9th Annual Conference on
Electronic Banking & Payment Systems

پژوهشکده پولی و بانکی
بانک مرکزی جمهوری اسلامی ایران

بانک مرکزی جمهوری اسلامی ایران

شرکت ملی انفورماتیک

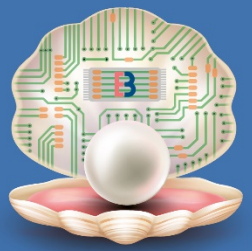
نهمین همایش سالانه
بانکداری الکترونیک
و نظام‌های پرداخت

(ارزش‌آفرینی دیجیتال)

تهران، مرکز همایش‌های بین‌المللی برج میلاد - ۱ و ۲ اسفند ۱۴۰۱

# نقش استنباط آماری در کشف تقلب در تراکنش‌های کارتی

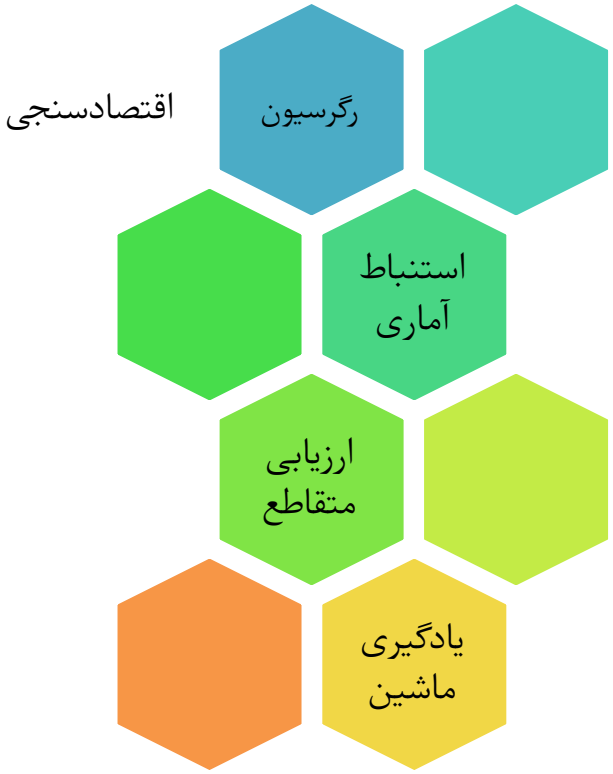
رامین مجاب
پژوهشکده پولی و بانکی



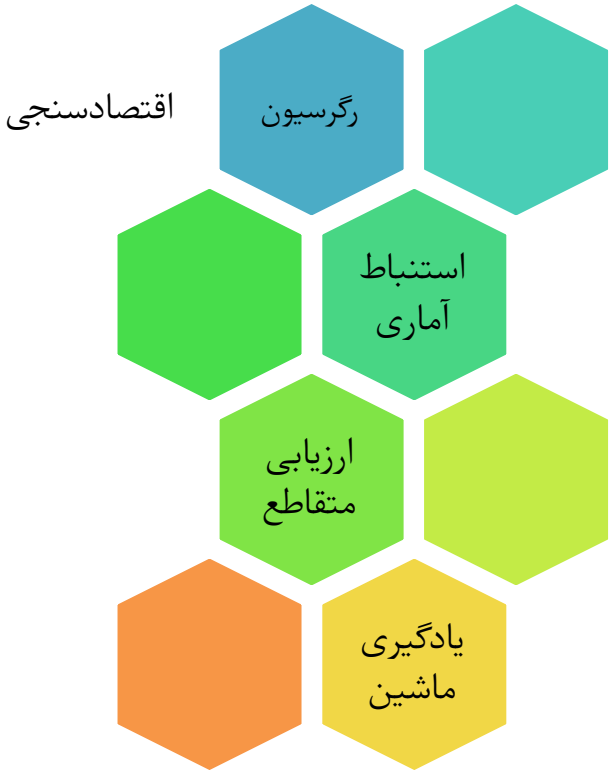
# مقدمه

- آمار کاربردی (رگرسیون)
  - تکیه بر استنباط آماری
  - نقش کمرنگ ارزیابی متقاطع (مثلاً در کتاب‌های آموزشی)
- یادگیری ماشین
  - نقش حیاتی ارزیابی متقاطع
  - بیش‌برآورد (مرحلهٔ آموزش)
  - مشکلی به نامِ حجم مشاهده عملاً وجود ندارد.
- آیا حجم بالای مشاهدات ما را از استنباط آماری بی‌نیاز می‌کند؟
  - استنباط آماری فقط مربوط به بررسی معناداری تخمین‌زننده نیست.
  - مثلاً آیا تفاوت زیر معنادار است؟

$$AUC_1 = 0.8 \text{ و } AUC_2 = 0.81$$



# مقدمه (ادامه)




- هدف: تأکید بر اهمیت استنباط آماری در کاربردهای مرتبط با الگوریتم‌های یادگیری ماشین و کارکردهای آن
  - کارکرد؟ مثلاً تغییر نوع نگاه اصلاحاتی را در محاسبات منجر می‌شود.
- مقایسهٔ قدرت کشف تقلب مدل‌های انتخاب گسسته و الگوریتم LightGBM با استفاده از داده‌های Vesta (2018)
  - این داده‌ها در قالب انجام مسابقه‌ای برای یافتن بهترین مدل پیش‌بینی‌کنندهٔ تقلب در سایت Kaggle منتشر شده است.
  - نقش الگوریتم LightGBM در بهترین مدل‌های پیش‌بینی‌کنندهٔ تقلب بسیار پررنگ است.

