

Churn the Data Around – A Machine Learning Approach to Understanding Why Customers Leave

Sean Ankenbruck, Zencos Consulting

ABSTRACT

Customer attrition is a common problem that affects companies across a variety of industries. Companies value profitable customers that are consistent and engage in long-term interactions that contribute to a high customer lifetime value. Identifying and understanding the factors that influence customer churn allows companies to gain a competitive advantage. This paper discusses a data-driven approach to identify important factors that influence churn and classify individuals based on their predicted likelihood to churn. It also provides a visual analysis of churn rates across various customer groups. Depending on whether the goal is purely accuracy, or a correct classification, there are several methods that can be used to identify customer churn. The entire process is accomplished using SAS® tools such as SAS Visual Analytics with SAS Visual Data Mining and Machine Learning® (VDMML).

INTRODUCTION

Customer attrition is the rate at which customers leave a company. Many companies that operate subscription models – banking, telecommunications, ecommerce, etc. – can use customer attrition as a metric to examine the health of their overall business. The actual rate at which customers churn varies by industry but for this analysis we looked specifically at customer churn for Oberweis Dairy (Oberweis).

Oberweis specializes in the home delivery of premium dairy products. They use a subscription model to distribute these products to their customers throughout the Midwest and now in certain areas in the Southeast. Customer retention is an important factor in their business and those customers who engage in a subscription for longer periods of time provide more overall value to the company. How does Oberweis classify churn? Their marketing department uses USPS carrier route information to determine which routes to target in direct mail campaigns. They have determined that for a direct mail campaign to be considered a success a customer that was acquired through the campaign must maintain their account for at least 6 months. This duration of time is considered the breakeven point for costs associated with implementing a marketing campaign. For this analysis, we used the variable *AcctOpen_6Mos* as the target variable. A value of 1 indicates that a customer was retained for at least 6 months and a value of 0 indicates that a customer churned, or left the company.

The purpose of this analysis was to use the data provided by Oberweis to do two things: identify those factors that are most important in predicting churn and classify individuals within a carrier route with their probability of churn. We used three different models in this analysis: logistic regression, decision trees, and supervised neural networks.

DATA PREPARATION

Overall, the ratio of retention (1) to churn (0) was 72.29% to 27.71%. The detailed data includes a row for each existing or former customer and a collection of attributes for the carrier route – or CART – associated with the customer. Since the demographic data was captured on a geographic scale by carrier route, we decided to aggregate the individual rows to the CART level. However, in doing so we had to develop a way to keep track of the number of customers who were retained as well as those who had churned within each CART using a frequency count variable. Frequency coding involves creating a unique row for each level of the target variable and within each row creating an additional variable that is an integer that represents the number of observations in the raw data that each new row represents. In Table 1, CART 123 is represented by two rows. Row 1 is for level 0 of the churn variable and represents 3 individuals from the raw data. Row 2 is for level 1 of the churn variable and represents 5 individuals from the raw data which results in 8 total people accounted for in CART 1.

CART	Churn	Count
123	0	3
123	1	5

Table 1. Example of Frequency Coding

Most models in SAS Viya®, with the exception of a neural network, allow for the use of the FREQ statement to specify integer frequencies for repeated observations. The FREQ statement allows the models to evaluate the total count of observations represented by each row rather than just evaluating each row as a unique observation. Since this feature was not yet available to the neural network procedure in Viya at the time of our analysis, we built the neural network in SAS EM®.

VISUAL ANALYSIS

One way to examine your data prior to modeling is to examine it visually. The figures below provide two visual aids that help us see what churn looks like across a couple of different variables within our data set.

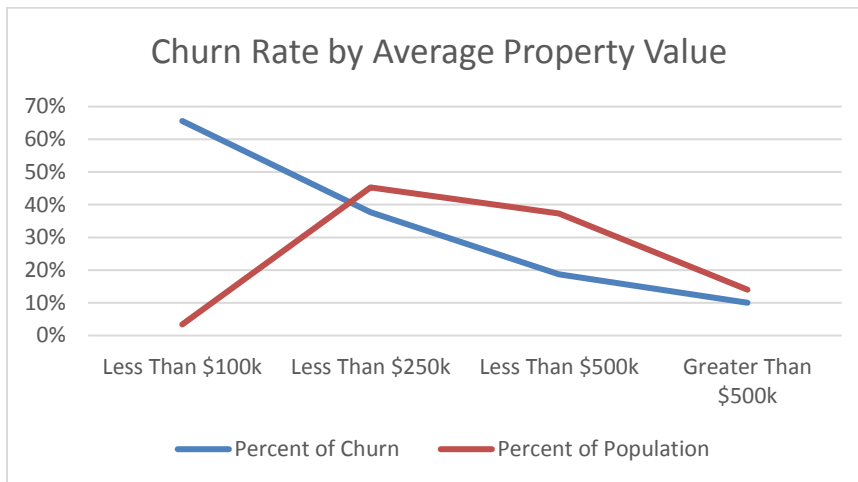


Figure 1: Churn Rate by Average Property Value

Figure 1 above shows the percent of the population represented by each level of the variable average property value as well as the churn rate for each level. The majority of homes fall within the range of \$100,000 to \$500,000. Notice how the churn rate decreases as the average property value of the carrier route increases. It is also possible to see that the largest difference between percent of population and churn rate occurs in the group with average property values between \$250,000 and \$500,000.

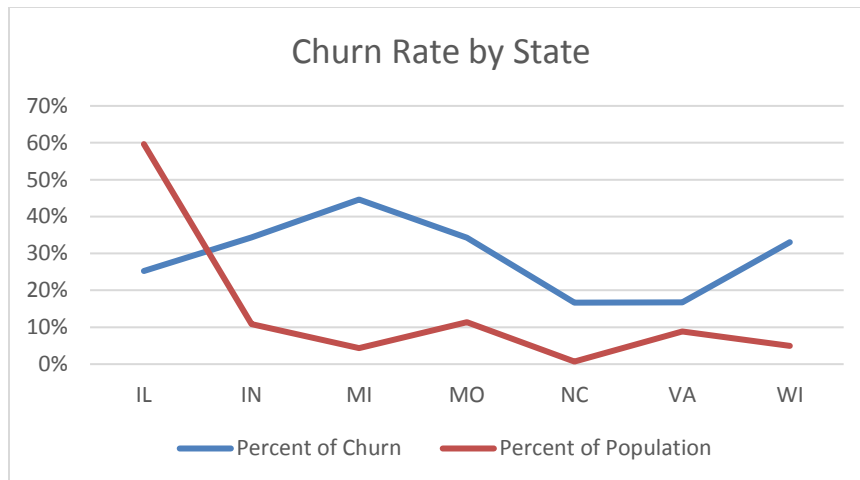


Figure 2: Churn Rate by State

Figure 2 above shows the percent of the population that resides in each state where Oberweis' service is offered and what the churn rate is for each of those states. The largest portion of customers are found in Illinois, where they are based. This chart makes it easy to see the churn rate across the different markets that Oberweis currently operates in.

VARIABLE IMPORTANCE

Variables importance is measured in different ways depending on the model that is used. Generally, variable importance ranks the variables in a model based on how much they are contributing to the predictive power of the model relative to each other. The importance helps to quickly visualize and assess which variables have the strongest relationships with the target variable. For this analysis, we used three different types of models:

- **Logistic Regression:** variable importance is determined by the Wald Chi-Square t-statistic, standardized estimate or partial correlation statistic. A larger value indicates a more useful variable, relative to the other variables included in the model, in the prediction of the target.
- **Decision Trees:** variable importance is determined by the gain in purity at each split of the tree. The variable with the largest potential purity gain is split first, followed by each subsequent variable in a decreasing order of importance.
- **Neural Networks:** variables are combined as inputs into one or more hidden layers and are converted to weights which are then used as predictions of the final output. They can be extremely accurate, but this accuracy comes at the expense of the loss of interpretability of the variables because they are not the same raw variables once they have been input into this type of black box model.

The variable importance for the logistic regression and decision tree models can be found in the following charts. Average property value was the strongest predictor across all the different models that were built. These charts include measurements for 3 different versions of the data – raw, undersampled, and oversampled. The specific differences for these 3 versions of the data are covered in a later section, but the relative variable importance does not vary significantly in many cases.

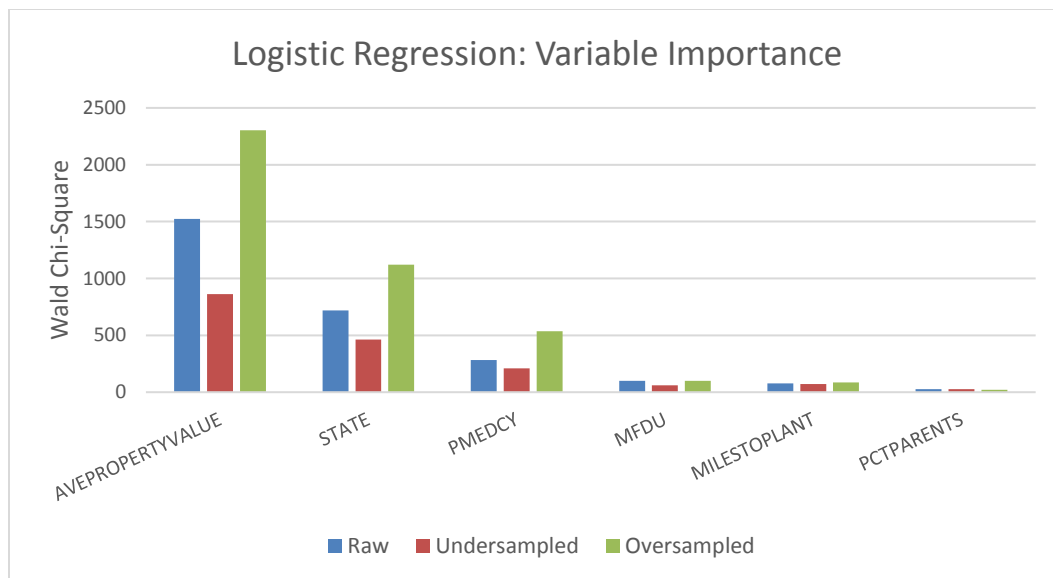


Figure 3. Variable Importance – Logistic Regression Models

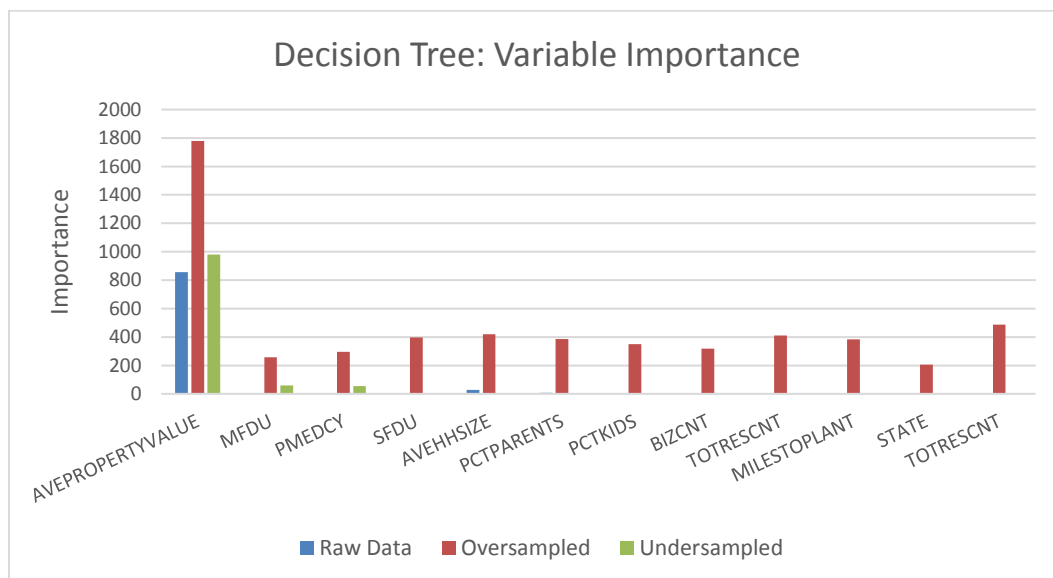


Figure 4. Variable Importance – Decision Tree Models

PREDICTING CHURN

We took several approaches in modeling to determine the best way to accurately predict customer churn. The initial method was to model the frequency data at its normal ratio of churn to retention; this resulted in rather accurate models overall, but they were biased toward the majority class – retention. Since the goal was to predict churn, we decided to try a few other techniques to examine how the predictive power of the models might be enhanced.

INITIAL MODEL

We modeled the frequency data using each of the three models. They were accurate overall, but we noticed a pattern in the results; the algorithms had no problem identifying retained customers, but they had an extremely difficult time accurately classifying churn. Regardless of the model used there were always many false negatives because most of the actual cases of churn had been incorrectly classified as retained. The initial accuracy results can be found in Table 2 below.

Statistic	Logistic Regression	Decision Tree	Neural Network
Accuracy Rate	73.5%	72.1%	73.9%
Misclassification Rate	26.5%	27.9%	26.0%
False Positive Rate	38.4%	N/A*	42.5%
False Negative Rate	25.7%	27.9%	23.6%
Correct Churn Classification	965	0	1,875

Table 2. Initial Model Accuracy

The base models were all quite similar in terms of accuracy. Notice how the false negative rates were all above 20% and the number of churn observations that were accurately classified was very low – with the decision tree model being the worst at 0 correct classifications. For a company that is interested in learning about customer churn, this type of modeling would not be very helpful in allowing them to analyze churn.

IMPROVED CLASSIFICATION

To generate a model that could more accurately identify churn, we tried other modeling techniques including:

- Adjusting the Classification Probability
- Imbalanced Modeling

We compared the results from each of these techniques to the initial modeling results to see how they improved the classification rate over the base models.

ADJUSTING THE CLASSIFICATION PROBABILITY

Youden's J Statistic can be used to capture the performance of a dichotomous diagnostic test. It is calculated as $J = \text{sensitivity} + \text{specificity} - 1$. The optimal probability for classification is then selected by finding the probability that is associated with the maximum value of J. You can obtain the table that contains all the different probability levels by adding the option *ctable* to *proc logistic* in Base SAS.

```
model &target. = &variables. / ctable;
```

Once the probability values have been obtained, you can then calculate the J statistic at each level to determine the probability with the best balance of sensitivity to specificity. The optimal value corresponds to the row that has the largest J statistic. For this model, the optimal value was determined to be 0.3 per Youden's J Statistic. Intuitively, it is reasonable to learn that the optimal probability is closer to the actual ratio of churn in the data than 0.5. The optimal value can be found in Table 3.

Probability	Sensitivity	Specificity	J Statistic
0.26	73.8	57.1	129.9
0.28	68.9	62.9	130.8
0.3	64.8	67.4	131.2
0.32	59.9	71.9	130.8
0.34	55	76.1	130.1

Table 3. Youden's J Statistic and Optimal Probability

This optimal probability did improved the ability of the models to identify churn in the test data set. Table 4 contains the results for each of the three models.

Statistic	Logistic Regression	Decision Tree	Neural Network
Accuracy Rate	66.5%	65.5%	68.3%
Misclassification Rate	33.4%	34.4%	31.6%
False Positive Rate	56.6%	57.7%	31.7%
False Negative Rate	16.8%	17.4%	17.8%
Correct Churn Class	4698	4624	4323

Table 4. Accuracy Statistics: Optimal Probability of 0.3

Although the accuracy dropped slightly across the board we observed a significant increase in the accurate classification of churn within the data.

IMBALANCED MODELING TECHNIQUES

The decreased probability cutoff value did improve the different models' abilities to predict churn, but we also wanted to see how imbalanced data modeling techniques could also be used to improve the overall classification of churn in the Oberweis data. Specifically, we decided to employ undersampling the majority event – retention and oversampling with replacement of the minority event – churn. These models were built on the under and oversampled sets and then a holdout sample from the original data was scored to evaluate their performance on the true population proportion of retention to churn.

Undersampling

Undersampling is the process of eliminating observations of the majority class at random to produce a more balanced ratio of the target class. For this analysis, we undersampled the majority class – retained – so that the training sample contained approximately 50% of each level of the target class.

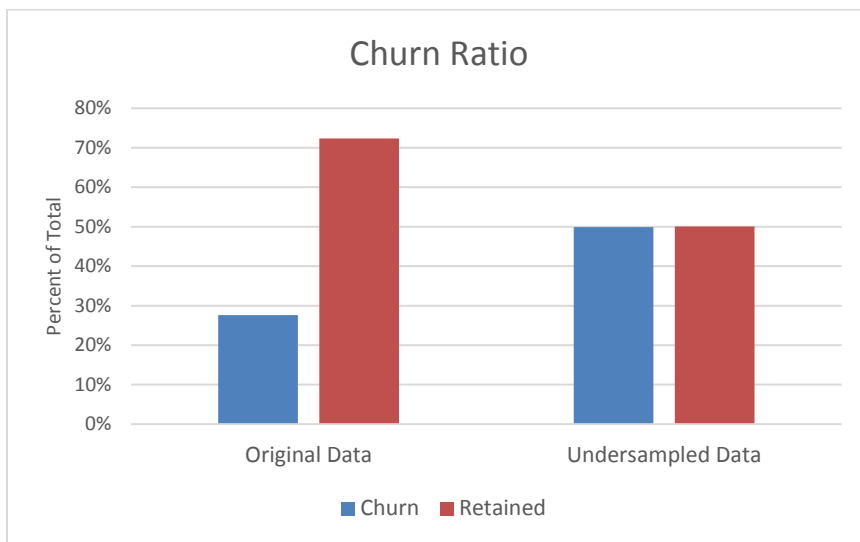


Figure 3. Distribution of Churn, Original vs. Undersampled Data

The results for the undersampled data can be found in Table 5.

Statistic	Logistic Regression	Decision Tree	Neural Network
Accuracy Rate	65.7%	61.3%	65.7%
Misclassification Rate	34.2%	38.7%	34.2%

False Positive Rate	57.1%	60.5%	57.5%
False Negative Rate	16.1%	15.9%	17.5%
Correct Churn Class	4892	5221	4593

Table 5. Accuracy Statistics: Undersampled Data

For the undersampled data the decision tree performed the best on the test data set. However, this was at the expense of accuracy which dropped almost 10% from the original baseline models.

Oversampling

Oversampling with replacement is another method to account for a binary target that is heavily skewed toward the majority class. It is accomplished by oversampling the rare event in your training data set so that the proportion of each level is more balanced. We sampled all the minority event observations twice so that our training data set was comprised of a more balanced ratio of retention to churn.

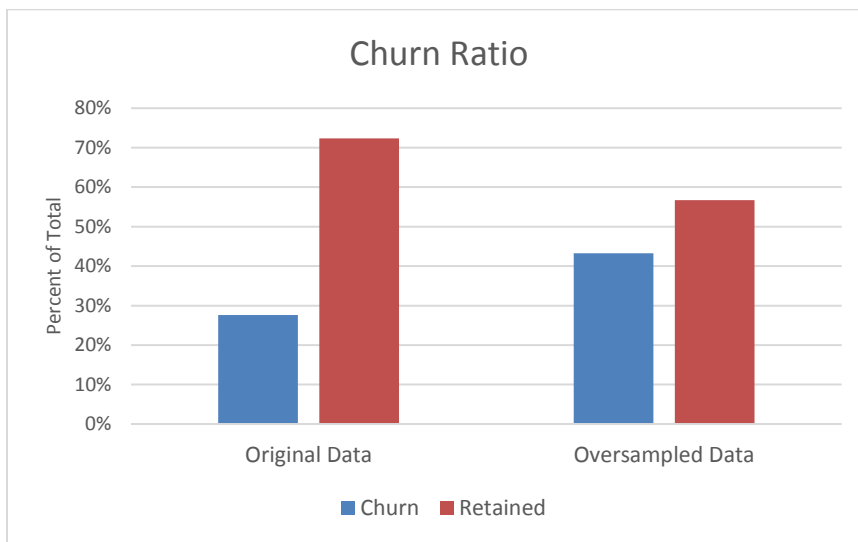


Figure 5. Distribution of Churn, Original vs. Oversampled Data

The results for the oversampled data can be found in Table 6.

Statistic	Logistic Regression	Decision Tree	Neural Network
Accuracy Rate	69.4%	59.8%	70.2%
Misclassification Rate	30.5%	39.8%	29.7%
False Positive Rate	53.9%	64.1%	52.8%
False Negative Rate	19.0%	22.1%	19.0%
Correct Churn Class	3929	3926	3864

Table 6. Accuracy Statistics: Oversampled Data

For the oversampled data, the neural network and logistic regression models performed almost equally as well on the test data set. Both models were also able to achieve this improved classification with higher accuracy rates than those from the undersampled models.

RESULTS

Figure 6 and Figure 7 provide an overview of the overall accuracy and churn classification rate for each of the models we built across all the different modeling techniques.

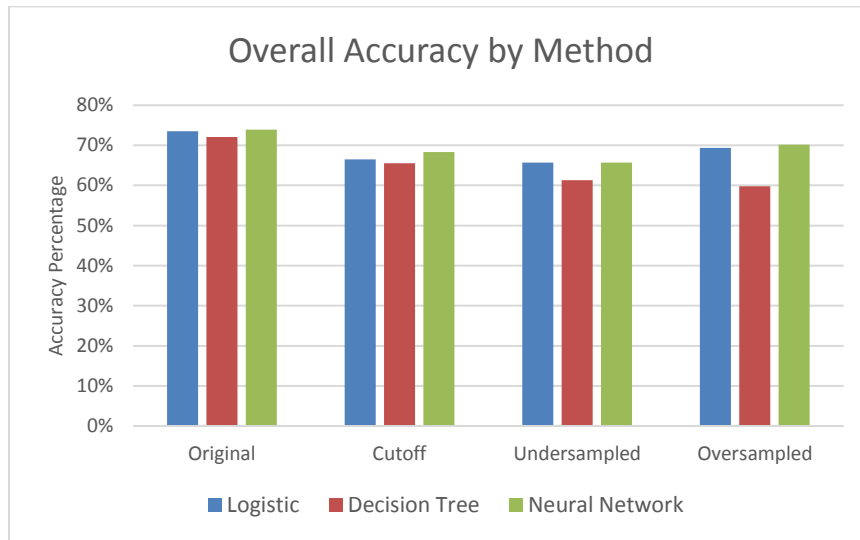


Figure 6. Accuracy of Each Model Across Methods

All of the techniques performed rather similarly in terms of overall model accuracy. However, the real differences between the techniques can be seen in Figure 7.

