UNIVERSITAT DE BARCELONA

E-Commerce Stock Prediction

Data Science and Big Data Postgraduate Course 20/21: Capstone Project

Autores: Joan Boronat Ruiz, Albert García López, Pep Martí Mascaro

Índice

1. Definición del problema
2. Análisis de los datos
3. Modelado
4. Conclusiones

- Análisis Exploratorio
- Preprocesado
  - Limpieza
  - Creación de nuevas características

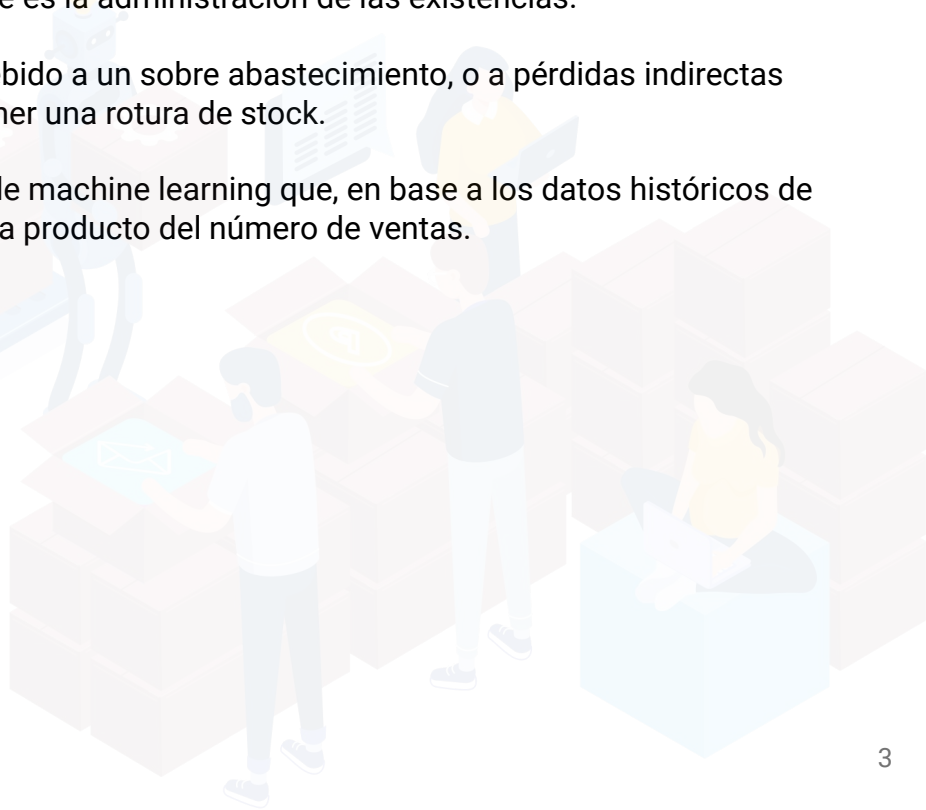
- Métrica de evaluación
- Planificación del modelo base y los modelos predictivos.
- Análisis de los modelos
- Blending

1. Definición del problema: ¿Cuántos Iphones voy a vender mañana?

Uno de los mayores retos a los que se enfrentan las e-commerce es la administración de las existencias.

Una mala gestión puede conllevar grandes pérdidas directas, debido a un sobre abastecimiento, o a pérdidas indirectas debido a las ventas potenciales que no se han conseguido al tener una rotura de stock.

Nuestro objetivo para este proyecto es el de crear un producto de machine learning que, en base a los datos históricos de nuestra tienda online, ofrezca una predicción día a día para cada producto del número de ventas.



El conjunto de datos con el que vamos a trabajar está formado por un dataset de train y uno de test con datos por cada producto y por cada día. En la siguiente tabla (Fig. 1) vemos una muestra aleatória de 5 registros. En datos totales el dataset contiene:

Registros	**Rango de fechas**	**Productos únicos**	**Categoria Uno**
4.045.032	2015-06-01 al 2016-12-31	4.168	13

fecha	id	visitas	categoria_uno	categoria_dos	estado	precio	dia_atipico	campaña	antiguedad	unidades_vendidas
2015-09-08	280286	2	A	127.0	No Rotura	NaN	0	0	NaN	0
2015-06-11	273536	31	A	34.0	No Rotura	15,38	0	0	855.0	3
2016-06-03	313452	75	A	85.0	No Rotura	NaN	0	0	652.0	0
2015-06-20	173088	37	H	316.0	No Rotura	32,93	0	0	1601.0	3
2016-09-25	333726	60	C	158.0	No Rotura	NaN	0	0	553.0	0

