Kaggle竞赛：汽车保险司机索赔行为预测

1. 项目介绍

Porto Seguro保险公司希望借助机器学习模型进行风险识别，更准确地预测投保司机下一年发起汽车保险索赔的概率，从而调整汽车保险的定价策略，以此降低赔付风险并扩大客户市场。项目地址

基于XGBoost模型的驾驶员安全索赔预测（金牌）

数据预处理；基于业务需求和逻辑，通过Python对100万条原数据进行离群值、缺失值处理。

数据分析：基于不平衡数据重采样方法实现数据重构，建立元数据框架，对驾驶员57个匿名特征进行数据分析&挖掘。

特征工程：通过特征交互、合并、增强、选取精英特征，缩减特征池规模（57维降至30维），极大节约高维特征研究、维护成本。

数据建模：基于XGBoost模型，经过参数调节、优化迭代进行分类预测，验证集评估指标达到0.285（0.5为最佳，竞赛最佳分数为0.286）。

Porto-Seguro Safe Driver Prediction With XGBoost (Gold Medal)

Wrangled data from Kaggle to create integrated insights by strategic planning that can be used to influence business decisions.

Established metadata framework to perform data-mining, EDA, and feature Engineering of anonymous feature extraction (from 57 to 23)

Developed tuned XGBoost classifier with 10-fold cv to estimate the probability of claim initialization, deriving the evaluation score: 0.285 (Second highest)

1.1 项目背景

1. Porto Seguro保险公司的产品线（针对个人/企业/财产/车辆）
2. Porto Seguro保险公司的汽车保险介绍（基本内容/个性化定制）
3. 《2019年中国保险行业智能风控白皮书》
4. 保险风控是什么?
5. 风控模型-WOE与IV指标的深入理解应用
6. 一文带你了解风控评分卡模型
7. 3分钟搞明白信用评分卡模型&模型验证
8. 模型评价指标-KS、ROC
9. 风控模型—群体稳定性指标(PSI)深入理解应用

2. 项目目标

2.1 核心问题

1. 对于保险公司
   a. 面临问题：识别客户索赔风险的准确度低，对于客户的用户画像不明晰，导致在现有定价策略下，保险公司的赔付损失大，且客户粘性低、获取难；
   b. 预期收益：
      i. 基于模型得到客户的关键特征，节约客户研究成本，有利于快速评估新用户索赔风险；
      ii. 根据客户信息评估索赔风险，以提供准确且合理的定价方案，降低公司赔付损失的同时提高客户体验，有利于减少客户流失，开拓客户市场。
2. 对于客户
   a. 面临问题：服务定价固化或不合理，投保或续保的意愿不高；
   b. 预期收益：客户可享受合理定价后的定制化方案，减少不必要的支出，获得性价比更高的服务内容。

3. 数据介绍

1. **train.csv**
   a. 数据量级：
      i. 595212 rows × 59 columns
2. **test.csv**
   a. 数据量级：
      i. 892816 rows × 58 columns

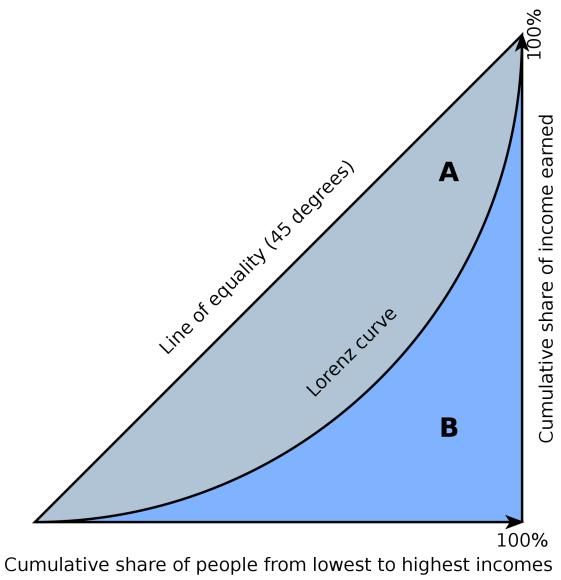
b. 数据类型：
   i. int64 - 49个 （包含id、target）
   ii. float64 - 10个
c. 序号： `id` - 用户id
d. 标签： `target` - 1索赔，0未索赔
e. 特征（×57）：
   i. `ps_ind_xx` - 个人信息（×18）
   ii. `ps_reg_xx` - 地区信息（×3）
   iii. `ps_car_xx` - 车辆信息（×16）
   iv. `ps_calc_xx` - 计算信息（×20）
f. 特征类型：
   i. Binary（后缀含bin） - 17个
   ii. Categorical（后缀含cat） - 14个
   iii. Continuous / Ordinal - 26个
g. 注意事项：
   i. 缺失值用-1表示

