



Caso 2 - Telco Co.

Universidad Austral, MEDGC | Cohorte 2020-2021 | Introducción a Data Mining

Grupo 5B

Agosto 24, 2020

INTEGRANTES

FARACH, VERONICA

HEREDIA BAEK, GABRIELA

KINA, HERNAN

LANZA, EZEQUIEL

LAXI, IGNACIO

MANZL, JUAN JOSE

Contents

| | |
|--|---|
| Introducción | 3 |
| Fase Exploratoria: Parte I (Gaby y Nacho?) | 3 |
| Fase Exploratoria: Parte II (?) | 3 |
| Modelos (Eze y Hernan?) | 3 |
| Conclusiones (Todos?) | 3 |
| Graficos multivariados para variables cualitativas- Observaciones sobre el grafico | 5 |

Introducción

Fase Exploratoria: Parte I (Gaby y Nacho?)

1. ¿Si Ud. tuviera que implementar un sistema de retención de clientes, considera que Data Mining puede ser de ayuda en la detección de clientes próximos a abandonar la compañía?
2. ¿Qué tipo de tarea o tareas de Data Mining se podrían aplicar?
3. ¿Qué tipo de datos requeriría para poder aplicar las tareas de Data Mining seleccionadas?
4. ¿Qué parámetros debería establecer para la tarea o tareas y qué valores asignaría a cada uno?
5. Si observa alguna particularidad en los datos de este problema con respecto al tiempo, proponga el tratamiento especial que considere adecuado.
6. Considere el caso de una compañía de telefonía fija con diversas alternativas de Churn (baja del cliente completo, líneas, Internet, paquetes urbanos, etc.). a. ¿Cómo aplicaría los resultados obtenidos del mining?

Fase Exploratoria: Parte II (?)

1. Explore si hay valores faltantes en alguna de las variables.
2. Compare los campos area code y state. Discuta cualquier aparente anomalía.
3. Emplee gráficos para determinar visualmente si hay algún valor extremo en la cantidad de llamadas a la línea de atención al cliente.
4. Identifique el rango de llamadas a customer service, que debieran considerarse outliers empleando: a. El método de puntaje z b. El método del RIC.
5. Transforme la variable day minutes empleando estandarización por puntaje Z.
6. Trabaje con los sesgos: a. Calcule el sesgo de la variable day minutes. b. Calcule el sesgo de la variable estandarizada por puntaje Z para day minutes. Comente. c. Basado en el valor del sesgo, ¿Considera que la variable se encuentra sesgada o es casi perfectamente simétrica? 7. Construya el normal probability plot de la variable day minutes. Comente sobre la normalidad de los datos. Análisis Descriptivo Previo del dataset churn 8. Trabaje con la variable international minutes: a. Construya el normal probability plot de la variable. b. ¿Qué evita que esta variable tenga una distribución normal? c. Construya una variable indicadora para lidiar con la situación anterior. d. Construya un normal probability plot de la variable derivada nonzero international minutes. Comente en relación a la normalidad de la variable derivada.
7. Transforme la variable night minutes empleando estandarización por puntaje Z. Empleando un gráfico, describa el rango de los valores estandarizados.

Modelos (Eze y Hernan?)

1. Observar las distribuciones de las variables
2. Buscar y eliminar variables correlacionadas.
3. Analizar las proporciones de churn para distintas variables.
4. Partitionar los datos.
5. Generar tres modelos de árbol.
6. Evaluar los modelos.

Conclusiones (Todos?)

APENDICE

Graficos multivariados para variables cualitativas- Observaciones sobre el grafico

1. La proporcion de subscriptores que no tienen Plan de Voice Mail es mayor, en linea con lo observado en los barplots
2. La cantidad de subscriptores que no tienen Plan Internacional es mayor independientemente de si tienen o no Plan de Voicemail.
3. La proporcion de churn es menor entre los subscriptores SIN Voice Mail y SIN Plan Internacional en comparacion con los subscriptores SIN Voice Mail y CON Plan Internacional donde la proporcion de churn y no churn son mas parejas
4. La proporcion de churn es menor entre los subscriptores CON Voice Mail y SIN Plan Internacional en comparacion con los subscriptores CON Voice Mail y CON Plan Internacional donde la proporcion de churn y no churn son mas parejas
5. Por los puntos 3 y 4 las proporciones de churn y no churn son mas similares en subscriptores con Planes internacionales, independientemente de si tienen o no Plan de Voicemail.

