**PROJECT DELIVERABLE #2**

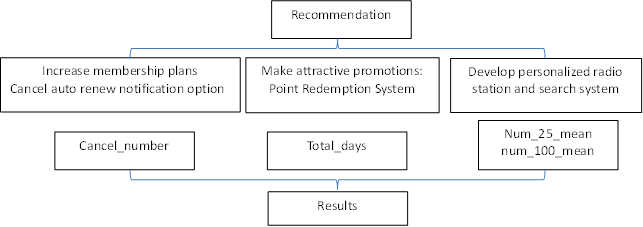
Siyang Guo & Zhenyu Qiu

* **BRIEF INTRODUCTION**

KKBOX (Taiwan) is Asia’s leading music streaming service, holding the world’s most comprehensive Asia-Pop music library with over 30 million tracks. They offer a generous, unlimited version of their service to millions of people, supported by advertising and paid subscriptions.

As a music streaming service provider, KKBOX has members subscribe to their service. When the subscription is about to expire, the user can choose to renew, or cancel the service. They also have the option to auto-renew but can still cancel their membership any time. We want to figure out the reasons users leave and make accordingly changes to retain users and improve competitiveness in the future. In this case, action that no new valid service subscription within 30 days after the current membership expires can be defined as churn. To solve this customer churn problem, **we need to predict whether a user with certain behavior properties will churn after subscriptions expire**. Also, according to our results we can find the reasons why users leave and new insights that can provide recommendations to KKBOX so that they can be proactive in keeping user dancing.

* **RECOMMENDATIONS & RESULTS**



We give KKBOX recommendations from three aspects. **First of all**, KKBOX should cancel auto renew notification option for users and add more diversified membership plans to decrease users’ total cancel number. For example, when a customer starts a new membership plan and system automatically choose auto renew method for him/her without reminder, the user will continue use membership and forget to cancel plans after using for several months. Also, we can encourage users to choose auto renew by giving reward regularly for those who continue paying in auto renew method. **Secondly**, to increase active day users stay in app, KKBOX can promote activities regularly. For example, if a user login in a day, and every minute he/she listens can accumulate 1 point which counts to get some priorities such as listening pop-singer’s latest album in advance. This kind of promotion can stimulate users to stay longer in using app as well as keeping long-term membership. **Finally**, it is important for KKBOX to develop a mature radio station which can recommend songs users would like according to their listening history because this platform conveys key drivers that attract and retain users, and people will find KKBOX is not only a good songs’ player and a songs’ searcher but also a songs’ recommender.

Actually, we give the above recommendations based on the following results which concluded from the churn predictive logistic regression model. **Firstly**, more times a user cancel memberships, more likely that he/she would churn after memberships expire, even if it is possible that the user cancel current membership just to transfer to another plan, which is not considered to be churn. This significant relationship between **cancel\_number** and is\_churn gives us suggestion that we may decrease churn rate by decreasing users’ cancel number. Usually, people cancel the plan because they are not satisfied with it or they just want to try several days but tend to cancel the plan at the beginning, and more planning options can give users more choices, at the same time, KKBOX can cancel auto renew notification option for users so that they will not be reminded to cancel the plan.

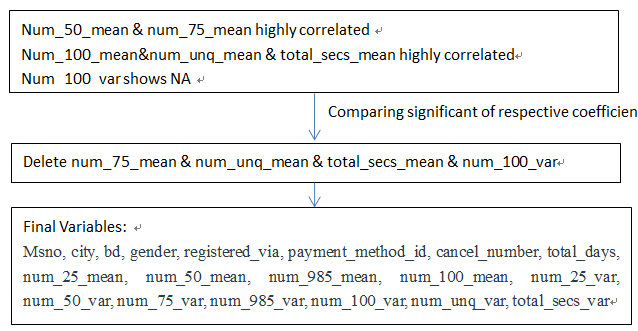
**Secondly**, we find that **total\_days** that users keep membership is highly correlated with is\_churn, which shows longer membership days users keep, less likely for them making a churn. What is more, the result also shows that **autorenew\_number** can significantly explain why users churn because actually higher the autorenew\_number is, more membership total days are. It provides insight that we should motivate users to choose auto renew option to keep their membership plans running. To keep users stay active in plan, KKBOX should make attractive promotions to catch users’ hearts.

