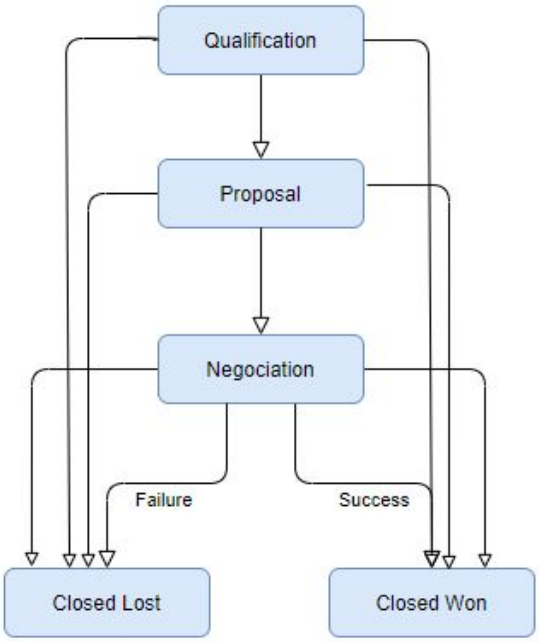
Organización de Datos (75.06/95.58)
Segundo Cuatrimestre 2020
Trabajo Práctico N°1

Curso 1: Argerich

Introducción:

Les presentamos a la empresa “Frío Frío”, dedicada a la venta e instalación de equipos de aire acondicionado para grandes superficies. Al ser una empresa B2B (“Business To Business”), es esencial para ellos optimizar los esfuerzos de los representantes comerciales, ayudándolos a priorizar las oportunidades en el pipeline.

Una “oportunidad” consiste en un proyecto de venta o instalación de equipos para un cliente. La venta se estructura alrededor de TRF (Toneladas de refrigeración) y puede estar compuesta por varios productos distintos. El "pipeline" hace referencia al flujo de oportunidades prospecto que la empresa está desarrollando. El equipo comercial asigna a distintos momentos, para cada oportunidad, un estado en la negociación. En la Ilustración se muestran los estados que las oportunidades tienen dentro del pipeline.



La variable que se está tratando de predecir es **“Probabilidad de éxito”** para cada oportunidad. ¿Cuál es la probabilidad de que la oportunidad se convierta en un caso *Closed Won*?

El dataset cuenta con información de cada oportunidad, como por ejemplo información sobre el vendedor a cargo de la venta, información geográfica de los clientes, TRF pedidas, fecha prevista de entrega de los equipos, etc. A partir de dichas variables es posible entrenar un modelo que prediga para un tiempo futuro el éxito o fracaso de cada oportunidad. Idealmente, “Frío Frío” podrá usar este modelo para predecir la probabilidad de éxito de cada oportunidad comercial, para mejorar el rendimiento y optimizar el esfuerzo de los vendedores.

Objetivo:

El TP consiste en realizar un análisis exploratorio de los datos provistos con el objetivo de determinar características y variables importantes, descubrir insights interesantes, y analizar la estructura de los mismos.
Tener en cuenta que el trabajo realizado podrá ser utilizado en el TP2 (que consistirá en predecir la probabilidad de que una oportunidad sea exitosa) y se puede considerar un paso previo al mismo.

Datos:

Los datos a analizar se pueden encontrar haciendo clic **aquí**.

La variable de interés es *“**Stage**”*, que indica en que estado se encuentra la operación(pueden ser 5: Closed Won, Closed Lost, Negotiation, Proposal, Qualification)

Requisitos:

Los requisitos de la entrega son los siguientes:

- El análisis debe estar hecho en Python o R.
- El análisis debe entregarse en formato PDF vía gradescope. En el informe no va código.
- Informar el link a un repositorio Github en donde pueda bajarse el código completo para generar el análisis.

Evaluación:

La evaluación del TP1 se realizará en base al siguiente criterio:

- Originalidad del análisis exploratorio.
- Calidad del reporte. ¿Está bien escrito? ¿Es claro y preciso?
- Calidad del análisis exploratorio
    - Qué tipo de preguntas se hacen y de qué forma se responden, ¿es la respuesta clara y concisa con respecto a la pregunta formulada?
    - ¿Se plantean hipótesis sobre lo observado?
    - ¿Se realiza un mínimo preprocesamiento o limpieza de los datos?
    - ¿Se profundiza en los datos más allá de un simple análisis estadístico?
- Calidad de las visualizaciones presentadas.
    - ¿Tienen todos los ejes su rótulo?
    - ¿Tiene cada visualización un título?
    - ¿Están numeradas las visualizaciones?
    - ¿Es entendible la visualización sin tener que leer la explicación?
    - ¿El tipo de plot elegido es adecuado para lo que se quiere visualizar?
    - ¿Es una visualización interesante?
    - ¿El uso del color es adecuado?
    - ¿Hay un exceso o falta de elementos visuales en la visualización elegida?
    - ¿La visualización es consistente con los datos?
    - ¿Presenta el grupo un listado de "insights" aprendidos sobre los datos en base al análisis realizado? ¿Es interesante?
- Conclusiones presentadas.

El grupo que realice el mejor análisis exploratorio obtendrá 10 puntos para cada uno de sus integrantes que podrán ser usados en el parcial además de ser publicado en el repositorio de la materia como ejemplo para los siguientes cuatrimestres.

Recursos:

Algunos recursos interesantes para visualizaciones en Python y R:

- https://python-graph-gallery.com/
- https://datavizproject.com/
- https://www.r-graph-gallery.com/

Referencia de los datos:

- **ID:** id único del registro (Entero).
- **Región:** región de la oportunidad (Categórica).

