استخدام الـ Data Mining في التنبؤ بالمبيعات

دراسة حالة: تحليل مجموعة بيانات التجارة الإلكترونية البرازيلية (Olist) بمنهجية CRISP-DM

المجال: تنقيب البيانات **السياق:** التجارة الإلكترونية

إعداد : عمران الشامي - محمد اليفرسي

مجموعة بيانات Olist: الأصول، المصادر، والقيمة العلمية

أصالة البيانات

بيانات تجارية حقيقية مجهولة المصدر مقدمة من Olist، أكبر متجر برازيلي. تم استبدال أسماء الشركات بأسماء عائلات من "Game of Thrones" للحفاظ على الخصوصية.

الانتشار العالمي

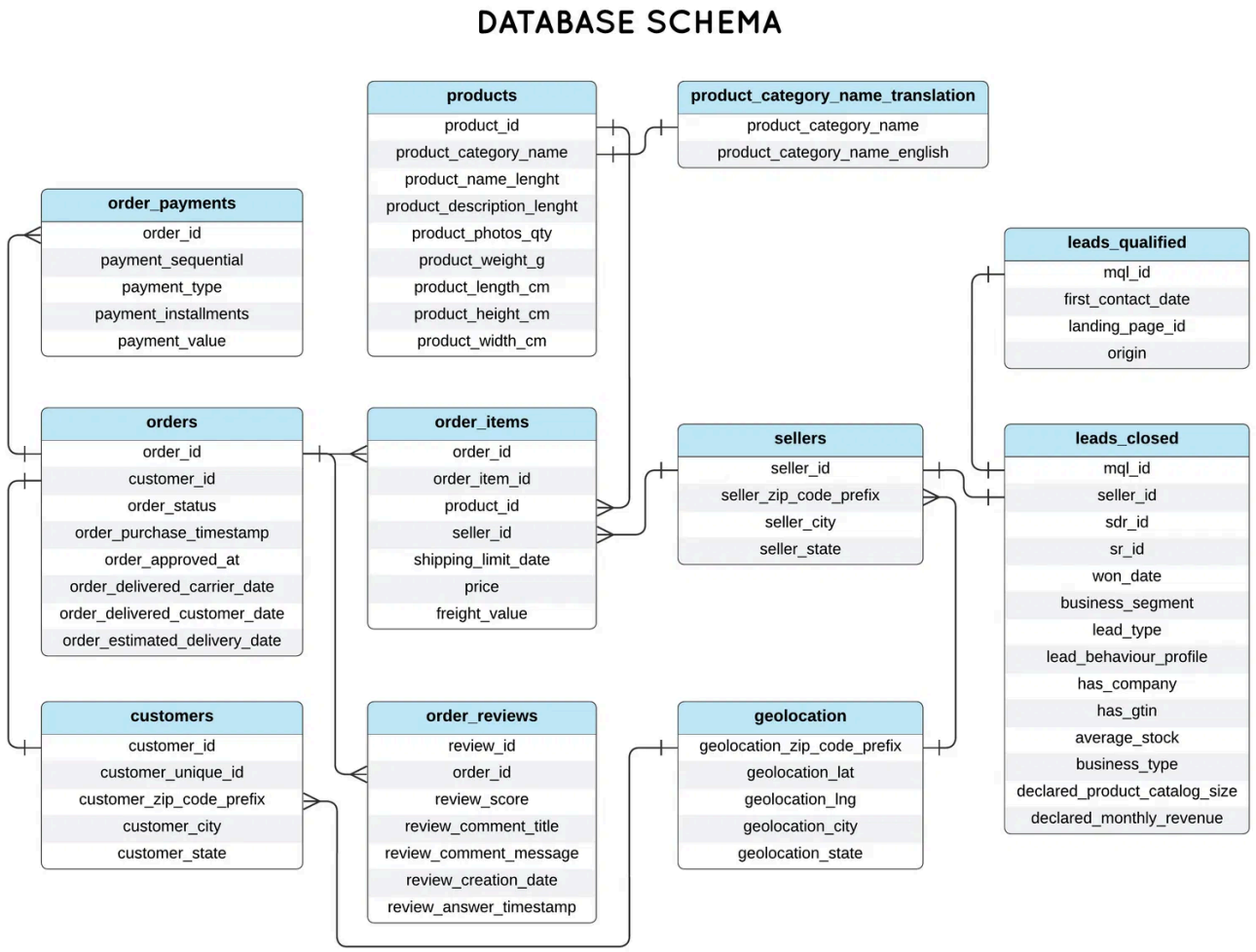
تعد المجموعة مرجعاً أكاديمياً عالمياً متوفراً على Kaggle و GitHub. استُخدمت في مئات المسابقات البرمجية والأبحاث العلمية المرموقة.