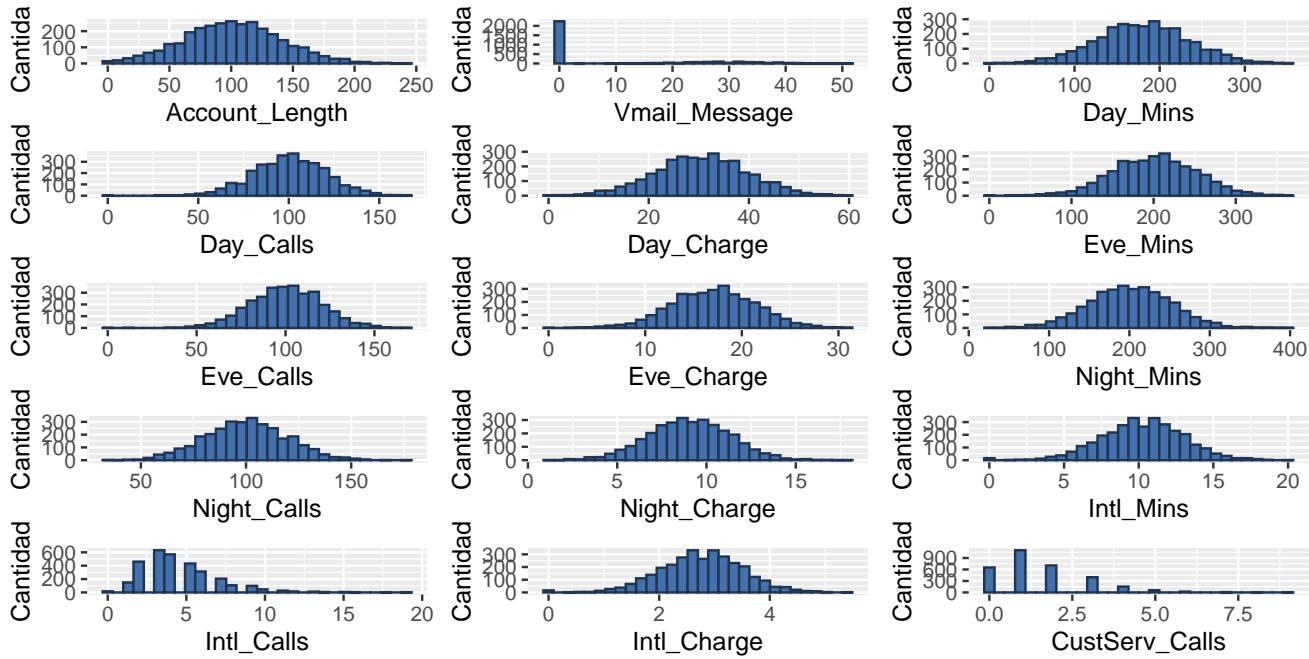


Figure 1: Histograma

De entre las 8 variables correlacionadas se pueden eliminar 4 variables. Las variables seleccionadas para eliminacion son las correspondientes a los cargos: Day_Charge, Eve_Charge, Night_Charge, Intl_Charge.

Creamos un nuevo dataset reducido en 4 columnas: subscriptores.red

Hago una matriz de dispersion indicando en rojo las observaciones que son outliers segun las distnacias de Mahalanobis

