Oakhurst Data Analysis

CDRL: A002 – Final Analysis

Performed by Group 19 Members:

Ashley Braun

Holly Kelly

Eric Tooke

Christopher Wezdenko

**Introduction:**

Group 19 was approached by Oakhurst financial advisors with a consultation request. Oakhurst is in the market for a new method of categorizing and acquiring new customers into their thriving business. Group 19 has been requested to help with the first part of the new mission statement, categorizing potential customers using historical data archived by Oakhurst. Oakhurst’s process for acquiring new customers requires advisors to expend significant amount of time and personal attention to each customer acquired. The cost of acquisition necessitates the categorization algorithm is as accurate as possible with the data provided.

**Background:**

Oakhurst has acquired a data set about residence in the state of California from a third-party data broker. Each of the records contains information pertaining to demographic information about the resident including households dwelling and mortgage. Oakhurst has acquired 40,000 records from the data broker and split the data into three parts:

1. 15,000 records for initial model creation
2. 15,000 records for model validation and tuning
3. 10,000 records for model verification

The first part of the data was provided to Group 19 for creation of the initial model and included a binary column labeled HiWorth which corresponds to records where the household income is greater than $150,000 and home value over $400,000. Group 19’s model must use the rest of the demographic information in the record to predict HiWorth. The second part of the data was provided to Group 19 after submission of the initial model to Oakhurst for validation and tuning. The third part of the data was retained by Oakhurst for verification of the model submitted by Group 19 after the validation phase of the project.

**Initial Model Generation:**

Group 19 started the process of model creation by importing the first part of the data set into JMP Pro 13.0.0. During the import process, JMP assigned the columns as continuous values. Based on the data dictionary provided by the third-party vendor, the following corrections where applied to the data:

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Initial Data Type | Corrected Data Type | Reason for Change |
| Lot | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| Bedroom | Continuous | Continuous | No Change |
| Units | Continuous | Nominal | Numbers refer to categories |
| Electric | Continuous | Nominal | Numbers refer to categories with embedded pricing in some cases |
| Internet | Continuous | Nominal | Numbers refer to categories |
| FiberOpt | Continuous | Nominal | Numbers refer to categories |
| HeatFuel | Continuous | Nominal | Numbers refer to categories |
| Rooms | Continuous | Continuous | No Change |
| Water | Continuous | Continuous | No Change |
| Built | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| Tenure | Continuous | Nominal | Numbers refer to categories |
| JuniorMtg | Continuous | Nominal | Numbers refer to categories |
| FamEmp | Continuous | Nominal | Numbers refer to categories |
| Lang | Continuous | Nominal | Numbers refer to categories |
| Family | Continuous | Nominal | Numbers refer to categories |
| HHpresence | Continuous | Nominal | Numbers refer to categories |
| Move | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| Npersons | Continuous | Continuous | No Change |
| Nchild | Continuous | Continuous | No Change |
| Under18 | Continuous | Nominal | Numbers refer to categories |
| Over60 | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| Over65 | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| Workers | Continuous | Ordinal | Numbers refer to categories with inherent ordering |
| WkExp | Continuous | Nominal | Numbers refer to categories |
| WkStatus | Continuous | Nominal | Numbers refer to categories |
| Vehicles | Continuous | Continuous | No Change |
| BroadBND | Continuous | Nominal | Numbers refer to categories |
| HiWorth | Continuous | Nominal | Numbers refer to categories |

The changes to the categorization of the columns was discussed and agreed to by all member of Group 19.

While investigating the data dictionary columns Electric, Family, FamEmp, WkExp, and WkStatus contains combined compress demographic information. Each of the columns do not contain the exact same information but in the case of Family, FamEmp, WkExp, and WkStatus some of the same information is referenced. Electric contains continuous values between 3 and 999 for cost of the monthly electric bill but reserve integers 1 and 2 for locations where the electric bill is included in the monthly payment or HOA fee. There was a decision to trying a split the columns to better understand the data. This was accomplished through a JSL script that imported the JMP file with the first part of the data set and created new columns from the original columns. This code can be found in the appendices under the title **JMP JSL code for column breakout**. The script file created the following columns:

|  |  |  |  |
| --- | --- | --- | --- |
| Parent Column | Child Column | Child Data Type | Description of changes |
| Electric | Electric Enumeration | Nominal | New Categories:  included in condo/HOA fee, none,  paid by occupant |
|  | Electric Value | Continuous | Dollar amount paid by resident, values of 1 and 2 replaced with 0 (no bill) |
| Family | Family Martial Status | Nominal | New Categories:  Married, Other Family,  Non Family |
|  | Family Head of Household | Nominal | New Categories:  Male, Female |
|  | Family Head of Household Living Alone | Nominal | New Categories:  Yes, No |
| FamEmp | Marital Status | Nominal | New Categories:  Married, Other Family |
|  | Male Present | Nominal | New Categories:  Yes, No |
|  | Male in Work Force | Nominal | New Categories:  Yes, No |
|  | Female Present | Nominal | New Categories:  Yes, No |
|  | Female in Work Force | Nominal | New Categories:  Yes, No |
| WkExp | Work Experience of HH | Nominal | New Categories:  FT, PT, Did Not Work |
|  | HH Sex | Nominal | New Categories:  UNK, Male, Female |
|  | Spouse Present | Nominal | New Categories:  Yes, No |
|  | Spouse Work Experience | Nominal | New Categories:  FT, PT, Did Not Work |
| WkStatus | Male in Labor Force | Nominal | New Categories:  LF, Not in LF, Not Present |
|  | Male Employment Status | Nominal | New Categories:  Armed Forces, Unemployed,  Not Present |
|  | Female in Labor Force | Nominal | New Categories:  LF, Not in LF, Not Present |
|  | Female Employment Status | Nominal | New Categories:  Armed Forces, Unemployed,  Not Present |

Changes to the data drove the following decisions by Group 19:

* Missing values in FamEmp, WkExp, and WkStatus are due to Family values that contain Non Family as part of the response. Our group believes the collection method did not request the rest of the information once that option was chosen by the household.
* Electric needs a cube root transformation applied for the distributions to be more normal. Also, the Electric Enumeration column is not needed.

Group 19 noticed other changes required in the data:

* Water needs to be binned to correct the distribution
* An Internet value of 3 represents no internet access. Our analysis showed that the missing values in FiberOpt and BroadBND correlated to an Internet value of 3 and thus FiberOpt and BroadBND were recoded with a value of 2 which represents a no response for those attributes.

The missing values in FamEmp, WkExp and WkStatus necessitated the use of the Informative Missing option in JMP. Multiple models were created with the columns transformed as described in the table above, but the misclassification rates were consistently higher than models created without the column breakouts. Group 19 decided to use a data set with the modifications made to FiberOpt and BroadBND, binning of water and the cube root transformation to the Electric columns.

The modeling process started once a validation column was added with 40% of the data coded to training, 30% coded to validation and 30% coded to test. Group 19 build numerous models using Decision Trees, Boosted Trees, Bootstrap Forest and Neural Nets. This process continued until the group was comfortable with the results being seen from these methods and lead to the creation of four best case modes from each category.

Eric’s Add

The data was analyzed for collinearity between variables. The initial data was imported into JMP and initially categorized as continuous variables. Many variables were recategorized as nominal or ordinal as mentioned previously. However, for the purposes of checking for collinearity, multivariate analysis was done with the variables categorized as continuous. Several variables exhibited collinearity and were flagged for removal from potential models as seen in the chart below. When possible candidates for a model were chosen, one variable from each of the collinear pairs or groups was removed from the dataset individually and the model rerun for comparison against the full dataset. Subsequently the model was rerun with data from each of the collinear pairs or groups removed collectively. In all cases tested, the full data set returned a lower test misclassification rate and thus it was determined that, despite collinearity between certain variables, the full data set would be used in creating the final model.

|  |  |  |  |
| --- | --- | --- | --- |
| Bedrooms | Rooms |  |  |
| Nchild | Npersons |  |  |
| Family | FamilyEmp | WkStatus | WkExp |
| Over65 | Over60 |  |  |
| BroadBND | Internet | FiberOpt |  |

