**Part I**

**SunnyBridge Data Science Case Study**

**Overview:** Enclosed is a case study that SunnyBridge Inc. asks candidates for its Data Science team to complete. The case study is designed to simulate a real project at SunnyBridge. You will be evaluated based upon:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Code quality and reusability
7. Writing

**Instructions**

1. Review this document entirety. Immediately ask any questions that come up.
2. Deadline is 12:00pm, May 31, 2018. You have up to one and half months (in the real interview, you only have 8 hours, and they think 4-6 hours should be sufficient) to complete the case.
3. Email back the entire set of case files and the files listed in the Output section.

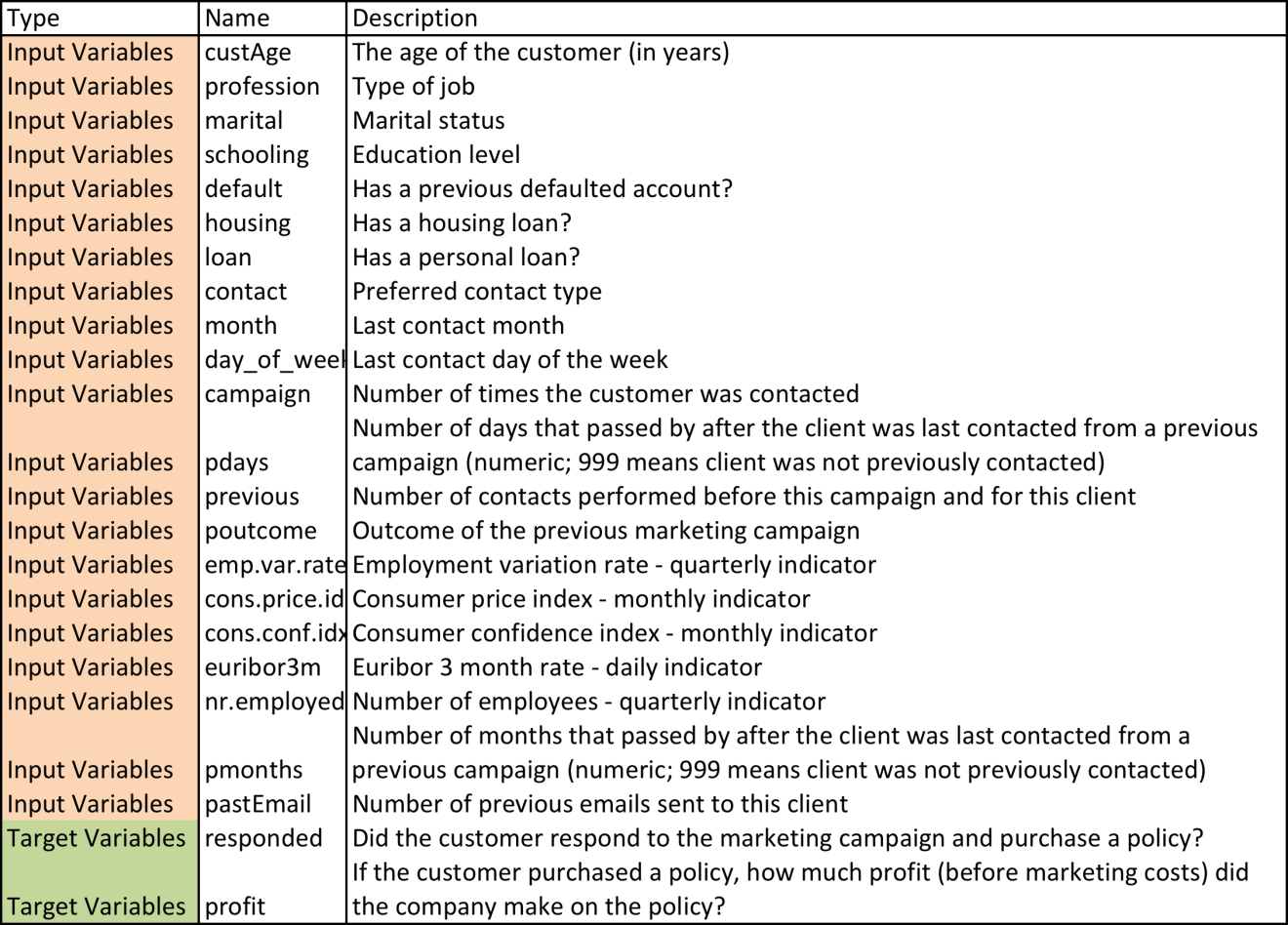
**Points to Consider**

1. If you have questions, you may email me or ask on piazza. I’ll do our best to respond ASAP. However, if I do not respond, then feel free to make a reasonable assumption. State this assumption in your report.
2. There isn’t meant to be any tricks here. This is just a straightforward data analysis problem.
3. Please do not discuss this case with anyone else. However, you may use any Internet resources for syntactical assistance only.
4. Please complete this case using Matlab, R, Python or whatever language you are familiar with.
5. Ensure you can explain all of the steps of your process and code. You will be asked to show your process, code and results to SunnyBridge employees as part of the interview process.
6. Cite any resource/reference you use in your report!

**Scenario:** You are working for SunnyBridge as a Data Scientist. SunnyBridge has been commissioned by an insurance company to develop a tool to optimize their marketing efforts. Your objective is to determine **which set of customers the marketing firm should contact** to maximize profit.

1. The cost of marketing to a particular customer is $30. This cost is paid regardless of whether the customer responds to our marketing or not.
2. Only if a customer responds to our marketing, do we earn a profit.
3. Profit does **NOT** include the marketing cost.
4. Total Profit = Average profit per responding Customer \* Number of customers responding – Number of customers to whom you marketed \* $30

**Data:** The insurance company has provided you with a historical data set (training.csv). The company has also provided you with a list of potential customers to whom to market (testingCandidate.csv). From this list of potential customers, you need to determine yes/no whether you wish to market to them.



**Output:** Please email back the following files:

1. All associated code files you used to complete your analysis.
2. Please add a column to the testingCandidate.csv file. In this column, for each observation indicate a 1 (yes) or a 0 (no) whether you wish to market to that candidate. Submit your testingCandidate.csv file.
3. Please prepare a 3-to-5-page (fontsize = 10pt, code files or appendices are NOT counted) write-up. This report should be written as if you were presenting your results to a non-technical audience at the client. You are free to use tables or/and figures in your write-up. You are recommended to prepare an English report. However, you are also allowed to write in Chinese if you feel painful to write in English.

**Part II**

**上证指数预测探讨**

**SSE Composite Index Case Study**

代号SHA:000001，上海证券交易所主要的综合股价指数，是反应挂牌股票总体走势的统计指标。其历史数据可在网易财经中免费下载。

链接：<http://quotes.money.163.com/trade/lsjysj_zhishu_000001.html>

附件000001.csv提供了1990年12月29日到2018年04月13日的历史数据，最新数据可以自行从以上链接中下载。

****

**任务描述**

1. 截止日期：2018年6月1日晚12:00。
2. 预测指标：6月4日开始的一周（5个交易日）的**最高价**、**最低价**。
3. 预测所用到的数据日期范围自行确定。
4. 预测所需的数据变量自行确定，也可自由通过其他渠道获取新的数据来源。
5. 报告正文主体不少于2页（不含代码、参考文献），应清晰阐述所用到的方法、模型以及数据处理过程。附件包含相关代码以及结果展示。
6. 成绩评价因素同PartI。

Emm pca