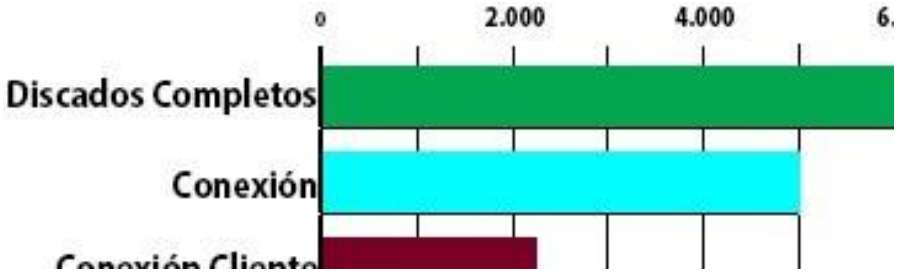

(Figura 1) – Flujo de trabajo actual.

Fuente: Elaboración propia.

## 3. Definición del Problema

El Discador realiza llamadas solamente siguiendo una nómina de clientes del FD de forma secuencial. En días normales el objetivo de discados completos es alcanzado. Sin embargo, eso no significa que se haya logrado un contacto directo con el cliente, el cual es fundamental para realizar la cobranza.

De los 120 mil clientes promedio en mora, solo 20% son seleccionados para el File diario, donde normalmente se completa un promedio de 2,6 discados completos para cada uno de los seleccionados. De estos discados completos solo 8% son de contacto con alguien y 45% de estos contactos son directos con el cliente. De los contactos directos con clientes 70% generan una promesa de pago. Poniéndolo en números, tenemos un File diario de 24 mil clientes donde se realiza 62.000 discados completos para obtener 5.000 conexiones, siendo 2.270 con el titular de la cuenta para así concluir 1.600 promesas de pago. O sea, con todo el trabajo realizado solo se logra el 2,6% de promesas de pago en referencia al total de clientes en mora y 7% del total de clientes trabajados, los que están en el FD.




(Figura 2) – Retorno llamadas.

Fuente: Elaboración propia.

## 4. Hipótesis

Si se distribuye a los clientes en franjas horarias donde existe una posibilidad más grande de alcanzar el contacto con el titular de la cuenta, generará un aumento en la tasa de promesas de pago.

## 5. Objetivo

### Objetivo General

El objetivo del trabajo reside en alocar los clientes del FD en franjas horarias. Esas están definidas por un modelo estadístico basado en el histórico de llamadas y perfil de los clientes para así intentar alcanzar una mejor probabilidad de contacto.

### Objetivos Específicos

1. Desarrollar una base histórica conteniendo información de los clientes, información financiera, llamadas realizadas y resultados de llamadas en los últimos 3 años.
2. Establecer los mejores rangos horarios, entre las 8 de la mañana y las 10 de la noche.
3. Verificar los status de respuesta de las llamadas y establecer lo que puede ser considerado como un contacto efectivo, ya que existen varios tipos de contacto con el cliente, y agruparlos como CONTACTO / NO-CONTACTO.
4. Aplicar técnicas de minería de las mejores variables del histórico que identifiquen el contacto positivo en cierta franja utilizada de este cliente.
5. Establecer una puntuación de contacto entre 0 y 1 para cada cliente en las franjas horarias, donde 0 es casi nula y 1 es muy posible.
6. Desarrollar el modelo analítico.
7. Distribuir el FD basado en los mejores puntajes en las franjas horarias y realizar las llamadas.
8. Verificar el desempeño de contacto con el cliente comparando el File diario con franjas versus el File diario sin franjas.

## 6. Metodología

Para obtener los resultados deseados serán utilizadas algunas técnicas y herramientas:

- ETL

ETL del inglés “Extract Transform Load” (Extraer Transformar Cargar) es un proceso formado por tres etapas, a saber, extracción de los datos de un sistema origen, transformación de los mismos siguiendo reglas que permitan garantizar calidad de los datos y agruparlos según el formato esperado y finalmente almacenar la información transformada en el repositorio destino.

- Minería de Datos

La minería de datos es una técnica de explotación y análisis de descubrimiento de conocimiento de un conjunto de datos.

Según Fayyad (1996), la minería de datos es “un proceso significativo de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles en los datos”.

Poniéndolo en términos más simples, la minería es sacar el oro del dato. Es sacar lo más valioso del conjunto de datos que se tienen, llevando al descubrimiento del conocimiento y patrones antes no vistos en el dato bruto.

- Modelo de decisión

El modelo de decisión será el “motor” que genera la puntuación de cada cliente.

Para realizar el estudio fueran elegidos dos modelos, regresión logística y árboles de decisión.

Se representa la dependencia entre una variable dependiente Y y las variables independientes X

El análisis de regresión es el proceso por el cual se observa como la variable Y está relacionada con la variable X, siendo y1,...,yn y x1,...,xn las observaciones correspondientes. Se utiliza para realizar predicciones en una gran variedad de situaciones, que pueden ser sociales, económicas, biológicas, geofísicas, etc.