Boosted Tree

A Boosted Tree model was chosen for it’s low test misclassification rate using a 60/20/20 split for the validation column. The model was subsequently rerun with several other breakdowns of training, validation, and test for the validation column to determine if a different percentage of training and validation would be more effective and improve the test misclassification rate. The results are listed below. It was determined that using a 60/20/20 split for the validation column resulted in the lowest test misclassification rate.

|  |  |
| --- | --- |
| Training / Validation / Test Percentage | Test Misclassification Rate |
| 60/60/20 | .1700 |
| 40/30/30 | .1713 |
| 40/40/20 | .1747 |
| 50/30/20 | .1725 |
|  |  |
|  |  |
|  |  |

Appendices:

**JMP JSL code for column breakout:**

//Import fuctions to allows for quick ingestion and changes to the data.

//Remember to provided the path to cali1.xlsx on the local machine.

dt = Open("D:\chris\Documents\GitHub\MSA-8220\_Final\_Project\Assignment Documents in Work\Project Initial Data Set - cali1 - 09182017.jmp");

//Converts Electric column into two columns, one for a character enumeration discribing how bill is paid and a

//second column with the payment value per month. The value is zero if included in HOA/condo fees

dt << New Column("Electric Enumeration", Character, Formula(

IF(

:Electric == 1, "included in condo/HOA fee",

:Electric == 2, "none",

:Electric >= 3, "paid by occupant"

)

));

dt << New Column("Electric Value", Numeric, Continuous, Formula(

IF(

:Electric < 3, 0,

:Electric >= 3, :Electric

)

));

dt << New Column("FiberOpt\_Corrected", Numeric, Continuous, Formula(

IF(

Is Missing(:FiberOpt), 2,

:FiberOpt

)

));

dt << New Column("BroadBND\_Corrected", Numeric, Continuous, Formula(

IF(

Is Missing(:BroadBND), 2,

:BroadBND

)

));

dt << New Column("JuniorMtg\_Corrected", Character, Formula(

IF(

Is Missing(:JuniorMtg), "Missing",

Char(:JuniorMtg)

)

));

dt << New Column("Workers\_Corrected", Character, Formula(

IF(

Is Missing(:Workers), "Missing",

Char(:Workers)

)

));

//Converts Family into the following columns (Family Marital Status, Family Head of Household, Family Head of Household living alone)

//Family Enumeration Values:

//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

dt << New Column("Family Marital Status", Character, Formula(

Choose(Family, "Married", "Other Family", "Other Family", "Non Family", "Non Family", "Non Family", "Non Family")

));

dt << New Column("Family Head of Household", Character, Formula(

Choose(Family, "Male", "Male", "Female", "Male", "Male", "Female", "Female")

));

dt << New Column("Family Head of Household living alone", Charactger, Formula(

Choose(Family, "No", "Yes", "Yes", "Yes", "No", "Yes", "No")

));

//Correct NPerson column using Family andNchild

dt << New Column("Npersons Corrected", Numeric, Continuous, Formula(

IF(

AND(OR(Is Missing(:Npersons), :Npersons == "Missing"), :Family == 4), (1+:Nchild),

AND(OR(Is Missing(:Npersons), :Npersons == "Missing"), :Family == 5), (2+:Nchild),

AND(OR(Is Missing(:Npersons), :Npersons == "Missing"), :Family == 6), (1+:Nchild),

AND(OR(Is Missing(:Npersons), :Npersons == "Missing"), :Family == 7), (2+:Nchild),

:Npersons

)

));

dt << New Column("Nperson Corrected and Binned", Character, Formula(

IF(

:Npersons Corrected == 1, "one",

:Npersons Corrected == 2, "two",

OR(:Npersons Corrected == 3, :Npersons Corrected == 4), "three or four",

"five or more"

)

));

dt << New Column("Nchild Binned", Character, Formula(

IF(

:Nchild == 0, "Zero",

OR(:Nchild == 1, :Nchild == 2), "One or Two",

"Three or More"

)

));

//Converts FamEmp to the following columns (Marital Status, Male Present, Male in Work Force, Female Present, Female in Work Force)

//FamEmp Enumeration Values:

//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

dt << New Column("FamEmp\_Corrected", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Char(:FamEmp)

)