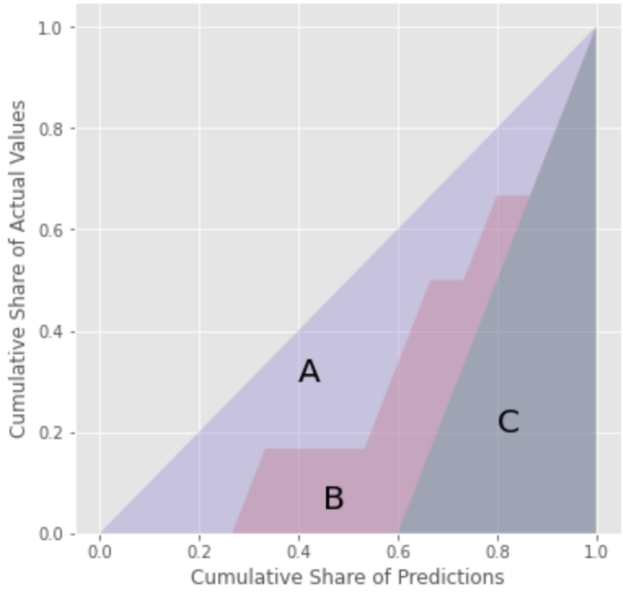
b. 数据类型：
   i. int64 - 48个（包含id、target）
   ii. float64 - 10个
c. 序号： `id` - 用户id
d. 无标签
e. 特征（×57）：
   i. `ps_ind_xx` - 个人信息（×18）
   ii. `ps_reg_xx` - 地区信息（×3）
   iii. `ps_car_xx` - 车辆信息（×16）
   iv. `ps_calc_xx` - 计算信息（×20）
f. 特征类型：
   i. Binary（后缀含bin） - 17个
   ii. Categorical（后缀含cat） - 14个
   iii. Continuous / Ordinal - 26个
g. 注意事项：
   i. 缺失值用-1表示

4. 评估方法

4.1 评估指标（Scoring Metric）

a. 模型的预测结果通过归一化基尼系数（Normalized Gini Coefficient）来评估；
b. 基尼系数（Gini coefficient），是20世纪初意大利学者科拉多·基尼根据洛伦兹曲线所定义的判断年收入分配公平程度的指标，是比例数值，在0和1之间；
c. 本项目中，基尼系数表明该模型在区分 "坏 "司机和 "好 "司机方面的有效性，前者将来会索赔，后者将来不会索赔。
d. 基尼系数的范围从0到0.5，归一化的基尼系数用理论最大值来调整分数，使最大的分数为1。





4.2 具体原理实现

a. 将prediction的值从最小到最大排序，并对应调整actual的顺序；

b. 对调整排序后的actual进行累加并绘制累加折线图（即洛伦兹曲线）；

c. 将 x 和 y 轴进行归一化（即转换范围在0到1之间）并绘制45°对角线；

d. 计算A的面积得到基尼系数a；

e. 理论上当prediction与actual完全匹配时，可以得到最大基尼系数；

f. 因此将actual的值从小到大排序，进行累加并绘图，计算A+B的面积得到最大基尼系数b；

g. 最后计算a/b，即为归一化基尼系数。

5. 项目总结

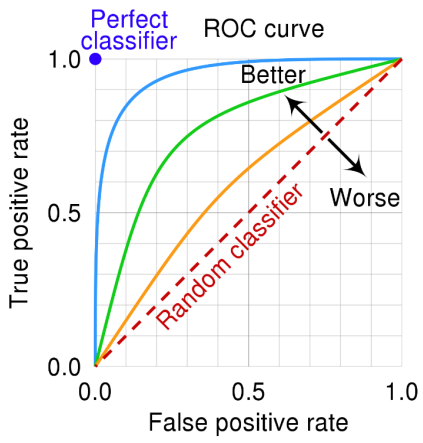
Phase1

1. **如何下载数据**
   a. 项目数据地址：Porto Seguro’ s Safe Driver Prediction
   b. 点击页面中的'Download All'即可下载
2. **如何写proposal**
   a. 明确项目背景、商业意义和目标
   b. 了解数据情况并明确评估方法
3. **风控的两个维度**
   a. 消灭或减少风险事件发生的可能性
   b. 减小风险事件发生所带来的损失

