

گزارش پروژه پیش‌بینی ریزش مشتریان

بهمن ۱۴۰۴

۱ خلاصه پروژه

این پروژه با هدف پیش‌بینی ریزش مشتریان در شرکت مخابراتی با استفاده از مجموعه داده Churn Customer Telco انجام شده است. هدف اصلی ساخت مدل طبقه‌بندی دودویی با تمرکز بر افزایش معیار Recall برای شناسایی مشتریان در معرض خطر ترک سرویس است.

۲ مراحل پیاده‌سازی

۱.۲ تحلیل اکتشافی داده (EDA)

- بارگذاری مجموعه داده با ۷۰۴۳ رکورد و ۲۱ ویژگی
- بررسی مقادیر گمشده و داده‌های تکراری
- تحلیل توزیع متغیر هدف: (Churn) عدم تعادل کلاس‌ها با نسبت ۷۳٪ عدم ریزش و ۲۷٪ ریزش
- تحلیل تک‌متغیره با هیستوگرام برای متغیرهای عددی و نمودار میله‌ای برای متغیرهای دسته‌ای
- تحلیل دومتغیره با ماتریس همبستگی و نمودارهای جعبه‌ای

۲.۲ پیش‌پردازش داده

- حذف ستون customerID (غیر مفید برای پیش‌بینی)
- رفع مشکل نوع داده در ستون TotalCharges و تبدیل به عددی
- جایگذاری مقادیر گمشده: مشتریان با tenure=۰ با مقدار صفر و بقیه با میانگین
- کدگذاری متغیرهای دودویی با Encoding Label

- کدگذاری متغیرهای چندکلاسه با Encoding One-Hot
- مقیاس‌بندی ویژگی‌های عددی با StandardScaler

۳.۲ مهندسی و انتخاب ویژگی

ویژگی‌های جدید ایجاد شده:

- AvgChargePerTenureMonth: میانگین هزینه ماهانه به ازای هر ماه وفاداری
- TenureGroup: دسته‌بندی مشتریان بر اساس مدت وفاداری (جدید، متوسط، قدیمی، بلندمدت)
- ServiceCount: تعداد کل سرویس‌های مشترک شده
- ChargeRatio: نسبت هزینه ماهانه به کل هزینه‌ها

روش‌های انتخاب ویژگی:

- روش فیلتر: F-test ANOVA برای ۱۵ ویژگی برتر
- روش مبتنی بر مدل: Importance Feature Forest Random
- ترکیب: اجتماع ۱۰ ویژگی برتر از هر دو روش

۴.۲ مدل‌سازی پیشرفته

مدیریت عدم تعادل داده:

- استفاده از SMOTE برای متعادل‌سازی کلاس‌ها در داده آموزش
- افزایش تعداد نمونه‌های کلاس اقلیت به تعداد کلاس اکثریت

مدل‌های پایه آموزش داده شده:

- Regression Logistic
- (KNN) Neighbors K-Nearest
- Forest Random
- Boosting Gradient
- XGBoost

بهینه‌سازی:

- تنظیم فرایارامترها برای Forest Random با RandomizedSearchCV
- تنظیم فرایارامترها برای XGBoost با RandomizedSearchCV
- معیار ارزیابی: Recall (برای حداکثرسازی شناسایی مشتریان در معرض ریزش)

روش گروهی:

- Classifier Voting Soft ترکیب سه مدل برتر: Forest Random، XGBoost (Tuned)، (Tuned) Regression Logistic

اعتبارسنجی:

- Validation Cross K-Fold Stratified با ۵ قسمت برای ارزیابی قابلیت تعمیم مدل‌ها

۵.۲ ارزیابی و نتایج

مدل برنده بر اساس Recall انتخاب شده است. معیارهای عملکردی شامل:

- Accuracy: دقت کلی مدل
 - Precision: دقت در پیش‌بینی ریزش
 - Recall: نرخ شناسایی صحیح مشتریان ریزش‌کننده (معیار اصلی)
 - Score: ۱F - میانگین هماهنگ Precision و Recall
 - ROC-AUC: سطح زیر منحنی ROC
- منحنی ROC برای مقایسه مدل‌ها و ماتریس درهم‌ریختگی برای تحلیل اشتباهات مدل ترسیم شده است.

