Quantum One 数据科学家笔试题目

附件Excel中是一张表名为LoanData的数据表及相关的数据字典DataDictionary， 请利用这张表完成以下问题：

1. 请以借款金额（loan_amt）为目标变量建立一个线性模型。
   * 注：请将Python/R/SAS程序和结果单独列出，并简单解释模型结果

结果： 在线性模型中，如下变量（按重要程度）可以用来Model loan_amnt （借款金额）：

1. installment 月供金额
2. term 借款周期
3. sub_grade 信用等级
4. mort_acc 房贷账户数目
5. verification_status 收入核验状态
6. revol_bal 客户当前授信总额度
7. purpose 借款目的
8. annual_inc 年收入
9. acc_open_past_24mths 去24个月内新开账户

Goodness of fit of the Linear Regression Model:

- Sample size: 20690
- R-Sqaured: 0.9937
- P-value of F-test <0.0001 F Value: around 60,000
- RMSE: around 1500

结论： 从各项指标来看，这个Model整体的准确度还是很不错的。选出的各项变量在现实生活中也能够合理解释。缺点是用的Classification Variables过多，有多种排列组合能给出相似的结果，导致了实际结果会有所偏差, 不如想象中的准确。

改进方法：

- 更好的处理missing value, 具体方法code里面有提
- 用更准确的Numerical Variables来代替Classification Variables. 比如信用等级可以用FICO scores代替
- 分析各个dependent variable之间的Collinearity，从而进一步剔除重复的Variables

2. 以借款利息（int_rate）高低区分不同信用等级的借款人，以此为目标变量建立逻辑回归模型。
   * 注：请将Python/R/SAS程序和结果单独列出，并简单解释模型结果

结果和分析： 用最好信用等级A1和最差的信用等级G5分别做参照做逻辑回归模型，从Chi-Sqaure的结果看两个模型都能证明借款利息和信用等级有明显的联系，但用A1做参照的时候明显的整体Goodness of fit 要好很多（G5做参照的时候，G2-G4的Model就开始不准了）。结合实际猜想原因的话，现在的信用记录评定是由好到坏逐步去扣。如果一种新的信用评定方法是假设人是坏的，并逐步变好增加信用，模型结果应该会相反。

3. 以借款利息（int_rate）高低区分不同信用等级的借款人，以此为目标变量建立一至俩个机器学习模型（GBM、Random Forest、Neural Network、SVM等）。
   * 注：请将Python/R/SAS程序和结果单独列出，并简单解释模型结果

结果和分析：用SAS做了K-Means Clustering（免费版算法选择有限） 在alpha = 0.05 时，分了7个信用等级；在alpha = 0.01 时，分了14个信用等级。当像原始数据中分成34个信用等级时，R-square 无限趋近于1。通过之前的Data Exploration，发现int_rate不是 continuously distributed，而是在2万多个数据中只有41个不同的int_rate，它本身就像信用等级一样可以作为一个categorial variable。