Paper 2590-2018

Analytics Applications in Targeted Marketing and Forecasting Demand - A Two Stage Model

Anurag Hardikar, Oklahoma State University

ABSTRACT

Identifying target customers for a product/service is one of the most important steps in developing a marketing plan for any business. A large electric utility company in Oklahoma with a customer base of more than 800,000 have recently launched its solar power program. The objective of this paper is twofold. In the first stage, build a classification model to identify; customers that have a higher propensity to enroll in the company's solar power program so as to drive savings on promotional mailing. Then, in the second stage, we see how we can forecast electricity consumption to predict the solar power capacity the company would need to fulfill the associated demand.

INTRODUCTION

This paper discusses the application of alternative approaches to take for improving the accuracy of a classification model. Often, there is the problem of rare events while dealing with a new product or service and choosing the right sampling method becomes an important factor in improving the classification model. Datasets used for such analyses typically have less than 1:99 ratio of events to non-events, even a plain assignment of non-event to all the cases can give us a 99% accurate model in terms of misclassification. Therefore selection of a performance measure that takes into consideration how well the model is separating the two classes is essential.

The dataset being studies has around 600,000 observations for residential customers and more than 60 variables that provide information about customer demographics, customer interactions with the company, utility usage data. There are different techniques (such as sampling the data to deal with issues brought up by rare cases, using feature engineering to derive significant features from the raw data, transforming variables with high skewness and kurtosis) to get the data distributions as close to normal as possible. After all of that, finally choosing of right performance measure.

After identifying the customers that have high propensity to enroll in the company's solar power program, it is possible to analyze customer electricity consumption with a time series model, decompose it into trend and seasonality, and build an autoregressive integrated moving average (ARIMA) forecasting model to forecast customer energy consumption for the next 18 months. Having a good estimate of future demand can help the company manage their resources.

STAGE ONE- IDENTIFYING CUSTOMERS

METHODOLOGY

Target here is a binary indicator indicating whether a customer is enrolled in the company's solar power program. Data is sampled, transformed, and partitioned in 70% training and 30% validation datasets. Random forest and logistic regression on various combinations of different sampling techniques and variable transformation techniques are built. Finally a model comparison node compares the performance of the models based on area under the ROC curve.

