Informe de Análisis y Modelado: Proyecto Superstore

Resumen Ejecutivo

El presente documento detalla los hallazgos y resultados del análisis del dataset "SampleSuperstore", el cual registra las operaciones de ventas y logística de una cadena minorista entre los años 2016 y 2017. El objetivo central del proyecto fue identificar las causas raíz de las pérdidas financieras y desarrollar modelos predictivos para optimizar la rentabilidad y la eficiencia logística.

Los hallazgos más críticos revelan una erosión significativa del margen de beneficio vinculada a políticas de descuentos agresivos, particularmente en la Región Central y en las categorías de Muebles y Suministros de Oficina. Asimismo, se identificó un patrón sistémico de retrasos en los envíos, altamente correlacionado con el modo de envío *Standard Class* y la proximidad de días feriados. Mediante el uso de modelos de *Machine Learning* (Random Forest), se logró una capacidad de explicación de la variabilidad financiera del 66% y una detección proactiva del 44% de los retrasos reales, permitiendo proponer estrategias basadas en datos para incrementar el beneficio neto y mejorar la comunicación con el cliente.

---

1. Alcance y Metodología del Proyecto

El análisis se basó en un dataset de 9.994 registros y 20 columnas, estructurado en diversas dimensiones críticas para la operación:

- **Geográfica:** Desglose por regiones (Central, East, South, West), estados y ciudades.
- **Segmentación de Productos:** Categorías de *Furniture* (Muebles), *Office Supplies* (Suministros de Oficina) y *Technology* (Tecnología).
- **Métricas Financieras:** Datos de ventas, cantidades, descuentos y ganancias (*Profit*).
- **Logística:** Fechas de orden y envío, y modalidades de transporte.
- **Segmentación de Clientes:** Clasificación en *Consumer*, *Corporate* y *Home Office*.

Proceso de Preparación de Datos

Para garantizar la integridad de los modelos de *Machine Learning*, se realizaron las siguientes intervenciones:

- **Limpieza y Auditoría:** Conversión de formatos de fecha y verificación de ausencia de valores nulos.

- **Prevención de Sobreajuste (Overfitting):** Eliminación de identificadores únicos (como *Customer Name* u *Order ID*) para que los modelos identifiquen patrones generales y no memoricen transacciones específicas.
- **Enriquecimiento mediante API:** Integración de datos de días feriados para analizar su impacto en los tiempos de entrega.
- **Ingeniería de Características:** Creación de las variables demora_envio (diferencia entre fecha de orden y envío) y es_feriado, además de la codificación de variables categóricas.

---

2. Diagnóstico de Rentabilidad (Profit)

El análisis identificó que la rentabilidad no es uniforme y presenta fugas críticas de valor en áreas específicas.

El Impacto del Descuento

Existe una correlación negativa sustancial (-0.48) entre el nivel de descuento otorgado y la ganancia obtenida. El uso de descuentos agresivos actúa como el motor principal de la erosión del margen.

Categoría	Umbral de Descuento Crítico	Impacto Observado
Muebles (Furniture)	> 30%	Generación de márgenes negativos extremos.
Suministros de Oficina	> 80%	Pérdidas individuales de hasta -$3,500.

Análisis Geográfico

La **Región Central** se destaca negativamente por presentar pérdidas significativas, con un rendimiento particularmente deficiente en la categoría de muebles. Por el contrario, la región **East** muestra una alta resiliencia y retornos promedio superiores en la categoría de Tecnología.

---

3. Eficiencia Logística y Gestión de Retrasos

El estudio de los tiempos de entrega revela que los retrasos no son eventos aleatorios, sino sistémicos y predecibles.

- **Correlación por Modo de Envío:** El modo *Standard Class* presenta una correlación de **0.73** con los retrasos en las entregas. En contraste, la modalidad

Same Day reduce significativamente el riesgo de incumplimiento (correlación de -0.53).

- **Efecto de los Feriados:** La cercanía de días feriados ensancha el margen de error en las entregas, sugiriendo la necesidad de ajustar las promesas de entrega en estos periodos.
- **Capacidad de Predicción:** El modelo de clasificación desarrollado permite prever 255 retrasos con una precisión del 52% y un *recall* del 44%, facilitando una gestión proactiva.

---

4. Modelado Predictivo y Simulación

Se implementaron dos enfoques de modelado utilizando el algoritmo *Random Forest*:

1. **Regresión (Predicción de Profit):**
    - **Resultado:** R^2 de 0.66.
    - **Conclusión:** El modelo valida que las ventas y los descuentos son los predictores con mayor peso en el resultado financiero final.
2. **Clasificación (Detección de Retrasos):**
    - **Resultado:** Identificación del 44% de los retrasos reales.
    - **Aplicación:** Permite activar protocolos de comunicación antes de que el cliente experimente el retraso.

Simulación de Escenarios

Una simulación predictiva demostró que si la cadena minorista limitara los descuentos máximos al **20%**, el beneficio neto total experimentaría un incremento drástico, eliminando las pérdidas extremas detectadas en la actualidad.

---

5. Recomendaciones Estratégicas

Basándose en la evidencia analítica, se proponen las siguientes acciones:

- **Restricción de Descuentos:** Implementar una política de límite de descuentos del 20% en la Región Central y eliminar las ofertas del 80% en suministros de oficina.
- **Focalización Comercial:** Priorizar y potenciar las ventas de la categoría **Tecnología** en la región **East**, debido a su alto retorno.

- **Optimización Logística:** Utilizar el modelo de clasificación para monitorear pedidos en *Standard Class*. Se recomienda activar alertas automáticas de demora cuando la probabilidad de retraso calculada por el modelo supere el 50%.
- **Gestión de Expectativas:** Ajustar automáticamente las fechas prometidas de entrega cuando el sistema detecte la proximidad de un feriado, mitigando el impacto negativo en la satisfacción del cliente.