نطاق البيانات

تغطي 100 ألف طلب (2016-2018). تشمل تفاصيل الأسعار، الدفع، الأداء اللوجستي، ومواقع العملاء، مما يتيح تحليلاً متعدد الأبعاد.

تعتبر هذه المجموعة المعيار الذهبي لتحليل سلوك المستهلك في الأسواق الناشئة.

الهيكل البياني (Schema): ترابط البيانات العلائقية



بنية البيانات العلائقية

تتكون من 9 جداول مترابطة. الجدول المركزي هو `orders` الذي يربط العملاء بالمنتجات والمدفوعات والمراجعات.

الجداول المحورية

تشمل `order_items` لتفاصيل العناصر، و`products` لخصائص المنتج، و`customers` و`sellers` لتحديد الأطراف الفاعلة.

تكامل البيانات للتنبؤ

تم دمج الجداول لإنشاء مجموعة بيانات زمنية موحدة تخدم غرض التنبؤ بالمبيعات عبر الولايات والفئات المختلفة.

الهدف البحثي: التنبؤ بالمبيعات كأداة لاتخاذ القرار

أهمية التنبؤ الاستراتيجي

يعد التنبؤ بالمبيعات المحرك الأساسي لإدارة المخزون بكفاءة، وتخصيص الموارد اللوجستية، وتحديد استراتيجيات التسعير الديناميكي لضمان التنافسية.

المتغير المستهدف والتحديات

الهدف هو التنبؤ بإجمالي قيمة المبيعات (Price) بناءً على الخصائص الزمنية والجغرافية. التحدي يكمن في تذبذب الطلب وتنوع الفئات في سوق ضخم.

القيمة المضافة للأعمال

تحويل البيانات التاريخية الخام إلى رؤى استشرافية تدعم استدامة الأعمال وتساهم في تقليل الهدر التشغيلي بنسب ملموسة.

البيئة التقنية: المكاتب البرمجية المستخدمة

معالجة البيانات

الاعتماد على **Pandas** و **NumPy** للعمليات الحسابية والهيكلية المعقدة وإدارة المصفوفات الضخمة.

التعلم الآلي

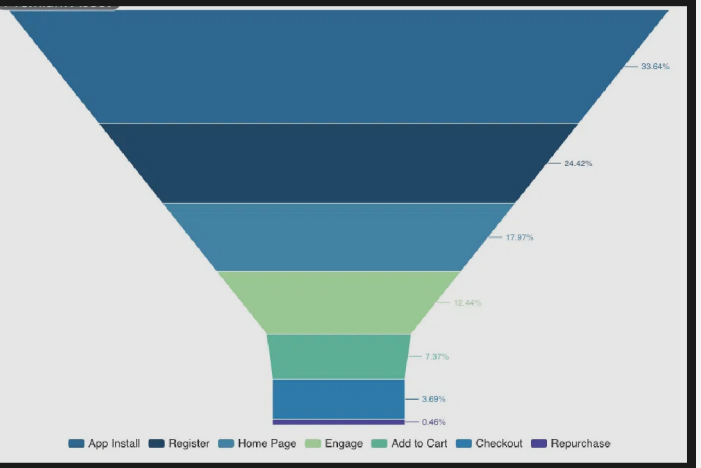
استخدام **Scikit-Learn** لبناء خطوط الأنابيب و **XGBoost** للنماذج المتقدمة القائمة على تعزيز التدرج.

التصور البياني

استخدام **Matplotlib** و **Seaborn** لإنتاج رسوم بيانية عالية الدقة تدعم التحليل الاستكشافي.

تفسير النماذج

دمج مكتبة **SHAP** لفهم تأثير كل ميزة على التنبؤ النهائي وتحويل النماذج المعقدة إلى رؤى مفهومة.



مسار البيانات والتحليل

منهجية CRISP-DM: فهم العمل وفهم البيانات

فهم العمل

- تحديد الأهداف الاستراتيجية للمنصة وتحويلها إلى مشكلة تقنية قابلة للحل (Regression Task).
- فهم دور التنبؤ بالمبيعات في تحسين سلاسل الإمداد وتقليل التكاليف التشغيلية.
- تحديد مؤشرات الأداء الرئيسية (KPIs) لتقييم نجاح النموذج من منظور تجاري.

