

KDD2020

背景：在海关报关的时候，如果人为操作商品的价格，那么就会使得收到的关税减少，同时商品的价格不一，这种欺诈行为带来的影响也是不一样的，如果对所有商品一一检查需要耗费大量的时间，会影响通关的效率和干扰正常的用户，因此如果能够利用机器学习模型同时预测出欺诈的概率和能额外增加税收的金额，配合相关的策略(control less but better,抓大头)就可以最大程度的减少关税损失和保证用户的通关体验。

本文：提出了一个双任务注意树感知嵌入模型，用于对非法贸易流进行分类和排序，这些非法贸易流在被抓获时对海关总收入贡献最大。DATE的优势来自于将基于树的可解释性和事务级嵌入模型与双重注意机制相结合。为了准确识别非法交易并预测税收，DATE同时从每笔交易的非法性和附加税中学习。

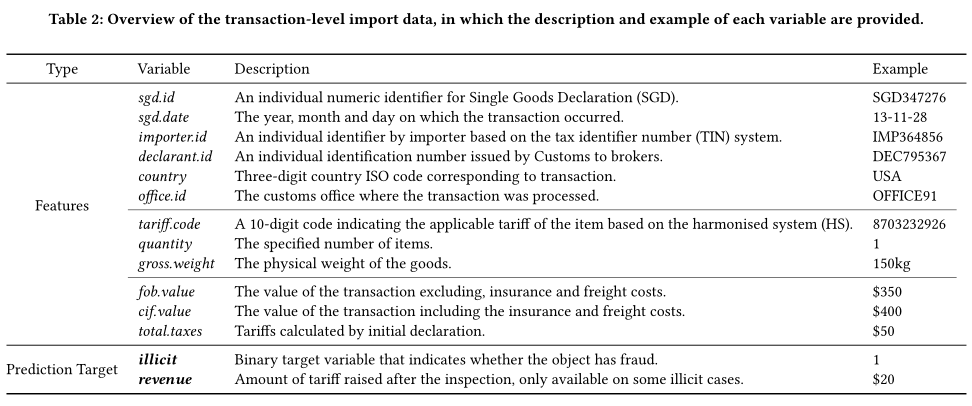
考虑到一项进口交易，以及相应的贸易商和贸易商品，我们的主要目标是预测其非法性，一个二元标签指示其是否非法。由于识别贸易欺诈将导致海关税收增加，我们还将预测增加的税收作为次要目标。

主要贡献：

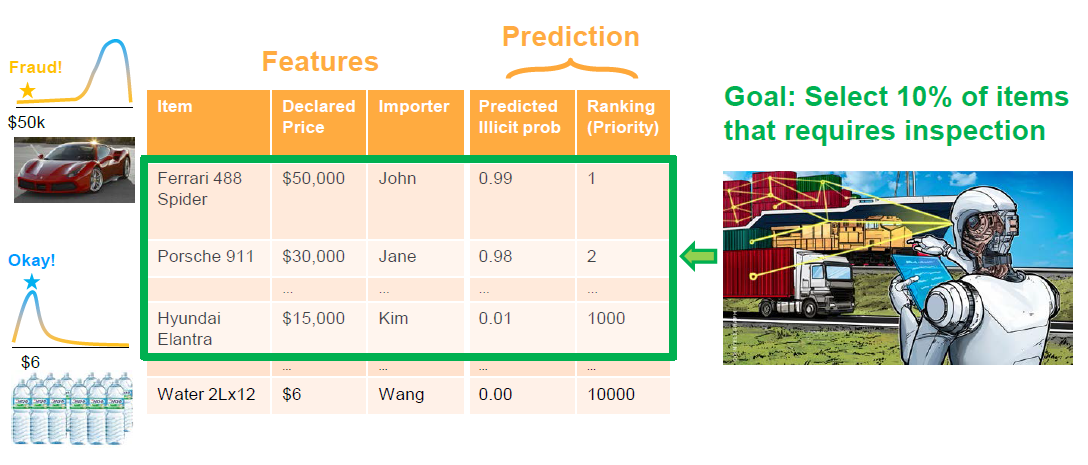
1. 我们预先训练一个基于树的模型来识别原始特征的最重要的组合，称为交叉特征。交叉特征不仅提供了处理结构化数据的有效表示能力，而且使我们的模型具有可解释性。
2. 通过学习交叉特征的嵌入，我们设计了一个双重注意机制来生成事务的表示。我们利用多头自我注意来学习交叉特征之间的相互作用，并利用注意网络编码交易者和交易商品（HS代码）如何与交叉特征相关。
3. 设计了一种双任务学习技术，该技术可以预测非法概率，并为海关当局联合最大化增加的海关税金额。双任务学习使分类任务更加有效，并帮助海关识别最有价值的非法交易。

