全国社会保险大数据应用创新大赛

一、团队与成绩

团队成员

许海城（Seaty）：华南理工大学数学学院 2018 届本科生，就读专业为信息管理与信息系统。

庄业广（Hans）：中山大学数据科学与计算机学院 2017 级硕士，就读专业为软件工程。

郭力（alik）：中山大学数据科学与计算机学院 2018 届本科生，就读专业为计算机科学与技术。

比赛成绩

初赛：第 8 名；复赛：第 6 名；决赛：第 4 名。

二、赛题背景

- **精准社保赛题：**

参赛队伍以医保卡历史消费数据为基础，完成数据算法模型设计，实现对各类医疗保险基金欺诈违规行为的准确识别，以进一步提高医保智能监控的针对性和有效性。

- **评选规则：**

$$Precision = \frac{|\cap(PredictionSet, ReferenceSet)|}{|\mathrm{Prediction}Set|}$$

$$Recall = \frac{|\cap(PredictionSet, ReferenceSet)|}{|ReferenceSet|}$$

$$F1 = \frac{2 \times Precision \times Recall}{|Precision + Recall|}$$

其中，PredictionSet 为算法预测的涉嫌造假人员的集合，ReferenceSet 为真实涉嫌造假人员集合。

三、赛题分析与理解

考虑到我们团队没有太多关于医疗保险以及医疗保险欺诈的知识，我们首先展开的工作便是查阅文献资料，快速了解医疗保险理赔的一般业务流程以及医疗保险欺诈的常用手段。

在此基础上，我们结合所给数据各个字段的具体含义，确定其价值以及如何有针对性的进行特征工程。

下面图 1 展示了简化了的医疗保险理赔的一般业务流程图，表 1 列举了一些医疗保险欺诈的常用手段及我们认为会产生的模式，并据此确定的重要字段。

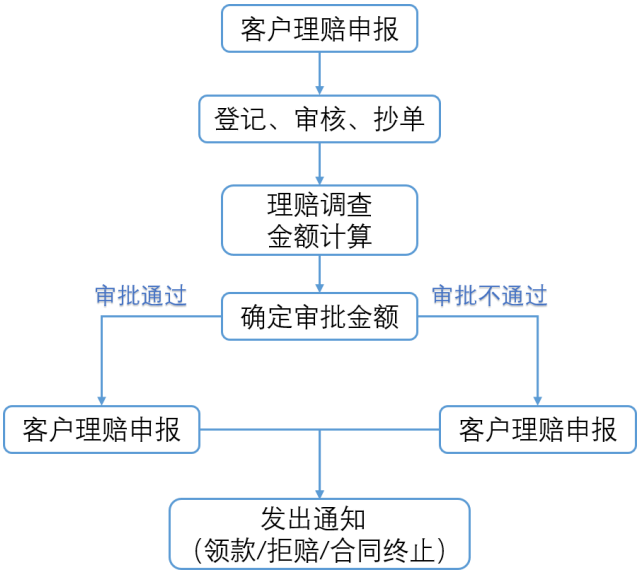


图 1 理赔业务流程

可以从这个简略的医疗保险理赔业务流程图看到有几个比较重要的步骤：用户申报、审批金额确定、领款和拒赔的通知，与人社数据中的申报金额、审批金额、拒付金额、自费金额一一对应，这对我们理解这些不同类别的金额的含义非常有帮助。

表 1：骗保行为分析表

骗保行为	潜在模式	重要字段
挂名诊疗	异常诊疗记录	消费日期、消费金额
医患串通	异常的诊疗医院历史	就诊医院
团伙伪造票据	集中的诊疗项目	就诊项目、诊断病种

