风险预测模型

1. 数据提取

* 将htn0817表、htn\_reg表合并，得到mpi\_person\_id、confirm\_date、visator\_date、icd字段，将出现实验组icd编码且最小并发症日期晚于确诊日期的人群提取出来作为实验组数据集。

|  |  |
| --- | --- |
| 疾病名称 | 最终数据集 |
| 严重脑神经疾病、意外死亡及后遗症 | 60050 |
| 心衰人群 | 936 |
| 心源性猝死 | 128 |
| 脑卒中 | 6349 |
| 急性心梗/急性冠脉综合症 | 757 |
| 肺栓塞 | 21 |
| 严重肾病人群 | 1167 |

实验组数据集存放位置：hdfs目录下/user/cj/output/dataset

将实验组、对照组的mpersonid和类标号放到一个文件中，该文件名为experment\_controlgroup, 共52794个mpi\_person\_id.

* 特征抽取

将experment\_controlgroup、htn\_reg、htn\_follow\_up、person\_info、liqi\_htn\_time\_gap表合并。

1. 数据预处理

1.无用特征删除

age(年龄)：系统时间-出生日期

心衰风险预测模型

目录

[一、数据抽取 2](#_Toc464914797)

[二、数据预处理 2](#_Toc464914798)

[1.特征处理： 2](#_Toc464914799)

[2.异常值处理： 2](#_Toc464914800)

[3.一位患者对应一条记录的抽取策略： 3](#_Toc464914801)

[4.缺失值处理： 3](#_Toc464914802)

[三、逻辑回归模型 3](#_Toc464914803)

[1.删除全为0的icd编码，k近邻填充缺失值k=20： 3](#_Toc464914804)

[2.完全删除严重并发症的icd编码,k近邻填充缺失值k=20： 3](#_Toc464914805)

[3.完全删除严重并发症的icd编码，k近邻填充k=10： 4](#_Toc464914806)

[四、进一步想法 4](#_Toc464914807)

[附 数据分析统计结果 4](#_Toc464914808)

[字段取值范围及空缺值统计： 4](#_Toc464914809)

[各字段空缺情况统计： 5](#_Toc464914810)

# 一、数据抽取

心衰的并发症日期根据门诊表中的门诊日期得出。

实验组：高血压人群中患有心衰、且确诊日期唯一、心衰并发症日期在确诊日期之后有694位患者。

对照组：除去300多个严重并发症、且三年以上患有高血压的健康人群有8552位。

实验组和对照组的特征从高血压登记表、高血压随访表、门诊表、个人基本信息表中抽取出来，共 52个特征，包括性别、年龄、出生日期、机构编码等等. 与心衰患者相关ICD编码有366个。初步所有特征共418个。

# 二、数据预处理

## 1.特征处理：

对实验组和对照组的特征初步筛选，筛掉明显无意义的特征，例如组织机构编码、档案流水号、终止管理日期、随访流水号、转诊标志、转诊原因、转诊机构、下次随访日期。

删除冗余特征，例如诊断编码和诊断名称，保留一个即可。由于诊断编码未找到说明，所以保留诊断名称列。

统计每一个特征的非缺失值（根据患者数目计算而非记录条数）的数目，删掉缺失值在70%以上的特征。

对于ICD编码列（出现该病的为1，没有记为0），统计每一列，所有值为0的列删除。由于对照组中的数据不包含严重并发症的ICD编码，所以将严重并发症的ICD编码删除。

## 2.异常值处理：

经过以上的处理，考虑到实验组数据量非常少，某些特征缺失值非常大，所以最终保留特征如下，除了日期数据异常进行删除操作之外，其他特征异常值视为缺失值处理。

## 3.一位患者对应一条记录的抽取策略：

由于一个患者标识对应多条记录（并发症日期唯一但随访 日期不唯一），优先提取随访日期距离并发症日期最近且在并发症日期之前的记录，即每个患者对应一条记录，次之提取随访日期距离并发症日期最近的记录。（有的患者没有随访日期在并发症日期后的记录。）