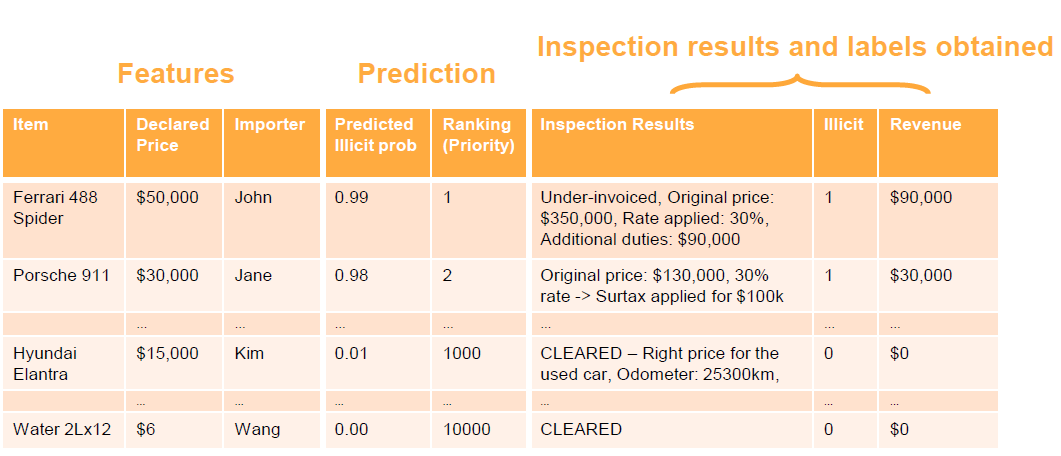
数据集：

本文采用了世界海关组织伙伴国尼日利亚的交易级进口数据。2013年至2017年，共有1932151项进口贸易流量构成了该数据。下表列出了关键数据字段：

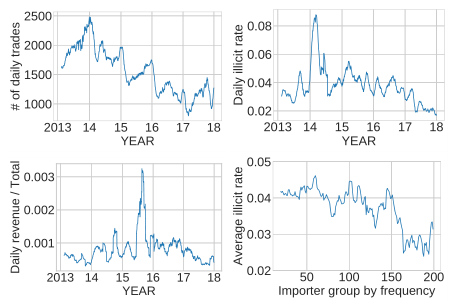


一些字段（如交易员ID）已被匿名化。贸易货物按关税分类。代码结合了全球使用的六位数HS代码和国家授予的四位数代码。底部两行（1. 是否非法 2. 检查后提高的关税金额，仅适用于某些非法案件）是要预测的目标变量，在检查结果之后生成。





数据观察：



如图所示，描述了我们使用的数据集的每日交易量、每日非法率和每日收入。2013-2017年有超过1.9m的输入货物。值得注意的是，根据尼日利亚海关规定，每个海关都要接受详细检查（即，目前检查率接近100%）。因此，除了走私案件外，非法和合法交易都会在完整的日志中准确标注。注意：我们的数据集只包括单个货物的日志记录，不包括多货物声明。因此，交易趋势不能代表该国的总进口量。总体非法率为3.83%，但每年略有变化。

侦查非法交易的每日税收收入全年也各不相同。根据数据，如右下角图所示，频繁进口商的欺诈行为比交易较少的进口商要少。同样，我们观察到某些HS代码在非法交易中使用得更频繁。基于这些观察，我们随后将进口商ID和HS代码视为欺诈检测中关键信号的一部分。

模型量化：

我们量化了模型中进口商、报关员、HS代码和原产国的风险指标。根据其不合规记录计算风险指标。例如，进口商的排名是根据欺诈性进口的具体数量除以其相应的总交易量。排名在第90百分位以上的进口商被视为高风险进口商，其风险指标的值为1；否则，0。

同时，我们增加了一些非线性关系，如unit.value（=到岸价格/数量）、value/kg（=到岸价格/毛重）、tax.ratio（=总税费/到岸价格），unit.tax（=总税款/数量）以及face.ratio（=离岸价/到岸价）。

最后，我们还添加了三个时间特征：一年中的每天，一年中的每周，一年中的每月进口的商品。

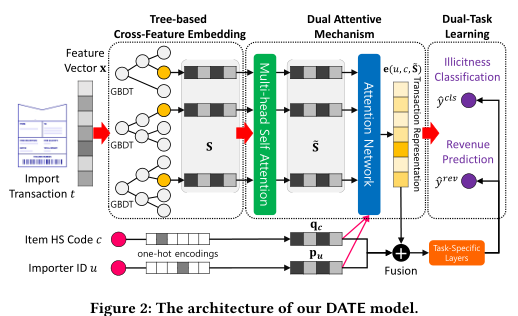
形式化定义：

考虑到进口贸易流量t，包括进口商u和货物的HS码c，目标是基于t预测欺诈分数和增加的收入。

基于cls和rev，海关可以根据他们的标准选择欺诈交易。有报告指出，由于天文数字的贸易量，发达经济体的海关检查率不超过5%。鉴于这一局限性，设计出用最少的检查来检测欺诈的算法至关重要。

模型：

双任务注意树感知嵌入（DATE）模型由三个阶段组成。第一阶段对基于树的分类器进行预训练，以生成每个事务的交叉特征。第二阶段是双注意机制，学习跨特征之间的交互以及进口商、HS代码和跨特征之间的交互。第三阶段是通过联合优化违法性分类和收入预测的双任务学习。总体架构如图2所示。



（一）基于树的分类器进行预训练

使用GBDT作为学习方法。GBDT是一种集成学习方法，通过顺序增长决策树来提高预测能力。GBDT的预测是通过汇总树的结果来完成的：



r是树的大小，是每个树的结果。是0-1之间的数，用于权重的控制。本文参考以下预训练的GBDT来生成交易的特征向量：

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al.2014. Practical lessons from predicting clicks on ads at facebook. InADKDD. 1–9.

Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. TEM: Tree-enhanced embedding model for explainable recommendation. InWWW. 1543–1552.