**Finally**, users’ **listening behavior variables** show that when people frequently quit listening under the 25% of a song’s length he/she is very likely to churn in the future. Generally speaking, this phenomenon can be explained by unreliable recommendation and searching systems which show the songs the users dismissed. In such a case, it is very necessary for KKBOX to concretize and improve the personalized radio station and searching function to satisfy listeners and decrease the number of songs they discard.

* **Feedback from deliverable 1——Multicollinearity**

When we finally merged all 24 variables into one table, we failed to take multicollinearity into consideration which can greatly impact the accuracy of model and we solved this problem by deleting highly correlated variables.

We think variables from 10 to 16 which are means of number of different songs’ lengths are very likely to linearly correlate to each other because people who usually quit listening halfway also tend to quit on ¾ way and unlikely to completely finish listening a song. Thus we use covariance matrix to help identify highly correlated variables and according to the result we cut out variables whose correlation index is higher than 0.8 and lower p-value in coefficient. In the end, we keep 19 independent variables.



* **MODELLING**

By reviewing relevant literatures, we find that Logistic Regression, Decision Tree, and Random Forest are commonly used in customer churn problem, so we decide to use these models to train our data and select the one that can help us predict customer churn rate of KKBOX most accurately and explore reasons behind it to the most extent. We begin with partitioning our data into two groups, 70% for training and the remaining 30% for testing.

**Decision Tree**

Decision tree is a nonparametric approach for building classification models. It recursively divides data set into two classes by evaluating all possible strategies to partition the data set and selecting the one that maximizes the evaluation metric. Then it forecasts a future event based on the rules discovered from the training data (Luo etc., 2007). We use “rpart” package in R to build decision tree model. We set the minimum leafs to be 15 and specify that the prediction objective is “is\_churn”　and other variables are all counted as predictor variables.

In the first run (confusion matrix is shown in the table below), the total prediction accuracy of the model is high, 92.1% for training group and 92.0% for testing group. But the result also shows that the accuracy of predicting those who churn actually is very low, 11.4% for training group and 6.4% for testing group. Class-1 accuracy is the most important and critical performance indicator because our goal is to predict whether a user will churn after his subscription expires and find solutions to retain users who are going to churn. We have to modify our algorithm or data to improve Class-1 accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training Data** | Prediction | | **Testing Data** | Prediction | |
| Reference | Not Churn (0) | Churn(1) | Reference | Not Churn (0) | Churn(1) |
| Not Churn(0) | 2753 | 13 | Not Churn(0) | 1181 | 4 |
| Churn(1) | 225 | 29 | Churn(1) | 102 | 7 |

**Imbalanced Data**

After examining our data again and reviewing related literatures, we find that imbalanced data is the cause of this problem. The number of churning sample is less than 10% of the total number of samples. The total accuracy is almost determined by Class-0 accuracy, we can fix this problem either by cutting over-represent not-churning sample or by adding under-represent churning sample. To guarantee the data set is big enough to train the decisions. We choose the latter method, duplicating churning samples and adding them to the original data set.

We build another decision tree with the new data set. (Confusion matrix is shown in the table below) In this run, although the total accuracy decreases, it is still acceptable. At the same time, Class-1 accuracy is improved dramatically, 91% for training data and 85% for testing data.

