Salesforce Predictive Modelling

Equipo Donkeys

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

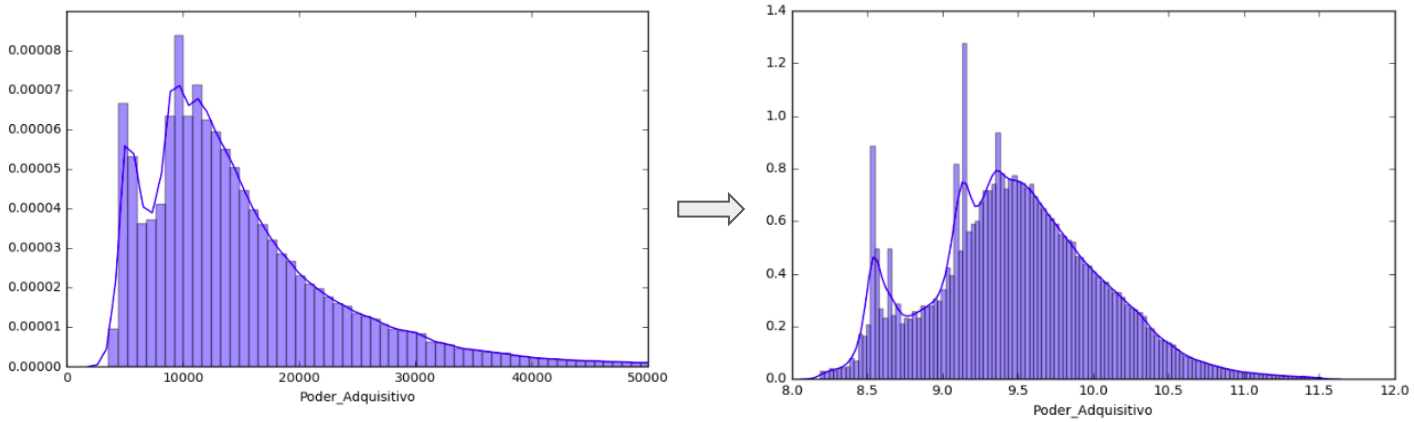
Florent
Micand

Jaume Puigbò
Sanvisens

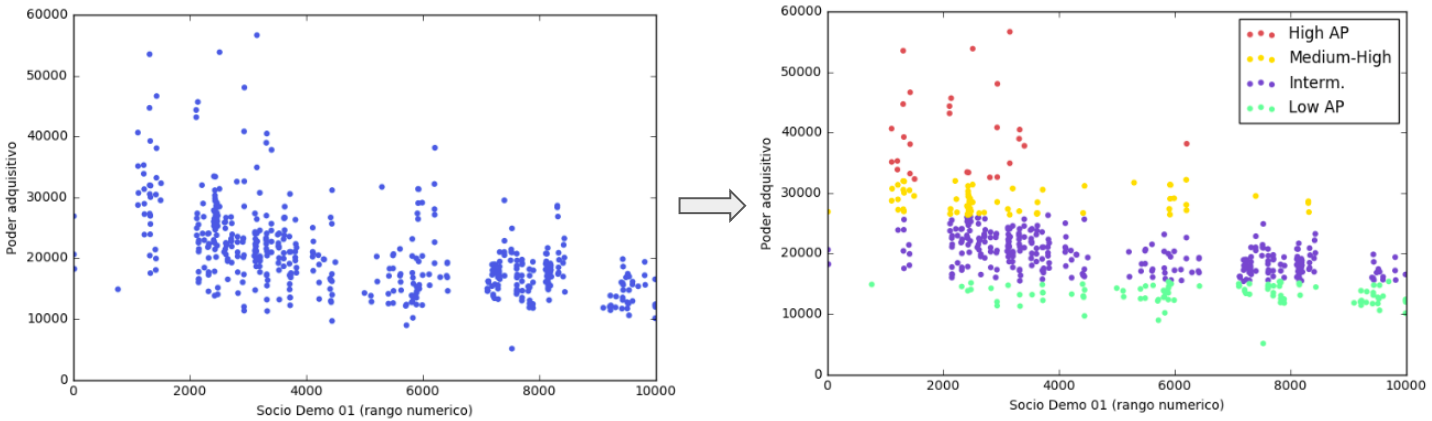
Análisis exploratorio

- Modificación de variables
- Transformación de variables
- Feature Engineering

Logaritmo del poder adquisitivo



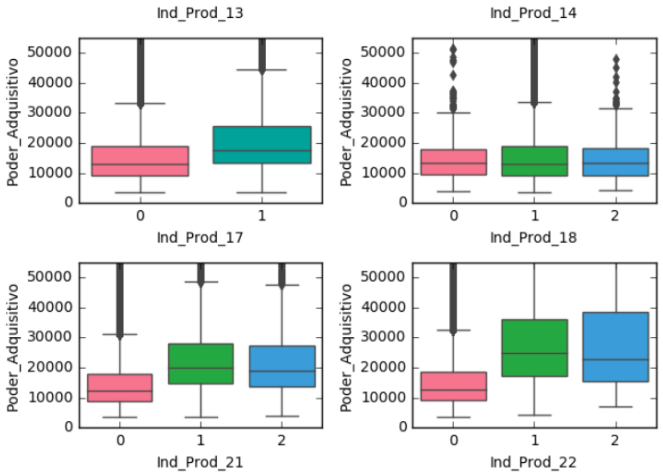
Bucketing de SocioDemo01