		、购买药品
虚假诊疗信息	**不合理的** **药品消费与诊疗项目**	**药品消费** **单价、数量、总额**

- **挂名诊疗**一般现象是，医保卡借用、冒名顶替、挂名住院等，**我们认为**多个不同人使用同一张医保卡，势必会导致该医保卡产生的消费历史存在异常，可能是诊疗项目上的、就诊医院上的等等。针对这种行为，**我们构造了**消费频率特征，消费金额统计特征。
- **医患串通**有很多种方式，我们这里主要是指非团伙式的，医生为主动方的诈骗方式，**我们认为**这在一定程度上会导致该医院的用户群与其他医院存在差异，进而这些用户群也会有不同的就诊医院潜在选择偏好。针对这种行为，**我们构造了**用户-就诊医院 TF-IDF 经 stacking 后的概率特征。
- **团伙式伪造票据**，顾名思义是团伙所为，由于大量的作案，因此作案手法一般会比较常规，并且有一定的频繁模式，如：诊疗项目一般是各式各样的常见项目，**我们认为**这会在一定程度上产生一定的集中的诊疗项目、诊断病种，我们也验证了一些常见病种在正负样本上有高置信水平的差异。针对这种行为，**我们构造了**文本的 one-hot 特征以及文本的 doc2vec 经 stacking 后的改率特征。
- **虚假诊疗信息**与团伙式伪造票据似乎有一些相似，虚假诊疗信息更多的是在患者要求下，医生将一些不可保的药品进行修改，或者对用量进行修改等等。相较之下，可能不如团伙作案那样，**我们认为**虚假诊疗信息会存在一些不合理的地方如：诊疗糖尿病，开了通心络胶囊或者过量开药等等。针对这种行为，**我们同样构造了**文本的 one-hot 特征以及文本的 doc2vec 经 stacking 后的改率特征。

四、数据探索与特征工程

1.消费日期字段

如图 2 所示，我们对每个用户，按交易时间进行排序后，发现有许多在时间

上成对出现的消费记录，详细统计发现：大约有 62.34%的记录在时间上是冗余的记录。因此，我们将在时间上冗余的记录过滤掉之后，对每个用户，计算相邻两次有效诊疗的天数，并计算其统计量作为**消费频率特征**。

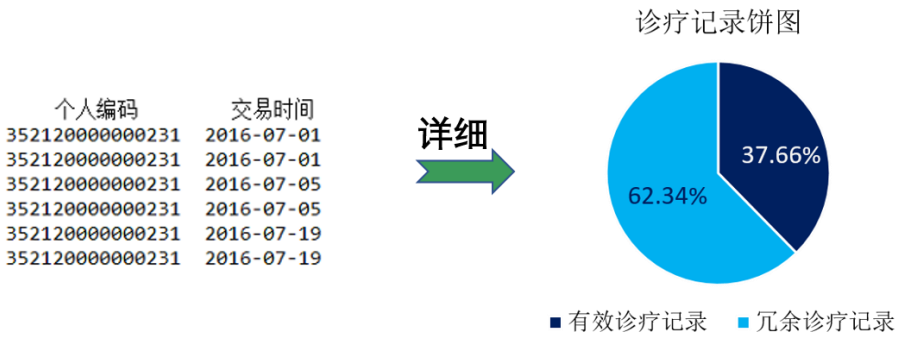
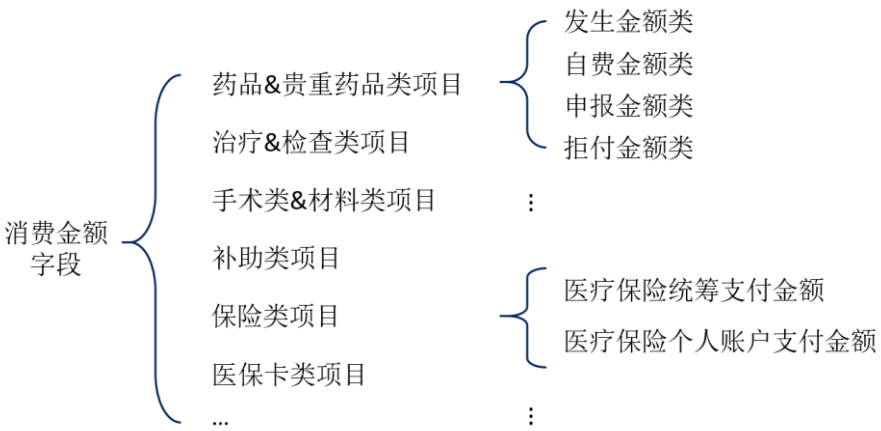


