Olist. Anàlisi exploratòria i models predictius de ML

*Anàlisi Exploratòria i Modelatge Predictiu de Vendes de **Brazilian E-commerce Dataset.**:*

Un Estudi sobre Estacionalitat, Preu i Satisfacció del Client, i models predictius de Machine Learning.

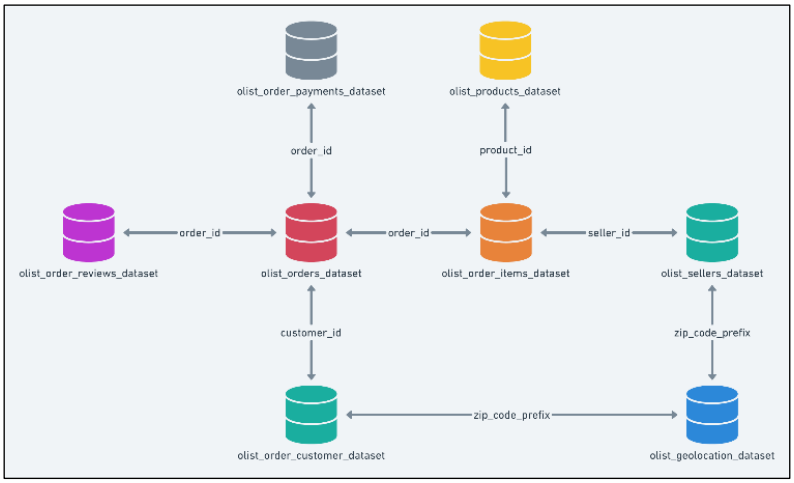

***Figura 1:** Estructura del DataSet: Brazilian E-commerce Dataset. Kaggle Olist & André Sionek. (2018).*

Resum

Aquest estudi presenta una anàlisi exhaustiva del comportament de compra dins d'un e-commerce brasiler mitjançant dades reals de més de 100.000 comandes. El projecte inclou la neteja, transformació i enriquiment del conjunt de dades per a l'estudi de variables temporals, comercials i qualitatives. Es duu a terme una anàlisi exploratòria per identificar patrons de vendes mensuals, categories més venudes, estacionalitat i la relació entre preu, nombre de fotos i satisfacció del client. A més, es desenvolupen models predictius (Prophet, XGBoost i SARIMA) per pronosticar la demanda setmanal de les principals categories, amb una comparació de mètriques com MAE, MAPE i R^2. Aquest treball proporciona un marc integral per a la presa de decisions basada en dades dins del comerç electrònic.

Introducció

El comerç electrònic ha esdevingut un dels motors econòmics més dinàmics dels últims anys, amb patrons de consum cada cop més digitals i influenciats per factors temporals, comercials i socials. En aquest context, entendre el comportament dels consumidors i predir la demanda en moments clau com el Black Friday o les Festes Junines resulta fonamental per a la presa de decisions estratègiques.

Aquest estudi se centra en l'anàlisi del dataset públic d'e-commerce brasiler Olist, amb l'objectiu de construir una pipeline completa de preprocessament, exploració i modelatge predictiu orientada a detectar patrons de compra, estacionalitats, factors que influeixen en la satisfacció del client i possibilitats de predicció de vendes per categoria. El projecte busca, així, oferir una visió de la aportació de la anàlisi qualitativa i de la aportació des models predictius de Machine Learning i el seu vincle amb la presa de decisions basada en dades dins del sector del comerç electrònic.