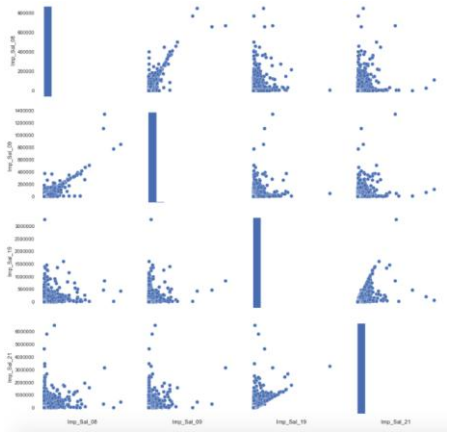
Análisis exploratorio

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

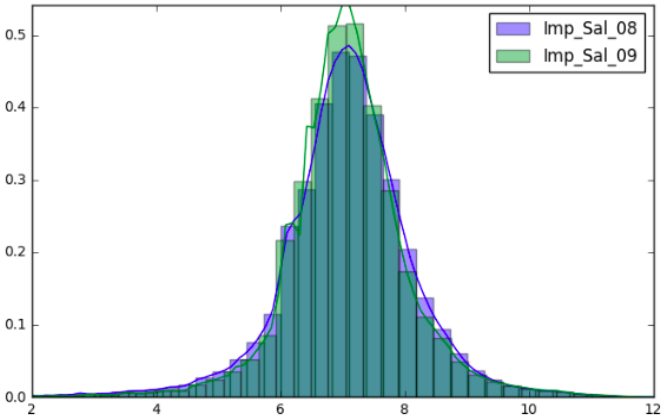
Relación de Ind Prod con el PA



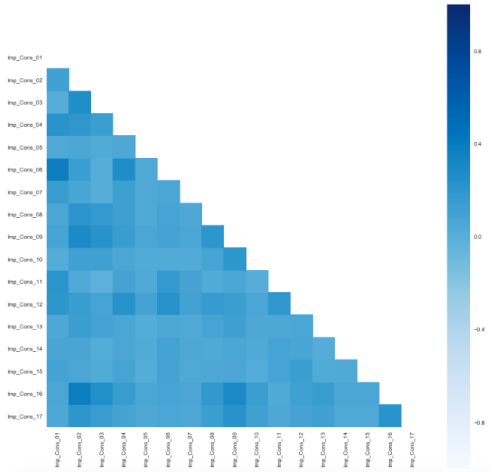
Correlación entre los Imp. de Salarios



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
- Eliminación de usuarios con PA > percentil 99.9
- 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~~

