**Learning Journal for Customer Analytics Project**

Norvin Chandra A0195335Y

**Scope of Work**

Data exploration, data visualisation, business understanding and building of churn prediction model.

**Data Understanding, Exploration and Preparation**

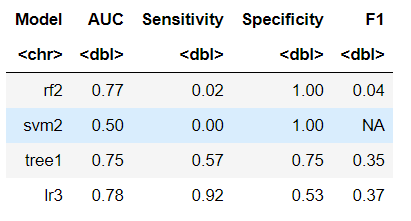
1. Dataset does not come with metadata or explanation on what the components in each variable means, for example what is the difference between Personal L1 and Personal L3? What is the difference among the 4 renewal offers? We do not have answer for this thus we cannot analyse and explain in detail. Although I come from insurance industry, each company has different product thus detailed product information is critical to a successful and meaningful analytics project
2. Since there is Customer Lifetime Value and Total Claim Amount but no clear information on how the CLTV is derived, we believe that value of a policyholder should be equal to CLTV – Total Claim Amount since claim amount is a cost to insurance company.
3. During exploration we realized that the data only contains 2 months renewal data. To create an effective churn prediction model, we definitely need more than 2 months data.
4. The data is imbalanced with “Yes” account for only about 16% of all the Responses. The insurance company has a very poor renewal ratio and losing too many renewal business
5. With the problem identified, we decided that we need to focus our approach on improving policyholder retention by:
   1. Deepening policyholder relation with Safe Insurance, discouraging policyholders to switch insurance company
   2. building a churn prediction model to allow business development team to identify and intervene churning policyholders before they actually churn

**Building Churn Prediction model**

1. Initial challenge was to establish understanding on the result of data exploration. We do know that Offer 3 and 4 are not popular but we cannot fully understand and correlate the result with the profile of the policy holder as we do not know what each offer type entails.
2. As we only have 2 months policy data, I decided to use Jan 2011 data as Training (80%) and Validation (20%). Testing is done using Out of Time test set, the Feb 2011 data.
3. We believe true value of Policyholder should equate to the difference between the CLTV and Total Claim Amount as claims is a cost to insurance company and created a new variable by the name of Net.CLTV
4. To classify policyholders into churner and non-churner, I use Logistic Regression, SVM, Decision Tree and Random Forest.
5. To evaluate the different models, I use AUC and Specificity (True Negative Rate) when model is run on OOT Test data since we are interested in the model’s ability to accurately detect negative response (churner)
6. Logistic Regression
   1. Converted the categorical variables to dummy variables using one hot encoding and removed irrelevant Customer variable as well as Customer.Lifetime.Value and Total.Claim.Amount which are already synthesized into Net.CLTV
   2. Balance training data set using Random Oversampling to increase the “Yes” response
7. Support Vector Machine
   1. Data for this model will be based on scaled, balanced and one hot encoded training dataset
8. Decision Tree
   1. Use balanced training data without one hot encoding as dummy variables has adverse impact to tree-based models
9. Random Forest
   1. Use balanced training data without one hot encoding as dummy variables has adverse impact to tree-based models

**Model Selection**

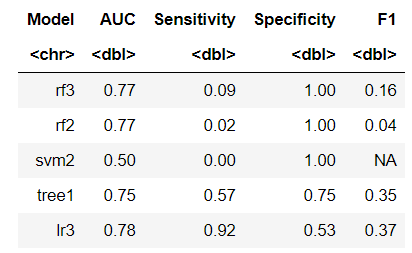
1. Compared the 4 models ran with standard parameter in terms of its AUC and Specificity.



1. Rf2 model has highest Specificity

**Model Tuning**

1. We run grid search on random forest to identify the optimal “mtry” (number of variables available for splitting at each tree node) parameter.
2. The best mtry value is 6 and with that we train rf3 model and obtained the following AUC Specificity parameters



**Implementation**

1. We use the random forest model to generate the churn probability and assign this as score to each policyholder.
2. Just knowing who will churn does not allow business team to effectively target the churners given limited time and resources.
3. We combine this churn probability with the Net.CLTV and sort data descending based on the score and the Net.CLTV to identify churner with high value to Safe Insurance. This allows business team to focus their limited resources on valuable churners to try retain them through renewal campaign.
4. The metrics to measure the implementation will be Renewal Ratio of expiring policyholders. Current baseline for renewal ratio is 16% and we are certain that improvement can be easily achieved if targeted effort is implemented.

**Conclusion and Individual Lessons**

* Customer Analytics require not only statistical and coding skills but also good understanding of the business, product, operation and customer
* Good data set is crucial for a successful customer analytics project.
* Our dataset does not allow us to do campaign analytics as such data is not available
* In this semester I refreshed what I learned in 2nd semester on Predictive Analytics module in terms of model building and combine this with the Customer Analytics concept that I learn in this semester
* There are other modelling techniques that are relevant to solve the churn prediction problem that I have yet to try, for example Gradient Boosted Tree or Neural Networks
* I am grateful to my cooperative team mates and all lecturers for their effort in imparting their knowledge to me in this semester despite the challenges in this extraordinary time.