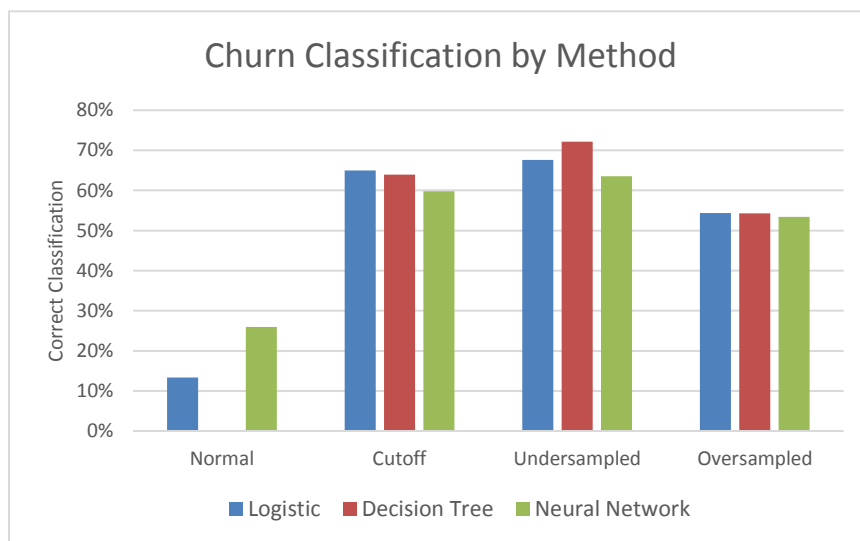


Figure 7. Percent of Churn Accurately Classified

The adjusted cutoff and imbalanced modeling techniques performed better at correctly identifying churn when compared to the original modeling results.

CONCLUSION

Identifying and understanding the factors that influence customer churn allows companies, like Oberweis, to gain a competitive advantage. This analysis attempted to provide insights into the types of variables that are most important in predicting customer churn. We then used various modeling methods to try and find the model that was most accurate in predicting customer churn. However, what we found is that it ultimately depends. It depends on whether the goal is purely accuracy, or a correct classification. The most accurate models lacked the ability to identify customer churn within the data. By applying various

modeling techniques, we drastically improved the ability of our models to identify customer churn and the overall accuracy only decreased slightly. In the end, we have seen that customer churn is not an easy problem to model but by using the advanced analytical tools available to us in SAS Viya, VDMML, and a little bit of creativity, it is possible to build a model that can identify and explain customer churn within your data.

REFERENCES

- SAS Online Documentation, SAS Support; “The LOGSELECT Procedure” Accessed August 2017. Available at:
http://documentation.sas.com/?docsetId=casstat&docsetTarget=viyastat_logselect_overview02.htm&docsetVersion=8.1&locale=en
- SAS Online Documentation, SAS Support; “The TREESPLIT Procedure” Accessed August 2017. Available at:
http://documentation.sas.com/?docId=vdmmlcdc&cdcVersion=8.1&docsetId=casstat&docsetTarget=viyastat_treesplit_syntax01.htm&locale=en
- SAS Online Documentation, SAS Support; “Getting Started with SAS Enterprise Miner 14.2” Accessed August 2017. Available at:
<https://support.sas.com/documentation/cdl/en/emgsj/70152/PDF/default/emgsj.pdf>
- Damadoran, R., Kumar, G, Mohan, K.R., Mohan, J. 2016. “Predicting Rare Events Using Specialized Sampling Techniques in SAS.” *Proceedings of the SAS Global Forum*. Las Vegas, NV. Available at:
<http://support.sas.com/resources/papers/proceedings16/11140-2016.pdf>
- Akosa, Josephine. 2017. “Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data”. *Proceedings of the SAS Global Forum*. Orlando, FL. Available at:
<http://support.sas.com/resources/papers/proceedings17/0942-2017.pdf>

ACKNOWLEDGMENTS

Zencos would like to thank Oberweis Dairy for the continued partnership and opportunities for collaboration.

Sean would like to thank his colleagues at Zencos Consulting for providing guidance and thorough review on this paper.

RECOMMENDED READING

- *Getting Started with SAS Visual Data Mining and Machine Learning*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sean Ankenbruck
Zencos Consulting
sankenbruck@zencos.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.