Next Best Offer

Caso práctico de venta cruzada

Daniel Marin Santonja

Índice

1. Presentación
2. Objetivos
3. Planteamiento del caso de uso
4. Limpieza y preparación de los datos
5. Modelos empleados
6. Calibración de los resultados
7. Referencias

1- Presentación

¿Quién soy?

- Daniel Marín Santonja 
- daniel.marinsant@gmail.com 
- Graduado en Matemáticas por la UAM.
- Exalumno del máster de Afi de Data Science y Big Data (2018-2019).
- Actualmente consultor de Data Science en Minsait.
- Atleta y friki en mi tiempo libre. 

¿Quiénes Somos?

Minsait es la unidad de Indra especializada en innovación y transformación digital, que aborda end to end el reto de la digitalización combinando capacidades especializadas de consultoría y tecnología junto con productos propios o de terceros.

minsait An Indra company

Áreas de actividad

Administración Pública y Sanidad

Energía e Industria

Servicios Financieros

Telco y Media

Consultoría Negocio

Tecnologías Avanzadas

Gestión Empresarial

Soporte y Aplicaciones

Desde la Práctica IA & Datos buscamos asegurar la excelencia tecnológica y mantener un alto grado de especialización.

Indra Open University

Un innovador modelo de Universidad Corporativa que se adapta a las nuevas necesidades de formación.

Udemy

Acceso gratuito con la cuenta corporativa a todos los cursos de Udemy.

Certificaciones oficiales de las plataformas Cloud

Azure

AWS

Databricks

Google Cloud

¿Qué hacemos en Prosegur?

Algunos de los proyectos desarrollados:

- Text Analytics
- Ageo
- Churn
- Armado Inteligente
- Segmentación de clientes
- **NBO – Venta cruzada**

2- Objetivos

Objetivos:

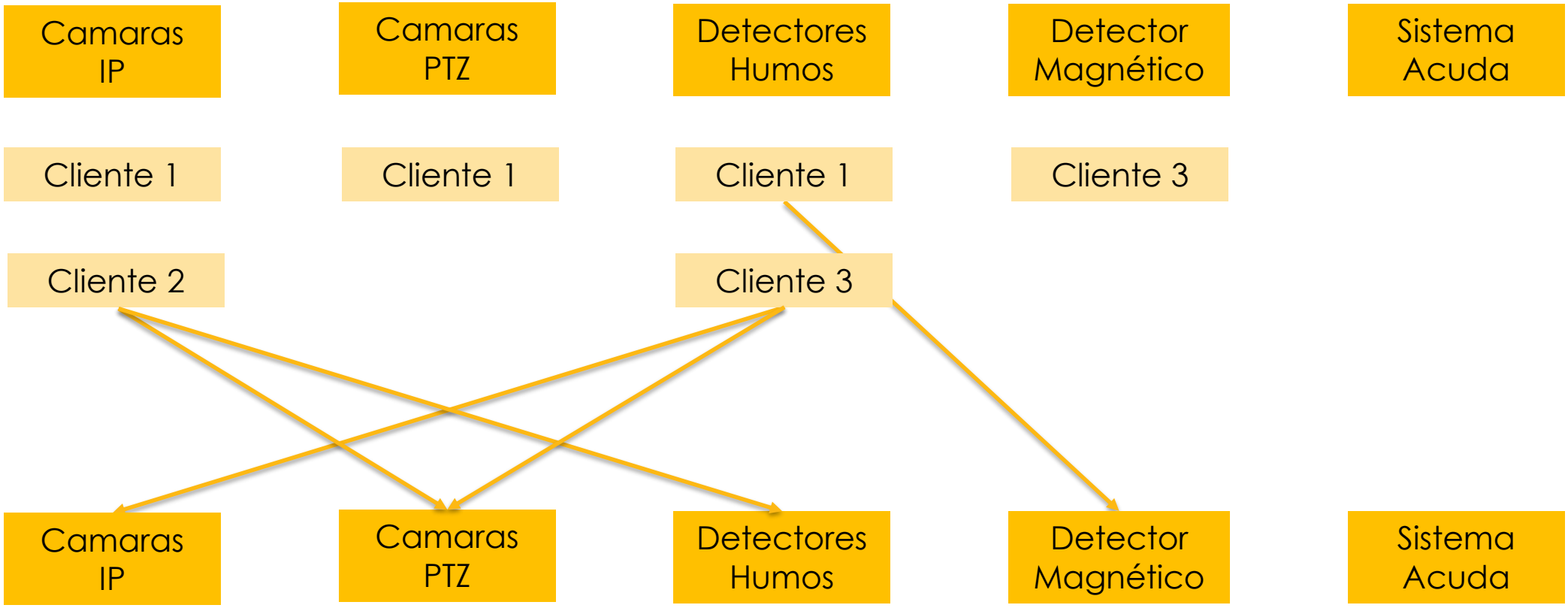
- Entender el proceso realizado para el desarrollo de un proyecto real de venta cruzada a clientes.

- Aprender algunas técnicas menos comunes dentro del mundo del Data Science, como son:
  - Optimal Binning.
  - Mean encoding.
  - Calibración de modelos.

