Identifying customers at risk of churn

Austin Hooper

Inge Koch, Berwin Turlach
Mathematics and Statistics
The University of Western Australia

Yuichi Yano
CEED Client: Synergy

Abstract

Synergy's contestable customers have the potential to close business with Synergy and enter a new contract with a competitor. This occurrence is referred to as churn. In order to minimise the risk of churn, Synergy's Retail Business Unit (RBU) currently relies on direct communication with existing customers to provide products and services that meet a customer's needs. As a result, there is not an existing model or quantitatively backed framework for assessing a customers' risk of churn.

This churn modelling project aims to bring value to Synergy through both the creation of a churn model that can be implemented to improve retention targeting and by producing insights that can be used to create a framework for assessing risk of churn. The project uses numerous distinct sources of data to create a churn modelling data set, then uses this data set to develop interpretable churn models that provide insight into factors contributing to customers' risk of churn.

1. Introduction

Synergy is Western Australia's largest electricity generator and retailer of gas and electricity with more than one million residential, business and industry customers. Synergy provides 52% of the electricity sold to households and business customers in the South West Interconnected System (SWIS) and provides 55% of the contestable gas load in the industrial and commercial market (Synergy, 2022).

Since the end of 2006 customers using more than 50MWh of electricity per year have been considered contestable in the SWIS retail energy market (AER, 2007). A contestable customer has the potential to close their business with Synergy and enter a new contract with a competitor. This occurrence is referred to as churn. Prevention of customer churn is of significant value to Synergy's RBU, as customers with the potential to churn represent a large portion of Synergy's retail revenue. To minimise the risk of churn, the RBU currently relies on proactive and frequent communication with existing customers to continuously provide products and services that meet customer needs.

Churn models have become a key technique used to reduce customer churn, with successful statistical models being used in the banking, telecommunication and European energy industry (Keramati et al., 2016; Balasubramanian & Selvarani, 2014; Moeyersoms & Martens 2015).

A work of significant relevance to our project is that of Marttinen (2021) who studied residential customer churn for a Finnish electricity company following a significant churn event. The study sought to apply decision tree and logistic regression modelling approaches found in the churn modelling literature in other industries (telecommunication, banking, etc.) and to verify the effectiveness of these techniques when applied to a relatively small data set in the energy industry. The paper used four variables to predict customer churn for 18,020 customers during the first half of 2020. Using the four variables in/out of network, product category, length of customer relationship and electricity consumption a decision tree model was found to be the best predictor of churn for their data set, with an 87% accuracy and a f-score of 74% while the logistic model scored of 79% and 45% respectively. The paper demonstrates that both logistic regression and decision tree models can be successfully applied to churn modelling in the energy industry, even with limited samples and a small variable pool. The success of this project is especially relevant due to the similarities in available data between the two projects, with Synergy recording data allowing for the construction of three out of the four variables that were used in Marttinen's work; product category, length of customer relationship and electricity consumption.

1.1 Project Aim

The project aims to utilise Synergy's recorded churn data to develop interpretable churn models that can be deployed across Synergy's commercial customers and that provide insight into factors contributing to customer risk of churn.

2. Process

2.1 Data Collection

On commencement of the project, there was not a pre-prepared dataset on which to conduct the churn analysis, the first action was to identify and extract information relevant to churn modeling from Synergy's databases. Based on consultation with Synergy subject matter experts, Synergy's CRM database was identified as a potential source of data. The database holds historical records of customer information on various administrative, customer contestability, and electricity consumption information. The following CRM database columns were used to identify relevant records:

Contestability – all records with a contestability value of 'non-contestable' were excluded. The project is only concerned with possible churns, and non-contestable customers cannot churn.

Annual Contributing Days – Only records with 365 annual contributing days were retained. This was done to ensure that all consumption and revenue data is comparable across records.

Annual Electricity Consumption kWh – If a record was classed as 'contestable', or if the record had 50,000 kwh or greater annual electricity consumption, the record was retained.

Churn Out Date and Move Out Date – All records without a churn out or move out date were retained**.** Customers with a move out date or a churn out date prior period of study were dropped from the data set. This is done as records for consumption and revenue are updated periodically, and old churn and moveout records would have data from a different time period to the active customers that are being compared against.

Main Product – Customers with certain main energy products are unable to churn due to their product. Records with any of these main products were dropped from the data set.

