|  |
| --- |
| **机器学习** |
| **法国巴黎银行心脏病索赔管理** |
| 主 研 人：刘一阳  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/04/19 | A | 初稿 | 刘一阳 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/04/19 | 法国巴黎银行索赔管理（BNP Paribas Cardif Claims Management） | 作为全球个人保险专家，法国巴黎银行卡迪夫为36个国家的9000万客户提供服务。Kagglers面临着根据流程早期可用功能预测索赔类别的挑战，帮助银行加快索赔流程，从而为客户提供更好的服务。 | 分类 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 5](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 5](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 6](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 6](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 6](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一数据预处理及特征工程部分方案 6](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案 6](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 6](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 6](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 7](#_Toc29922_WPSOffice_Level2)

[… 8](#_Toc4787_WPSOffice_Level2)

[4. 算法比较 8](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 8](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 8](#_Toc16174_WPSOffice_Level1)

[5.1 总结 8](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 8](#_Toc16619_WPSOffice_Level2)

1. 背景描述

法国巴黎银行在全球85个国家和地区拥有700多个办事处，20.5万名雇员中有2.8万名从事国际银行业务，为100家国际大型企业中的80多家提供服务；法国巴黎银行在全球范围从事出口信贷方面的业务。2018年7月19日，《财富》世界500强排行榜发布，法国巴黎银行位列44位。 [1] 2018年12月，巴黎银行入围2018世界品牌500强。

**1.1 竞赛赛题描述**

作为全球个人保险专家，法国巴黎银行卡迪夫为欧洲，亚洲和拉丁美洲36个国家的9000万客户提供服务。

在一个由新用途和生活方式的出现塑造的世界中，一切都变得越来越快。当遇到意外事件时，客户希望他们的保险公司尽快为他们提供支持。但是，索赔管理可能需要不同级别的支票才能批准索赔并且可以付款。随着数字经济产生的新实践和行为，这一过程需要通过数据科学进行调整，以满足客户的新需求和期望。

在这一挑战中，法国巴黎银行卡迪夫提供了一个匿名数据库，其中包含两类索赔。Kagglers面临着根据流程早期可用功能预测索赔类别的挑战，帮助法国巴黎银行加快索赔流程，从而为客户提供更好的服务。

<https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/overview/description>

**1.2 评估指标描述**

竞赛中评估模型优劣的指标。

选手的提交的评估指标是 Log Loss

N ：观测次数

Log：自然对数

：二进制目标

：预测的 = 1 的可能性

注意：实际提交的预测的可能性会被替代

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

描述大赛数据的来源

大赛数据提供法国巴黎银行，为大赛提供历史销售数据，从软硬件环境诸多方面提供大赛支撑。

这里是数据的超链接。

链接：https://pan.baidu.com/s/1C8O6UG78fAnhg0chJ2U-qQ

提取码：0z07

**2.2 数据的描述性统计**

* + 1. **数据基本情况描述：**

您将获得一个匿名数据集，其中包含法国巴黎银行收到索赔时可用的分类和数字变量。所有字符串类型变量都是分类的。没有序数变量。

列车组中的“目标”列是要预测的变量。对于适合加速批准的索赔，它等于1。

任务是预测测试集中每个声明的概率（“PredictedProb”）。

* + 1. **数据字段介绍：**

**文件说明：**

train.csv – 包含目标的训练集

test.csv – 没有目标测试集

sample\_submission.csv – 格式正确的示例提交文件

**表2-1 train.csv数据表字段介绍**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| Id | 唯一Id字段 | 连续 | 0% |
| Target | **目标的分类** | **离散** | **0%** |
| V1~V131 | **变量** | **离散，连续都有** |  |

