|  |  |
| --- | --- |
|  | **Introduction to Business Data Analytics** |

|  |  |
| --- | --- |
| **Homework #1 (Part 2)** |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(put your names above (including any nicknames))

Note: This is a team homework assignment. Discussing this homework with your classmates outside your MSBA team is a **violation** of the Honor Code. If you **borrow code** from somewhere else, please add a comment in your code to **make it clear** what the source of the code is (e.g., a URL would sufficient). If you borrow code and you don’t provide the source, it is a violation of the Honor Code.

Total grade for Part 2: \_\_\_\_\_\_\_ out of \_\_\_50\_\_\_ points

***ATTENTION: HW1 has two parts. Please first complete the Quiz “HW1\_Part1” on Canvas. Then, proceed with Part 2 in the following page. You will need to submit (a) a PDF file with your answers and screenshots of Python code snippets as well as Rapidminer repositories and (b) the Python code and Rapidminer repositories.***

**(50 points) [Implement this exercise with both Python (30 points) RapidMiner(20 points). Use the decision tree classification technique on the *HW1* dataset. This dataset is provided on the course website and contains data about consumers and their decisions to terminate a contract (i.e., consumer churn problem).**

**Data description:**

Col. Var. Name Var. Description

----- ---------- --------------------------------------------------------------

1 revenue Mean monthly revenue in dollars

2 outcalls Mean number of outbound voice calls

3 incalls Mean number of inbound voice calls

4 months Months in Service

5 eqpdays Number of days the customer has had his/her current equipment

6 webcap Handset is web capable

7 marryyes Married (1=Yes; 0=No)

8 travel Has traveled to non-US country (1=Yes; 0=No)

9 pcown Owns a personal computer (1=Yes; 0=No)

10 creditcd Possesses a credit card (1=Yes; 0=No)

11 retcalls Number of calls previously made to retention team

12 churndep Did the customer churn (1=Yes; 0=No)

**30 Points (Python):**

1. **(5 points) Build a decision tree model that predicts whether a consumer will terminate his/her contract. In particular, I would like for you to create a decision tree using entropy with no max depth. Some possible issues / hints to think about: using training vs. test datasets.**
2. **(5 points) Explore how well the decision trees perform for several different parameter values (e.g., for different splitting criteria).**
3. **(5 points) Discuss the model (decision tree) that provides the best predictive performance from experimenting with different parameter values in question (b).**
4. **(15 points) Present a brief overview of your predictive modeling process, explorations, and discuss your results. That is, you need to lay out the steps you have taken in order to build and evaluate the decision tree model. For instance, how did you explore the data set before you built the model? Write this report in a way that the upper level management of the team would understand what you are doing. Why is the decision tree an appropriate model for this problem? How can we evaluate the predictive ability of the decision tree? If you build decision trees with different splitting criteria, which decision tree would you prefer to use in practice? Make sure you present information about the model “goodness” (please report the confusion matrix, predictive accuracy, classification error, precision, recall, f-measure).**

**20 Points (Rapidminer):**

**As you discuss the results please make sure you provide screenshots of your corresponding Python code at the same time. At the end, also please provide the Rapidminer screenshots as well (Screenshot on how you split the data, how you built and evaluated the model, the Parameters panel for the Decision tree operator, all the corresponding performance metrics as well as the visualization of the decision tree).**