Cajamar UniversityHack 2021

Daniel Corral Ruiz, Antonio Pascual Hernández, Diego Senso González

16/3/2021

Contents

1. Breve resumen del trabajo desarrollado.

El objetivo del trabajo desarrollado es predecir las ventas de diferentes productos de PcComponentes para cada día a partir de los datos de ventas históricos.

A continuación se explicarán los contenidos de los dos notebooks realizados: el notebook "exploratorio", en el cual se ha realizado un análisis exploratorio de los datos; y el notebook "predicción", en que se llevan a cabo distintos modelos para predecir las ventas de los productos.

2. Resumen del análisis exploratorio llevado a cabo.

El objetivo del presente notebook es realizar una primera exploración y análisis descriptivo de los datos, así como realizar la ingeniería de variables necesaria para preparar los datos y trabajar con ellos en el notebook de predicción.

Se ha comenzado estudiando el total de productos que forman parte de cada categoría. El mayor número de productos pertenecen al grupo A, seguido de los grupos K y F. Los grupos con un menor número de productos son N, O y D.

En cuanto al "estado" de los productos, la mayoría de los productos están en estado de "No Rotura". Los productos cuyo estado es "Rotura" o "Transito" representan una menor cantidad de productos.

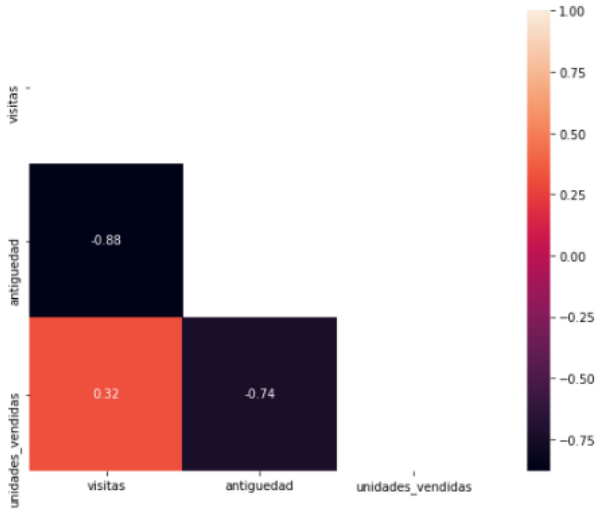
Tras distinguir los datos entre días por la demanda de productos, se puede observar que 208659 días tuvieron alta demanda, mientras que 70907 tuvieron baja demanda. Si se estudia si los productos se encontraban en campaña o no, 9146 días había productos en campaña.

Podemos observar que la mayoría de los productos cuya antigüedad se encuentra por debajo de los 1200 días. Podemos observar que también existen productos con una antigüedad superior a los 5000 días.

Se ha podido observar como la gran parte de los productos obtienen menos de 15000 visitas.

Tras estudiar la correlación entre las variables, se ha podido observar un correlación positiva entre las variables "visitas" y "unidades_vendidas", por lo que a más visitas mayor número de unidades vendidas. La

segunda relación que podemos comprobar es cómo la antigüedad del producto se relaciona negativamente con el número de unidades venidas, por lo que a más antigüedad menor número de unidades vendidas.



Una vez concluido el análisis exploratorio, se ha continuado con la ingeniería de variables.

3. Resumen de la manipulación de variables y su argumentación.

3.1. Tratamineto inicial y NA.

Para comenzar, se han cambiado de tipo algunas de las variables para trabajar con ellas posteriormente. A continuación se han pasado a tratar los NA. En cuanto a los NA del precio, se han completado con los datos del precio del mismo producto el día anterior. Los NA de la variable "antiguedad" han sido sustituidos por la media de la antigüedad de su misma categoría_dos. Las observaciones que no tenían un registro anterior y su precio era NA, han sido eliminados.

3.2. Generación de nuevas variables.

Se ha generado un nueva variable denominada "atípico_campaña" que tendrá el valor de 1 cuando un día sea de alta demanda y haya campaña. El motivo para generar esta nueva variable es que se entiende que un producto puede ver muy incrementadas sus ventas si el día es de alta demanda y, además, si hay campaña.

También se ha creado otra variable denomida "cobro". El valor de esta variable tomará valor 0 si el día del mes se encuentra entre el día 5 y 25.

Finalmente, se han generado variables dummies a partir de las variables categóricas "categoría_uno", "estado" y "dia_atípico". Cada una de las opciones de estas variables genera nuevas variables que tomarán valores 1 y 0 si las observaciones cumplen con las características.

4. Justificación de la selección del modelo.

En primer lugar, se han separado los datos de modelar en train y test. Se ha escogido el train con un 1% de los datos disponibles, para que el modelo tenga más datos para entrenar.

Se ha realizado un Lasso para estudiar la importancia de las variables en el modelo. Se ha obtenido que la variable creada "atipico_campaña" es aquella con un coeficiente más elevado, seguido de "campaña". Algunas de las dummies creadas también tienen cierta importancia a la hora de explicar las unidades vendidas.

A continuación, se ha procedido a construir y entrenar diferentes modelos. Los modelos que se han probado para intentar estimar el número de unidades vendidas son:

- Gradient Boosting,
- XGBoost,
- Arboles de decisión,
- Random Forest

Todos los modelos se ha entrenado con diferentes hiperparámetros y configuraciones. Posteriormente, se han utilizado estos modelos para predecir sobre el 1% de test que se había reservado previamente, de cara a observar el desempeño de cada uno de ellos.

Tras entrenar los modelos, se han calculado tanto el rRMSE como el porcentaje de casos favorables de cada uno, que son los dos aspectos elegidos para evaluar la calidad de los modelos. Una vez evaluado todo esto, se ha decido elegir el modelo Gradient Boosting con 500 estimadores para predecir las ventas sobre los nuevos datos ofrecidos.

El motivo del haber escogido el modelo Gradient Boosting con 500 estimadores han sido las métricas rRMSE y porcentaje de casos favorables.