

Informe de desarrollo

Predicción de éxitos en oportunidades comerciales

<https://metadata.fundacionsadosky.org.ar/competition/20/>

Fundación Sadosky - AlixPartners - ECI

Participante :

Esp. Ing. Dawoon Choi

E-mail : dawoon.choi.330@gmail.com

Twitter: [@dauny90](https://twitter.com/dauny90)

Linkedin: <https://www.linkedin.com/in/dawoon-choi-85230054/>

Introducción

El presente informe consiste en describir todos los aspectos relacionados al desarrollo de un modelo basado en aprendizaje automático supervisado con capacidad de predecir la probabilidad del éxito de las oportunidades comerciales de la empresa "Frio Frio".

Para la resolución se propone una metodología de trabajo de 4 etapas, que a grandes rasgos incluye, desde análisis descriptivos y estadísticos de los datos, preprocesamiento, feature engineering y desarrollo del modelo predictivo.

Los datos fueron provistos por la competencia y se pueden descargar desde el siguiente link:

<https://metadata.fundacionsadosky.org.ar/competition/20/>

Para llevar a cabo el desarrollo del modelo se utilizó principalmente:

- Anaconda
- Python 3.7
- JupyterLab
- GoogleColab
- Pandas;Numpy;PandasProfiling;Seaborn
- MIFlow
- LightGbm;Xgboost

Metodología de trabajo¹:

La metodología empleada para el desarrollo del proyecto se divide en 4 etapas:

1. Exploración de datos
2. Limpieza/pre procesamiento de datos
3. Ingeniería de variables (Feature engineering)
4. Entrenamiento y Evaluación de modelos

¹ Se adjunta el script para validar el proceso descrito en este informe (**Script.ipynb**)

Exploración de datos y Limpieza/pre procesamiento de datos:

El trabajo de exploración de datos consiste en conocer el contenido y la magnitud de los datos que utilizaremos para el modelo predictivo. En esta etapa se debe poder contestar a preguntas como:

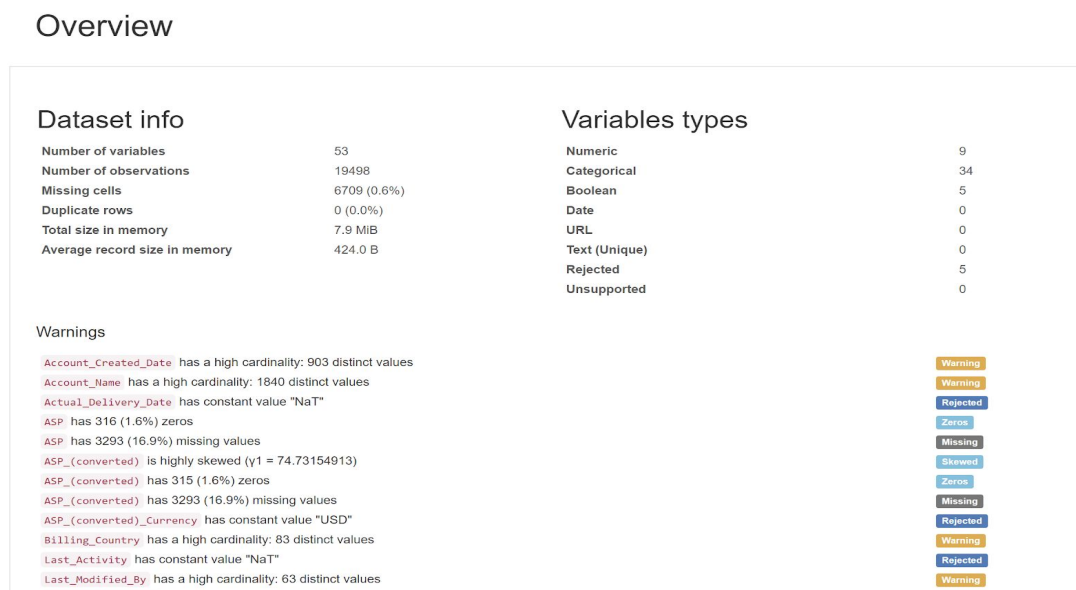
- A. ¿Cuántas filas tienen las tablas?
- B. ¿Cuántas columnas tienen las tablas?
- C. ¿Hay valores nulos?
- D. ¿Cómo son las correlaciones entre las variables?
- E. ¿Qué tipos de datos tienen las tablas?
- F. ¿Cómo son las distribuciones de las variables?
- G. ¿La variable de respuesta (target) buscado está balanceada?