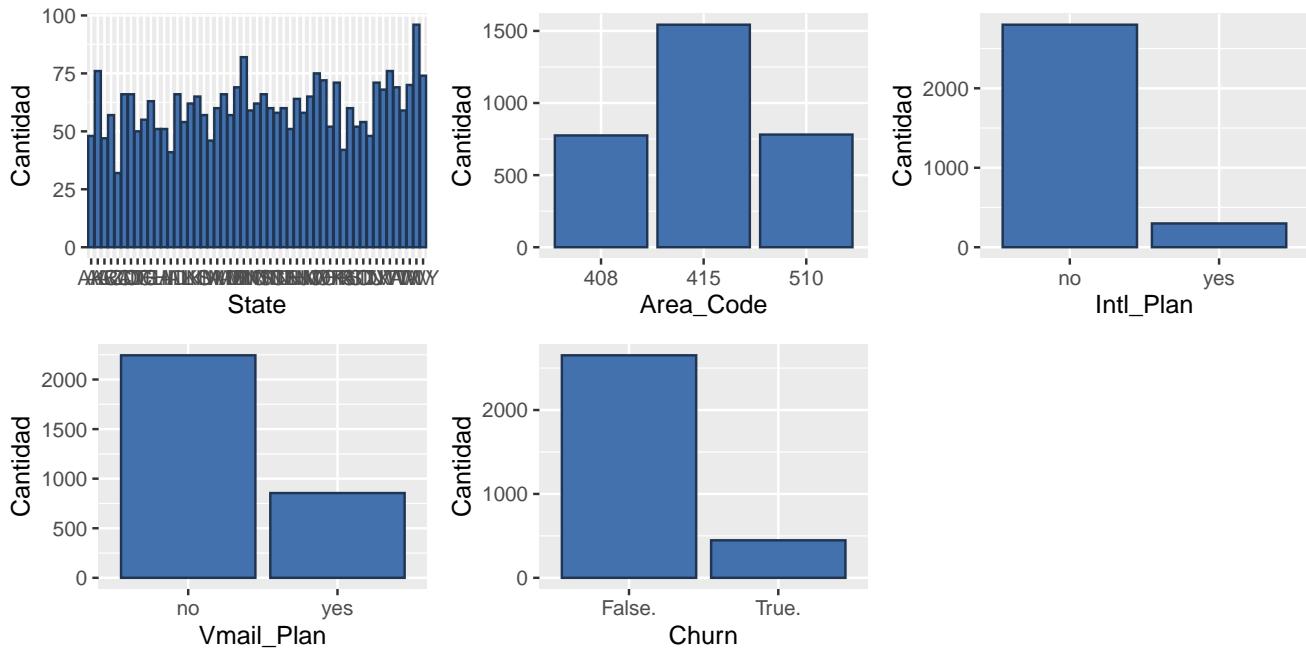


Figure 2: Bar Plots

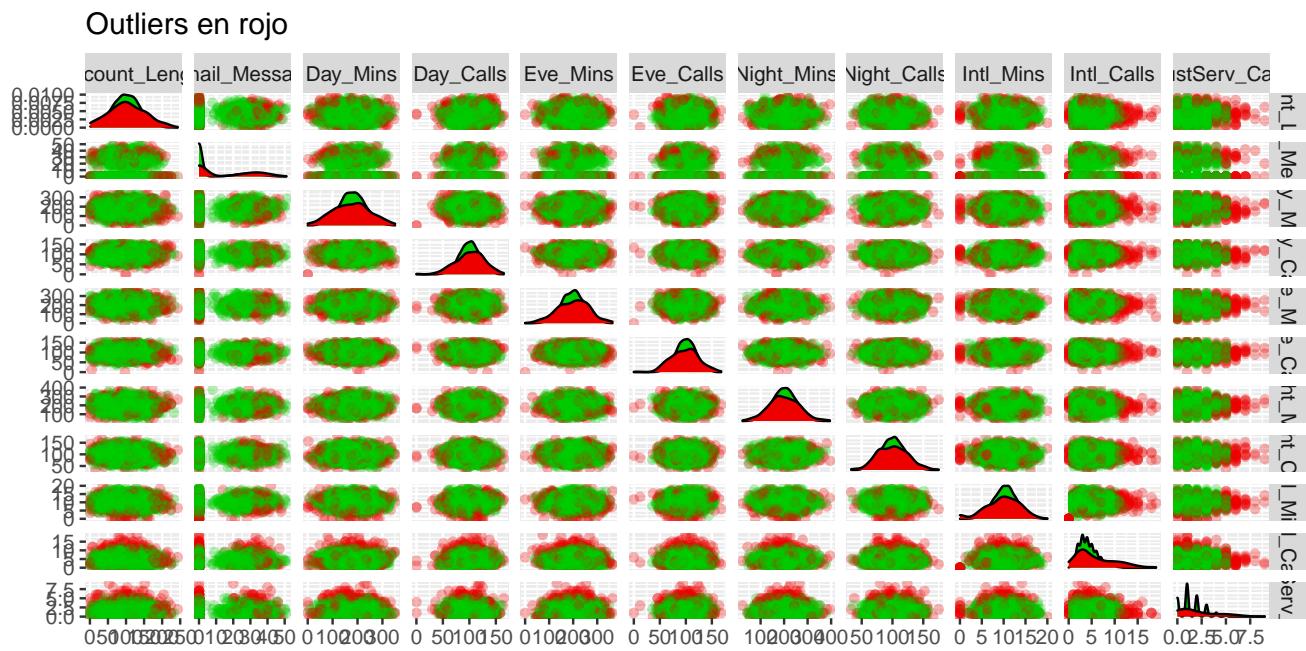


Figure 3: Bar Plots