- **Territory:** territorio comercial de la oportunidad (Categórica).
- **Pricing, Delivery_Terms_Quote_Approval:** variable que denomina si la oportunidad necesita aprobación especial de su precio total y los términos de la entrega (Binaria).
- **Pricing, Delivery_Terms_Approved:** variable que denomina si la oportunidad obtuvo aprobación especial de su precio total y los términos de la entrega (Binaria).
- **Bureaucratic_Code_0_Approval:** variable que denomina si la oportunidad necesita el código burocrático 0 (Binaria).
- **Bureaucratic_Code_0_Approved:** variable que denomina si la oportunidad obtuvo el código burocrático 0 (Binaria).
- **Submitted_for_Approval:** variable que denomina si fue entregada la oportunidad para la aprobación (Binaria).
- **Bureaucratic_Code:** códigos burocráticos que obtuvo la oportunidad (Categórica).
- **Account_Created_Date:** fecha de creación de la cuenta del cliente (Datetime).
- ***Source*:** fuente de creación de la oportunidad (Categórica).
- **Billing_Country:** país donde se emite la factura (Categórica).
- **Account_Name*:*** nombre de la cuenta del cliente (Categórica).
- **Opportunity_Name**: nombre de la oportunidad (Categórica).
- **Opportunity_ID**: id de la oportunidad (Entero).
- **Sales_Contract_No**: número de contrato (Entero).
- **Account_Owner**: vendedor del equipo comercial responsable de la cuenta cliente (Categórica).
- **Opportunity_Owner**: vendedor del equipo comercial responsable de la oportunidad comercial (Categórica).
- **Account_Type**: tipo de cuenta cliente (Categórica).
- **Opportunity_Type**: tipo de oportunidad (Categórica).
- **Quote_Type**: tipo de presupuesto (Categórica).
- **Delivery_Terms**: términos de entrega (Categórica).
- **Opportunity_Created_Date**: fecha de creación de la oportunidad comercial (Datetime).
- **Brand**: marca del producto (Categórica).
- **Product_Type**: tipo de producto (Categórica).
- **Size**: tamaño del producto (Categórica).
- **Product_Category_B**: categoría 'B' del producto (Categórica).

- **Price**: precio (Decimal).
- **Currency**: moneda (Categórica).
- **Last_Activity**: fecha de la última actividad (Datetime).
- **Quote_Expiry_Date**: fecha de vencimiento del presupuesto (Datetime).
- **Last_Modified_Date**: fecha de ultima modificación en la oportunidad (Datetime).
- **Last_Modified_By**: usuario responsable de la última modificación en la oportunidad (Categórica).
- **Product_Family**: familia de producto (Categórica).
- **Product_Name**: nombre del producto (Categórica).
- **ASP_Currency**: moneda del precio promedio (Categórica).
- **ASP**: (Average Selling Price) precio promedio a la venta (Decimal).
- **ASP_(converted)_Currency**: moneda del precio promedio convertido en la variable (Categórica)
- **ASP_(converted)**: precio promedio a la venta convertido a otra moneda (Decimal).
- **Planned_Delivery_Start_Date**: límite inferior del rango previsto para la fecha de entrega (Datetime).
- **Planned_Delivery_End_Date**: límite superior del rango previsto para la fecha de entrega (Datetime).
- **Month**: mes-año de Planned_Delivery_Start_Date (Fecha).
- **Delivery_Quarter**: trimestre de Planned_Delivery_Start_Date (Categorica).
- **Delivery_Year**: año de Planned_Delivery_Start_Date (Fecha).
- **Actual_Delivery_Date**: fecha real de la entrega (Datetime).
- **Total_Power**: potencia del producto (Entero).
- **Total_Amount_Currency**: moneda del monto total (Decimal).
- **Total_Amount**: monto total (Decimal).
- **Total_Taxable_Amount_Currency**: moneda del monto gravado total (Categórica).
- **Total_Taxable_Amount**: monto gravado total (Decimal).
- **Stage**: variable target. Estado de la oportunidad (Categórica).
- **Prod_Category_A**: categoría 'A' del producto (Categórica).
- **TRF**: Toneladas de refrigeración (Entero). Es una unidad de potencia.

Preguntas frecuentes

Como este cuatrimestre no tenemos contacto con la empresa no podemos evacuar todas las dudas con certeza. Parte del trabajo es indagar sobre los datos y armar posibles explicaciones en base a ellos.

¿Por qué Total_Taxable_Amount es a veces mayor que Total_Amount?

Cada fila del dataframe es un **item** de una **oportunidad**. Si agrupan por Opportunity_Name y suman los Total_Amount deberían llegar a Total_Taxable_Amount. Si esto no se cumple en todos, es algo para analizar en el TP.

¿Qué hacemos con las monedas distintas?

Las pueden convertir. Si hacen esto de manera estática o dinámica (según la fecha del registro; por ejemplo, el mes y año) depende de ustedes.

¿Qué pasa con las columnas que tienen faltantes de datos?

Hay que ver cuántos faltan, si imposibilita el análisis, si son importantes para el problema, si se puede rellenar con algún valor que tenga sentido, si es necesario descartarla. Es algo a analizar como parte del TP1.

¿Cada fila sería como un "producto" que se vende dentro de cada oportunidad?

Sí.

¿Qué es exactamente el TRF?

Frigorías (una "medida" de energía).