UNIVERSITYHACK 2018®
DATATHON

salesforce

GRUPO COOPERATIVO CAJAMAR

wefferent

Deloitte.

minsait by Indra

VIEWNEXT AN IBM SUBSIDIARY

Hewlett Packard Enterprise

KABEL

UNIVERSITAT DE BARCELONA

Donkeys

Àlex Escolà Nixon

ToolsGroup

Florent Micand

AMERICAN & BRITISH ACADEMY ABA ENGLISH

Jaume Puigbò Sanvisens

UNIVERSITYHACK 2018®
DATATHON

La competición de analítica de datos más grande de España.
Del 31 de enero al 12 de abril.

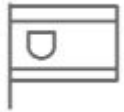
2 RETOS

19 CENTROS

10 EQUIPOS x CENTRO

Inscripciones
Del 15 al 29 de enero de 2018

Fase LOCAL
Del 31 de enero al 21 de febrero de 2018



Fase NACIONAL
Del 1 de marzo al 14 de marzo de 2018

Presentación de mejores trabajos y Fallo del Jurado Nacional
12 de abril de 2018

¿Eres capaz de estimar el **poder adquisitivo** de un cliente?

Objetivo

Encontrar el mejor modelo de regresión mediante el desarrollo de un modelo predictivo que defina con precisión el **poder adquisitivo** de los clientes del Grupo Cajamar.

GRUPO COOPERATIVO CAJAMAR

BCC GRUPO CAJAMAR

cajamar CAJA RURAL

"… puedes utilizar las distintas técnicas de Machine Learning disponibles para este tipo de problemas."

¿Qué hicimos?


Obtención de datos
Análisis exploratorio
Modelizar
Manipulación de variables
Diagnosis / Evaluación de Resultados
Interpretación del Modelo

Condicionantes

1. No hay contacto con el **usuario del modelo**.
    a. Falta de conocimiento del negocio.
    b. Casos de uso limitados.
2. No tenemos especificaciones de las variables.
    a. No es posible clasificar en categorías concretas.
    b. Con tan alto nivel de abstracción no es posible orientar el análisis.

Dataset

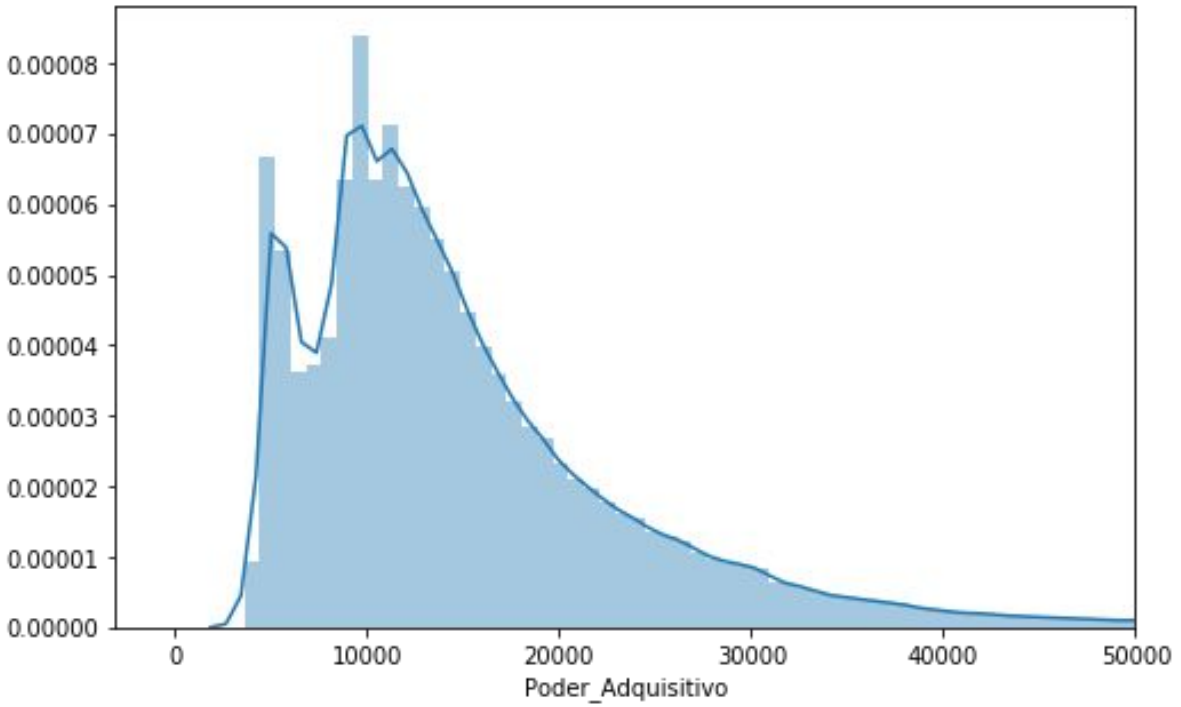
Datos históricos de un grupo clientes, **particulares y autónomos**, del Grupo Cajamar con **88 variables** de productos incluyendo atributos socio demográficos.