由于对照组的数据量远大于实验组数据量，为了更好建立预测模型，首先，统计分析age年龄分布、sex性别分布。按照实验组比例，从对照组中选取了976个数据缺失量最少的患者的记录。最终保留了694个实验组和976个对照组作为最终的数据集。保留特征26+178 ICD编码。特征见附件：保留特征.xlsx

## 4.缺失值处理：

经过统计， 缺失值主要集中在Symptom\_Name、GLU（空腹血糖）、weight(体重)、height（身高）、heart\_rate\_times（心率）、mind\_adju\_code(心理描述)、behavior\_code(行为指导)、followup\_type\_code(随访评价结果)、age。

Diag\_name\_inhos用icd编码替换，所以该列不存在缺失值。

1）症状编码（Symptom\_Name)为离散属性，所以维度扩充，扩充后，对于每一个值，出现该特征的记为1，否则记为0，这样就不存在缺失值的情况。

2）其他列，一个患者对应的多条记录中非缺失值的平均值进行填充。

3）对于一个患者对应的多条记录均为缺失值的情况，体重、身高等特征按照性别分类，取平均值填充。其他列使用k近邻法填充。

# 三、逻辑回归模型

Spark mlib库中的Logistic regression 初步训练数据集，10折交叉验证评价结果。

## 1.删除全为0的icd编码，k近邻填充缺失值k=20：

数据集：Nullvaluefill.csv：

参数权重见”模型参数.xlsx” 中子表”原始”；

10折交叉验证：99.95%

## 2.完全删除严重并发症的icd编码,k近邻填充缺失值k=20：

数据集：icd\_data\_control\_group\_final\_dataset.csv；

参数权重见”模型参数.xlsx” 中子表”删除后ICD编码”

10折交叉验证结果：96.92%

## 3.完全删除严重并发症的icd编码，k近邻填充k=10：

数据集：icd\_data\_control\_group\_final\_dataset.csv；

参数权重见”模型参数.xlsx” 中子表”删除后ICD编码1”

10折交叉验证结果：96.92%

# 四、进一步想法

特征之间的相关性分析，数据规范化处理、参数调优。

# 附 数据分析统计结果

## 字段取值范围及空缺值统计：

1. 部分字段取值范围：

|  |  |  |  |
| --- | --- | --- | --- |
| 异常值定义 | 实验组 | 对照组 | 值类型 |
| ﻿MPI\_PERSON\_ID | 无 | 无 | 标称 |
| VISITOR\_DATE | 在确诊日期之前 | 无 | 序数 |
| CONFIRM\_DATE | 在出生日期之前 | 在出生日期之前 | 序数 |
| DATE\_OF\_BIRTH | 在确诊日期之后 | 在确诊日期之后 | 序数 |
| SEX\_CODE | 非0、1、2、9 | 非0、1、2、9 | 标称 |
| RISK\_STRATIFY | 非1、2、3、4、 | 非1、2、3、4 | 序数 |
| SBP | 96-180取值正常 | 90-220取值正常 | 数值 |
| DBP | 48-128取值正常 | 50-160取值正常 | 数值 |
| FOLLOWUP\_DATE | 在确诊日期之前 | 在确诊日期之前 | 序数 |
| SYMPTOM\_NAME | 正常 | 正常 | 非对称二元 |
| GLU | 4-11取值正常 | 2-14取值正常 | 数值 |
| WEIGHT | 37-96取值正常 | 19-120取值正常 | 数值 |
| HEIGHT | 137-182取值正常 | 16-191取值正常 | 数值 |
| BMI | 17-35取值正常 | 15.4-42取值正常 | 数值 |
| HEART\_RATE\_TIMES | 54-130取值正常 | 18-138取值正常 | 数值 |
| MIND\_ADJU\_CODE | 非1、2、3 | 非1、2、3 | 序数 |
| BEBAVIOR\_CODE | 非1、2、3 | 非1、2、3 | 序数 |
| FOLLOWUP\_TYPE\_CODE | 非1、2、3、4 | 非1、2、3、4 | 序数 |
| AGE | 33-115取值正常 | 4-115取值正常 | 数值 |
| DIAG\_NAME\_INHOS | 无 | 无 | 序数 |