4. **什么是类别不平衡**
    a. 分类任务的数据集中来自不同类别的样本数目相差悬殊
5. **为什么本数据不能用accuracy来作为metric**
    a. 因为该项目的数据集是不平衡的，正负样本数量差距悬殊，分类器对于minority class的判别将会十分困难，majority class的准确率会拉升accuracy的值，给我们造成我们极大的误导
6. **什么是precision和recall**
    a. Precision（精确率/查准率）：TP / (TP+ FP) - 在预测为正的样本中，实际为正的样本比例
    b. Recall（召回率/查全率）：TP/(TP+FN) - 在实际为正的样本中，预测为正的样本比例
7. **什么是第一类、第二类错误**

	A	B	C
1		第一类错误（FP）	第二类错误(FN)
2	发生概率	α	β
3	含义	推翻正确的原假设	接受错误的原假设
4	对应实际	误认为（预测）某一改动能带来收益，而（实际）并不能带来收益	误认为（预测）某一改动不能带来收益，而（实际）能带来收益
5	接受程度	更不能接受第一类错误	

8. **roc曲线的横纵坐标是啥**
    a. 横坐标：FPR = FP/(FP+TN) - 在实际为负的样本中，预测为正的样本比例，越小越好
    b. 纵坐标：TPR = TP/(TP+FN) - 在实际为正的样本中，预测为正的样本比例，越大越好
    c. TPR和FPR的范围都在0到1
    d. TPR在数值上等于positive class的recall，FPR在数值上等于(1 - negative class的recall)
9. **roc曲线的对角线意义**
    a. AUC（Area under the curve of ROC）表示ROC曲线下方的面积，用于评估分类模型
    b. 对角线表示随机猜测分类器的AUC（AUC=0.5）,作为参考线 参考资料



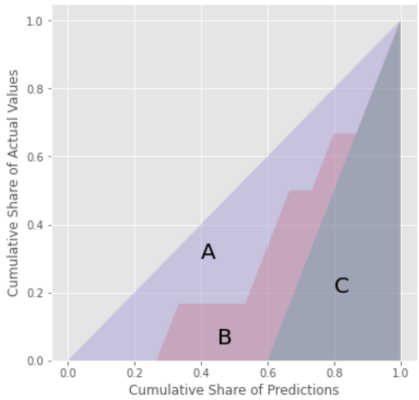
10. **什么是洛伦兹曲线**

a. 在经济学中：
    i. 通过把人口从最穷的到最富的排序，画出累积的收入配比，形成了洛伦兹曲线
b. 在本机器学习项目中：
    i. 将prediction的值从最小到最大排序，并对应调整actual的顺序
    ii. 对调整排序后的actual进行累加并绘制累加折线图（即洛伦兹曲线）

11. **什么是基尼系数**
    a. 在经济学中：
        i. 基尼系数（Gini coefficient），是20世纪初意大利学者科拉多·基尼根据洛伦兹曲线所定义的判断年收入分配公平程度的指标，是比例数值，在0和1之间
    b. 在本机器学习项目中：
        i. 基尼系数表明该模型在区分 "坏 "司机和 "好 "司机方面的有效性，前者将来会索赔，后者将来不会索赔

12. **什么是归一化基尼系数**
    a. 基尼系数的范围从0到0.5，归一化的基尼系数用理论最大值来调整分数，使最大的分数为1

13. **归一化基尼系数的计算逻辑**



    a. 将prediction的值从最小到最大排序，并对应调整actual的顺序；
    b. 对调整排序后的actual进行累加并绘制累加折线图（即洛伦兹曲线）；
    c. 将 x 和 y 轴进行归一化（即转换范围在0到1之间）并绘制45°对角线；
    d. 计算A的面积得到基尼系数a；
    e. 理论上当prediction与actual完全匹配时，可以得到最大基尼系数；
    f. 因此将actual的值从小到大排序，进行累加并绘图，计算A+B的面积得到最大基尼系数b；
    g. 最后计算a/b，即为归一化基尼系数。