Arboles de decisión es un método que divide la información en varios subconjuntos binarios llamados nodos. Cada nodo consiste en una decisión binaria llevando a una rama de opciones hasta alcanzar la mejor posibilidad en base a los datos disponibles, o estableciendo límites que mejoren la precisión del objetivo.

Si bien los dos modelos fueran testados, y el de árboles resultó en un mejor porcentaje de aciertos. El objetivo del modelo es encontrar la posibilidad de contacto con el cliente en cada una de las franjas horarias disponibles. Esta posibilidad será dada a través de una puntuación entre 0 y 1 generada por el modelo de árboles, siendo cerca del 0 poco probable y cerca del 1 muy probable.

- SAS

SAS es una herramienta que ofrece una plataforma analítica completa e integrada que maneja cada paso del ciclo de vida analítico interactivo. Tiene una gran capacidad de interacción con los datos y cuenta con herramientas de minería de datos y creación de modelos analíticos.

# 7. Datos

La información utilizada en el presente estudio proviene del trabajo realizado en el sector de cobranzas de un banco. Esta información se compone por las llamadas realizadas, resultado de la llamada e información complementaria de los clientes para mejorar la posibilidad de previsión en el caso de no contener informaciones de llamadas.

- Fuentes principales de información

Serán utilizadas 3 principales fuentes de información para realizar el estudio. Las siguientes variables son las seleccionadas para utilizar en la construcción del modelo:

## CTA

Data WareHouse, que contiene la información de clientes en mora.

Variables CLIENTE utilizadas:

| Nombre | Tipo | Descripción |
|---|---|---|
| CD_CUST_NBR | String | Código único del cliente. |
| DT_BASE | Date | Fecha que la información fue generada en el aplicativo. |
| CD_STATE | Int | Código de la provincia de origen del cliente. |
| CD_COUNTRY | Int | Código del país de origen del cliente. |
| CD_PHONE_IND | Char | Status del número de teléfono del cliente. |
| ST_EMPLOYER | String | Nombre del local de trabajo del cliente. |
| CD_BEHAVIOUR_SCORE | Int | Valor de puntuación de riesgo del cliente. |

Volumen diario:

El volumen medio de la información (en un mes) es de 93.200 líneas por día, donde no existen informaciones del cliente duplicadas.

## DISCADOR

Sistema que realiza llamadas a los clientes. Extrae la información diaria de lo trabajado en un archivo crudo separado por punto y coma.

Variables DISCADOR utilizadas:

| Nombre | Tipo | Descripción |
|---|---|---|
| NR_CUST_NBR | String | Código único del cliente. |
| DT_BASE | Date | Fecha que la información fue generada en el aplicativo. |
| TM_DISPOSITION | Time | Horario en que el llamado fue realizado. |
| TM_MANAGEMENT | Time | Tiempo total del llamado. |
| TP_APL_DIALER | String | Código con el resultado del retorno de la llamada. Esta será la variable predictor. |
| NR_ACD | Int | Código del operador que realizo el contacto. |
| TP_LISTA | Char | Nombre de la lista que están los clientes. |
| NR_PHONE | String | Número de teléfono llamado. |

Volumen diario:

El volumen medio de la información (en un mes) es de 110.000 líneas por día, donde no existen informaciones del cliente/Horario duplicadas.

### SMG3

Información provista por el Sistema de decisión del banco, que contiene información complementaria del cliente.

Variables DISCADOR utilizadas:

| Nombre | Tipo | Descripción |
|---|---|---|
| NR_TP_DOC | String | Código del tipo de documento del cliente. |
| NR_CUST_NBR | String | Código del DNI del cliente. |
| DT_BASE | Date | Fecha que la información fue generada en el aplicativo. |
| CD_PAIS_RESIENCIA | String | Código del país de residencia del cliente |
| CD_PROVINCIA_RESI | String | Código de la provincia de residencia del cliente. |
| CD_ESTADO_CIV | Char | Código del estado civil del cliente |
| CD_SEXO | Char | Sexo del cliente. |
| CD_TIT_TIENE_AUTOMOVIL | Int | Informa si el cliente posee automóvil. |
| NR_VALOR_AUTOMOVIL | Int | Valor del automóvil del cliente. |
| NR_SUELDO_DECLARADO | Int | Valor del sueldo declarado por el cliente. |
| NR_SUELDO_VALIDADO | Int | Valor del sueldo verificado del cliente. |

Volumen diario:

El volumen medio de la información (en un mes) es de 300.000 líneas por día, donde pueden existir informaciones duplicadas del cliente en el día.

- Tratamiento de datos

Para atender a la necesidad del modelo, son realizados algunos tratamientos utilizando las buenas prácticas de extracción, importación y tratamiento de los datos.

## Extracción del DataWarehouse

Para extraer información de una tabla, por ejemplo, la tabla de clientes que contiene más de 350 millones de registros, son necesarios ciertos cuidados. Así lo mejor es utilizar las variables indexadas. En SAS puede hacerse utilizando el PROC CONTENTS, como filtro principal en la extracción.

```
proc contents data = gcmcm;
```

**Alphabetic List of Indexes and Attributes**

| # | Index | # of Unique Values |
|---|---|---|
| 1 | CD_CUST_NBR | 1030223 |
| 2 | DT_BASE | 1750 |

(Figura 3) – Variables indexadas.

