## Assignment 2 specification and instructions

*Credits*: 35% of the module

**Aim / Learning objectives**

The main objective – written by the client Maydup Bank plc – is:

“*The goal is to develop a predictive model for customer defaults. Data Scientists are free to choose any modelling approach. The only requirement is that the model gives a decision on whether to accept or decline new applicants. Possibilities include - for example - a classification model or a probability model with a specified cut-off. The decision should only be based on variables provided in this dataset.*”

You should treat this an end-to-end machine learning project, where your aim is to (i) make good predictions and (if possible) (ii) provide useful insights to the client about the quality and usefulness of the data they are collecting.

主要目标（由客户 Maydup Bank plc 撰写）是：

“目标是开发一个预测客户违约的模型。数据科学家可以自由选择任何建模方法。唯一的要求是该模型可以决定是否接受或拒绝新申请人。可能性包括 - 例如 - 具有指定截止值的分类模型或概率模型。决策应仅基于此数据集中提供的变量。”

您应该将其视为端到端机器学习项目，您的目标是 (i) 做出正确的预测和（如果可能）（ii）向客户提供有关他们正在收集的数据的质量和实用性的有用见解。

**Data**

You will use the data in **P2data????.csv**. This is obtained by emailing the module lecturer(s) on  will consist of 9,000 rows of data.  See **Appendix A** below for meanings behind the entries.

Maydup Bank plc has holdout data (about 4,500 instances). These data will be used to evaluate the performance of your predictive model when applied to unseen data.

您将使用 P2data????.csv 中的数据。这可以通过向模块讲师发送电子邮件获得，包含 9,000 行数据。请参阅下面的附录 A 了解条目背后的含义。

Maydup Bank plc 有保留数据（约 4,500 个实例）。这些数据将用于评估您的预测模型在应用于未见数据时的性能。

**Appendix A: Data Dictionary**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| app\_date | Date of application |
| app\_id | Unique application ID |
| cust\_age | Customer's age at application |
| GOOD | Indicator if account turned bad |
| BAD | Indicator if account turned bad |
| occ\_code | Profession |
| time\_emp | Years in employment |
| res\_indicator | Residential status |
| CA\_01 | Applicant - Worst status of all current accounts in the last 6 months |
| CA\_02 | Applicant - Total balance of all live current accounts |
| CA\_03 | Applicant - Number of live current accounts |
| ER\_01 | Household - Length of residence at current address |
| ER\_02 | Applicant - Length of residence at current address |
| S\_01 | Applicant - Number of credit searches in the last 3 months |
| S\_02 | Applicant - Minimum number of months since credit search last 12 months |
| disp\_income | Applicant - Disposable income |
| I\_01 | Applicant - Number of accounts opened in the last 3 months |
| I\_02 | Applicant - Total credit limit for all live cards |
| I\_03 | Applicant - Age in months of most recently opened account |
| I\_04 | Applicant - Total number of accounts (including loans, cards, contracts, etc) |
| D\_01 | Applicant - Total balance of all utility bills last month |
| D\_02 | Applicant - Cards total balance (live) |
| I\_05 | Household - Total number of accounts (including loans, cards, telecoms, etc) |
| I\_06 | Household - Average Number of CCJs |
| P\_01 | Average % of households with a live credit card |

|  |  |
| --- | --- |
| **Variable:** | **Occupation\_code** |
| **Value** | **Meaning** |
| SA | SELF-EMPLOYED |
| SB | MANAGERS, DIRECTORS AND SENIOR OFFICIALS |
| SC | PROFESSIONAL OCCUPATIONS |
| SD | PUBLIC SECTOR |
| SE | ADMINISTRATIVE AND SECRETARIAL OCCUPATIONS |
| SF | ASSOCIATE PROFESSIONAL AND TECHNICAL OCCUPATIONS |
| SG | CARING, LEISURE AND OTHER SERVICE OCCUPATIONS |
| SH | SKILLED TRADES OCCUPATIONS |
| SJ | AGRICULTURAL SECTOR |
| SK | JOB SEEKER |
| SL | SALES AND CUSTOMER SERVICE OCCUPATIONS |
| SM | ELEMENTARY OCCUPATIONS |
| SN | CARING, LEISURE AND OTHER SERVICE OCCUPATIONS |
| SO | CASUAL/PART-TIME |
| FT | UNEMPLOYED/EDUCATION |

|  |  |
| --- | --- |
| **Variable:** | **Homeowner\_Indicator** |
| **Value** | **Meaning** |
| H | Home Owner |
| P | Parental |
| R | Rented |

|  |  |
| --- | --- |
| **Variable:** | **CA\_01** |
| **Value** | **Meaning** |
| 1. --- | payments 3-6 months overdue |
| 2. --- | payments 1-2 months overdue |
| 3. --- | payments current / inactive / not yet received |
| 4. --- | bureau data missing - address not found |
| 5. --- | bureau data missing - not visible or not qualified |

**Instructions**

You will see that the data is sparse (i.e. contains many missing values). In particular, the responses are sparse (i.e. the data do not supply a complete separation of customers into GOOD and BAD). In light of this, one approach – which you can ignore – would be to treat the problem as a three-class classification into GOOD, BAD and PASS, where PASS denotes “insufficient data to make a firm classification into GOOD or BAD”. PASS predictions could then be saved by the client for use when sufficient data has been obtained.