树有多个（设为L个），每个树又有不同的叶子结点，每个树的其中一个叶子节点是激活的（正确分类），其余不激活。因此，我们将他们串接在一起，生成词向量（有L个元素为1，其余为0）。将该方法称为Tree‐based **Cross‐Feature** Embedding。

为了获取高阶语义，映射cross-feature到可学习的embedding vector（d维向量）。具体的，将每个p都增加一个来完成这一转换（r是树的数量），如下：



是移除矩阵中为0的元素的函数（相当于稀疏矩阵到密集矩阵的转换）。

学习密集嵌入的好处：

1. 可以对不同叶片之间的潜在相关性进行建模，并将相似的交叉特征映射到嵌入空间中的邻近点。
2. 由于嵌入矩阵在培训过程中是可学习的，因此它允许我们加入额外的信息，例如进口商和商品id。换句话说，采用可学习向量而不是静态向量为我们的模型提供了一定的定制灵活性，因为不同国家的海关记录了各种信息。

（二）双重注意机制

我们提出了一种双重注意机制，基于叶子的自注意力和注意网络。前者是对来自不同视图的交叉特征之间的相关性进行建模。后者是为了了解每个交叉特征如何构成进口商和HS代码。然后我们融合嵌入以获得每个事务的表示。

将FQ、FK、FV从原来的向量拓展到矩阵，可以得到点积版的注意力机制。



注意到，因为是矩阵，我们映射了n\_h次不同的线性映射，最终输出d\_v维的向量。

再进一步的，为了实现对叶（即交叉特征）嵌入之间交互的不同方面建模的目标，我们利用了多头（MH）自我关注：

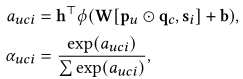


每一个head都是一个SDPA。

本文的基于叶子的自注意力，即Q=K=V的MH。

前文是学习来自不同树的交叉特征的贡献权重，以利用交叉特征的交互作用进行智能预测。接下来的注意网络，指提出了一个考虑进口商ID和项目标识符（即HS代码）的注意网络，以模拟给定交易中交叉特征、进口商和项目之间的三维交互。给一对(u,c)，u是进口商和c是HS代码，与交叉特征嵌入S一起，目标是生成注意权重a\_ucl。注意权重可以反映哪种交叉特征F\_i在确定基于u和c，特定进口商的那些非法行为方面更为重要。

这里的网络是传统版本，不是SDPA。



对不可见的u和c初始化为-。使用点乘获得交叉关系。。生成向量。S是上文self-attn结果。

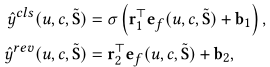
最后，由于交易是否非法在很大程度上取决于进口商是谁以及里面的物品是什么，因此我们结合e(u,c,S)与进口商合作和项目用于预测（one-hot）。通过连接这三个向量以及一个隐藏层，我们可以生成一个融合的向量库，由以下公式给出：



W是学习参数，。

（三）双任务学习

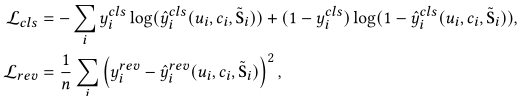
识别非法交易自然会增加海关税收。然而，只有当交易被发现非法时，才能确定增加的税收数额。这表明，如果我们能够准确估计增加的税收数额，就有可能有助于预测非法分类。多任务学习技术已用于训练同时优化多个目标的模型。我们提出了一种双任务学习方法来使用事务信息用于二进制非法分类和增加收入预测两项任务。给定两个不同的事务, 我们有特定于任务的层：



r1,r2是隐藏层，是sigmoid函数，cls和rev指预测欺诈分数和增加的收入。

最后，损失函数：





实验：

根据时间将数据分为训练集、验证集和测试集。去年（2017年）的数据用作测试集。过去四年（2013-2016年）的最后一个月数据作为验证集进行了验证。相应地，训练、验证和测试集的规模为1635157（84.4%）、25948（1.3%）、276440（14.3%），相应的非法比例为3.94%, 2. 51%和2.24%。

评价指标：

Precision@n%：该指标解释了在前n%的交易中有多少交易是非法的。

Recall@n%：该指标表示已成功从总非法交易量中筛选出的已检查交易量（即所选流量的n%）。

Revenue@n%：它是由模型确定的前n%交易的总收入除以所有交易的总收入。该指标解释了与检查整个交易产生的收入相比，前n%的交易可以产生多少关税。（相当于前n名次的占比）

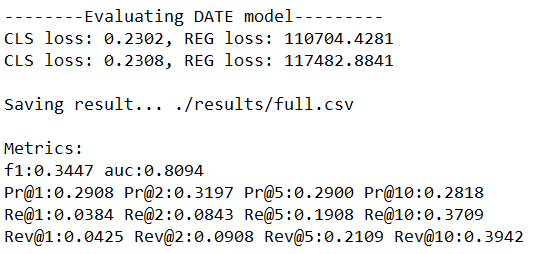
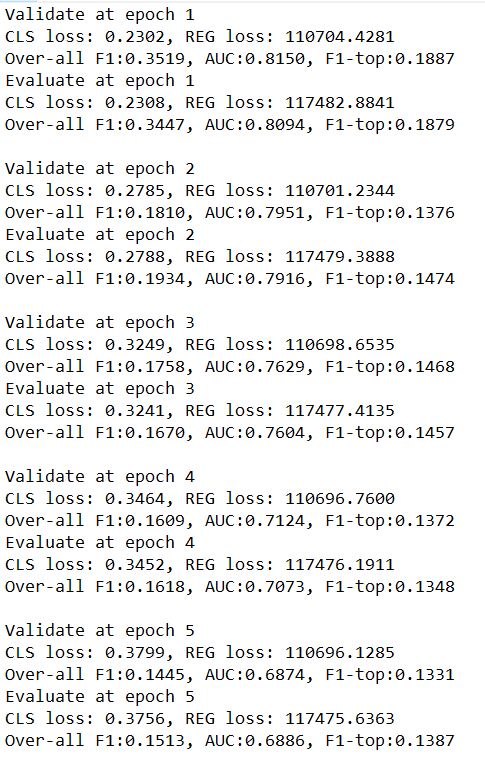
AUC：范围为0到1。它用于评估整个数据的预测算法的结果。