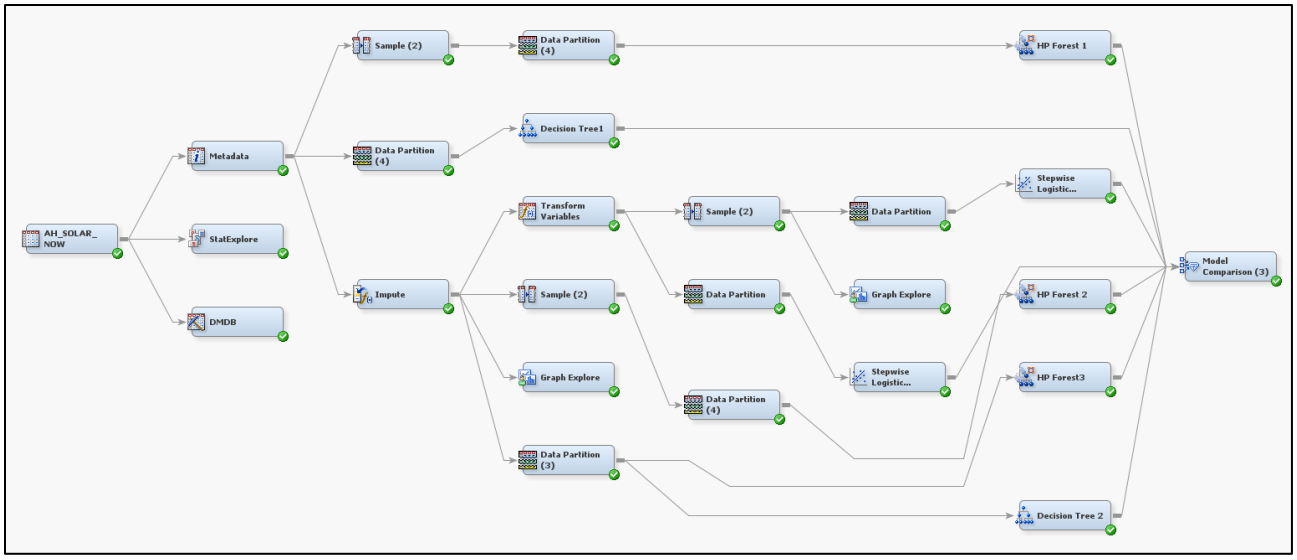

Figure 1. Enterprise Miner diagram for classification

SAMPLING

The dataset built for this research has less than 1% events. In such cases training gets biased towards non-event cases. We deal with this issue by incorporating various sampling techniques. We consider three different techniques:

- **Minority oversampling:** This technique increases the number of rare class observations by repeatedly sampling its observations. We chose 1:1 ratio for oversampling.
- **Majority undersampling:** This technique keeps the number of rare class observations the same while reducing the number of majority class observations by random sampling. We chose 1:1 ratio for underdamping.
- **Original data** :No sampling is performed

FEATURE ENGINEERING

Feature engineering is the process of deriving features from raw data in an attempt to improve the predictive power of the model. Feature engineering is one of the most important steps in building a model, and can be the difference between a good model, and a bad one.

- **Electricity Usage:** Electricity usage data is available for each customer for every 5 minute time interval, and is not usable as it is. Using this data it is possible to calculate a customer's daily electricity usage, average daily usage, average monthly usage, and variance in daily usage.
- **Billing:** Billing data is available for each customer for all previous months. This can be used to calculate a customer's average monthly bill, the maximum bill for the last year, minimum bill for the last year, and variance in the bills.
- **Demographic:** This data helps to consider a customer's age, gender, marital status, and duration the customer has been with the company.
- **Enrollment:** Data on customer enrollments in companies varies throughout the different customer programs. The motivation here is to find if there is any association between a particular enrollment and customer's inclination towards the new solar power program.

DATA TRANSFORMATION

In order to make the standard error unbiased, it is necessary to have the input integer variables as normally distributed as possible. For this project log transformation, square root transformation, and power series

transformation were conducted on the inputs that have either high kurtosis or high skewness. The following figures show how the distribution of average monthly kWh moved towards a normal distribution after performing square root transformation.

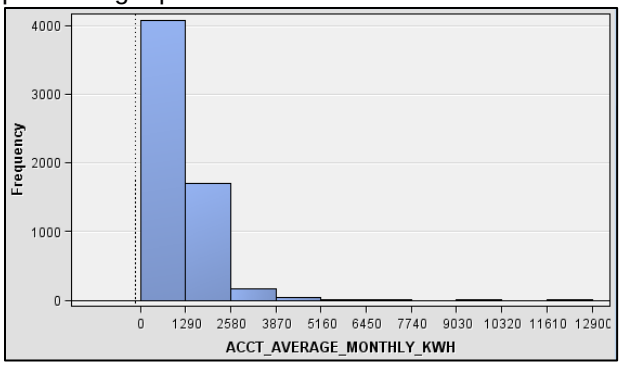

Figure 2. Before transformation

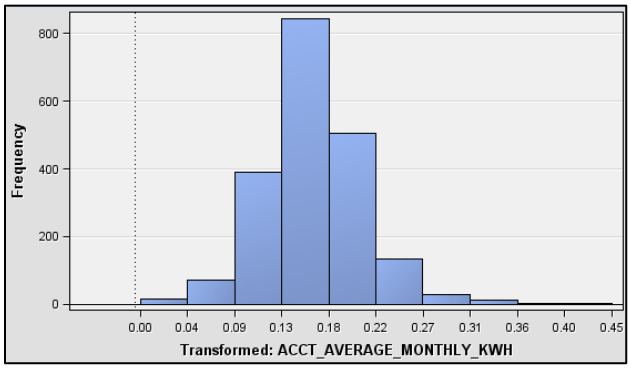

Figure 3. After transformation

PERFORMANCE MEASURE

Misclassification rate is not the best performance measure for classification problems, especially with rare events. For instance, in the given case, even if we blindly mark all the cases as non-events, we will be right 99% percent of the time. So misclassification rate doesn't give the whole picture.