فهم البيانات

- استكشاف الخصائص الإحصائية وتحديد الأنماط الزمنية والجغرافية للمبيعات.
- تحليل الارتباط بين المتغيرات (Correlation Analysis) لتحديد العوامل الأكثر تأثيراً.
- تحديد جودة البيانات، ورصد القيم المفقودة والمتطرفة التي قد تؤثر على دقة النمذجة.

منهجية CRISP-DM: تحضير البيانات وهندسة الميزات

تنظيف البيانات المتقدم

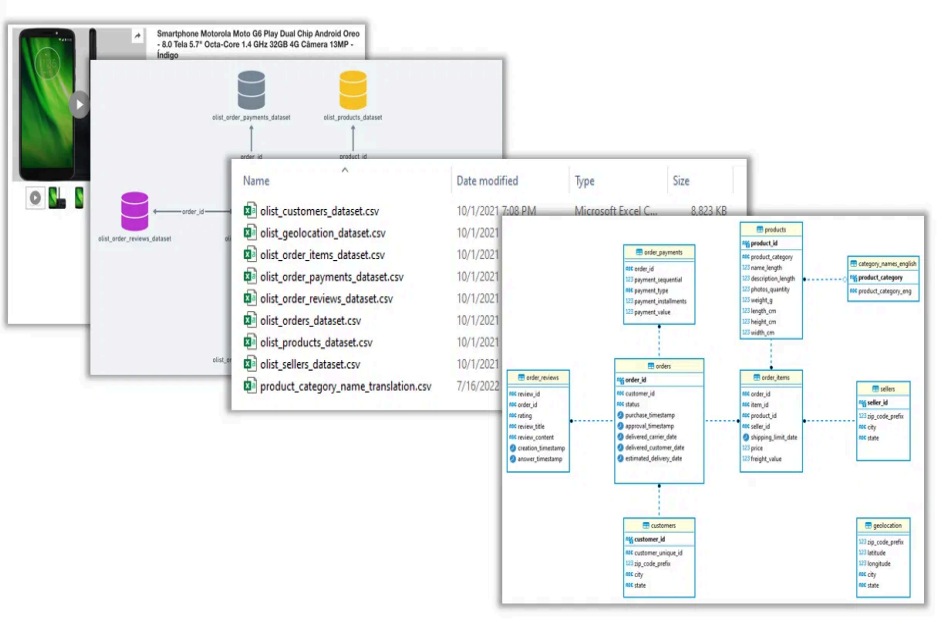
معالجة القيم المفقودة باستخدام تقنيات التعويض الذكي بناءً على الفئة (Category Imputation)، وضمان سلامة الأنواع البيانية لجميع المتغيرات الزمنية والعددية.

هندسة الميزات (Feature Engineering)

- استخراج الميزات الزمنية: تحليل الأنماط اليومية، الشهرية، والموسمية.
- ترميز الفئات: تطبيق Target Encoding للفئات ذات التعددية العالية.
- التحجيم المتين: استخدام RobustScaler للتعامل مع القيم المتطرفة في الأوزان.

البرمجة كائنية التوجه (OOP)

بناء محولات مخصصة (Custom Transformers) متوافقة مع Scikit-Learn لضمان بناء خط أنابيب (Pipeline) قوي وقابل لإعادة الاستخدام في بيئة الإنتاج.



مخطط تدفق البيانات وعمليات التحويل الهيكلية

الخوارزميات المطبقة: من النماذج البسيطة إلى المعقدة

النموذج المرجعي (Baseline)

