Inductrial Accident Causal Analysis

陈佳禾，曹栋承，邵嘉豪

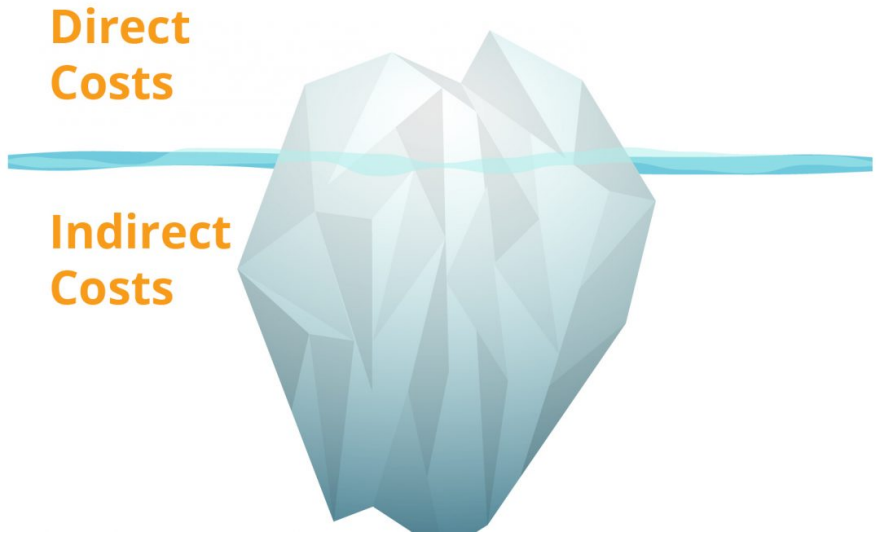
ZJUCSE

Oct. 26, 2022

Background

工业安全事故

工业生产过程中，造成操作人员受伤的事故的总称



- 直接损失
  - 人员伤亡、医疗费用
  - 停工期间工资、新雇员工工资
- 间接损失
  - 劳动仲裁机构罚款、企业信誉降低
  - 工作人员花费的额外行政、公关时间

Dataset

数据由巴西工业巨头IHM Stefanini提供，该公司的业务包含造纸、采矿、冶炼、石油工业等, 其提供的数据来源于 3 个国家的 12 个生产基地，共记 439 条.

- 数据条目

数据条目	相关信息
Date	事故发生时间
Countries	事故发生国家
Local	事故发生城市
Industri sector	事故发生部门
Accident level	事故等级：事故严重程度
Potential accident level	潜在事故等级：事故可能的严重程度
Genre	发生事故人员的性别
Employee or Third Party	正式工/临时工
Critical Risk	事故的简短描述
Description	事故的细节描述

- 目标
  从事故数据中**解析**出有效信息，并**应用**于生产中，降低损失

Pre-Processing NLP

数据分析

- 原始数据

原始类别	Date	Countries	Local···	Description

- 明确目标 “**Accident Level（事故等级）**”**与其余数据关系**
- 数据划分

划分种类	客观属性	文本
事故等级		

- 处理方法
  - 客观属性 -EDA- 事故等级
  - 文本 -NLP- 事故类别 - 事故等级

Part2： 数据解析

客观属性变量预处理

- **time related**
  - add datetime features: ‘year’ ‘month’ and ‘day’
- **temprature related**
  - add ‘season’ feature
- **workday related**
  - add workday feature:’weekday”weekofyear’
- **处理结果**

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description	Year	Month	Day	Weekday	WeekofYear
0	2016-01-01	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...	2016	1	1	Friday	53
1	2016-01-02	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...	2016	1	2	Saturday	53
2	2016-01-06	Country_01	Local_03	Mining	I	III	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...	2016	1	6	Wednesday	1

Part2：数据解析

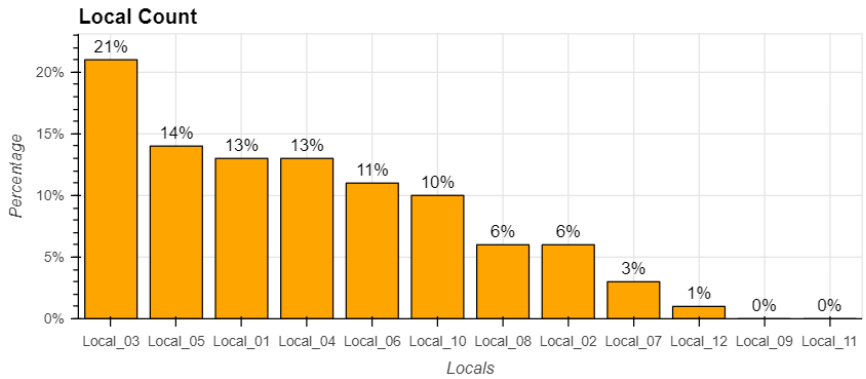
EDA:

