**关于候诊时间预测模型训练的初步阶段总结**

广东易生活信息科技有限公司莫世玉

2020年8月11日

1. **原始数据**

表头内容：['His流水号', '急诊系统流水号', '门诊号', '就诊时间', '年龄', '性别', '分诊科室', '分诊等级', '分诊去向', '分诊时间','抢救上床时间', '体温类型', '体温', '脉搏', '收缩压', '舒张压', '呼吸', 'SpO2', '微量血糖', '疼痛评分', 'GCS评分', 'PHI评分', '来诊主诉', '现病史', '既病史', '药物过敏史', '诊断', '处理意见', '注意事项', '时长', '没有时长的主诉', '主诉分词', '编码']。

数据量共350626条，涵盖了2016年、2017年、2018年的数据。根据项目需要和现有数据对项目研究完整性的具体需求，从中筛选出2018年的数据，数据量共有12万8645条。

1. **数据预处理**

针对所选的2018年数据，进行特征的选择与预处理如下：

（1）数据中【'His流水号', '急诊系统流水号', '门诊号'】这部分数据是属于医院相关的，专属病人的代号，具有强烈的隐私的性质，并且在数据模型训练中利用不上，已做去除/清洗/掩盖等处理。

（2）目标：候诊时间，为患者实际等候的时长，是模型训练的target，是由【'就诊时间'】减去【‘分诊时间’】得出，时间单位为分钟。其中，最大值为：1439（分钟），最小值0分钟，平均数32.48(分钟)，中位数15(分钟）。

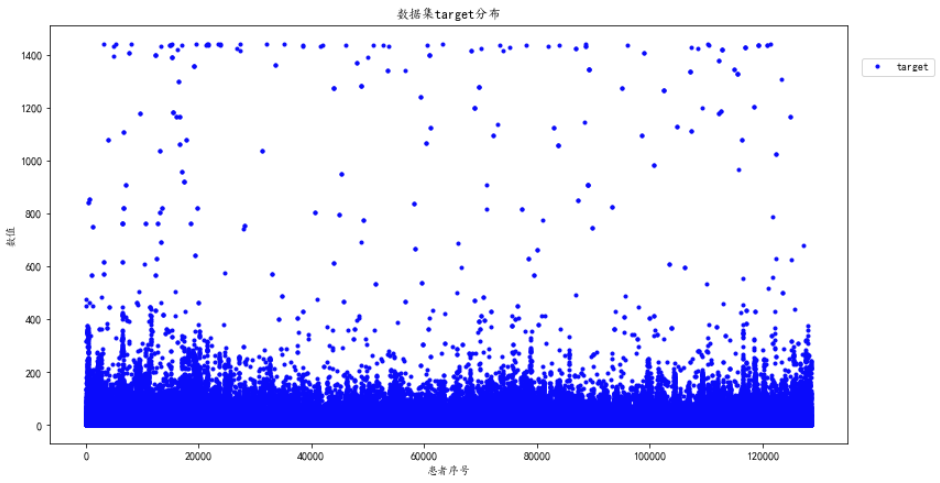


图1 数据集target的数值分布

从数据分布来看，大部分的target都基本集中在200以下，非常密集，400（Y轴）以上的数据分布非常稀疏。

（3）数据中特征【‘年龄’】的样式是字符串的类型，有数值的同时也标注了年(‘Y’)和月(‘M’) 的区别。在处理中，有‘Y’的标志直接取其数值作为年龄，对于‘M’标志的取其数值后除以12，统一换算成以年为单位的数值。

