4.10 model card + evaluation report

کارت مدل و گزارش ارزیابی - ProDecks
نسخه: 1.0
تاریخ: 1403/11/17
وضعیت: مدل‌های MVP - ارزیابی شده

فصل ۱: مرور کلی مدل‌ها

۱.۱. هدف این سند

ارائه اطلاعات جامع درباره مدل‌های یادگیری ماشین استفاده‌شده در ProDecks، ارزیابی، شامل نحوه آموزش، محدودیت‌ها و شرایط عملکرد بهینه.

۱.۲. مدل‌های موجود در MVP

۱. مدل پیشنهاد کارت (Card Recommender) - CR-MODEL-v1
۲. مدل پیش‌بینی زمان تکمیل (Completion Time Predictor) - CTP-MODEL-v1
۳. مدل تشخیص blockage (Blocker Detector) - BD-MODEL-v1

۱.۳. اصول اخلاقی و مسئولیت

• شفافیت: افشای کامل قابلیت‌ها و محدودیت‌ها
• عدالت: بررسی و کاهش bias
• مسئولیت‌پذیری: تعیین مسئول برای هر مدل
• امنیت: محافظت از مدل‌ها در برابر سوءاستفاده

فصل ۲: مدل پیشنهاد کارت (CR-MODEL-v1)

۲.۱. هدف مدل

پیشنهاد خودکار کارت‌های مشابه هنگام ایجاد کارت جدید، بر اساس الگوهای تاریخی کاربر و تیم.

۲.۲. معماری مدل

• نوع: Collaborative Filtering با Neural Network Enhancement
• لایه‌ها:
- Input Layer: ویژگی ۵۰۰
- Embedding Layer: بعد ۱۰۰

- Hidden Layers: ۲ لایه (۶۴، ۳۲ نورون)
- Output Layer: احتمال تطابق کارت‌ها

• پارامترها: ۱۵۰,۰۰۰ پارامتر قابل آموزش
• حجم: ۱۲ MB

۲.۳. داده‌های آموزشی

• منبع: داده‌های ناشناس‌شده ۱۰۰,۰۰۰ کارت از ۱,۰۰۰ کاربر
• دوره: ۶ ماه گذشته
• ویژگی‌ها:
- متن کارت (TF-IDF vectorized با)
- برچسب‌های کاربر
- تاریخ ایجاد
- کاربر ایجادکننده
- تیم مربوطه

• تقسیم داده:
- Training: ۷۰٪
- Validation: ۱۵٪
- Test: ۱۵٪

۲.۴. ارزیابی عملکرد

• دقت (Accuracy): ۷۸.۵٪
• Precision@10: ۶۵.۲٪
• Recall@10: ۵۸.۷٪
• F1-Score: ۶۱.۷٪
• AUC-ROC: ۰.۸۲

۲.۵. شرایط بهینه عملکرد

• حداقل داده: ۱۰۰ کارت تاریخی
• زبان: فارسی و انگلیسی
• زمینه: مدیریت پروژه نرم‌افزاری
• کیفیت داده: متن‌های توصیفی حداقل ۱۰ کلمه

۲.۶. محدودیت‌ها و خطاها

۱. خطای رایج: پیشنهاد کارت‌های نامرتبط در پروژه‌های جدید
• علت: عدم داده کافی برای پروژه جدید
• راه‌حل: استفاده از داده‌های عمومی تا جمع‌آوری داده کافی

۲. خطای رایج: bias به سوی کاربران فعال

• علت: داده‌های بیشتر از کاربران فعال
• راه‌حل: نمونه‌گیری متعادل

۳. خطای رایج: عدم درک context پیچیده
• علت: محدودیت مدل در درک nuance
• راه‌حل: افزودن contextual features

۲.۷. تست robustness

• تست adversarial: مقاوم در برابر ورودی‌های نامربوط
• تست out-of-distribution: کاهش عملکرد در دامنه‌های کاملاً جدید
• تست stress: عملکرد پایدار تا ۱۰۰ درخواست/ثانیه

فصل ۳: مدل پیش‌بینی زمان تکمیل (CTP-MODEL-v1)

۳.۱. هدف مدل

پیش‌بینی زمان واقعی تکمیل کارت‌ها بر اساس ویژگی‌های کارت و تاریخچه تیم.

۳.۲. معماری مدل

• نوع: Gradient Boosting Regression (XGBoost)
• درختان: ۲۰۰ درخت
• عمق حداکثر: ۶
• نرخ یادگیری: ۰.۱

۳.۳. داده‌های آموزشی

• منبع: ۵۰,۰۰۰ کارت تکمیل‌شده
• ویژگی‌ها:
- طول متن توصیف
- تعداد subcards
- اولویت
- مسئول (تجربه قبلی)
- تاریخ‌های سررسید قبلی
- فصل و روز هفته

• هدف: زمان واقعی تکمیل (ساعت)

۳.۴. ارزیابی عملکرد

• RMSE: ۱۲.۵ ساعت
• MAE: ۸.۲ ساعت
• R^2 Score: ۰.۷۴

• Mean Absolute Percentage Error: ۲۸٪

تفسیر اهمیت ویژگی‌ها ۳.۵.

تجربه مسئول: ۳۵٪ اهمیت ۱.
طول متن: ۲۰٪ اهمیت ۲.
اهمیت ۱۵٪ :subcards تعداد ۳.
اولویت: ۱۰٪ اهمیت ۴.
عوامل دیگر: ۲۰٪ اهمیت ۵.

محدودیت‌ها ۳.۶.

(> ۵ subcards) دقت پایین برای کارت‌های بسیار پیچیده •
عدم در نظر گرفتن عوامل خارجی (مرخصی، تعطیلات) •
نیاز به به‌روزرسانی مداوم با داده‌های جدید •

bias تست ۳.۷.

(Δ < ۲٪) ناچیز bias :بر اساس جنسیت bias بررسی •
به نفع کاربران با تجربه bias بر اساس تجربه: کمی bias بررسی •
بر اساس زبان: عملکرد مشابه فارسی و انگلیسی bias بررسی •

فصل ۴: مدل تشخیص blockage (BD-MODEL-v1)

هدف مدل ۴.۱.

.دارند قبل از وقوع blockage تشخیص خودکار کارت‌هایی که احتمال

معماری مدل ۴.۲.

Neural Network با Binary Classification :نوع •
Dense Layers + برای دنباله‌های زمانی LSTM :لایه‌ها •
پارامترها: ۸۵,۰۰۰ پارامتر •

داده‌های آموزشی ۴.۳.

blocked کارت ۲,۰۰۰ :(blocked) مثبت •
کارت عادی ۱۸,۰۰۰ :(not blocked) منفی •
کلاس مثبت oversampling :توازن •

ارزیابی عملکرد ۴.۴.

٪دقت: ۸۷.۳ •
• Precision: ۷۶.۵٪
• Recall: ۷۲.۱٪

• F1-Score: ۷۴.۲٪
• AUC-ROC: ۰.۸۸

۴.۵. ماتریس اشتباه (Confusion Matrix)

	پیش‌بینی Not Blocked	پیش‌بینی Blocked
واقعی Blocked	۱۱۵	۲۸۵
واقعی Not Blocked	۱۵۱۰	۹۰

۴.۶. خطاهای بحرانی

۱. False Negative (عدم تشخیص blockage واقعی):
• تأثیر: تأخیر در پروژه
• دفعات: ۱۱۵ از ۴۰۰ (۲۸.۷٪)
• راه‌حل: کاهش threshold

۲. False Positive (تشخیص blockage نادرست):
• تأثیر: هشدارهای بی‌دلیل
• دفعات: ۹۰ از ۱۶۰۰ (۵.۶٪)
• راه‌حل: افزودن ویژگی‌های contextual

۴.۷. تست adversarial

• مقاوم در برابر دستکاری عمدی ویژگی‌ها
• عملکرد پایدار با داده‌های نویزی
• تشخیص تلاش برای فریب مدل

فصل ۵: ارزیابی جامع مدل‌ها

۵.۱. ارزیابی روی سناریوهای واقعی

سناریو ۱: تیم جدید (کم‌داده)
• مدل CR: عملکرد ضعیف (Accuracy: ۴۵٪)
• مدل CTP: خطای بالا (MAPE: ۵۰٪)
• مدل BD: عملکرد قابل قبول (Accuracy: ۷۰٪)

سناریو ۲: تیم بزرگ (داده غنی)
• مدل CR: عملکرد عالی (Accuracy: ۸۵٪)
• مدل CTP: دقت بالا (MAPE: ۲۰٪)
• مدل BD: عملکرد عالی (Accuracy: ۹۰٪)