14. **归一化基尼系数和auc的关系是什么**
    a. $\text{Normalized Gini} = 2 \times \text{AUC} - 1$

Phase2

1. **变量类型有哪些？** 参考资料
    a. binary ==> 0 or 1
        i. 这一列只含有两种值：不是0就是1
    b. nominal ==> categorical 不含顺序
        i. 最常用的类别型变量，不含任何排序
        ii. 举例：中国、美国、英国等国家就属于nominal变量，他们之间不存在顺序关系
    c. ordinal ==> categorical 含顺序
        i. 顾名思义，跟nominal相比，含顺序
        ii. 举例：成绩排名；小明第一，小红第二，小王第三等等，虽然但看这三个任命不存在任何排序关系，但是在这里就代表着成绩高低，所以是ordinal变量。
    d. interval ==> continuous 无绝对零点
        i. 连续型数据，可以参与统计计算，比较大小，量化差距
            1. 举例：温度；温度是衡量冷热而存在的物理量，40度比20度高20度，但是我们不能说40度是20度的两倍，这是因为interval数据没有绝对零点！
    e. ratio ==> continuous有绝对零点
        i. 连续型数据，可以参与统计计算，比较大小，加减乘除运算都可以
        ii. 举例：身高；200cm比100cm长100cm，也可以说200cm是100cm的两倍。这是因为ratio数据有绝对零点！

如果满足：	类别变量 (nominal)	等级变量 (ordinal)	等距变量 (interval)	等比变量 (ratio)
可排序	✗	✓	✓	✓
可比较大小	✗	✓	✓	✓
可计算频率	✓	✓	✓	✓
可计算众数	✓	✓	✓	✓
可计算中位数	✗	✓	✓	✓
可计算均值	✗	✗	✓	✓
可以量化差值	✗	✗	✓	✓
可以做加减法	✗	✗	✓	✓
可以做乘除法	✗	✗	✗	✓
有真正的零点（0代表没有）	✗	✗	✗	✓

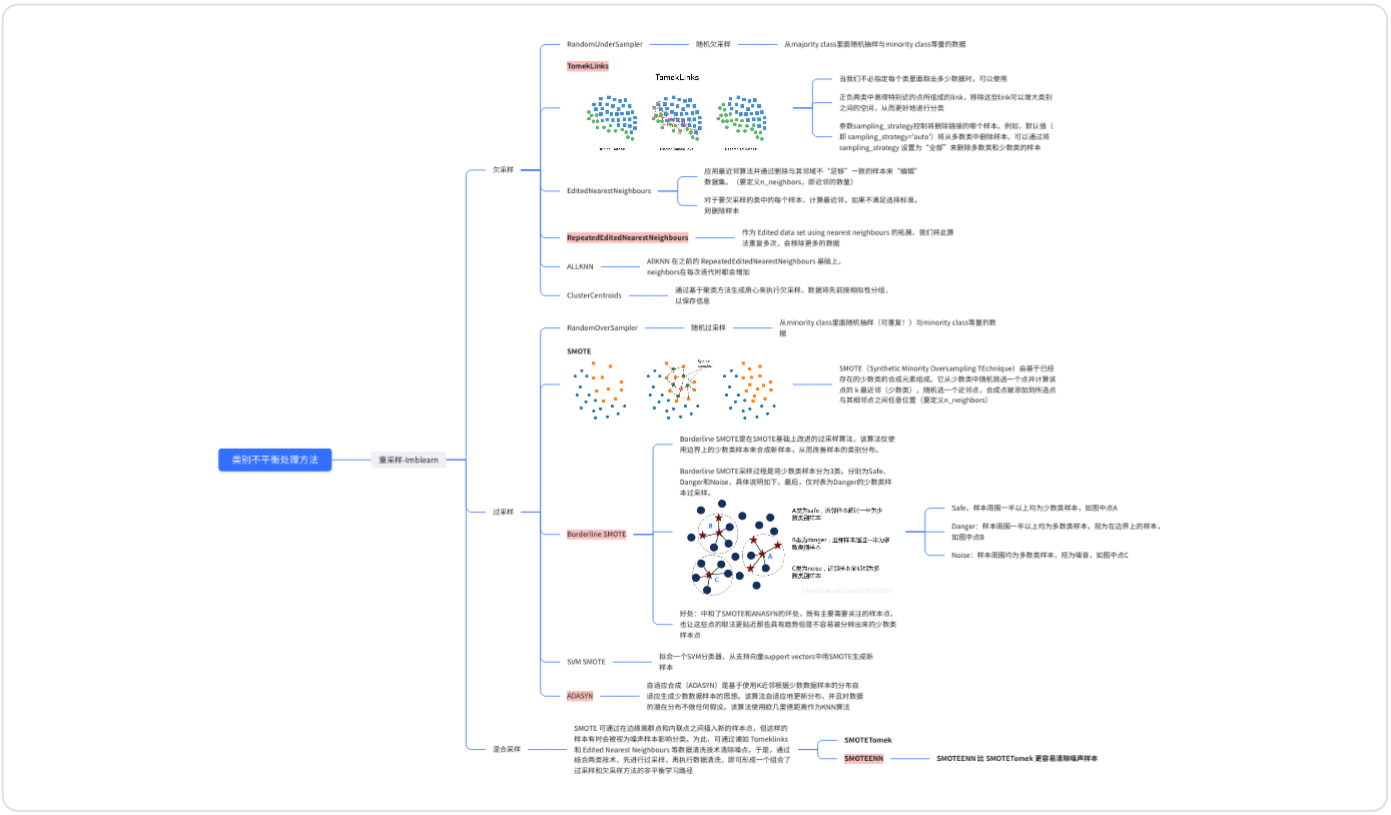
2. **nominal和ordinal的区别？**
    a. nominal不含顺序，ordinal含顺序
    b. nominal不可以比较大小，ordinal可以比较大小