图 2 交易时间字段探索

2.消费金额字段



可以看到消费金额类字段，我们按项目、类型进行了分组，我们在计算字段的统计量时会分组进行。具体来说：假设总共有 3000 条记录，补助类项目有字段 col1、col2、col3，我们在计算 col1、col2、col3 的统计量时，会过滤掉那些 col1、col2、col3 都取 0 值的记录，因为 col1、col2、col3 都取 0，我们并不认为该条记录属于在该项目上消费的记录（除了这种直观的解释，如图 3 展示了部分字段取非 0 值的比例情况，有些项目取非 0 值的记录相对较少，如果不这样处理，那在该项目包含的字段上计算的统计量都会受到影响，例如：平均值偏小）。

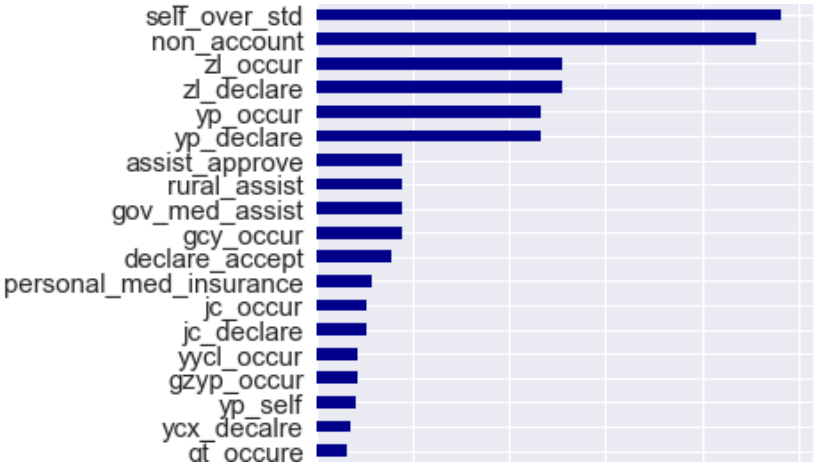


图 3 消费金额类字段 0 值比例

3.就诊医院字段

如图 4，对于原始数据用户 k 在医院 m 的就诊序列，我们将其类比作文档 k 中出现了单词 m，基于此再将“词频”转换成能更适合的表达重要性 TF-IDF 统计量。

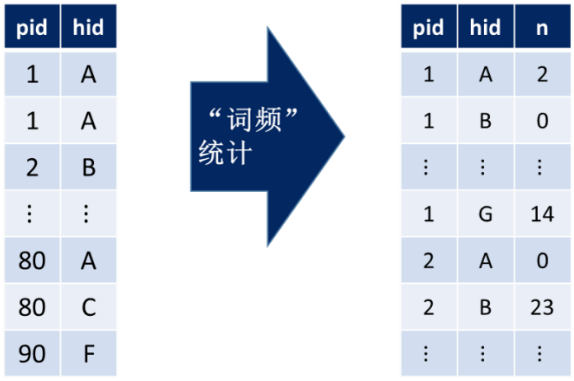


图 4 就诊医院序列处理

这里有必要展开详细的解释， TF-IDF 统计量是为了减小文档中频繁出现的常见词的重要性而定义出来的统计量，那么对人社数据使用这样一个统计量是否合适呢？

再强调一次：用户对应文档、医院对应单词。由于欺诈用户只占少数，因此我们认为主要是由正常用户决定一个医院是否是“常用单词”，在这样的假设下，一个医院如果是“常用单词”，即很多**正常**用户都会去的医院，那么他一般是正规的、口碑好的大医院。因此，这类“单词”对于预测欺诈用户没有过多的信息量，因此适当减小这类医院的重要性是合理的。这说明在人社数据上使用 TF-IDF 统计量进行重要性修正是合适的。

