**Grupo 3. Avances del proyecto(Entrega2)**

**YOSELIN NIETO GIL, FELIPE DIEGO FELIPE DIEGO LOBATO DA SILVA, NICOLAS ALEJANDRO YEPES JOVEN y ANA MARIA SOTO OROZCO.**

1. RESUMEN SOBRE EL PROBLEMA

1.1 Contexto y Definición del Problema:

El proyecto aborda un desafío común en la industria de productos de consumo masivo: la necesidad de anticipar los precios de venta de los productos durante las distintas temporadas del año (otoño, invierno, primavera y verano). Estas variaciones estacionales afectan los hábitos de compra y el precio de los productos, lo que puede complicar la planificación estratégica de las empresas. Si no se prevén adecuadamente, estos cambios pueden provocar problemas de inventario, como exceso o falta de productos, pérdidas de ventas y costos adicionales. El objetivo de este proyecto es desarrollar un modelo que, utilizando datos históricos de ventas, estime un precio de venta óptimo por temporada, ayudando a las empresas a mejorar su planificación estratégica y eficiencia operativa.

1.2 Pregunta de Negocio y Alcance del Proyecto:

La pregunta central del proyecto es: ¿Cómo predecir el precio de venta estacional de productos de consumo masivo para optimizar la gestión de inventarios y mejorar la eficiencia operativa?

Este proyecto se limita a productos de consumo masivo vendidos a través de plataformas de e-commerce, en particular Olist Store, durante el año 2018. El modelo debe prever cuáles productos tendrán un mayor precio de venta en cada temporada, analizando patrones estacionales y de comportamiento del consumidor. La solución está dirigida a empresas que buscan ajustar sus inventarios y maximizar la disponibilidad de productos, considerando el precio estacional estimado y la demanda anticipada, para una mayor efectividad en sus estrategias comerciales.

* 1. Datos a Emplear y Preparación Inicial:

El proyecto empleará datos históricos de Olist Store, que incluyen información detallada sobre transacciones, productos, clientes, vendedores y reseñas. Las fuentes de datos principales se obtuvieron de Kaggle, complementadas con datos de códigos postales (CEP) adquiridos de un proveedor para enriquecer los análisis geoespaciales. La preparación de datos incluyó varios pasos:

1. **Estandarización y Transformación**: Se unificaron los formatos de los CEP y se convirtieron los archivos a Parquet para mejorar el rendimiento.
2. **Geolocalización**: Se integraron y ajustaron datos de geolocalización mediante un cruce de archivos y el uso de la API de Nominatim para obtener coordenadas precisas.
3. **Consolidación:** Los datos enriquecidos se consolidaron en un archivo optimizado para análisis espaciales, con las coordenadas ajustadas en formato de alta precisión.
4. **Transformación final:** Con los datos consolidados, se homologan los nombres de los campos al español para facilitar su interpretación. A continuación, utilizando el campo de fecha de compra del pedido, se filtra la base de datos para incluir únicamente el año más reciente disponible, en este caso, 2018. Posteriormente, se crea el campo temporada en función del mes de la fecha de compra, categorizando cada registro en una de las cuatro estaciones del año: otoño, invierno, verano o primavera. Esta transformación permitirá analizar y prever variaciones estacionales en los precios de venta de forma más precisa. Se obtiene una base con 63.215 registros y 33 columnas.

A continuación, se detalla la información de la base final:

Tabla " Tabla de metadatos de base final"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre de la Columna | Descripción | Tipo de Dato | Ejemplo del Dato | Clave |
| id\_pedido | Identificador único del pedido | string | "995392413cee616c1..." | Clave primaria |
| id\_cliente | Identificador único del cliente | string | "4bf2490c4245cdb25a..." | Clave foránea (customers.customer\_id) |
| estado\_pedido | Estado del pedido (ej. entregado, cancelado) | string | "entregado" |  |
| fecha\_pedido | Fecha en que se realizó la compra | timestamp | "2017-09-04" |  |
| hora\_compra\_pedido | hora en que se realizó la compra | string | "22:43:54" |  |
| fecha\_aprobacion\_pedido | Fecha y hora en que se aprobó el pedido | timestamp | "2017-09-04 22:43:54" |  |
| fecha\_entrega\_estimada | Fecha estimada de entrega del pedido | timestamp | "2017-09-27 00:00:00" |  |
| id\_item\_pedido | Identificador del ítem dentro del pedido | integer | 1 | Clave primaria compuesta junto a id\_pedido |
| id\_producto | Identificador único del producto | string | "4lebbrb7a41c44632..." | Clave foránea (products.product\_id) |
| id\_vendedor | Identificador único del vendedor | string | "7a67c85e85bbc2e85..." | Clave foránea (sellers.seller\_id) |
| fecha\_limite\_envio | Fecha límite para el envío | timestamp | "2017-05-22 16:05:44" |  |
| precio | Precio del producto | doble | 109.99 |  |
| valor\_flete | Valor del envío | doble | 18.02 |  |
| secuencia\_pago | Secuencia del pago dentro del pedido | integer | 1 | Clave primaria compuesta junto a id\_pedido |
| tipo\_pago | Tipo de pago utilizado | string | "tarjeta\_credito" |  |
| cuotas\_pago | Número de cuotas del pago | integer | 8 |  |
| valor\_pago | Valor del pago | doble | 99.33 |  |
| id\_unico\_cliente | Identificador único y persistente del cliente | string | "15ee900ec703c9a10" |  |
| codigo\_postal\_cliente | Prefijo del código postal del cliente | string | "68590" | Clave foránea (geo.cep\_prefix) |
| ciudad\_cliente | Ciudad del cliente | string | "jacunda" |  |
| estado\_cliente | Estado del cliente | string | "PA" |  |
| nombre\_categoria\_producto | Categoría del producto | string | "perfumaria" |  |
| longitud\_nombre\_producto | Longitud del nombre del producto | integer | 40 |  |
| longitud\_descripcion\_producto | Longitud de la descripción del producto | integer | 287 |  |
| cantidad\_fotos\_producto | Cantidad de fotos del producto | integer | 1 |  |
| peso\_producto\_g | Peso del producto en gramos | integer | 225 |  |
| largo\_producto\_cm | Longitud del producto en centímetros | integer | 16 |  |
| altura\_producto\_cm | Altura del producto en centímetros | integer | 10 |  |
| ancho\_producto\_cm | Ancho del producto en centímetros | integer | 14 |  |
| codigo\_postal\_vendedor | Prefijo del código postal del vendedor | string | "13023" | Clave foránea (geo.cep\_prefix) |
| ciudad\_vendedor | Ciudad del vendedor | string | "campinas" |  |
| estado\_vendedor | Estado del vendedor | string | "SP" |  |
| temporada | Estación del año de acuerdo al mes de compra del pedido (Otoño, Invierno, Verano, Primavera) | string | "Verano" |  |

1. MODELOS DESARROLLADOS Y SU EVALUACIÓN

Con la intensión de determinar el mejor modelo para resolver las necesidades de la empresa Olist Store, de definió desarrollar modelo que predice el precio de productos de acuerdo con la temporada en la que se compran, la ubicación del cliente, la categoría del producto y sus características físicas. Esta información resulta valiosa para la empresa como una herramienta que complementa la planeación contable, financiera y logística de la empresa, así como para la definición de estrategias de marketing ajustadas a los precios de venta de sus productos a lo largo del año.

**Proceso de Entrenamiento**:

* Cada modelo fue implementado dentro de un pipeline que integra preprocesamiento y entrenamiento, facilitando la escalabilidad y reproducibilidad del flujo de trabajo.
* Los datos se dividieron en conjuntos de entrenamiento y prueba utilizando una proporción de 80-20, con una semilla (random\_state=42) para garantizar la consistencia en cada iteración.
* Se aplicó una búsqueda de hiperparámetros para cada modelo usando GridSearchCV o RandomizedSearchCV, asegurando así la selección óptima de parámetros para minimizar errores y mejorar la precisión.

**Selección de Características:**

* Las características utilizadas incluyen tanto atributos numéricos (como el peso, largo, altura y ancho del producto) como categóricos (temporada, categoría del producto, ciudad y estado del cliente, y el ID del producto).
* El preprocesamiento incluyó la estandarización de variables numéricas mediante StandardScaler y la codificación de variables categóricas con OneHotEncoder, facilitando una entrada balanceada y normalizada para los modelos.
* La selección de estas características se basó en su relevancia teórica para la predicción del precio del producto según factores como la estacionalidad y las propiedades físicas del producto.

**Evaluación de Modelos:**

* Los modelos fueron evaluados utilizando métricas de error y precisión, seleccionando la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R²) para capturar diferentes aspectos del rendimiento.
* Se aplicó validación cruzada para obtener un RMSE promedio que refleje la robustez del modelo en diferentes particiones del conjunto de datos, disminuyendo así la posibilidad de overfitting o underfitting.
* Los modelos probados incluyen XGBRegressor, RandomForestRegressor, LinearRegression, ElasticNet y KNeighborsRegressor. En términos de rendimiento, el modelo de regresión lineal obtuvo el mejor RMSE (118.01) y el mayor R² (0.6336), lo cual indica que es el modelo más adecuado para este caso​(Documentación de entren…)​(modelos).