Fig. 1 - Muestra aleatória de 5 registros del training dataset

En este gráfico (Fig. 2) podemos apreciar que el número de visitas medio se multiplicó por un factor 5 a principios de 2016 y la gran influencia de los periodos atípicos como Black Friday, Navidad o los Happy Days (periodo de rebajas específico de este e-commerce) que tienen lugar en Julio.

Debido a la gran estacionalidad de los datos deberemos usar modelos de predicción de series temporales como ARIMA o Prophet.

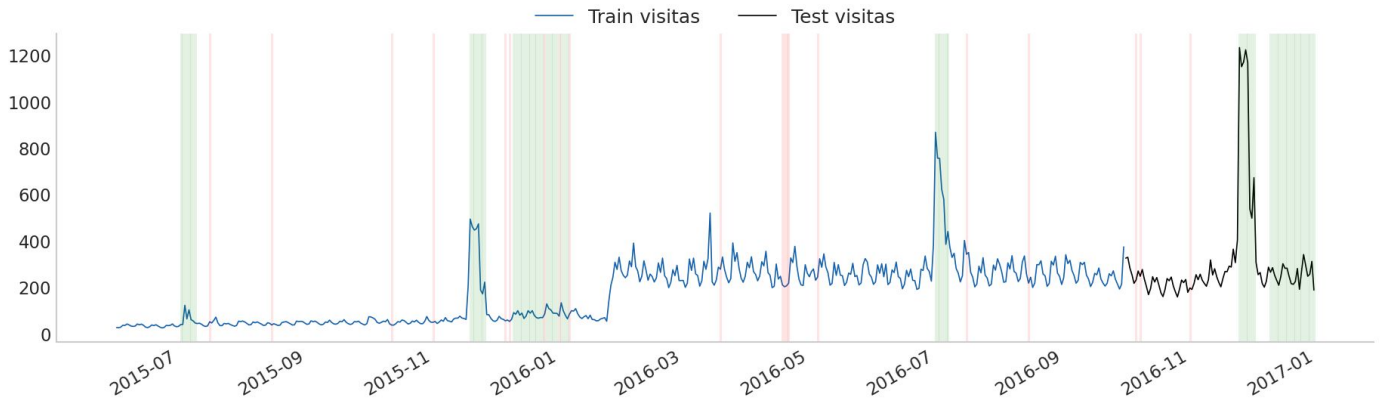


Fig. 2. Gráfico con la media de visitas por dia para el conjunto de train (azul) y de test (negro). Las franjas verdes indican días atípicos con valor 1. Estos pueden tener descuentos y un mayor volumen de ventas es esperado. Las franjas rojas indican días atípicos negativos.

2. Análisis de los datos > Preprocesado

Analizando el dataset encontramos algunas anomalías que debían ser tratadas para realizar el modelado. El preprocesado se resume en los siguientes pasos:

- Eliminar entradas duplicadas: Suponían casi el 50% del dataset de entrenamiento. Las eliminamos para evitar agregar confusión en los modelos predictivos ya que no nos aportan información adicional.

- Definición de los tipos de datos: Para la predicción final, tratamos precio y antigüedad como características numéricas. El resto, como categóricas.

- Estimar missing values: Algunos de los productos no tenían ningún valor en la variable antigüedad. Para ello usamos un K-nearest neighbours para predecir la antigüedad a partir de las características que, durante el análisis, habíamos observado que guardan relación con la antigüedad; principalmente el ID.

- Propagación de valores: Tal y como se describe en el dataset, la variable precio no siempre tiene un valor, para estos casos hay que usar el valor anterior más cercano. El mismo procedimiento fue aplicado para los valores de antigüedad en aquellos productos que tenían algún registro de antigüedad disponible.

2. Análisis de los datos > Preprocesado

- Normalización: Debido a un cambio en la página web a principios de 2016 observamos que las visitas pasan de incrementarse de 1 en 1 a incrementarse de 5 en 5. Por este motivo, hemos multiplicado por 5 la variable visitas para los registros anteriores al 25 de Enero de 2016, teniendo así una magnitud constante en la variable.