F1分数：这是六种不同检查率之间测试准确性的衡量标准。调整非法阈值0.1-0.6(step=0.1)，找到最佳F1。

超参数：

树的大小100，深度4。嵌入维度16，4头注意力，16维注意力层。Lr=0.005，batch=128，loss中的alpha=10，lambda=0.01，第五个epoch后验证的效果最好，使用该结果做测试。

结果（引用原来的ranger-git包）：



Baseline：

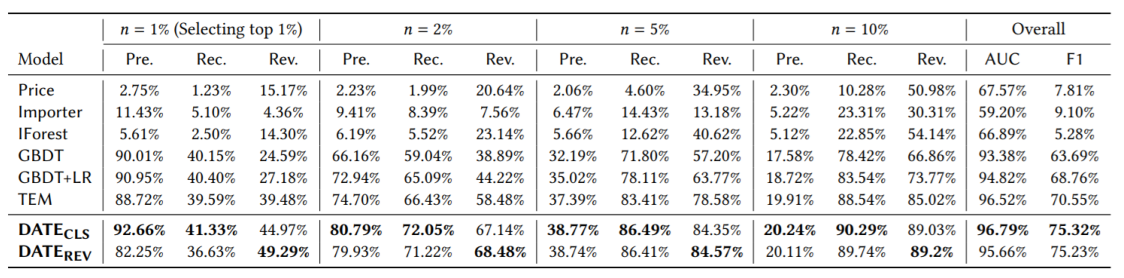
价格/进口商：按进口商申报价格/欺诈率的顺序选择

IForest：基于树的异常检测算法

GBDT：具有交叉特征的XGBoost，在二进制标签上训练

GBDT+LR：基于交叉特征的逻辑回归

TEM（Wang等人，2018）：关注的树增强模型。我们将其目标函数从贝叶斯个性化排序（BPR）改为交叉熵分类。

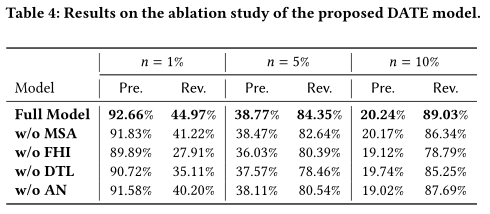


结论：

1. 基于树的模型显示出相当好的结果，并且比Price和Importer性能要好得多。
2. GBDT+LR比GBDT好，这证明了模型在为不同交叉特征赋予权重方面的灵活性。
3. TEM与其他基线方法相比，n分别为2%、5%和10%时效果比较好。因此，利用稠密向量将交叉特征嵌入到低维空间，并结合注意机制为交叉特征赋予动态权重是有效的。
4. 尽管TEM实现了很好的性能，但它无法捕获交叉特征之间的交互。此外，没有一种基线方法同时考虑分类损失和收益预测的优化。在我们的研究中，我们利用自我注意来捕捉交叉特征之间的相互作用，并通过在自我注意中连接不同的头部来进一步获得方面层面的嵌入。因此，DateCls的性能在除收入外的所有指标上都远远领先于其他基线。我们还可以确认，DateRev是保证海关最大收入的最有效方法。
5. 由于数据的严重不平衡性，强基线的AUC在90%以上，因此结果表明F1分数可以更好地区分模型的整体预测性能。对于检查率为n%的海关检测问题，即使在最先进的模型之间，收入差异也很大，这证明了我们提出的算法的有效性。

消融实验：

* DATECLS(Full Model): 完整模型.
* w/o multi-head self-attention (MSA): 没有多头注意力.
* w/o fusion with HS & Importer embeddings (FHI): 不使用HS和Importer ID作为第二阶段的补充嵌入
* w/o dual task leaning (DTL): 仅在二进制标签上训练模型后，使用预测的概率值选择事务
* w/o attention network (AN): 在不考虑动态注意权重的情况下，平等对待每个交叉特征。

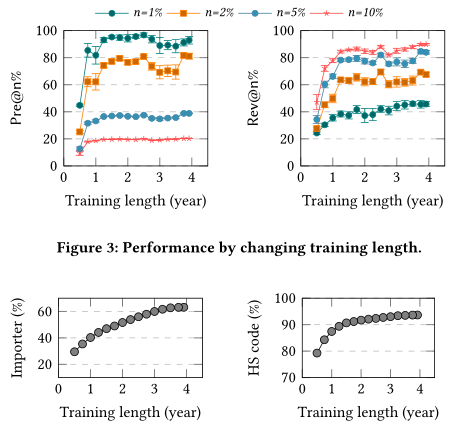


1. 在我们的模型中删除融合操作时，所有度量都会大幅下降。这表明只使用交叉功能而不使用Importer和item info会损害性能。融合步骤是绝对必要的，因为它同时学习交叉特征、Importer和item info之间的相关性，
2. Revenue@n%显然在不使用双任务学习的情况中下降。这样的结果表明，同时优化收入预测和分类任务可以相互增强这两个任务的有效性。
3. 移除注意力网络会恶化所有指标的性能。这证明了根据进口商和货物对交叉特征的关注程度为交叉特征赋予动态权重的有效性。
4. 虽然消除自我注意几乎可以维持Precision@n%，未能取得较好的业绩Revenue@n%. 结果表明，通过多头自我注意学习方面级特征有利于收益预测。