1. **Modelo con XGBRegressor:**
   * **Descripción**: Se implementa un pipeline que incluye preprocesamiento y el modelo XGBRegressor. Se realiza una búsqueda de hiperparámetros con GridSearchCV.
   * **Parámetros óptimos**: {'colsample\_bytree': 0.8, 'gamma': 0.1, 'learning\_rate': 0.1, 'max\_depth': 9, 'n\_estimators': 300, 'reg\_alpha': 0.01, 'subsample': 0.8}
   * **Métricas**:
     + **RMSE**: 132.67
     + **MAE**: 54.75
     + **R²**: 0.5368
     + **RMSE (validación cruzada)**: 127.56
2. **Modelo con RandomForestRegressor:**
   * **Descripción**: Similar al anterior, este pipeline utiliza RandomForestRegressor y RandomizedSearchCV para ajustar hiperparámetros.
   * **Parámetros óptimos**: {'n\_estimators': 100, 'min\_samples\_split': 5, 'min\_samples\_leaf': 1, 'max\_features': 'log2', 'max\_depth': None, 'bootstrap': False}
   * **Métricas**:
     + **RMSE**: 135.53
     + **MAE**: 43.09
     + **R²**: 0.5167
     + **RMSE (validación cruzada)**: 132.04
3. **Modelo con LinearRegression:**
   * **Descripción**: Modelo lineal básico, sin ajuste de hiperparámetros. Incluye preprocesamiento mediante un pipeline.
   * **Métricas**:
     + **RMSE**: 118.01
     + **MAE**: 25.55
     + **R²**: 0.6336
     + **RMSE (validación cruzada)**: 114.60
4. **Modelo con ElasticNet:**
   * **Descripción**: Modelo de regresión penalizada (ElasticNet) que también se entrena en pipeline.
   * **Métricas**:
     + **RMSE**: 183.53
     + **MAE**: 81.10
     + **R²**: 0.1137
     + **RMSE (validación cruzada)**: 171.00
5. **Modelo con KNeighborsRegressor:**
   * **Descripción**: Modelo KNeighborsRegressor con ajuste de hiperparámetros mediante GridSearchCV.
   * **Parámetros óptimos**: {'n\_neighbors': 5, 'p': 1, 'weights': 'distance'}
   * **Métricas**:
     + **RMSE**: 142.54
     + **MAE**: 39.61
     + **R²**: 0.4654
     + **RMSE (validación cruzada)**: 137.98

