Data Mining Competition
Fall 2018

You and your team have been hired by a financial advisory group (WestCap) to help improve its efforts at acquiring investors for 529 college savings plans. WestCap runs and operates these plans exclusively in Mountain and Pacific states; Montana, New Mexico, Nevada, Oregon, Utah, Washington, and Wyoming. Historically the firm has achieved higher acquisition rates when it markets to households most likely to desire college savings plans (i.e., households with income over $100K and children under 18 years old). Focused outreach is particularly important to WestCap because its agents' employ a high touch sell that can be time consuming and costly. The file WestCap1.xlsx contains a subset of data collected on households in the Mountain and Pacific states. Each record provides information on a household's dwelling, mortgage, and demographic characteristics (data that the company has obtained from a third party data broker). Although household income and ages are not explicitly available, WestCap hopes the additional variables in the dataset may be useful in the prediction process. For clarity the 26 variables in the dataset are defined below.

Dwelling data:

1 – State: state where the household resides (MT, NM, NV, OR, UT, WA, WY)
2 – Internet: access to the internet (1= yes with subscription, 2=yes without subscription, 3=no)
3 – Lot Size: lot size (1=less than one acre, 2=one to ten acres, 3=more than ten acres)
4 – Bedroom: number of bedrooms
5 – Units: units in structure (bb=N/A, 1=mobile home, 2=one family house detached, 3=one family house attached, 4=two apartments, 5=three to four apartments, 6=five to nine apartments, 7=10-19 apartments, 8=20-49 apartments, 9=50+ apartments, 10=boat or RV
6 – Vehicles: number of vehicles owned (b=N/A, 0, 1, 2, 3, 4, 5, 6=six or more)
7 – BroadBND: mobile broadband plan (b=N/A, 1=Yes, 2=No)
8 – Electricity: monthly electricity cost (1=included in condo/HOA fee, 2=none, 3-999 = $3-$999)
9 – FiberOPT: fiber-optic internet service (b=N/A, 1=Yes, 2=No)
10 – HeatFuel: house heating fuel (b=N/A, 1=utility gas, 2=tank, 3=electricity, 4=fuel oil, 5=coal, 6=wood, 7=solar, 8=other, 9=none used)
11 – Rooms: #rooms
12 – Water: annual water cost (0=$0, 1=$1,..., 9999 = $9999+)
13 – Built: year structure was built (1: <=1939, 2: 1940-1949, 3: 1950-1959, 4: 1960-1969, 5=1970-1979, 6=1980-1989, 7=1990-1999, 8=2000-2004, 9=2005, 10=2006, 11=2007,…19=2015)
14 – PropValue: estimated property value from Zillow (in dollars)
15 – PropTax: property tax (01: none, 02: $1 – 49,…, 22: $1000 – 1099,…, 32: $2000–2099,…, 42: $3000-3099,..., 52: $4000-4099,…, 62: $5000-5499, 63: $5500-5999,…, 64: $6000-6999,…, 67: $9000-9999, 68: $10000+)

Mortgage data

16 – Tenure: status of ownership (b=N/A, 1=owned with mortgage, 2=owned free and clear, 3=rented, 4=occupied without payment of rent)
17 - JuniorMtg: second mortgage or home equity status (1=second mortgage, 2=home equity loan, 3=no, 4=both second mortgage and home equity loan)

Demographic data for head of household:

18 – FamEmp: family type and employment status (1=Married-couple family: Husband and wife in labor force (LF), 2=Married-couple family: Husband in labor force, wife.not in LF, 3= Married-couple family: Husband not in LF,wife in LF, 4 =Married-couple family: Neither husband nor wife in LF, 5=Other family: Male householder, no wife present, in LF, 6=Other family: Male householder, no wife present, not in LF, 7=Other family: Female householder, no husband present, in LF, 8=Other family: Female householder, no husband present, not in LF)

19 – LANG: household language (1= English only, 2=Spanish, 3=Other Indo-Europen languages, 4=Asian and Pacific Island languages, 5= Other languages)

20 – Family: household and family type (1=Married couple household, 2=Other family household:male head of household (HH), no wife present, 3=Other family household: female HH, no husband present, 4=Nonfamily household:Male HH:Living alone, 5 =Nonfamily household: Male HH:Not living alone, 6 =Nonfamily household:Female HH:Living alone, 7=Nonfamily household:Female HH: Not living alone)