注:除了日期异常处理方式是剔除之外，其他视为缺失值

## 各字段空缺情况统计：

1. 实验组：共有695位患者

|  |  |
| --- | --- |
| 字段名称 | 非缺失值 |
| VISITOR\_DATE | 695 |
| CONFIRM\_DATE | 695 |
| DATE\_OF\_BIRTH | 695 |
| SEX\_CODE | 694 |
| RISK\_STRATIFY | 695 |
| SBP | 695 |
| DBP | 695 |
| FOLLOWUP\_DATE | 695 |
| SYMPTOM\_NAME | 323 |
| GLU | 322 |
| WEIGHT | 569 |
| HEIGHT | 323 |
| BMI | 323 |
| HEART\_RATE\_TIMES | 546 |
| MIND\_ADJU\_CODE | 678 |
| BEBAVIOR\_CODE | 680 |
| FOLLOWUP\_TYPE\_CODE | 323 |
| AGE | 682 |
| DIAG\_NAME\_INHOS | 671 |

1. 对照组：共8552位患者

|  |  |
| --- | --- |
| 字段名称 | 非缺失值 |
| VISITOR\_DATE | 8552 |
| CONFIRM\_DATE | 8552 |
| DATE\_OF\_BIRTH | 8552 |
| SEX\_CODE | 8523 |
| RISK\_STRATIFY | 8552 |
| SBP | 8548 |
| DBP | 8548 |
| FOLLOWUP\_DATE | 8552 |
| SYMPTOM\_CODE | 6209 |
| GLU | 5082 |
| WEIGHT | 7996 |
| HEIGHT | 6207 |
| BMI | 6203 |
| HEART\_RATE\_TIMES | 7615 |
| MIND\_ADJU\_CODE | 8402 |
| BEBAVIOR\_CODE | 8414 |
| FOLLOWUP\_TYPE\_CODE | 6209 |
| AGE | 6432 |

1. 各字段取值情况说明：

经过统计， 缺失值主要集中在Symptom\_Name、GLU（空腹血糖）、weight(体重)、height（身高）、heart\_rate\_times（心率）、mind\_adju\_code(心理描述)、behavior\_code(行为指导)、followup\_type\_code(随访评价结果)、age。

SYMPTOM\_NAME的编码情况（对应的SYMPTOM\_CODE）：

|  |  |
| --- | --- |
| SYMPTOM\_NAME | SYMPTOM\_CODE |
| 无症状 | 1 |
| 头晕 | 2 |
| 四肢发麻 | 3 |
| 头痛 | 4 |
| 眼花耳鸣 | 5 |
| 心悸胸闷、心悸 | 6 |
| 恶心呕吐 | 7 |
| 呼吸困难 | 8 |
| 下肢水肿 | 9 |

|  |  |  |
| --- | --- | --- |
| mind | behavior | follow |
| 1、良好 | 1、良好 | 1、控制满意 |
| 2、一般 | 2、一般 | 2、不满意 |
| 3、差 | 3、差 | 3、不良反应 |
|  |  | 4、并发症 |

实验组并发症日期和确诊日期最长的时间间隔20773天

对照组的并发症日期应该在确诊日期和最长时间间隔之间

1. 实验组缺失值填充：

初步提取不同的M\_person\_id( 695 位), 距离并发症日期最近的记录提取出来作为该m\_person\_id的记录标准（先提取离并发症日期最近且大于该日期的随访日期的记录）， 对出现空值的字段进行同一m\_person\_id内均值填充。

