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

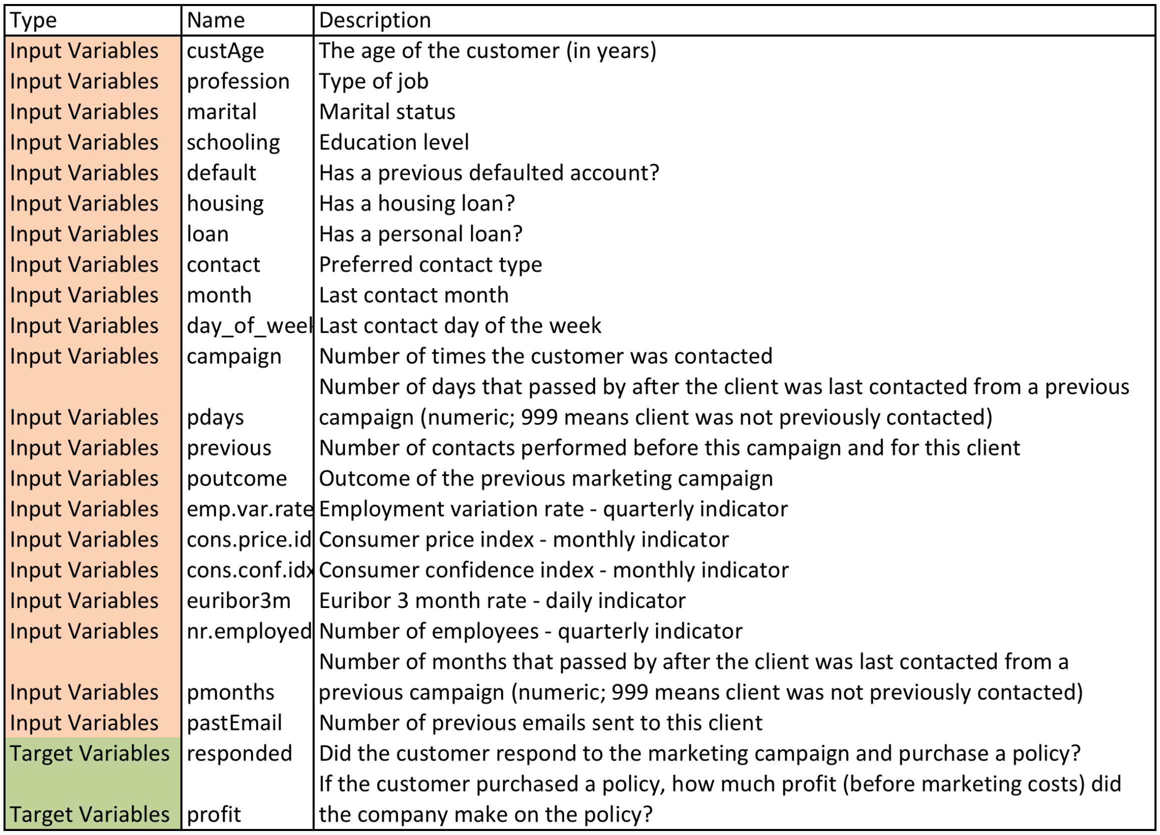
1. If you have questions, you may email me or ask on piazza. I'll do our best to respond ASAP. However, if I do not respond, then feel free to make a reasonable assumption. State this assumption in your report.
2. There isn't meant to be any tricks here. This is just a straightforward data analysis problem.
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.



Type	Name	Description
Input Variables	custAge	The age of the customer (in years)
Input Variables	profession	Type of job
Input Variables	marital	Marital status
Input Variables	schooling	Education level
Input Variables	default	Has a previous defaulted account?
Input Variables	housing	Has a housing loan?
Input Variables	loan	Has a personal loan?
Input Variables	contact	Preferred contact type
Input Variables	month	Last contact month
Input Variables	day_of_week	Last contact day of the week
Input Variables	campaign	Number of times the customer was contacted
Input Variables	pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Input Variables	previous	Number of contacts performed before this campaign and for this client
Input Variables	poutcome	Outcome of the previous marketing campaign
Input Variables	emp.var.rate	Employment variation rate - quarterly indicator
Input Variables	cons.price.id	Consumer price index - monthly indicator
Input Variables	cons.conf.idx	Consumer confidence index - monthly indicator
Input Variables	euribor3m	Euribor 3 month rate - daily indicator
Input Variables	nr.employed	Number of employees - quarterly indicator
Input Variables	pmonths	Number of months that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Input Variables	pastEmail	Number of previous emails sent to this client
Target Variables	responded	Did the customer respond to the marketing campaign and purchase a policy?
Target Variables	profit	If the customer purchased a policy, how much profit (before marketing costs) did the company make on the policy?

