2020 Fall MS CISBA Comprehensive Exam - Data Analytics Problem

Background

Assume that you are a data scientist working at the loan department of a regional bank. Your boss has asked you to set up a targeted marketing campaign for a loan offer. Your main goal is to identify 20 people who are likely to accept the loans and contact them. You are given a training dataset, which provides the information about 601 bank customers that have already responded (Yes or No) to a previous loan campaign in 2020 Spring. Your assumption is that *people, who have responded positively to the previous campaign share similar characteristics with the people, who will respond positively to this campaign*. You are going to develop data mining models from the training data and then apply them to prediction dataset with 200 customers to identify 20 people who are most likely to accept the loans.

For the bank, there are monetary benefits associated with someone accepting the loan offer. However, the amount of such benefits is not the same and depends on where the person is living ("Region" attribute). Specifically, the returns are:

- $10,000 if the person lives in INNER_CITY
- $9,000 if the person lives in SUBURBAN
- $7,200 if the person lives in TOWN
- $4,800 if the person lives in RURAL

You are allowed to contact only 20 people for loan offerings. The main objective is to maximize the expected return, which considers the probability of accepting the offer as well as the return amount. For example, if a customer living in suburban is predicted to have a 80% probability to accept the offer [i.e., confidence(Yes)= 0.80], the expected return from this customer is $9,000*80%=$7,200.

In addition, your boss wants you to deploy a secure automation process for identifying a potential customers in the future. With these in mind, your main goals will include:

- Understanding the business problem
- Building multiple predictive models
- Applying those models for prediction to identify top 20 customers and compute the highest expected return for each model.
- Designing a conceptual automation process for data mining so that the whole process can be automatically applied to the same needs in the future. The access to the automation process will be available to particular users via the bank's intranet.

Datasets: Please download the datasets titled "loan_train.csv" and "loan_predict.csv" from WT Class (see the appendix for variable definition).

Instructions: Please follow the CRISP-DM procedure to complete this project and then write a 12-page report.

1. Business Understanding, including but not limited to:
    a. Enrich to the background information.
    b. Include any public reports or academic articles to show the importance of this problem and the value of a data mining approach to solving the problem.
    c. Formulate the business problem or research question.
    d. Briefly and clearly describe how you are going to approach the business problem or research question.
2. Data Understanding, including but not limited to:
    a. How many attributes and records are included in the dataset?
    b. What does each attribute mean?
    c. How those attributes are related?
    d. What is the data type of each attribute?
    e. What is the target attribute in this case?
3. Data Preparation, including but not limited to:
    a. Identify any data quality problems such as duplicated records, missing values, and outliers and describe how you tackle them one by one. Justify what you are doing.
    b. Organize and/or visualize all the attributes except id with the cleaned data in appropriate tables or charts. Interpret each of them briefly.
    c. Explore the relationship between the target attribute and each of the other the attributes except id in appropriate tables or charts. Discuss your preliminary findings briefly.
    d. Check if there is any redundant or irrelevant attribute before modeling and describe how are you deal with them one by one. Justify what you are doing.
4. Modeling
    a. Use both decision tree and logistic regression to develop your classification models. For decision tree model, please use two different sets of parameters such as minimal gain and minimal leaf size. In total, you need to have three models: one logistic regression and two decision tree models.
    b. Prepare the data slightly differently for each method because data mining methods may have different requirement for the data.
    c. Discuss at least one advantage and disadvantage for each method you use.
    d. Present and discuss the results of your classification models. What conclusions can you find from each model?
5. Prediction and Deployment
    a. Apply three models to the prediction dataset.
    b. Please compute how many customers are predicted to Yes by each model.
    c. Compare the prediction results by the three models. Are they the same?
    d. Identify the top 20 customers by listing their ID and then compute the maximum expected return for each model.
    e. Discuss which model will generate the highest expected return and provide your recommendation.

f. Design a conceptual automation for data mining so that the whole process can be automatically applied to the same or similar needs in the future. You may draw a chart to show your automation process and then describe it clearly.
g. In addition, you need to consider security concerns or issues when automating your data mining process via the intranet. Please consider at least two security concerns or issues and discuss how you are going to address each of them.

6. Reflection & Discussion, including but not limited to the following questions:
   a. Do you encounter any problems in this project? How do you solve them?
   b. Are there any limitations (such as data, model, assumptions, etc.) in this project report? Discuss clearly at least two limitations.
   c. Are there any other important factors that are not considered here? Discuss clearly at least two factors.
   d. How can you make the project report better in the future?

Software

Use whatever software that you feels the most comfortable and appropriate about, even though Excel, RapidMiner, Python, and R are the analytical tools that you may learn in your graduate program. You are allowed to combine multiple tools.

Deliverable

1. A Project Report with the following format and content (you may refer to your group project report in CIDM 6308 or CIDM 6355 as the template of your final report).
   - Format
     - ✓ A Microsoft Word file
     - ✓ Single spaced
     - ✓ Font: Times New Roman
     - ✓ Font size: 12
     - ✓ 1 inch margins (top, bottom, left and right).
     - ✓ Align your text left except report title and table/chart captions.
     - ✓ Display the page number at the center of footer
   - Content: with the following order
     1) Cover page: an informative title, full name, and date.
     2) Table of Content
     3) Executive Summary: one page, including a brief summary of your goal, data, models, and findings, and your recommendation.
     4) Main text: 9-12 full pages of written text, including all the six sections mentioned in the instruction. You may include appropriate tables (with an appropriate table caption above and centered) and/or figures (with an appropriate table caption above and centered) as needed in your main text. You are required to provide in-text citation with the author-date method with the APA style.
     5) Reference: at least one page, including at least ten references with the APA style.
     6) Appendices: you may include any relevant appendices to provide details about your data and models such as screenshots of RapidMiner processes, R or Python output, and diagrams.

2. Processing files
    - ✓ If you are using Excel, please save and submit your Excel file(s).
    - ✓ If you are using RapidMiner, please use File → Export Process to save and submit your process file(s)
    - ✓ If you are using R or any other programming tool such as Python, please save and submit your script(s).
    - ✓ If you are using any other analytics tool for this task, please let the exam coordinator know and s/he will determine what processing files you need to submit.
    - ✓ If you are using multiple tools, please save and submit all the processing files.

Appendix

Variable definition

Field	Description
id	A unique ID to identify each customer
age	A customer's age (years old) when receiving the loan offer
sex	A customer's gender
region	The region in which a customer is living when receiving the loan offer
income	A customer's annual taxable income when receiving the loan offer
married	Is this customer married when receiving the loan offer?
children	The number of children that a customer has when receiving the loan offer
car	Does this customer own a car when receiving the loan offer?
save_act	Does this customer have a saving account when receiving the loan offer?
current_act	Does this customer have a current loan account with the bank?
mortgage	Does this customer have a mortgage when receiving the loan offer?
response	Does this customer respond to the loan offer