# چرا داده‌های پرداخت و بحث تقلب؟

- از منظر هدف
  - حجم بالای مشاهدات

- ارزش کل کلاهبرداری تبادلات کارتی در سال ۲۰۱۹ به ۱/۰۳ میلیارد یورو می‌رسد. این ۰/۰۳۲ درصد ارزش کل تراکنش‌هاست.
  - بانک مرکزی اروپا ، ۲۰۲۱
  - ارزش کل تراکنش‌های شاپرک در سال ۱۴۰۰ برابر با ۷۰۰۰ همت است (شاپرک، ۱۴۰۱)
  - تعمیم: ۲/۲۴ همت
- موضوع این پژوهش بخش کوچکی از اقدامات مرتبط با کشف تقلب است.
  - بحث آموزش و ایجاد انگیزه برای فروشندگان و صاحبان کارت در حفاظت بهتر اطلاعات
  - استفاده از تکنولوژی‌های به روز مثلاً در صدور کارت و غیره
  - وضع قوانین برای افزایش هزینهٔ دسترسی به تکنولوژی‌های تولید کارت‌های تقلبی،
  - **استفاده از الگوریتم‌های کامپیوتری**

# استفاده از الگوریتم‌های کامپیوتری

- طبقه‌بندی الگوریتم‌ها ساده نیست.
  - در برخی صرفاً از اطلاعات تراکنش استفاده می‌شود.
    - الگوریتم‌های طبقه‌بندی بانظارت  نقش محوری دارند
  - در برخی داده‌های تاریخی از رفتار مشتریان مورد استفاده قرار می‌گیرد
    - نقش الگوریتم‌های بدون نظارت نیز پررنگ‌تر می‌شود
  - در برخی با توجه به حجم بسیار زیاد درخواست‌ها، مباحثی در خصوص زمان-حقیقی بودن یا پردازش موازی مطرح می‌شود.
  - اهداف عملیاتی همچنین باعث می‌شود که نوع نگاه به کشف تقلب سیستماتیک‌تر و مقیاس‌پذیرتر باشد.
- کشف تقلب در تراکنش‌های مالی معمولاً در قالب یک مدل طبقه‌بندی دودویی صورت‌بندی می‌شود.
- مطالعات تجربی متنوع هستند. از منظر،
  - نوع داده‌های مورداستفاده
  - مهندسی ویژگی
  - نوع الگوریتم
  - نوع ارزیابی در فرایند مقایسه (contribution)