Esta fase es crucial ya que ayuda a entender el esfuerzo necesario en la limpieza y pre-procesamiento de datos.

Para poder contestar las preguntas anteriores se utilizó una librería muy interesante llamada pandas profiling (<https://github.com/pandas-profiling/pandas-profiling>) que genera como output un archivo .pdf interactivo.

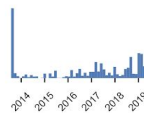


Dicha herramienta es muy útil, dado que con una simple línea de código genera un archivo .pdf interactivo que permite entender la dimensión de la base de datos de manera **gráfica y estadística**.²

Algunas capturas:

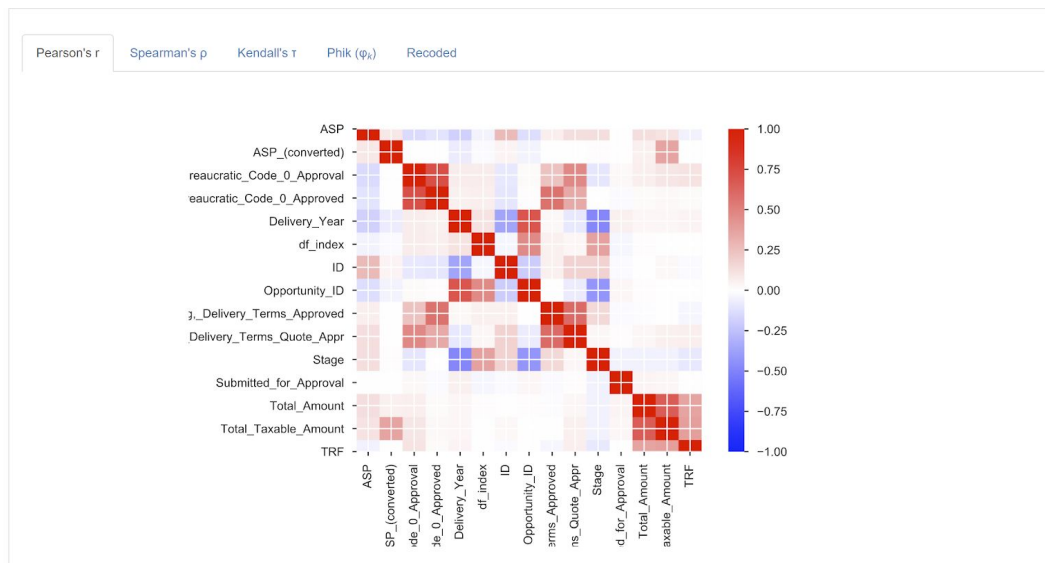


² Se adjunta los los informes interactivos resultantes de este punto (**EDA_TRAIN.html** y **EDA_TEST.html**)

Variables

Account_Created_Date Date	Distinct count 343 Unique (%) 13.4% Missing (%) 0.0% Missing (n) 0 Infinite (%) 0.0% Infinite (n) 0	Minimum 2013-07-27 00:00:00 Maximum 2019-04-23 00:00:00	
Account_Name Categorical	Distinct count 508 Unique (%) 19.9% Missing (%) 0.0% Missing (n) 0	Account_Name_25 121 Account_Name_266 91 Account_Name_1613 60 Other values (505) 2279	
Account_Owner Categorical	Distinct count 35 Unique (%) 1.4% Missing (%) 0.0% Missing (n) 0	Person_Name_64 262 Person_Name_8 246 Person_Name_43 210 Other values (32) 1833	