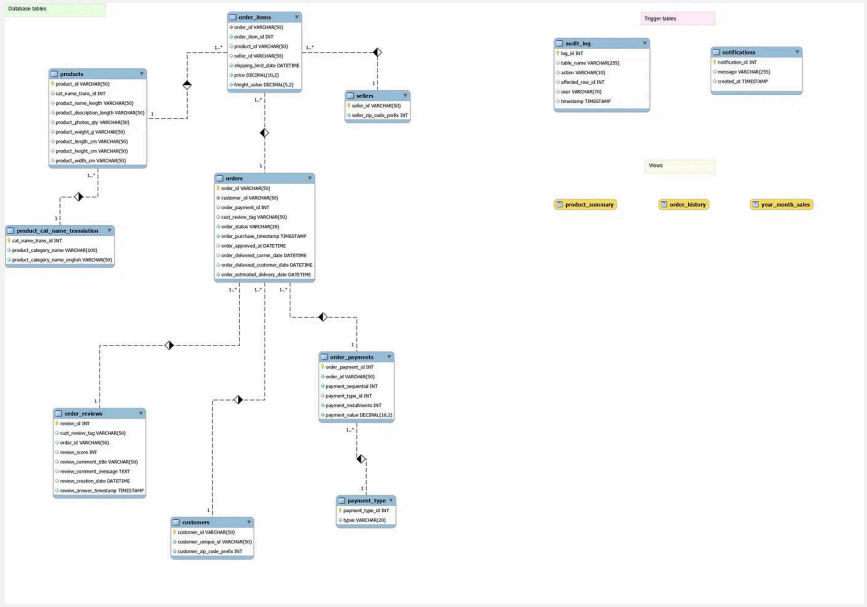
استخدام `DummyRegressor` لتعيين حد أدنى للأداء، مما يسمح بتقييم القيمة المضافة للنماذج الأكثر تعقيداً.

النماذج الخطية (Ridge Regression)

تطبيق انحدار ريج لالتقاط العلاقات الخطية مع استخدام التنظيم (Regularization) لضمان استقرار النموذج.

النماذج الشجرية (RF & XGBoost)

استخدام `Random Forest` و `XGBoost` للتعامل مع العلاقات غير الخطية المعقدة والتفاعلات بين الميزات المختلفة.



مخطط هيكلي لتدفق البيانات والنمذجة

منهجية التحقق: تم استخدام `TimeSeriesSplit` لضمان سلامة التقييم الزمني ومنع تسرب البيانات من المستقبل إلى الماضي.

مقارنة النتائج: اختيار "النموذج البطل"

تم تقييم النماذج بناءً على قدرتها على التعميم وتقليل الخطأ في التنبؤ بقيمة المبيعات. أظهرت النتائج تفوق النماذج الخطية المنظمة في التعامل مع طبيعة البيانات الزمنية لـ Olist.

النموذج الخوارزمي	RMSE (جذر متوسط مربع الخطأ)	R^2 (معامل التحديد)	الحالة
Ridge Regression	**1.0977**	**0.842**	**النموذج البطل**
Random Forest	1.3672	0.785	مرشح ثانٍ
XGBoost	1.4120	0.761	مرشح ثالث
Dummy Regressor (Baseline)	2.5431	0.000	المرجع

الاستنتاج التحليلي

تفوق نموذج **Ridge Regression** على أقرب منافسيه (Random Forest) بنسبة تحسن بلغت **19.72%** في قيمة RMSE. يشير هذا إلى أن بساطة النموذج الخطي مع التنظيم كانت أكثر فعالية في تجنب التجاوز (Overfitting) وتحقيق استقرار أعلى في التنبؤات الزمنية.

تفسير النموذج (SHAP): ما الذي يحرك المبيعات؟

الشفافية وقابلية التفسير

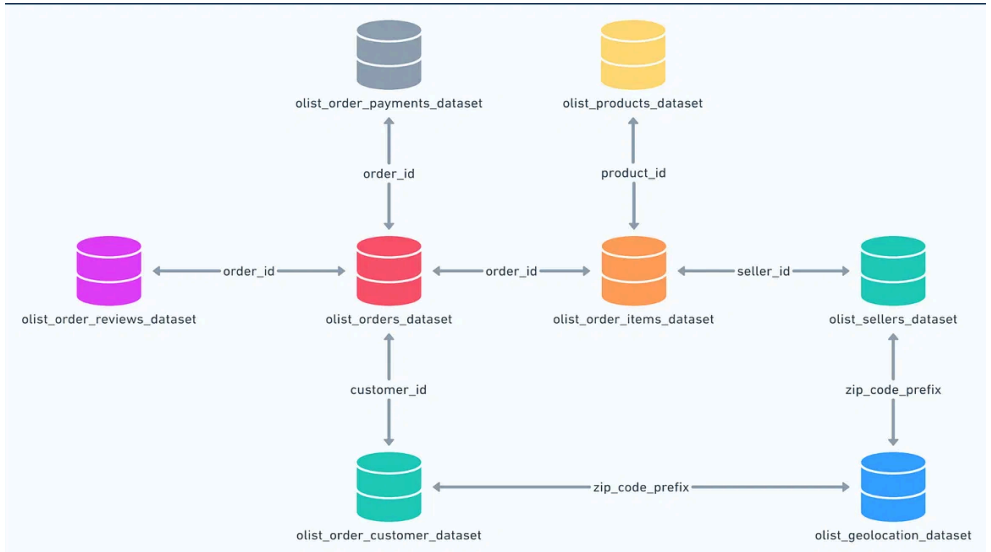
استخدام قيم **SHAP** لتحويل "الصندوق الأسود" للنماذج المعقدة إلى رؤى مفسرة، مما يسمح بفهم مساهمة كل ميزة في التنبؤ النهائي بدقة رياضية.