Some better measures of performance of a classification are:

- **F1 score:** The F1 score is the harmonic mean of precision and recall. Precision is a measure of the accuracy of positive predictions and recall is the indicator of true positive rate of the classifier, and it is the ratio of positive instances that are correctly detected by the classifier. The classifier will only get a high F1 score if both recall and precision are high.
- **AUC ROC:** The ROC curve is another common performance measure used for binary classification. Instead of plotting precision versus recall in case of F1 score, ROC plots true positive rate against false positive rate. And the Area under the curve AUC ROC is a measure of how well a binary classification model is performing and it is good practice to use this measure for comparing models.

We see from the comparison of models that the random forest algorithm built on oversampled data performs the best with the highest AUC ROC.

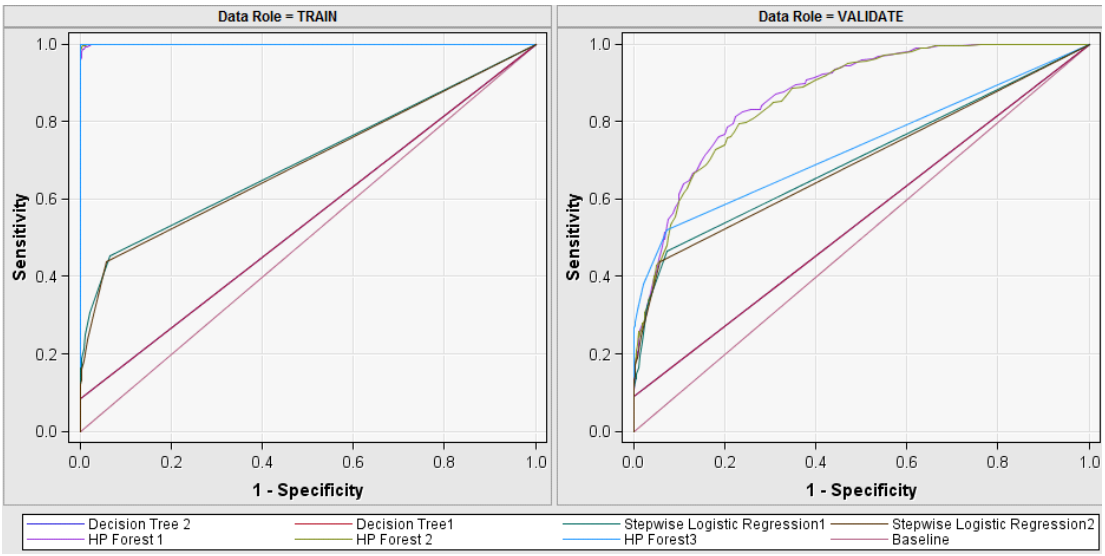

Figure 4. Comparison of ROC of all the classification models

SIGNIFICANT VARIABLES

The research gives us various important variables that proved to be significant in the prediction of whether a customer will enroll in solar power or not.

- **Account age:** analysis shows that new customers, who have not been with the company have higher propensity to enroll into solar power program.
- **Average monthly electricity usage:** Solar power subscribers have more average monthly usage than non-subscribers.
- **Variance in the electricity usage:** Solar power subscribers tend to have low variance in the electricity usage than non-solar power subscribers
- **Marital status:** Solar power subscribers have more percentage of singles than married customers.

FORECASTING DEMAND

TIME SERIES ANALYSIS

In the second stage of the model, it is time to consider customers who have been identified by the model as those with a high propensity to enroll in the solar power program and look at their daily electricity usage. We aggregate their usage by month and analyze the time series. Figure 5 shows the time series plot of their usage. With visual inspection, it is possible to get a sense for the strong seasonal component with little or no trend.

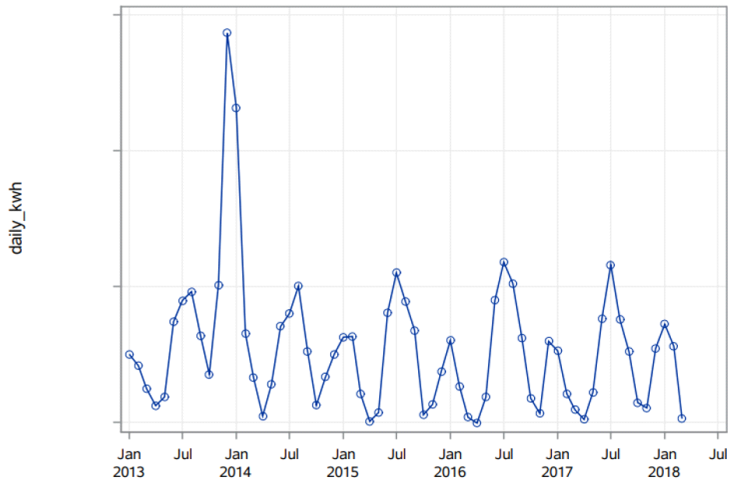

Figure 5. Electricity usage time series plot (numbers are hidden)

DECOMPOSITION ANALYSIS

We further decompose the time series to look for a hidden seasonal component and a trend component. Figure 6 shows a strong seasonal component of the time series. Typically the usage increases during winter and summer and this is captured in figure 6.

The trend component of the time series is shown in figure 7. We don't have a significantly strong trend in the time series

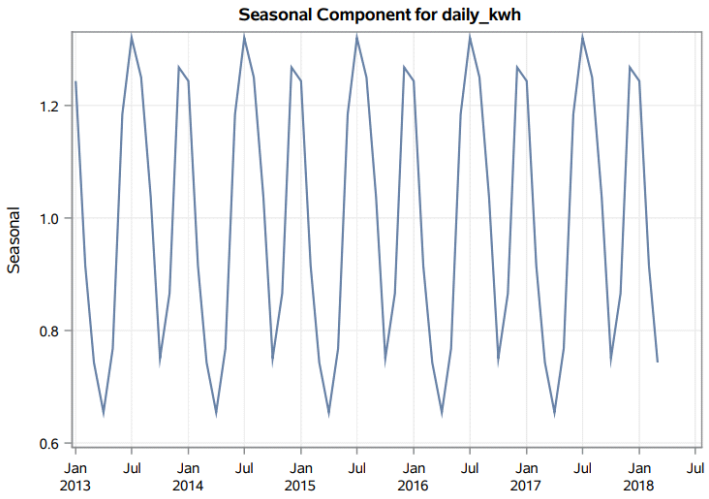

Figure 6. Seasonal component

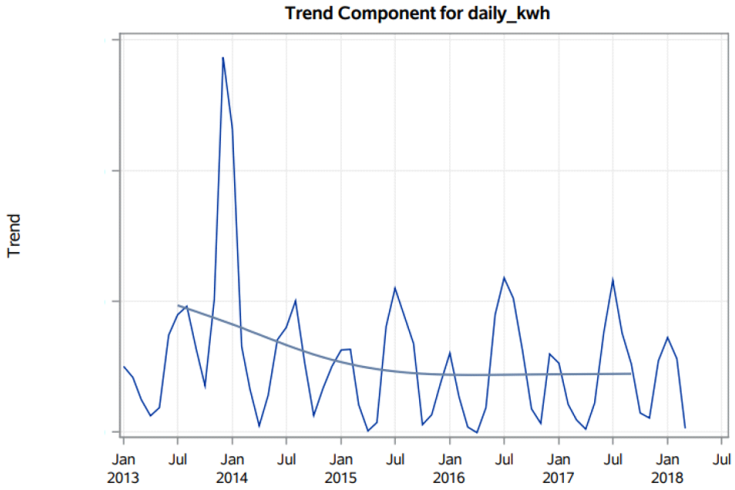

Figure 7. Trend component (numbers are hidden)