Fuente: SAS.

Las tablas utilizadas tienen como campo indexado las fechas y el documento del cliente, para el presente trabajo utilizaremos solamente la fecha como filtro de extracción.

Con las tablas extraídas en ambiente SAS se realiza un tratamiento de la información, manteniendo únicamente la más actualizada de cada uno de los clientes.

Las variables que contienen el lugar de trabajo del cliente reciben un tratamiento sobre los caracteres, donde son removidos todos los caracteres que estén fuera del padrón ASCII (Del inglés “American Standard Code for Information Interchange”) y también arreglando espacios duplicados. Para eliminar la posibilidad de tener empresas duplicadas en el catastro por motivos de error de tipeo del nombre se ha usado la función SAS soundex(). Soundex es un algoritmo utilizado para indexar nombres de forma fonética conforme sonidos de la pronunciación Inglesa. Abajo puede ver un ejemplo en SAS:

```
data sondexExample;
    a = "ExxonMobil";
    b = "ExonMobil";
    aS=soundex(a);
    bS=soundex(b);
    PUT "Soundex A=" aS ", soundex B=" bS;
run;
```

```
Soundex A = E2514 Soundex B= E2514
```

El valor correcto es “ExonMobile”. El encontrado es “ExxonMobil”. Sin embargo, ambos recibirán el mismo valor fonético, donde al final las empresas serán agrupadas por ese valor indexado y no por el nombre.

Usaremos como información principal las variables de Edad, País de Origen, Provincia, indicativo del teléfono, nombre de la empresa donde trabaja (Privada o pública) y la puntuación del cliente en la institución.

## Importación del Discador

Antes de utilizar la información provista por el Discador, es necesario convertir los archivos crudos en tablas de SAS, donde las llamaremos Data. SAS contiene un óptimo sistema de importación automática de archivos, pero como será necesaria la importación de muchos archivos diarios, será utilizada una macro que reconoce los archivos en una carpeta utilizado el data step de SAS. Para su ejecución se provee la dirección del archivo, el delimitador de campos, la posición inicial del contenido y los nombres de cada campo con su formato de origen y de salida.

Debido a que el sistema del Discador es utilizado por otras áreas, es necesario filtrar un código de área para mantener solamente la información necesaria, así como borrando las que no contienen el documento del cliente.

En el momento de la importación son necesarios cambiar algunos campos, como por ejemplo el documento del cliente que no viene con tamaño fijo. Abajo podemos verificar la forma de importación de un archivo texto separado por “,” en SAS.

```
filename disca "/Directorio del archivo/";
data importDiscador;
    /*lrecl    - Limita el tamaño de la línea del archivo a importar
      dlm      - Determina el caracter que delimita los campos
      firstobs - línea inicial de la importacion*/
    infile disca lrecl=202 dlm=',' firstobs=2;
    /*input de las variables*/
    input   varName01 $
            varName02 $
            ...
            varNasmen $;
    /*Tratamiento auxiliar*/
    /*estandariza los DNIs*/
    DNI = put(input(varName1,8.),z19.);
    ...
run;
```

Después de filtrado y ajustado el dato, las tablas diarias son almacenadas en una Library de datasets.

Usaremos como información principal de ese archivo las acciones tomadas por el cobrador y el Horario de la llamada.

## Unión de los Datos

Para que la función del modelo de árboles funcione como se espera es necesario juntar todas las tablas en una única, una tabla final, conteniendo toda la información de los clientes, llamadas e información de soporte. Será considerando como registro único el Cliente + Franja Horaria. El proceso total fue separado en 4 pasos o steps:

- Step 1

Es realizada la unificación de todos los archivos diarios importados del Discador. En ese momento también es creada la variable que representa la franja horaria del momento de la llamada. Estarán agrupadas de la siguiente manera:

| DE | HASTA | FRANJA |
|---|---|---|
| 09:00 | 10:59 | 1 |
| 11:00 | 12:59 | 2 |
| 13:00 | 14:59 | 3 |
| 15:00 | 16:59 | 4 |
| 17:00 | 18:59 | 5 |
| 19:00 | 21:30 | 6 |

Adicionamos el tipo de contacto utilizando el campo de acción anotada por el cobrador. Para eso utilizamos una tabla auxiliar que contiene todas las acciones y se representa un intento, un Contacto o un Contacto con el cliente.

Por fin, agrupamos la información por Cliente, Franja Horaria y fecha para saber la cantidad de contactos. Abajo es posible ver una agrupación utilizando códigos SQL en SAS:

```
proc sql; /*Si hace necesario abrir una llamada sql*/
    /*codigo sql*/
    create table junccionSQL as select
        dni,
        franja,
        fecha ,
        sum(conexion) as cant_conex,
        sum(rpc) as cant_rpc
    from tpLlamadas
    group by 1,2,3
    order by 1,3,2;
quit;
```

- Step 2

Utilizando la tabla generada en el Step 1, se realiza la transposición de la información, manteniendo una línea por cliente/franja horaria distribuyendo la información de llamada de los meses en nuevas variables. Para eso es utilizado el método de arrays, que es muy eficiente para la cantidad de datos utilizados y por su simplicidad de uso en SAS.