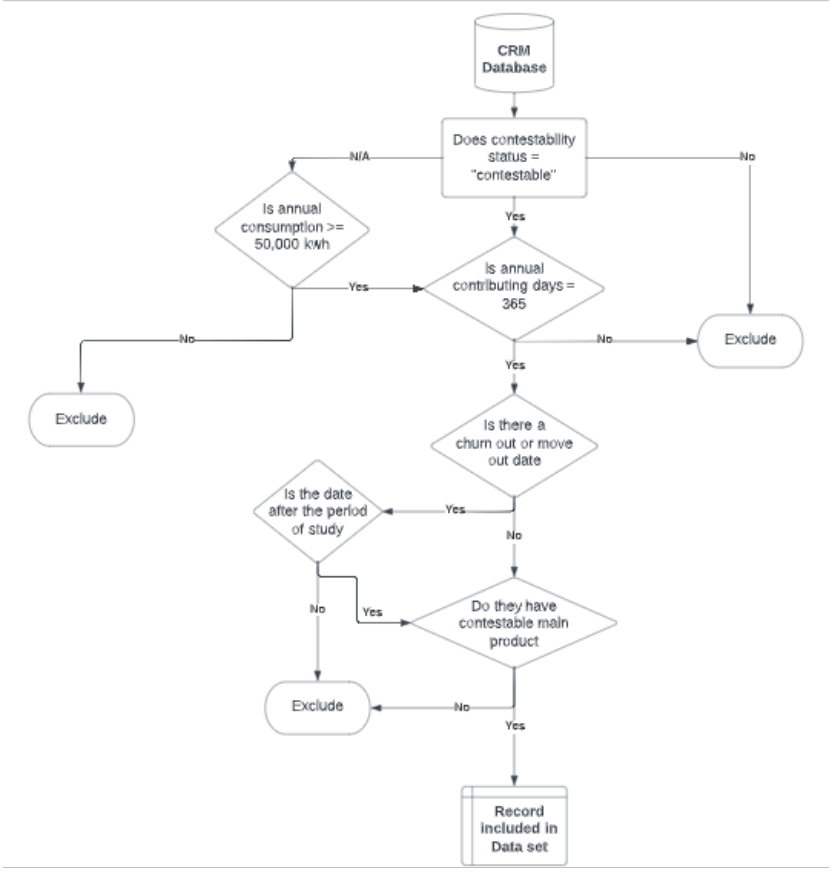

Figure 1 Data collection process

Following the application of these rules to the database, A set of records were identified with which to create the dataset. The features summarised in table 1 were then extracted from the CRM database for each identified record.

Feature	Description
Annual Electricity Consumption kWh	The amount of electricity consumed in the last 365 days measured in kilowatt hours.
Annual Electicity Revenue	Synergy's revenue from the customer's electricity bills in the last 365 days recorded in AUD.
Main Product	The main Synergy product that the service account is using. This has been semantically grouped into several larger categories.

Table 1 CRM features

2.2 Data Processing

Further features were then identified as potentially significant to the project and constructed from data sourced from both Synergy's CRM database and a variety of other Synergy databases. These features are summarised in Table 2.

Constructed feature	Description
Class	A feature that categories records into 3 categories, active, churn, and move_out_only.
Is_churn	A re-coded binary representation of the Class feature's category churn.
Master industry	The industry that the customers parent account belongs in. This has been semantically grouped into several larger categories.
Number of contract refresh	The number of times a contract has been refreshed by the customer of a given record.
Avg contract refresh	The average length of a record's contracts
Length of customer relationship	The length in days that the customer has been a customer of synergy.
System size	The size of the solar panel system a customer has. If the customer has no solar panel system, this feature is 0.

Table 2 Constructed features

Next, in discussion with data experts, observations with relevant missing variables were identified and removed from the dataset. Following this, for the purpose of modelling, both "Annual Electricity Consumption Kwh" and "Annual Electricity Revenue" were natural log transformed to reduce variables with large values, outliers and overall skew. The final dataset following feature engineering and data cleaning contains 8 explanatory variables and one response variable, "is_churn".

Less than 5% of customers during the period of study churned, as such, we have a highly imbalanced data set. One key problem that can occur due to having a small number of churns is that of complete separation, which is defined in the literature as a problem for "regression models with a discrete outcome (such as logistic regression) where the covariates perfectly predict the outcome" (Mansournia, 2018). Complete seperation results in infinite coefficient estimates, and thus will cause our model to be a poor predictor of churn. To combat this, the 2 categorical variables, "Main Product", and "Master industry" have had any categories with less than 9 occurances of churn reassigned following consultation with Synergy experts and similar database columns.

2.3 Data Analysis and Modelling

Logistic regression has been chosen to model churn. Logistic regression uses a logistic function to model the probability that a variable belongs to a specific category. By then applying a decision rule to this probability, a binary prediction can be made predicting whether a customer will or will not churn. The choice to use logistic regression has been made due to the method's relative success in the literature in predicting churn, combined with the methods interpretability, noting Synergy's desire for models that provide insight into factors contributing to customers' risk of churn. Because of the high imbalance between churns and non-churns we will sample

from the non-churn to create a balanced set. This approach has been implemented to improve the training of our logistic models.

To prevent data leakage, where data not available during prediction is used to train the model and thus contaminates it, 10-fold cross validation was selected. As part of this process, the data is split into 10 folds, with each model run once for each fold (10 iterations). For each iteration models are trained on 90% of the data (9 non-current folds) and tested on the remaining 10% (current fold). The mean accuracy, precision and recall are then reported for each prediction on the 10 test folds. This allows verification of the predictive quality of trained models.