（4）数据中特征【‘性别’】，按照惯例，采用0/1来代替，‘男’为0，‘女’为1。其中性别人数的分布情况统计如下，可见性别属于均匀分布：

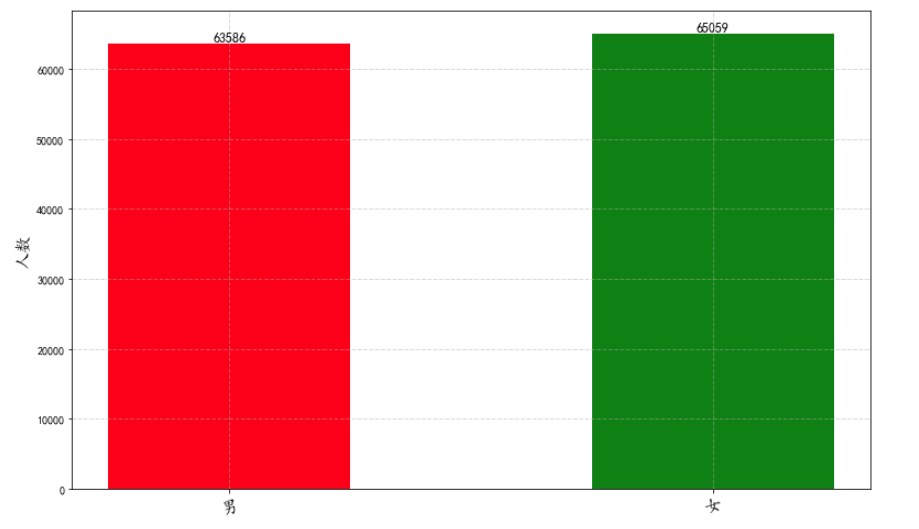


图1 数据中性别人数的分布

（5）特征【‘分诊科室’】，含有的数据为['急诊外科', '急诊内科', '急诊儿科', '急诊产科', '内科夜诊', '发热门诊', '急诊耳鼻喉科', '急诊复诊', '急诊眼科', '急诊妇科', '急诊口腔科', '急诊感染科', '急诊新生儿科']等13种数据样式，分别以1到13的数值进行置换处理。

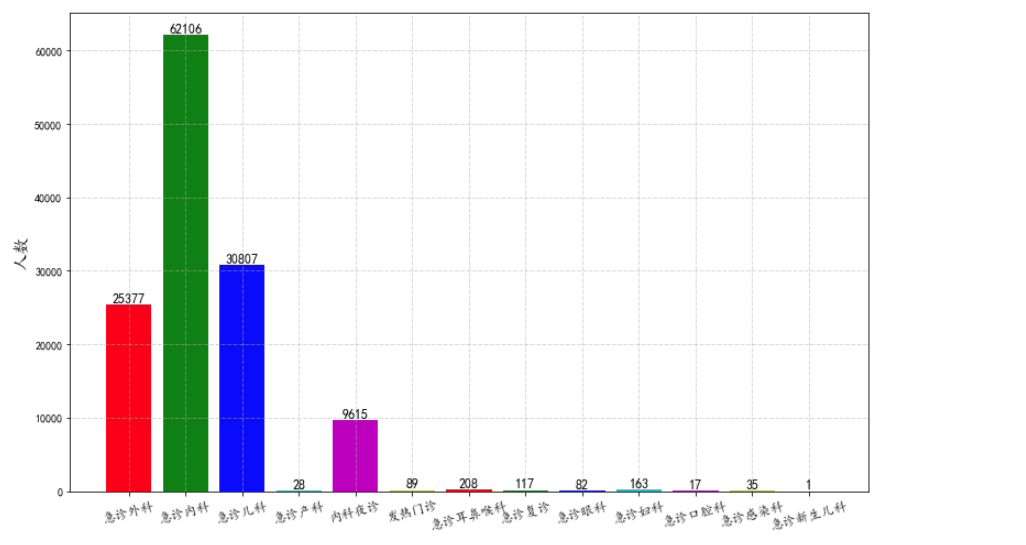


图2 特征【‘分诊科室’】的人数分布

从图2来看，分诊科室的人数都基本集中在‘急诊外科’，‘急诊内科’，‘急诊儿科’和‘内科夜诊’，其他科室也有人数，但在这四科的人数面前，可以说忽略不计。在目前的数据探索中，依然保留该部分的不平衡的数据。