Metodologia

L'estudi s'ha dut a terme a partir del *Brazilian E-Commerce Public Dataset*, compost per més de 110.000 registres de comandes entre 2016 i 2018, juntament amb dades sobre productes, clients, venedors, pagaments i ressenyes. El procés metodològic ha inclòs:

- **Preprocessament avançat**: tractament de valors nuls, unificació i transformació de timestamps, traducció i categorització dels productes al català, i optimització de la memòria. S'han incorporat variables derivades com dies de lliurament, festius brasilers, períodes comercials i esdeveniments clau com el Black Friday o el Dia das Mães.
- **Anàlisi exploratòria**: visualitzacions temporals de la facturació, mapes de calor horaris, estacionalitats per categoria i relacions entre preu, fotos de producte i satisfacció del client.
- **Modelatge predictiu**: entrenament de tres models per predir vendes setmanals per categoria —*Prophet*, *XGBoost* i *SARIMA*. S'ha aplicat validació temporal i comparació de mètriques (MAE, MAPE, R^2) per identificar els millors models en les 15 categories principals que representen més del 75% de la facturació total del període analitzat.

Totes les anàlisis s'han desenvolupat en Python, utilitzant biblioteques com *pandas*, *matplotlib*, *seaborn*, *prophet*, *xgboost, sarima* i *statsmodels entre d'altres*. S'ha utilitzat també Power BI per a alguna de les visualitzacions.

Resultats

L'anàlisi s'ha basat en 113.425 registres de comandes, amb una mitjana d'1,14 ítems per comanda. El període analitzat comprèn de setembre de 2016 a octubre de 2018. Els principals resultats obtinguts són:

- **Facturació mensual**: Tendència general de creixement amb un pic de facturació a novembre de 2017 (BRL 1.010K), coincidint amb el Black Friday.
- **Categories amb més impacte**:
  - *Facturació (2017-2018)*: salut_i_bellesa, rellotges_regals i llit_bany_taula.
  - *Unitats venudes*: llit_bany_taula i accessoris_informatics.
- **Estacionalitat**:
  - Categories com *ordinadors* o *aire_condicionat* mostren màxims clars en mesos específics.
  - Els mesos amb més demanda global han estat maig, juliol i agost.
- **Satisfacció i qualitat del servei**:
  - Correlació negativa entre dies de lliurament i puntuació mitjana.
  - S'ha detectat un impacte positiu de tenir entre 3 i 6 fotos per producte en la satisfacció i facturació.
- **Modelatge predictiu**:
  - El model XGBoost ha ofert els millors resultats generals en 10 de les 15 categories analitzades.
  - La categoria *accessoris infomàtics* ha obtingut una MAPE del 17,66%.

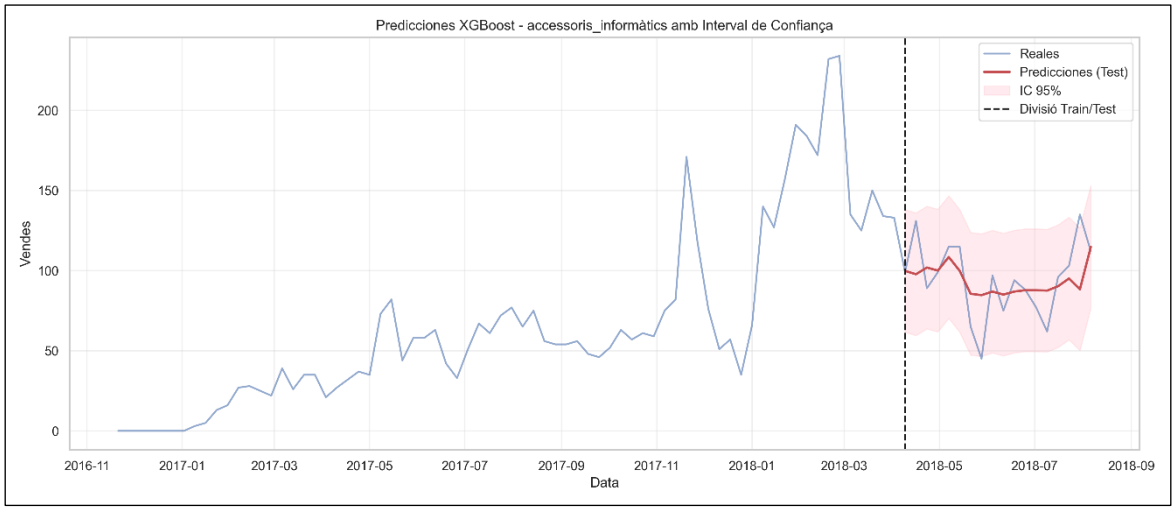

Figura 2: *Predicciones XGBoost – accessoris_informatics*
Creació pròpia . Dades de Olist & André Sionek. (2018).