简言之，我们的DateModel的每个组件在执行这两项任务时都真正起作用。验证了设计方案的正确性。

Q：模型的性能随训练数据量的变化有多大？

我们通过改变6个月训练数据的长度来进行实验。验证集和测试集保持完整，直到大约两年的过去数据增量不大。学习数据越丰富，获得的新信息就越少。因此，性能改进逐渐减少。

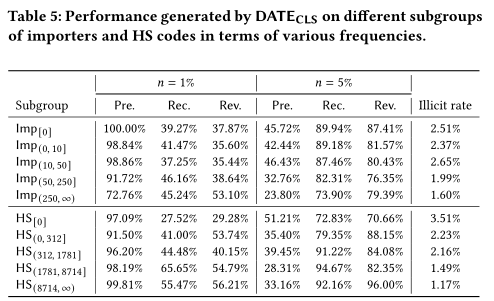


子群性能：

在海关申报的交易中，一些进口商ID和HS代码比其他更为频繁。为了研究发生频率如何影响性能，我们将测试集分为几个子组，如下所述：

Importer：我们根据Importer在训练集中出现的次数将其分为五个子组。例如，Imp[0]表示新的进口商，而Imp（0,10）表示出现但少于或等于10次的进口商组。

HS：我们以类似的方式将HS代码分为五个子组。



积极交易者（Imp>250) 与不活跃的交易者（即Imp<10和Imp[0]）相比，欺诈率最低。这可能表明进口商可能倾向于使用新ID以避免留下非法记录并欺骗海关。

HS代码的结果显示出类似的趋势。与不受欢迎的项目的交易往往被预测为非法交易。这一结果可能与HS代码的操纵有关。

研究结果还表明，DATE对不可见进口商的召回率接近90%，而对活跃进口商的召回率则较低，因为他们的非法率较低。

同时，与进口商的结果不同，DATE在常见的HS代码中导致了显著的召回率。由于世界海关组织每隔一段时间为新产品创建一次HS代码，因此DateOnHS的性能较差。

案例分析：

尽管我们的DATE在发现欺诈方面取得了显著的成绩，但我们仍然不知道DATE发现了哪些证据来预测违法性。为了证明DATE的可解释性，下表列出了非法和合法案件之间的比较及其相应的交叉特征。



我们根据最高的注意力分数选择前2个重要的交叉特征。在数据中的交易中，二手车占了最大的比例，进口商有更高的可能性报告较低的价值，以逃避额外的税收。因此，我们选择了两个二手车交易并分析它们的差异。

对于非法交易，交叉特征表明其具有较高的贸易价值（即cif.value、fob.value）和较低的毛额。报告可获得的税收的一小部分（即税率<0.18%）时的权重。此外，每公斤价值相对高于合法价值。这些变量表明二手车的交易价值较高，但海关只收取少量税款。结果恰好是贸易商品低估的一个令人信服的例子。世界海关组织海关局也对从Date获得的交易进行了审查，指出该交易的单位价值较低。世界海关组织领域专家的这一声明与Date确定的交叉特征形成了共鸣，这证明了Date在实现人类层面的可解释性方面具有一定的潜力。

与非法案件相比，合法交易显示了正常的交易信息。即使cif.value值大于$1640，则税率将在可接受的区间内下降。此外，value/kg相对较小，表明可能不存在低估问题。

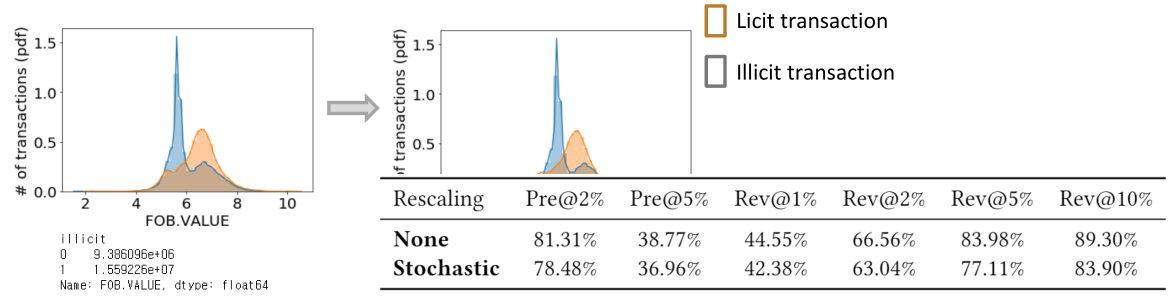
讨论：

问题：某些初始特征值在检查后进行了调整，这可能会影响性能

我们注意到平均到岸价、离岸价和总价。非法交易的税收价值比典型交易高66%。假设非法交易和合法交易的对应值没有差异。我们可以推测，两组之间的差异部分是由于信息更新。因此，我们有两种方法来回滚cif, fob和total.taxes：

1. 确定性：乘以标量值0.6。
2. 随机化：乘以高斯随机变量。

更好的实验结果：将平均价格乘以高斯随机变量𝑋=Ν(0. 6,0. 12)



采取行动的修订数据集在非法交易和合法交易之间具有相似的价值分布，这更难预测。表格显示Date在修订后的数据集上表现相对较好。这样的结果意味着Date可以从导入事务中学习各种模式，而不会因为错误而被有偏差的分布所支配。通过这个实验，我们希望数据管道固定后，Date将顺利执行，并且我们可以接收更干净的数据。

