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 file named Final Project Script.jsl. 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 without a random seed set 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 = 0.1813
* Boosted Tree – misclassification test rate = 0.1617
* Bootstrap Forrest – misclassification test rate = 0.173

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 records began to look at changes that could be made to the initial model which would result in increased accuracy. With the combined data, 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
* Validation\_1Seed5 – 60,20,20 split with a 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 or 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 shows 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 collinear 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.

The results of the boosted tree modeling effort end with two models having the exact same misclassification rate (7 and 18). The difference between these boosted trees is evident when comparing how many records are predicted to be HiWorth = 1. Latin Hypercube Model 7 predicts a total of 530 records with 329 records correctly identified and 201 records misclassified. Latin Hypercube Model 18 predicts a total of 598 records with 363 records correctly identified and 235 records misclassified. The actual percentage of HiWorth = 1 records in the data set is approximately 20%, our models are predicting 8.8% and 9.96% records with HiWorth = 1 for our test sets. Based on these results, our models are still under-predicting the number of individuals with HiWorth scores but Latin Hypercube Model 18 is more accurate than Latin Hypercube Model 7. Group 19 then proceeded to determine if a change in the cutoff value for the tree model would lead to a model with a prediction rate closer to the 20/80 split that exists in the combined part one and part two data set. Using a JMP add-in called altcutoffconfusionmtx, the cut off values were varied from .4 to .6 in increments of .01. The result of this analysis is shown in the table below:

|  |  |  |
| --- | --- | --- |
| Cut-Off Value | Misclassification Rate | Correct HiWorth = 1 Rate |
| 0.4 | 17.74% | 8.74% |
| 0.41 | 17.50% | 8.53% |
| 0.42 | 17.32% | 8.33% |
| 0.43 | 17.28% | 8.05% |
| 0.44 | 17.13% | 7.82% |
| 0.45 | 17.07% | 7.58% |
| 0.46 | 17.15% | 7.18% |
| 0.47 | 17.18% | 6.80% |
| 0.48 | 17.20% | 6.47% |
| 0.49 | 17.05% | 6.28% |
| 0.5 | 17.00% | 6.05% |

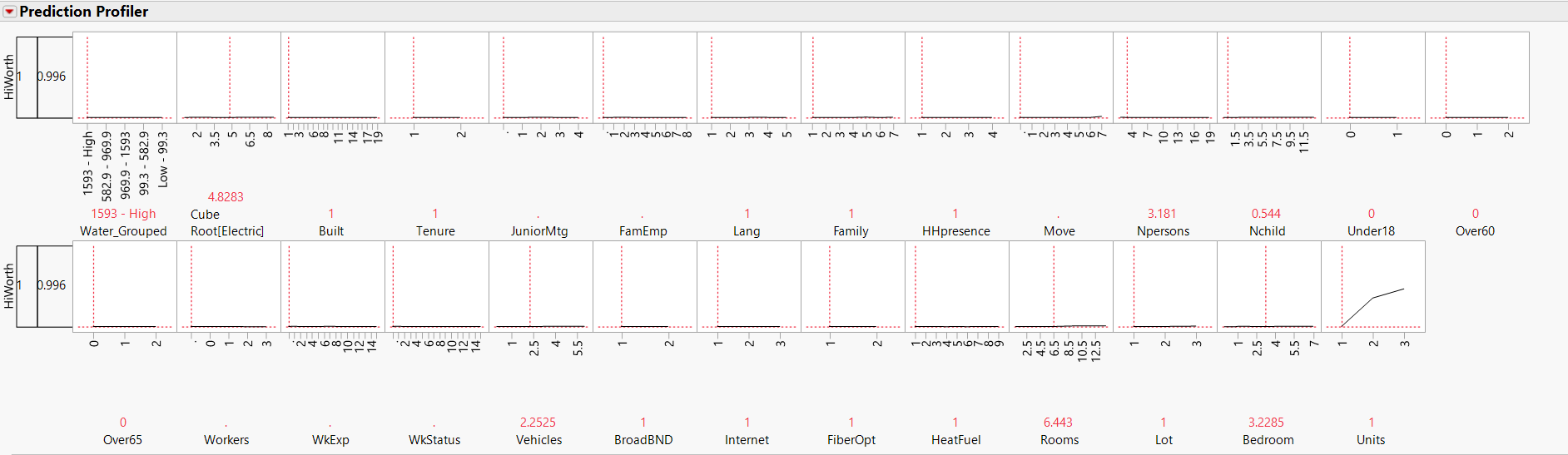
Based on the goal of optimizing Oakhurst outreach campaign, Group 19 recommends changing the cut off value to 0.45 from 0.50. The misclassification rate of the data increases to 17.07% from 17.00% but the percentage of HiWorth = 1 predictions that are correct go from 6.05% to 7.58%. This increase in correct HiWorth = 1 predictions is well worth the .0007% increase in the misclassification rate.

**Final Model:**

Based on the results from the second phase of modeling, Group 19 has chosen a Boosted Tree model using the Latin Hypercube Model 18 tuning table with a cut-off point of 0.45. This model has a test misclassification rate of 17.07% 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).