Primero se contabiliza la cantidad de fechas existentes en el archivo, para crear un array utilizando solamente la cantidad de memoria necesaria, así facilitando el mantenimiento del código en caso de que exista un cambio en el rango de fechas.

En segundo lugar, se crean cuatro conjuntos de arrays, dos temporarios y dos de salida.

Tercero, es creado el MOB, del inglés “Months On Book”, que hace referencia a la posición donde la información será colocada en el Array.

Cuarto, utilizando el MOB se coloca informaciones de contacto en su casilla del Array.

Quinto y último: es realizado un tratamiento en el momento del “output” de la información. Eso es necesario cuando no existe ninguna información en alguno de los rangos horarios, es importante saber que en determinado rango horario es posible que no exista la información. Así es creado el rango faltante con todas las casillas vacías.

** Después de realizar los primeros tests se observó que es mejor crear variables de suporte en vez de utilizar todos los meses. Así son creadas las variables binarias de 1 mes, 3 meses, 6 meses y 12 meses siendo 1 cuando existe un contacto con el cliente en uno de los periodos, y 0 cuando no.

- Step 3

Con los datos extraídos del Data WareHouse se indentifica por el nombre de la empresa el tipo de trabajo del cliente, como:

| TIPO TRABAJO | CATEGORIA |
|---|---|
| Estatal | 1 |
| Privada | 2 |
| Jubilado | 3 |
| No trabaja | 0 |

Es mantenida solamente la información más reciente del cliente.

- Step 4

En este step es creada la tabla final, realizando la unión de las 3 tablas descritas en los steps anteriores, así como las del SMG3 (la última información disponible del cliente), que ya están disponibles en SAS. La clave de la relación es el documento del cliente. De esta manera, los datos del cliente estarán duplicados por rango horario, pero eso no es considerado un problema ya que la información de llamadas no estará.

# 8. Modelo

## Modelo Analítico

El modelo consiste en analizar cuál es el mejor momento para entrar en contacto con dicho cliente, entonces la variable objetivo del modelo será el éxito de contacto con el cliente (RPC).

Para eso usaremos el modelo de árboles de decisión, donde es analizado algo en común entre todos los clientes para definir cuál es la rama principal y la secundaria y así con cada rama creada.

**Fue realizado con el modelo analítico lineal porque se obtuvo los mejores resultados.

Para eso la tabla final será dividida en otras 6 tablas, una para cada rango horario, aplicado la función hpslipt para cada una, adicionando el score generado y generando los decíles con el proc Rank para análisis posterior. Usando una macro SAS es posible generar las 6 tablas y correr todos los pasos descritos de una vez.

```
%macro horarios(_nr);
    %do _i = 1 %to &_maxHorarios;
        horario_&_i;
    %end;
    %if &_nr = 1 %then %do;
        if TP_HORARIO = &_i then output horario_&_i;
    %end;
    %if &_nr = 3 %then %do;
        proc hpsplit data = horario_&_i seed = 15531;
            target RPC;
            input _variables_;
            criterion entropy;
            partition fraction (validate=0.2);
            prune costcomplexity;
            score out = scored_time&_i;
        run;

        data scored_time&_i._j;
            merge scored_time&_i (in= a keep= P_RPC1 V_RPC1)
            horario_&_i (in=b keep = DNI TP_HORARIO);
            P_1_a=round(P_RPC1*1000);
        run;

        proc rank data=scored_time&_i._j
            group=10
            ties=mean
            out=rankedScored_time_&_i._j;
            var P_1_a;
        run;
    %end;
    %if &_nr = 4 %then %do;
        rankedScored_time&_i._j
    %end;
%mend;
```

Data: Data set de entrada;
target: Variable binaria base para el análisis;
input: Variables de apoyo para el target;
Partition fraction: Como la información de entrada va a ser particionada;