（6）特征【‘分诊等级’】的数据样式均为正整数数值型，如【1,2,3,4】，其人数分布如下图所示：

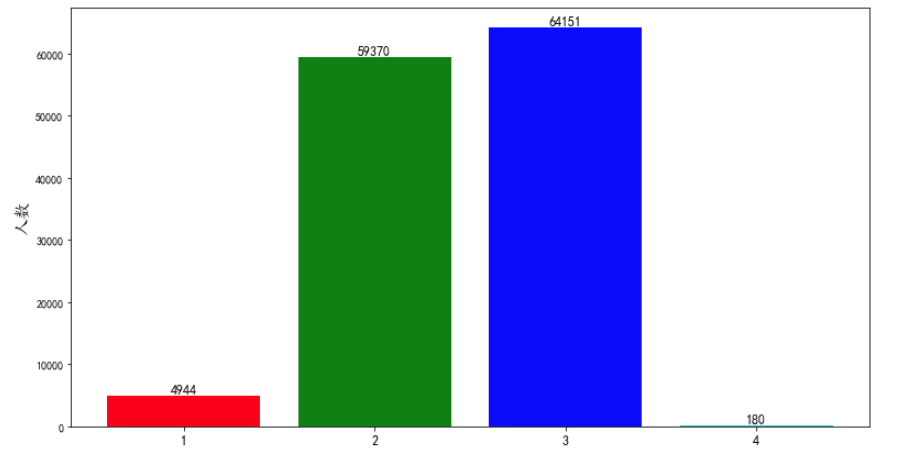


图3 特征【‘分诊等级’】的人数分布

从图3来看，患者的‘分诊等级’基本都在2级和3级的分布上，同样属于不平衡的数据。

1. 特征【‘分诊去向’】，数据样式仅为中文样式【‘医疗室’，‘抢救室’】，‘医疗室’的数量占比最多，有12万5195条，‘抢救室’有3450条数据。由于是中文样式，转换数字目前分别用【1,2】来进行替换处理。
2. 特征【‘分诊时间’】不直接作特征选项，但对该特征进行分裂处理，形成新的以下【‘分诊-月’，‘分诊-日’，‘分诊-时’，‘分诊-分’】四个新特征进行如模型训练的。说明：这里没有‘分诊-年’，由于年份的数据在机器学习中是不会有循环的，年份的数值只会往后走，机器学习模型意义是学到过往的数值；没有‘分诊-秒’是因为‘分诊时间’中的原始数值中，其‘秒’的部分全部为0。
3. 特征【'抢救上床时间'】为穿越数据的一部分，因此不作特征考虑。
4. 特征【'体温类型'】的数据样式为中文字符串的['耳温', '腋温']，数据中只有7条数据是描述为‘腋温’，分布似乎有些单一，且由于是中文样式，转换数字目前分别用【1,2】来进行代替。
5. 特征【‘体温’】数据中，数据样式中除了有整型的数据外，也有其他不符合的字符存在，可能是技术人员原始录入时出错导致的。在进行归一化预处理时，按照逻辑合理的原则进行相应的修正，例如：

这一类的数据样式为带有其他不符合的字符存在时，但却能够直观的看到起数值的大小，故按照逻辑合理的规则进行修正。除此之外对于一些空值，和一些异常值的存在，统一修正为正常体温37。

1. 特征【'脉搏'】对空值的进行填充平均值89.96。
2. 特征【'收缩压'】对空值的进行填充平均值138.404。
3. 特征【'舒张压'】对空值的进行填充平均值81.03
4. 特征【'呼吸'】对空值的进行填充平均值19.3。

（16）特征【'SpO2'】，数据样式中除了有整型的数据外，还有其他不符合的字符存在，对其进行合理性原则修正，对于空值或者没有值的数据补充为0（在树模型才可用None表示，对于其他无法使用None表示的模型进行补充0）。

（17）特征【‘疼痛评分’】都为一些不合理、且无法理解的数据存在，因此对这个特征，不作模型训练特征。

（18）特征【'GCS评分'】对空值的进行填充平均值15。

1. 特征【'PHI评分'】对空值的进行填充0。
2. 特征【‘来诊主诉’】参照‘分级’的项目模型的做法，

对主诉中的词语进行提取，按照项目中的统一编码来进行数值代替，形成新特征【‘编码’】。例如下图显示：

上图中，‘恶心呕吐’是来诊患者的主诉（属一句话），在分级项目中的处理方法：首先，分词，分成‘恶心’和‘呕吐’；然后进行词语同义，‘恶心’和‘呕吐’属同义词，替换成‘呕吐’，只有一个词语；最后，进行词语编码（模型需要数值型的数据），通过部位的分类升序来进行判断该词语所在的位置，因此得到一个编码数值‘70101’。

1. 特征【'现病史', '既病史', '药物过敏史', '诊断', '处理意见', '

注意事项', '时长', '没有时长的主诉'】：这方面的数据样式均为中文句子表示，且大多数为空值。由于其中间或有穿越数据，比如：‘处理意见’，‘注意事项’明显是已经属于就诊后的，并不在候诊的时刻中，且中文句子中需要进行相当大量的词语提取、词语替换、词语编码等NLP（自然语言处理）的处理，因此目前不考虑作为特征参与进行训练模型。

**综上所述**，经过数据分析与处理后，符合进行模型训练的基本特征为：['年龄', '性别', '分诊科室', '分诊等级', '分诊去向', '体温类型', '体温', '脉搏', '收缩压', '舒张压', '呼吸', 'SpO2', '微量血糖', 'GCS评分', 'PHI评分', '分诊-月', '分诊-日', '分诊-时', '分诊-分', '脉压差']；模型训练的target为：【‘候诊/分钟’】。

1. **模型训练**

采用上述中数据预处理方法后，对所形成的特征数据进行模型搭建，目前已采用比较常用的模型lightGBM模型进行了初步尝试，模型参数暂时采用常用的默认参数，往后将根据实际训练和测试的结果再进行参数调整，或者选择其它神经网络的模型继续尝试和比较结果。