El número total de registros es de **363.834** con 89 variables por registro.

Variables en el Dataset

- **ID_Customer**: Identificador de cliente.
- **Socio_Demo_01-05**: Variables sociodemográficas relacionadas con el cliente.
- **Imp_Cons_01-17**: Importe de consumos habituales del cliente en base a sus operaciones con tarjetas y domiciliaciones más comunes.
- **Imp_Sal_01-21**: Importe de los saldos de los distintos productos financieros.
- **Ind_Prod_01-24**: Tenencia de los distintos productos financieros.
- **Num_Oper_01-20**: Número de operaciones a través de los distintos productos financieros.
- **Poder_Adquisitivo**: Variable objetivo, variable numérica que define el poder adquisitivo del cliente.

Distribución del **poder adquisitivo**



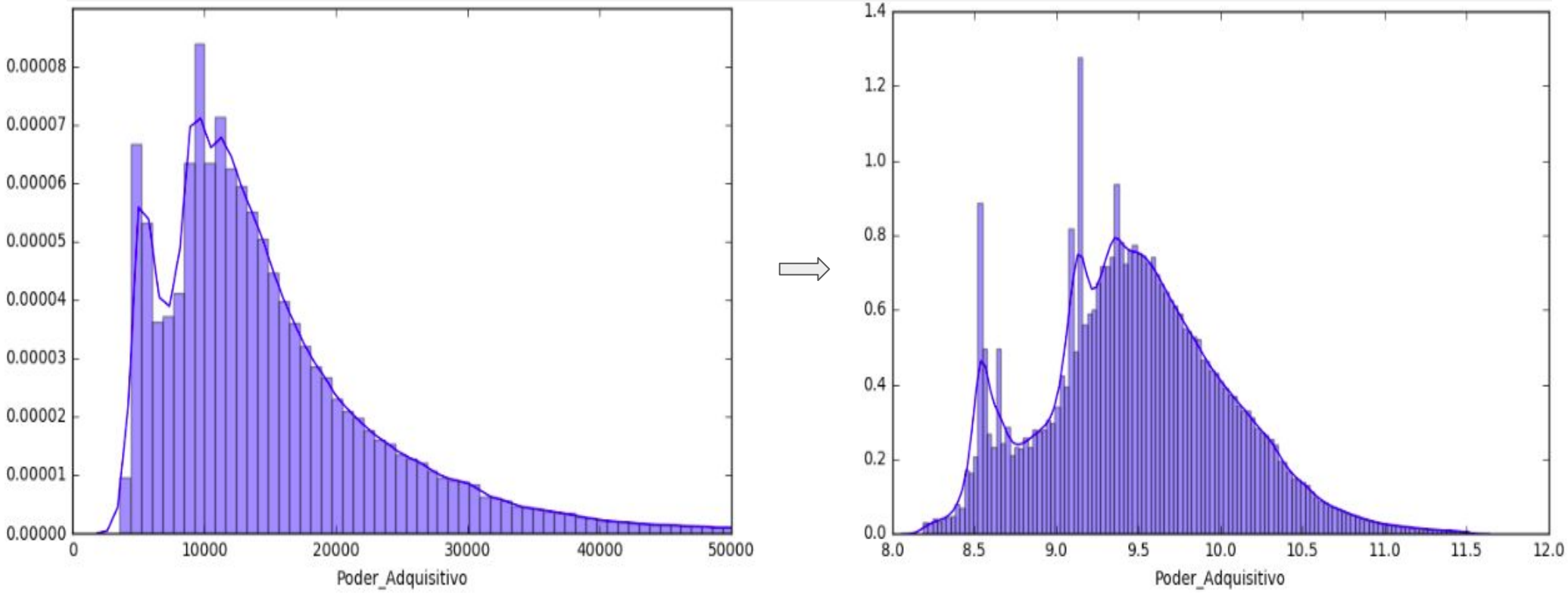
Poder Adquisitivo en Euros

Min	5%	10%	25%	**Mediana**	Media	75%	90%	95%	**99%**	**Max**
3.600	5.140	5.961	9.300	**12.925**	16.421	18.948	27.812	35.013	**63.924**	**5.040.000**