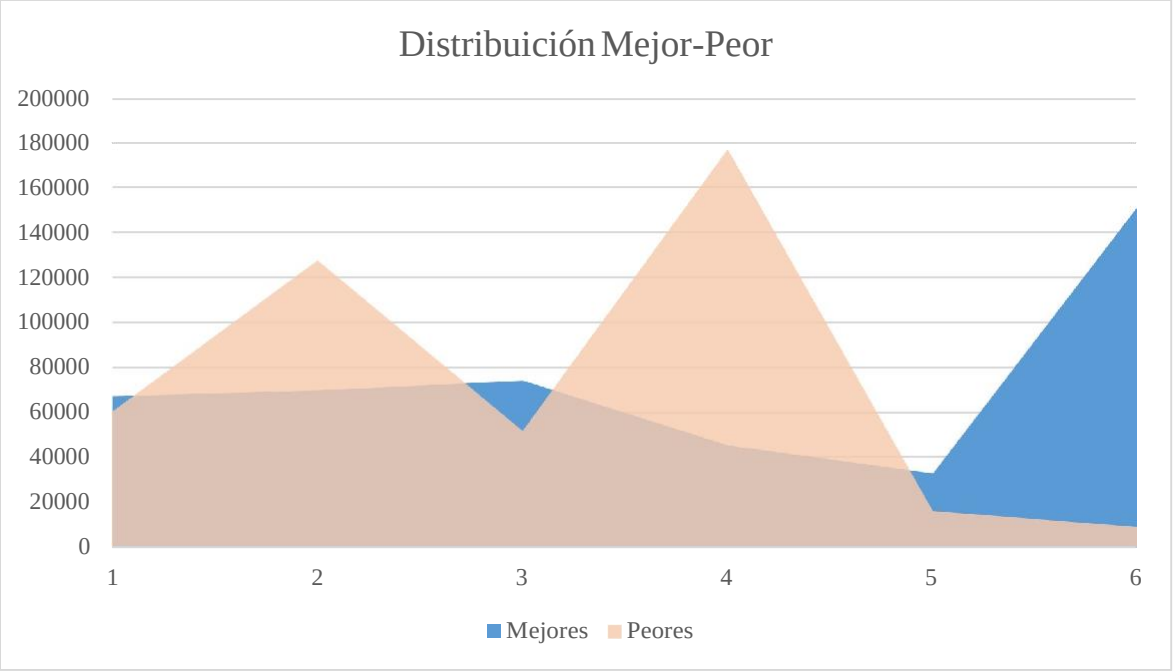
Criterion: Especifica el criterio de como el árbol va crecer;
Prune: Como será realizada la poda del árbol;
Score out: Nombre del data set de salida.

Y la ejecución de las macros, que puede ser considerado el Main():

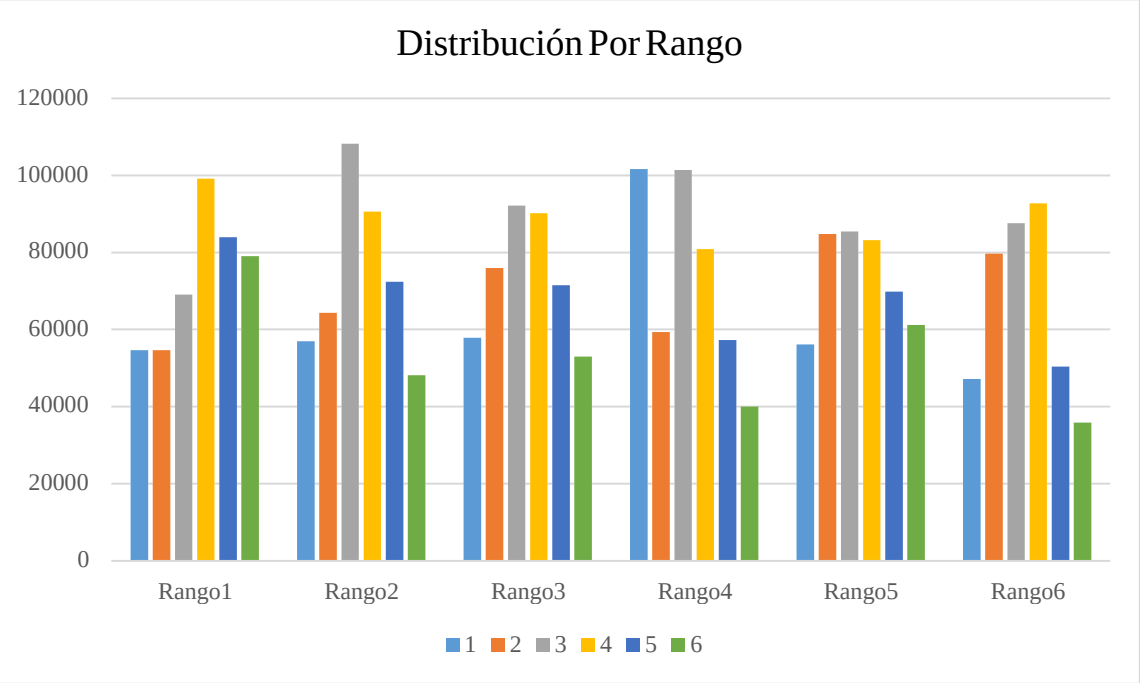
```
data %horario(0);
set histLlamadasClientesTrat;
    %horarios(1);
run;

%horarios(3);

data juntaHorarios;
    set %horarios(4);
run;
```

