Data Mining – Final Project

Due Date: Tuesday, December 17

You and your team have been hired by a new private school near Villanova (ClearPath) to help improve its efforts at acquiring new students. The school serves children in preschool through grade 12 and would like to promote its programs to those households most likely to enroll (i.e., those with college educated parents earning over $100K income and children under 18 years old). The file households_1.xlsx contains a subset of data collected on households in regions where ClearPath expects to draw students. Each record provides information on a household's demographic characteristics, location, dwelling, and mortgage (data that the school believes it can readily obtain from third party data brokers). Although education level, household income, and children ages are not explicitly available, ClearPath hopes the additional variables in the dataset may be useful in the prediction process. For clarity the 32 variables in the dataset are defined below.

Demographic data for head of household:

1 – SEX: gender (1=male, 2=female)
2 – MSTATUS: marital status (1=married with spouse present, 2=married with spouse absent, 3=widowed, 4=divorced, 5=separated, 6=never married)
3 – CHILD: number of children (any age)
4 – AGE: age in years
5 – BIRTH: country of birth (57=US, 72=Puerto Rico, 109=France, 110=Germany,…)
6 – RACE: race (letting W=White, B=Black, I=American Indian, A=Asian, and H=Hawaiian: 1=W only, 2=B only, 3=I only, 4=A only, 5=H only, 6=W/B, 7=W/I, 8=W/A, 9=W/H, 10= B/I, 11=B/A, 12=B/H, 13= I/A, 14=A/H, 15=W/B/I, 16=W/B/A, 17=W/I/A, 18=W/A/H, 19=W/I/A, 21= other combination of 4 or 5 races)
7 – SP: spanish origin (1=yes, 2=no)
8 – OWN: whether the occupant ever owned a home before (1=yes, 2=no, D= don't know, R=not answered)
9 – VEH: number of vehicles kept for use by household

Location data:

10 – SLOC: geographic location of school relative to household (1=north, 2=east, 3=south, 4=west)
11 – DWELLRate: rating of dwelling as a place to live (10= best, 1=worst)
12 – NBHDRate: rating of neighborhood as place to live (10= best, 1=worst)
13 – VLoINC: limit on what constitutes "very low income" for the surrounding region
14 – LoINC: limit on what constitutes "low income" for the surrounding region
15 – MedINC: median income for the surrounding region

Dwelling data:

16 – NSQFT: square footage of unit
17 – NBATHS: number of bathrooms
18 – NBEDS: number of bedrooms
19 – NROOMS: # of rooms in unit
20 – STORIES: # of floors in building

21 – LOTSIZE: square footage of lot
22 – CONDO: whether unit is condominium or cooperative (1=yes, 2=no)
23 – GARAGE: garage or carport included (1=yes, 2=no)
24 – SEWER: unit connected to public sewer (1=yes, 2=no)
25 – YRBUILT: year unit was built
26 – ESTIMATE: Zillow value of the unit

Mortgage data:

27 – PurchYR: year unit bought/obtained/received
28 – MTGd: down payment percentage (0=none, 1=0-2%, 2=3-5%, 3=6-10%, 4=11-15%, 5=16-20%, 6=21-40%, 7=41-99%, 8=100%, B=not applicable, D= don't know)
29 – MTGdS: main source of down payment (1=sale of previous home, 2=savings, 3=sale of other investment, 4=borrowing, 5=gift, 6=land used for financing, 7=other 8=none, B=not applicable, D= don't know, R=not answered)
30 – HELOC: has a home equity line of credit (1=yes, 2=no)
31 – HELUMP: has a lump sum home equity loan (1=yes, 2=no)

Dependent variable:

32 – HiEdInCh: household with college educated parent(s), income over $100K, and at least one child under 18 years old (0=no, 1=yes)

Project

ClearPath has commissioned your team to develop and compare a variety of data mining models with the goal of accurately predicting whether a household has college educated parent(s), household income over $100K, and at least one child under 18 years old. This information will help the group to more effectively utilize its marketing resources in the future, letting it focus on warm leads (i.e., those households which the school believes are most likely to enroll children in its programs). ClearPath has indicated that their goal is to maximize the monetary value of its predictions (i.e., maximize revenue from correct predictions less the cost of prediction errors). The revenue and costs are tied to the confusion matrix. The school estimates that only 2% of warm lead households marketed to will enroll a student. Each new student enrolled has an average lifetime value of $90,000 to ClearPath, so the average revenue for each warm lead is $1800 ($90k * 2%). Since the school is new and needs to fill its roster, it is willing to spend significant marketing budget on its warm leads, including a champagne brunch reception and golf outing for its warm leads at an average cost of $500 per warm lead. Therefore, the profit for correctly predicting a warm lead is $1300 ($1800 revenue - $500 cost = $1300 profit for a true positive). The cost for incorrectly predicting someone is a warm lead when in fact they are not (i.e., a false positive) is $500 since these folks are happy to drink champagne and play golf but do not enroll their kids. Accurately predicting a non-warm lead (true negative) and incorrectly predicting someone is a non-warm lead (false negative) have a profit/cost of $0 since neither of these types of households are marketed to, and as a result do not enroll even if they might actually be warm leads.

The school has given your team an initial dataset with 10,000 records. In four weeks (on November 11) your team will receive another 10,000 records (without dependent variable values). Teams will generate and send their (monetary maximizing) predictions for these

records to Dr. Strandberg alicia.strandberg@villanova.edu) by the end of the day on Monday, November 25 to compile a mid-competition leaderboard. Specifically, each team will submit one Excel workbook with its "best" predictions for each of the 10,000 households in the 2nd round dataset in one column (indicating whether each of those households is or is not a "warm lead" for enrolling children at ClearPath). The following week (on Monday, December 2) you will receive actual dependent variable values for the 2nd round data AND the final hold out sample of 8,000 records (without dependent variable values).

Each team will submit a final report detailing its analysis on Tuesday, December 17. The final report will contain the following:

1. A section that describes any changes you made to the data set – for example, transforming variables, binning variables, or eliminating outliers.
2. A section that summarizes each broad class of modeling types that you conducted – for example, regression-based techniques, decision trees, bootstrap forest, boosted trees, neural nets, and any other types of data mining models.
3. A section that describes the actual model that you selected to make your "best" predictions.
4. Three actionable insights that you learned from doing this analysis that the company could use to improve their operations.
5. A table that lists all the models that you tried. The table is contained in the Word file called Data Mining Project Results Table Summary.

In addition to your formal report each team will submit one Excel workbook with its "best" (i.e., expected profit maximizing) predictions for each of the 8,000 households in the final hold out sample (again, in one column indicating whether each of those households is or is not "HiEdInCh") and one JMP file with its "best" model. Your grade on the project will be based on the scope, depth, and clarity of your analysis, the quality of your write-up, and the total expected revenue less cost of errors under your predictions.

Timeline

Monday, 11/11:	Second dataset distributed (10K records)
Monday, 11/25:	Send your teams' "best" predictions for each of the 10,000 households in the 2nd round dataset in one column (indicating whether each of those households is or is not "HiEdInCh") to Dr. Strandberg (alicia.strandberg@villanova.edu)
Monday, 12/2:	Third dataset distributed (8K records)
Tuesday, 12/17:	Final report and predictions due