1. 背景及相关研究（已发现的风险因子及风险模型）:

Framingham研究

1. 数据来源及处理（按照医学类论文的规范组织，重点叙述病人来源、病人纳入及排除标准、分类标准，特征字段的选取和排除标准，人群基本情况统计）:

* 此次心衰数据来源于深圳市各大医院、社康等真实门诊数据， 定位于高血压人群。
* 提取心衰发病之前高血压发病后的人群作为实验组，三年以上未患严重病症的高血压健康人群作为对照组，研究心衰的影响因素。
* 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774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774个对照组作为最终的数据集。

1. 方法简述（预处理、建模（包括特征选择），交叉验证，基线对比方法简述）：
2. 根据数据集，删除无意义或者冗余性特征。
3. 考虑到心衰数据数据量有限，异常值检测并视为缺失值处理。
4. 统计分析数据分布，按照实验组比例，从对照组中选取了774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774个对照组作为最终的数据集。
5. 缺失值处理，均值填充，最近邻填充。
6. 由于特征数量大，为防止过拟合，使用F值检验，保留p值<0.05的特征。
7. 使用逻辑斯特回归模型预测数据，并将实验结果同Framingham评分进行比对。
8. 结论及讨论（模型测试/验证结果，与基线方法的结果比较，主要特征因素及趋势，与相关文献结果的对比，特征因素在全体高血压人群上的分布趋势及与疾病的关系，特征因素和目标疾病之间的时间关系分布等等）