# روش‌شناسی

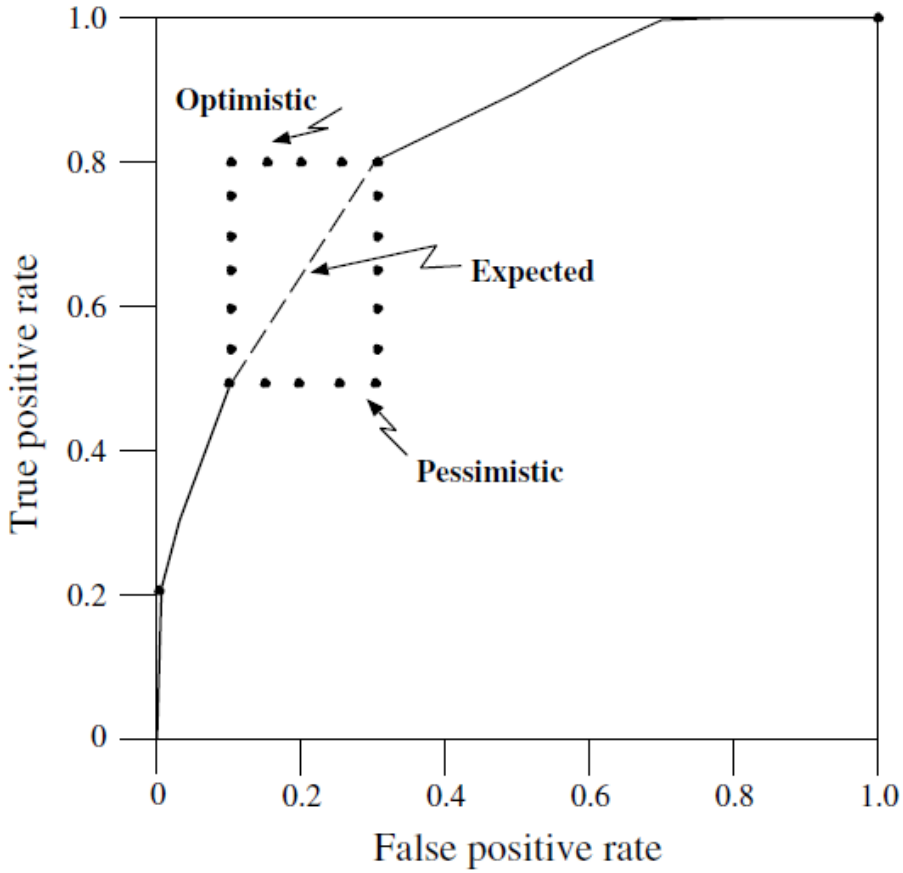



Fig. 6. The optimistic, pessimistic and expected ROC segments resulting from a sequence of 10 equally scored instances.

- بخشی از روش‌شناسی به تخمین رگرسیون انتخاب گسسته و مبانی نظری الگوریتم LightGBM مربوط می‌شود.
- Fawcett (2006)
- Muschelli (2020) انتخاب حالت «موردانتظار» در کارهای عملی را گمراه‌کننده معرفی می‌کند و توصیه به انتخاب حالت بدبینانه دارد.
- **پیشنهاد: مقایسه با لحاظ معناداری آماری یا اقتصادی صورت گیرد.**