- Nuevas características: Tras observar la estacionalidad semanal de las ventas así como comportamientos atípicos en algunos meses del año creamos nuevas características que nos permitan reflejar esta información. Asimismo, extraemos el componente tendencia de las series temporales para cada “categoría uno” que utilizaremos como regresor en el entrenamiento del modelo.

3. Modelado > Métrica de evaluación

Para el entrenamiento de los distintos modelos que hemos usado los datos de los 12 primeros meses del dataset y los 3 meses posteriores para la validación. En cuanto a la métrica, hemos usado una variación de Root Mean Square Error para penalizar las roturas de stock.

$$(0.7 * rRMSE) + (0.3 * (1 - CF))$$

$$rRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}}{\bar{y}}$$

Fig. 3 Métrica a minimizar. Donde CF es el porcentaje de casos favorables.

3. Modelado > Prophet

Prophet es un modelo de predicción de series temporales basado en un modelo aditivo en el cual tendencias no lineales se ajustan a la estacionalidad anual, semanal y diaria además de las fechas concretas, cómo las vacaciones o las rebajas. Hemos escogido este modelo por su versatilidad a la hora de predecir problemas de series temporales como el que nos encontramos, con fuertes incrementos en fechas concretas como los Happy Days en julio, Black Friday en noviembre y Navidades.

Para una mayor precisión hemos creado un modelo distinto para cada uno de los productos. Por contra, hemos perdido la visibilidad sobre las tendencias globales y por categorías.

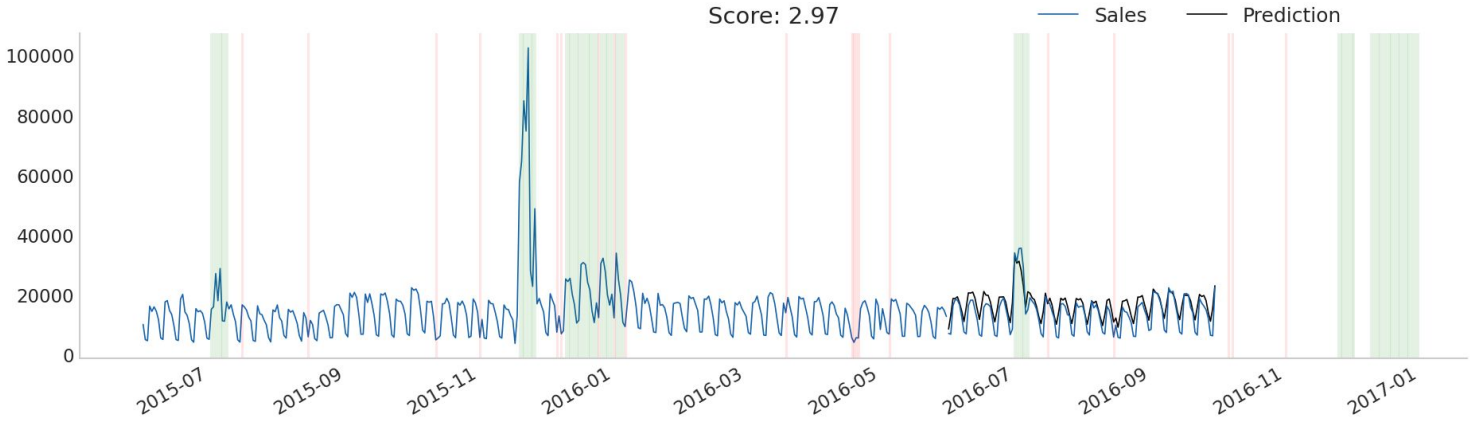


Fig. 4 - Resultado de validación del modelo Prophet

3. Modelado > XGBoost

XGBoost es un modelo de Gradient Boosting muy versàtil y eficiente que nos ha permitido crear una predicción basada en las características del producto. Lo hemos complementado añadiendo como variables categóricas el mes y el día de la semana entre otros. Para entrenar este modelo hemos usado únicamente los mismos meses del año anterior del conjunto de test para evitar que periodos con muchas ventas no presentes en el conjunto de test influyan en la predicción.