Análisis exploratorio

- Modificación de variables
- Transformación de variables
- Feature Engineering

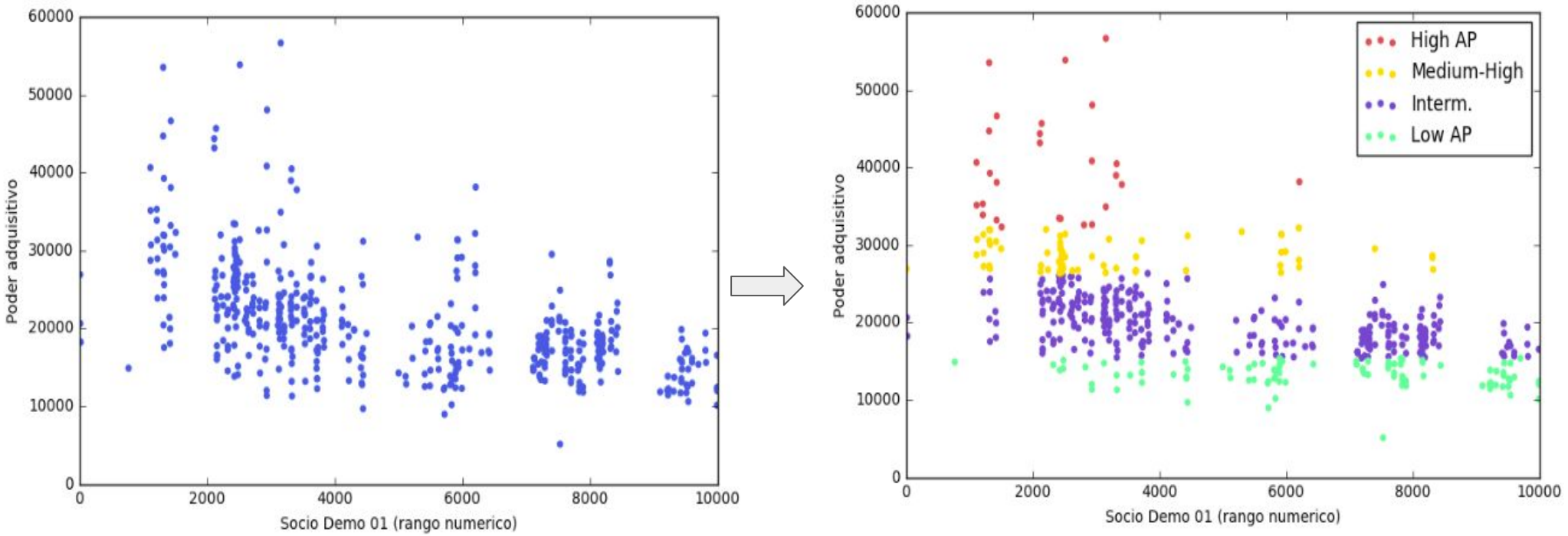
Logaritmo del poder adquisitivo



Análisis exploratorio

- Modificación de variables
- Transformación de variables