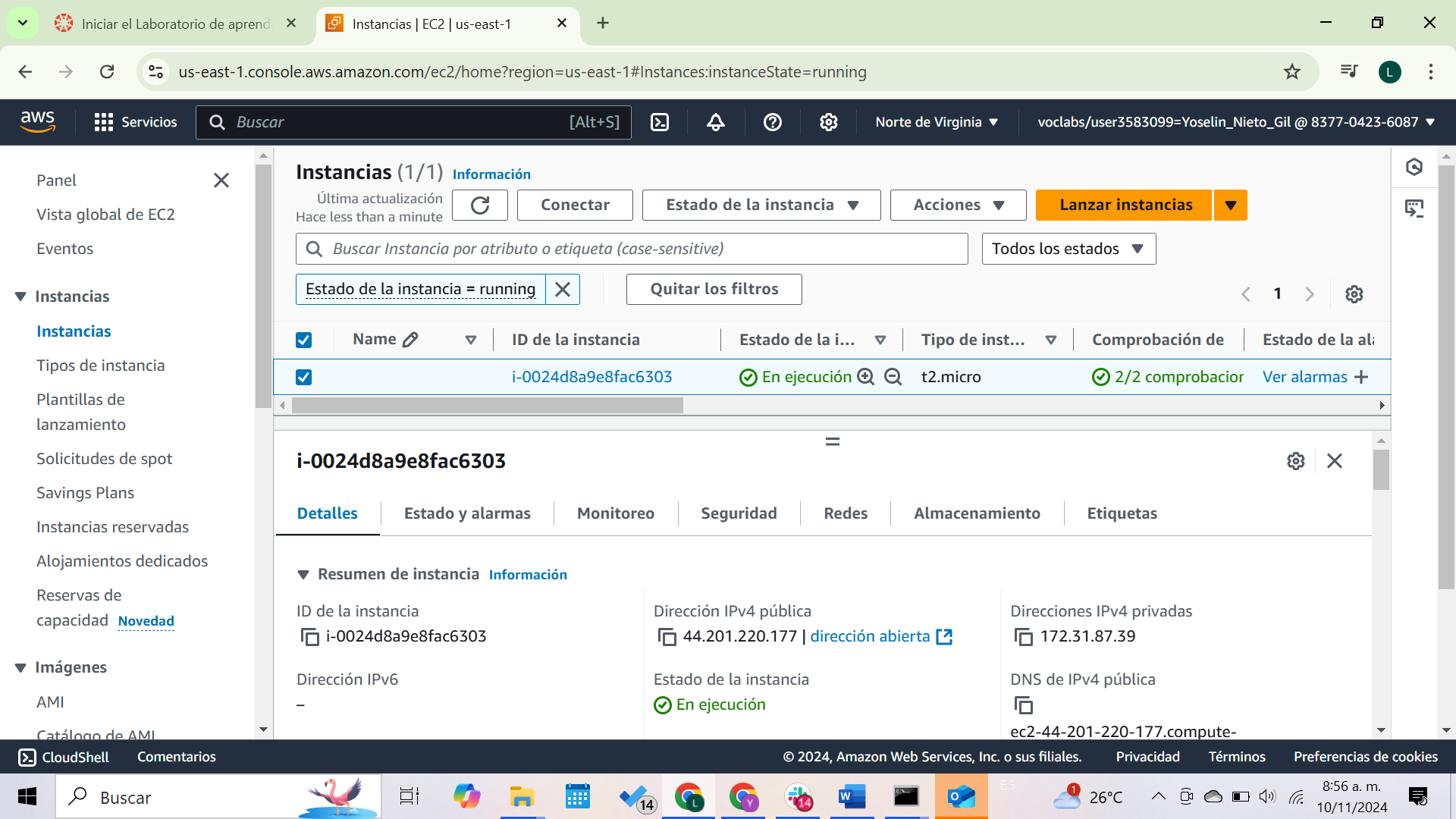
**Modelo con Mejor Desempeño**

Basado en el RMSE y el R², el modelo de **regresión lineal (LinearRegression) tiene el mejor desempeño** con un **RMSE de 118.01** y un **R² de 0.6336**. Este modelo proporciona las predicciones más precisas en este conjunto de pruebas y validación cruzada.

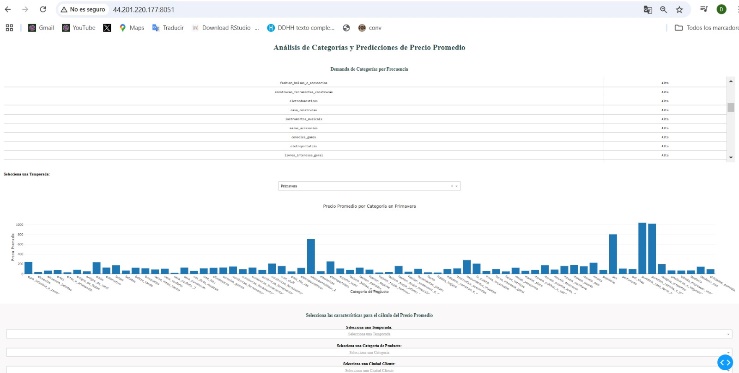
1. OBSERVACIONES Y CONCLUSIONES SOBRE LOS MODELOS
2. TABLERO DESARROLLADO

Para la elaboración del Dash inicial del proyecto, se creó un archivo jupyther notebook para el dash el cual se basa temporalmente en el archivo base\_dash.csv, que contiene las características y predicciones realizadas por el modelo de regresión lineal sobre df\_baseFinal.parquet, seleccionado para las predicciones del precio. El tablero muestra de acuerdo con la temporada seleccionada, la tabla con las categorías de producto y su demanda en esa temporada también muestra un gráfico de barras con el precio promedio por categoría según la temporada, otra parte del Dash sobre las predicciones que se conectará al modelo, busca que el cliente (vendedores) seleccione algunas características usadas como la features del modelo y se estime un precio aproximado.

Se creó una instancia en AWS EC2



Se configuró security group para permitir el acceso al dash. Se creó la carpeta “mi\_proyecto dash” en la máquina virtual y se cargó la base de datos “base\_dash.csv” y se creó el “Dash\_despliegue\_sem5.py” con la configuración del dash, se ejecutó y el dash quedó operativo para uso con el enlace http://44.201.220.177:8051



Parte 2. Prediccion del precio

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente