تقرير تحليل بيانات المبيعات والتنبؤ بها باستخدام هندسة الميزات المتقدمة وتعلم الآلة

مقدمة

يهدف هذا التقرير إلى عرض عملية تحليل بيانات مبيعات التجزئة بهدف التنبؤ بكميات المبيعات وعدد المعاملات المستقبلية. تم تحقيق ذلك من خلال تطبيق تقنيات هندسة الميزات المتقدمة وبناء نماذج تعلم آلة قوية. يستند التحليل إلى مجموعة بيانات تاريخية للمبيعات، مع التركيز على استخلاص رؤى قيمة وتحقيق تنبؤات دقيقة يمكن أن تدعم عمليات اتخاذ القرار في قطاع التجزئة. تم استخدام مجموعة متنوعة من المكتبات البرمجية الشائعة في مجال علم البيانات، بما في ذلك Pandas لمعالجة البيانات، و NumPy للعمليات العددية، و LightGBM لبناء نماذج التدرج المعزز، و Scikit-learn لتقييم النماذج، بالإضافة إلى Matplotlib و Seaborn للتصور البياني.

استكشاف البيانات الأولي وهندسة الميزات

بدأت عملية التحليل بتحميل البيانات من ملف بصيغة Parquet. تم فحص البيانات الأولية للتعرف على هيكلها وخصائصها. أظهر الفحص الأولي للبيانات عدم وجود قيم مفقودة أو صفوف مكررة، مما يشير إلى جودة البيانات الأولية. تم عرض معلومات مفصلة حول أنواع البيانات لكل عمود وحجم الذاكرة المستخدمة، مما ساعد في فهم طبيعة المتغيرات المتاحة.

بعد ذلك، تم تنفيذ مرحلة مكثفة من هندسة الميزات لإنشاء متغيرات جديدة يمكن أن تحسن من أداء نماذج التنبؤ. شملت هذه المرحلة الخطوات التالية:

1. **إنشاء ميزات متعلقة بالوقت:** تم استخلاص مكونات زمنية متعددة من عمود التاريخ الرئيسي، مثل اليوم، وأسبوع السنة، والشهر، والسنة، ويوم الأسبوع. كما تم تحديد ما إذا كان اليوم يقع في عطلة نهاية الأسبوع، أو بداية الشهر، أو نهاية الشهر. هذه الميزات الزمنية ضرورية لالتقاط الأنماط الموسمية والدورية في المبيعات والمعاملات.
2. **إنشاء ميزات التأخر والمتوسطات المتحركة للمبيعات (Sales Lag and Rolling Features):** لتمكين النموذج من التعلم من الاتجاهات السابقة في المبيعات، تم إنشاء ميزات تأخر (lag features) لوحدات المبيعات لفترات زمنية مختلفة (يوم واحد، 7 أيام، 14 يومًا، و 28 يومًا). بالإضافة إلى ذلك، تم حساب المتوسطات المتحركة (rolling means) لوحدات المبيعات على نوافذ زمنية متعددة (7 أيام، 14 يومًا، و 28 يومًا). تساعد هذه الميزات في التقاط الاعتماد الذاتي في السلاسل الزمنية للمبيعات.
3. **إنشاء ميزات التأخر والمتوسطات المتحركة للمعاملات (Transaction Lag and Rolling Features):** بشكل مشابه لميزات المبيعات، تم إنشاء ميزات تأخر ومتوسطات متحركة لعدد المعاملات. تم استخدام فترات تأخر (يوم واحد، 7 أيام، 14 يومًا) ونوافذ متوسطات متحركة (7 أيام، 14 يومًا) لمتغير المعاملات. تهدف هذه الميزات إلى نمذجة الديناميكيات الزمنية لعدد المعاملات في المتاجر.

4. **إنشاء ميزات إحصائية مجمعة (Grouped Statistical Features)**: تم حساب إحصائيات مجمعة، مثل متوسط المعاملات، بناءً على مجموعات مختلفة مثل عائلة المنتج (family)، ورقم المتجر (store_nbr)، والجمع بينهما. هذه الميزات تساعد في التقاط الخصائص الفريدة لكل متجر أو عائلة منتجات.

بعد إنشاء هذه الميزات الجديدة، تم فرز البيانات بناءً على رقم المتجر، وعائلة المنتج، والتاريخ لضمان الترتيب الزمني الصحيح قبل تطبيق عمليات مثل `shift` و `rolling`. نتج عن عمليات إنشاء ميزات التأخر والمتوسطات المتحركة ظهور قيم مفقودة (NaN) في بداية السلاسل الزمنية لكل مجموعة. لمعالجة ذلك، تم حذف الصفوف التي تحتوي على أي قيم مفقودة في هذه الميزات المشتقة حديثًا، مما يضمن أن البيانات المستخدمة في تدريب النماذج كاملة.

بناء نموذج التنبؤ بمبيعات الوحدات

تم التركيز في هذا القسم على بناء نموذج لتوقع متغير "unit_sales" (مبيعات الوحدات). تم اختيار مجموعة من الميزات التي يُعتقد أنها الأكثر تأثيرًا على مبيعات الوحدات، وشملت هذه الميزات المتغيرات الأصلية مثل رقم المتجر، عائلة المنتج، حالة العرض الترويجي، قابلية المنتج للتلف، سعر النفط، درجة الحرارة، هطول الأمطار، بالإضافة إلى الميزات المشتقة المتعلقة بالوقت، وميزات التأخر والمتوسطات المتحركة للمبيعات.

قبل تدريب النموذج، تم التعامل مع أي قيم مفقودة متبقية (إن وجدت بعد الخطوة السابقة) عن طريق تعويضها بالقيمة -1، وكذلك تم استبدال أي قيم لانهائية (inf, -inf) بالقيمة -1. تم تقسيم مجموعة البيانات إلى مجموعتي تدريب واختبار بناءً على التاريخ؛ حيث أُعتبرت البيانات قبل تاريخ 1 يوليو 2017 كبيانات تدريب، والبيانات من هذا التاريخ فصاعدًا كبيانات اختبار (تحقق).

نظرًا لأن توزيع متغير مبيعات الوحدات قد يكون منحرفًا، تم تطبيق تحويل لوغاريتمي (log1p) على متغير الهدف (unit_sales) لكل من مجموعتي التدريب والاختبار. هذا التحويل يساعد في جعل التوزيع أقرب إلى التوزيع الطبيعي وتحقيق استقرار التباين، مما قد يحسن أداء النموذج.

تم اختيار نموذج LightGBM، وهو تطبيق فعال لخوارزمية التدرج المعزز (Gradient Boosting Decision Tree - GBDT)، لتدريب نموذج التنبؤ بالمبيعات. تم تحديد معلمات النموذج بعناية، بما في ذلك هدف الانحدار (regression)، ومقياس الخطأ التربيعي المتوسط للجذر (rmse) كمقياس للتقييم، ومعدل تعلم منخفض (0.02)، وعدد أوراق مناسب (256)، وعمق أقصى للشجرة (8)، بالإضافة إلى معاملات تنظيم (L1 و L2) لتجنب التجهيز المفرط (overfitting). تم تدريب النموذج على بيانات التدريب مع استخدام بيانات الاختبار كمرجع للتحقق من الأداء وإيقاف التدريب مبكرًا (early stopping) إذا لم يتحسن أداء النموذج على بيانات التحقق لعدد معين من الجولات (300 جولة في هذه الحالة).

أظهرت نتائج تدريب النموذج توقفًا مبكرًا بعد 3593 جولة، مما يشير إلى أن النموذج قد وصل إلى أفضل أداء له على بيانات التحقق. تم بعد ذلك استخدام النموذج المدرب للتنبؤ بمبيعات الوحدات على مجموعة بيانات التحقق. تم عكس التحويل اللوغاريتمي (باستخدام np.expm1) للحصول على التنبؤات بالقيم الأصلية لمبيعات الوحدات، مع التأكد من أن القيم المتوقعة ليست سلبية (باستخدام np.maximum(0, y_pred_sales)).

تم تقييم أداء النموذج باستخدام عدة مقاييس. بلغ مقياس الخطأ اللوغاريتمي التربيعي المتوسط للجذر (RMSLE) على بيانات التحقق حوالي 0.37087. كما تم حساب مقاييس أخرى مثل الخطأ التربيعي المتوسط للجذر (RMSE) الذي بلغ 42.25، ومتوسط مبيعات الوحدات الفعلي في بيانات التحقق الذي كان 140.58. بناءً على ذلك، كان الخطأ التربيعي المتوسط للجذر النسبي حوالي 30.06%. وأخيرًا، بلغ معامل التحديد (R^2 Score) قيمة 0.9588، مما يشير إلى أن النموذج يفسر نسبة عالية جدًا من التباين في مبيعات الوحدات.