实验结果：

|  |  |  |
| --- | --- | --- |
| 特征 | 模型权重系数，L2正则化的结果 | 特征名 |
| MARITAL\_CODE | 0.05916918 | 婚姻状况 |
| I25\_103 | 0.65169399 | 冠状动脉粥样硬化性心脏病 |
| N18 | 0.62116044 | 慢性肾病 N18 |
| I42 | 0.44352033 | 心肌病 I42 |
| I48\_X01 | 0.42795755 | 心房颤动 I48\_X01 |
| BIT16 | 0.37978592 | 代谢紊乱 |
| BIT38 | 0.33453695 | 肺水肿 |
| BIT35 | 0.3220253 | 肺气肿 其他慢性阻塞性肺病 |
| I50\_906 | 0.31441969 | 心肌损害 I50\_906 |
| BIT30 | 0.30567783 | 风湿性心脏病 BIT30 |
| BIT46 | 0.3012815 | 咳嗽 呼吸异常 咽痛和胸痛 |
| I11\_900 | 0.27338346 | 高血压性心脏病 I11\_900 |
| R60 | 0.27328484 | 水肿，不可归类在他处者 |
| BIT33 | 0.26300687 | 肺炎 |
| BIT20 | 0.24251666 | 其他的基底核变性疾病 张力失常 |
| G45 | 0.20535874 | 短暂性大脑缺血性发作和相关的综合征 |
| RISK\_STRATIFY2 | 0.19891541 | 危险分层 |
| E14\_900 | 0.19726178 | 糖尿病 |
| G47 | 0.19432168 | 睡眠障碍 G47 |
| BIT47 | 0.18810263 | 胃肠气胀及有关情况 大便失禁 肝大和脾大，不可归类在他处者 等等 |
| BIT08 | 0.17998013 | 缺铁性贫血 |
| BIT05 | 0.17912847 | 皮肤癣菌病 |
| BIT02 | 0.17893988 | 梅毒 |
| BIT24 | 0.16912358 | 神经根和神经丛疾患 |
| I48\_X02 | 0.16309198 | 阵发性心房纤颤 I48\_X02 |
| BIT31 | 0.16272638 | 静脉炎和血栓性静脉炎 |
| BIT36 | 0.16007277 | 哮喘 哮喘持续状态 BIT36 |
| I66 | 0.15661406 | 大脑动脉的闭塞和狭窄，未造成脑梗死 I66 |
| I27\_9 | 0.15020322 | 慢性肺源性心脏病 I27\_9 |
| I70\_0 | 0.14963662 | 主动脉的动脉粥样硬化 |
| BIT14 | 0.14817911 | 维生素A缺乏病 |
| I49\_900 | 0.14495528 | 心律失常 I49\_900 |
| BIT41 | 0.14154916 | 化脓性关节炎 |
| E11\_601 | 0.13634294 | 2型糖尿病足病 E11\_601 |
| I77\_6 | 0.13513294 | 未特指的动脉炎 I77\_6 |
| BIT45 | 0.13193571 | 心脏搏动异常 心脏杂音和其他心音 |
| BIT07 | 0.13071184 | 口腔、食管和胃原位癌 |
| BIT48 | 0.12667509 | 嗜眠、木僵和昏迷 |
| R65 | 0.12451452 | 全身炎症反应综合征 R65 |
| R56 | 0.12405482 | 惊厥，不可归类在他处者 |
| R74 | 0.11554616 | 血清酶水平异常 R74 |
| I67\_900 | 0.11342932 | 脑血管病 I67\_900 |
| DBP2 | 0.10771338 | 收缩压 |
| I38 | 0.102684 | 瓣膜未特指的心内膜炎 I38 |
| R09 | 0.10239591 | 累及循环和呼吸系统的其他症状和体征 R09 |
| BIT06 | 0.092859 | 恶性疟原虫疟疾 |
| R10 | 0.0892114 | 腹部和盆腔痛 R10 |
| I70\_9 | 0.08860318 | 全身性和未特指的动脉粥样硬化 I70\_9 |
| I51\_600 | 0.08774981 | 心血管疾病 I51\_600 |
| ABO\_CODE | 0.08513063 | 血型 |
| I27\_0 | 0.07953465 | 原发性肺动脉高压 I27\_0 |
| BIT01 | 0.07592394 | 霍乱 伤寒和副伤寒 其他沙门菌感染 |
| I25\_102 | 0.07547719 | 冠状动脉粥样硬化 I25\_102 |
| I70\_2 | 0.07409942 | 四肢动脉的动脉粥样硬化 |
| BIT27 | 0.07258123 | 中毒性脑病 |
| BIT29 | 0.07192957 | 外耳炎 |
| R54 | 0.06729754 | 衰老 R54 |
| R69 | 0.06502818 | 原因不知和原因未特指的发病 R69 |
| AGE\_NEW | 0.06502474 | 年龄 |
| E16\_803 | 0.05493761 | 代谢综合征 E16\_803 |
| I51\_700 | 0.05121474 | 心脏扩大 I51\_700 |
| R11 | 0.04997942 | 恶心和呕吐 R11 |
| E16\_200 | 0.04931574 | 低血糖症 E16\_200 |
| I25\_901 | 0.0443375 | 冠状动脉性心脏病 I25\_901 |
| BIT09 | 0.03636215 | 播散性血管内凝血[去纤维蛋白综合征] |
| BIT34 | 0.03566581 | 急性支气管炎 急性细支气管炎 |
| I67\_805 | 0.03308385 | 慢性缺血性脑血管病 I67\_805 |
| BIT40 | 0.03280203 | 葡萄球菌性烫伤样皮肤综合征 脓疱病 皮肤脓肿、疖和痈 蜂窝织炎 急性淋巴结炎 藏毛囊肿多形性红斑 结节性红斑 其他红斑性情况 |
| BIT13 | 0.02870023 | 夸希奥科病[恶性营养不良病] 营养性消瘦 消瘦性夸希奥科病 |
| R72 | 0.02364995 | 白细胞异常，不可归类在他处者 R72 |
| I20\_900 | 0.02302899 | 心绞痛 I20\_900 |
| BIT15 | 0.01687336 | 局部多脂症 肥胖症 其他营养过度 营养过度后遗症 BIT15 |
| BIT28 | 0.01331958 | 睑腺炎和睑板腺囊肿 眼睑的其他炎症 眼睑的其他疾患 分类于他处的疾病引起的眼睑疾。 |
| BIT42 | 0.01134754 | 肾炎 |
| I49\_400 | 0.01123961 | 早搏 I49\_400 |
| BIT37 | -0.00134117 | 支气管扩张(症) 煤炭工肺尘埃沉着病 石棉和其他矿物纤维引起的肺尘埃沉着病 |
| BIT44 | -0.00313107 | 无脑儿和类似畸形先天性肌肉骨骼变形 |
| BIT39 | -0.00812819 | 口腔、涎腺和颌疾病 埋伏牙和阻生牙 龋(牙) 牙齿硬组织的其他疾病及口腔粘膜的其他疾病 舌疾病 食管炎 胃-食管反流性疾病 |
| BIT04 | -0.00930346 | 疱疹病毒[单纯疱疹]感染 水痘 带状疱疹 天花 猴痘 麻疹 风疹 |
| BIT11 | -0.01960997 | 甲状旁腺功能减退症 甲状旁腺功能亢进症和甲状旁腺的其他疾患 |
| R53 | -0.02124408 | 不适和疲劳 R53 |
| I67\_2 | -0.02216529 | 大脑动脉粥样硬化 I67\_2 |
| E11\_401 | -0.02716066 | 2型糖尿病性周围神经病 E11\_401 |
| BIT10 | -0.03352078 | 先天性碘缺乏综合征 碘缺乏相关性甲状腺疾患和有关情况 |
| G90 | -0.03931451 | 自主神经系统的疾患 G90 |
| R50 | -0.04554916 | 其他和原因不明的发热 R50 |
| HEART\_RATE\_TIMES | -0.05905261 | 心率 |
| I51\_900 | -0.0871607 | 心脏病 I51\_900 |
| I67\_803 | -0.09483765 | 脑动脉供血不足 I67\_803 |
| BMI | -0.09750807 | 身高体重指数 |
| BIT17 | -0.10640502 | 脑部疾病、损害和功能障碍引起的人格和行为障碍 |
| LIMBS\_NUMB | -0.11126914 | 四肢发麻 |
| BIT32 | -0.13082035 | 急性鼻咽炎[感冒] 急性鼻窦炎 急性咽炎 急性扁桃体炎 急性喉炎和气管炎 |
| NEW\_DISEASE | -0.46810102 | 发病时间 |
| DIZZINESS | -0.81221313 | 头痛头晕 |

>保留心衰发病之前的特征，将发病后的特征置为0，随机选取对照组，运用逻辑回归模型得到的结果：

>训练集auc 0.87

>测试集auc 0.77

特征数据在Framingham表现：

男性每位患者的得分，按照对照组、对照组类别排序，从下图可以看出前半部分（对照组）得分总体没有后半部分（实验组）得分高，女性得分同理。

男性：

Auc: 0.65

女性：

Auc:0.69