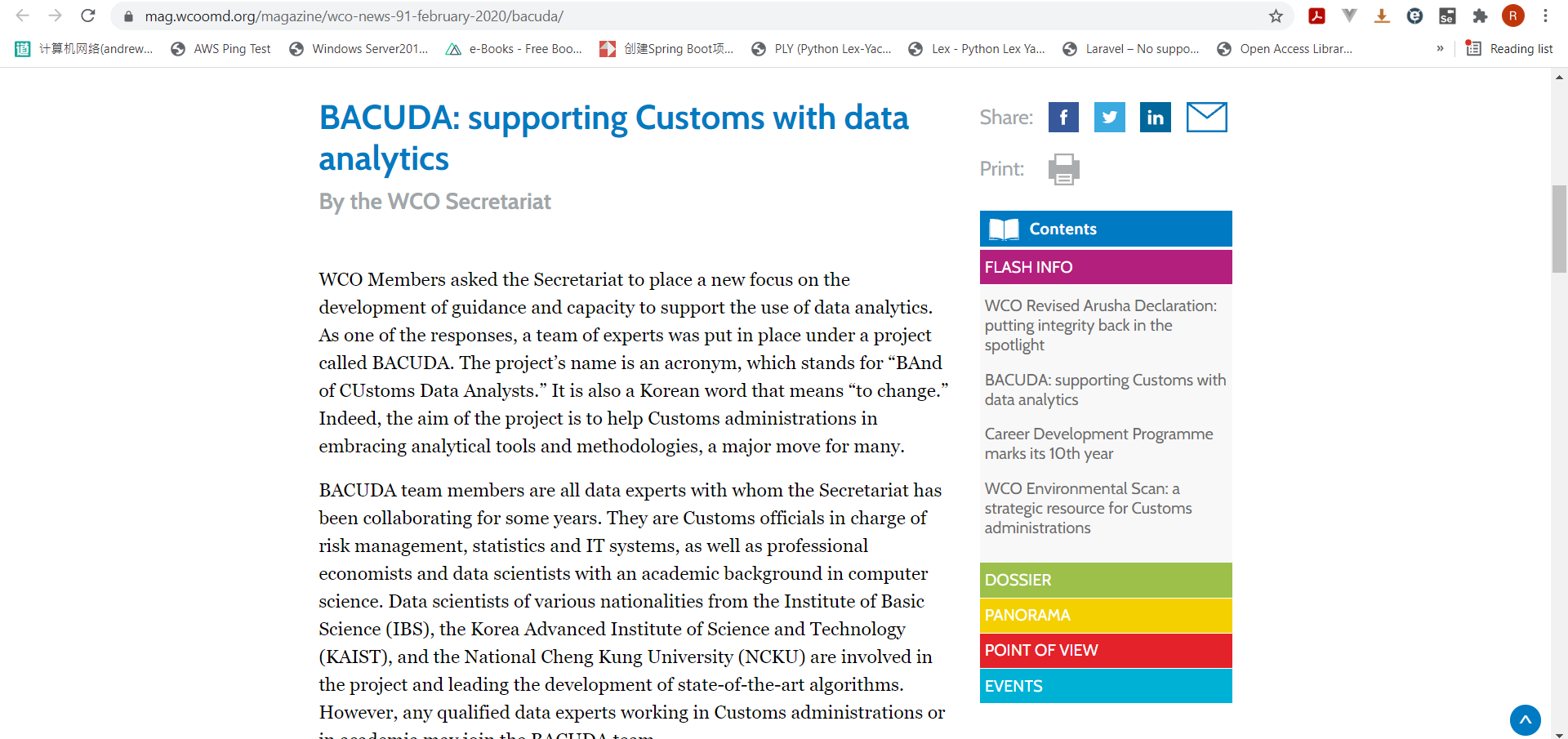
小结：

什么是DATE模型？DATE是一种海关选择模型，它按照欺诈风险的顺序对贸易流进行排序，并使海关收入最大化。

它是可解释的吗？是的，因为它的决策规则来自GBDT，权重来自注意机制。

有效吗？是的，我们确认DATE优于包括TEM在内的最先进模型，并且对输入数据中的噪声具有鲁棒性。

它将用于海关吗？是的，我们公开了代码和教程的来源，并且我们正在积极参加世界海关组织及其巴布达倡议举办的研讨会。几个国家加入了BACUDA（https://mag.wcoomd.org/magazine/wco-news-91-february-2020/bacuda/），我们正在尼日利亚海关测试我们的算法。



Tree‐based Cross‐Feature Embedding: 先利用GBDT树模型(论文中采用XGBoost)训练出一个树模型，保证模型具有一定的可解释性，将树模型的叶子节点通过要学习的Embedding层，得到各个叶子节点的Embedding向量，而叶子节点Embedding向量是通过Transformer来学习，主要是学习树模型中各叶子节点的相似性；

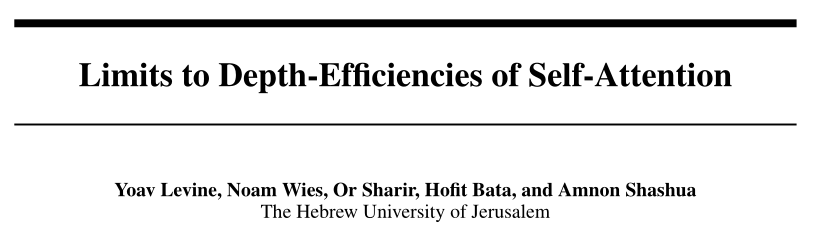
Dual Attentive Mechanism: 使用Attention机制学习商品货物进口商(Importer ID u)和进口商品(Item HS code c), 以及上述各个叶子节点Embedding向量的权重，根据通过权重值，对每条输入的交易的叶子节点Embedding向量进行加权聚合；

Dual‐Task Learning: 即多任务学习，通过将Importer，Item，以及叶子节点Embedding加权求和的向量，通过拼接(Fusion)后，送入到多任务网络中学习；

在Tree‐based Cross‐Feature Embedding部分，我们之前熟悉的架构为GBDT + LR的形式，而在此文中通过Transformer学习叶子节点的Embedding矩阵，一方面用于挖掘各叶子节点之间的相关性，另外在学习Embedding矩阵的参数过程中，可以融合其他知识，比如作者发现进口频繁的进口商存在欺诈的可能性相对较低，同时欺诈检测中进口商品的HS编码重要性也较高，因此可以通过设计Dual Attentive Mechanism来融合这些信息；