Correlations



Esta etapa permitió entender rápidamente el contenido de la base y tomar algunas decisiones:

- Se decide eliminar las siguientes variables dado que no aportan información (Variables correlacionadas entre sí; Variables que contienen el mismo valor para toda la columna):
 - ID
 - Actual_Delivery_Date
 - ASP_(converted)_Currency
 - Last_Activity
 - Prod_Category_A

- Month
 - Total_Amount_Currency
 - Delivery_Year
 - Opportunity_Name
2. Se observó que hay una cantidad mínima de registros donde la variable Stage tiene valores distintos a “Won” y “Lost”. Para el cual se probó dos alternativas:
 - a. Eliminar dichos registros
 - b. Considerarlos como “Lost” dado que no hay garantía de que sean “Won”.

Se decide avanzar con la opción “b” dado que performa mejor para el modelo final.

3. Se detectan algunas variables de tipo “String” que contienen información sobre fechas.
 - a. Account_Created_Date
 - b. Opportunity_Created_Date
 - c. Quote_Expiry_Date
 - d. Last_Modified_Date
 - e. Planned_Delivery_Start_Date
 - f. Planned_Delivery_End_Date
4. Se detectan dos variables numéricas que contienen strings.
 - a. Price
 - b. Sales_contact_no

Feature Engineering:

En esta etapa se realiza la tarea de enriquecimiento de la estructura de datos recibida en el punto anterior, esto significa, generar nuevas variables convirtiendo las ya existentes. Ejemplos:

- A. A partir de las fechas, generar columnas como, mes, día de la semana, semana del año, etc.
- B. A partir de las variables categóricas, generar columnas dummies de 0/1
- C. A partir de las variables continuas, discretizarlas.
- D. A partir de las distribuciones de las variables, agrupar aquellos valores de poca frecuencia.

Acciones llevadas:

1. Dividir el set en dos grandes conjuntos.
 - a. Aquellas variables que tienen los mismos valores en la misma oportunidad.
 - b. Aquellas variables que cambian por producto sobre la misma oportunidad.
2. Para el primer conjunto se realizan las siguientes acciones:
 - a. Crear una variable que represente la cantidad de productos por oportunidad
 - b. Para todas las variables fechas se abre por:
 - i. Mes
 - ii. Día

- iii. Dia de la semana
 - iv. Quarter
 - v. Semana del año
 - vi. Bimestre
 - c. Se crean variables que representan restas entre fechas:
 - i. Last_Modified_Date - Account_Created_Date
 - ii. Opportunity_Created_Date - Account_Created_Date
 - iii. Last_Modified_Date - Opportunity_Created_Date
 - iv. Quote_Expiry_Date - Opportunity_Created_Date
 - v. Last_Modified_Date - Account_Created_Date
 - vi. Quote_Expiry_Date - Last_Modified_Date
 - d. Para tratar los valores no numéricos de la variables **Price** se crean dos variables:
 - i. La primera donde los valores no numericos fueron reemplzados por un default (-1)
 - ii. La segunda donde se reemplza por la media de los valores numericos.
 - e. Para la variable Sales_Contract_no se siguió la misma estrategia que el punto 4.
 - f. Para las variables de tipo “objctct” que vendrían a ser string se crean las variables dummies
3. Para el segundo conjunto se realizan las siguientes acciones:
- a. Para todas las variables fechas se abre por:
 - i. Mes
 - ii. Dia
 - iii. Dia de la semana
 - iv. Quarter
 - v. Semana del año
 - vi. Bimestre
 - b. Se crean variables que representan restas entre fechas
 - c. Se crean mínimos,máximos,medias,medianas,sumas agrupadas por “opportunity id” de las variables numéricas
 - i. TRF,Total_Amount,ASP,ASP_(converted),Price,Total_Taxable_Amount
 - d. Se crean variables dummies para variables categoricas:
 - i. Product_family
 - ii. Product_name
 - iii. Delivery_Quarter
4. Se concatena los dos conjuntos nuevamente y se obtiene un solo dataframe donde una oportunidad está resumida y representada en un solo registro.
5. Finalmente del dataframe del punto 4, se eliminan aquellas variables que presentan multicolinealidad,se probaron distintos cortes y el valor que mejor perforó fue de 0.9.
6. Como producto del proceso los datos para train quedan con esta cantidad de filas y columnas:
- a. **Train: 9841;2838**