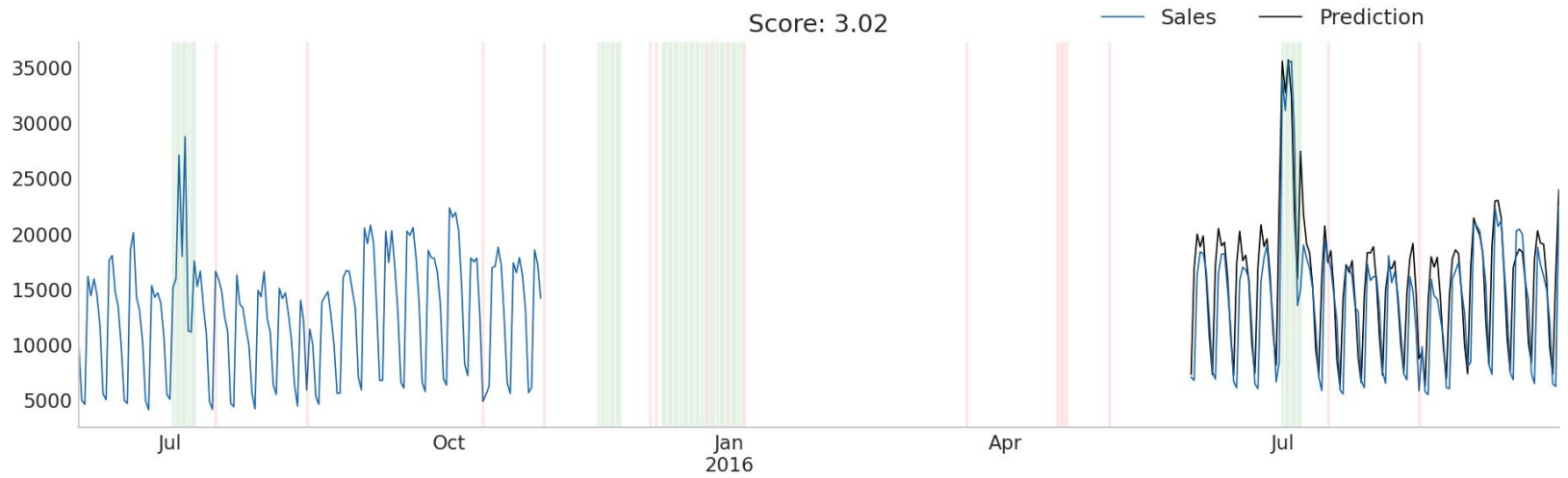


Fig. 5 - Resultado de validación del modelo XGBoost

3. Modelado > LightGBM

LightGBM es un modelo de Gradient Boosting muy versàtil y eficiente que nos ha permitido crear una predicción basada en las características del producto de forma más optimizada y ágil que el XGBoost al añadir un paso previo de inspección de aquellas muestras de datos que aportan mayor información dejando de lado las muestras menos informativas.

Nota: Hemos probado diversas configuraciones de hiperparámetros, siendo el modelo por defecto el que mayor resultado de predicción ha obtenido