- **Discussió**

Els resultats mostren una clara incidència de l'estacionalitat en les vendes, amb pics previsibles a festius o les vacances. Això confirma la importància de calendaritzar estratègies comercials. A més, la relació inversa entre temps de lliurament i satisfacció reforça la necessitat d'optimitzar la logística per mantenir l'experiència del client.

La presència d'un nombre intermedi de fotos (3-6) i millorar la qualitat de les imatges sembla associada amb una millor resposta del consumidor, en línia amb estudis previs que suggereixen que massa informació pot generar fatiga visual.

Aquest projecte complementa estudis previs en e-commerce i aporta una visió localitzada dins del context brasiler, tenint en compte esdeveniments culturals i comercials nacionals, sovint ignorats en models generalistes.

Conclusió

Aquest estudi ofereix una visió integrada i pràctica del comportament de consum en l'e-commerce brasiler, mitjançant tècniques de preprocessament, visualització i predicció de dades. Les troballes demostren la importància d'adaptar l'estratègia comercial a l'estacionalitat, millorar els temps de lliurament, i generar incentius per potencia la repetició de compra per part dels clients.

De l'anàlisi exploratòria cal destacar:

- Cal monitorar problemàtiques logístiques contínuament ja que afecten greument a la satisfacció dels clients.
- La major part de les compres es produeixen des dels grans centres de població.

Els models predictius desenvolupats proporcionen una base per a la planificació estratègica de vendes i inventari. Es recomana:

- Focalitzar campanyes en períodes com novembre (Black Friday) i maig (Dia das Mães).
- S'ha de destacar la importància d'analitzar i ajustar les estratègies comercials als diferents cicles de demanda de cada categoria.

Futures línies de recerca podrien aprofundir en l'anàlisi de cohorts de clients.

1. Optimització de campanyes de màrqueting

Els pics de vendes al novembre (Black Friday) o al maig (Dia das Mães) confirmen quan cal reforçar la inversió publicitària i les promocions. Les empreses poden anticipar-se amb ofertes personalitzades i campanyes multicanal durant aquests períodes.

2. Millora logística i de servei

La correlació negativa entre temps de lliurament i satisfacció del client indica que les empreses haurien de prioritzar la reducció de terminis d'enviament o, si no és possible, ajustar les expectatives amb comunicació clara i serveis de seguiment.

3. Estratègia de contingut visual

S'ha demostrat empíricament que tenir entre 3 i 6 fotos per producte millora la satisfacció i, de retruc, la facturació. Això dona una recomanació molt concreta: invertir en fotografia de qualitat en lloc de quantitat excessiva.

4. Polítiques de preus adaptades per categoria

La normalització de preus per categoria permet identificar segments on es poden oferir productes "premium" i d'alt marge, versus d'alt volum però baix preu. Això ajuda a definir estratègies com els paquets promocionals o els productes d'entrada.

5. Planificació d'estoc i aprovisionament

Els models predictius amb errors relativament baixos (MAPE < 25% en moltes categories) es poden utilitzar per estimar la demanda setmanal futura i preparar estocs amb antelació, evitant trencaments i costos logístics innecessaris.

6. Segmentació de clients i personalització

Amb una futura extensió del projecte, aquestes troballes poden alimentar sistemes de recomanació, fidelització i personalització de l'oferta segons el comportament de compra estacional o la sensibilitat al preu.

REFERÈNCIA

Olist & André Sionek. (2018). Brazilian E-Commerce Public Dataset by Olist [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/195341