21 - Move: Length of time at current residence (1: <=12 months, 2= 13-23 months, 3=2-4 years, 4=5-9 years, 5=10-19 years, 6=20-29 years, 7=30+ years)

22 - Npersons: Number of persons in family

23 - Workers: Workers in family during the past 12 months (0= none, 1, 2, 3=3+ workers in family)

24 - WkExp: Work experience of head of household (HH) and spouse (1 =HH and spouse worked FT, 2 =HH worked FT; spouse worked < FT, 3 =HH worked FT; spouse did not work, 4 =HH worked < FT; spouse worked FT, 5 =HH worked < FT; spouse worked < FT, 6 =HH worked < FT; spouse did not work, 7 =HH did not work; spouse worked FT, 8 =HH did not work; spouse worked < FT, 9 =HH did not work; spouse did not work, 10 =Male HH worked FT; no spouse present, 11 =Male HH worked < FT; no spouse present, 12=Male HH did not work; no spouse present, 13 =Female HH worked FT; no spouse present, 14 =Female HH worked < FT; no spouse present, 15 =Female HH did not work; no spouse present)

25 - WkStatus: Work status of HH or spouse (1 =Husband and wife both in LF, both employed or in Armed Forces, 2 =Husband and wife both in LF, husband employed or in Armed Forces, wife unemployed, 3 =Husband in LF and wife not in LF, husband employed or in Armed Forces, 4 =Husband and wife both in LF, husband unemployed, wife employed or in Armed Forces, 5 =Husband and wife both in LF, husband unemployed, wife unemployed, 6 =Husband in LF, husband unemployed, wife not in LF, 7 =Husband not in LF, wife in LF, wife employed or in Armed Forces, 8 =Husband not in LF, wife in LF, wife unemployed, 9 =Neither husband nor wife in LF, 10 =Male HH with no wife present, HH in LF, employed or in Armed Forces, 11 =Male HH with no wife present, HH in LF and unemployed, 12 =Male HH with no wife present, HH not in LF, 13 =Female HH with no husband present, HH in LF, employed or in Armed Forces, 14 =Female HH with no husband present, HH in LF and unemployed, 15 =Female HH with no husband present, HH not in LF)

Dependent variable

26 – WarmLead: household income over $100K and children under 18 (1=yes, 0=no)

Project

WestCap has commissioned your team to develop and compare a variety of data mining models for predicting whether a household has income over $100K and at least one child under 18 years old (i.e., a "warm lead" household). This information will help the group to more effectively utilize its marketing resources in the future, letting it focus on warm leads which the company believes are most likely to open a 529 college savings plan. WestCap 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 and therefore the corresponding profit matrix. The firm estimates that only 2% of warm lead households marketed to will open a 529 plan. Each new plan has an average lifetime value of $75,000 to WestCap, so the average revenue for each warm lead is $1500 ($75k * 2%). Since the firm wishes to brand itself in these regions, it is willing to spend significant marketing budget on its warm leads, including a 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 $1000 ($1500 revenue - $500 cost = $1000 profit for a true positive). The cost for incorrectly predicting a warm lead, a false positive, is $500 since these folks are happy to attend the brunch and play golf but do not open a 529 plan. Accurately predicting a non-warm lead (true negative) and incorrectly predicting a non-warm lead (false negative) has 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 firm has given your team an initial dataset with 10,000 records. In one month (on November 12) 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 Thursday, November 29 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 WestCap). The following week (on Monday, December 3) 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 turn in a one page final report detailing its analysis on Tuesday, December 18. The final report will contain the following:

1. A section that summarizes the software used and 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.
2. A section that describes the actual model that you selected to make your "best" predictions.

In addition to your formal report each team will submit one Excel workbook with its "best" (i.e., expected revenue less error cost 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 a "WarmLead"). The winning teams will be selected by the total expected revenue less cost of errors under your predictions. The winning team will have the highest calculated value here. See FitnessFunctionProfitNonEqualErrorCosts.pptx for calculation details.

Timeline

Wednesday, 10/10:	Initial dataset distributed (10K records)
Monday, 11/12:	Second dataset distributed (10K records)
Thursday, 11/29:	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 "WarmLead") to Dr. Strandberg (alicia.strandberg@villanova.edu)
Monday, 12/3:	Third dataset distributed (8K records)
Tuesday, 12/18:	Final report and predictions due