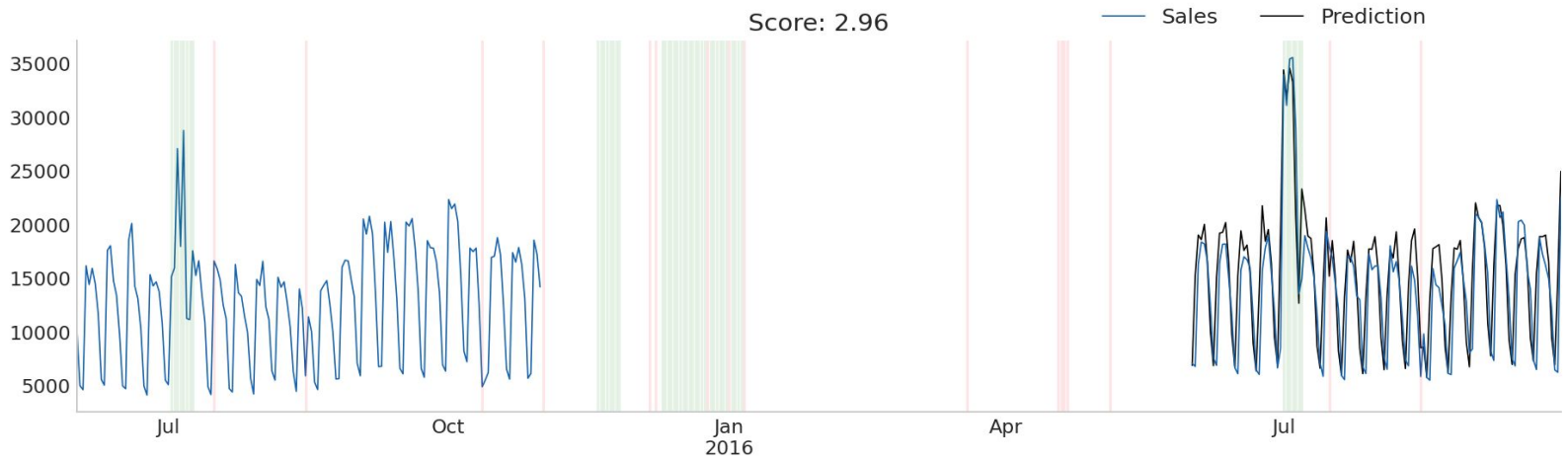


Fig. 6 - Resultado de validación del modelo LightGBM

3. Modelado > CatBoost

CatBoost es un modelo de gradient boosting sobre árboles de decisión. Es un modelo muy parecido a XGBoost siendo su principal diferencia el tratamiento de los datos categóricos y precisamente de ahí viene su nombre. En este caso sin embargo, al tener unas variables categóricas con poca correlación con la variable objetivo, este modelo no ha conseguido mejorar el resultado de XGBoost.

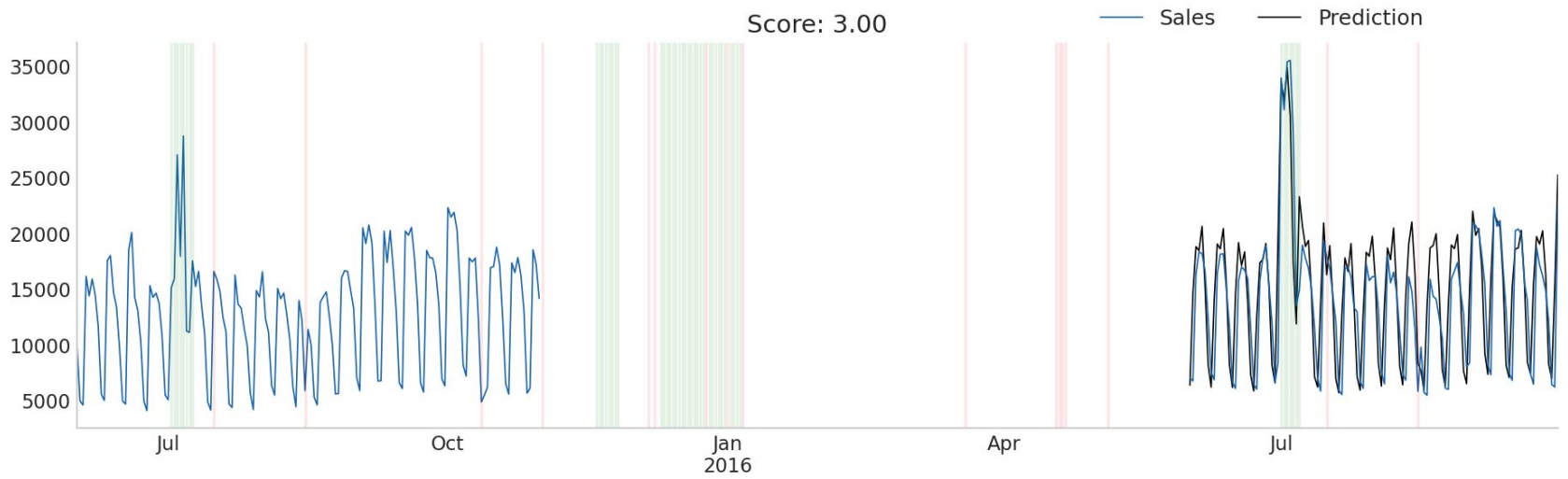


Fig. 7 - Resultado de validación del modelo XGBoost

3. Modelado > Análisis del modelo

Una vez el modelo ha sido entrenado, usamos la librería SHAP para analizar los pesos de las distintas variables en la predicción. En el gráfico (Fig. 8) observamos en el eje de las abscisas el impacto de la variable y estas tienen un gradiente de color de azul a rojo en función del valor de las variables.