Univariate Analysis

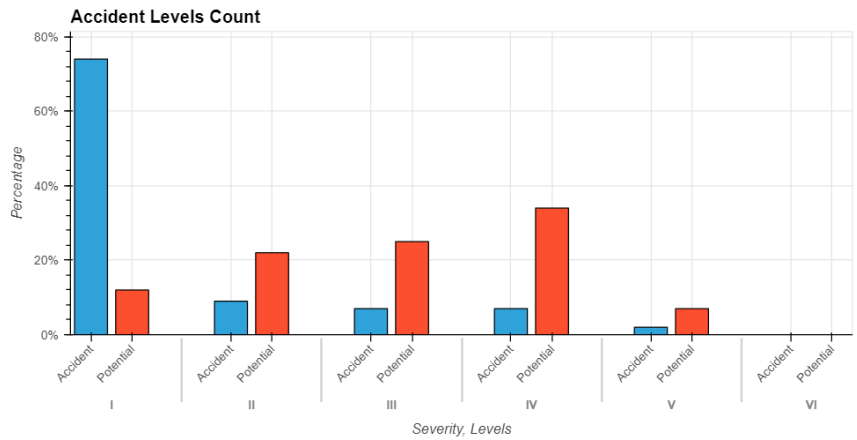
- 分析变量

国家	地区	工业部门	性别	雇员类型	主要风险项
时间	季节				

- 可视化展示：
  - 地区



  - 事故等级

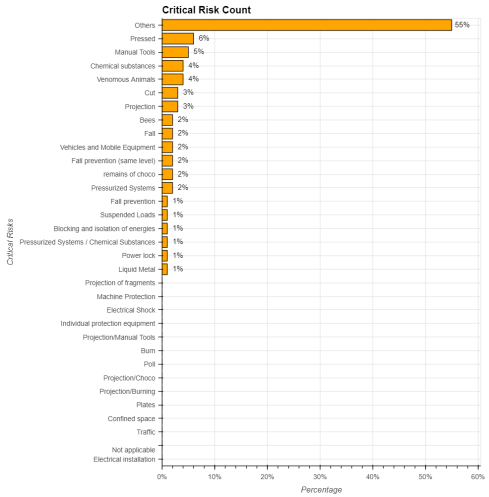


Part2： 数据解析

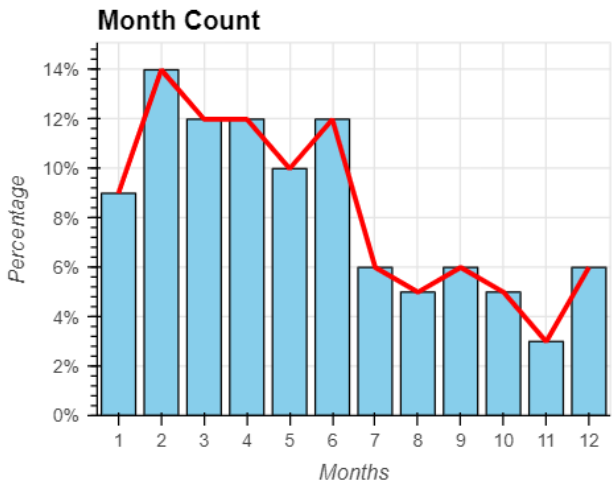
EDA:

Univariate Analysis

- 可视化展示：
  - 主要风险项



  - 月份



Part2： 数据解析

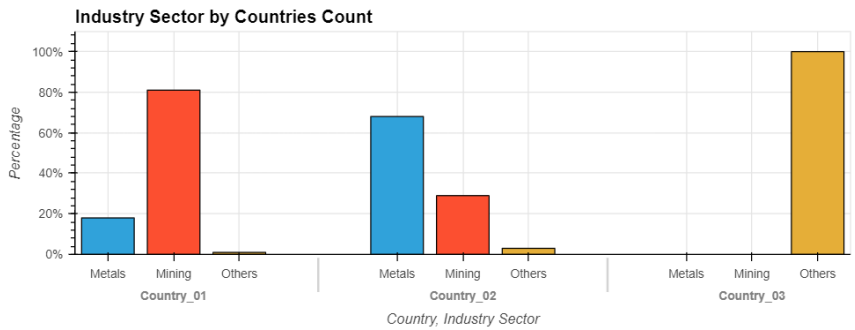
EDA:

Multivariate Analysis

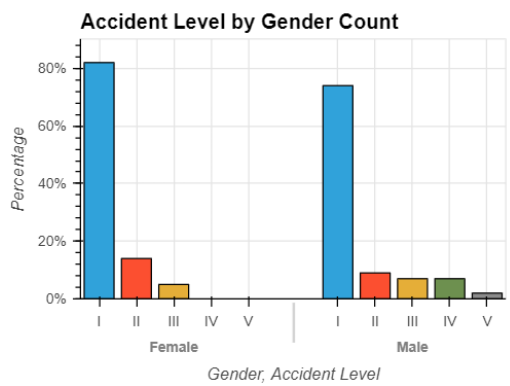
- 变量间关系

工业部门 by 国家	雇员类型 by 性别	工业部门 by 性别	事故等级 by
事故等级 by 雇员类型	事故等级 by 月份	事故等级 by 季节	

- 可视化展示：
  - 工业部门 by 国家



  - 事故等级 by 性别

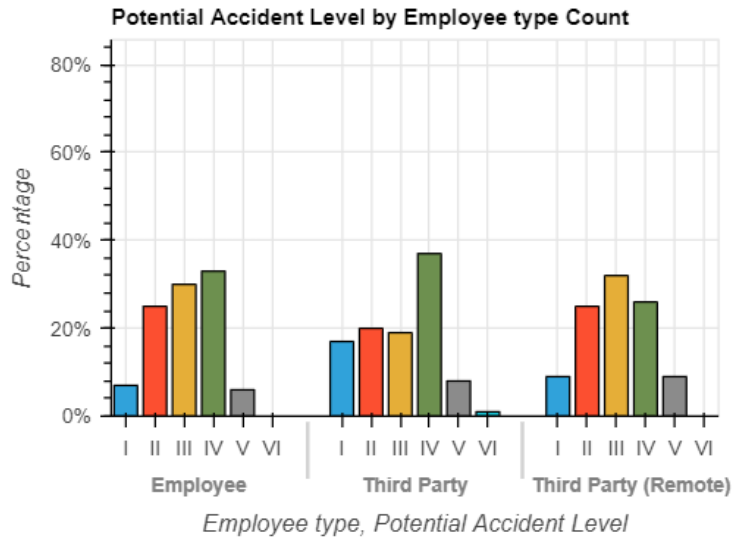


Part2： 数据解析

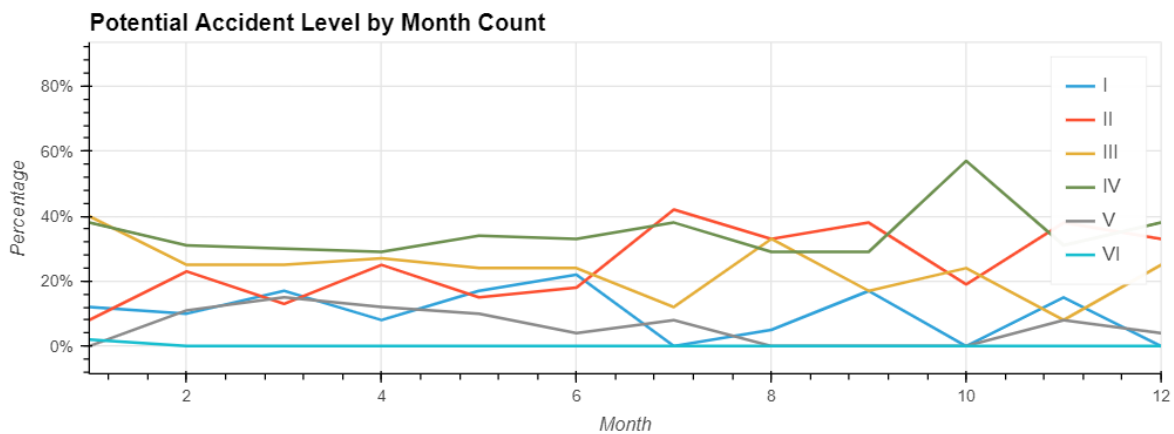
EDA:

Univariate Analysis

- 可视化展示：
  - 潜在事故等级 by 雇员类型



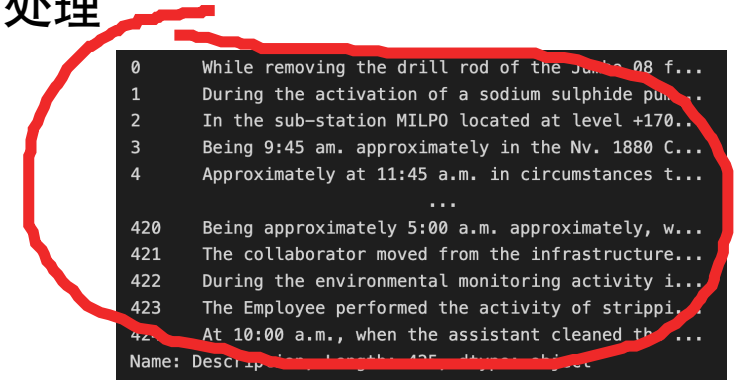
  - 潜在事故等级 by 月份



NLP

自然语言处理

- 为什么需要 NLP 处理



- 停用词
  - 在某些特定的 NLP 处理任务中，一些词语不能提供有价值的信息，例如一些助词、标点等等，
  - 会使得问题搜索的复杂度提高

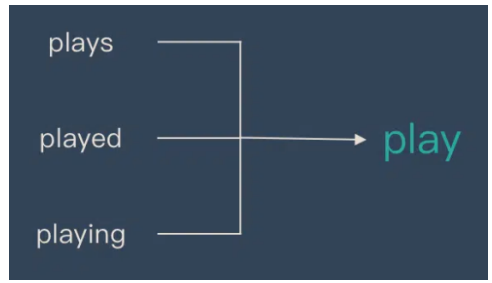
While removing the drill rod of the Jumbo 08 for maintenance, the supervisor proceeds to loosen the support of the intermediate centralizer to facilitate the removal, seeing this the mechanic supports one end on the drill of the equipment to pull with both hands the bar and accelerate the removal from this, at this moment the bar slides from its point of support and tightens the fingers of the mechanic between the drilling bar and the beam of the jumbo.

- 构建停用词表
  - 采用 wordcloud 库中的 STOPWORDS 作为停用词表

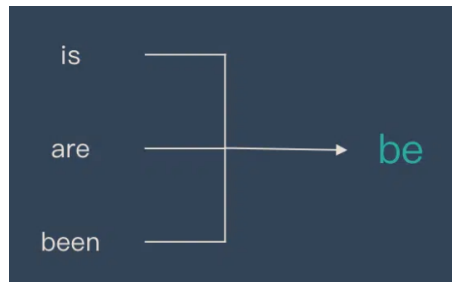
NLP

自然语言处理

- 处理步骤
  - 分词
    - 简单的英文分词，为后面的处理做准备
  - 词干提取



  - 词形还原



  - 对词干提取、词形还原之后的词若为停用词则丢弃
  - 删除非字母单词

NLP

自然语言处理

- 处理结果
  - 处理之前

```
'While removing the drill rod of the Jumbo 08 for maintenance, the supervisor
the support of the intermediate centralizer to facilitate the removal, seeing
supports one end on the drill of the equipment to pull with both hands the bar
removal from this, at this moment the bar slides from its point of support and
fingers of the mechanic between the drilling bar and the beam of the jumbo.'
```

  - 处理之后

```
'remov drill rod jumbo mainten supervisor proce loosen su
see mechan support one end drill equip pull hand bar acce
tighten finger mechan drill bar beam jumbo'
```

Feature Engineering

特征工程

- 利用 TFIDF 算法提取重要的词语
  - TFIDF 算法
    - 字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降
    - 词频 $TF = \frac{\text{在某类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$
    - 逆向文件频率 $IDF = log(\frac{\text{语料库的文档总数}}{\text{包含词条 w 的文档数}+1})$
    - $TF - IDF = TF \times IDF$
  - 提取结果