这样我们就得到了每个用户的 TF-IDF 特征向量，为了防止过拟合，我们选用了最简单但有效的 Logistic 模型，结合 3 折交叉验证，预测出一列新的概率值，作为新的特征。具体的产生过程如下图 5：

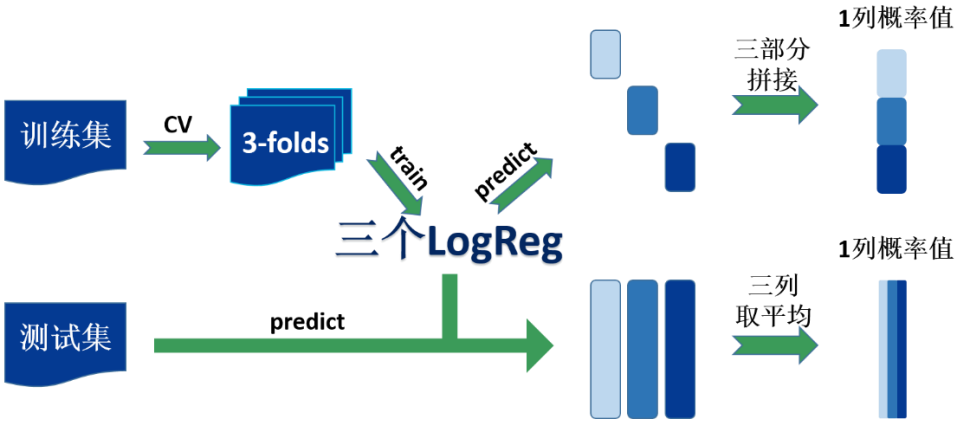


图 5 Logistic + CrossValidation

4.文本类字段（诊断病种 & 就诊项目 & 所购药品）

由于这些文本都是由病种名、项目名、药品名组成，很直观的想法就是人为选取一些病种名、项目名、药品名，通过判断文本是否含有相应的名称，来构造一系列的二元变量。值得注意的，文本中的分隔符也包含信息（可以构造一个新的字段，指示文本中含有的分隔符数量）!

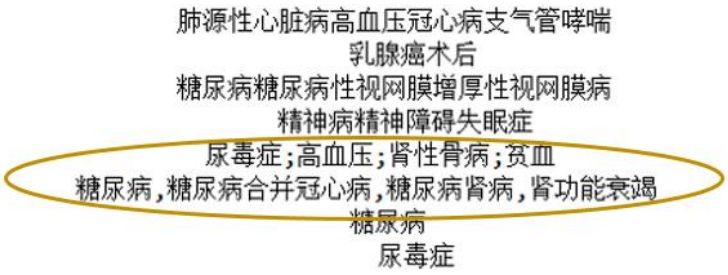


图 6 诊断病种示例

不仅如此，考虑到仔细的挖掘这些不一致、不规整的文本需要耗费大量精力，同时也为了能挖掘出更多的信息，我们还采用了文本挖掘的方法（Doc2Vec），最终构造了用户特征矩阵，具体构造步骤见图 7，同样的，我们再次使用 Logistic 模型，结合 3 折交叉验证，预测出一列新的概率值，作为新的特征。

值得一提，但是不知道有没有效果（没有验证），我们在构造历史就诊文档的步骤是既粗糙又精细的——粗糙在于我们是直接粗暴的将不规整的文本利用分隔符直接拼接起来的；精细在于我们在拼接时，先在字段维度上拼接，拼接结果

