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 one household’s demographic information. 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 is required to 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. During the course of investigation of the data provided, and attempt was made to split the columns to better understand the data stored in the column. This split was also an attempt to create a more accurate model. This split 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
* The missing values in FiberOpt and BroadBND correlated to an Internet value of three. This value states the household does not have internet access. The missing values are recoded with a value of 2 in the FiberOpt and BroadBND columns

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. During the modeling process, a validation column with 60% training, 20% validation and 20% test was tried and found to decrease the misclassification rate of the models. This change lead to the creation of the four best models found by our group:

* Neural Net – misclassification test rate = 0.1623
* Decision Tree – misclassification test rate = 0.1841
* Boosted Tree
* Bootstrap Forrest

The final model was a Boosted Tree that used a tuning table to achieve a misclassification rate of .1617 and in testing achieved a misclassification rate of .1742 against the second part of the data provided by the Oakhurst.

**Model Creation, Phase 2:**

Group 19 armed with 30,000 began to look at changes that could be made to the initial model which would result in increased accuracy. With the data combined, five validation columns were created for modeling purposes:

* Validation\_1 – 60,20,20 split with no random seed set
* Validation\_2 – 40,30,30 split with no random seed set
* Valdiation\_1Seed5 – 60,20,20 split with no random seed of 5 set
* Validation\_2Seed5 – 40,30,30 split with a random seed of 5 set

During our evaluation of models, it was noted that changing the validation column and seed was causing changes to the misclassification rate by as much as 2 percent. This phenomenon is not well understood by Group 19, the result should not very that much with the size of data set being used. Group 19 cannot determine if the phenomenon is an artifact of the JMP software of due to the data set. Group 19 made the decision to try all the validation columns and determine the best misclassification rate compared to the first model. Group 19 selected Validation\_1\_Seed5 and all future models will be compared against each other using the same validation column. This allows for the removal of error due to changing validation columns by inducing the same error in all the models (“All models are wrong, some are just useful” -George Box).

With this decision made, Group 19 proceeded to create new models. Group 19 decided to run more neural nets to determine if a lower misclassification rate was achieved. These neural nets took a long amount of time to run but did not achieve a lower misclassification rate than the boosted tree. Group 19 then decided to create multiple tuning tables for the boosted tree model to determine the changes in misclassification rate due to variation of each variable. This process resulted in the creation of 16 different tuning tables for the boosted trees with the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| Validation Column | Hypercube | Random Seed | Misclassification Rate |
| Validation\_1\_Seed5 | None | 5 | 0.1762 |
| Validation\_1\_Seed5 | 2 | 5 | 0.1742 |
| Validation\_1\_Seed5 | 5 | 5 | 0.1768 |
| Validation\_1\_Seed5 | 6 | 5 | 0.1702 |
| Validation\_1\_Seed5 | 17 | 5 | 0.1708 |
| Validation\_1\_Seed5 | 7 | 5 | 0.1700 |
| Validation\_1\_Seed5 | 8 | 5 | 0.1737 |
| Validation\_1\_Seed5 | 9 | 5 | 0.1732 |
| Validation\_1\_Seed5 | 10 | 5 | 0.1743 |
| Validation\_1\_Seed5 | 11 | 5 | 0.1718 |
| Validation\_1\_Seed5 | 12 | 5 | 0.1740 |
| Validation\_1\_Seed5 | 13 | 5 | 0.1773 |
| Validation\_1\_Seed5 | 14 | 5 | 0.1713 |
| Validation\_1\_Seed5 | 15 | 5 | 0.1712 |
| Validation\_1\_Seed5 | 16 | 5 | 0.1740 |
| Validation\_1\_Seed5 | 18 | 5 | 0.1700 |
| Validation\_1\_Seed5 | 19 | 5 | 0.1727 |
| Validation\_1\_Seed5 | 20 | 5 | 0.1712 |
| Validation\_40\_40\_20\_Seed5 | None | 5 | 0.1805 |
| Validation\_40\_40\_20\_Seed5 | 18 | 5 | 0.1772 |

The boosted tree changes achieved a misclassification rate of .1700. With these results, Group 19 decided to look at the assumptions made during the initial modeling effort. The current model uses all the base columns except Electric and Water, these columns required transformation to make the distribution more normal. Electric was changed to a cube root of Electric and Water was binned. Group 19 discussed the changes and decided that it still applied. The conversation then moved to collinearity between the variables. The following sets of columns contain mutual information:

* Family, FamEmp, WkStatus, WkExp
* Internet, FiberOpt, BroadBND
* Rooms, Bedrooms
* Over65, Over60

For determining collinearity, these columns where changed back to continuous so that a multivariate analysis could be performed against them. During the analysis, collinearity was found between the following sets of columns:

* Family, FamEmp, WkStatus, WkExp
* Rooms, Bedrooms
* Over65, Over60

Internet, FiberOpt and BroadBND surprised Group 19 because the missing data in FiberOpt and BroadBND was filled in using the values found in Internet but the analysis but the columns are not collinear. Using the results from the boosted tree models created with combined part one and part two data, the following changes were made to the model:

* Removal of one of the collinear columns at a time and re-run the model
* Remove complete sets of the colinear columns and re-run the model
* Removal of all the collinear columns and re-run the model

The results of the models did not decrease the test misclassification rate seen in the boosted trees. Group 19 believes that the columns do have some collinearity but not enough to require removal from the model.

**Final Model:**

Based on the results from the second phase of modeling, Group 19 has chosen a Boosted Tree model using the 7 Latin Hypercube Model created. This model has a test misclassification rate of 17% and based on the approximately 50 to 100 models created, Group 19 believes this model will play a part in achieving the stated goals of the Oakhurst group to assist their advisors with acquisition of new customers with the correct demographic background ($150,000 annually salary and home value $400,000 dollars).

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")

)

));