سناریو ۳: پروژه پیچیده
• همه مدل‌ها: کاهش عملکرد ۱۵-۲۰٪
• نیاز به ensemble با قوانین دستی

۵.۲. ارزیابی robustness

• تغییرات کوچک در داده: عملکرد پایدار
• تغییرات بزرگ در توزیع: نیاز به retraining
• ورودی‌های 极端: handling مناسب

۵.۳. ارزیابی bias

• بررسی بر اساس زیرگروه‌ها:
- تجربه کاربران: bias متوسط
- اندازه تیم: bias کم
- نوع صنعت: bias قابل توجه

• اقدامات کاهش bias:
- داده‌های متعادل
- regularization
- post-processing

فصل ۶: تفسیرپذیری و توضیح‌پذیری

۶.۱. روش‌های تفسیر

۱. SHAP Values: برای مدل‌های tree-based
۲. LIME: برای مدل‌های neural
۳. Feature Importance: برای تمام مدل‌ها
۴. Attention Weights: برای مدل‌های sequence

۶.۲. نمونه تفسیر

CTP: برای مدل
"۳ subcard، این کارت به دلیل طولانی بودن توضیحات (۴۰۰ کلمه) و داشتن
و اختصاص به کاربری با تجربه متوسط، احتمالاً ۴۰ ساعت زمان خواهد برد."

۶.۳. محدودیت‌های تفسیر

• تفسیرهای تقریبی، نه قطعی
• عدم توانایی در تفسیر تعاملات پیچیده
• نیاز به تخصص برای درک برخی تفسیرها

فصل ۷: امنیت و حریم خصوصی

۷.۱. حفاظت از مدل‌ها

• Watermarking: برای تشخیص سرقت

• Model Encryption: در حالت استراحت
• Access Control: دسترسی محدود

۷.۲. حریم خصوصی داده‌ها

• Differential Privacy: ε = ۱.۰
• Federated Learning: برای آینده برنامه‌ریزی شده
• Anonymization: قبل از آموزش

۷.۳. حملات احتمالی و دفاع

۱. حملات Evasion:
• روش: تغییر ورودی برای فریب مدل
• دفاع: adversarial training

۲. حملات Poisoning:
• روش: تزریق داده بد برای تخریب مدل
• دفاع: outlier detection

۳. حملات Extraction:
• روش: استخراج مدل با های زیادquery
• دفاع: rate limiting، query auditing

فصل ۸: کنترل‌های ایمنی

۸.۱. کنترل‌های قبل از استقرار

• Review توسط کمیته اخلاق
• تست‌های جامع
• تأیید stakeholders

۸.۲. کنترل‌های حین اجرا

• Human-in-the-loop برای تصمیمات مهم
• امکان override توسط کاربر
• Thresholdهای متغیر بر اساس ریسک

۸.۳. کنترل‌های نظارتی

• لاگ تمام تصمیمات
• امکان audit
• گزارش‌های منظم

فصل ۹: برنامه بهبود

بهبودهای کوتاه‌مدت (۳ ماه) .۹.۱

- افزایش دقت مدل CR ۸۵ به٪
- کاهش خطای مدل CTP ۲۰ به٪
- بهبود recall مدل BD ۸۰ به٪

بهبودهای میان‌مدت (۶ ماه) .۹.۲

- افزودن مدل‌های جدید
- بهبود تفسیرپذیری
- کاهش bias

بهبودهای بلندمدت (۱۲ ماه) .۹.۳

- مدل‌های multimodal
- یادگیری reinforcement
- پیشرفته personalization

فصل ۱۰: نتیجه‌گیری

مدل‌های ProDecks با استانداردهای بالای اخلاقی، امنیتی و عملکردی توسعه یافته‌اند. در حالی که عملکرد خوبی دارند، محدودیت‌های آنها به‌طور شفاف بیان شده و برنامه‌ای برای بهبود مستمر وجود دارد. کنترل‌های ایمنی و تفسیرپذیری تضمین می‌کنند که مدل‌ها به صورت مسئولانه‌ای مورد استفاده قرار گیرند.

ضمیمه‌ها

ضمیمه ۱: نتایج کامل ارزیابی

ضمیمه ۲: کد آموزش مدل‌ها

ضمیمه ۳: گزارش‌های تفسیرپذیری

ضمیمه ۴: نتایج تست‌های امنیتی

تهیه‌کنندگان:

تیم یادگیری ماشین ProDecks

حامد کوهی (سرپرست ML)

مشاوران اخلاق AI