	TFIDF_accid	TFIDF_finger	TFIDF_left	TFIDF_assist
0	0.0	0.203563	0.000000	0.000000
1	0.0	0.000000	0.000000	0.000000
2	0.0	0.000000	0.209264	0.000000
3	0.0	0.000000	0.268575	0.390256
4	0.0	0.000000	0.000000	0.000000

```
'TFIDF_accid', 'TFIDF_activ', 'TFIDF_area', 'TFIDF_assist',
'TFIDF_caus', 'TFIDF_collabor', 'TFIDF_cut', 'TFIDF_employe',
'TFIDF_equip', 'TFIDF_fall', 'TFIDF_finger', 'TFIDF_floor',
'TFIDF_hand', 'TFIDF_hit', 'TFIDF_injuri', 'TFIDF_left', 'TFIDF_level',
'TFIDF_mesh', 'TFIDF_moment', 'TFIDF_oper', 'TFIDF_perform',
'TFIDF_pipe', 'TFIDF_place', 'TFIDF_remov', 'TFIDF_right',
'TFIDF_support', 'TFIDF_time', 'TFIDF_use', 'TFIDF_work',
'TFIDF_worker', 'TFIDF_accid employe', 'TFIDF_allerg reaction',
'TFIDF_caus cut', 'TFIDF_caus injuri', 'TFIDF_circumst worker',
'TFIDF_describ injuri', 'TFIDF_employe perform', 'TFIDF_employe repor',
'TFIDF_employe use', 'TFIDF_finger left', 'TFIDF_finger right',
'TFIDF_fragment rock', 'TFIDF_gener injuri', 'TFIDF_hand caus',
'TFIDF_hit right', 'TFIDF_injur person', 'TFIDF_injuri describ',
'TFIDF_injuri time', 'TFIDF_left foot', 'TFIDF_left hand',
'TFIDF_medic center', 'TFIDF_right hand', 'TFIDF_right leg',
'TFIDF_safeti glove', 'TFIDF_support mesh',
'TFIDF_time event', 'TFIDF_took bite', 'TFIDF_use safeti',
'TFIDF_wear safeti'],
```

Feature Engineering

特征工程

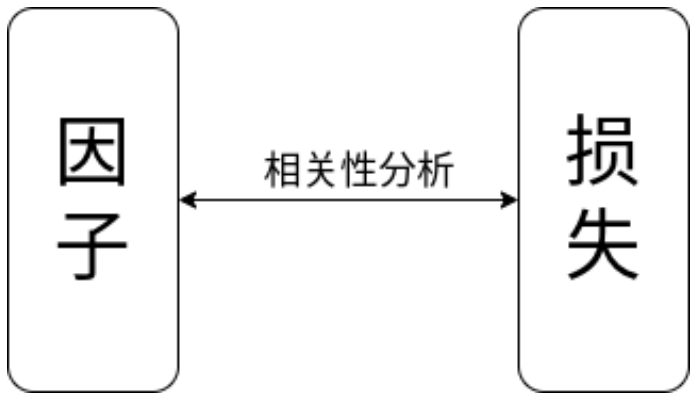
- 事故类别归类
  归类为'Accident_left hand', 'Accident_right hand', 'Accident_right leg', 'Accident_left foot', 'Accident_being cut', 'Accident_being hit', 'Accident_falling', 'Accident_fragment rock'
- 特征编码
  - 事故类别，不关心提取出的重要性，只需 0-1 编码
  - 客观因素，直接编码

	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Year	Month	...	Season	D
0	0	0	1	0	3	1	1	20	2016	1	...	2	
1	1	1	1	0	3	1	0	21	2016	1	...	2	
2	0	2	1	0	2	1	2	15	2016	1	...	2	

Application
应用与指导

目标：降低工厂损失
方法：找出与损失相关性大的因子，针对性提出建议

- 任务
  - 定义”损失”
    从为数不多的数据中，怎样定义损失？
  - 定义”因子”
    哪些方面可能与损失有关？
  - 定义”相关性”
    怎样衡量”因子”与”损失”的相关程度？
  - 找出”与相关性大的因子”，进行分析



Loss Function

定义” 损失”

- 怎样衡量” 损失” ?

  事故等级、潜在事故等级、发生概率

  - 损失函数

$$L_i = [\lambda_1 \cdot V_i(\text{事故等级}] + \lambda_2 \cdot V_i(\text{潜在事故等级})) \cdot P_i(\text{发生概率})$$

  - 统计近似

$$P_i(\text{发生概率}) \approx P(\text{该等级事故发生概率}) \approx \frac{\text{该等级事故发生次数}}{\text{总事故次数}}$$

- 将损失分级

Loss Value	(0.08, 0.34]	(0.34, 0.59]	(0.59, 0.84]	(0.84, 1.09]	(1.09, 1.34]
Loss Level	I	II	III	IV	V

Factors

定义” 因子”

- 客观因素

数据条目	相关信息
Date	事故发生时间
Countries	事故发生国家
Local	事故发生城市
Industri sector	事故发生部门
Genre	发生事故人员的性别
Employee or Third Party	正式工/临时工

- 事故类别
  - 按受伤部位分

Accident_Part	Accident_left leg	Accident_right leg	Accident_left hand	Accident_right hand
Description	左腿	右腿	左手	右手