- Feature Engineering

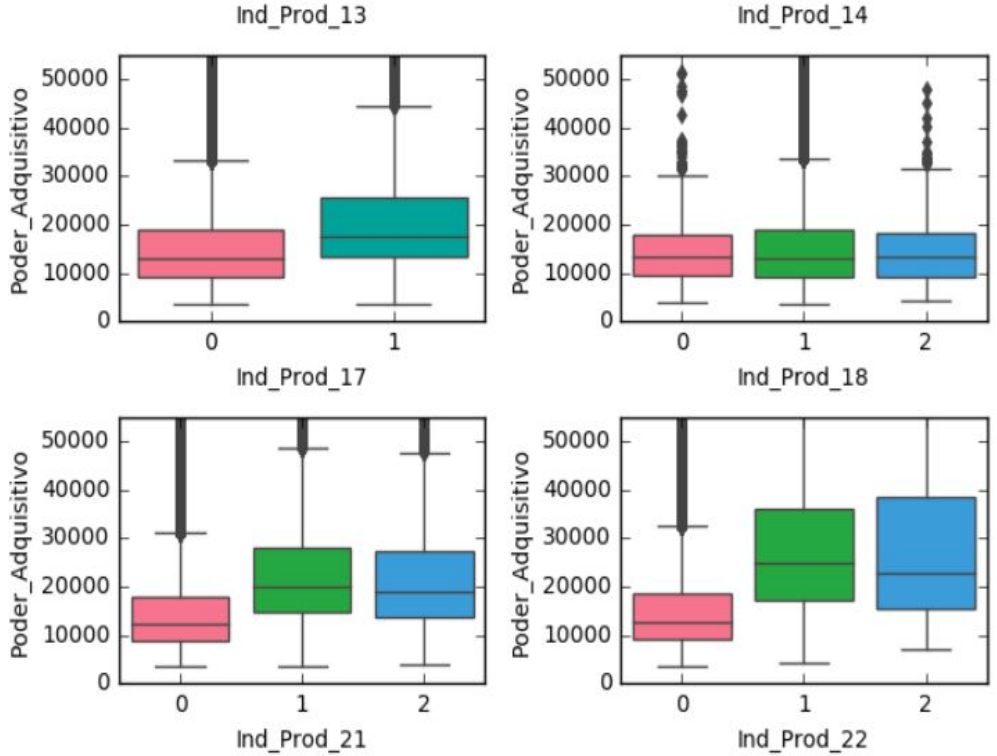
Bucketing de *SocioDemo01* (Categórica)



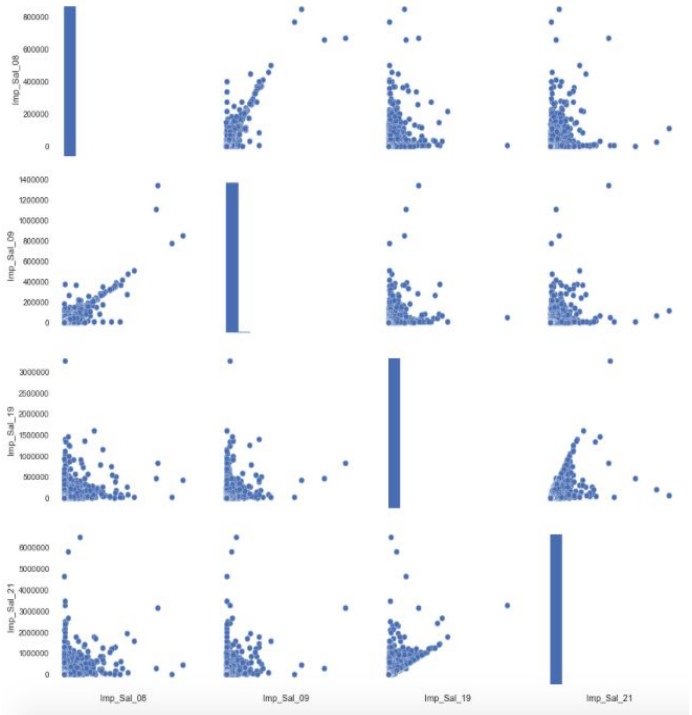
Análisis exploratorio

- Dependencia de las variables con el poder adquisitivo
- Correlación entre variables