بناء نموذج التنبؤ بالمعاملات

بشكل موازٍ لنموذج المبيعات، تم بناء نموذج آخر للتنبؤ بمتغير "transactions" (عدد المعاملات). تم اختيار مجموعة ميزات مشابهة لتلك المستخدمة في نموذج المبيعات، مع استبدال ميزات التأخر والمتوسطات المتحركة للمبيعات بمثيلاتها الخاصة بالمعاملات.

تم اتباع نفس خطوات معالجة البيانات الأولية، بما في ذلك التعامل مع القيم المفقودة واللانهائية. بالإضافة إلى ذلك، تم تطبيق خطوة لمعالجة القيم المتطرفة (outliers) في متغير المعاملات. تم تحديد الحد الأعلى للقيم المقبولة عند النسبة المئوية 99.5 لتوزيع المعاملات، وتم قص (clip) أي قيم تتجاوز هذا الحد إلى هذا الحد الأعلى، مع ضمان عدم وجود قيم سلبية.

تم تقسيم البيانات إلى مجموعتي تدريب واختبار بنفس الطريقة المتبعة في نموذج المبيعات، وتم تطبيق تحويل لوغاريتمي (log1p) على متغير الهدف (transactions).

استُخدم نموذج LightGBM مرة أخرى لتدريب نموذج التنبؤ بالمعاملات، مع استخدام نفس مجموعة المعلمات الأساسية التي أثبتت فعاليتها في نموذج المبيعات. تم تدريب النموذج مع آلية الإيقاف المبكر.

أظهرت نتائج تدريب نموذج المعاملات توقفًا مبكرًا بعد 5319 جولة. عند التنبؤ على بيانات التحقق وتقييم الأداء، بلغ مقياس RMSLE للمعاملات حوالي 0.31791. أما بالنسبة للمقاييس الأخرى، فقد بلغ الخطأ التربيعي المتوسط للجذر (RMSE) للمعاملات 580.06، وكان متوسط عدد المعاملات الفعلي في بيانات التحقق 1725.03. وبالتالي، كان الخطأ التربيعي المتوسط للجذر النسبي حوالي 33.63%. وبلغ معامل التحديد (R^2 Score) قيمة 0.8409، مما يشير إلى قدرة جيدة للنموذج على تفسير التباين في عدد المعاملات.

الخلاصة والتوصيات

نجح هذا التحليل في بناء نماذج تعلم آلة قوية باستخدام LightGBM للتنبؤ بكل من مبيعات الوحدات وعدد المعاملات في متاجر التجزئة. أظهرت عملية هندسة الميزات المتقدمة، التي شملت إنشاء ميزات زمنية، وميزات تأخر، ومتوسطات متحركة، وإحصائيات مجمعة، أهميتها في تحسين دقة النماذج. حقق نموذج التنبؤ بمبيعات الوحدات أداءً ممتازًا مع قيمة R^2 Score بلغت 0.9588، بينما حقق نموذج التنبؤ بالمعاملات أداءً جيدًا أيضًا مع R^2 Score بقيمة 0.8409.

تشير هذه النتائج إلى أن النماذج المطورة يمكن أن تكون أدوات قيمة لدعم عمليات التخطيط وإدارة المخزون وتحسين الكفاءة التشغيلية في قطاع التجزئة. يمكن استخدام التنبؤات الدقيقة

للمبيعات والمعاملات لتحسين جدولة الموظفين، وتخطيط الحملات الترويجية، وضمان توافر المنتجات.

لتحسينات مستقبلية، يمكن استكشاف ميزات إضافية، مثل تأثير العطلات والأحداث الخاصة بشكل أكثر تفصيلاً، أو دمج بيانات اقتصادية أوسع. كما يمكن تجربة خوارزميات تعلم آلة أخرى أو تقنيات تجميع النماذج (ensembling) لمحاولة تحقيق دقة أعلى. بالإضافة إلى ذلك، يمكن النظر في تحليل أهمية الميزات بشكل مفصل لكل نموذج لفهم العوامل الرئيسية التي تؤثر على المبيعات والمعاملات بشكل أعمق. وأخيرًا، يُوصى بحفظ النماذج المدربة (باستخدام مكتبة مثل joblib) لتسهيل إعادة استخدامها في تطبيقات التنبؤ المستقبلية دون الحاجة لإعادة التدريب من الصفر في كل مرة.