۳ توصیه‌های کسب‌وکاری

۱.۳ استقرار مدل

- استفاده از مدل برنده در محیط تولید برای پیش‌بینی ریزش مشتریان
- نمره‌دهی به تمام مشتریان فعلی و ایجاد بخش "خطر بالا" (احتمال ریزش بالای ۷۰٪)
- پیاده‌سازی سیستم هشدار زودهنگام ۳۰-۶۰ روز قبل از ریزش پیش‌بینی شده

۲.۳ بینش‌های کلیدی از تحلیل ویژگی‌ها

۱. نوع قرارداد و مدت وفاداری:

- مشتریان با قرارداد ماه‌به‌ماه خطر ریزش بالاتری دارند
- اقدام: تشویق به قراردادهای بلندمدت با تخفیف ۱۰-۱۵ درصدی
- تمرکز تلاش‌های حفظ مشتری بر افراد با کمتر از ۱۲ ماه وفاداری

۲. هزینه‌های ماهانه و کل:

- هزینه‌های ماهانه بالاتر با افزایش ریزش همبستگی دارد
- اقدام: بازنگری استراتژی قیمت‌گذاری برای مشتریان پرهزینه
- ارائه تخفیف‌های وفاداری یا سرویس‌های بسته‌ای

۳. روش‌های پرداخت و صورتحساب الکترونیکی:

- برخی روش‌های پرداخت نرخ ریزش بالاتری نشان می‌دهند
- اقدام: تبلیغ روش‌های پرداخت پایدار (مانند انتقال خودکار بانکی)
- ساده‌سازی فرایند صورتحساب و شفافیت در صورتحساب

۴. اینترنت و سرویس‌های جانبی:

- مشتریان با optic fiber بدون سرویس‌های اضافی بیشتر ریزش می‌کنند
- اقدام: بسته‌بندی سرویس‌های اضافی (امنیت آنلاین، پشتیبانی فنی، پشتیبان‌گیری)
- ایجاد پکیج‌های جذاب برای استفاده جامع از سرویس‌ها

۳.۳ استراتژی حفظ مشتری

اقدامات فوری (۰-۳ ماه):

۱. نمره‌دهی به تمام مشتریان فعلی با مدل استقرار یافته
۲. ایجاد بخش "خطر بالا" بر اساس احتمال پیش‌بینی شده
۳. راه‌اندازی کمپین حفظ مشتری هدفمند با پیشنهادهای شخصی‌سازی شده

استراتژی میان‌مدت (۳-۶ ماه):

۱. پیاده‌سازی پیش‌بینی ریزش در زمان واقعی در سیستم CRM

۲. آموزش تیم خدمات مشتری در تاکتیک‌های حفظ مشتریان در معرض خطر

۳. تست A/B پیشنهادهای مختلف حفظ مشتری

استراتژی بلندمدت (۶-۱۲ ماه):

۱. توسعه سیستم هشدار زودهنگام پیش‌بینی‌کننده

۲. ایجاد نقشه‌های سفر مشتری شخصی‌سازی شده

۳. بازآموزی مستمر مدل با داده‌های جدید

۴.۳ تأثیر مالی مورد انتظار

با فرض ۱۰۰۰ مشتری در معرض خطر و موفقیت ۳۰ درصدی در حفظ آنها:

- درآمد بالقوه نجات یافته: با توجه به ارزش طول عمر مشتری
- هزینه کمپین: تخمین ۵۰٪ به ازای هر مشتری
- سود خالص: مثبت و قابل توجه (ROI بالای ۱۰۰٪)
- هزینه جذب مشتری جدید ۵ تا ۲۵ برابر هزینه حفظ مشتری فعلی است

۵.۳ محدودیت‌ها و ملاحظات

- نرخ مثبت کاذب: برخی پیش‌بینی‌های ریزش ممکن است اشتباه باشند، اما پیشنهادهای ملایم به مشتریان وفادار آسیبی نمی‌رساند
- نرخ منفی کاذب: درصدی از ریزش‌های واقعی ممکن است از دست بروند، لذا برنامه‌های حفظ پایه باید ادامه یابد
- بازآموزی مدل: بازآموزی ماهانه با داده‌های جدید برای انطباق با تغییرات رفتار مشتری
- عوامل خارجی: در نظر گرفتن روندهای فصلی، اقدامات رقبا، و تغییرات بازار

۴ بهبود مستمر

- نظارت بر عملکرد مدل: پیگیری ریزش واقعی در مقابل پیش‌بینی شده به صورت ماهانه
- حلقه بازخورد: ترکیب نتایج کمپین‌های حفظ برای بهبود پیش‌بینی‌ها
- انصاف: اطمینان از عدم تبعیض مدل علیه هیچ بخشی از مشتریان

۵ مشارکت اعضای تیم

نام عضو تیم	وظایف و مسئولیت‌ها
مهید خالتي	پیا‌ده‌سازی فاز ۴ و ۵
باران حسینی	نوشتن گزارش و پیا‌ده‌سازی فاز ۳
بنیامین کریمی‌زاده	پیا‌ده‌سازی فاز ۱ و ۲