b. Test : 1567;2838 ³

Entrenamiento y Evaluación de modelos

Esta etapa consiste en entrenar los modelos variando los hiperparametros que tiene cada uno , con el objetivo de encontrar el algoritmo con mejor performance, sobre los datos preparados en las fases anteriores.El algoritmo elegido es **Lightgbm** dado que es el mas utilizado en las competencias internacionales y comprobado su eficacia.En ese contexto también se probó el modelo **XGboost** pero se decide descartar dado que el tiempo de corrida es mayor a lightgbm y un performance similar.

Para la evaluación del modelo,como la competencia indica, se usa **logloss** para medir la predictibilidad del modelo.

Con el objetivo de encontrar los mejores hiperparámetros se decide entrenar el modelo utilizando **grid search** con **cross validation** de 5 folds.Tambien se probó utilizar el método típico de dividir el dataset en 70% para train y 30% para test, el performance no fue más alentador que la estrategia de CV.

Los parámetros en cuestión son:

- max_depth
- lambda
- num_leaves
- feature_fraction
- bagging_fraction
- bagging_freq
- learning_rate

El entrenamiento es corrido en **Google Colab**,con el objetivo de aprovechar el procesamiento en la nube.Cuando se identifica la combinación de hiperparametros que mejor performa se entrena el modelo con todo el set completo de train y se hace la evaluación sobre el set sin target.

La mejor combinación encontrada de hiperparametros es la siguiente:

```
{ 'bagging_fraction': 0.9500000000000002, 'bagging_freq': 16, 'feature_fraction': 0.2, 'lambda': 0.15000000000000002, 'learning_rate': 0.1, 'max_depth': 10, 'metric': 'binary_logloss', 'num_leaves': 17, 'objective': 'binary'}4
```

También se probó la estrategia de “votación” entre los mejores modelos obtenidos.

Además,cabe mencionar que esta última etapa fue iterativa junto a la anterior (feature engineering) donde se fue probando distintas creaciones/eliminaciones de variables para obtener un performance competitivo.

³ Se adjunta el dataset final utilizado para entrenar el modelo(df_final.csv)

⁴ Se adjunta el output final subido a la plataforma para la validación,cuyo logloss resulta 0.07738 (0.07738.csv)

Con el objetivo de tener trazabilidad de código, parámetros y resultados se utiliza la librería **MIFlow** que permite el registro de tales características.

Conclusiones

Desde en lo personal fue una competencia divertida donde pude poner en práctica distintas técnicas que fui desarrollando a lo largo de estos años. Como comentario general de estas competencias, dado por el tiempo incluso, se deja muy bien preparada la base de train y testing para los competidores, siendo el trabajo previo de limpieza y preparación, el trabajo que más dedicación y horas demanda para un cientista de datos.

Definitivamente hay puntos que no pude explorar que podrían convertirse en oportunidades interesantes para mejorar la métrica. En lo particular explorar fuentes de datos que no necesariamente salen de las bases transaccionales de la empresa sino de otras para enriquecer las variables. En este sentido creo que se podría relacionar la base de la competencia con enriqueciendo con:

1. La situación económica de cada país del cliente
2. Condiciones climáticas de cada país del cliente
3. Situación financiera de cada cliente
4. Proyección de crecimiento de cada cliente
5. Magnitud de la empresa del cliente

También sería interesante tener otras variables a la operación de la empresa “frio”:

1. Cantidad de contactos realizados desde la empresa al cliente.
2. Información del empleado que lleva la cuenta (años de experiencia, historial previo de oportunidades concretadas, etc)
3. Información acerca de los productos, cómo fue su historial de compra.

Otra idea interesante sería hacer la votación con los resultados de los otros competidores para aún mejorar la métrica.