**Model Observations:**

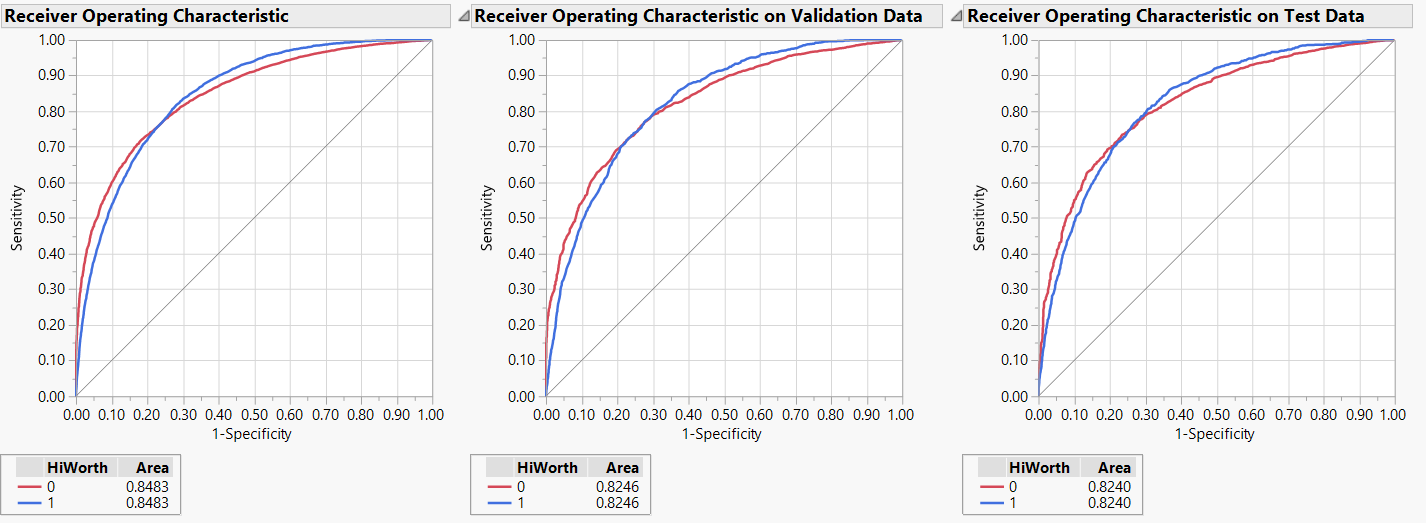
The Group 19 final model has some interesting characteristics that can be used for insight into the household demographic. Using the predication profiler with the default settings, the only column variable that makes probability changes to the HiWorth prediction model is Unit. Unit requires a value of 2 or 3 to have an effect. A value of Unit equal to two corresponds to a single-family home detached and the value of three corresponds to single family home attached. The prediction profiler is shown here:



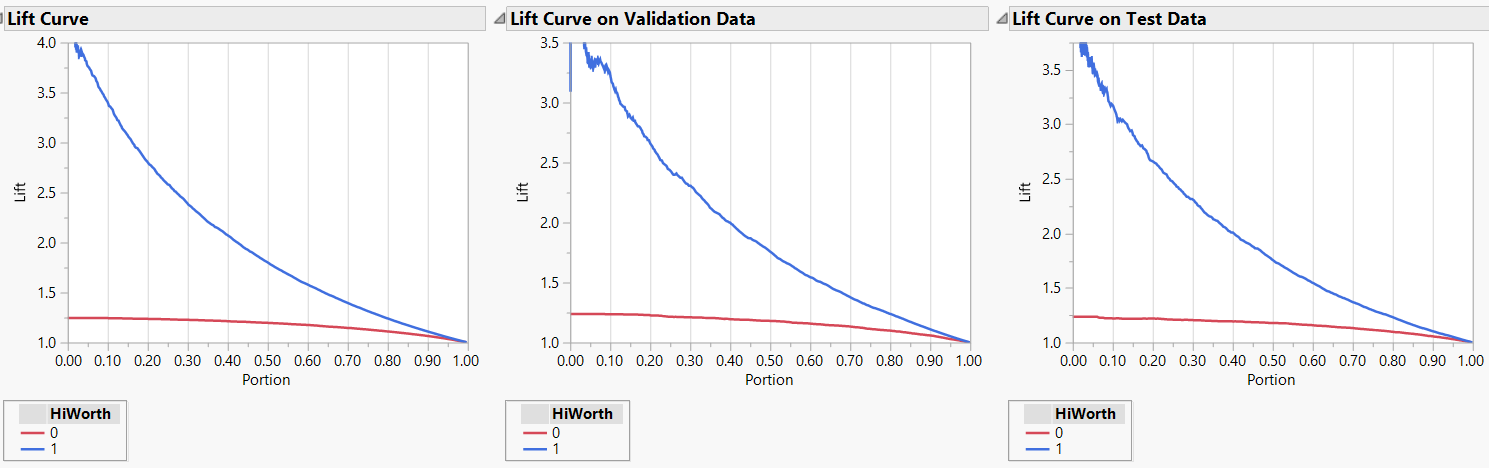
When the value of two is selected in the unit selector, all the variables except HHpresence and Tenure have an effect on the result of the prediction. The same behavior is observed when Unit equal to three is selected except HHpresence and Tenure influence the result of the prediction. The prediction profiler is shown below:



Group 19 also reviewed the Receiver Operating Characteristic (ROC) for the final model. The area under the curve (AUC) for the training, validation and test set is 0.8483, 0.8246 and 0.8240 respectively. The AUC value should be as close to one as possible, a value of one corresponds to a perfect fit of the model to the data.



The final model was also evaluated using the Lift Curves. Based on the test data, using a 30% proportion of the data, the model is 2.3 time more accurate than by picking the values at random. The lift curves are shown below:



What does this information mean for Oakhurst’s advisors when it comes to bringing new customers into the business? The final model has a decent fit to the data and it provided more accurate insights than randomly choosing the households to acquire. The advisors should be focusing on single-family home, ether detached or attached, as a starting point. From that data, the possibility of a HiWorth household can be determined by considering a combination of factors from the prediction profiler for the type of unit in question. All this data provided to a trained advisor pool should result in a lower cost of acquisition of new accounts for Oakhurst.