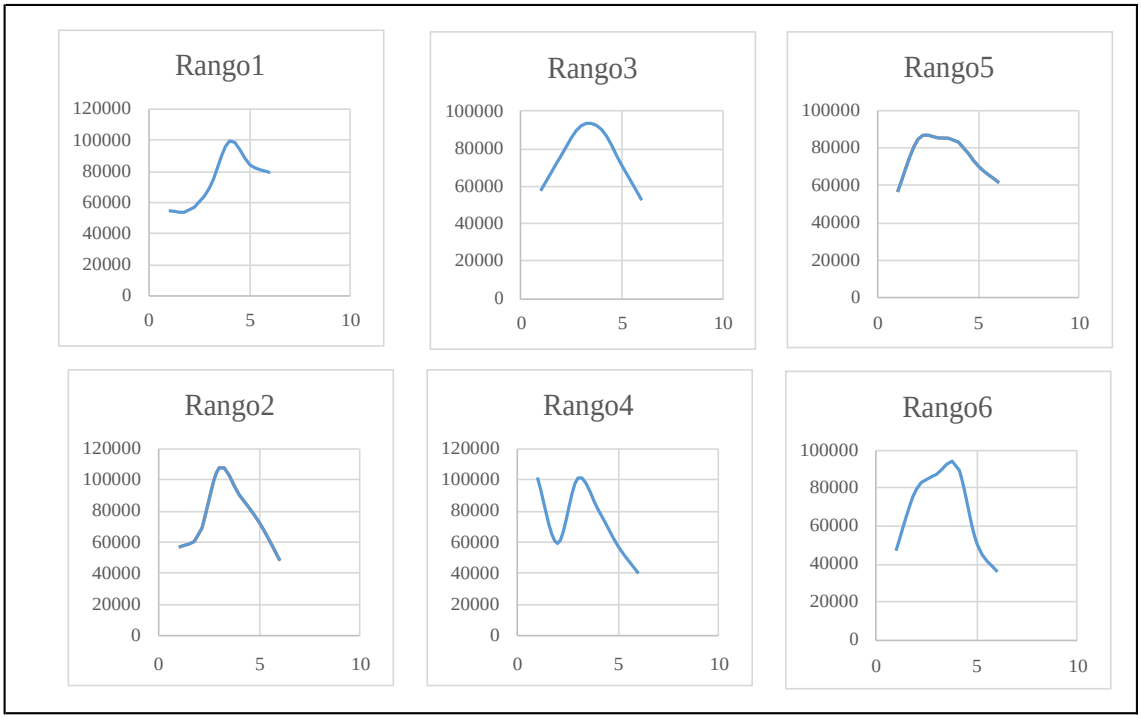
- Resultados Obtenidos

La tabla creada juntaHorarios contiene la puntuación final de cada cliente para cada uno de los 6 rangos horarios. Extraemos el mejor y el peor horario de cada cliente, manteniendo solamente dos rangos por cliente.



Es posible observar que, utilizando ese método, la mayor cantidad de clientes con un buen score está en el rango 6 y los peores scores están concentrados en el rango 4. Pero ese método no es el mejor, ya que varios clientes presentan una puntuación igual en varios rangos, así las puntas acaban siendo numerosas.




Puede verse que el modelo que mejor alineó su distribución fue el generado por el rango 3, que contiene una curva casi perfecta.

Con el dato fuente creado por el proceso hpsplit utilizado en cada uno de los 6 rangos horarios, es creada una “fuente modelo”. Así con esa fuente será posible aplicar el modelo específico en la totalidad de los datos y no solamente en su rango, creando la posibilidad de obtener mejoras en los resultados. Como obtuvimos del rango 3 la mejor distribución, utilizaremos solamente éste entre los 6 para hacer las pruebas, así como el modelo principal.

## 9. Prueba inicial del Modelo

Para verificar la eficacia inicial del modelo se realiza una verificación de los scores con su mejor momento de llamada con días futuros, o sea, los que no utilizamos para generar el score de prueba.

### Muestra de Prueba

Para la prueba serán utilizados datos provenientes del Discador, conteniendo la respuesta de todos los llamados realizados. Esa muestra será de apenas un día de trabajo.

Para realizar la prueba también se necesitó la creación del rango horario del llamado, utilizando la variable que contiene el horario/momento en que el cliente contestó el llamado.

### Conceptos de Prueba

Con los datos de la prueba vamos a analizar la veracidad del score que hemos creado. Para eso vamos a juntar a los clientes que tienen un RPC con sus 5 mejores rangos basados en los puntajes del score (SCORE_1 – SCORE_5), donde la primera variable contiene el mejor rango para llamar, la segunda el segundo mejor rango y así sucesivamente.

Como cada cliente tiene 6 scores, uno para cada uno de los rangos horarios, es posible que contengan un valor igual, mejor o peor en cada uno de ellos. Por eso mantenemos los 5 mejores scores para quitar posibilidad de falta de acierto en caso de scores repetidos, como por ejemplo la existencia de un RPC en el rango 4 cuando teníamos el rango 1 como mejor probabilidad de acierto, siendo que el cliente tiene el mismo valor de score en el rango 4, que está en la segunda mejor variable. Así se realiza un escaneo para cuando existan puntuaciones iguales subyacentes.

El resultado será presentado dividido en 3 categorías:

RPC Aciertos: Todos los altos scores acertarán el momento de llamar obteniendo un acierto.

Conexión Sin RPC: No hablamos con el cliente, pero tenemos un llamado atendido en ese momento.

Conexiones posibles aciertos: No realizamos llamadas para el cliente en el rango donde ya existía su mejor potencial de acierto.