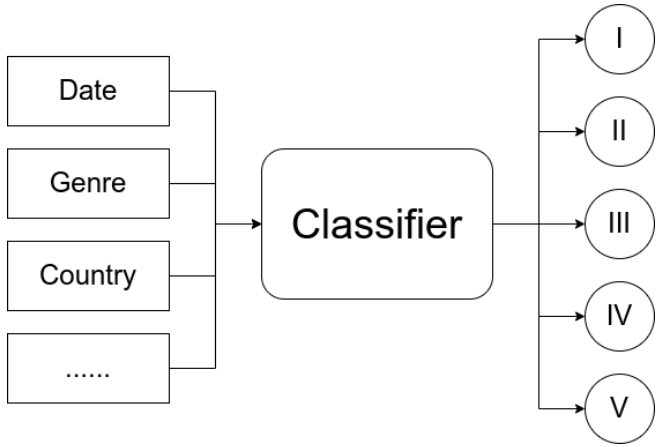
  - 按受伤类型分

Accident_Type	Accident_falling	Accident_being cut	Accident_being hit
Description	坠落	切伤	撞伤

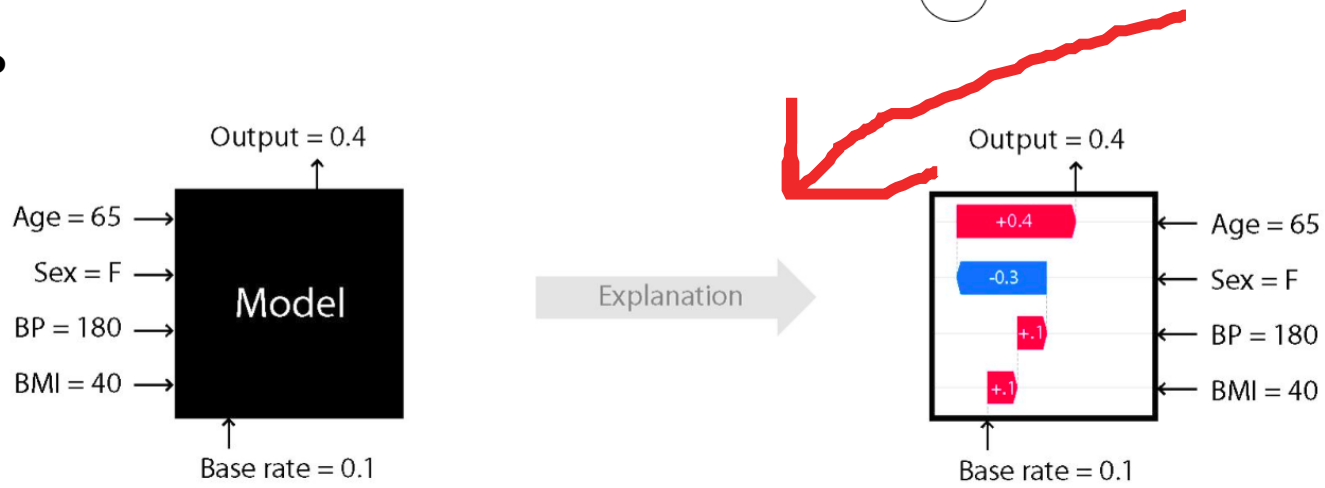
Relevance

定义” 相关性”