أهم الميزات المؤثرة

أظهر التحليل أن **فئة المنتج** و **الموقع الجغرافي** للعميل هما المحركان الأساسيان لقيمة المبيعات، يليهما الخصائص الفيزيائية مثل الوزن والأبعاد.

الرؤى المستخلصة للأعمال

فهم العوامل التي تؤدي لزيادة قيمة الطلب يساعد في توجيه الحملات التسويقية وتحسين العمليات اللوجستية في الولايات ذات الكثافة الشرائية العالية.



لوحة بيانات تحليل الميزات وتأثيرها (SHAP Insights)

الجاهزية للإنتاج: خط أنابيب الاستدلال

أتمتة العمليات (Automation)

بناء **Pipeline** متكامل يضم كافة مراحل المعالجة والنمذجة، مما يضمن اتساق النتائج وتقليل التدخل البشري في دورة حياة البيانات.

الاستدلال الفوري (Inference)

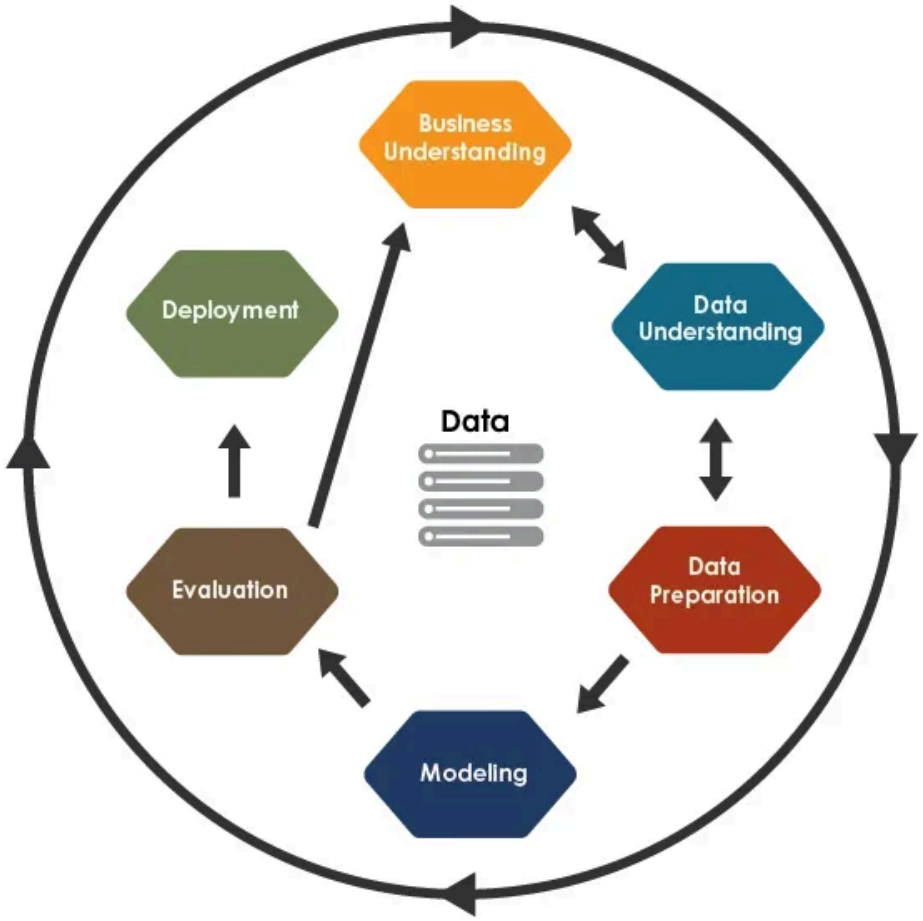
تصميم النظام ليكون قادراً على استقبال بيانات جديدة وتوليد تنبؤات فورية، مما يدعم اتخاذ القرارات اللحظية في بيئة التجارة الإلكترونية.