));

dt << New Column("Marital Status", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Choose(FamEmp, "Married", "Married", "Married", "Married","Other Family", "Other Family", "Other Family", "Other Family")

)

));

dt << New Column("Male Present", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Choose(FamEmp, "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No")

)

));

dt << New Column("Male in Work Force", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Choose(FamEmp, "Yes", "Yes", "No", "No", "Yes", "No", "No", "No")

)

));

dt << New Column("Female Present", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Choose(FamEmp, "Yes", "Yes", "Yes", "Yes", "No", "No", "Yes", "Yes")

)

));

dt << New Column("Female in Work Force", Character, Formula(

IF(

Is Missing(:FamEmp), "Missing",

Choose(FamEmp, "Yes", "No", "Yes", "No", "No", "No", "Yes", "No")

)

));

//Converts WkExp into the following columns (Work Experience of Head of House Hold, Head of Household Sex, Spouse Present, Spouse Work Experience)

//WkExp Enumeration Values:

//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

dt << New Column("WkExp\_Corrected", Character, Formula(

IF(

Is Missing(:WkExp), "Missing",

Char(:WkExp)

)

));

dt << New Column("Work Experience Head of Household", Character, Formula(

IF(

Is Missing(:WkExp), "Missing",

Choose(WkExp, "FT", "FT", "FT", "PT", "PT", "PT", "Did Not Work", "Did Not Work", "Did Not Work", "FT", "PT", "Did Not Work", "FT", "PT", "Did Not Work")

)

));

dt << New Column("Head of Household Sex", Character, Formula(

IF(

Is Missing(:WkExp), "Missing",

Choose(WkExp, "UNK", "UNK", "UNK", "UNK", "UNK", "UNK", "UNK", "UNK", "UNK", "Male", "Male", "Male", "Female", "Female", "Female")

)

));

dt << New Column("Spouse Present", Character, Formula(

IF(

Is Missing(:WkExp), "Missing",

Choose(WkExp, "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No", "No")

)

));

dt << New Column("Spouse Work Experience", Character, Formula(

IF(

Is Missing(:WkExp), "Missing",

Choose(WkExp, "FT", "PT", "Did Not Work", "FT", "PT", "Did Not Work", "FT", "PT", "Did Not Work", "Did Not Work", "Did Not Work", "Did Not Work", "Did Not Work", "Did Not Work", "Did Not Work")

)

));

//Converst WkStatus into the following columns (Male in Labor Force, Male Employment status, Female in Labor Force, Female Employment Status)

//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)

dt << New Column("WkStatus\_Corrected", Character, Formula(

IF(

Is Missing(:WkStatus), "Missing",

Char(:WkStatus)

)

));

dt << New Column("Male in Labor Force", Character, Formula(

IF(

Is Missing(:WkStatus), "Missing",

Choose(WkStatus, "LF", "LF", "LF", "LF", "LF", "LF", "Not in LF", "Not in LF", "Not in LF", "LF", "LF", "Not in LF", "Not Present", "Not Present", "Not Present")

)

));

dt << New Column("Male Employment Status", Character, Formula(

IF(

Is Missing(:WkStatus), "Missing",

Choose(WkStatus, "Armed Forces", "Armed Forces", "Armed Forces", "Unemployed", "Unemployed", "Unemployed", "Unemployed", "Unemployed", "Unemployed", "Armed Forces", "Unemployed", "Unemployed", "Not Present", "Not Present", "Not Present")

)

));

dt << New Column("Female in Labor Force", Character, Formula(

IF(

Is Missing(:WkStatus), "Missing",

Choose(WkStatus, "LF", "LF", "Not in LF", "LF", "LF", "Not in LF", "LF", "LF", "Not in LF", "Not Present", "Not Present", "Not Present", "LF", "LF", "Not in LF")

)

));

dt << New Column("Female Employment Status", Character, Formula(

IF(

Is Missing(:WkStatus), "Missing",

Choose(WkStatus, "Armed Forces", "Unemployed", "Unemployed", "Armed Forces", "Unemployed", "Unemployed", "Armed Forces", "Unemployed", "Unemployed", "Not Present", "Not Present", "Not Present", "Armed Forces", "Unemployed", "Unemployed")

)

));

**Continuous and Electric Distribution Data:**