En el caso de CatBoost observamos que la variable con mayor peso ha sido las visitas normalizadas y estas tienen un impacto positivo cuando el valor es alto y negativo cuando el valor es bajo.

En este gráfico podemos observar como el precio tiene un impacto negativo en las ventas cuando este es elevado y positivo cuando es bajo.

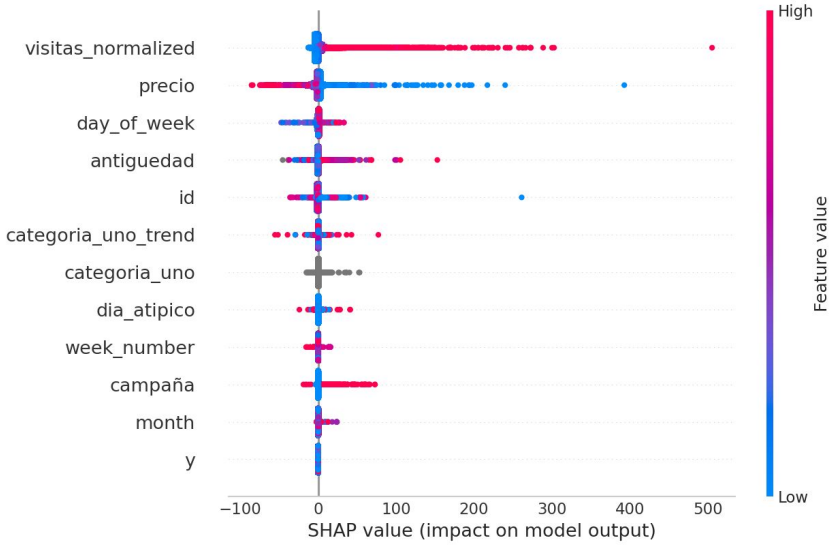


Fig. 8 - Análisis del impacto de las variables en la predicción del modelo CatBoost.

3. Modelado > Blending (XGBoost, Catboost y LightGBM)

Cada uno de los modelos de regresión han explotado diferentes aspectos del dataset con distintas metodologías y resultados muy parecidos. Para minimizar la varianza de estos modelos y mejorar el resultado hemos entrenado los modelos XGBoost, Catboost y LightGBM y los hemos combinado entrenando un meta-model (un modelo de los modelos) cuyas predicciones están basadas en múltiples predicciones obtenidas de cada uno de los modelos.

Para realizar el blending hemos usado una red neuronal de dos capas con función de activación RELu con lo que evitamos predecir valores negativos.

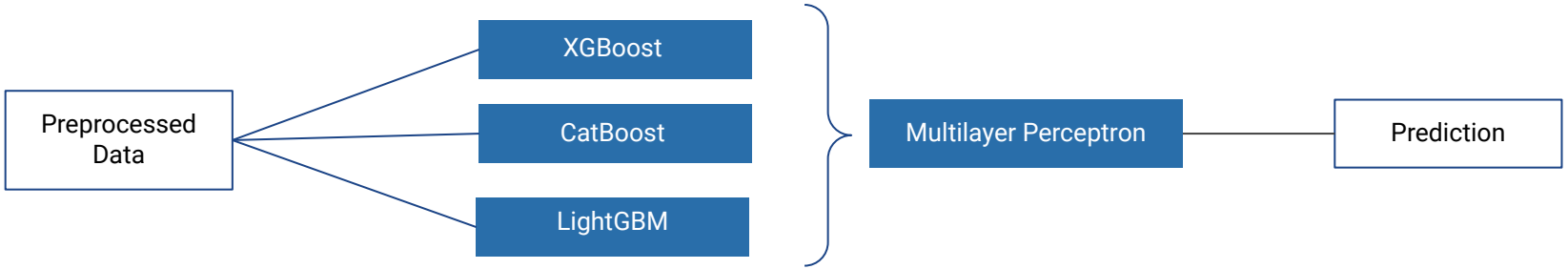


Fig. 9 - Esquema del Blending de los modelos de regresión

3. Modelado > Predicción final con Blending

Hasta ahora hemos visto el modelo de series temporales Prophet, capaz de capturar la tendencia i la estacionalidad, así como los periodos de vacaciones y rebajas. También hemos analizado diferentes modelos de regresión que nos permiten capturar los patrones encontrados mediante el estudio de las variables exógenas y de las nuevas categorías creadas a partir del análisis exploratorio.

