"Exploración Anual de Ventas, Productos y Consumidores y Predicción de Envíos con *Machine Learning*"

RESUMEN

Este estudio presenta un análisis exploratorio de datos sobre ventas, productos y consumidores utilizando una base de datos obtenida a través de la API de *Kaggle.* Tras la recolección y limpieza inicial de los datos, se identificaron varios *outliers* que fueron considerados en el análisis debido a su origen de ventas reales.
Se realizaron análisis exploratorios para extraer patrones clave en las ventas a lo largo de los 4 años de nuestro *dataset*, segmentando los resultados según geografía, temporalidad, productos y consumidores. Todo ello quedó reflejado en 3 *dashboards* interactivos donde se visualizaron los principales *KPI's (Key Performance Indicators*) y métricas relevantes, permitiendo un análisis dinámico a los usuarios.
En la segunda parte del estudio, se implementaron modelos predictivos utilizando finalmente una regresión logística, con el objetivo de predecir el tipo de envío basándose en las características de las ventas y las categorías de productos. Aunque el modelo presentó una precisión del 37,85 %, lo que limita su efectividad, permitió identificar algunos patrones y relaciones entre estas tres variables. Los resultados obtenidos proporcionan una base para futuras mejoras en cuanto a los modelos predictivos y aportan un conocimiento ventajoso sobre la evolución de las ventas.

INTRODUCCIÓN

La elección de utilizar una base de datos de ventas para este estudio se fundamenta en mi experiencia de más de 20 años en el sector *retail,* trabajando estrechamente con *KPI's* relacionados con ventas. He sido testigo directo del impacto positivo que tiene el análisis detallado de los datos en la toma de decisiones estratégicas.

En el sector del *retail* y comercio, el análisis de datos es una herramienta clave para optimizar estrategias de negocio, mejorar la experiencia del cliente y maximizar los beneficios, lo que resulta esencial en una industria tan dinámica y con alta demanda de adaptabilidad como esta. ("Los KPI's del retail,"2024).

El objetivo de este proyecto es doble. Primero, realizar un análisis exploratorio exhaustivo para identificar patrones clave en las ventas, los productos y los consumidores y segundo: desarrollar un modelo predictivo basado en regresión logística para determinar el tipo de envío en función de variables como la categoría del producto y las ventas.

Específicamente, el objetivo de este estudio se enfoca en responder las siguientes preguntas:

- ¿Cuáles son las principales tendencias de ventas durante los 4 años de nuestros datos? ¿Existe una constante de crecimiento en el tiempo?
- ¿Qué zonas de la geografía estadounidense aportan más beneficio a la empresa?

- ¿Cómo actúan los consumidores?
- ¿Qué productos y categorías destacan en términos de desempeño?

Y, por último:

- ¿Cómo se pueden predecir características de envío basadas en datos históricos de ventas y productos?

Este trabajo destaca la importancia de analizar datos históricos y aplicar técnicas de *Machine Learning* para anticipar necesidades logísticas y mejorar estrategias empresariales.

METODOLOGÍA

Recolección de Datos:

Empecé la investigación explorando diversas fuentes estadísticas, como el Instituto Nacional de Estadística, el *Portal de Dades de Barcelona* e IDESCAT (*Institut d'Estadística de Catalunya*). Finalmente, seleccioné una base de datos que se encontraba en *Kaggle* la cual contiene información sobre las ventas de una empresa proveedora de material de oficina. Decidí obtener la base de datos mediante la API de *Kaggle,* frente a la descarga directa del archivo CSV, porque además de permitirme aumentar conocimientos y practicar en este tema, nos aseguramos que la obtención de datos fuera más eficaz, asegurándonos también las posibles actualizaciones automáticas que pudieran haber con esta base de datos.

Análisis Exploratorio y chequeo de Datos:

El análisis exploratorio lo realicé íntegramente en *Python*, utilizando bibliotecas como *Pandas* para la manipulación de datos y *Seaborn* y *Matplotlib* para la visualización. Los pasos principales fueron:

1. Verificar la estructura de los datos: traté los valores nulos y ajusté los tipos de datos que fueron necesarios.

2. Creación de nuevas variables: añadí las columnas que extraen el año y el mes de las fechas para facilitar el análisis temporal.

3. Decisión de incluir todas las columnas: realicé una valoración de todas las variables, y decidí no descartar ninguna porque podía obtener información sobre dimensiones geográficas, temporales, de productos, consumidores y de logística.

4. Tratamiento de outliers: aunque se detectaron valores atípicos, decidí incluirlos en el análisis exploratorio ya que se trata de ventas reales y aportan un valor crucial al análisis. No fue lo mismo para la parte del modelo de predicción como explicaremos más adelante.