# داده‌ها

- محدود کردن مطالعه به نمونهٔ یادگیری به این دلیل است که متغیر دودویی هدف که نشان‌دهنده تقلب بودن یا نبودن اطلاعات است با نام isFraud گزارش شده است و نمونهٔ تست فاقد این ستون است.
- برگزارکننده مسابقه عنوان کرده است که برچسب‌های ستون isFraud بر اساس وجود تراکنش اصلاحی بر کارت انجام می‌شود.
  - اگر این برچسب بر یک حساب کاربری زده شود، بر این تراکنش و بقیهٔ تراکنش‌های پیشینی که ایمیل یا آدرس مشترک دارند نیز چنین برچسبی زده می‌شود.
- در این پژوهش متغیرهای طبقه‌ای به متغیر موهومی تبدیل شده‌اند.
- وجود NA؟
  - می‌توان مدل‌های مختلف تخمین زد و نتایج را ترکیب کرد.
  - با توجه به هدف پژوهش، دو استراتژی در حذف NA دنبال می‌شود:
    - در استراتژی اول حجم مشاهدات به‌شرطی حداکثر می‌شود که حذف تعداد مشاهدات تا حد ممکن زیاد باشد. این انتخاب به ۵۹۰۵۳۹ مشاهده و ۳۷ ویژگی می‌انجامد.
    - در استراتژی دوم، حجم مشاهدات به شرطی حداکثر می‌شود که تعداد بیشتر ویژگی‌ها در اولویت باشد. این انتخاب به ۴۴۵۴۰ مشاهده و ۲۱۵ ویژگی می‌انجامد.

# نتایج

آماره‌های توصیفی معیار AUC تحت فروض مختلف و در سناریوی اول حذف NA

| نسبت یادگیری: | | | ۶۰ درصد | | ۸۰ درصد | |
|---|---|---|---|---|---|---|
| مدل: | | | DCR | LightGBM | DCR | LightGBM |
| آستانهٔ معناداری و رویکرد محاسبه: | مورد | $10^{-16}$ | 0.80 (0.03) | 0.83 (0.05) | 0.79 (0.02) | 0.83 (0.04) |
| | انتظار | $10^{-4}$ | 0.80 (0.02) | 0.81 (0.04) | 0.78 (0.02) | 0.82 (0.03) |
| | بدبینانه | $10^{-16}$ | 0.78 (0.03) | 0.80 (0.05) | 0.78 (0.02) | 0.81 (0.04) |
| | | $10^{-4}$ | 0.69 (0.02) | 0.75 (0.04) | 0.69 (0.01) | 0.74 (0.03) |

آماره‌های توصیفی معیار AUC تحت فروض مختلف و در سناریوی دوم حذف NA

| نسبت یادگیری: | | | ۶۰ درصد | | ۸۰ درصد | |
|---|---|---|---|---|---|---|
| مدل: | | | DCR | LightGBM | DCR | LightGBM |
| آستانهٔ معناداری و رویکرد محاسبه: | مورد | $10^{-16}$ | 0.84 (0.02) | 0.89 (0.05) | 0.85 (0.02) | 0.88 (0.03) |
| | انتظار | $10^{-4}$ | 0.83 (0.01) | 0.87 (0.03) | 0.84 (0.01) | 0.87 (0.03) |
| | بدبینانه | $10^{-16}$ | 0.80 (0.02) | 0.81 (0.05) | 0.81 (0.02) | 0.82 (0.04) |
| | | $10^{-4}$ | 0.76 (0.01) | 0.80 (0.03) | 0.76 (0.01) | 0.79 (0.03) |

# بحث و نتیجه‌گیری

- رویکرد عملیاتی باعث کمرنگ بودن نقش استنباط آماری می‌شود.
  - این رویکرد بیشتر توسط پژوهش‌ها در حوزهٔ یادگیری ماشین دنبال می‌شود.
  - سهم بالای نمونه دردسترس عامل مهمی در پیگیری این رویکرد است.
  - تکیه اصلی بر عملکرد خارج از نمونهٔ مدل‌هاست.
    - از منظر نظری و در صورت محدود کردن مطالعه به مدل‌های خطی، نتایجی در رابطه با معادل بودن برخی از انواع اعتبارسنجی متقابل و معیارهای انتخاب مدل نظیر AIC بدست می‌آید (مثلاً از جمله مطالعات اولیه Stone (1976) و Shao (1993, 1997))
  - استنباط آماری نتایج انتخاب نوع الگوریتم را تحت تأثیر قرار می‌دهد.

با تشکر