Relación de Ind Prod con el PA



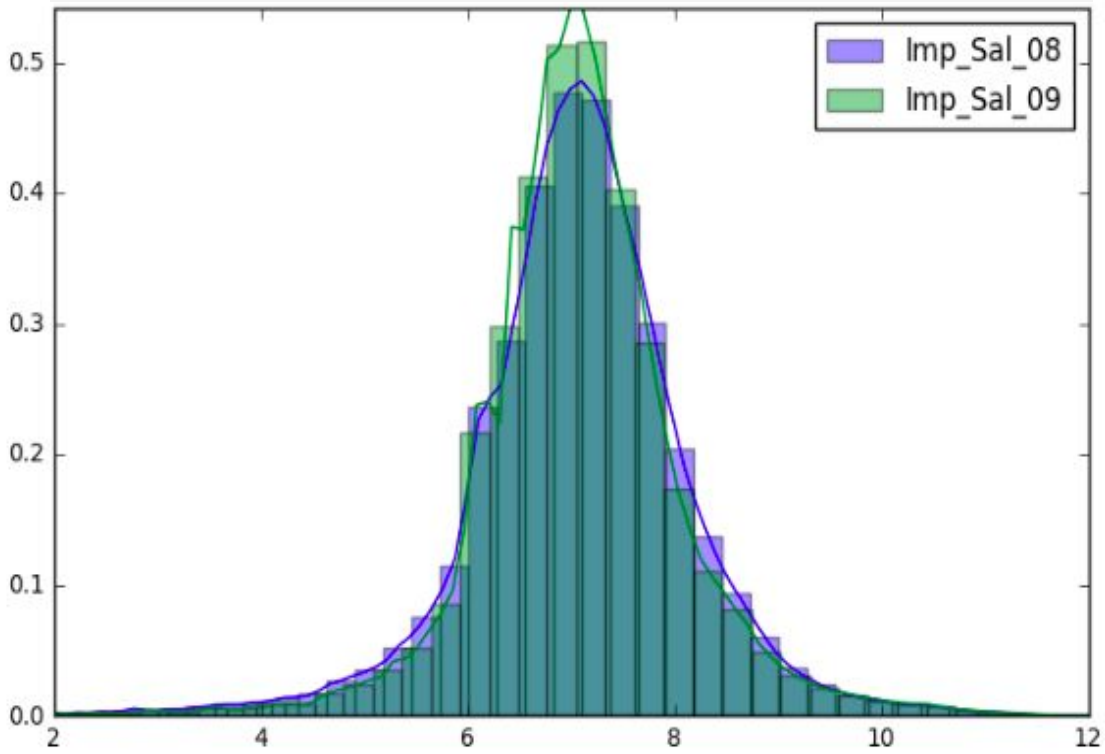
Correlación entre los Imp. de Saldos



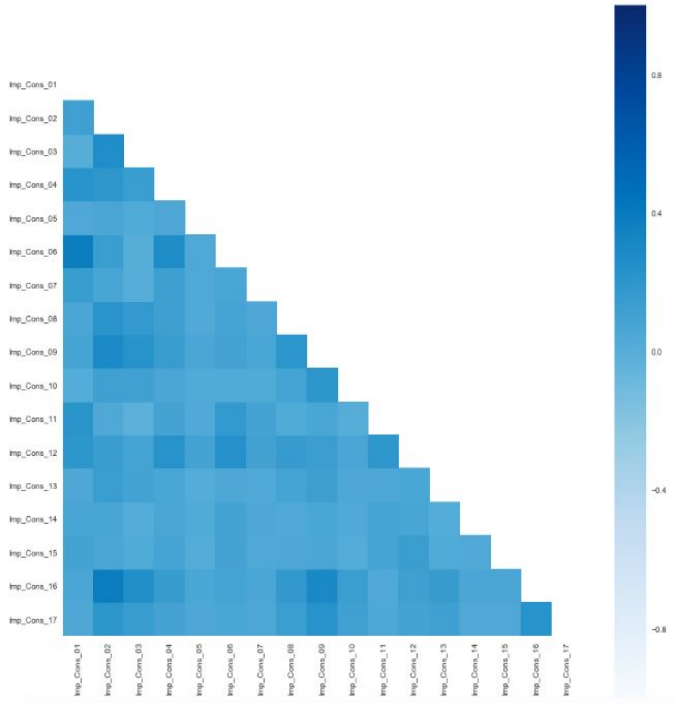
Análisis exploratorio

- Dependencia de las variables con el poder adquisitivo
- Correlación entre variables

