Atmira Stock Prediction

Grupo: JAP Consulting

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

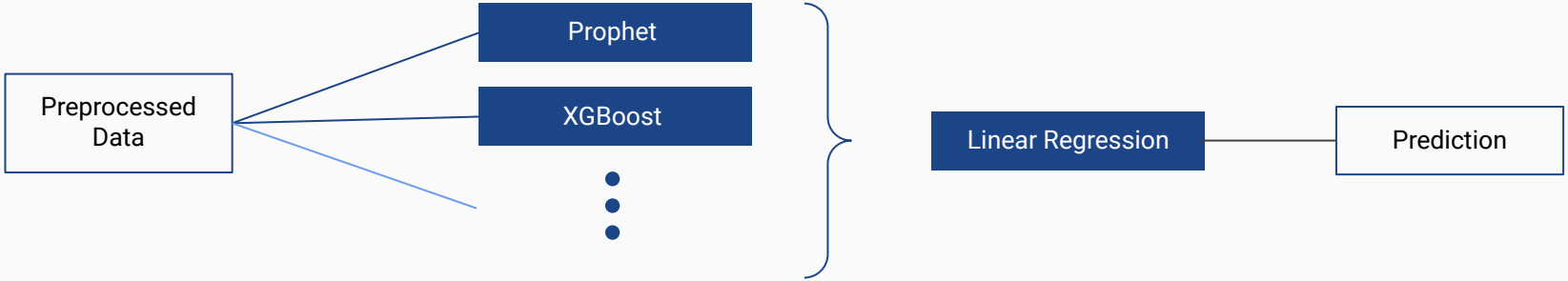
Resumen

Este reto consiste en la predicción de ventas de la eCommerce PCComponentes para los meses de octubre, noviembre y diciembre de 2016 a partir de los datos de los 15 meses anteriores. Como particularidad en este reto contamos con el número de visitas para los meses a predecir. Debido a esto, hemos enfocado el problema desde el punto de vista de una serie temporal y desde el punto de vista de regresión.

- La predicción de la serie temporal nos permite pesar con mayor precisión la estacionalidad de los datos con periodos, como por ejemplo Black Friday o Navidad, con un gran incremento de ventas. El modelo usado en este caso ha sido Prophet. Para una mejor predicción, hemos creado un modelo para cada producto en el dataset.

- Para la regresión hemos usado XGBoost que nos ha permitido predecir las ventas a partir de las características de los productos. Para evitar que los periodos con descuentos influencien en exceso la predicción, hemos entrenado el modelo únicamente con los datos de los mismos meses del año anterior.

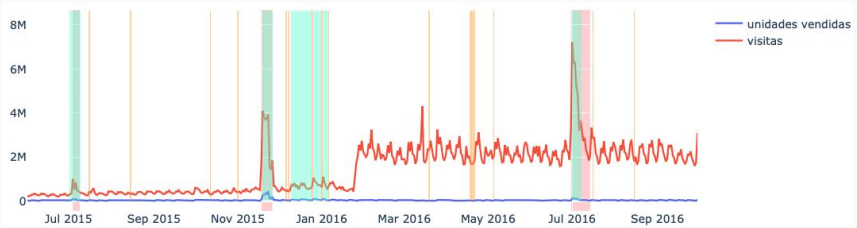
Finalmente hemos aplicado una técnica de blending con los dos anteriores resultados para obtener un resultado final. Para ello hemos usado un regresor lineal. En el siguiente esquema se puede observar una representación de la arquitectura usada en nuestra solución.



Análisis exploratorio

Nuestro principal objetivo para el análisis exploratorio ha sido buscar la mejor manera de **entender** los datos, **probar diferentes estrategias** para nuestro preprocesado final, **extraer** las principales características de las variables, y **definir** cúal es la mejor manera de enfocar el problema a partir de la creación de dos modelos baseline. Para esta parte se ha seguido la siguiente estructura:

1. Visualizar el problema. Una vez dado el primer vistazo al conjunto de modelado y predicción se ha visualizado el problema a través de distintos gráficos. Por ejemplo, en el siguiente gráfico se puede ver la evolución de las suma de ventas y visitas a la página web.



2. Preprocesado. En esta sección se comentan las diferentes estrategias seguidas que nos han llevado a nuestro preprocesado final. Por ejemplo, en esta sección se podrá encontrar los diferentes enfoques que se intentaron para abordar el aumento en las visitas a la página web que se observa en el gráfico.
3. Análisis exploratorio de los datos. Esta sección nos permite investigar cada una de las variables por separado y las principales relaciones entre ellas.
4. Modelos propuestos y no utilizados. Con el objetivo de explorar diferentes maneras de abordar el problema y tener una referencia para futuros modelos se han creado:
    - Un primer modelo muy simple basado en las ventas del año anterior, y características por dia de la semana y crecimiento anual.
    - Un modelo lineal Ridge para agrupaciones de registros en base a “categoría uno”, día de la semana y mes.

Preprocesado

El preprocesado se divide en los siguientes pasos:

1. 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.

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

3. 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.

4. 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.

5. 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.

6. 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.

Selección de modelos

Nuestra propuesta de solución es un Blended de dos modelos:

1. **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 PCDays 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.

2. **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.

A partir de las predicciones de los modelos anteriores, hemos hecho un blending con un modelo de regresión lineal que nos ha permitido obtener la predicción final. En este caso la elección del modelo ha sido principalmente por su simplicidad ya que su única función es la de encontrar el equilibrio entre las predicciones de los modelos anteriores.

Siguientes pasos

Actualmente para el análisis de la serie temporal estamos usando el modelo Prophet cuya sencillez nos ha permitido integrarlo rápidamente, pero también hemos encontrado algunas limitaciones. Por este motivo vamos a probar otros modelos más versátiles como el ARIMA o el seq2seq. De igual modo también probaremos otros modelos para añadir al blending o para sustituir a los actuales.

Aparte de probar otros modelos, también queremos mejorar los existentes usando técnicas de normalización como el Box-Cox sobre nuestra variable objetivo para mejorar la distribución de los errores que actualmente se focaliza, sobre todo, en los productos con pocas ventas. También vamos a explorar la creación de una loss function personalizada para que tenga en cuenta que la rotura de stock penaliza más que el excedente.

Seguir explorando la creación de variables a partir de las variables actuales, así como trabajar en la mejora de la evaluación de nuestros modelos, tanto en la adición de nuevas métricas como MAPE y MAE como en la mejora de los gráficos utilizados.

Finalmente tenemos el modelo de blending. Dicho modelo tiene como objetivo entender en qué casos cada uno de los modelos de nivel 0 predice mejor las unidades vendidas. El modelo que usamos actualmente es un regresor lineal, pero vamos a explorar posibles alternativas como un random forest.

Como resumen de los siguientes pasos:

1. Métodos de normalización. Mejorar la distribución de las features.
2. Feature engineering. Explorar la creación de nuevas categorías o agregaciones de las existentes en función de las características de los productos
3. Nuevas métricas de evaluación y visualización de los errores.
4. Custom Loss function. Aumentar el peso del error en caso de predicciones a la baja que nos puedan llevar a rotura de stock.
5. Explorar nuevos modelos para el posterior stacking. En concreto el seq2seq y el ARIMA.
6. Probar otros modelos para el blending p.ej. Random Forest.


UNIVERSITYHACK 2021
DATATHON