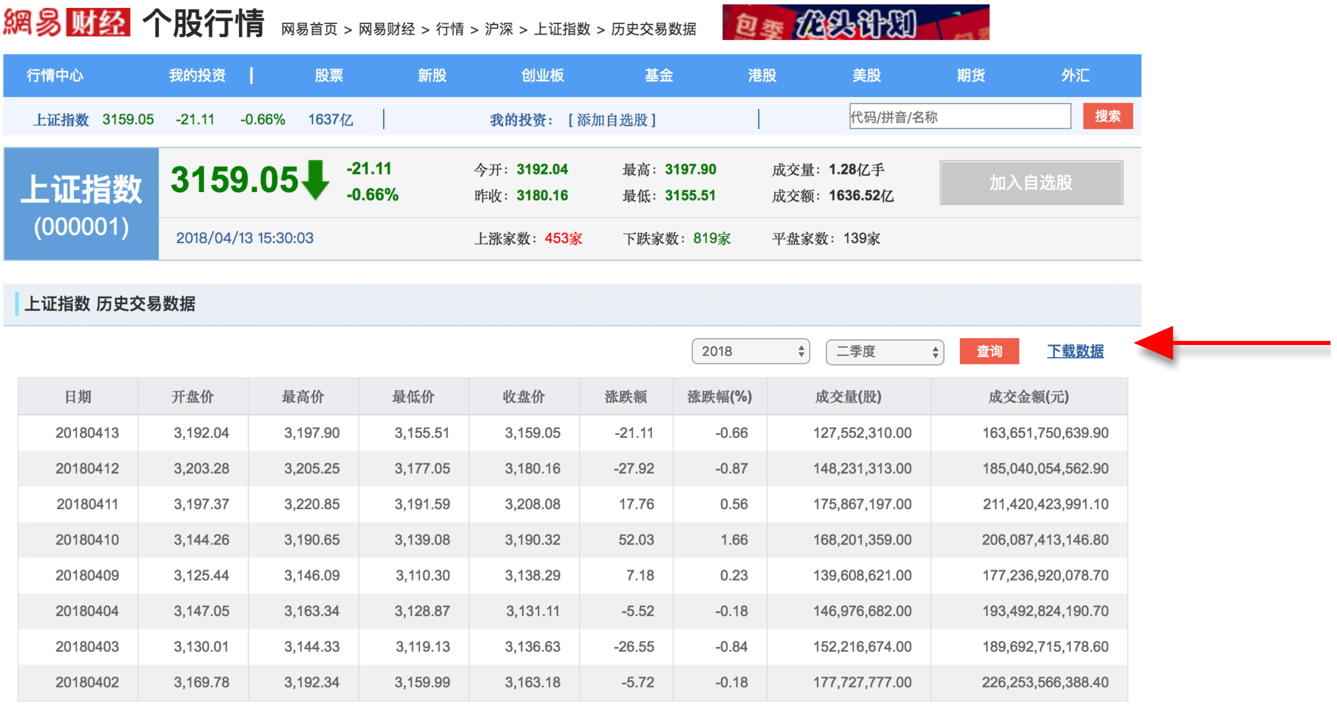
Output: Please email back the following files:

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) writeup. 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

<u>上证指数预测探讨</u>
<u>SSE Composite Index Case Study</u>

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。



上证指数 3159.05 -21.11 -0.66% 1637亿

上证指数 (000001) 3159.05 -21.11 -0.66%
今开：3192.04 昨收：3180.16 最高：3197.90 最低：3155.51 成交量：1.28亿手 成交额：1636.52亿
2018/04/13 15:30:03 上涨家数：453家 下跌家数：819家 平盘家数：139家

上证指数 历史交易数据

日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(股)	成交金额(元)
20180413	3,192.04	3,197.90	3,155.51	3,159.05	-21.11	-0.66	127,552,310.00	163,651,750,639.90
20180412	3,203.28	3,205.25	3,177.05	3,180.16	-27.92	-0.87	148,231,313.00	185,040,054,562.90
20180411	3,197.37	3,220.85	3,191.59	3,208.08	17.76	0.56	175,867,197.00	211,420,423,991.10
20180410	3,144.26	3,190.65	3,139.08	3,190.32	52.03	1.66	168,201,359.00	206,087,413,146.80
20180409	3,125.44	3,146.09	3,110.30	3,138.29	7.18	0.23	139,608,621.00	177,236,920,078.70
20180404	3,147.05	3,163.34	3,128.87	3,131.11	-5.52	-0.18	146,976,682.00	193,492,824,190.70
20180403	3,130.01	3,144.33	3,119.13	3,136.63	-26.55	-0.84	152,216,674.00	189,692,715,178.60
20180402	3,169.78	3,192.34	3,159.99	3,163.18	-5.72	-0.18	177,727,777.00	226,253,566,388.40

<u>任务描述</u>

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<u>**最高价**</u>、<u>**最低价**</u>。
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。