Alta correlación Imp Sal 8-9



Matriz de correlación de Imp- Cons



Manipulación de variables

- Manipulaciones mencionadas previamente
  - Log PA
  - Eliminación de variables
  - “2” a “1” en productos
  - Buckets Socio_Demo_01

- Outliers
- Regresión lineal como característica de entrada
- Normalización de variables numéricas
- OneHot Encoding de variables catégoricas no binarias

Modelo

Extreme Gradient Boosting Regressor

- Split train (70%) / test (30%)
- Transformación de variables (formato sparse)
- Ajuste de parámetros mediante Grid Search
- Evaluación
- Entrenamiento sobre el conjunto completo
- Generación de predicciones

~~Random Forest~~ , ~~Dense Neural Networks~~

dmlc XGBoost

Evaluación

1. Mean Absolute Error

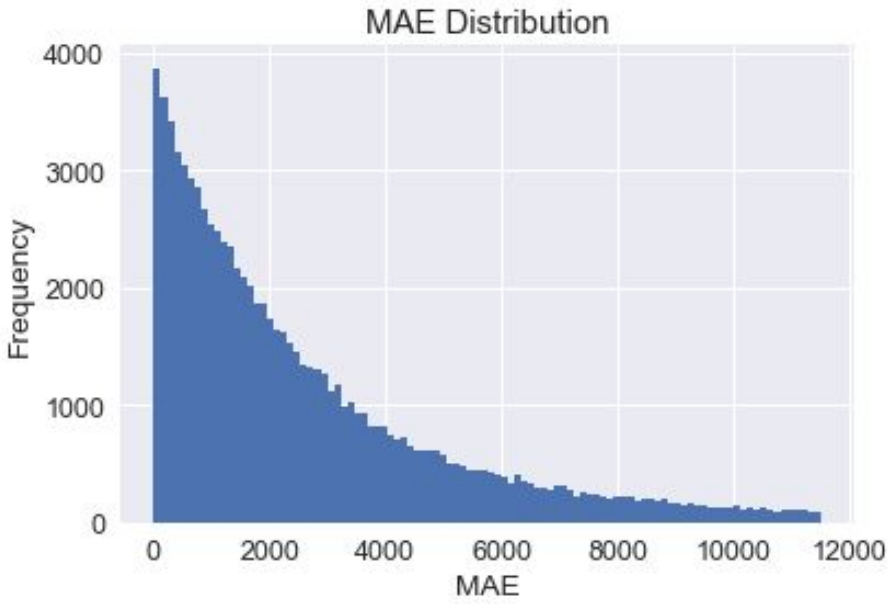
$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{real} - y_{pred}\right|$$

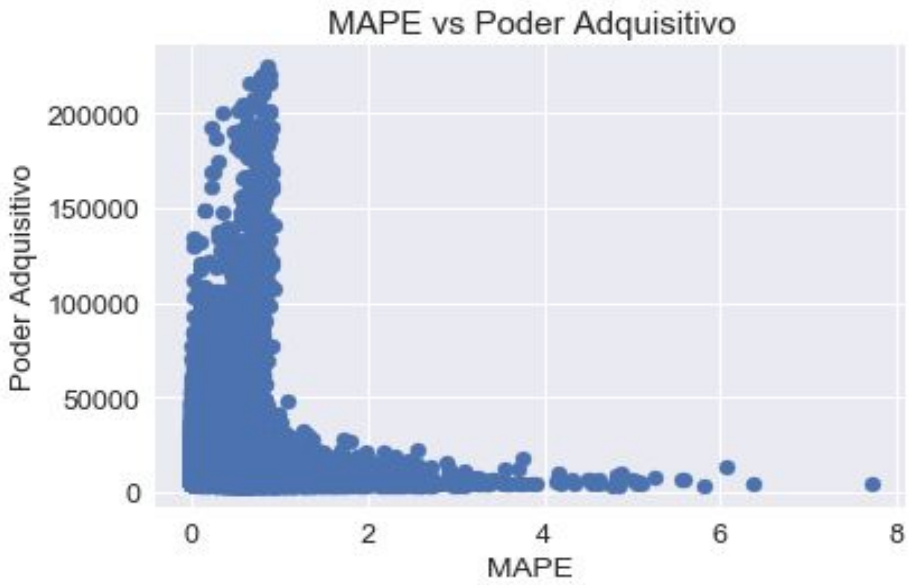
2. Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{real} - y_{pred}\right|}{\left|y_{real}\right|}$$

