Salesforce Predictive Modelling

Equipo Always Learning Deeper

UNIVERSITYHACK 2018®
DATATHON

salesforce

GRUPO COOPERATIVO CAJAMAR

wefferent

Deloitte.

minsait by Indra

VIEWNEXT AN IBM SUBSIDIARY

Hewlett Packard Enterprise

KABEL

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

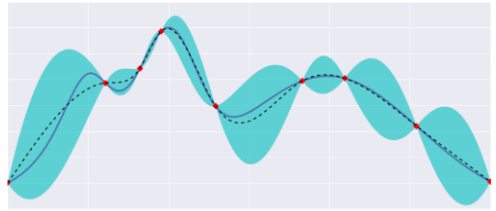
Cajamar UniversityHack 2018

Reto: Salesforce Predictive Modelling

- **Jaime Ferrando Huertas** *Data Scientist @ Polystar, Stockholm*
- **Javier Iranzo Sánchez** *Becario @ MLLP Research Group, UPV*
- **Javier Rodríguez Domínguez** *Becario @ ITI, UPV*

Introducción

Herramientas: Visualización + Desarrollo de modelos + Optimización

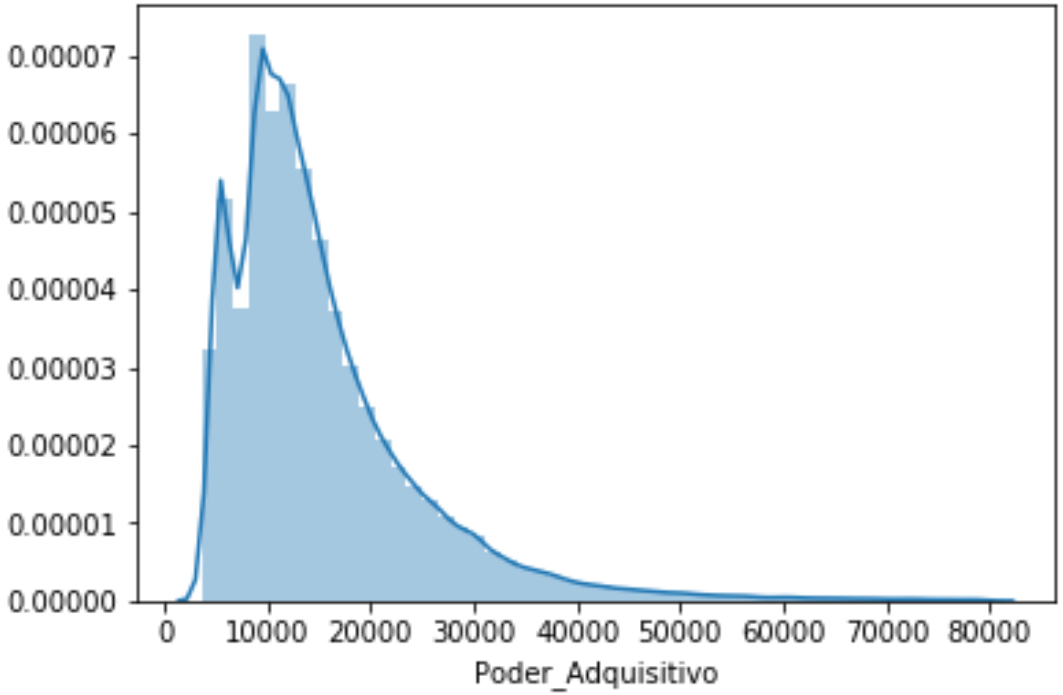


Pandas, Tensorflow, Sklearn, XGBoost, Bayesian Optimization, ...

Análisis de Variables y Preprocesado de datos

Variable a predecir : Poder adquisitivo del cliente.

Distribución con larga cola, algunos clientes tienen un valor que está muy lejos de la media → Fuente de problemas



Análisis de Variables y Preprocesado de datos

Utilizados

- Variables categóricas con muchos valores : Considerar sólo los más comunes
- Transformación variables categóricas -> one−hot
- Eliminar ID Customer

Indiferentes

- No considerar clientes con alto poder adquisitivo
- Diversas técnicas de escalado
- Clustering

Elección de Modelos

Una selección :

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost
- Diferentes Ensembles de los anteriores

Otros modelos considerados : Deep Neural Networks, SVM, KNN, Lasso, Ridge, ElasticNet . . .

Optimización de parámetros

Optimizar hiperpárametros = Optimizar una función.
Problemas:

- Coste
- Espacio
- Interacción entre argumentos

Búsqueda aleatoria: Costoso, muy poco eficiente
Solución: **Optimización Bayesiana**

Ensemble - Modelo final

Por estas razones, nos interesa usar una combinación de modelos

+ Diverso, + Robusto

Los errores de un modelo se compensan gracias al otro

Predicción

Validación cruzada en 5 bloques.

Medidas de rendimiento:

- Raíz del error cuadrático medio (RMSE) → Muy sensible
- Media del error absoluto (MAE)
- Mediana del error absoluto (MAD)

Conclusiones

Error medio de **4318**€
Error mediano de sólo **2028**€

- Hay que estudiar la distribución de los datos
- Hay que seleccionar medidas de evaluación adecuadas
- Optimización inteligente de modelos
- Los modelos basados en árboles obtienen mejores resultados frente a modelos tradicionales y redes neuronales

UNIVERSITYHACK 2018®
DATATHON

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

Muchas gracias.