1. 对照组缺失值填充：

初步提取不同的m\_person\_id（8552位），确诊日期和并发症日期的时间间隔在实验组得出的最长时间间隔之间，随机选择同一m\_person\_id的一条记录，对出现空值的字段进行同一m\_person\_id内均值填充。

1. 为更好的建模，防止数据分布不平衡带来的模型预测失真问题，对对照组进行数据抽取，首先统计实验组和对照组的性别、年龄分布：

实验组：

性别分布：369位男， 324 位女

年龄分布：<18岁 0位，18 – 29岁 0位，29 -40岁 12位（1.73%），41-65岁 218位（31.37%），〉=65岁 424 位（61%）

对照组：

性别分布：4789 位男，3733 位女

年龄分布：<18岁 3位，18 – 29岁 183位，29 -40岁 1069位，41-65岁 3966位，〉=65岁 1210 位

按照实验组年龄分布的比例，从对照组中选取了976个数据缺失量最少的患者的记录。

7．最终数据集的缺失值率（实验组695，对照组976）：

|  |  |
| --- | --- |
| 字段名称 | 缺失值率 |
| VISITOR\_DATE | 0 |
| CONFIRM\_DATE | 0 |
| DATE\_OF\_BIRTH | 0 |
| SEX\_CODE | 0.000599 |
| RISK\_STRATIFY | 0 |
| SBP | 0 |
| DBP | 0 |
| FOLLOWP\_DATE | 0 |
| SYMPTOM\_CODE | 0.455689 |
| GL | 0.326946 |
| WEIGHT | 0.08024 |
| HEIGHT | 0.281437 |
| BMI | 0.282036 |
| HEART\_RATE\_TIMES | 0.111976 |
| MIND\_ADJ\_CODE | 0.015569 |
| BEBAVIOR\_CODE | 0.013174 |
| FOLLOWP\_TYPE\_CODE | 0.281437 |
| AGE | 0.023952 |

1. 剩余缺失值填充

Diag\_name\_inhos用icd编码替换，所以该列不存在缺失值。

1）症状编码（Symptom\_Name)为离散属性，所以维度扩充，扩充后，对于每一个值，出现该特征的记为1，否则记为0，这样就不存在缺失值的情况。

2）对于一个患者对应的多条记录均为缺失值的情况，体重、身高等特征按照性别分类，取平均值填充。其他列使用k近邻法填充。

1. K最近距离邻法（K-means clustering）

|  |  |  |
| --- | --- | --- |
| 字段名称 | 缺失率 | 填充方法 |
| SEX\_CODE | 0.00059880239521 | 删除 |
| SYMPTOM\_CODE | 0.455688622754 | 不存在，向量化 |
| GL | 0.326946107784 | k近邻（20） |
| WEIGHT | 0.0802395209581 | 均值，按性别计算 |
| HEIGHT | 0.281437125749 | 均值，按性别计算 |
| BMI | 0.282036 |  |
| HEART\_RATE\_TIMES | 0.111976047904 | k近邻（20） |
| MIND\_ADJ\_CODE | 0.0155688622754 | k近邻（20） |
| BEBAVIOR\_CODE | 0.0131736526946 | k近邻（20） |
| FOLLOWP\_TYPE\_CODE | 0.281437125749 | k近邻（20） |
| AGE | 0.0239520958084 | k近邻（20） |

D:\Documents\Tencent Files\1611328786\Image\Group\O6_K~7XIN}MFPB5}60SKF_9.png这是门诊数据icd编码后的表，依次为全体人群、糖尿病人群、高血压人群

D:\Documents\Tencent Files\1611328786\Image\Group\)7FT~09PE]M}3WQ$IT~8ZL8.png这个是门诊icd编码后转化为二维表，0-1格式，同时带有日期列