3. Evaluación por umbrales

Resultados





Métrica	100%	95%
MAE	**3720.5 euros**	**2596.5 euros**
MAPE	**23.7%**	**18.4%**

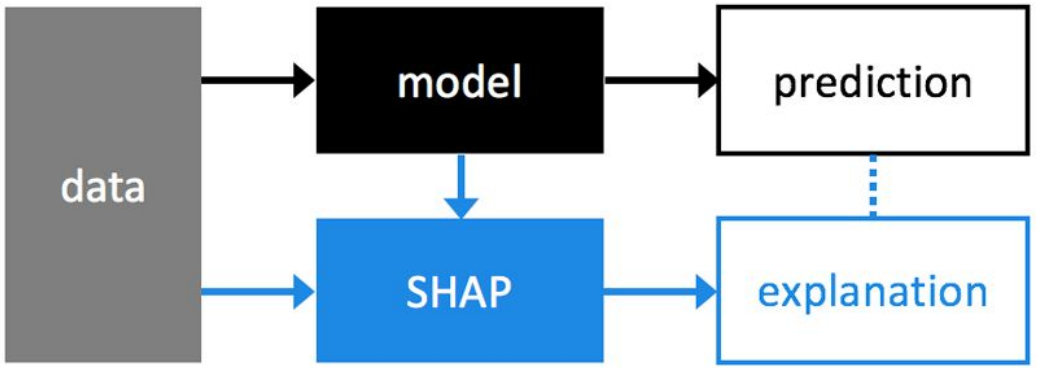
Resultados

Evaluación por umbrales

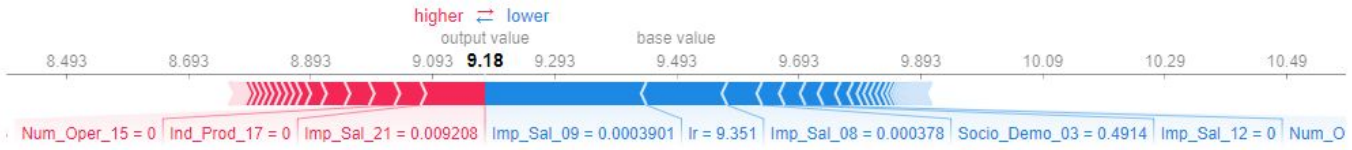
	Muestras de Test	Muestras de Pred correspondientes	Precisión por Umbral (%)
< 5000€	2314	38	1.64
5000€ - 7000€	10245	5074	49.53
7000€ - 9000€	8786	4045	46.04
9000€ - 11000€	12799	8510	66.49
11000€ - 13000€	11743	8317	70.83
13000€ - 15000€	9583	7020	73.25
15000€ - 17000€	7237	5327	73.61
17000€ - 19000€	5571	4211	75.59
19000€ - 21000€	4174	3174	76.04
21000€ - 23000€	3356	2609	77.74
23000€ - 25000€	2708	2163	79.87
25000€ - 27000€	2270	1833	80.75
> 29000€	8104	4278	52.79