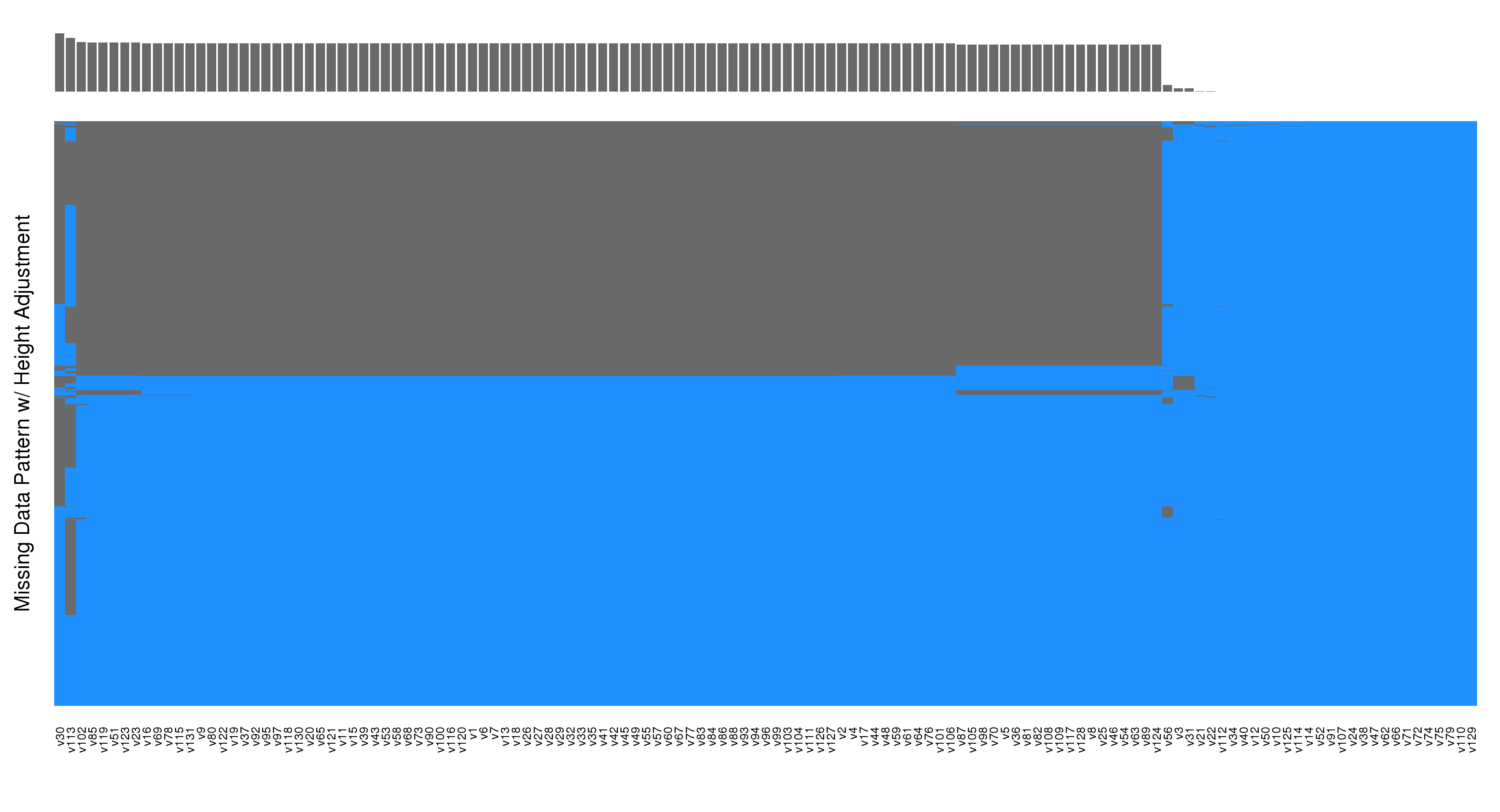
**表2-2 test.csv数据表字段介绍**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| Id | 唯一Id字段 | 连续 | 0% |
| Target | **目标的分类** | **离散** | **0%** |
| V1~V131 | **变量** | **离散，连续都有** |  |