Modelos Predictivos:

Tras el análisis exploratorio, pasé a aplicar los algoritmos de *Machine Learning* de aprendizaje supervisado y trabajé con dos modelos de regresión: lineal y logística. Utilicé *Scikit Learn* para ello.
Se trataron primeramente los *outliers* mediante un *winsorizado* con *Scipy*. Elegimos esta opción porque de esta manera no eliminamos datos, pero sí que reducimos el ruido y nos aseguramos una mejor precisión del modelo("Winsorize: Definition, Examples in Easy Steps."2024).

- Regresión Lineal: Realizamos las primeras pruebas iniciales con una regresión lineal. Se hicieron pruebas con diferentes variables y luego comprobamos sus resultados generando las predicciones y métricas de precisión para evaluar el modelo. Sin embargo, los resultados no fueron satisfactorios debido a la naturaleza de las variables y su falta de ajuste al modelo.

- Regresión Logística: Se optó por este modelo, que es más adecuado para la predicción de variables categóricas. Se preprocesaron nuevamente los datos y se configuró un modelo para predecir el tipo de envío según las ventas y la categoría de los productos. Aunque la precisión del modelo fue del 37%, el proceso permitió aprender sobre la relación entre las variables y ajustar futuros enfoques predictivos.

Visualización de Resultados:

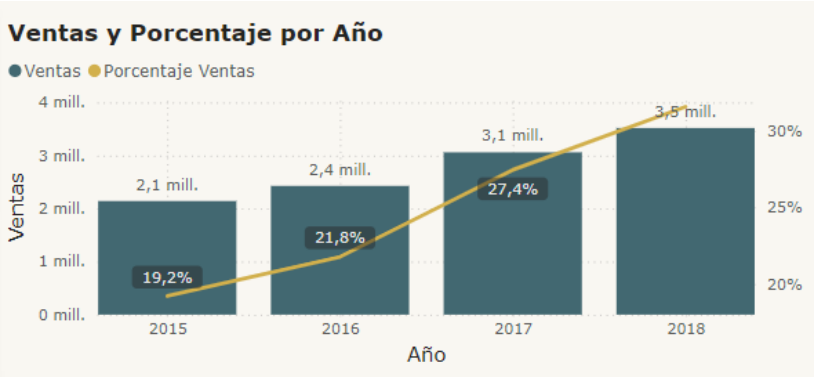
Finalmente, los resultados del análisis exploratorio y las predicciones se integraron en cuatro *dashboard* interactivos para los que utilicé la herramienta de *Power BI*.

Lo dividí en 4 áreas principales:

- Ventas: análisis temporal (años y meses) y geográfico.
- Productos: identificación de los productos más relevantes para la empresa.
- Clientes: estudio de los clientes más rentables.
- Predicciones de envío: visualización de los patrones identificados en el modelo de regresión logística.

RESULTADOS

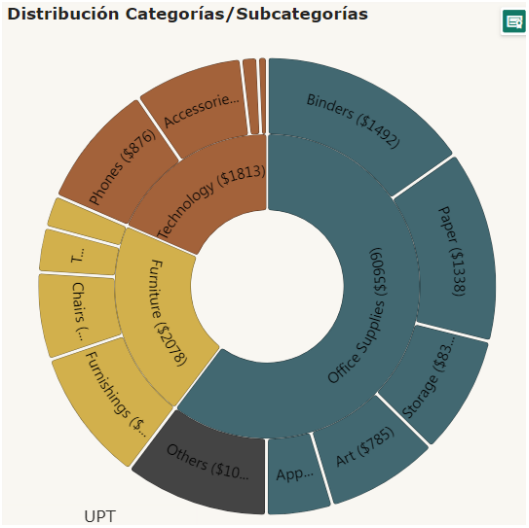
Análisis Exploratorio:



Los gráficos nos muestran una clara constante de crecimiento tanto de manera anual como mensual por año en cuanto a las ventas.

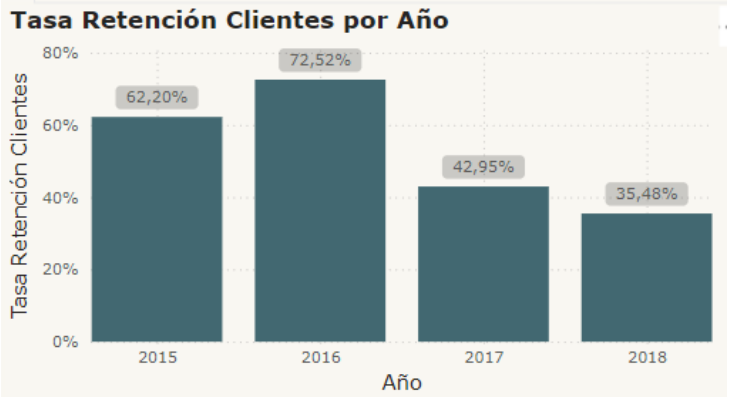
Tenemos un tique medio de unos 2.300$ y un total de entre 950 y 1600 ventas por año. Entre las ciudades que generan más beneficio están California, Texas y Nueva York.