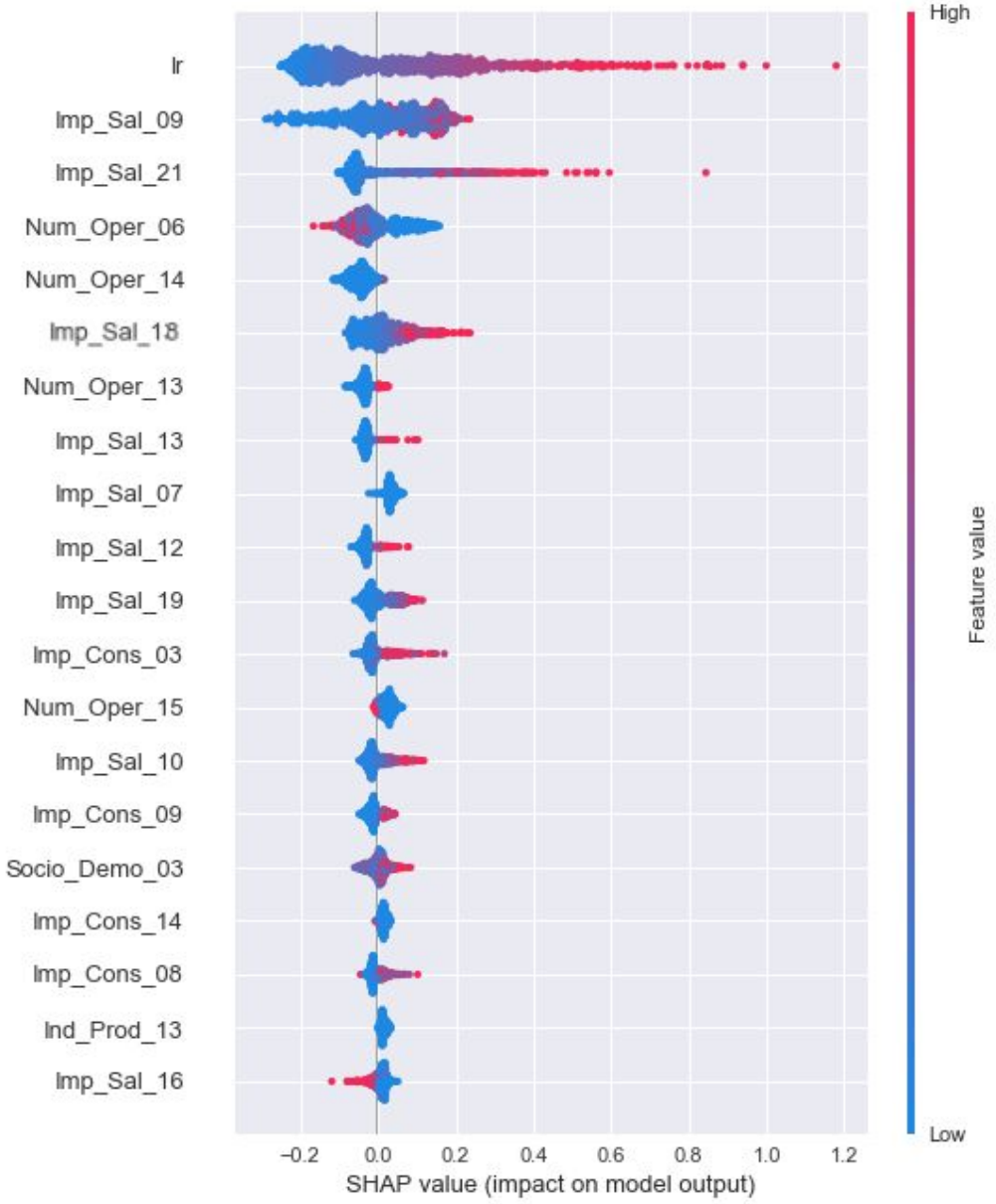
Interpretación

- SHAP: Teoría de juegos + explicaciones locales
- Gain, Split Count, Permutation son inconsistentes.
- Resultados individualizados.
- Mide el efecto al quitar ciertas variables y compara los resultados con los valores esperados usando la teoría de juegos.



Interpretación





Mejoras futuras

- Ampliar el **Grid Search** del modelo
- Clasificador del PA por **umbrales** como modelo alternativo
- Obtener bondad de cada predicción mediante un clasificador cuyo target es el MAPE

Mas información

Notebook: https://github.com/jpuigbo88/Universityhack2018

SHAP: https://github.com/slundberg/shap

Ganadores: http://www.cajamardatalab.com/datathon-cajamar-universityhack-2018/ganadores/

UNIVERSITYHACK 2018®
DATATHON

La competición de analítica de datos más grande de toda España.

Muchas gracias.