- 多分类



- SHAP

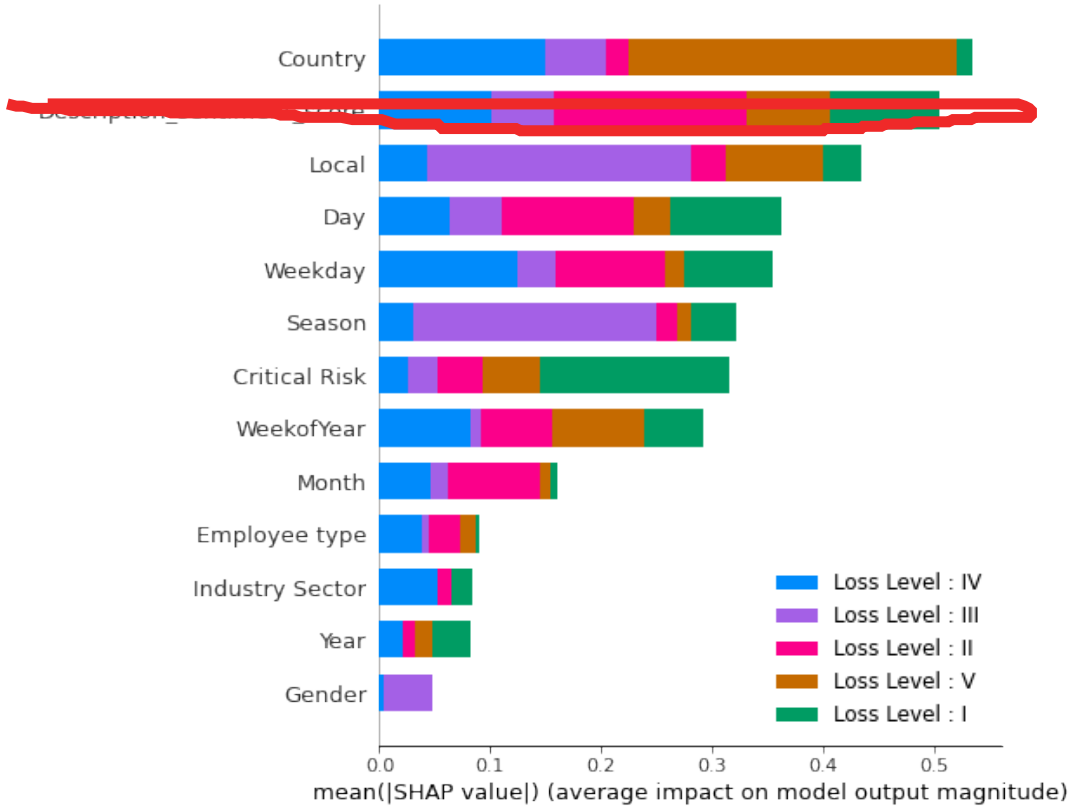


为 ML 算法提供可解释性

Results

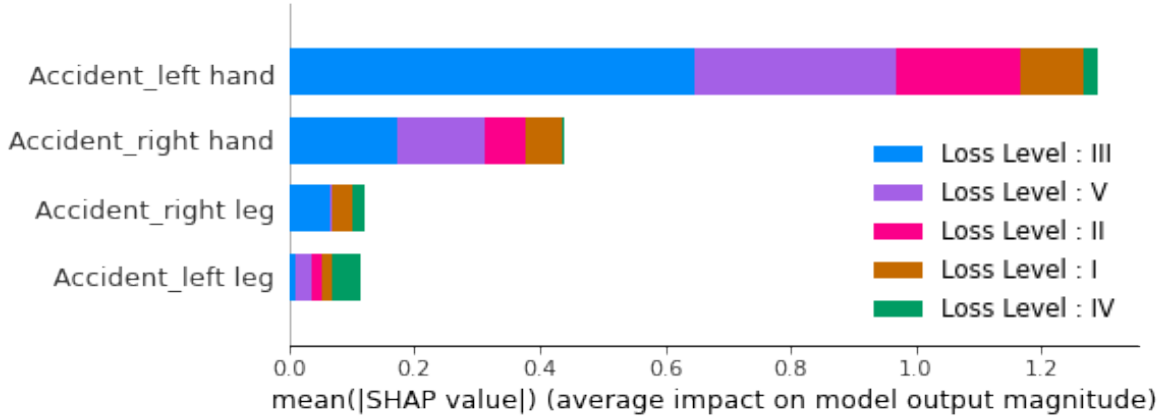
将因子分成 3 组，使用随机森林算法对损失进行多分类

- 客观因素

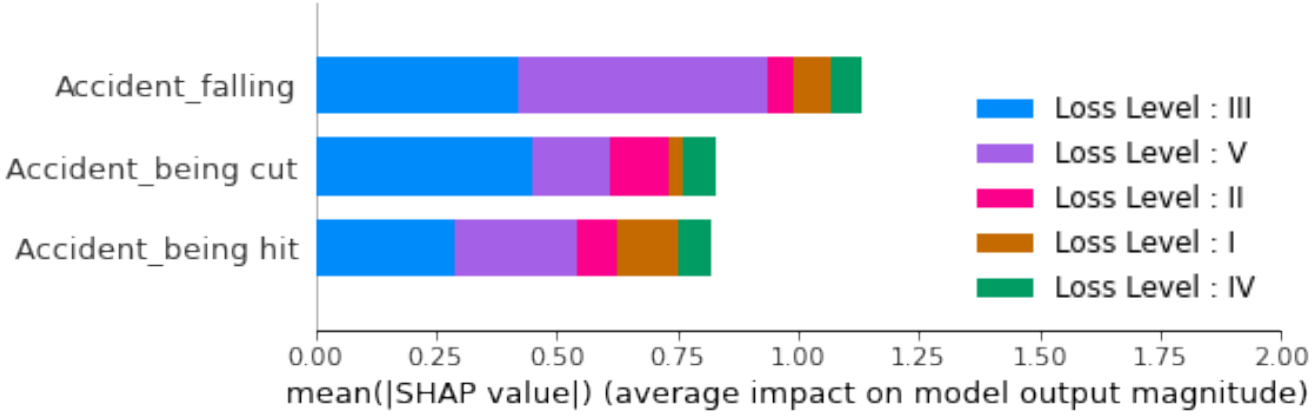