修改：

参数设置文章：



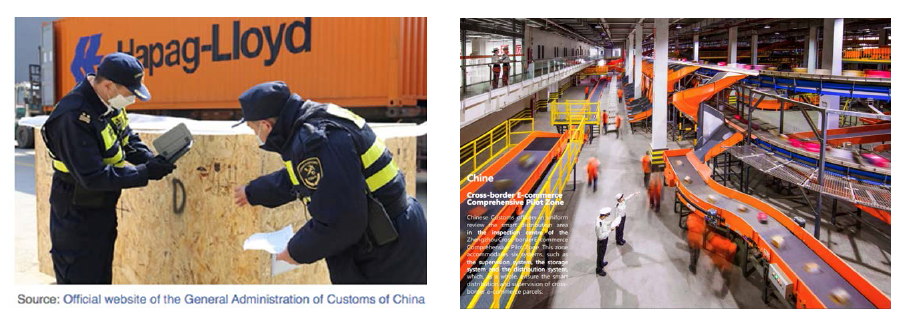
希伯来大学（The Hebrew University of Jerusalem），简称希大，是犹太民族的第一所大学

迁移学习：

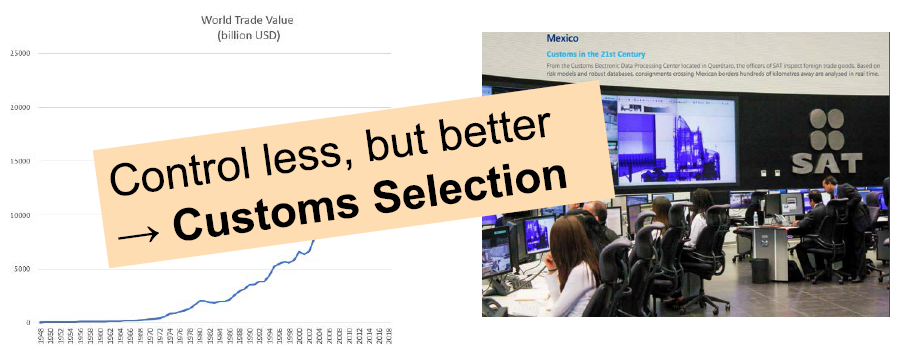
机器所需要的一种能力是将其对一个领域的知识转移到数据（通常）稀缺的另一个领域的能力。尽管迁移学习在各种深度学习应用中得到了充分的适应，但我们仍不了解是什么促成了成功的迁移，以及网络的哪个部分对此负责。在本文中，我们提供了新的工具和分析来解决这些基本问题。通过对传输到块混洗图像的一系列分析，我们将特征重用的效果与学习数据的低级统计信息分开，并表明传输学习的一些好处来自于后者。我们提出，当使用预先训练好的权重进行训练时，该模型在损失景观中保持在同一盆地中，并且该模型的不同实例在特征空间中相似，在参数空间中接近。



关于海关的背景故事：



海关是负责控制跨境货物和旅客流动以及向贸易商征收关税和税款的政府机构。根据世界海关组织（WCO），海关当局清关19.7美元的进口，1.4百万的乘客，占全球税收收入的30%。鉴于跨境流量的天文数字，如何控制更少但更好是海关管理部门的主要任务。另一方面，贸易商可能会试图操纵和忽略一些申报细节，以逃避关税、税收和法规。



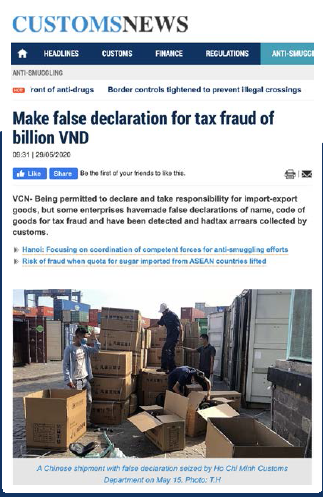


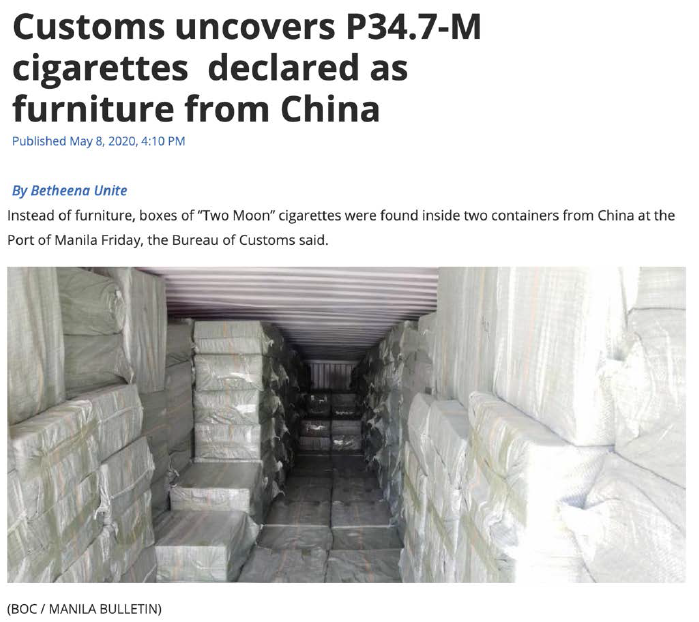
上表总结了具有相应非法动机的海关欺诈类型。低估是最常见的海关欺诈类型，进口商或出口商以低于实际价格的价格申报贸易货物的价值，主要是为了逃避从价关税和税款。然而，在广义的定义中，它包括HS编码中的错误分类——一种对全球贸易产品进行分类的标准化国际协调制度——以及出于逃避关税和税收的动机操纵原产国。例如，将电视（HS 852859，8%关税）申报为PC显示器（HS 852852，0%关税）。本文使用了低估的广义定义。还有其他类型的海关税收欺诈，包括试图完全避免开具发票的走私货物。