c. nominal不可以计算中位数，ordinal可以计算中位数

3. **interval和ratio的区别?**
   a. interval无绝对零点，ratio有绝对零点
   b. interval不可以做乘除法，ratio可以
4. **什么是元数据?**
   a. 元数据（metadata），也叫做中介数据，或者叫数据的数据（data about data）
5. **元数据的作用是什么?**
   a. 主要来描述数据的属性（property）关于数据的组织、结构梳理，以及为以后的数据分析、可视化、建模都有很重要的意义
6. **元数据是否能应用到别的数据场景下？how?**
   a. 元数据应用场景
7. **元数据怎么帮助我们更好地数据分析?**
   a. 遇到维度很高，特征很多的数据，元数据可以帮助我们对数据进行结构化的梳理
8. **元数据报告函数**
   a. data_management.py.pdf

Phase3

1. 类别不平衡的处理方法汇总
   a. 参考文章：使用 imblearn 应对类别不均衡
   b. 参考文章：数据处理中的上采样、下采样、联合采样和集成采样



2. **上下采样的缺点是什么，改进方法有哪些？为什么？**
   a. 欠采样直接从原数据删除了很多样本，这回导致原数据的信息丢失（information loss）
   b. 过采样由于从minority class的样本进行了复制，所以会出现过拟合的问题
   c. 改进方法（引入imblearn）：
      i. 对majority class的数据进行聚类，并通过从每个cluster中删除一些数据来进行欠采样，从而减少信息丢失
      ii. 在过采样中，我们不再直接复制那些从minority class取出的样本，而是加一些小小的变化，一些合理的噪声进去，从而增加数据的复杂度和多样性
3. **TL 的原理简述**
   a. 正负两类中离得特别近的点所组成的link，移除这些link可以增大类别之间的空间，从而更好地进行分类
4. **TL 和 ENN 的区别**
   a. EditedNearestNeighbours要求正负两类其中一个样本必须有一个相反类的样本作为其最近的邻居，以便将其删除。另一方面，Tomek Links要求正负两个样本都是对方的近邻。总之，Tomek Links使用了一个更严格的条件，使得被删除的样本更少。
5. **SMOTE的原理简述**
   a. SMOTE（Synthetic Minority Oversampling TEchnique）由基于已经存在的少数类的合成元素组成。它从少数类中随机挑选一个点并计算该点的 k 最近邻（少数类），随机选一个近邻点，合成点被添加到所选点与其相邻点之间任意位置（要定义n_neighbors）
6. **SMOTE和ADASYN的区别**

a. SMOTE为每个原始少数群体样本生成相同数量的合成样本。
b. ADASYN使用密度分布作为标准，通过自适应地改变不同少数样本的权重来补偿偏态分布，从而自动决定必须为每个少数样本生成的合成样本数。
c. 参考文章：不平衡数据处理之SMOTE、Borderline SMOTE和ADASYN详解