ARIMA FORECASTING

After analyzing the time series, it is time to build a forecasting model. The assumption of stationarity is not satisfied by the time series. Therefore, it is necessary to consider 1st order differencing with 12 periods of differencing. This yields a stationary time series. Next perform an autocorrelation analysis to see if there is any significant autocorrelation in the data. There is significant autocorrelation for up to 2 lags.

Using a grid search algorithm to get the combination of (p, d, q) that minimizes the standard error contributes the following parameters: Autoregressive order = 1, Differencing order = 1, Moving average order = 0

The following split is used for building this model:

- **Train:** 30 months
- **Holdback:** 18 months
- **Forecast** 18 months

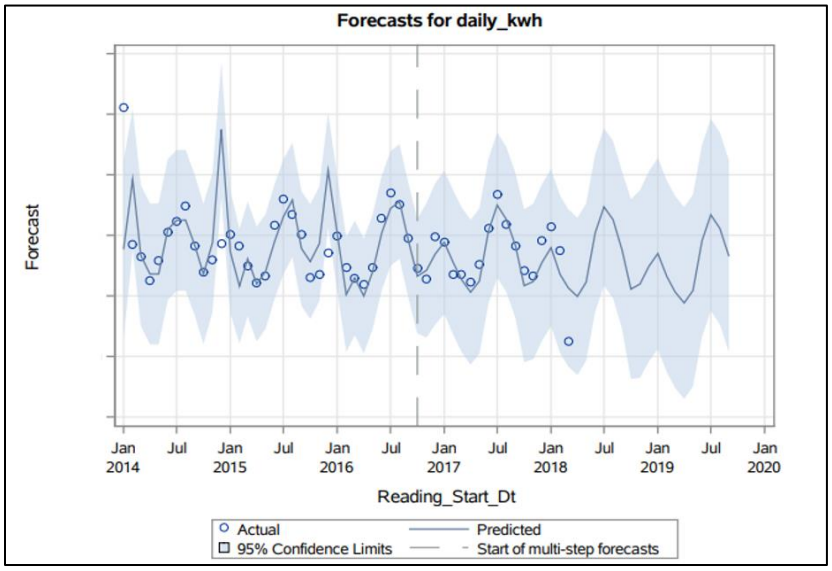

Figure 8.ARIMA forecasting graph (numbers are hidden)

CONCLUSION

For any industry trying to identify customers for a new product/ service, targeted marketing can save costs by focusing promotional efforts on a specific category of customers. A very common issue of rare events encountered while trying to identify target customers can be solved using various sampling techniques dissed in the paper. We also saw that sampling along with feature engineering and data transformation, choice of appropriate performance measure can lead to improved classification model. Paper analyzed the demand time series for the target customers and forecast demand fairly accurately using ARIMA forecasting model.

REFERENCES

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical LearningJournal

Alice Zheng and Amanda Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists

Ruey S. Tsay, Multivariate Time Series Analysis: With R and Financial Applications

SAS Notes Available at http://support.sas.com/kb/22/601.html

ACKNOWLEDGMENTS

This paper utilized data from a major utility company in Oklahoma. We are thankful to them for providing us the data for analysis purpose.

CONTACT INFORMATION

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

Anurag Hardikar
Master of Science in Business Analytics
+1 405-780-3844
anurag.hardikar@okstate.edu
https://www.linkedin/com/in/anuraghardikar