贸易欺诈检测、贸易相关金融交易和跨境旅客是海关管理的支柱。一些海关管理部门在其欺诈检测系统中成功地采用了机器学习模型，许多管理部门正在计划或请求国际支持。最近，世界海关组织开始研究机器学习和数据挖掘方法用于欺诈检测的潜力。尽管如此，机器学习在海关领域的应用仍然有限，特别是对发展中经济体而言，而且仍然存在一些挑战，例如可解释性、历史数据的可用性、不断变化的欺诈模式、标记数据的可用性、不平衡数据和隐私问题。

构建海关欺诈检测模型需要考虑以下因素。首先，可解释性是海关管理部门的一项基本要求。在实践中，检查进口商品涉及审查数十份文件，从仓库中数百个包裹中找到欺诈线索，甚至安抚因检查而遭受额外成本和延误的愤怒贸易商。如果我们的模型未能告知检查员其瞄准目标的任何原因，他们可能会抵制模型的预测。其次，贸易商和贸易货物的识别码是侦查海关欺诈的基本特征。在我们的随机森林模型初步结果中，进口商和HS代码的不合规记录是最关键的变量。Date旨在充分利用这两个变量，克服其极高的基数（165000个进口商和6000个HS代码）。最后，我们期望我们的模型在海关欺诈检测方面比当前和传统的机器学习方法提供更好的性能。

欺诈举例：





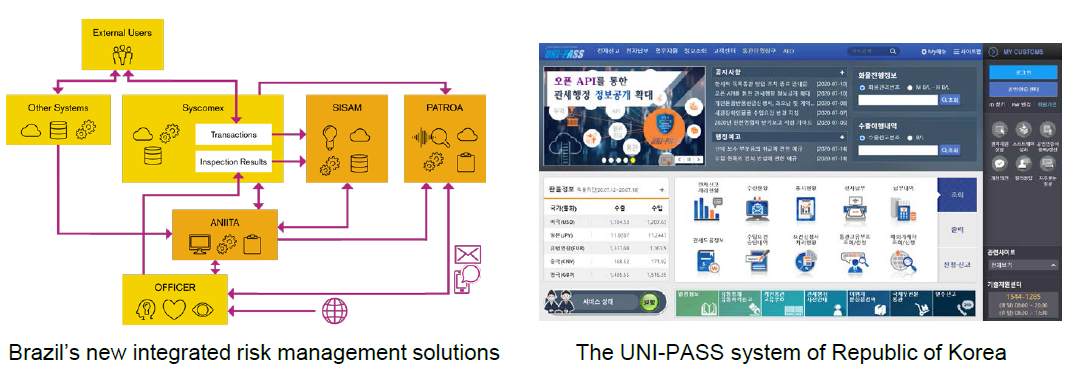


不同地区的欺诈事件：

欺诈检测是一项与许多行业相关的普遍任务。随着数据挖掘和机器学习的出现，已经取得了许多进步，例如基于决策树的方法。对于海关管理（即本研究的目标领域），迄今为止，大多数税务机关都在使用基于规则的方法。基于规则的系统具有可解释性和直截了当性，但对任何新的行为和变化都很脆弱，受专家知识的影响，并且维护起来很麻烦。基于机器学习的系统可以克服这些限制。

就海关欺诈侦查而言，已发表的文献主要由于任务的专有性质而受到限制。然而，也有一些工作可以概括为有监督、无监督和半监督学习技术。有时，为了获得更好的性能，需要部署这些技术的集成。例如，比利时海关在置信度提升算法中测试了基于支持向量机的学习者的集成方法。哥伦比亚海关已经证明了在检测标签有限的税务欺诈时使用无监督光谱聚类。最近，基于树的方法、支持向量机和神经网络的集成已在印度尼西亚的海关数据仓库上进行了测试。

另一项研究揭示了荷兰国际航运记录中的海关欺诈。在这项工作中，建立了一个基于贝叶斯网络和神经网络的模型，将货物清单上货物的存在与货物的随附文件进行比较，以确定是否存在文件欺诈行为。还有其他研究利用Benford定律等方法来检测海关审计中的欺诈行为。另一项研究采用深度学习模型，根据2017年尼泊尔海关随机选取的200000份数据，将高风险和低风险货物分开。



一些国家已致力于开发其欺诈检测系统。**韩国**海关总署开发了一个电子清关系统Uni-PASS，该系统最近发展为部署名为IRM pass的基于AI的风险管理模块。新西兰于2016年建立了联合边境分析（JBA），以利用数据分析和挖掘，获得对边境和海关风险的新见解。虽然他们的方法不是公开的，但JBA使用来自不同来源的数据，如货物、乘客和邮件流以及开源数据。JBA开始关注一系列海关风险和问题，包括低估、药品进口，甚至是黑暗药品市场。巴西联邦税务局开发了一个名为SISAM的复杂选择系统，该系统报告各种欺诈场景的错误率，并输出评估报告以协助税务审计员。定位系统的算法是专有的，并利用了计算机视觉和自然语言处理的最新进展。已知SISAM是从分层模式贝叶斯发展而来的。

大多数没有欺诈检测系统的国家使用海关数据自动化系统，这是一个由联合国设计的计算机化海关管理系统。目前，全世界有90多个国家使用该系统。4海关数据自动化系统有助于纳入简单的基于规则的欺诈检测方法，我们的算法旨在以此为基础。一些欺诈检测方法适用于海关数据自动化系统。例如，可以使用镜像数据分析等传统欺诈检测技术来识别贸易进出口方之间的任何差异。然而，贸易方很少有交易层面的数据，从进口报关单中确定货物的确切类型及其零售价格也不容易。