- Resultados Obtenidos

| RPC Aciertos | total obs | Resultado | % |
|---|---|---|---|
| Fuente del modelo 3 | 2886 | 293 | 10,2% |
| Utilizando los 6 rangos | 2886 | 378 | 13,1% |
| | | | |
| Conexiones Sin RPC | | | |
| Fuente del modelo 3 | 54531 | 8186 | 15,0% |
| Usando los 6 rangos | 54531 | 6625 | 12,1% |
| | | | |
| Conexiones Posibles acierto | | | |
| Fuente del modelo 3 | 54531 | 46345 | 85,0% |
| Usando los 6 rangos | 54531 | 47906 | 87,9% |

Es posible observar que el modelo de los “6 rangos” sale victorioso cuando lo comparamos con el “modelo 3”, no es expresivo, siendo ~3% mejor.

Con ese resultado es posible afirmar que existe una gran posibilidad de mejora aplicado el modelo en producción, tenemos una posibilidad de 87% de mejora para aquellos que no obtuvieron un contacto directo con el cliente. Donde vamos a comprobar eso será en el testeo del modelo en ambiente productivo.

## 10. Prueba del Modelo en ambiente productivo

Para verificar si existe una mejora real en el nuevo modelo debemos comparar con el modelo de llamadas que está en producción actualmente. Para eso hay que separar una muestra de la población total de clientes a llamar y aplicar el modelo de llamadas solamente en esta muestra, dejando el restante bajo el modelo actual. Al fin del día hacemos la comparación de performance porcentual entre las dos muestras.

### Muestra de Prueba

La muestra de análisis será de 10% del total de clientes entre los días 14 y 18 de enero de 2019. Abajo es posible ver el total de cuentas de cada uno de los días con sus respectivas muestras:

| Día | Clientes Total | 10% | 90% |
|---|---|---|---|
| 14/01/2019 | 26.881 | 2.688 | 24.193 |
| 15/01/2019 | 25.858 | 2.586 | 23.272 |
| 16/01/2019 | 27.815 | 2.782 | 25.034 |
| 17/01/2019 | 25.885 | 2.589 | 23.297 |
| 18/01/2019 | 26.039 | 2.604 | 23.435 |

Serán seleccionados 10% de los mejores scores en cada uno de los rangos, totalizando seis veces el total de la muestra. Puede pasar que un cliente tenga más de un buen rango horario entre aquellos seleccionados. Para alcanzar el total necesario de clientes se dividido el 10% total en 6 rangos. Tomemos como ejemplo el día 14:

1) 2.688 mejores scores para cada uno de los rangos, totalizando 16.129 observaciones.
2) Sacando los clientes duplicados nos quedamos con la siguiente distribución:

| Rango | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| Clientes | 2654 | 880 | 1635 | 1106 | 2373 |

3) Restamos los clientes excedentes hasta que alcancemos los 2.688 / 6 = 448 de cada uno de los rangos.

## Conceptos de Prueba

La muestra será insertada manualmente en el Discador durante el transcurso de cada día, donde en cada cambio de rango entran los casos seleccionados del día/rango. El resto de casos, que no están en el 10%, será insertado al principio del día y las llamadas ocurrirán sin orden programado, así como está establecido hoy.

Al final del día, será realizada la comparación porcentual de RPCs obtenidos sobre la cantidad de discados completos en el rango horario en el correr del día.

## Resultados Obtenidos

Abajo es posible ver en el cuadro el porcentaje de RPCs sobre la cantidad de discados completos en cada uno de los días en que la prueba fue ejecutada.

| | | | | | PRUEBA | | | NORMAL | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Fecha | Clientes Total | Prueba | Normal | Rango | Discados | RPC | % | Dicado | RPC | % |
| 14/01/2019 | 26.881 | 2.688 | 24.193 | 1 | 684 | 54 | 7,9% | 33.996 | 660 | 1,9% |
| | | | | 2 | 721 | 46 | 6,4% | 22.436 | 359 | 1,6% |
| | | | | 3 | 672 | 55 | 8,2% | 14.164 | 179 | 1,3% |
| | | | | 4 | 703 | 49 | 7,0% | 3.361 | 158 | 4,7% |
| | | | | 5 | 606 | 57 | 9,4% | 1.368 | 89 | 6,5% |
| | | | | 6 | 703 | 43 | 6,1% | 1.358 | 104 | 7,7% |
| 15/01/2019 | 25.858 | 2.586 | 23.272 | 1 | 643 | 57 | 8,9% | 25.891 | 814 | 3,1% |
| | | | | 2 | 589 | 33 | 5,6% | 14.065 | 422 | 3,0% |
| | | | | 3 | 703 | 46 | 6,5% | 17.139 | 317 | 1,8% |
| | | | | 4 | 567 | 43 | 7,6% | 20.703 | 542 | 2,6% |
| | | | | 5 | 743 | 39 | 5,2% | 13.987 | 389 | 2,8% |
| | | | | 6 | 689 | 55 | 8,0% | 10.537 | 559 | 5,3% |
| 16/01/2019 | 27.815 | 2.782 | 25.034 | 1 | 905 | 57 | 6,3% | 22.820 | 805 | 3,5% |
| | | | | 2 | 705 | 65 | 9,2% | 10.662 | 367 | 3,4% |
| | | | | 3 | 896 | 54 | 6,0% | 7.353 | 286 | 3,9% |
| | | | | 4 | 853 | 69 | 8,1% | 30.991 | 477 | 1,5% |
| | | | | 5 | 743 | 57 | 7,7% | 19.286 | 330 | 1,7% |
| | | | | 6 | 632 | 47 | 7,4% | 22.647 | 492 | 2,2% |
| 17/01/2019 | 25.885 | 2.589 | 23.297 | 1 | 743 | 63 | 8,5% | 25.054 | 723 | 2,9% |
| | | | | 2 | 597 | 53 | 8,9% | 19.111 | 379 | 2,0% |
| | | | | 3 | 607 | 37 | 6,1% | 12.977 | 243 | 1,9% |
| | | | | 4 | 602 | 56 | 9,3% | 15.912 | 564 | 3,5% |
| | | | | 5 | 743 | 72 | 9,7% | 11.319 | 295 | 2,6% |
| | | | | 6 | 664 | 63 | 9,5% | 14.679 | 416 | 2,8% |
| 18/01/2019 | 26.039 | 2.604 | 23.435 | 1 | 684 | 45 | 6,6% | 26.561 | 724 | 2,7% |
| | | | | 2 | 706 | 63 | 8,9% | 22.879 | 372 | 1,6% |
| | | | | 3 | 643 | 54 | 8,4% | 9.175 | 239 | 2,6% |
| | | | | 4 | 595 | 36 | 6,1% | 20.584 | 582 | 2,8% |
| | | | | 5 | 574 | 59 | 10,3% | 16.456 | 396 | 2,4% |
| | | | | 6 | 609 | 45 | 7,4% | 20.160 | 457 | 2,3% |