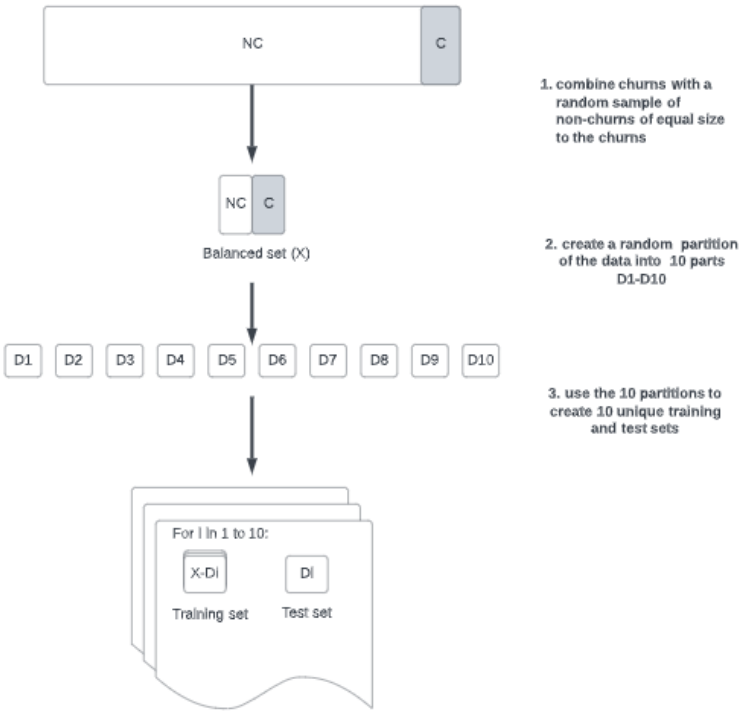

Figure 2 Cross validation process

Throughout the project two areas of research Synergy expressed specific interest in was obtaining insights into the impact that a customer's product or industry had on their propensity to churn. To meet this need, separate models will be fit solely on the categories of "Main product" and "Master Industry" to investigate whether used individually, a specific main product or master industry can be indictive of churn. Additionally, a model combining all the variables in the data set (with the exception of class) will be evaluated. Best subset variable selection (also known as all subset selection) will be run on all three models, and the scores of each subset from 1 variable to all model variables will be evaluated. This will be done to identify the most important variables for churn, to evaluate the predictive power of individual variables, and to identify a subset of variables that prevents overfitting, as we are training on a data set of limited size.

A summary of the modelling process is as follows

1. Create a balanced set by combining a random sample of N observations of the non-churn with the N churns.
2. Using this balanced sample, conduct randomized 10-fold cross validation
3. For each k fold from 1-10 conduct the following:
   3.1. Train on complement of the k_{th} fold (90% of the data)
      3.1.1. all subset variable selection on main product

3.1.2. all subset variable selection on master industry
3.1.3. all subset variable selection on the master industry and main product top 4 each combined with all numeric variables
3.1.3.1. where top 4 is selected to meet the subset package's max of 15 variables
3.2. Test each subset for each model on the k_{th} fold (10% of the data)
3.2.1. For each variable subset (from 1 to the max number of variables) record the variables selected as well as the test set accuracy
4. For all models report the mean accuracy of the 10 folds and number of variables selected

3. Conclusion and Future Work

The next step in the project is to complete the modelling process and to report the findings. As the project has progressed, there has been many avenues identified for further work at Synergy regarding churn prediction. The most notable of which are related to the limited historical data resources for the variables used as part of the churn modelling process, as the standard updating procedure of the operational database our data was sourced from is to overwrite pre-existing data. This limits how far back in time data could be sourced. One proposed piece of work is to create and implement a process for accessing stored historical data relevant to churn, so that in future a time series analysis could be conducted to get further insights from the data. A time series analysis could answer pertinent questions such as whether change over time in a customer's energy consumption or the annual revenue Synergy recieves from a customer is indictive of churn. A further benefit of using historical data would be an increase in the number of churns present in the data set, allowing the use of more granular cateogrisations of "Main Product" and "Master Industry".

4. Acknowledgements

I would like to thank Yuichi Yano, Inge Koch, and Berwin Turlach for their consistent and generous mentorship and supervision throughout the project. I would also like to thank Jeremy Leggoe and Kimberlie Hancock for their assistance throughout the year, as well as for the organisation of this phenomenal opportunity. Finally, I would like to thank Andras Varga, Michael Osan, and Joshua Allen at Synergy for assisting in the creation of the churn data set.

5. References

AER. (2022). State of the Energy Market 2007 (pp. 204-213). *Melbourne: Australian Energy Regulator*.

Balasubramanian, M., & Selvarani, M. (2014). Churn prediction in mobile telecom system using data mining techniques. *International Journal of scientific and research publications*, 4(4), 1-5.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(1), 1-13.

Marttinen, J. (2021). *Modelling customer churn with private electricity customer data*. [Master Thesis, Lappeenranta-Lahti University of Technology], LUT Pub.https://lutpub.lut.fi/handle/10024/163003

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, 72, 72-81.

Synergy. Who we are. Synergy. (2022). Retrieved 1 June 2022, from https://www.synergy.net.au/About-us/Who-we-are.