 Another option – which you can ignore – is to try VERY GOOD, GOOD, PASS (maybe), BAD, VERY BAD.

 A third – which you can ignore and which is clearly related to the second – is not to classify but to return a probability of being GOOD, together with – if possible – a level of confidence in the value of an individual prediction.

 For all these options, and for any other approach you take, bear in mind the “*specified cut-off*” part of the client’s instructions.

 You can use your choice of internal validation method and metric(s) to provide your estimate of the predictive performance when the client runs your code on the holdout data. Make sure when developing code that processes and modifies your data that the client can run the same code on their holdout data.

 If you use any randomisation (in, say, shuffling and/or splitting data), use **????** as the random key where **????** denotes the last four digits in your matriculation number.

 Write a short report summarising your methods and findings. In particular: Can you identify any variables that are important (or unimportant) for predicting new instances of this data class? Which specific model would you advise the client to deploy based on analyses analyses, and why?

您将看到数据很稀疏（即包含许多缺失值）。特别是，响应很稀疏（即数据无法将客户完全分为好客户和坏客户）。鉴于此，一种方法（您可以忽略）是将问题视为三类分类，即好客户、坏客户和合格客户，其中合格表示“数据不足，无法将客户明确分为好客户或坏客户”。然后，客户可以保存合格预测，以便在获得足够数据后使用。

另一种选择（您可以忽略）是尝试非常好、好客户、合格（可能）、坏客户、非常坏客户。

第三种选择（您可以忽略，它与第二种选择明显相关）不是分类，而是返回好客户的概率，以及（如果可能）单个预测值的置信度。

对于所有这些选项以及您采取的任何其他方法，请记住客户指示中的“指定截止”部分。

您可以使用您选择的内部验证方法和指标来提供客户在保留数据上运行代码时的预测性能估计。在开发处理和修改数据的代码时，请确保客户可以在其保留数据上运行相同的代码。

如果您使用任何随机化（例如，改组和/或拆分数据），请使用 ???? 作为随机密钥，其中 ???? 表示您的入学编号的最后四位数字。

撰写一份简短的报告，总结您的方法和发现。特别是：您能否确定任何对预测此数据类的新实例重要（或不重要）的变量？根据分析，您会建议客户部署哪种特定模型，以及为什么？

**Key points**

* Your final model may not perform well, since the data is (a) not clean, (b) sparse in places, and (c) is not real data (the data were randomly generated, and correlations between variables were used to mimic what an actual application and performance snapshot might look like). We are looking to see if you can build a sensible model and provide a sensible estimate of its generalisation error, and also if you can clearly document your effort in a manner that would be useful to the client.
* The holdout data were simulated using the same process, so ideally the error rates for your internally validated model(s) will be similar to those when applied to holdout data.
* Presentation counts.​ A clear and concise Jupyter notebook will earn more marks than an unstructured notebook full of opaque and poorly commented code.  For the report,
* If you do not understand something or have questions, you are encouraged to discuss it with your peers (say, via the Teams channel) or the module lecturer(s). However, the deliverables that you submit must comply with the policy on good academic practice.
* Note that any discussions involving other students taking x will be affected by the inherent differences in the underlying data, the model classes chosen, and any randomization performed as part of your workflow. Hence discussions will involve options for techniques and reporting, rather than comparing actual results.
* • 您的最终模型可能表现不佳，因为数据 (a) 不干净，(b) 在某些地方稀疏，并且 (c) 不是真实数据（数据是随机生成的，变量之间的相关性用于模拟实际应用程序和性能快照可能的样子）。我们希望看到您是否可以构建一个合理的模型并提供其泛化误差的合理估计，以及您是否可以以对客户有用的方式清楚地记录您的努力。
* • 使用相同的过程模拟保留数据，因此理想情况下，内部验证模型的错误率将与应用于保留数据时的错误率相似。
* • 演示很重要。清晰简洁的 Jupyter 笔记本将比充满晦涩难懂且注释不清的代码的非结构化笔记本获得更多分数。对于报告，
* • 如果您不理解某些内容或有疑问，我们鼓励您与您的同学（例如，通过 Teams 频道）或模块讲师讨论。但是，您提交的可交付成果必须符合良好学术实践政策。
* • 请注意，涉及其他参加 学生的任何讨论都将受到基础数据、所选模型类别以及作为工作流程一部分执行的任何随机化的固有差异的影响。因此，讨论将涉及技术和报告选项，而不是比较实际结果。

**Submission**

1.     The code of your solution in a Jupyter notebook. The client should be able to replace your data with the holdout data and perform the same data cleaning, feature engineering, internal validation and model selection procedures in order to get a full **external validation** of your results.

 2.     A brief, clear and concise summary describing your models, your comparative measures of performance, and your final result, in a maximum five page PDF file.

1. Jupyter 笔记本中的解决方案代码。客户应该能够用保留数据替换您的数据，并执行相同的数据清理、特征工程、内部验证和模型选择程序，以便对您的结果进行完整的外部验证。

2. 简短、清晰、简洁的摘要，描述您的模型、性能比较指标和最终结果，最多五页 PDF 文件。