按照8:2 的比例进行随机分割训练集和测试集，训练集的数量为10万2916，测试集为是2万5729条。

1. 模型结果

特征工程探索：根据分诊的时间（即假设为病人的挂号后的时刻）进行堆叠特征，采用日期的工作日判断的接口，对每条数据的日期进行判断是否‘工作日’、‘节假日’、‘休息日’。 经判断，总数据中‘工作日’共有8万1908条（训练集：65552，测试集：16356），‘节假日’有10187条（训练集：8190，测试集：1997），‘休息日’有3万6550条（训练集：29174，测试集：7376）。

因此，添加了三个特征‘工作日’、‘节假日’、‘休息日’。

目前最优的模型的分数表现如下：

MAE（平均绝对误差）：22.7320（分钟，越小越好），

说明：平均绝对误差是所有单个观测值与算术平均值的偏差的绝对值的平均。平均绝对误差可以避免误差相互抵消的问题，因而可以准确反映实际预测误差的大小。

R方决定系数：0.28（越接近1越好），说明：表示一个随机变量与多个随机变量关系的数字特征，用来反映回归模式中说明因变量变化可靠程度的一个统计指标

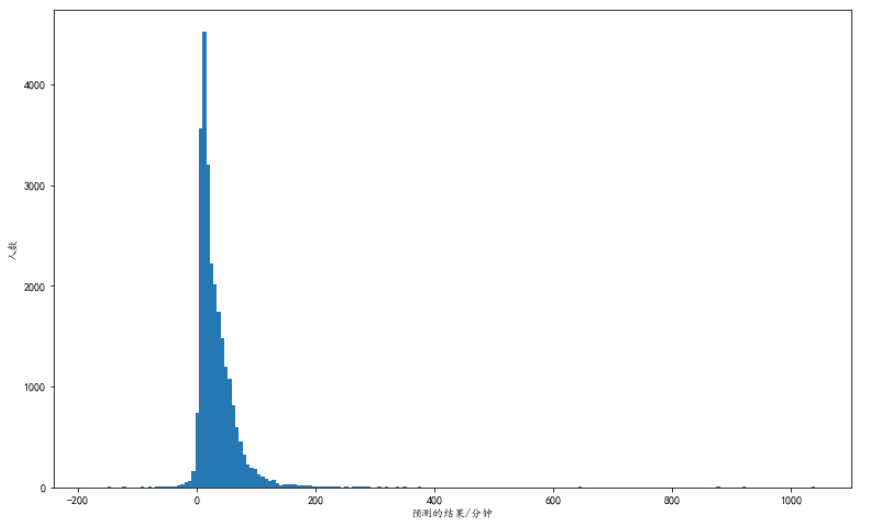


图4 预测结果的频数直方图

从上图来看，预测的结果绝大部分都落在【0-100】分钟的区间中，解释为：在当前的误差情况下，模型预测出来的患者的候诊时间绝大部分都在100分钟以内，另从图中看，有115条数据的预测结果在0以下，为负数，可作为异常值处理。

因此在这里定义了一个标准，预测值最好不超过真实值的15分钟，即数据模型给出的预测结果较实际患者的候诊时间误差不超过15分钟。

①在汇总真实值与预测值的数据进行统计，符合标准的[0<(预测值-真实值)<=15，没有绝对值] 数据量在9929条，只占了测试集数据量的约39%。

②在回归模型中，预测值有大于真实值的误差，也会有小于真实值的误差，因此统计 [-15<=(预测值-真实值)<0]的情况下的数据有4508条，占数据量的17.5%；

如果从绝对值来看，[-15<=(预测值-真实值)<15] 的数据就有了 14437条，占56.5%（39%+17.5%）。

③统计误差在30 分钟以内（绝对值，[-30<=(预测值-真实值)<30]）的数据，则符合的数据量有1万9197条，占测试集数据量的约75%。这18.5%（75%-56.5）的部分数据有望能够通过模型的优化来降低误差，因此这部分的数据存在有一定的模型优化空间。

1. 讨论和建议

初步探索预测患者候诊时间的初步模型的误差在22.73分钟（MAE），从数值结果分布来看具有可以模型优化的空间，尽最大可能把误差降低。当下用于模型训练的数据是来源于病人本身，即通过病人的自述主诉以及生化指标的数据呈现；但候诊时间的本身对于医疗地点相关的资源配置同样具有相关的联系，如院方的出诊情况、医疗设备、资源配置等等，如有这方面的数据，可以更进一步的探索能够够提高相关数据模型的特征方法来优化模型。

除此之外，模型优化的特征工程目前仍在探索中，力求从目前存在的特征作添加、筛选等处理的方法来寻找出最合适模型训练的数据特征。

另外，从模型选择、算法设计的方法入手，可以从常用的方法来选择最优的模型，如‘k折交叉验证法’，‘特征交叉法’，‘参数调节’等等。