为一次交易的文本，随后按时间排序后再进行拼接，拼接结果为交易序列文本，我们在两次拼接使用的是不同的分隔符，期望文本挖掘方法能利用到这一信息。

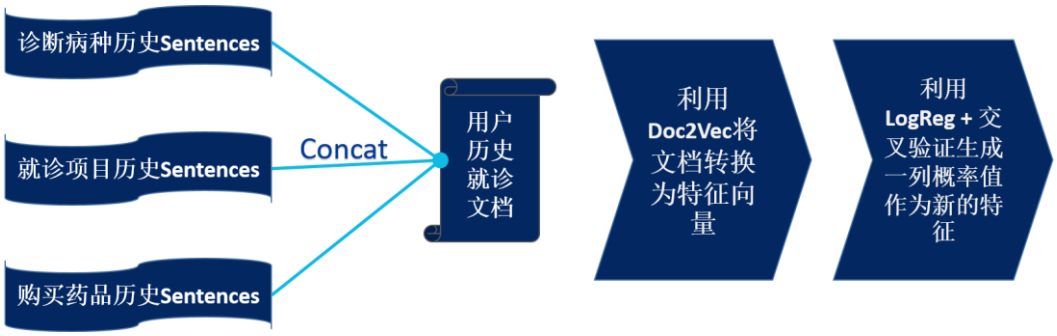


图 7 Doc2Vec + Logistic + CrossValidation

五、算法模型

由于预测结果的融合对预测结果的多样性有一定要求，主流一般使用不同模型、不同参数、不同训练集三种方式来提高预测结果的多样性。

由于人社数据具有数据不平衡的特点（欺诈类样本占 3%-5%左右），因此相比随机森林这样的 Bagging 方法，XGBoost、GBDT、LightGBM 这类 Gradient Boosting Machine 会有更好的效果，因此我们主要是选用了这三类模型作为后续模型融合的基模型。不仅如此，我们还为样本设置了权重，以增大单个欺诈样本的 Loss，从而使模型更注重把欺诈样本划分出来。

另外，因为我们选用的融合方法对基模型的效果要求比较高，因此我们选用的都是经过交叉验证确定的最优参数，于是我们结合了人社数据的特点在训练集多样性又进行了一些工作——我们为每个模型训练都构造一份训练集。如图 8，具体来说，考虑到欺诈类样本非常珍贵，我们予以全部保留，非欺诈类样本我们随机去除 10%的样本，以此构造训练集。这样既考虑到欺诈类样本的宝贵性，也兼顾了模型融合对多样性的需求。

图 8 训练集构造

我们的融合策略简单，但有效，如图 9，首先按图 8 流程构造 20 份训练集，作为 10 个 XGBoost、10 个 GBDT 的训练数据，最终训练得到 20 个 GBM 模型，利用这些模型对测试集进行预测，得到 20 组预测为欺诈的概率，对每组取概率最大的前 520 个用户作为欺诈用户，将 20 组欺诈用户再取并集，最终得到 920 个左右的欺诈用户，以此作为我们最终的预测结果。值得一提的是，这种融合方式的主要优点是能稳定模型结果，使线下线上尽可能一致，缺点是需要基模型有较好的性能，比较容易达到上限，无法超越基模型的效果。

图 9 模型融合

六、总结与反思

从我们上面的比赛介绍，不难看出，在这次类别不平衡的二分类比赛中，对于最核心的特征构造方面，我们团队更多的是根据业务逻辑驱动的方式找到构造特征的方向，而且总体方向基本都涵盖到了。实际上，与其他团队相比，在特征种类方面我们并没有落后，但是与数据分析驱动的团队相比，我们这种方式缺少了对哪些种类的特征更有效的一个认识。例如，其他团队经过分析，发现消费频率类特征能和金额类特征有很好的互补作用，因此他们在消费频率类特征上经过**滑窗**做了更多时间粒度的特征，最大限度的利用了这一信息，而我们没意识到这一特征的重要性，因此当时只简单做了一个维度。这一问题，一定程度上可以通过分析树模型产出的特征重要性来发现，但是很遗憾的是，当时金额类特征的重要性比较高，因此尽管我们观察了特征重要性结果，还是没发现这个特征的重要程度。除此之外，考虑到消费频率这类波动特征非常重要，而我们只利用方差来捕获金额的波动信息，因此应该考虑对金额波动也做一些更细致的特征。

七、参考文献

http://d.wanfangdata.com.cn/Periodical/zgshbz201502041

http://news.vobao.com/zhinan/jiankangxian/805719998669264361.shtml

《基于医疗大数据的病患医保欺诈行为的数据挖掘》- 张家宝 李昂 鄢然

《数据挖掘在医疗保险理赔分析中的应用》- 李娜娜