Evaluación

Mean Absolute Error

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

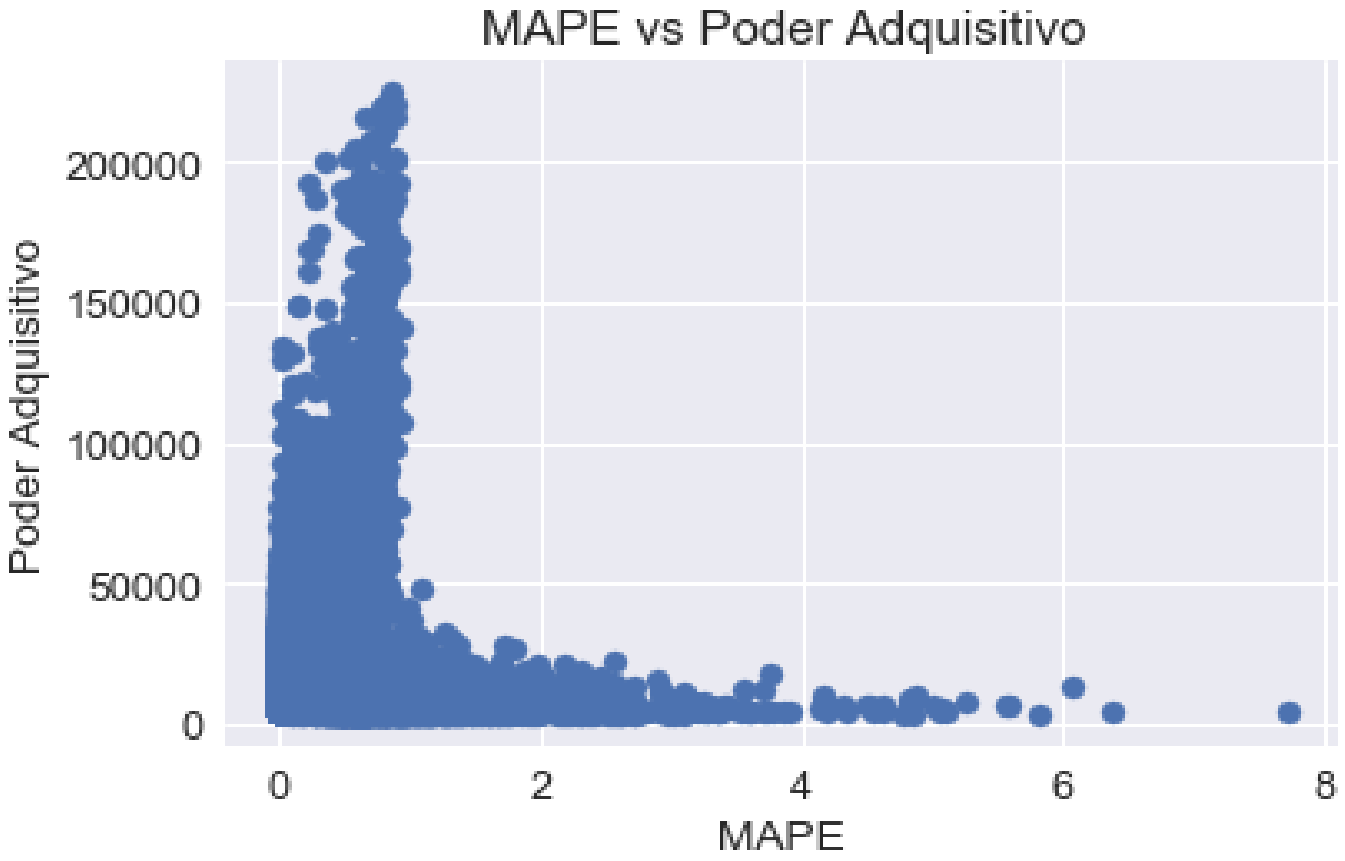
Mean Absolute Percentage Error

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

Evaluación por umbrales

Resultados

MAE (100%)	3720.475520
MAE (95%)	2596.515064
MAPE (100%)	0.237092
MAPE (95%)	0.184063



Resultados

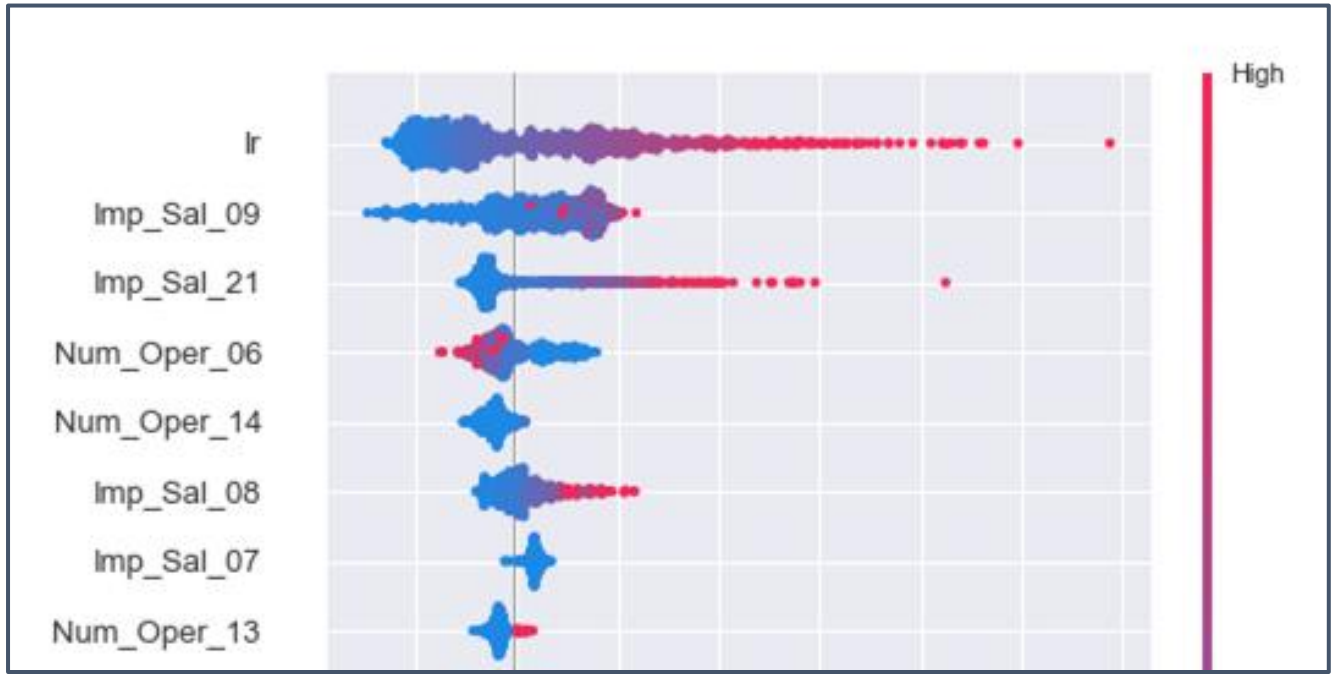
Evaluación por umbrales

	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

Interpretabilidad

Nueva normativa europea de protección de datos (GDPR)

SHAP: Teoría de juegos



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

UNIVERSITYHACK 2018®
DATATHON

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

Muchas gracias.