الوظيفة الثانية

تحدي سيارات الأجرة

الملخص

1. الموضوع: تحليل نزعة رحلات سيارات الأجرة في مدينة نيويورك
2. الوحدة: السلاسل الزمنية واستكشاف البيانات الكبيرة
3. الأهداف:

إجراء تحليل لرحلات سيارات الأجرة اليومية في مدينة نيويورك باستخدام مجموعة كبيرة من السجلات، حيث يؤول التحليل إلى تحديد الأنماط والاتجاهات في البيانات وإجراء توقعات حول الطلب على سيارات الأجرة والإيرادات المستقبلية بناءً على تلك الأنماط التاريخية.

5. عدد الطلاب: من ثلاثة إلى خمسة طلاب
6. الوقت المخصص: اثنا عشر يومًا

قيود التحقيق

1. لغة البرمجة: Python
2. بيئة العمل: Jupyter Notebook, Colab
3. مجموعة البيانات: NYC Taxi Trips

المتطلبات

1. تنظيف ومكاملة البيانات

a. جوهر الوقت

i. لدينا بيانات تاريخية تحتوي رحلات سيارات الأجرة الصفراء التي جرت بين عامي 2019 و2022 أي خلال الأربع سنوات الماضية في مدينة نيويورك والتي جمعتها عدد من الوكالات المحلية. يطلب إنشاء سلسلة زمنية (Time Series) لعدد الرحلات ومجموع الإيرادات الكلي (Total Revenue) خلال كل ساعة من ساعات النهار من أجل كل وكالة ضمن مجموعة البيانات.

ii. قبل البدء بعملية البناء يجب عليك تنظيف مجموعة البيانات أي يجب عليك حل مشكلات القيم الفارغة والمتطرفة وكذلك المتناقضة بطريقة تراها مناسبة.

iii. يجب أن تحوي مجموعة البيانات 120 مليون سجل على الأقل قبل القيام بعملية بناء السلسلة الزمنية.
iv. حاول العمل ضمن بيئة موزعة (استخدم Docker إن استطعت وذلك لسهولة الضبط)
v. قم بإجراء استطلاع بسيط لمجموعة البيانات من أجل تسهيل عملية التنظيف.
vi. قم بالاحتفاظ بمجموعة البيانات النظيفة على القرص الصلب بصيغة parquet لاستخدامها في المراحل التالية.

توجيه: يجب أن تكون كافة عمليات التلاعب بالبيانات باستخدام Dask بسبب الحجم الهائل لعدد السجلات، فلا تحاول استخدام Pandas إلا لتخزين النتائج الإحصائية والجداول صغيرة الحجم.

b. ضبط آلة الزمن

i. قم باستعمال تقنيات معالجة السلاسل للتأكد من أن المجموعة النهائية تحقق شروط بيانات السلاسل الزمنية.
ii. قم بتخزين مجموعة البيانات النهائية في جدول من الشكل الطويل (Long-Format) على القرص الصلب بصيغة csv.

2. هندسة السمات

- قم بإنشاء سمة source_zone للدلالة على منطقة البداية
- قم بإنشاء سمة destination_zone للدلالة على منطقة النهاية
- قم بإنشاء سمة source_borough للدلالة على بلدية البداية
- قم بإنشاء سمة destination_borough للدلالة على بلدية النهاية
- قم بإنشاء سمة location_pair والتي تتضمن منطقة (zone) البداية والنهاية بغض النظر عن الترتيب (اعتبر السمات تبادلية) بشكل comma separated.
- قم بإنشاء سمة payment_type_name للدلالة على طريقة الدفع
- قم بإنشاء سمة vendor للدلالة على اسم الوكالة
- قم بإنشاء سمة rate_code للدلالة على آلية حساب الأجرة
- قم بإنشاء سمة trip_class والتي تتضمن صنف الرحلة بالاعتماد على تكرارات location_pair ضمن مجموعة المعطيات على النحو التالي
  (rare, less-common, common, more-common) وعين حدود كل صنف بطريقة تراها مناسبة موضحًا ذلك ضمن المفكرة.
- قم بإنشاء سمة trip_duration للدلالة على المدة الزمنية للرحلة (بالدقائق)
- قم بتحويل سمة trip_distance إلى الكيلومترات بدلًا من الأميال المقطوعة.

توجيه: تخلص من سمة التراتبية للبداية والنهاية ضمن سمة location_pair ثم استخدم تقنية التقطيع لإنشاء سمة trip_class.

تذكرة: يمكن أن تحتاج إلى تخزين مجموعة المعطيات على القرص الصلب أكثر من مرة لتخفيف العمليات الحسابية وتصغير حجم Dask Computation Graph.

3. الاستكشاف والتحليل

a. ما وراء حافة الأرقام

i. قم بعرض حصة كل وكالة (Vendor) باستخدام Pie chart
ii. قم بعرض حصة كل بلدية (Borough) باستخدام Pie chart
iii. قم بعرض طرق الدفع المستخدمة باستعمال Bar chart
iv. قم بعرض طرق الدفع المستخدمة ضمن كل صنف من أصناف الرحلة باستعمال sunburst chart.
v. قم بدراسة الارتباط **الخطي** بين الكلفة الإجمالية ومبلغ الإكرامية ومقدار الأجرة وعدد الركاب ومدة الرحلة والمسافة المقطوعة مبينًا رأيك بالنتائج.
vi. قارن بين المتوسط والوسيط والانحراف المعياري للكلفة الإجمالية ومدة الرحلة والمسافة المقطوعة من أجل كل صنف من أصناف الرحلة (statistics per trip_class) باستخدام Bar chart مبينًا رأيك بالنتيجة (استفد من خاصية Facet Row/Col ضمن مكتبة Plotly)
vii. قم بدراسة ترابط السلاسل الزمنية Autocorrelation, Partial autocorrelation
viii. قم بعرض نسخة ملساء من كل سلسلة زمنية باستخدام Rolling Window ومخطط Line Chart
ix. قم بإجراء أي عملية تحليل تراها مناسبة للسلاسل الزمنية.

b. اصطياد الأنماط

i. قم بنمذجة مكونات السلاسل الزمنية لكل وكالة محلية باستعمال نموذج إحصائي تراه مناسبًا ثم استخدم Prophet لبناء النموذج النهائي وقم بعمليات التوليف tuning المناسبة وقارن بين النماذج.
ii. ماذا تستنتج من كل مكون؟
iii. قارن بين الوكالات بناءً على ما وجدته.

c. الوقت قيم

i. تنبأ بالطلب المستقبلي لرحلات وإيرادات سيارات الأجرة باستخدام النماذج النهائية بعد عملية التوليف tuning، ويجب تقييم النماذج ومقارنتها باستخدام المقاييس المناسبة، تفسير النتائج في سياق البيانات وأي عوامل خارجية قد تؤثر على طلب سيارات الأجرة.
ii. وازن بين جودة ملاءمة النموذج وتعقيده، بحيث تُفضل النماذج الأبسط ما لم تكن أقل من أداء النماذج الأكثر تعقيدًا بشكل ملحوظ.

d. عشوائية القصص

i. قم بإجراء عملية sampling للبيانات إلى درجة تستطيع تحليلها ومعالجتها بحيث تكون العينات شاملة قدر الإمكان.
ii. قم بإجراء عملية Clustering باستخدام تقنيات التعلم التلقائي (**ثلاث خوارزميات** على الأكثر و**اثنين** على الأقل) ثم حاول وصف العناقيد وقارن بينها.
iii. قم بإجراء عمليات ما بعد المعالجة المناسبة ثم حدد أهم السمات التي أثرت على قرار أفضل نموذج دربته لكل خوارزمية استخدمتها مبينًا رأيك.

4. التوثيق

كما تعلم، يعد تقرير التوثيق مكونًا حاسمًا في أي مشروع أو مهمة، لأنه بمثابة سجل شامل لما تم إنجازه وكيف تم تحقيقه، مما يسمح لأعضاء الفريق وأصحاب المصلحة في المستقبل بفهم أهداف المشروع والجدول الزمني والتحديات والحلول.

ستكون مسؤولاً عن إنشاء تقرير باللغة العربية يلخص ما قمت به وما توصلت إليه، وتَذَكر أن جودة التقرير المكتوب لا يفيدك أنت وفريقك فحسب، بل يخدم أيضًا كأصل قيم للمقارنة بين عملك وعمل أقرانك، وعليه تحرى الدقة والترتيب في كتابة تقرير التوثيق.

خذ وقتك وكن دقيقًا في وصفك وتأكد من تضمين جميع المعلومات ذات الصلة وقدم تفسيرات وأمثلة واضحة لدعم نتائجك واستخدم الأدوات المناسبة لتصدير صور المخططات ونسق النتائج ضمن جدول إن لزم الأمر ولا تضع أي لقطة شاشة أو جزء من الكود ضمن التقرير!

قيود وقواعد تنظيمية

1. يجب أن يكون تقرير التوثيق بما لا يتجاوز الستة صفحات.
2. عند التقدم كفريق ثلاثي تعتبر عملية المقارنة بين الوكالات طلبًا إضافيًا (يتم إجراء التحليل لبيانات الوكالة ذات الحصة السوقية الأكبر) واجباريًا عدا ذلك.
3. يعد الطلب (عشوائية القصص) طلبًا إضافيًا عند التقدم كفريق ثلاثي واجباريًا عدا ذلك.
4. سيتم الإعلان عن نموذج التقديم في موعد لاحق.
5. سيتم اجراء مقابلة بالوظيفة في موعد لاحق.
6. يحدد يوم الجمعة 16/7 الموعد النهائي لتسليم الوظيفة.
7. سيتم معاقبة الوظائف المتأخرة على النحو التالي:
    a. تأخير يوم واحد: خصم 15٪ من العلامة
    b. تأخير يومين: خصم 30٪ من العلامة
    c. ثلاث أيام متأخرة أو أكثر: لن يتم قبول الوظيفة ويحصل الطالب على درجة الصفر كاملةً!
8. خصص وقتًا كافيًا لإكمال الوظيفة دون تسرع أو طرق ملتوية، ابدأ مبكرًا وخطط مسبقًا للتأكد من أن لديك الوقت الكافي للحصول على البيانات والمعالجة المسبقة والتحليل، ولا تنتظر حتى اللحظة الأخيرة فقد يؤدي ذلك إلى تسريع العمل وزيادة احتمالية سوء السلوك الأكاديمي.

9. يجب أن يكون كل العمل خاص بك. لا تنسخ أو تعيد صياغة أعمال الآخرين دون **اقتباس** مناسب لتوضيح **المصدر**، إذ أن الغش لا يؤدي إلى تقويض عملية التعلم فحسب، بل ينتهك أيضًا الثقة بين الطالب والمدرس ويمكن أن يؤدي إلى عواقب وخيمة، لذلك تجنب مشاركة عملك مع الآخرين أو السماح للآخرين باستخدام عملك.
10. افهم الفرق بين التعاون والغش، يُسمح بالتعاون طالما أنه لا يتضمن جزءًا من عمل زملائك.
11. كن صريحًا وشفافًا بشأن عملك ونظم المفكرة قدر الإمكان.
12. إذا كانت لديك أسئلة بشأن الوظيفة فاطلب التوجيه من المدرس.
13. تأكد من الاطلاع على دليل مكتبة Dask حتى لا تقع في أخطاء فادحة لسوء الاستخدام، ولا تحصر نفسك بما تم اعطائه ضمن المحاضرة!
14. لا تترك أي رسالة خطأ أو رسالة طباعة طويلة جدًا ضمن المفكرة!
15. عند مراجعة كود الوظيفة في حال تبين أنه مأخوذ بالكامل من الانترنت سوف تنال الوظيفة درجة الصفر كاملةً!
16. لا يجب عليك التكلف في استنتاجاتك ضمن المفكرة، اكتبها بأي طريقة تناسبك وبأي لغة تريدها (العربية أو الإنجليزية)، موضحًا المصطلحات إن وجدت باللغة الإنجليزية وذلك ضمن قوسين.
17. في حالة عدم اقتباس مصدر الكود هذا يعني أنك صاحب الفكرة، وهنا انتبه إلى العبء الممكن أن تتحمله في حال تبين خلاف ذلك!، وتذكر أن أي مخالفة للضوابط والقواعد المذكورة وسوء الممارسة الأكاديمية يستلزم العقوبة وفق ما يراه المدرس مناسبًا.