**PA Final Assessment Project**

**Executive Summary**

The ability of ABC Insurance Ltd to accurately forecast future cash flows is dependent on assumptions on the likelihood of a policyholder voluntarily forfeiting their policy (without additional premium due or future insurance liability). The predictive modeling team is selected to create a model to predict the probability a policy holder will terminate.

Data is given to the team in a cvs file with around 5,000 records, together with a data dictionary file. The team has decided to build a logistic GLM model and tested it again a regression Random Forest (RF) model. In a comparison, the GLM model has similar prediction power to that of the RF model.

The team think a logistic GLM is an appropriate model since we think that there are linear relationships among variables and that it is easy to interpret results. However, the quality of prediction is not very strong. As a result, we are recommending the following considerations:

* 5,000 observations may be not enough. We should collect more data.
* We should consider collect more variables such as income, occupation, …
* We need to address data issues as mentioned in the Data section.
* We may need to explore other models such as boosting methods.

**Problem**

The ability to correctly predict if policyholder will forfeit each year is important to project revenue as well as liability. Building such a model requires fair amount of good data. The model chosen needs to have both prediction power and is interpretable.

Data given to the team is small (5,000 rows) and many variables have missing values. After cleaning data, there are less than nine meaningful variables to use as predictors.

The lack of quality data will significantly impact outcomes. If the data does not present the true population, predictions will not be valuable.

**Data Preparation**

Data file of4,916 policyholders has 14 columns including an indicator if a policy was forfeited. We have made the following changes:

1. Drop variable sg\_prem\_type and sg\_status due to missing values

Text

Description automatically generated

1. Remove 8 records that has 0 value for face\_amount\_at\_issue. Also we make log transformation of this variable. The resulting distribution looks better.
2. Drop the substandard\_indicator variable as it has two values and one of which accounts 4%.
3. Drop the weight variable since it has one value for all records
4. Group the product variable to few values: Term, WL, and OTHER.
5. Smoker has three values. One of them is labeled as US (Unismoker) which we don’t know how to interpret.

The resulting data, after cleaning, has 4,907 records and 10 variables.

**Data Exploration**

We first run a mutual information analysis and ranked values against the target variable.

Table

Description automatically generated with medium confidence

The result shows sex, uw\_requirements, and study\_year may have weak relationships with target. Since the number of remaining variables is small, we decided to keep all of them.

We then randomly partitioned data into 70% for training and 30% for testing. The distribution of the target variable looks similar in training and testing data sets.

**Modeling Approach**

To predict a probability of an event, a GLM model with binomial (or Bernoulli) distribution and logit link function is a common approach. Predicted values are numbers between 0 and 1.

We built a GLM model including all variables and tested results against training and testing data sets.

glm(formula = f, family = binomial(link = "logit"), data = data.training)

* For training data, the accuracy is 0.58 with an AUC curve value of 0.58
* For testing data, the accuracy is 0.59 with an AUC curve value of 0.58

Testing results show consistency. However, 58% accuracy indicate the model is not a strong enough.

We then tested the interaction between duration and attained\_age, target ~ duration + attained\_age + sex + smoker + uw\_key + uw\_requirements + product + log\_face\_amount + duration \* attained\_age

The accuracy is slightly improved but not significant. We also tried to drop a few variables but there are little improvements in accuracy. As a result, we kept the original model.

Next, we built a regression RF model.

Text

Description automatically generated

* For training data, the accuracy is 0.81 with an AUC curve value of 0.81
* For testing data, the accuracy is 0.60 with an AUC curve value of 0.60

The results from this model clearly shows the model overfitted. We tried a few different changes in parameters. Overfitting persisted and accuracy did not improve. We decided not to use this model.

**Model Results**

The final model selected is listed below:

Table

Description automatically generated

Coefficients – These are the key parameter produced by the model. All variables contributes to the model.

P-Value – ProductWL and productTerm has lowest values which indicate they are significant. Since we grouped this value in the data preparation step. More research is needed.

Testing results:

Text

Description automatically generated

**Conclusions/Next Step**

Due to low prediction power value, we recommend further research as outlined below:

* Collect more data and more variables
* Clarify on smoker value descriptions
* Address missing values and propose solutions
* Explore feature generation such as interaction between predictors or variable transformation (log, reciprocal, ….)
* Consider other model models – different GLM models, different parameters for decision trees, bagging and boosting models

**Appendices**

Data dictionary

Original data

Data after cleaning and transformation (data.csv)

Code file