3- Planteamiento del caso de uso

Planteamiento del caso de uso:

Productos



Producto recomendado

4- Extracción, tratamiento y limpieza de los datos

Tratamiento y limpieza de los datos

- Generación del tablón de acuerdo a las necesidades de Negocio.
- Imputamos los NA.
- Tratamiento de los outliers y las categorías poco frecuentes empleando Optimal Binning.

Optimal Binning

EL binning es una técnica de preprocesado que se usa para disminuir los efectos de los outliers, en el caso de las variables numéricas, y de las categorías minoritarias, en el caso de las variables categóricas.

Esto se hace segmentando las variables en función de la Target del dataset. Si son numéricas, en tramos, y si son categóricas, uniendo varias categorías. Esto se puede hacer con diferentes tipos de targets.

Ventajas:

- Evita que los outliers desbalanceen las variables. Nos genera unos datos más robustos.

Inconvenientes:

- Hay que saber cuando usarlo. En algunos casos, puede ser contraproducente y hacernos perder información.

Tratamiento y limpieza de los datos

- Generación del tablón de acuerdo a las necesidades de Negocio.
- Imputamos los NA.
- Tratamiento de los outliers y las categorías poco frecuentes empleando Optimal Binning.
- Se balancean las clases de la variable objetivo para que haya 2/3 de ceros y 1/3 de unos.
- Partición de los datos en train y test (80-20)
- Tratamiento de las variables categóricas con mean encoding
- Se eliminan las variables con una correlación muy alta.

Mean encoding

Es una alternativa al one-hot encoding. En vez de partir las variables categóricas en dummies, reemplaza las categorías por el valor medio de la variable objetivo para dicha categoría.

Ventajas:

- Controla la dimensionalidad del dataset.
- Visualiza de forma más clara la información.

Inconvenientes:

- Si no se tiene cuidado, se puede provocar sobreajuste al modelo.

Variable
Cat 1
Cat 2
Cat 1
Cat 3
Cat 2

Opción 1 | Opción 2

Variable	Mean enc	Dummy 1	Dummy 2	Dummy 3
Cat 1	0.27	1	0	0
Cat 2	0.45	0	1	0
Cat 1	0.27	1	0	0
Cat 3	0.1	0	0	1
Cat 2	0.45	0	1	0

5- Modelos empleados

Modelos empleados

Se probaron diversos algoritmos supervisados:

- **XGBoost Classifier.**
- Random Forest Classifier.
- Regresión logística.
- Gradient Boosting Classifier.
- Extra Tree Classifier.

6- Calibración de los resultados

Calibración de los resultados

Un **modelo calibrado** es aquel en el que el valor estimado de probabilidad puede interpretarse directamente como la confianza que se tiene de que la clasificación predicha es correcta.

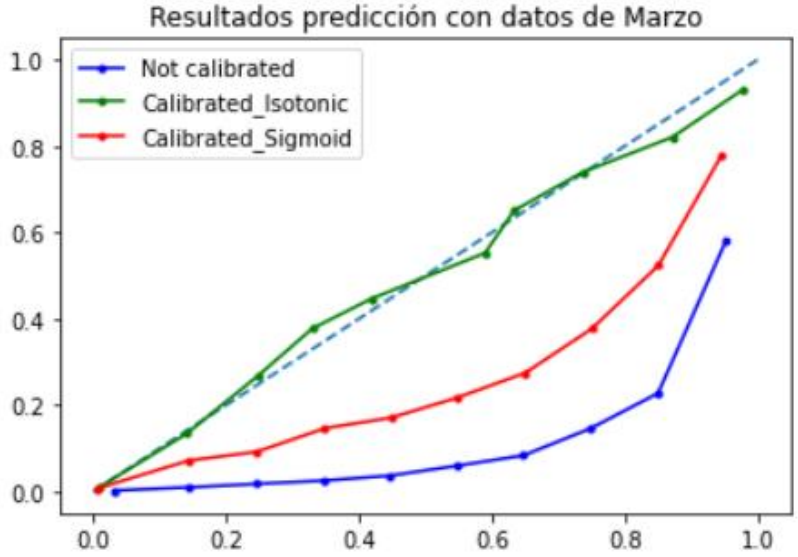
Los modelos por defecto no devuelven unos resultados calibrados, por lo que empleamos un calibrador, que permite que las probabilidades obtenidas de un modelo de Machine Learning para se acerquen a las que se pueden observar en la realidad.

Ventajas:

- Comparar probabilidades entre dos targets distintas.
- Resultados más fiables.

Desventajas:

- La calibración no siempre funciona.
- Depende mucho del volumen de los datos y de la distribución de la target.



7- Ordenación de los resultados

Ordenación de los resultados

En último lugar, una vez tenemos los modelos de cada producto para cada cliente desarrollados, empleamos una función para ordenar a dichos clientes y saber cuáles priorizar, y qué ofertarles:

Cliente	Producto	Prob Orig	Prob Calib	Orden
932985	Acuda	0,887	0,801	1
932985	Magnetico	0,902	0,793	2
932985	Humos	0,765	0,502	3
932985	Cam PTZ	0,205	0,170	4
932985	Cam IP	0,952	0,903	YA

Referencias

Referencias

- http://gnpalencia.org/optbinning/
- http://gnpalencia.org/optbinning/tutorials/tutorial_binary.html
- https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
- https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/

Afi Escuela