المستند التقني - نظام التنبؤ بانسحاب العملاء

المحتويات

نظرة عامة

تم تطوير نظام متكامل للتنبؤ بانسحاب العملاء من منصة بث الموسيقى باستخدام بيانات سلوك المستخدمين. النظام يتعامل مع تحديات متعددة مثل عدم توازن البيانات وصعوبة تحديد تعريف دقيق للانسحاب.

الأهداف المحققة:

- تطوير نموذج تنبؤ بدقة F1 Score تبلغ 82.4%
- بناء API متكامل باستخدام FastAPI
- تطبيق نظام مراقبة للكشف عن انحراف البيانات
- إنشاء لوحة معلومات تفاعلية باستخدامStreamlit
- تطبيق MLflow لتتبع التجارب

تحليل البيانات وتنظيفها

1. استكشاف البيانات الأولي

تم تحليل مجموعة البيانات المصغرة التي تحتوي على 286,500 سجل لأحداث المستخدمين:

- حجم البيانات: 286,500 سجل
- عدد الأعمدة: 18
- الفترة الزمنية: من 2018-10-01 إلى 2018-12-03
- عدد المستخدمين الفريدين: 225 (بعد التنظيف)

2. التحديات في البيانات

أ. معرفات المستخدمين المفقودة

المشكلة: وجود 8,346 سجل بدون معرف مستخدم (قيمة فارغة) السبب: المستخدمون غير المسجلين أو الذين سجلوا خروج
الحل المبتكر: تطوير خوارزمية لاستنتاج معرف المستخدم بناءً على:

- تسلسل itemInSession يزداد بمقدار 1 لكل إجراء
- معرف الجلسة (sessionId)
- القرب الزمني للأحداث

نتائج الخوارزمية:

- تم استرجاع 8,183 معرف مستخدم (98% نجاح)
- بقي 163 سجل فقط بدون معرف
- تم حذف السجلات المتبقية

ب. إعادة استخدام الجلسات
الاكتشاف: معرفات الجلسات يتم إعادة استخدامها بين مستخدمين مختلفين التأثير: 466 جلسة تحتوي على أكثر من مستخدم
الحل: استخدام مزيج من sessionId و itemInSession لتتبع المستخدمين بدقة

ج. البيانات المفقودة
- length: 20.4% بيانات الأغاني - طبيعي للصفحات غير (NextSong)
- song: 20.4%
- artist: 20.4%

3. معالجة البيانات
تم تطبيق خط معالجة شامل:
1. تحويل التواريخ: من timestamps إلىdatetime
2. تنظيف معرفات المستخدمين: تحويل القيم الفارغة إلى NaN
3. استنتاج المعرفات المفقودة: باستخدام الخوارزمية المطورة
4. استخراج معالم الموقع: فصل المدينة والولاية
5. ملء البيانات المفقودة: ربط خصائص المستخدم بمعرفه

هندسة العناصر
تم إنشاء 40 عنصر مقسمة على 7 فئات رئيسية:
1. معالم النشاط (Activity Features)
- total_events إجمالي الأحداث
- num_sessions عدد الجلسات
- total_interactions مجموع التفاعلات
- events_per_session متوسط الأحداث لكل جلسة

2. معالم الاستماع (Listening Features)
- songs_played عدد الأغاني المشغلة
- total_listening_time إجمالي وقت الاستماع
- avg_song_length متوسط طول الأغنية
- unique_artists عدد الفنانين الفريدين
- unique_songs عدد الأغاني الفريدة
- days_since_registration الأيام منذ التسجيل
- avg_daily_listening_time متوسط وقت الاستماع اليومي
- avg_daily_songs متوسط الأغاني اليومية
- artist_diversity تنوع الفنانين (نسبة الفنانين الفريدين)

3. معالم التفاعل (Engagement Features)
- thumbs_up عدد الإعجابات
- thumbs_down عدد عدم الإعجابات
- total_feedback مجموع التقييمات
- positive_feedback_ratio نسبة التقييمات الإيجابية
- playlist_adds إضافات قوائم التشغيل
- add_friend إضافة أصدقاء
- advert_roll مشاهدة الإعلانات

4. معالم الاشتراك (Subscription Features)

- is_paid هل المستخدم لديه اشتراك مدفوع
- subscription_changes تغييرات الاشتراك
- downgrades عدد تخفيضات الاشتراك
- upgrades عدد ترقيات الاشتراك

5. معالم المشاكل التقنية (Technical Issues Features)

- error_count عدد الأخطاء
- help_visits زيارات صفحة المساعدة
- settings_visits زيارات الإعدادات
- logout_count عدد تسجيلات الخروج
- has_issues مؤشر وجود مشاكل

6. معالم زمنية (Temporal Features)

- days_since_last_activity الأيام منذ آخر نشاط
- days_used_in_period عدد أيام الاستخدام
- days_available_in_period الأيام المتاحة في الفترة
- usage_frequency تكرار الاستخدام

7. معالم الجلسات (Session Features)

- avg_session_length متوسط طول الجلسة
- session_length_std انحراف طول الجلسة
- max_session_length أقصى طول جلسة
- avg_session_duration_mins متوسط مدة الجلسة بالدقائق
- session_duration_std_mins انحراف مدة الجلسة
- max_session_duration_mins أقصى مدة جلسة
- session_consistency ثبات الجلسات

8. معالم ديموغرافية (Demographic Features)

- gender_F, gender_M الجنس(one-hot encoded)
- state_CA, state_TX, إلخ: الولاية (top 3 + Other)

بناء النموذج

1. تعريف الانسحاب (Churn Definition)

تم تعريف الانسحاب بناءً على زيارة صفحة "Cancellation Confirmation

- إجمالي المستخدمين: 225
- المستخدمون المنسحبون: 52
- معدل الانسحاب: 23.1%

2. النماذج المختبرة

تم اختبار 4 نماذج مختلفة مع ضبط معاملاتها:

أ. Logistic Regression

- الأداء F1 Score = 0.824 :
- المعاملات المثلى :

- feature_selection__k: 25
- model__C: 1
- model__class_weight: {0: 1, 1: 2}
- model__penalty: 'l1'

ب. Random Forest

- الأداء F1 Score = 0.687 :
- المعاملات المثلى :
  - n_estimators: 200
  - max_depth: 10
  - class_weight: 'balanced'

ج. Gradient Boosting

- الأداء F1 Score = 0.766 :
- المعاملات المثلى :
  - n_estimators: 200
  - learning_rate: 0.05
  - max_depth: 3

د. XGBoost

- الأداء F1 Score = 0.791 :
- المعاملات المثلى :
  - n_estimators: 100
  - max_depth: 4
  - scale_pos_weight: 2

3.معالجة عدم توازن البيانات

تم استخدام عدة تقنيات:

- class_weight في النماذج
- scale_pos_weight في XGBoost
- التركيز على F1 Score كمقياس أساسي
- استخدام StratifiedKFold للحفاظ على توزيع الفئات

4.اختبار الاستقرار

تم اختبار النماذج عبر 5 بذور عشوائية مختلفة:

- Logistic Regression: الأكثر استقراراً (انحراف معياري 0.0058)
- متوسط الأداء عبر البذور: 0.824
- النطاق: 0.819 - 0.835

5.أهم المعالم

المعالم الأكثر تأثيراً في التنبؤ:

1. days_since_last_activity: 3.022 أقوى معامل
2. session_consistency: -1.363
3. artist_diversity: -1.286
4. session_length_std: -0.721
5. add_friend: -0.461

ملاحظة: بعض المعالم لها معامل صفر بسبب L1 regularization

البنية التقنية

1. هيكل المشروع

```
customer-churn-prediction/
├── src/
│   ├── data/
│   │   ├── preprocessing.py
│   │   └── feature_engineering.py
│   ├── models/
│   │   ├── train.py
│   │   └── predict.py
│   ├── api/
│   │   ├── main.py
│   │   └── schemas.py
│   ├── monitoring/
│   │   ├── drift_detection.py
│   │   └── performance_tracking.py
│   └── utils/
│       └── config.py
├── models/
├── mlruns/
├── tests/
├── notebooks/
├── dashboard/
└── docker/
```

2. التقنيات المستخدمة

- Python 3.13.5
- FastAPI للواجهة البرمجية
- MLflow لتتبع التجارب
- Streamlit للوحة المعلومات
- Docker للحاويات
- PostgreSQL لقاعدة البيانات
- Redis للذاكرة المؤقتة

3. واجهة برمجة التطبيقات (API)

تم تطوير API باستخدام FastAPI يوفر:

- /predict: التنبؤ لمستخدم واحد
- /batch_predict: التنبؤ لعدة مستخدمين
- /model/info: معلومات النموذج الحالي
- /update_user_events: تحديث أحداث المستخدم

4. لوحة المعلومات

لوحة معلومات تفاعلية (mock up) لهدف العرض فقط؛ تعرض:

- معدلات الانسحاب الحالية
- توزيع مستويات المخاطر

- أداء النموذج بمرور الوقت
- كشف انحراف البيانات
- التنبؤات الحديثة

التحديات التقنية

1. استنتاج معرفات المستخدمين المفقودة

التحدي الأكبر كان في تطوير خوارزمية ذكية لاستنتاج معرفات المستخدمين المفقودة. الخوارزمية تعتمد على:

- فهم أن itemInSession يزداد تسلسلياً
- استخدام القرب الزمني كعامل حاسم
- التحقق من صحة التسلسل قبل الإسناد

2.إعادة استخدام الجلسات

اكتشاف أن الجلسات يُعاد استخدامها تطلب إعادة تفكير في كيفية تتبع المستخدمين وربط الأحداث.

3.تعريف الانسحاب

كان من الصعب تحديد تعريف دقيق للانسحاب. تم اختبار:

- الانسحاب الصريح: زيارة صفحة إلغاء الاشتراك
- الانسحاب الضمني: عدم النشاط لفترة طويلة
- تم اعتماد التعريف الصريح فقط لدقته

4.عدم توازن البيانات

معدل انسحاب 23.1% يمثل عدم توازن معتدل. تم معالجته باستخدام:

- أوزان الفئات المناسبة
- اختيار المقاييس المناسبة(F1 Score)
- التحقق من الأداء على كل فئة منفصلة

5.تسرب البيانات

تم تجنب تسرب البيانات عبر:

- عدم استخدام معالم مستقبلية
- استخدام cross-validation صحيح
- فصل البيانات زمنياً عند الحاجة

استراتيجية إعادة التدريب

1.معايير إعادة التدريب

النظام مصمم لإعادة التدريب بناءً على:

أ. معيار زمني

- إعادة تدريب كل 30 يوم تلقائياً

ب. معيار الأداء

- إذا انخفض F1 Score عن 0.75

ج. معيار انحراف البيانات

- عند اكتشاف انحراف كبير في توزيع المعالم

2.عملية إعادة التدريب

1. جمع البيانات الجديدة
2. معالجة وهندسة المعالم
3. تدريب النموذج مع تتبعMLflow
4. التحقق على مجموعةholdout
5. نشر النموذج إذا تجاوز معايير الأداء

3.الجدولة

تم كتابة كود الجدولة باستخدام:schedule library

- فحص يومي في الساعة 2:00 صباحاً
- تقييم المعايير
- إعادة التدريب إذا لزم الأمر

ملاحظة: الجدولة غير مفعلة حالياً في البيئة التطويرية

المراقبة والأداء

1.كشف انحراف البيانات(Data Drift)

تم تطبيق نظام لكشف الانحراف باستخدام:

- اختبار Kolmogorov-Smirnov للمعالم الرقمية
- عتبة p-value = 0.05
- مراقبة كل معلم على حدة

2.كشف انحراف المفاهيم(Concept Drift)

مراقبة أداء النموذج عبر نوافذ زمنية:

- تقسيم التنبؤات إلى 10 نوافذ
- حساب الأداء لكل نافذة
- كشف الاتجاه باستخدامlinear regression

3.تتبع الأداء

- تسجيل كل تنبؤ مع الوقت
- حساب المقاييس بشكل دوري
- تنبيهات عند انخفاض الأداء

4.لوحة المراقبة

لوحة معلومات في الوقت الفعلي تعرض:

- معدل الانسحاب الحالي مقابل المتوقع
- درجات انحراف المعالم
- أداء النموذج بمرور الوقت
- التنبؤات الحديثة ومستويات المخاطر

التحسينات المقترحة

1.تحسينات على مستوى البيانات

- إضافة معالم خارجية (مثل الموسم، العطلات)
- تتبع سلوك المستخدم على مستوى أعمق (تفضيلات الموسيقى ونوعها)

2.تحسينات على مستوى النموذج

- تجربة تقنيات ensemble أكثر تطوراً
- استخدام deep learning للتعامل مع التسلسلات الزمنية
- تطبيق AutoML لاستكشاف مساحة أكبر من النماذج
- تجربة تقنية SMOTE لمعالجة عدم التوازن

3.تحسينات على مستوى النظام

- نقل البيانات والنماذج إلى السحابة(AWS S3, Azure Blob)
- تطبيق CI/CD pipeline كامل
- إضافة A/B testing للنماذج الجديدة

4.تحسينات على مستوى المراقبة

- إضافة تنبيهات في الوقت الفعلي(Slack, Email)
- تطوير dashboards أكثر تفصيلاً
- تتبع business metrics بجانبmodel metrics

5.تحسينات على مستوى الأعمال

- تطوير نظام توصيات لمنع الانسحاب
- تحديد أسباب الانسحاب لكل مستخدم
- إنشاء segments للمستخدمين حسب خطر الانسحاب
- تطوير استراتيجيات تدخل مخصصة

الخلاصة

تم تطوير نظام متكامل للتنبؤ بانسحاب العملاء باستخدام البيانات الmini فقط؛ يحقق نتائج ممتازة = F1 Score)
(82.4% مع إمكانيات للتطوير المستقبلي. النظام يتضمن جميع المكونات المطلوبة من معالجة البيانات إلى النشر والمراقبة.

النقاط الرئيسية:

- حل مبتكر لمشكلة البيانات المفقودة
- نموذج مستقر وقابل للتفسير
- بنية تقنية قابلة للتوسع
- نظام مراقبة شامل

التحديات المتبقية:

- نقل النظام إلى بيئة الإنتاج الفعلية
- تفعيل الجدولة الأوتوماتيكية
- ربط النظام بقواعد البيانات الحقيقية
- تحسين أداء Docker containers