Para la predicción final hemos decidido realizar un blending de las predicciones de los modelos que han obtenido mejor resultado. Para el blending hemos usado un modelo de regresión lineal que además de tener como entrada las predicciones de los otros modelos, hemos usado el día de la semana.

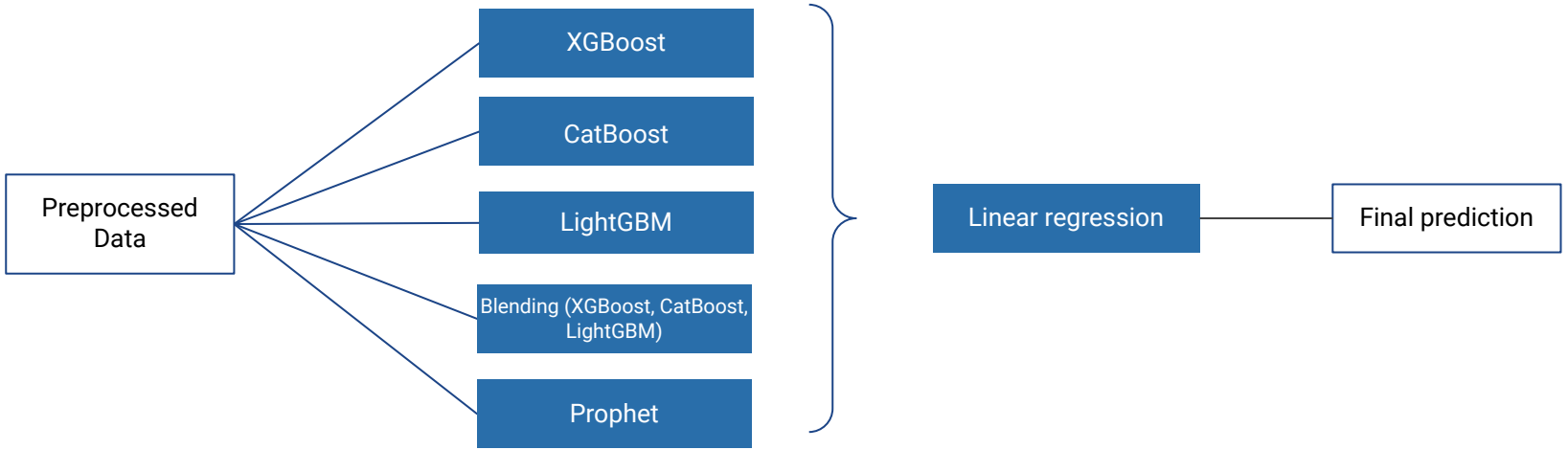


Fig. 9 - Esquema del producto final

4. Conclusiones

Nuestro objetivo principal era el de crear un **producto de machine learning** que, en base a los datos históricos de nuestra tienda online, ofreciera una **predicción día a día para cada producto del número de ventas**.

Nos hemos encontrado un dataset muy informativo y con muchas variables a explorar. Así, nuestro primer objetivo ha sido maximizar la extracción de información de los datos, para ello hemos llevado a cabo un extenso trabajo exploratorio.

Para la planificación del modelado hemos utilizado modelos con diferente naturaleza, se ha estudiado el modelo de series temporales Prophet y diferentes modelos de regresión. Hemos entendido las diferencias principales entre estos modelos, así como sus puntos fuertes y sus debilidades.

Se ha podido comprobar mediante la evaluación de los modelos que todos hacen un trabajo similar a la hora de identificar las variables que influyen más en las ventas de cada producto. Se ha confirmado el buen funcionamiento de la técnica de blending. Esta técnica de blending ha sido utilizada para la predicción final ya que nos ayuda a diluir posibles errores por sobreajuste que cada uno de los modelos individuales tenga.


LIBERTAS PERFVNDET
OMNIA LVCE
UNIVERSITAT DE
BARCELONA