Results

- 受伤部位

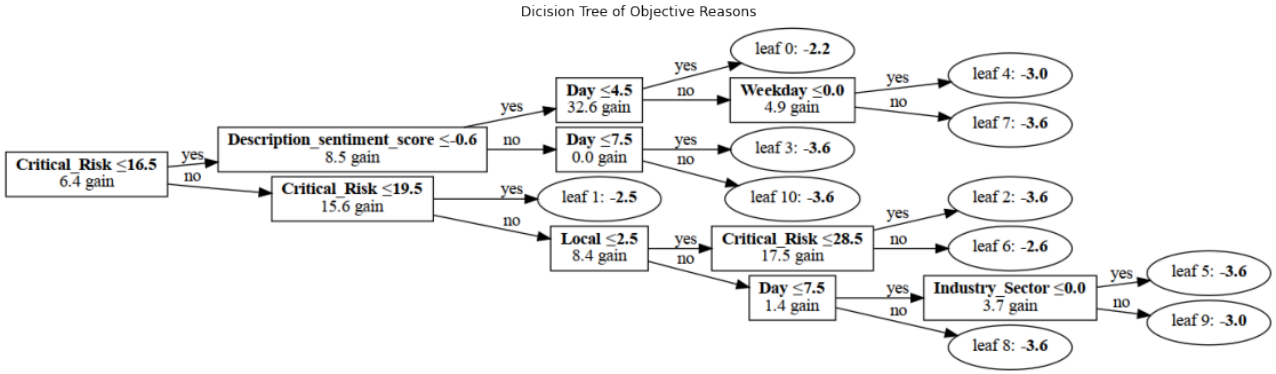


- 受伤类型

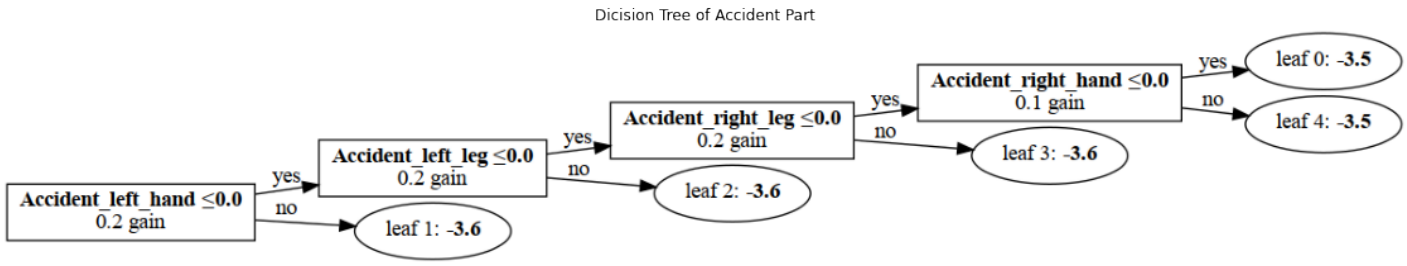


Results

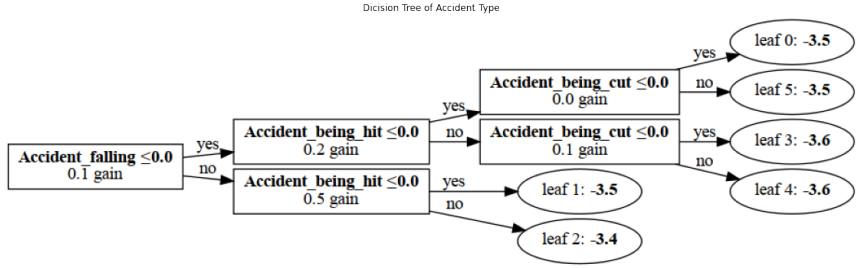
- 客观因素



- 受伤部位



- 受伤类型



Suggestion

建议