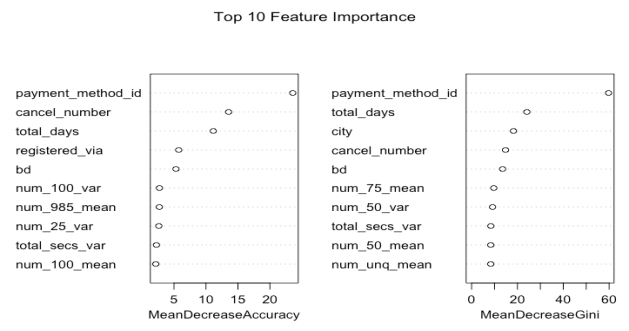
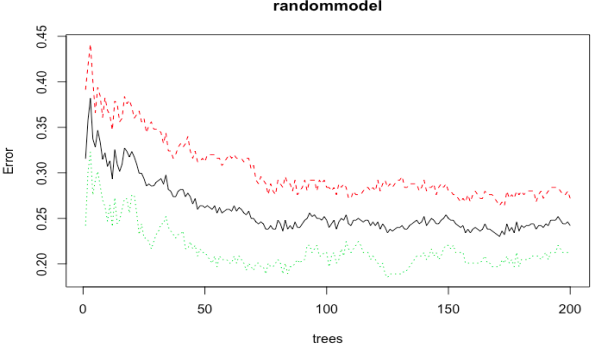
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training Data** | Prediction | | **Testing Data** | Prediction | |
| Reference | Not Churn (0) | Churn(1) | Reference | Not Churn (0) | Churn(1) |
| Not Churn(0) | 1858 | 908 | Not Churn(0) | 792 | 393 |
| Churn(1) | 260 | 2494 | Churn(1) | 16 | 93 |

**Random Forest**

To avoid overfitting problem that might occur in the decision tree model, we apply random forests to our project. Random forest is also widely used in feature selection, so we can fix multicollinearity problem automatically in RFM. A random forest is consisted of a set of decision trees. Each tree is different from each other, so every tree can introduce diversity to the forest. The predicted value is generated by selecting the majority among the predictions of the individual trees (Johansson etc., 2014). We use “randomForest” package in R to build random forests model.

The random forest model gives us a satisfied result. The total accuracies of training data (76.2%) and testing data (72.1%) are closer to each other compared to the result in decision tree, showing that it fix overfitting problem to some extent. Class-0 accuracy (71.2%) and Class-1 accuracy (81.7%) in testing data are lower than the ones in decision tree, but still acceptable. The plot of error with number of trees shows that we are not able to decrease the error rate after about 100 trees, so we set the number of tree in our RFM to be 100.

The variable importance plot below shows that payment\_method\_id, cancel\_number and total\_days are important factors that determine whether a user will churn in the future. This result not only provides us solid evidence in rising up above recommendations but also gives us indications of what features to select out in the multicollinearity situation.



**Logistic Regression**

Both decision tree and random forest perform well in predicting customer churn behavior, solving a part of our problem. We still want to know how these variables affect the churn rate and how to decrease the churn rate. Logistic regression can generate the quantitative relationships between churning probability and each variable, what decision tree and random forest can’t realize, so we apply logistic regression model to our project.

Logistic regression is a regression model used to deal with binary dependent variable, and it cannot be more suitable to predict is\_churn by using logistic regression in this case. By summarizing generated logistic regression we can find out which variable has significant positive or negative coefficient and predict is\_churn by selecting key drivers.

Since we have already deal with multicollinearity problem before modeling and we have got top 10 important variables in decision tree, we finally include 18 variables by combining the two considerations in regression model. In test data set, we got accuracy 76.61% and balanced accuracy 75.14% which show a good fitness of the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training Data** | Prediction | | **Testing Data** | Prediction | |
| Reference | Not Churn (0) | Churn(1) | Reference | Not Churn (0) | Churn(1) |
| Not Churn(0) | 2264 | 438 | Not Churn(0) | 909 | 29 |
| Churn(1) | 496 | 2316 | Churn(1) | 273 | 80 |

**Reference**

[1] Bin, Luo & Peiji, Shao & Juan, Liu. 2007. Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. Proceedings - ICSSSM'07: 2007 International Conference on Service Systems and Service Management. 1 - 5. 10.1109/ICSSSM.2007.4280145.

[2] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. 2014. Regression conformal prediction with random forests. *Mach. Learn.* 97, 1-2 (October 2014), 155-176. DOI=http://dx.doi.org/10.1007/s10994-014-5453-0