Es posible observar una mejora de ~5% promedio sobre el total de RPCs totales comparando con el modelo de trabajo actual.

Es importante referenciar que los discados completos están distribuidos entre los dos modelos, y no existe una “prioridad” para llamada en cada uno los rangos horarios.

## 11. Conclusión

Con todo lo que fue presentado sabemos que se necesita de una inteligencia por detrás de los operadores que hablan con los clientes. Es necesario un planeamiento y varios controles para que los objetivos sean alcanzados. Siempre es necesario evolucionar y las mejoras alcanzadas en estudios como este, muestran que existe margen de mejora y beneficios para el negocio.

Mirando los resultados acá obtenidos, con la integración de mayor control e inteligencia para el negocio, se obtendría una mejora no solamente con el porcentaje de contacto sino también en la oportunidad de analizar cómo debemos tratar a cada uno de los clientes con la información adicional sobre la segregación del día en franjas horarias.

La mejora de 5% presentada puede parecer no muy significativa, y difícilmente ese valor será alcanzado cuando el modelo entre en producción. Pero eso es un cambio de paradigma ya que desde que fue implementado el Discador automático, no importaba el esfuerzo realizado y nunca fue posible subir la tasa de RPC. En cobranzas el objetivo final es obtener los pagos atrasados y hablar con el cliente es el primero paso para realizar esa cobranza

Existe mucho más para hacer, como por ejemplo instalar una inteligencia artificial sobre ese score quitando la necesidad de interacción manual para realizar los cambios de listas, pero eso todavía no es posible porque serían necesarias inversiones ya que no es posible hacerlos con los equipos y aplicaciones disponibles en el momento.

## 12. Referencias

Acosta, Juan Carlos (2001). Gestion de Creditos y Cobranzas.

Fayyad U.M., Piatetskiy-Shapiro G., Smith P., & Ramasasmy U. (1996). Advances in Knowledge Discovery and Data Mining

Fayyad, U.M., Grinstein, G., Wierse, A. (2001). Information Visualization in Data Mining nd Knowledge Discovery

Hand, D.J., Mannila, H., & Smyth, P. (2000). Principles of Data Mining, The MIT Press.

Pyle, D. (1999). Data Preparation for Data Mining. Morgan Kaufmann, Harcourt Intl.

Berthold, M., & Hand, D.J. (2002). Intelligent Data Analysis.

Tan, P., Steinbach, M., & Kumar V. (2016). Introduction to Data Mining.

Ames, J., Abbey, R., & Thompson, W. (2013). Big Data Analytics: Benchmarking SAS, R, and Mahout. SAS Institute, Inc.

Hall, P., Chien, A., Kabul, I., & Silva, J. (2016). An Efficient Pattern Recognition Approach with Applications. SAS Institute, Inc.

Czika, W., & Liu, Y. (2016). Ensemble Modeling: Recent Advances and Applications. SAS Institute, Inc.

Murphy C., & Peter F. (2010). Building Decision Trees from Decision Stumps. SAS Global Forum.

Manoj I., & Chakraborty, G. (2012). Kass Adjustments in Decision Trees on Binary/Interval Target. SAS Global Forum.

SAS (2015). SAS/STAT14.1 User's Guide. The HPSPLIT Procedure. https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf

Duhigg, C. (2012). The power of Habit.

Gambini, J. (2017). Fundamentos de Análisis de Datos. ITBA Class.

Russell,Kevin - SAS (2015). How to perform a fuzzy match using SAS functions https://blogs.sas.com/content/sgf/2015/01/27/how-to-perform-a-fuzzy-match-using-sas-functions/