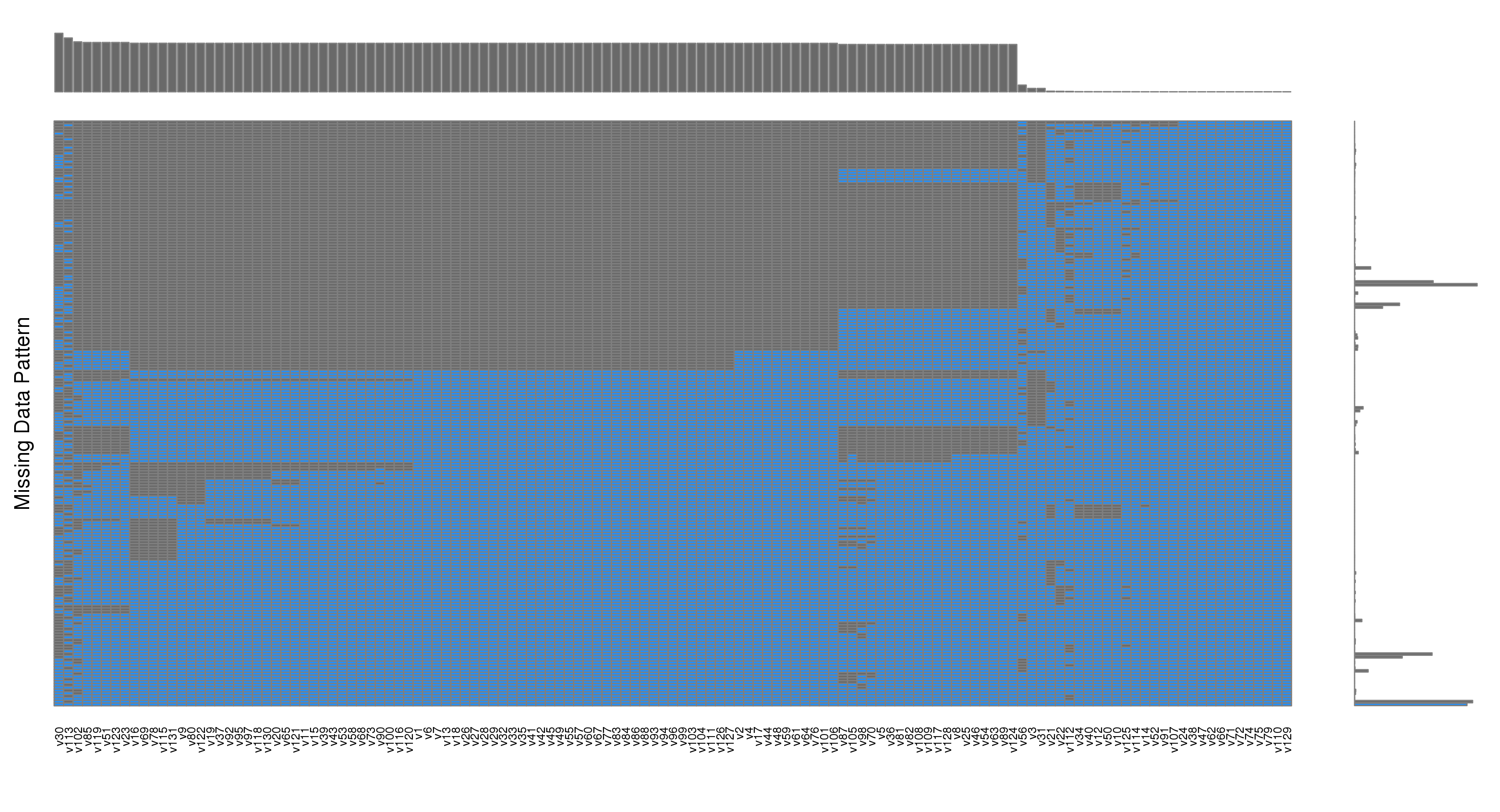
**表2-5 sample\_submission.csv数据表字段介绍**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| Id | 唯一Id字段 | 连续 | 0% |
| PredictedProb | 预测的可能性 | 连续 | 0% |

* + 1. **数据描述性统计**



**图2-2 缺失数据统计**



**图2-2 缺失数据统计**

3. 优秀算法思路

**3.1 方案一**

**3.1.1** 方案一数据预处理及特征工程部分方案

很明显，这场比赛并不是关于集合而是特色工程，我们很高兴我们发现了早期赢得这场比赛的主要关键。

因此，我们有充足的时间从最大限度地利用我们的发现并保持最高位置。

诀窍非常简单，我甚至在竞争初期就提到了一个暗示（当时没有我理解它的重要性）：<https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/forums/t/19240/analysis-of-duplicate-variables-correlated-variables-large-post/110095#post110095>

我们认为，数据匿名化非常重要，最后我们能够找出一些变量的含义：

v40 - 观察日期

v40-v50 - 合同开始日期

v50 - 合同开始后的天数（换句话说，索赔日期）

v10 - 以月为单位的合同期限

合同剩余天数的其他几个：

V12 = V10 \*（365.25）/ 12-V50

V34 = V10 \*（365.25）/ 12-V40

现在将这些知识添加到假设v22是客户并且v56 / v113是产品类型的假设中，

您会看到非常明显的模式，一个合同在合同生命期内可能有多个索赔（R: group\_by(v22,v40-v50) %>% arrange(v50)）：

[v22,v40-v50] [v50] [target]