الاستدامة والقابلية للتوسع

تطوير الكود البرمجي ليكون **Scalable** وسهل الصيانة، مع ضمان توافق مؤشرات الأداء التقنية مع أهداف العمل الاستراتيجية.

التوثيق والجاهزية التشغيلية

ضمان توثيق كافة مراحل خط الأنابيب لسهولة النشر (Deployment) والمراقبة المستمرة لأداء النموذج في بيئة الإنتاج الحقيقية.



مخطط تدفق البيانات في بيئة الإنتاج

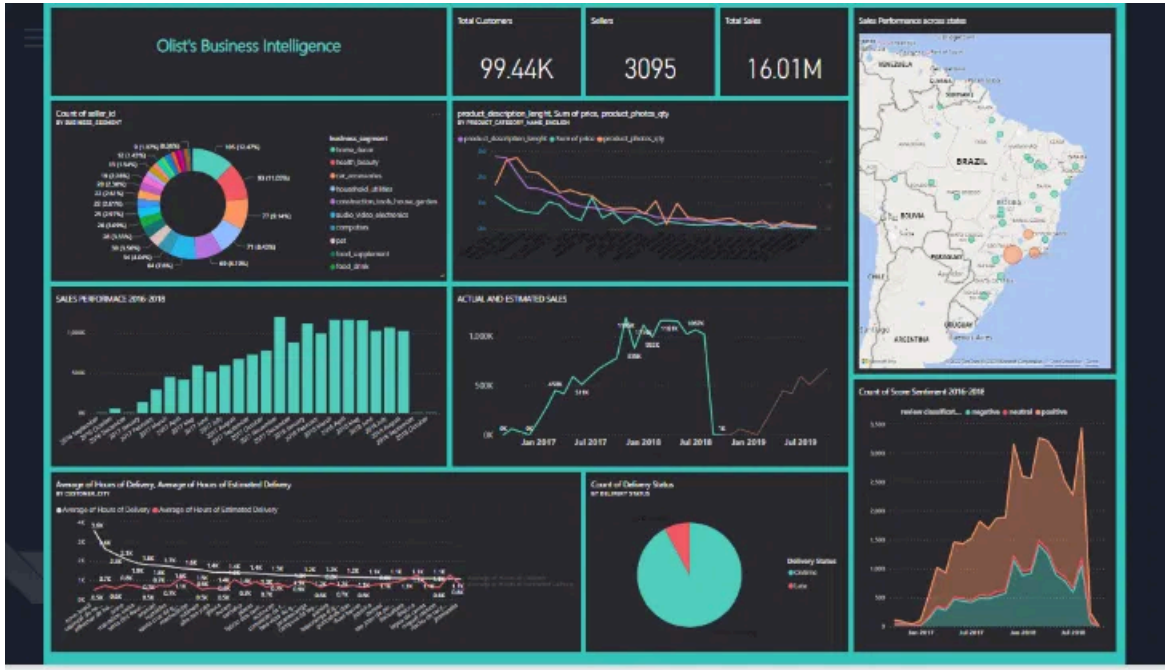
الخلاصة والتوصيات المستقبلية

خلاصة المشروع

نجح المشروع في تطبيق دورة حياة CRISP-DM كاملة، مما أدى إلى بناء نموذج تنبؤي متين لمبيعات Olist. أثبتت النتائج أن النماذج الخطية المنظمة (Ridge) توفر توازناً مثالياً بين الدقة وقابلية التعميم في هذا السياق الزمني.

التوصيات المستقبلية

- دمج بيانات خارجية مثل المؤشرات الاقتصادية البرازيلية ومعدلات التضخم لتحسين دقة التنبؤ طويل الأمد.
- تجربة نماذج السلاسل الزمنية العميقة مثل LSTM أو النماذج الاحتمالية مثل Prophet للتعامل مع الموسمية المعقدة.
- توسيع نطاق النموذج ليشمل التنبؤ على مستوى البائعين الأفراد لتعزيز الدعم اللوجستي المخصص.



تحليل المبيعات الجغرافي والزمني لمنصة Olist

تحويل البيانات إلى قرارات: نحو تجارة إلكترونية أكثر ذكاءً.