7. **我们的数据集适合用哪种方法？为什么？**
   a. 混合采样
      i. SMOTE 可通过在边缘离群点和内联点之间插入新的样本点，但这样的样本有时会被视为噪声样本影响分类。为此，可通过诸如 Tomeklinks 和 Edited Nearest Neighbours 等数据清洗技术清除噪点。于是，通过结合两类技术，先进行过采样，再执行数据清洗，即可形成一个组合了过采样和欠采样方法的非平衡学习路径。
      ii. SMOTEENN 比 SMOTETomek 更容易清除噪声样本

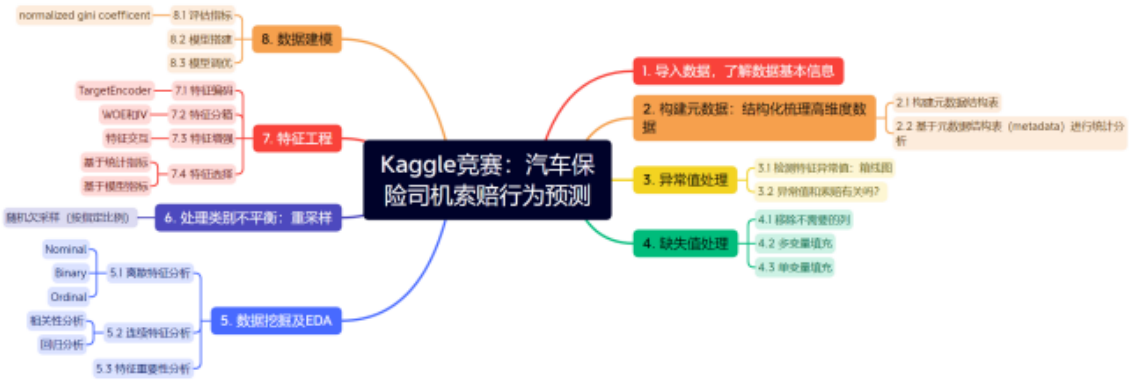
Phase4

1. **缺失值的类型有哪些**
   a. MCAR (Missing completely at Random)
      i. 意思是“缺失与任何值之间没有关系”。因此，在这种情况下，可以删除缺失值。为此，可以删除列缺失值；最小化丢失数据或删除行缺失值。
   b. MAR (Missing at Random)
      i. 意思是“缺失与其他观察到的数据之间存在系统关系，但与缺失数据无关”。
   c. MNAR (Missing not at Random)
      i. 意思是“缺失与它的值之间存在关系，缺失或非缺失”。
2. **缺失值填充方法**
   a. 单变量填充（univariate imputation）
      i. 就是当前列的缺失值只能通过当前列的信息填充，跟其他列无关。
   b. 多变量填充（Multivariate imputation）
      i. 每一列填充缺失值可以借助其他列的信息进行填充。
      ii. Multivariate Imputation By Chained Equations (MICE)：通过一系列迭代的预测模型来“填充”（插补）数据集中的缺失数据。在每次迭代中，将使用数据集中的其他变量来估算数据集中的每个指定的变量，这些迭代持续运行，直到满足收敛为止。
      iii. KNN

---

6. 项目代码

Kaggle竞赛：汽车保险司机索赔行为预测



In [1]:
```
# import packages
# data processing
import pandas as pd
import numpy as np
from datetime import timedelta, datetime
import re
```

📎Kaggle竞赛：汽车保险司机索赔行为预测 - Jupyter Notebook.pdf

Porto Seguro’s Safe Driver Prediction

Project Process:

- 1. Read Data
- 2. Build metadata
- 3. Outlier detection
- 4. Missing Value handling
  - 4.1 Multivariate imputation
  - 4.2 Univariate imputation
- 5. EDA
  - 5.1 Discrete feature analysis
    - 5.1.1 Nominal
    - 5.1.2 Binary
    - 5.1.3 Ordinal
  - 5.2 Continuous feature analysis
    - 5.2.1 Correlation Analysis
    - 5.2.2 Regression Analysis
- 6. Feature Engineering
  - 6.1 Encoding

📎Porto Seguro’ s Safe Driver Prediction - Jupyter Notebook.pdf