[ZLS 12840] 197 1

[ZLS 12840] 962 1

[ZLS 12840] 1437 1

[ZLS 12840] 1498 NA

[ZLS 12840] 1501 1

[ZLS 12840] 1726 1

[ZLS 12840] 1788 NA

[ZLS 12840] 1882 NA

[ZLS 12840] 2418 1

[ZLS 12840] 3352 NA

[ZLS 12840] 3370 NA

[ZLS 12840] 3909 NA

因此，我们可以删除分类{v22}级别iid假设，并转到面板数据结构化分类级别，例如：

{v22,v40-v50} => sort(v50), {v22,v40-v50,v56} => sort(v50), {v22} => sort(v40)等;

只需以这种方式查看数据，很明显目标值对于每个{v22，v40-v50}级别都是非常持久的 -

即如果时间序列以target = 1开始，则通常以target = 1结束; 如果以target = 0开头，它通常以target = 0结束; 声称目标转移0-> 1是非常罕见的（1-> 0只有少数情况）;

所以最终这一切都导致正确地为每个级别输入目标序列 - 我们使用滞后（目标）和领导（目标）变量;

令我们惊讶的是，这些变量并没有过度拟合LB，最终我们制作了如此多的滞后/导致变量，有可能放弃v22列，

甚至没有使用论坛中讨论的v22平均目标技术，这可能很多顶级球队无论如何都做了。

**3.1.2** 方案一模型设计、建立部分方案

模型设计：

达武特：我们得到了不同的特征集，并训练了各种各样的模型，如knn，额外的树分类器，随机森林和神经网络。我们还在XGBoost中尝试了不同的目标。

Song：在我们的最终模型中，我们将XGBoost作为一个集合模型，其中包括20个XGBoost模型，5个随机森林，6个随机决策树模型，3个正规化贪婪森林，3个逻辑回归模型，5个ANN模型，3个弹性网模型和1个SVM模型。

部分方案：

首先，由于给定数据集中的每个特征都被缩放并且引入了一些随机噪声，我认为识别如何处理噪声和解除数据扩展可能很重要。我想到了一种简单但快速的方法来检测整数类型特征的缩放因子。花了一些时间，但最终它是我们获胜解决方案的关键部分。

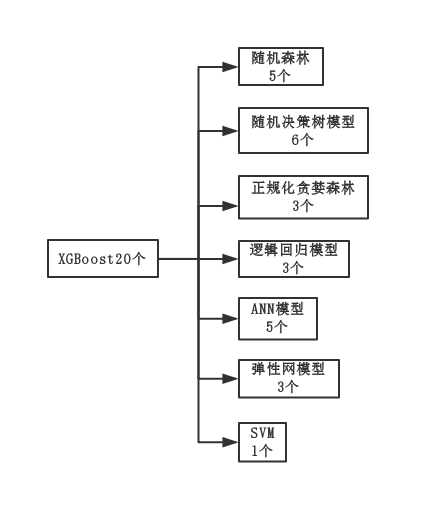
其次，考虑到我们对变量意义的假设，我们构建了有效的特征交互。除了其他想法之外，我们根据我们处理面板数据的印象设计了滞后和引导功能。最后，我们对面板数据和变量含义的假设是不现实的（实际上，这意味着同一个客户可能面临数百或数千个索赔）。但是，我们的滞后和主要功能确实为我们的解决方案带来了重要价值，这当然是因为它是编码交互的有效方式。这与其他前两个团队的解决方案一致，这也得益于编码v22和其他变量之间的一些交互，除了滞后和导致之外还有不同的方法。

**3.1.3** 方案一结果、排名等

最后结果：0.42193

排名：1/2926

**3.1.4** 方案一算法流程图



**图3-1 流程图**

<http://blog.kaggle.com/2016/05/13/bnp-paribas-cardif-claims-management-winners-interview-1st-place-team-dexters-lab-darius-davut-song/>

4. 算法比较

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **0.42193** | **Word2vec,TFIDF, LSA** | **Xgboost，梯度增强，随机森林，SVR，extratrees。** | **Sklearn, xsboost** |

制胜之处还是在参赛者的特征工程，排名第一的参赛者显然做了更加独特的特征工程。

5. 总结与展望

**5.1 总结**

这个竞赛题目本身其实具有一定的特殊性，其针对的是银行的环境下，而第一名的团队在前期的准备过程中吸收了一位从事相关工作的职员因而具有较大的优势。因此在将机器学习应用到实际应用领域时，将其与专业知识的结合尤为关键。

在收集整理的过程中也学习到了这些优秀参赛者的一些方法和经验，为今后的学习和工作提供了宝贵的经验。

**5.2 建模思路**

在学习了选手的思路后，我倾向于采用第一名 的方案，对数据进行尽可能详细的分析。