En cuanto a producto podemos ver en la tabla de la derecha las categorías y subcategorías que nos generan un mayor retorno. En el número uno se encuentra *office supplies* y como subcategoría los *binders*.



Entre otros KPI's hemos, podemos ver el Top 10 de productos más vendidos y menos vendidos.

En cuanto a los clientes, hemos destacado la información relacionada con la tasa de retención, la cual parece estar disminuyendo con el tiempo.
Sin embargo, a pesar de esta caída en la retención, las ventas totales siguen aumentando, al igual que el valor medio por cliente y el promedio de artículos adquiridos por cada uno. Esto sugiere que la estrategia de negocio está siendo efectiva, particularmente en lo que respecta a la captación de nuevos clientes.



Modelo predictivo:

```
Precisión del modelo: 0.37854363535297386
Informe de clasificación:
                precision    recall  f1-score   support

   First Class       0.00      0.00      0.00       287
      Same Day       0.05      0.23      0.08       110
  Second Class       0.20      0.16      0.18       345
Standard Class       0.59      0.57      0.58      1057
```

El resultado del *Score* en nuestro modelo de regresión logística fue de 0.37, lo cual nos proporciona una base sólida para evaluar el desempeño del modelo. Aunque el informe de clasificación revela que el modelo predice tres de las cuatro categorías existentes, esto nos ofrece valiosas oportunidades de mejora. En la matriz de confusión, por su parte, queda resaltado como el modelo tiene una alta precisión en la predicción del envío más utilizado por los consumidores, lo que refleja su capacidad para identificar patrones clave de comportamiento.

Matriz de Confusión

Actual Ship Mode	Same Day	Second Class	Standard Class
First Class	74	56	157
Same Day	25	13	72
Second Class	94	55	196
Standard Class	301	155	601

DISCUSIÓN

En cuanto a los resultados obtenidos destacan la eficiencia de las estrategias comerciales implementadas por la empresa, ya que las ventas, tanto mensuales como anuales, muestran una tendencia constante al alza. En cuanto a la estacionalidad, no se observa un patrón claro, lo que sugiere que el comportamiento de las ventas no está ligado a este factor, sino que más bien se caracteriza por picos en áreas geográficas específicas. Además, el análisis de los tiques medios y el consumo promedio por cliente sugiere que el modelo de negocio se enfoca en maximizar la cantidad de productos adquiridos por cliente. Aunque no disponemos de información sobre los precios unitarios de los productos, estas métricas elevadas nos permiten deducir que los beneficios provienen de la compra recurrente de productos por parte de los clientes, exceptuando grandes compras puntuales como demuestran los *outliers*. Finalmente, resalta la concentración geográfica de los clientes, que coincide con las zonas de mayor rentabilidad. Y la capacidad estratégica en cuanto a obtener nuevos clientes cada año.

Por otro lado, aunque este estudio ha proporcionado valiosas perspectivas sobre las ventas, productos y consumidores, así como una aproximación a la predicción de envíos mediante regresión logística, existen algunas limitaciones que deben ser consideradas: Durante el proceso de modelado, no se exploraron otros enfoques predictivos que podrían haber mejorado la precisión de las predicciones. Nuestro modelo permitió identificar algunas relaciones clave entre las variables, pero su precisión limitada sugiere que hay margen para seguir ajustando y mejorándolo, tanto en términos de modelado predictivo como en el análisis de nuevas fuentes de datos.

CONCLUSIONES

El análisis exploratorio permitió identificar patrones significativos en las ventas y tendencias de consumo. Además, la implementación de un modelo predictivo demostró ser útil para predecir el tipo de envío con base en las características del producto y las ventas. Estos resultados destacan la importancia de integrar el análisis de datos con técnicas de *Machine Learning* en estrategias comerciales.

REFERENCIAS

- Los KPI's del retail esenciales para garantizar un crecimiento sostenido. (2024). https://www.wappingweb.com/los-kpis-del-retail-esenciales-para-garantizar-un-crecimiento-sostenido/
- Winsorize: Definition, Examples in Easy Steps. (2024). https://www.statisticshowto.com/winsorize/